UNIVERSITY OF CALIFORNIA
SANTA CRUZ

TRANSIENTLY EXPRESSED LNC-RNAS FINE-TUNE GENE REGULATION DURING PRIMATE NEURAL DIFFERENTIATION

A dissertation submitted in partial satisfaction
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

MOLECULAR, CELL AND DEVELOPMENTAL BIOLOGY

by

Andrew R. Field

December 2017

The Dissertation of Andrew R. Field is approved:

_____
Professor David Haussler, chair

_____
Sofie Salama, Ph.D.

_____
Professor Richard Ed Green

_____
Professor Jeremy Sanford

_____
Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures iv

Abstract vii

Acknowledgements ix

1. Background 1

2. Reprogramming Great Ape Pluripotent Stem Cells 13

3. Identifying and Cataloguing Transiently Expressed lncRNAs During Primate Cortical Neuron Differentiation 32

4. Verifying Gene Regulatory Network Correlation from Single Cell RNA-Sequencing with CRISPR-Activation 54

5. Discussion and Conclusions 77

A. Appendix 83

Supplemental Files 95

Bibliography 97

List of Figures

Figure 1.1: Reprogramming factor delivery methods 6

Figure 2.1: Generalized protocol for Sendai Virus fibroblast reprogramming 15

Figure 2.2: Immunofluorescence staining of Sendai Virus transduced fibroblasts after passage 0 17

Figure 2.3: Early chimpanzee iPSC colonies derived by Sendai Virus reprogramming on MEFs 18

Figure 2.4: Efficiency of Sendai Virus transduction in orangutan fibroblast cell lines 20

Figure 2.5: Initial Sumatran orangutan iPSC colonies 20

Figure 2.6: Immunofluorescence staining of early Jos-3C1 orangutan iPSC colony 21

Figure 2.7: Persistence of Sendai Virus in reprogrammed orangutan iPSCs 23

Figure 2.8: Jos-3C1 confirmed clear of Sendai Virus particles 23

Figure 2.9: Jos-3C1 karyotype analysis 25

Figure 2.10: Establishment of Jos-3C1 on feeder-free conditions 25

Figure 2.11: Jos-3C1 teratoma analysis 25

Figure 2.12: Initial chimpanzee iPSC colonies 27

Figure 2.13: Immunofluorescence of chimpanzee Epi-8919-1A colonies 27

Figure 2.14: Epi-8919-1A transfer to feeder-free conditions 29

Figure 2.15: Epi-8919-1A karyotype analysis 31

Figure 2.16: Epi-8919-1A teratoma analysis 31

Figure 3.1: Cortical neural epithelium differentiation protocol 35

Figure 3.2: Embryoid body formation and differentiation 35

Figure 3.3: Immunofluorescence staining of week 5 neurospheres 37

Figure 3.4: Transcriptomic analysis 39

Figure 3.5: On-target differentiation by cell marker expression 40

Figure 3.6: Expression of canonical cell markers over the time course in each species 42

Figure 3.7: Classification of lncRNA transcript types 42

Figure 3.8: Human-specific TrEx5700 44

Figure 3.9: Human-chimpanzee specific TrEx2174 46

Figure 3.10: Primate conserved TrEx4039 47

Figure 3.11: Transcript structure conservation 48

Figure 3.12: Novel detected Cufflinks transcripts 52

Figure 3.13: Conservation of the TrEx expression pattern 53

Figure 4.1: Human neurosphere single cell RNA-sequencing time course 56

Figure 4.2: Cell type detection in week 2 neurosphere single cell RNA-seq 58

Figure 4.3: Selected TrEx lncRNAs 60

Figure 4.4: TrEx lncRNAs are cell cluster specific 60

Figure 4.5: Persistence of TrEx5008 in immature radial glia cell 61

Figure 4.6: Single cell RNA sequencing gene correlations 64

Figure 4.7: CRISPR activation assay 67

Figure 4.8: CRISPRa of TrEx5008 69

Figure 4.9: CRISPRa of TrEx2819 and TrEx6514 70

Figure 4.10: CRISPRa of TrEx4039 72

Figure 4.11: CRISPRa of TrEx8168 73

Figure 4.12: TrEx8168 in week 2 cortical organoids 75

Figure 4.13: CRISPRa of TrEx108, 2174, and 2578 76

Figure A.1: 783 plasmid map 94

Abstract

Transiently expressed lncRNAs fine-tune gene regulation during primate neural differentiation

Andrew R. Field

The cerebral cortex has undergone rapid changes in size and complexity in the primate lineage, yet the molecular processes underlying primate brain development are poorly understood. I have developed a common protocol for generating cortical organoids from human, chimpanzee, orangutan, and rhesus pluripotent stem cells that recapitulates early events in cortical development and enables comparative molecular analysis of this process. Here I focus on long non-coding RNAs (lncRNAs), which as a class have been implicated in gene regulation, differentiation of pluripotent cells into specific tissues, and can play a role in the fine-tuning of developmental processes. Despite their potential importance in driving the development of tissues, studies focusing on lncRNAs have been impeded by the low sequence conservation and extremely tissue-specific expression patterns of functionally relevant lncRNAs. For this reason, we developed a new approach focusing on the sequence, gene structure, and expression conservation of lncRNAs in equivalent tissues among closely related primate species. To use these aspects of conservation in concert, we collected RNA for high throughput total transcriptome sequencing at weekly time points

during the differentiation protocol and identified thousands of multi-exonic lncRNAs in each species. Of the 2,975 expressed multi-exonic lncRNAs in human, 2,143 were conserved in gene structure to chimpanzee, 1,731 to orangutan, and 1,290 to rhesus. Among these were 386 human transiently expressed (TrEx) lncRNAs that were primarily induced at one time point during differentiation and off by week 5. This pattern was observed to be well conserved in 60-68% of transcripts among great apes but significantly diminished in rhesus macaque with only 39% of human TrEx lncRNAs with conserved structure retaining a transiently expressed pattern in that species. Many of these transiently expressed transcripts were also associated with specific cell subtypes in single cell RNA-sequencing and 8 were found to influence key transcription factors in correlated gene networks by endogenous locus activation via CRISPRa. Identification of lncRNAs expressed during cortical development and this initial functional analysis is a first step towards mechanistic studies that will evaluate the full extent of their importance during neurogenesis, provide insight to primate-specific and human-specific features of cortical development, and give insight to the role of many disparate genetic lesions that contribute to human neurological diseases. This study provides a framework for identifying new potentially functional transcripts by their expression conservation in closely related species.

Acknowledgements

I thank my advisor David Haussler and our wet lab director Sofie Salama for their support through a long and risky, yet ultimately rewarding, project in their lab.

I thank my thesis committee. Ed Green seeded the ideas for what lead to the final experiments of this project. Jeremy Sanford kept our meetings grounded in the scientific questions and offered awkward life advice over a duck sandwich.

There were significant contributions to this work from Frank Jacobs, Ian Fiddes, Sol Katzman, Edwin Jacox, and Rachel Harte. Frank Jacobs began cross-species experiments with human and rhesus macaque neural differentiation and generously provided me with mentorship in cell culture as I began in the lab and RNA-sequencing data from his own experiments to mine for novel non-coding elements. Ian Fiddes tackled gene structure conservation among primates and provided invaluable input and time for our approaches to transcriptomic analysis. Sol Katzman mapped all of the sequencing data and provided the initial differential expression analysis. Finally, Rachel Harte and Edwin Jacox assisted in lncRNA annotation and visualization of the transcriptomic analysis necessary for this work.

I thank the other Haussler lab members for their support, friendship, and mentorship throughout my studies at UCSC. In particular, I thank Adam Ewing for his patience in providing me with mentorship in “dry lab” analysis and David Greenberg for his words of wisdom, sagely advice, and extended metaphors.

I had the great fortune of mentoring and working with a group of extremely talented undergraduate and high school students during my studies at UCSC. Alex Phillips, Andrea Reyes-Ortiz, Vincent Meng, Kacey Fang, Lila Whitehead, Erin LaMontagne, and Jethro Marisagan made it possible to do the sheer amount of cell culture required for this project and, without them, the work with orangutan cells would not exist.

I thank the Susan Carpenter and Sergio Covarrubias for their guidance and providing the materials for CRISPR-activation. Pablo Cordero and Henry Gong assisted with batch effect analysis and general sanity checks for the single cell RNA-sequencing data. Prof. Alex Pollen and Prof. Tom Nowakowski at the University of California, San Francisco provided their invaluable expertise in the interpretations of the cell clusters produced in single cell analysis.

I thank Ollie Ryder and the San Diego Frozen Zoo for providing primate fibroblasts for our reprogramming experiments and Josephine for being our unwitting orangutan donor. Robert Diaz and Karen Shaff established our chimpanzee pluripotent stem cell lines.

I thank Ben Abrams for his maintenance and training in microscopy and Bari Nazario for facility support and the continual stream of baked goods.

I thank Melanie Day, Greg Roe, Krishna Roshkin, Jonathan Casper, Jeltje van Baren, Yarry Gonzalez, Cricket Sloan, Olena Morozova, Charlie Vaske, Karen Miga, and Ed Miga for friendship and support through the hardest times. You helped more than you could know.

This work was supported in part by Howard Hughes Medical Institute, CIRM predoctoral Fellowship T3-00006, StemPath NIH/NIGMS R01 GM109031, and UCSC TA-ships.

Chapter 1

Background

1. Background

1A. Human brain evolution

The human brain has grown remarkably in size and complexity even compared to our closest living evolutionary cousins the great apes. In particular, the size of the cerebral cortex has doubled since our divergence with chimpanzee 4.5 to 6 million years ago (Locke et al., 2011; Hill and Walsh 2005; DeFelipe, 2011). The mechanisms that enacted this drastic phenotypic change are still largely unknown. The protein-coding genes that define tissues are highly conserved and virtually unchanged among mammalian species (Locke et al., 2011). Neural tissues in particular seem to evolve at a significantly slower rate than other organs, suggesting that the changes in size, structure, and cellular composition of the brain observed in the hominid lineage represents an extreme outlier in what is otherwise an incredibly fine-tuned expression network (Brawand, 2011).

With that, the changes seen in the primate lineage are rather remarkable among mammals. It has been known for some time that a sustained period of neurogenesis in primates, specifically in the upper cortical layers, has likely

contributed to human and primate specific attributes in cognition and mental function (Marin-Padilla et al., 1992). Cortical progenitors, which undergo a total of 11 divisions in mouse fetal development (Takahashi et al., 1995), skyrocket to at least 28 in rhesus macaque (Kornack et al., 1998) and presumably many more in human. This increased cortical growth is theorized to allow for the sub-specialization of regions of the cortex (Orban et al., 2004; Semendeferi et al., 2002). This functional asymmetry is thought to allow for human-specific attributes including handedness (Hopkins et al., 2004) and frontal temporal specification for language (Cantalupo and Hopkins, 2001).

One theory at the center of this increased brain size and functional capacity is the radial unit hypothesis (Noctor et al., 2001). It states that a small change in the number of radial glia in the developing brain, the angle at which they radiate outwards while dividing, the radii of their expansion, or the time in which they continue to divide can have profound effects on the resulting size of the cortex. All of these changes are likely to have their beginnings in the timing and mechanisms of gene regulation in the developing fetal brain (King and Wilson 1975; Sun et al., 2005) during which a small change can lead to drastic phenotypic difference in adult brains between species (Enard et al., 2002; Caceres et al., 2003; Uddin et al., 2004) though the mechanism by which these changes are enacted are still unknown.

1B. Pluripotent stem cells

Studying cortical brain development *in vivo* is difficult if not impossible in human and endangered species in fetal tissue which preclude the use of many cell biology and molecular biology techniques. Luckily, recent advances in the use of pluripotent stem cells have allowed researchers to recapitulate many of the early processes of developing tissue *in vitro*.

A stem cell is broadly defined as a cell capable of self-replicating and producing a daughter cell of a different cell identity in a process called differentiation. Embryonic stem cells (ESCs) represent a special case in which a small population of short lived cells normally restrained to the inner cell mass of a blastocyst are capable of differentiating into all three germ layers of an adult organism: endoderm, mesoderm, and ectoderm (Tuch et al., 2006). This state, referred to as pluripotency, is normally fleeting *in vivo*, but if provided with the appropriate growth substrate and media, this cell state can be maintained indefinitely. By removing this maintaining factor and adding chemical inhibitors, the cells can be influenced to differentiate into specific cell lineages in a fashion that recapitulates many of the features of *in vivo* fetal development.

1C. Generating induced pluripotent stem cells

In the case of endangered species such as chimpanzee and orangutan, embryonic tissue is unavailable. For this purpose, researchers have turned to generating induced pluripotent stem cells (iPSCs) which can be generated from somatic tissues that do not require the destruction of an embryo or the sacrifice of an animal's life. Takahashi and Yamanaka demonstrated in 2006 that mouse fibroblasts can be reprogrammed into a pluripotent state similar to ESCs through the over-expression of four transcription factors, OCT3/4, SOX2, c-MYC, and KLF4, termed Yamanaka Factors, via viral vectors. Cells produced in this way were capable of differentiating into the three germ layers *in vitro* and reconstitute all tissues in mouse through a chimera assay. Later, two groups demonstrated that this reprogramming process can be carried out in human cells (Takahashi et al., 2007; Yu et al., 2007).

Methods for generating iPSCs largely differ in the vehicle used to introduce the Yamanaka Factors. Early studies utilized integrating vectors such as retrovirus (Takahashi et al., 2007; Yu et al., 2007), which could potentially lead to undesired genetic aberrations in the cell genome affecting gene regulation, genome stability, or decreased differentiation efficiency by continued production of the viral genes. For this reason, non-genomically integrating techniques were developed where reprogramming factors are introduced transiently and removed upon further sub-culturing (Ban et al., 2011). Two

popular strategies revolve around episomal transfection of lentiviral vectors (Okita et al., 20011) and transduction by the single-stranded RNA virus, Sendai Virus (Fusaki et al., 2009), as depicted in Figure 1.1. Episomal reprogramming utilizes expression plasmids that are nucleofected into the cells. Since these plasmids do integrate into the host genome, they eventually dilute out of the cell population through expansion of the stem cell line. Sendai Virus reprogramming uses an RNA virus whose entire life cycle occurs within the cytoplasm. These viruses are mutated such that they lack the ability to replicate, but still produce capsid proteins for easy detection by immunofluorescence (IF) and reverse-transcriptase polymerase chain reaction (RT-PCR). These viral elements will also dilute over time and eventually be removed from the cells over time.

Even though most iPSC lines seem to behave similarly to ESCs in most respects, there has been some scrutiny of their true nature. Concerns have been raised of whether the cells retain a "memory" of their source somatic cell type (Ji et al., 2012). There is also concern that the intensely selective nature of sub-culturing clones of iPSC lines could lead to artificial selection of non-biological cell types that perform best in artificial culturing conditions or accumulate mutations due to culturing stress (Ji et al., 2012). For these reasons, iPSC lines are subjected to numerous tests including karyotyping, teratoma assay and *in vitro* undirected differentiation to confirm their efficacy as pluripotent stem cells.

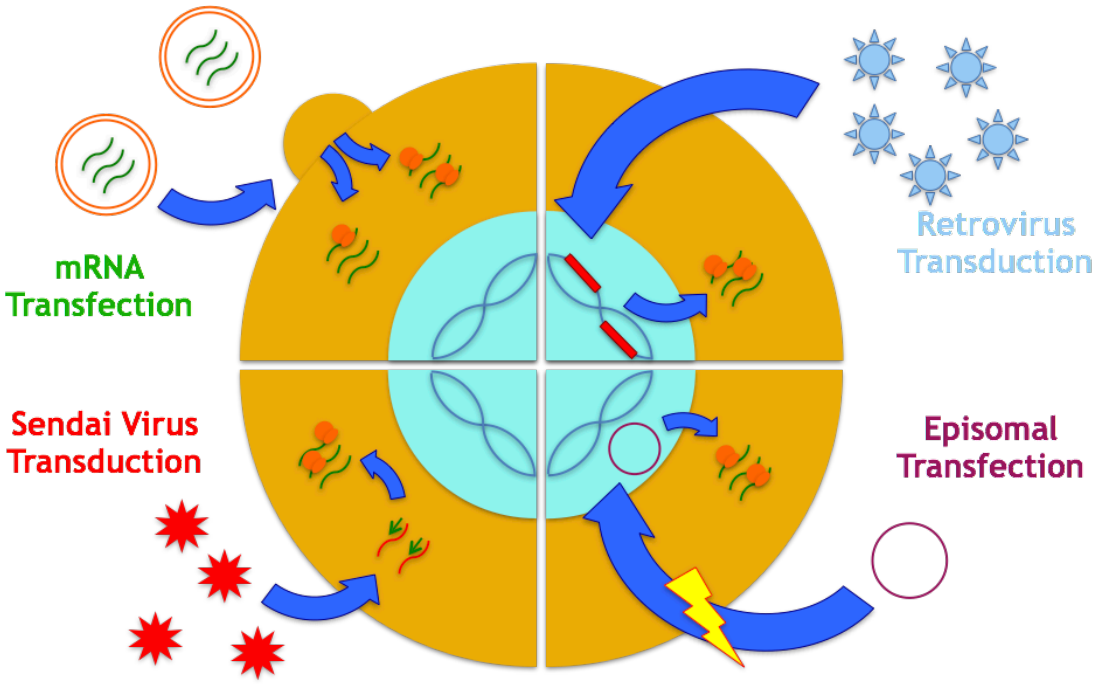

Figure 1.1: Reprogramming factor delivery methods. This schematic of a cell depicts four popular strategies for introducing reprogramming factors to somatic cells for iPSC generation. The original method utilizing retrovirus involved integrating expression cassettes into the cell genome which could lead to undesirable genomic aberrations. Episomal transfection involves introducing expression plasmids by nucleofection. Sendai Virus is an RNA virus that undergoes its replication in the cytosol and never enters the nucleus, allowing for eventual dilution of the viral factors. mRNA transfection involves directly introducing reprogramming gene messenger RNA into the cytoplasm through liposomes. The factors are continually added over the period of a week or more until the endogenous gene regulatory networks can sustain themselves.

1D. Using pluripotent stem cell differentiation as a model of tissue development

The self-organizing capacity of primary mammalian cells has long been known to be maintained in cell culture with the ability to form complex 3-dimensional structures called organoids that recapitulate their organization and functions observed within tissues. Perhaps the most characterized of these organoid systems is that of developing epithelial tissue which form highly structured layers of stem, progenitor, and differentiated cell types (Sato et al., 2009) which has been successfully adapted to form models of intestine, colon, stomach, and liver (Sato et al., 2011; Dekkers et al., 2013; Barker et al., 2010; Stange et al., 2013; Huch et al., 2013).

Organoid culturing techniques have also been applied to stem cell differentiation assays to co-culture multiple cell types simultaneously that more closely recapitulate early developing tissues. Ectoderm lineage organoids, in particular, have been successfully generated from spherical aggregates of pluripotent stem cells called embryoid bodies (EBs). After these EBs are formed from either aggregating single cell suspensions or self-formed after manual lifting of colonies, cocktails of chemical inhibitors and signaling proteins are introduced to guide the differentiation of these cells to a desired neural lineage. Using this method, groups have been able to generate retinal tissue that form optical cups, telencephalon directed organoids that recapitulate the inside-out

layering of the cerebral cortex, and multiple brain regions within one organoid using spinning bioreactors (Eiraku et al., 2011; Mariani et al., 2012; Lancaster et al., 2013). All these organoid methods produce highly structured layered tissue recapitulating early neural development including proliferative zones of radial glia, projecting intermediate progenitors, and functional early neurons capable of excitation in stratified layers over the course of 5 to 50 weeks of culturing (Eiraku et al., 2011; Mariani et al., 2012; Lancaster et al., 2013). These studies have opened the way for directly comparing disease states and multiple species in a parallelizable and controlled fashion.

1E. Long non-coding RNAs in tissue specification and development

Devastating diseases associated with human neural development, including autism spectrum disorders, schizophrenia, and bipolar disorder, have SNP markers that lie outside protein-coding gene bodies (Barnet and Smoller, 2009; O'Donovan et al., 2009; Williams et al., 2009). And despite drastic phenotypic differences, protein-coding sequences remain nearly identical in mammalian species (Britten, 2002; Chimpanzee Sequencing and Analysis Consortium, 2005; Mouse Genome Sequencing Consortium, 2002; Ruvolo, 1997; Wildman et al., 2003). It is clear from these findings that there are significant non-protein coding elements involved in central gene regulatory events during

brain development. A detailed molecular understanding of these elements and their impact on human cortical neuron development and function will be invaluable for understanding evolutionary differences between primate species as well as the many disparate genetic lesions that contribute to human diseases. Furthermore, this information could enable the generation of specific subpopulations of neurons for regenerative therapies.

Where protein coding genes only account for ~1.2% of the human genome (International Human Genome Sequencing Consortium 2004), RNA sequencing studies have estimated that up to 90% of mammalian genomes are actively transcribed with about 98% of the cell's transcriptional output at any given time being non-protein-coding (Bertone et al., 2004; J. Cheng et al., 2005; ENCODE Project Consortium 2007; Kapranov et al., 2007; Mattick 2001). A significant portion of this transcription has been ascribed to the still largely mysterious class of long non-coding RNA (lncRNA). LncRNAs have been defined as transcripts longer than 200nt that are transcribed from genomic regions outside of protein-coding genes. These transcripts can also be spliced and poly-adenylated (Mattick 2004; Ponting, Oliver, & Reik 2009). Further, lncRNAs offer a potential target for evolutionary studies because while they are less conserved than protein-coding genes they are more evolutionarily constrained than would be expected for neutral sequence (Khalil et al., 2009; Mu et al., 2011; Ørom et al., 2010b). This supports a potential mechanism for more rapid adaptability between species than protein sequences while still implying functional

importance. Studies have implied roles for lncRNAs in maintenance of pluripotency (Guttman et al., 2009; Huarte et al., 2010; Hung et al., 2011; Loewer et al., 2010) by regulation of protein gene expression and chromatin remodeling (Khalil et al., 2009; Nagano et al., 2008; Pandey et al., 2008; Rinn et al., 2007). In addition, their expression is highly cell type specific, even more so than protein coding genes (Cabili et al., 2011), which suggests importance in cell fate specification.

Here I focus on the degree of conservation of lncRNAs which often show tissue specific expression (Pontig et al., 2009; Cabili et al., 2011; Derrien et al., 2012; Pauli et al., 2012), account for a significant proportion of Pol II output (Carninci et al., 2005; Harrow et al., 2006; Derrien et al., 2012), and show particular enrichment in neural tissues (Ravasi et al., 2006; Cabili et al., 2011; Derrien et al., 2012; Ramos et al., 2013). LncRNAs have been shown to have diverse roles in gene regulation including chromosome inactivation (Penny et al., 1996; Zhao et al., 2008), imprinting (Lighton et al., 1995; Camprubi et al., 2006; Buiting et al., 2007; Pandey et al., 2008; Martins-Taylor et al., 2014), and developmental processes (Rinn et al., 2007; Heo and Sung, 2011) and many more have been implied in establishment of pluripotency (Guttman et al., 2009; Guttman et al., 2011), stem cell maintenance (Rani et al., 2016), reprogramming (Loewer et al., 2010), and differentiation (Guttman et al., 2011). But still, most of the tens of thousands of identified lncRNAs in human have undetermined function (Hon et al., 2017; Lagarde et al., 2017) and lack sequence conservation

among vertebrate species (Wang et al., 2004; Church et al., 2009; Cabili et al., 2011; Ulitsky et al., 2011; Kutter et al., 2012). Their tissue specific expression patterns, low level of sequence conservation, and implication in cell type specification make lncRNAs an attractive target for species-specific gene regulation during development.

Despite their potential importance in driving the development of tissues, functional studies have been impeded by low sequence conservation and extremely tissue-specific expression patterns of individual lncRNAs. For this reason, I utilized an approach focusing on both the gene structure and expression conservation of lncRNAs in equivalent developing tissues among closely related primate species to identify potential human regulatory elements. An adapted protocol for cortical neuron generation from human, chimpanzee, orangutan, and rhesus pluripotent stem cells was used to recapitulate early events in cortical development and enable us to do comparative molecular analysis of this process. Bulk strand-specific total-transcriptome RNA-sequencing was performed on weekly time points to assess lncRNA expression conservation and retention of intron boundaries among primates. Particular attention was paid toward lncRNA transcripts that were transiently expressed (TrEx lncRNAs), having max expression during the time course and were diminished by week 5, then looking for retention of this expression pattern across species. Single Cell RNA-sequencing on a subset of time points relevant to major differentiation events was used in human to identify the cell

subpopulations associated with the expression of candidate TrEx lncRNAs. Finally, CRISPR activation (CRISPRa) in HEK293FT cells was used to express these transcripts out of context as an initial step to probe gene regulatory function between transiently expressed lncRNAs and their associated gene networks from single cell data. In all, I identified 8 TrEx lncRNAs whose expression had significant effects on protein-coding genes associated with the transcripts in single cell data.

Chapter 2

Reprogramming Great Ape Pluripotent Stem Cells

2. Reprogramming Great Ape Pluripotent Stem Cells

2A. Introduction

Human, chimpanzees, and orangutans share much of their genomic sequence and genomic analysis alone is insufficient to understand the source of the phenotypic differences between the species (Carroll et al., 2008; Locke et al., 2011). It has long been supposed that gene regulation during development is at the core of human and great ape-specific traits (King and Wilson, 1975). Our current understanding of molecular phenotypes in these species is largely limited to post-mortem tissues which are difficult to study in parallel, are rare in relevant developmental stages, and do not allow for cell biology techniques for perturbation. While human embryonic stem cells (ESCs) have allowed for a backdoor into the molecular underpinnings of human development, it remains illegal to destroy embryos of endangered species as would be necessary to create such lines from chimpanzees and orangutans. For this reason, I sought to create induced pluripotent stem cell (iPSC) lines from these animals to subject to neural differentiation and allow a window into the molecular landscape of great ape tissue development.

2B. Sendai Virus Reprogramming

Three female chimpanzee fibroblast lines, S003647, S008919, and S008933 (Yerkes Primate Center), two female Sumatran orangutan fibroblast lines, PR01110 (Coriell) and 11045-4593 "Josephine" (San Diego Frozen Zoo®), and one female Bornean orangutan fibroblast line, 13692-6363 "Chelsea" (San Diego Frozen Zoo®), were subjected to Sendai Virus reprogramming. The culturing conditions for these experiments is outlined in Figure 2.1.

Fibroblasts were plated at day -2 at a density such that they would reach 50-80% confluence on day 0. Four Sendai Virus vectors containing Oct3/4, Sox2, Klf4, and c-Myc (CytoTune, ThermoFisher) respectively were introduced at varying titers on day 0 and media was changed to fresh fibroblast media on Day 1. Sendai Virus efficiency was evaluated by immunofluorescence imaging (IF) during this time. At day 7, the transduced cells were passaged at low density to mouse embryonic fibroblast (MEF) feeder cells and allowed to grow for up to 21 days in iPSC media. During this time, the plates were constantly monitored for colony growth. If a colony was spotted, it was manually picked, passaged, and clonally expanded on MEFs. The primary variables that were changed in each iteration of this protocol was the multiplicity of infection (MOI), or the concentration of virus used, and the plating density onto MEFs at day 7.

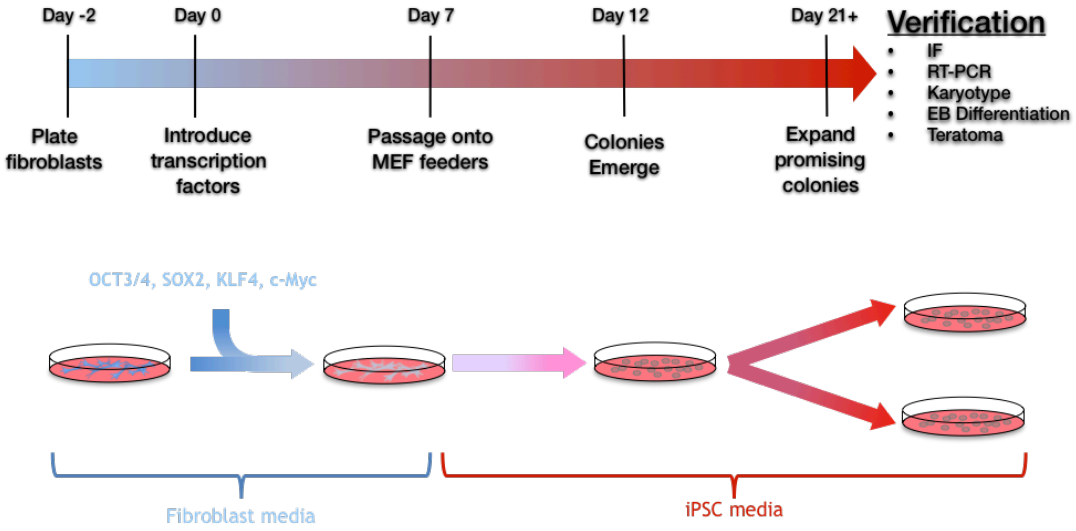

Figure 2.1: Generalized protocol for Sendai Virus fibroblast reprogramming. Fibroblasts are plated at a low density such that they achieve 50%-80% confluence in two days (Day -2). Sendai Virus was introduced at varying titers on Day 0 and the media was changed to fresh fibroblast media on Day 1. At Day 7, transfected fibroblasts are passaged at low density on a mouse embryonic fibroblast (MEF) feeder layer. Colonies can emerge as early as Day 21, but plates are kept growing in iPSC growth media conditions for up to 30 days. Promising colonies are selected from each plate and expanded.

2D. Chimpanzee Sendai Virus Reprogramming Attempts

Sendai Virus reprogramming was attempted on three female chimpanzee fibroblast cell lines from the Yerkes Primate Center, S003647, S008919, and S008933. Human "HuSk" fibroblasts were used as a control as I had previously efficiently reprogrammed the cell line. Reprogramming was attempted a total of four times, but here I will describe the most successful attempt.

All four virus vectors were added at 3 MOI. Media was changed on day 1 then every other day until passage at day 7 when they were passaged onto MEFs at a density of 100 thousand to 400 thousand per well of a 6-well dish with iPSC media supplemented with β-FGF. RNA was harvested from transduced fibroblasts at this stage for later RT-PCR analysis. Additional 24-well plates were seeded with transduced fibroblasts for immunofluorescence imaging (IF) at day 9. Clusters of cells expressing OCT3/4 indicated potentially successful reprogramming and initial colony growth (Figure 2.2). Media was changed daily and cultures were monitored for the appearance of colonies. Colonies began appearing at day 12 and were selected through day 28 (Figure 2.3). The control HuSk fibroblast line generated the most reprogrammed colonies. S008933 showed the best colony formation efficiency among the chimpanzee cell lines. Around this time, the episomal method, which was used in parallel, yielded many more colonies for both S008919 and S008933 chimpanzee lines, so colonies derived from Sendai Virus were frozen into cell stocks and not expanded further.

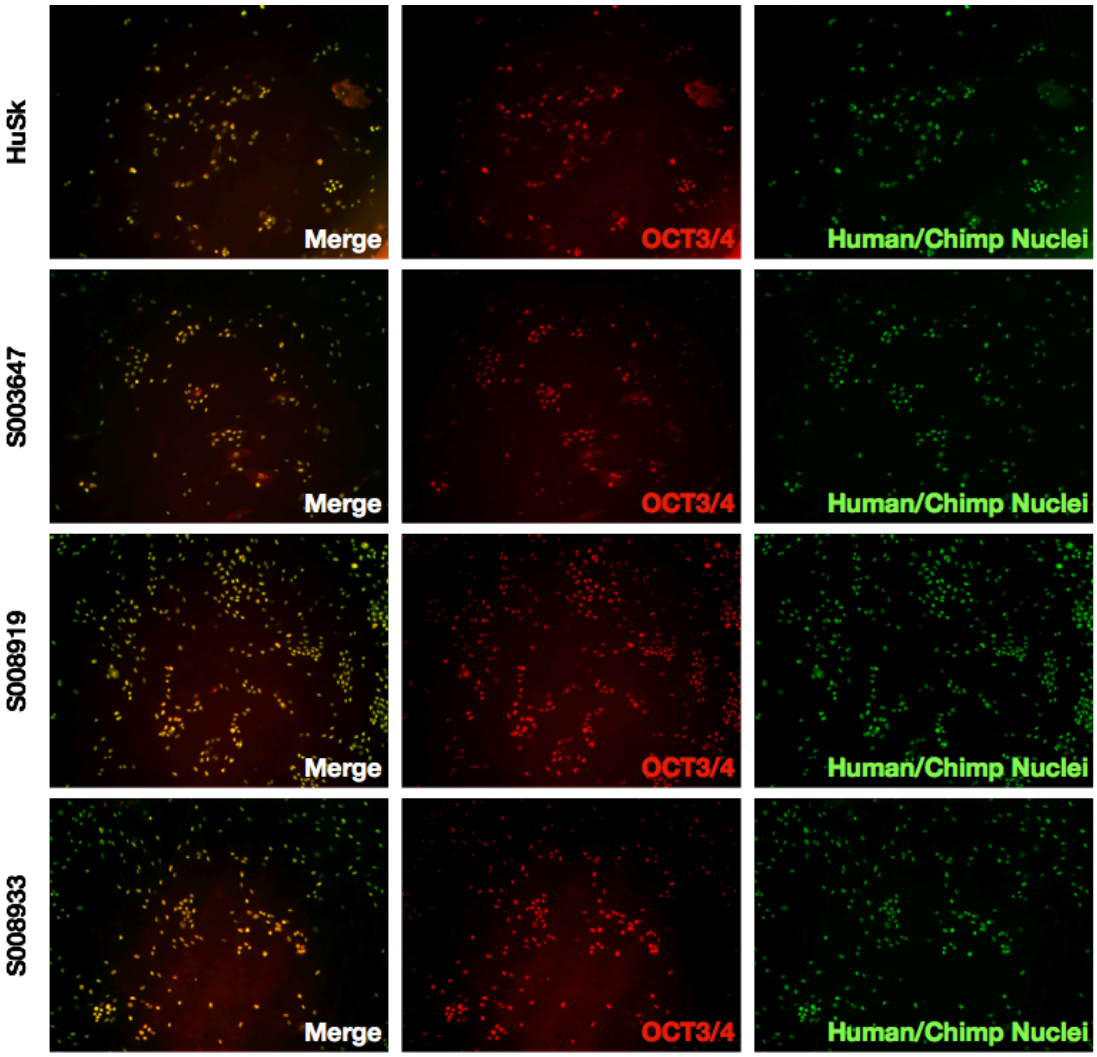

Figure 2.2: Immunofluorescence staining of Sendai Virus transduced fibroblasts after passage 0. Transduced fibroblasts from each cell line were seeded onto MEF feeders and stained for OCT3/4 and a great ape-specific nuclei marker to distinguish them from the feeder layer. Tight clusters of OCT3/4 positive cells were early indications of colony growth and successful reprogramming.

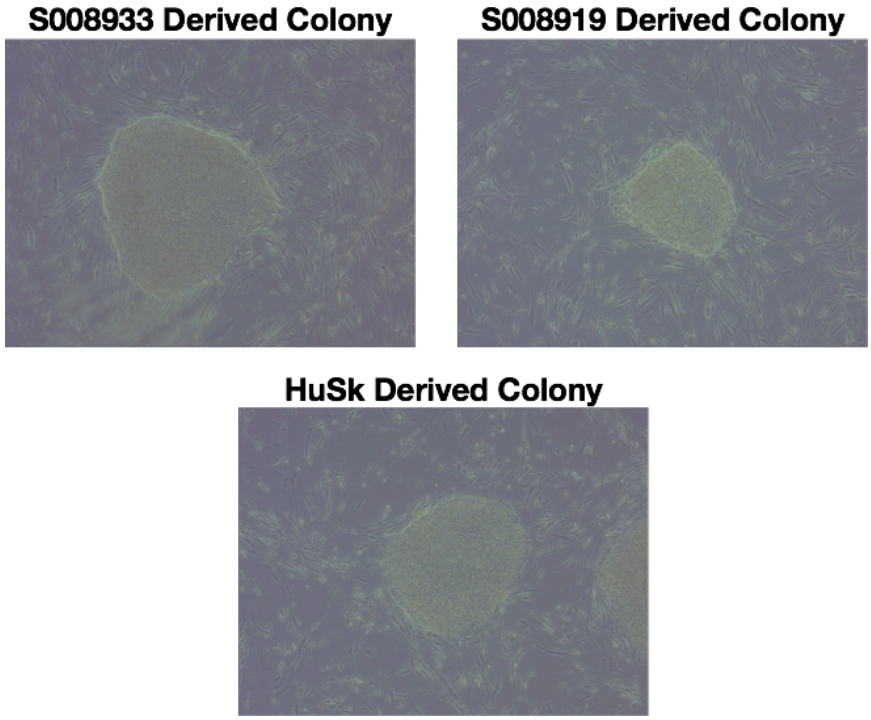

Figure 2.3: Early chimpanzee iPSC colonies derived by Sendai Virus reprogramming on MEFs. These are brightfield images of the first colonies that arose from each successfully reprogrammed cell line used in these initial experiments. They exhibited the canonical dense "cobblestone" cell morphology with phase-bright edges to the colonies. They were all picked and grown on 6cm plates prior to freezing cell stocks.

2E. Orangutan Sendai Virus Reprogramming

Karyotipically stable orangutan fibroblast cell lines were selected from those obtained from the San Diego Frozen Zoo® for Sendai Virus reprogramming. 11045-4593 "Josephine" and 13692-6363 "Chelsea" were chosen and tested for their efficiency of transduction by Sendai Virus using the green fluorescent protein (GFP) expression vector CytoTune-EmGFP (ThermoFisher). Maximal expression of GFP was found using an MOI of 6 for both cell lines (Figure 2.4).

The CytoTune 2.0 Sendai Reprogramming kit (ThermoFisher) was used to reprogram the two Sumatran orangutan fibroblast lines. This kit differed from our initial attempts with chimpanzee as it featured three viral vectors: a KOS vector containing a polycistronic mRNA containing KLF4, OCT3/4, and SOX2, a separate c-MYC vector, and an additional KLF4 vector. In addition, the c-MYC and KOS vectors are heat sensitive to aid with removal from reprogrammed cells.

Both 11045-4593 and 13692-6363 fibroblast cells were transfected at 6 MOI of each virus at about 50-80% confluency. After re-plating on MEFs, four promising colonies appeared derived from the 11045-4593 cell line called Jos-3C1, Jos-2A3, Jos-2A4, and Jos-1B4. Efforts in this study were focused on Jos-3C1 (Figure 2.5). After initial colony expansion, samples were collected for immunofluorescence staining (Figure 2.6). Colonies from Jos-3C1 showed

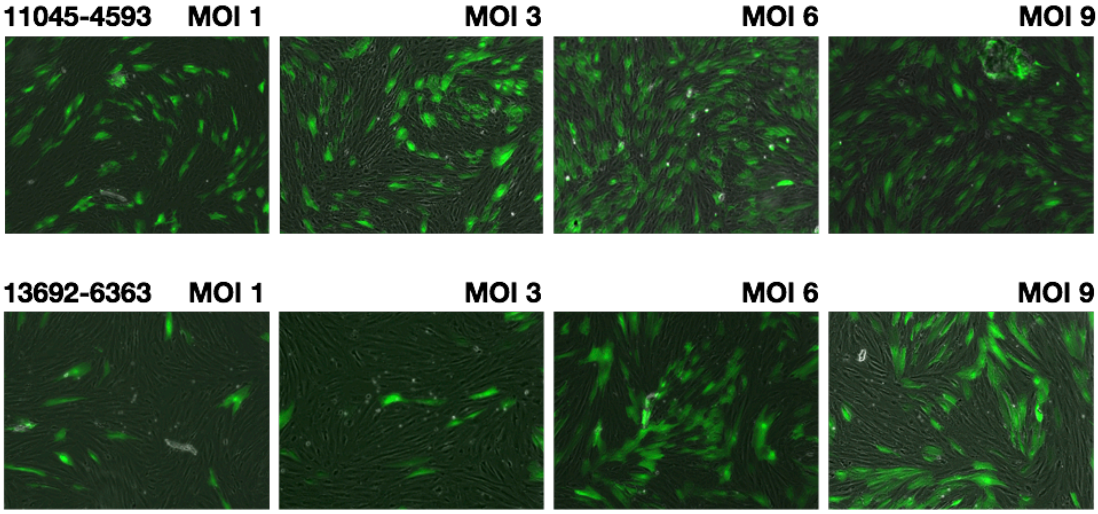

Figure 2.4: Efficiency of Sendai Virus transduction in orangutan fibroblast cell lines. Cells were transfected with MOIs of 1, 3, 6, and 9 with CytoTune-EmGFP (ThermoFisher) and imaged after 48 hours with IF spectroscopy. Maximal efficiency was observed at MOI 6 in both cell lines.

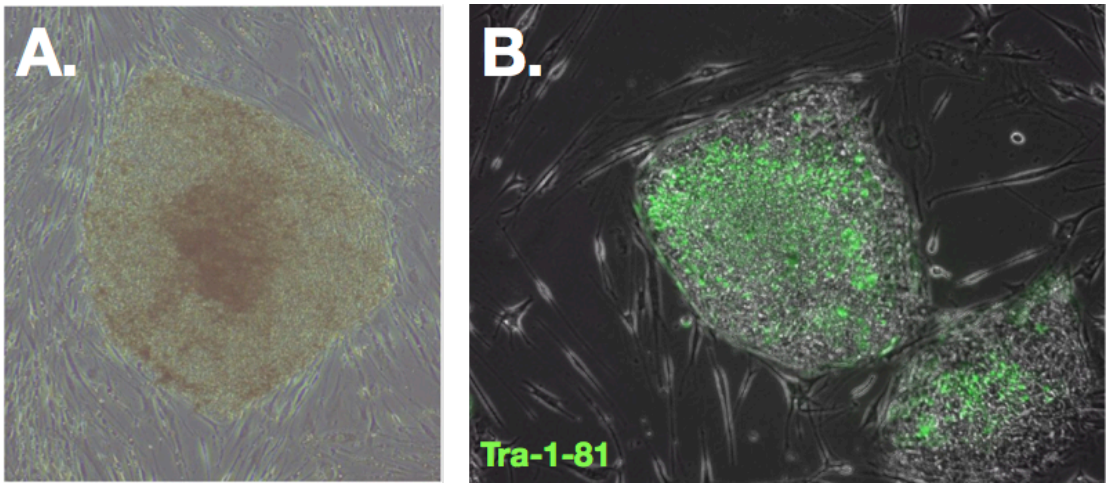

Figure 2.5: Initial Sumatran orangutan iPSC colonies. (A) Initial Jos-3C1 colony picked from passage 0 plate. (B) Passage 1 of Jos-3C1 live-stained for Tra-1-81.

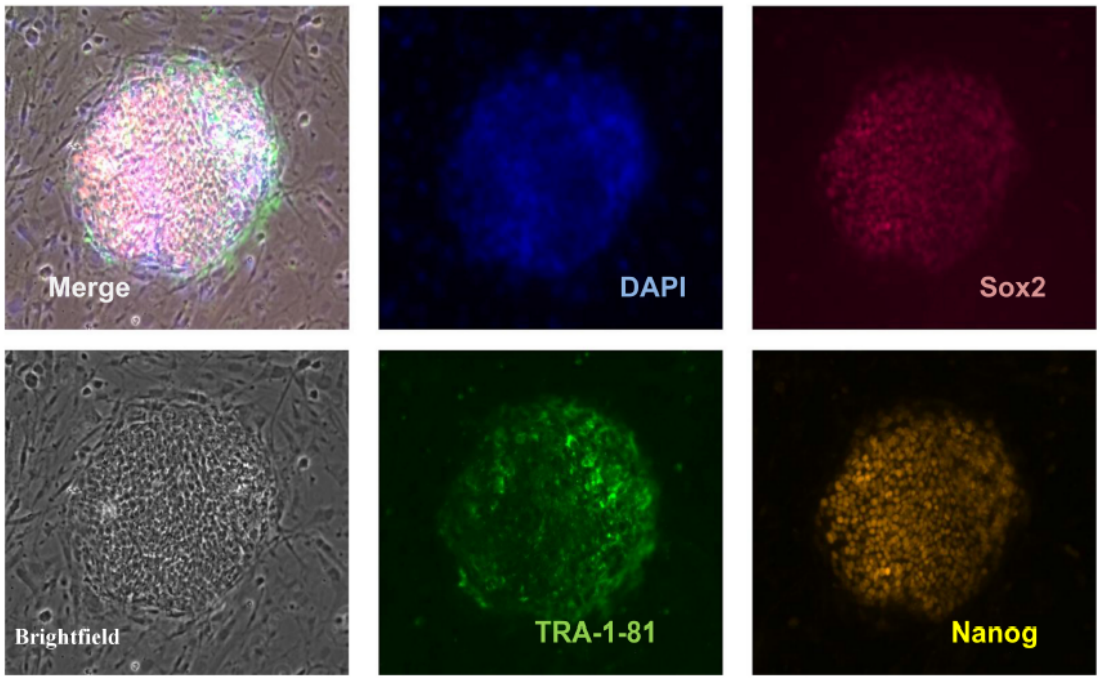

Figure 2.6: Immunofluorescence staining of early Jos-3C1 orangutan iPSC colony. Staining of this early colonly on MEFs shows homogeneous staining for SOX2 as well as the endogenously expressed Tra-1-81 and Nanog indicating complete reprogramming.

homogeneous expression of the exogenously supplied gene SOX2 as well as endogenous Tra-1-81 and Nanog throughout each colony indicating complete reprogramming.

Next, to confirm the removal of the Sendai Virus genome from the cell lines, RT-PCR was performed RT-PCR on total RNA collected from the iPSC clones. Surprisingly, most were found to have retained Sendai Virus gRNA well beyond 10 passages (Figure 2.7A). In particular, 2A3 clone retained the c-Myc vector and 3C1 which cleared detectable traces of all of the vectors. In an attempt to rescue some of the 2A3 clone by removing the recalcitrant vector, they were heat treated for a full passage at 39°C to disrupt the heat sensitive elements incorporated into the KOS and c-Myc vectors, though those attempts were unsuccessful (Figure 2.7B). For this reason, I focused on Jos-3C1 which naturally cleared the viral vectors. Complete removal of Sendai Viral particles was confirmed by immunofluorescence staining of Sendai coat protein in Jos-3C1 colonies (Figure 2.8).

Long periods of *in vitro* culturing of cell lines carries the inherent risk of propagating genomic damage or genetic rearrangements that are advantageous to cell growth. During subsequent passaging, cell culturing conditions could select for unnatural genetic variants that will overwhelm the culture. As a matter of course in establishing cell lines, karyotype analysis was used to ensure genomic stability. Jos-3C1 was confirmed to retain the same karyotype at

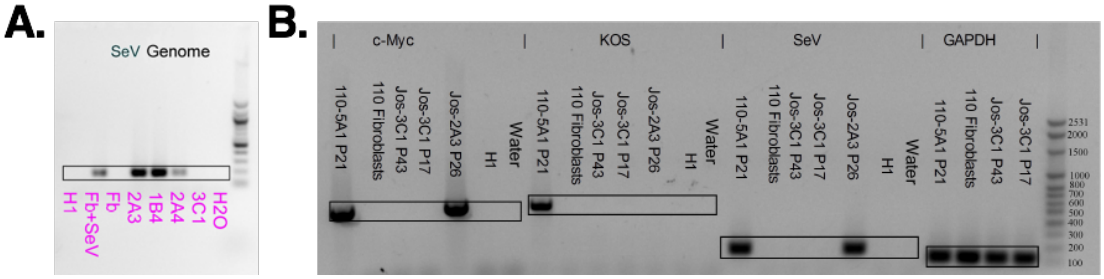

Figure 2.7: Persistence of Sendai Virus in reprogrammed orangutan iPSCs. (A) RT-PCR for Sendai Virus genomic RNA of early passages of 4 orangutan iPSC clones 2A3, 1B4, 2A4, and 3C1 compared to H1 human embryonic stem cells (negative control), orangutan fibroblasts (Fb, negative control), and orangutan fibroblasts transduced with Sendai Virus (Fb+SeV, passage 0, positive control). Sendai Virus genome appears retained in 2A3, 1B4, and 2A4 but not 3C1. (B) A later passage of both 3C1 and 2A3 after heat treatment was tested by RT-PCR for Sendai Virus genome with primers specific to the c-Myc and KOS viral sequence in addition to the general SeV primer set that detects all 3 used viruses. 110-5A1 represents an early sample from transduced fibroblasts (positive control), 110 fibroblasts are prior to transduction (negative control), Jos-3C1 is represented at early and late passage, and Jos-2A3 is shown after heat treatment. Jos-2A3 shows retention of the c-Myc Sendai Virus genome.

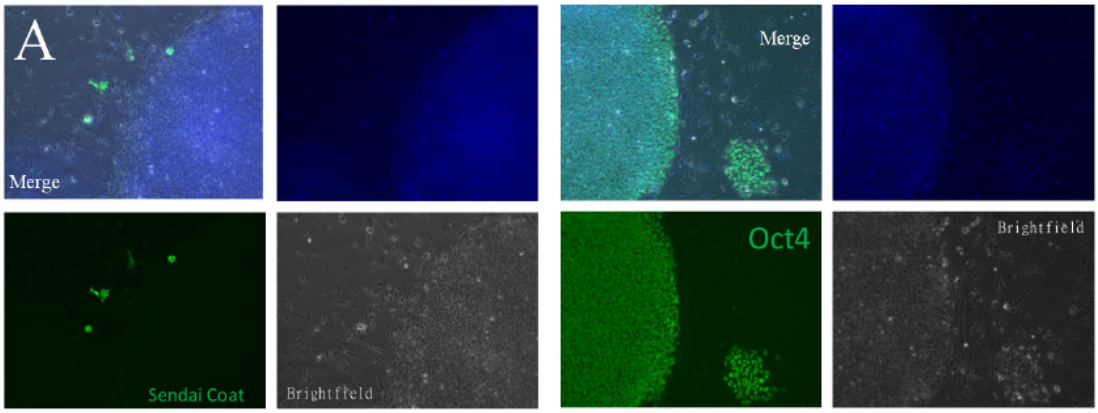

Figure 2.8: Jos-3C1 confirmed clear of Sendai Virus particles. Immunofluorescence staining for Sendai Virus coat protein shows a lack of expression in Jos-3C1 colonies in panel A. The fluorescent signal observed is likely auto-fluorescence from MEFs. OCT4 staining of colony in the same passage shows maintenance of homogenous expression throughout iPSC colonies.

passage 36 as the source fibroblasts ensuring that no gross rearrangements had accrued over the course of differentiation (Figure 2.9). A pericentric inversion in chromosome 10 was observed in both the starting fibroblasts and Jos-3C1 which has been reported as naturally occurring within wild populations of Sumatran orangutans (Locke et al., 2011).

In order to better perform in the cortical organoid differentiation protocol, Jos-3C1 was adapted to feeder-free conditions on vitronectin (ThermoFisher) and E8-Flex (ThermoFisher) (Figure 2.10). Though colonies appeared morphologically different than their feeder culture counterparts, colonies maintained homogeneous expression of pluripotency factors (Figure 2.10).

These feeder-free cultures were used as input for a teratoma assay to confirm their differentiation potential. Mice were anesthetized by intraperitoneal injection with 100mg/kg ketamine. 4 subcutaneous injections of 1 million cells suspended in 30% Matrigel (Corning) were made in the ventral lateral areas of NOD-SCID mice (NOD.CB17-Prkdc^{scid}/NCrCrl, the Jackson Laboratory) similar to Prokhorova et al., 2009. Mice were observed for 9 weeks for the appearance of tumors in the injected areas. The animals were euthanized by cervical dislocation and teratomas were harvested, fixed in 4% paraformaldehyde, saturated in 30% sucrose in PBS, embedded in Tissue Freezing Medium™ (Triangle Biomedical Sciences), and frozen for cryostat sectioning. Sections of the tumors were stained with hematoxylin

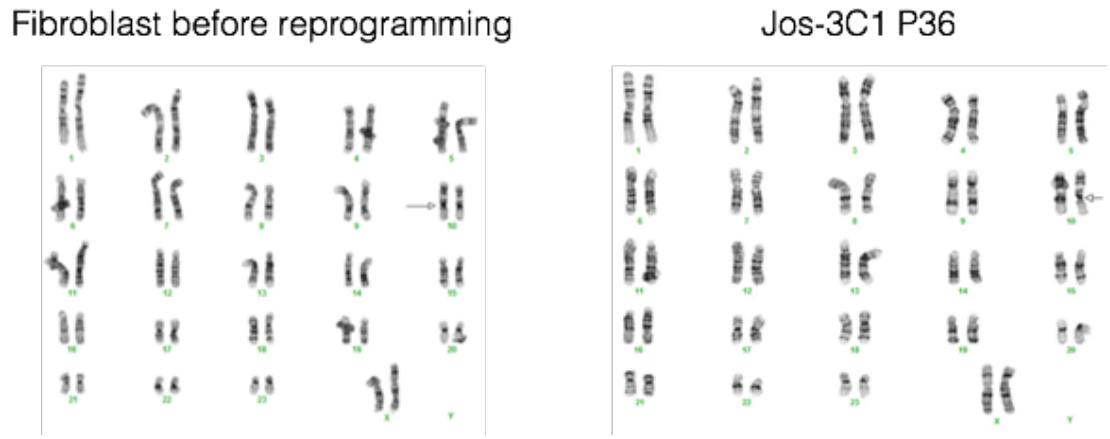

Figure 2.9: Jos-3C1 karyotype analysis. Wildtype 48, XX karyotype was confirmed in orangutan iPSCs at passage 36 after reprogramming. An inversion in chromosome 10 is naturally occurring in the wild Sumatran orangutan population (Locke et al., 2011) and was present in the original fibroblasts.

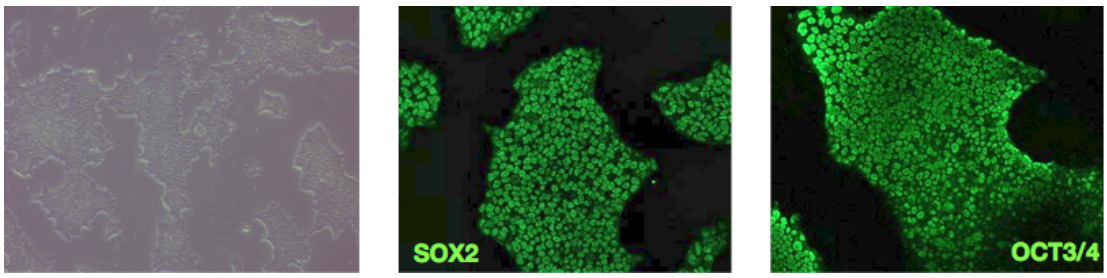

Figure 2.10: Establishment of Jos-3C1 on feeder-free conditions. Jos-3C1 was established in feeder free conditions on vitronectin (ThermoFisher) and E8-Flex media (ThermoFisher) shown in brightfield in the left panel. Maintenance of pluripotency in these new conditions was tested by immunofluorescence staining of SOX2 (middle) and OCT3/4 (right).

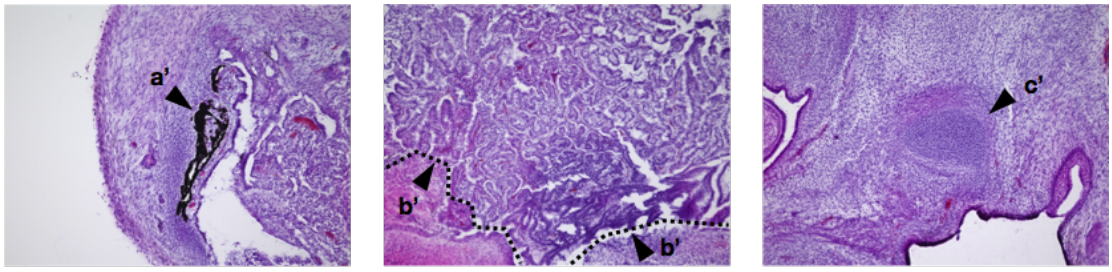

Figure 2.11: Jos-3C1 teratoma analysis. Haemotoxylin and Eosin staining of teratomas derived from Jos-3C1 iPSCs shows the generation of all three germ layers: (a') ectoderm (pigmented cells), (b') endoderm (gut), and (c') mesoderm (cartilage).

(Mayer's Hematoxylin Solution, Sigma) & eosin (Eosin Y solution, Sigma) and analyzed for the generation of all three germ layers (Figure 2.11). The generation of pigmented cells confirmed the ectoderm lineage, gut cells indicated formation of endoderm, and cartilage showed the generation of mesoderm.

2F. Chimpanzee Episomal Reprogramming

Episomal transfections for the reprogramming of chimpanzee cell lines S008919 and S008933 were performed at Applied StemCell (https://www.appliedstemcell.com). They used the Y4 combination of plasmids as described in Okita et al., 2011. Applied Stem Cell provided me with twelve 6-well plates and I isolated 24 clones derived from S008919 and 21 from S008933 (examples in Figure 2.12). From here, the most promising clones were expanded and the rest frozen down for possible revival at a later time. Clones were selected by persisting with densely packed cells in a "cobblestone" morphology with phase-bright colony edges. Picking colonies that maintained a pluripotent state for passaging was aided by live staining for Tra-1-81. Efforts in this study were focused on a clone derived from S008919 named Epi-8919-1A (Figure 2.12A).

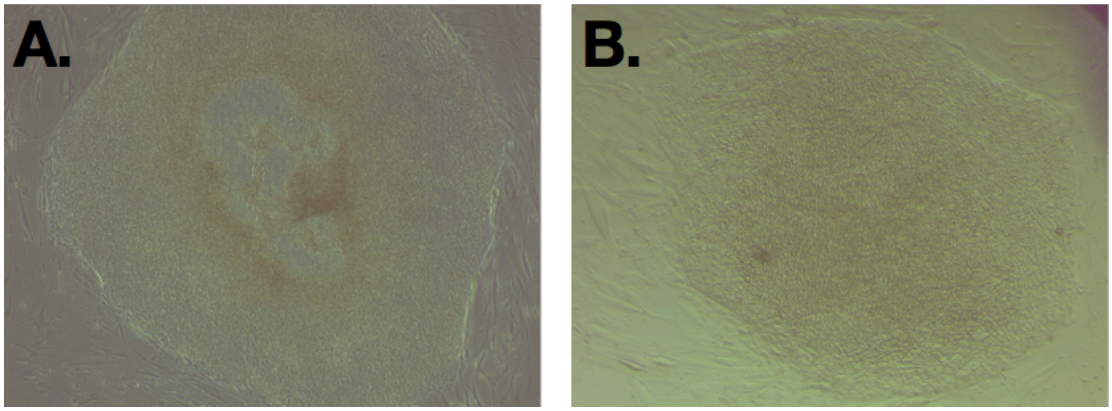

Figure 2.12: Initial chimpanzee iPSC colonies. Initial picked colonies derived from S008919 (A) and S008933 (B) at passage 0.

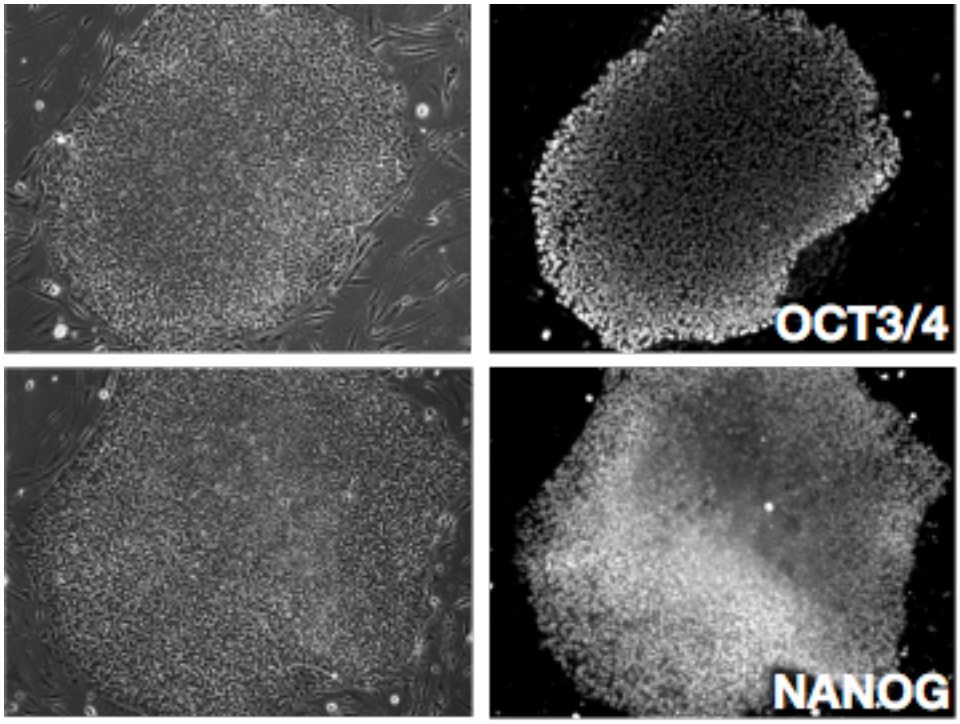

Figure 2.13: Immunofluorescence of chimpanzee Epi-8919-1A colonies. Immunofluorescence staining of Epi-8919-1A chimpanzee iPSC colonies displays OCT3/4 and NANOG expression.

To confirm the acquisition and maintenance of pluripotent stem cells, periodic immunofluorescence staining for pluripotency factors was performed on our iPSC cultures (Figure 2.13). Colonies exhibited homogenous expression of OCT3/4 throughout the colony indicating robust expression of the reprogramming factor. The expression of Nanog indicates an endogenous activation of gene networks associated with pluripotency suggesting complete reprogramming.

Epi-8919-1A cells were transferred to feeder-free conditions on Matrigel (Corning) with mTeSR-1 (ThermoFisher) (Figure 2.14A) to aid in our neural differentiation protocol. Pluripotency was further verified by RT-PCR on total RNA harvested from feeder-free cultures. Epi-8919-1A was shown to have similar levels of expression of OCT3/4, Nanog, SOX2, and L-MYC to human H9 embryonic stem cells (Figure 2.14B). A stable wildtype chimpanzee karyotype of 48, XX was also confirmed through passage 32 in these conditions (Figure 2.15).

Teratoma assays were conducted similarly to orangutan iPSCs using NOD/SCID mice. 2 subcutaneous injections of 1 to 5 million cells suspended in 30% Matrigel (Corning) were made in the dorsal lateral areas of NOD-SCID mice (NOD.CB17-Prkdc^{scid}/NCrCrl, Charles River) similar to Prokhorova et al., 2009. Mice were observed for 11 weeks for the appearance of tumors in the injected areas. One animal formed 2 tumors and euthanized by cervical dislocation. One tumor was largely fluid filled, but the other had a solid dense mass. That teratoma was harvested, fixed in 4% paraformaldehyde, saturated in 30%

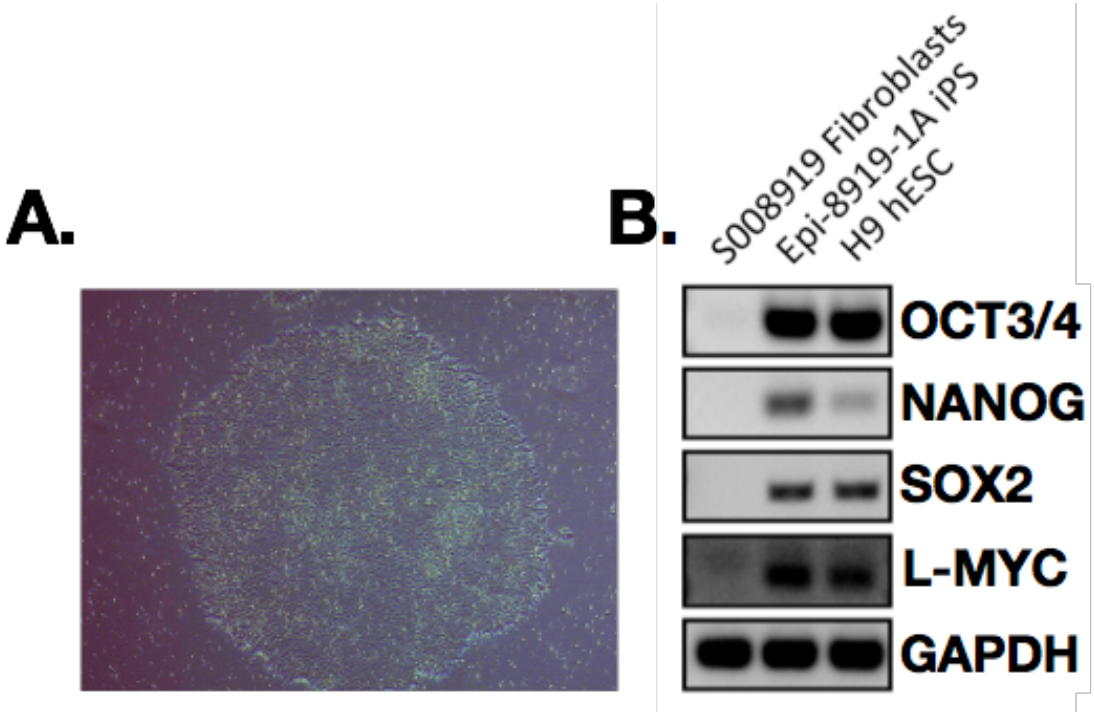

Figure 2.14: Epi-8919-1A transfer to feeder-free conditions. (A) Brightfield image of a chimpanzee iPSC colony on Matrigel feeder-free conditions. (B) RT-PCR products visualized on agarose gel comparing expression of OCT3/4, NANOG, SOX2, L-MYC, and GAPDH in S008919 starting chimpanzee fibroblasts, reprogrammed Epi-8919-1A chimpanzee iPSCs, and human H9 ESCs.

sucrose in PBS, embedded in Tissue Freezing Medium™ (Triangle Biomedical Sciences), and frozen for cryostat sectioning. Sections of the tumor were stained with hematoxylin (Mayer's Hematoxylin Solution, Sigma) & eosin (Eosin Y solution, Sigma) and analyzed for the generation of all three germ layers (Figure 2.16). The generation of pigmented cells and neural rosettes confirmed the ectoderm lineage. Gut epithelial cells indicated formation of endoderm. Abundant cartilage throughout the tumor showed the generation of mesoderm.

2G. Reprogramming Discussion

I have successfully generated pluripotent stem cell lines from chimpanzee and orangutan fibroblasts allowing for an unprecedented look into the development of early tissues in these species. I have applied these cells to a common protocol for generating cortical organoids which the Haussler lab has already demonstrated in human and rhesus embryonic stem cells. This differentiation protocol recapitulates early events in cortical neuron development and enables comparative molecular analysis of this process. Adding these new great ape species provides a rich resource of expression data in early cortical development that could shed light on expression events that happen during the earliest stages of brain development and what makes humans unique from our closest living evolutionary cousins.

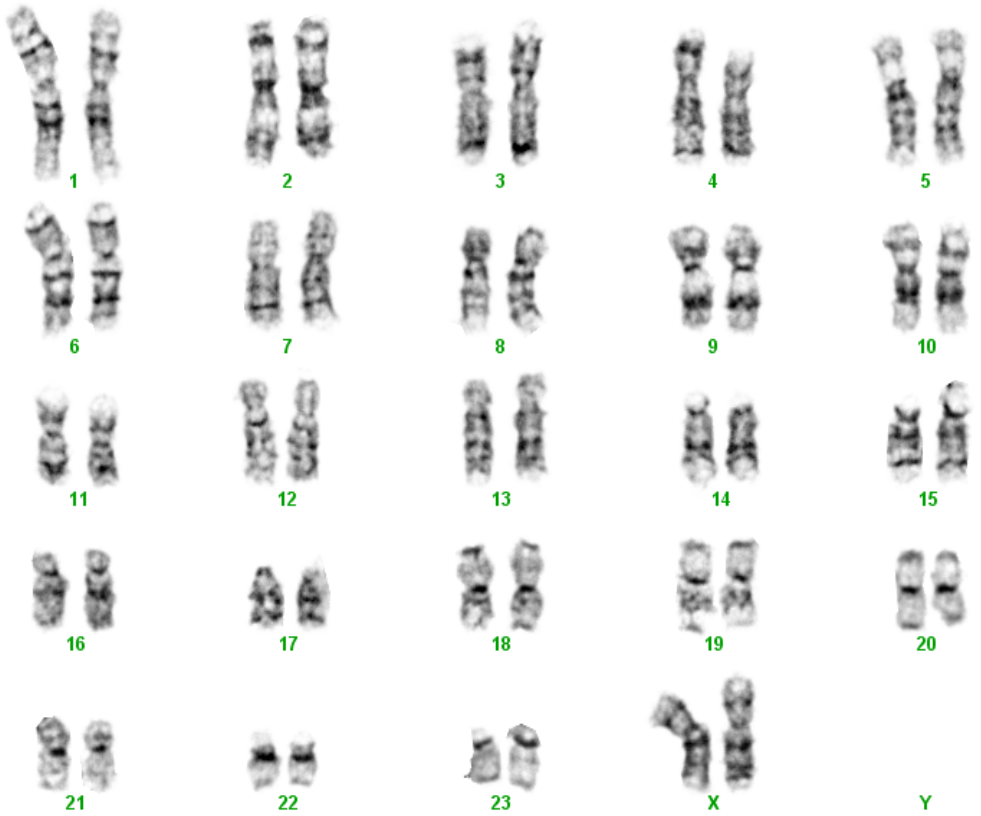

Figure 2.15: Epi-8919-1A karyotype analysis. Wildtype 48, XX karyotype was confirmed in chimpanzee iPSCs at passage 32 after reprogramming.

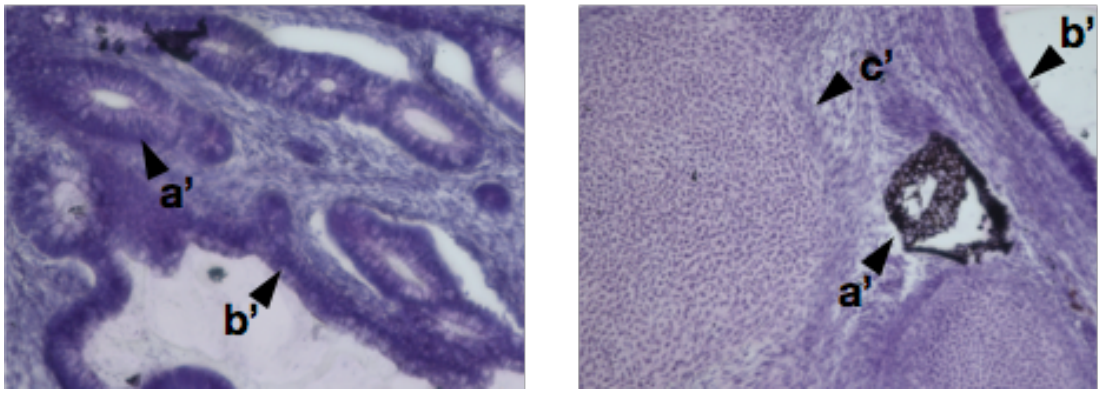

Figure 2.16: Epi-8919-1A teratoma analysis. Haemotoxylin and Eosin staining of teratomas derived from Epi-8919-1A iPSCs shows the generation of all three germ layers: (a') ectoderm (neural rosettes and pigmented cells), (b') endoderm (gut), and (c') mesoderm (cartilage).

Chapter 3

Identifying and Cataloguing Transiently Expressed lncRNAs During Primate Cortical Neuron Differentiation

3. Identifying and Cataloguing Transiently Expressed lncRNAs During Primate Cortical Neuron Differentiation

3A. Introduction

Advances in pluripotent stem cell technology have allowed researchers to probe gene regulatory events that occur during the differentiation of early neocortical cell types using cell lines that model both normal and disease states (Eiraku et al., 2008; Eiraku and Sasai, 2012; Lancaster et al., 2013; Qian et al., 2016). These protocols have been shown to closely recapitulate cellular organization and gene expression events observed in fetal tissue (Camp et al., 2015; reviewed in Fatehullah et al., 2016). Comparisons between primate organoids have revealed differences in the timing of cell divisions and differentiation events when compared to humans (Mora-Bermudez et al., 2016; Otani et al., 2016), though the mechanisms by which these changes are enacted are unknown.

Here I focus on one class of gene regulatory element, long non-coding RNA (lncRNA) which often show tissue specific expression (Pontig et al., 2009;

Cabili et al., 2011; Derrien et al.,2012; Pauli et al., 2012), account for a significant proportion of Pol II output (Carninci et al.,2005; Harrow et al., 2006; Derrien et al., 2012), and show particular enrichment in neural tissues (Ravasi et al., 2006; Cabili et al., 2011; Derrien et al., 2012; Ramos et al., 2013). LncRNAs have been shown to have diverse roles in gene regulation including chromosome inactivation (Penny et al., 1996; Zhao et al., 2008), imprinting (Lighton et al., 1995; Camprubi et al., 2006; Buiting et al., 2007; Pandey et al., 2008; Martins-Taylor et al., 2014), and developmental processes (Rinn et al., 2007; Heo and Sung 2011) and many more have been implicated in establishment of pluripotency (Guttman et al., 2009; Guttman et al., 2011), stem cell maintenance (Rani et al., 2016), reprogramming (Loewer et al., 2010), and differentiation (Guttman et al., 2011). But still, most of the tens of thousands of identified lncRNAs in human have undetermined function (Hon et al., 2017; Lagarde et al., 2017) and lack sequence conservation among vertebrate species (Wang et al., 2004; Church et al., 2009; Cabili et al., 2011; Ulitsky et al., 2011; Kutter et al., 2012). Their tissue specific expression patterns, low level of sequence conservation, and implication in cell type specification make lncRNAs an attractive target for species-specific gene regulation during development.

Here, I utilized an approach focusing on both the gene structure and expression conservation of lncRNAs in equivalent developing tissues among closely related primate species to identify potential functional human lncRNAs. A common protocol for cortical neuron generation from human, chimpanzee,

orangutan, and rhesus pluripotent stem cells was used to recapitulate early events in cortical development and enable comparative molecular analysis of this process. Bulk strand-specific total-transcriptome RNA-sequencing was performed on weekly time points to assess conservation of lncRNA transcript structure and expression among primates. Particular attention was paid toward lncRNA transcripts that were transiently expressed (TrEx lncRNAs), those with max expression after neural induction and with diminished expression by week 5. Single Cell RNA-sequencing on a subset of time points relevant to major differentiation events was used to identify the cell subpopulations associated with the expression of candidate TrEx lncRNAs.

3B. Generation and RNA-sequencing of primate cortical organoids

To study the transcriptional landscape of early cell type transitions during primate cortical neuron differentiation, I subjected human, chimpanzee, orangutan, and rhesus macaque pluripotent stem cells to a cortical neurosphere differentiation protocol based on Eiraku et al., 2008 and optimized for use with our cell lines (Figure 3.1 & 3.2). Embryonic stem cell lines were used for human (H9) and rhesus (LYON-1) time courses, but, since embryonic stem cells are not available for great apes, I generated integration-free induced

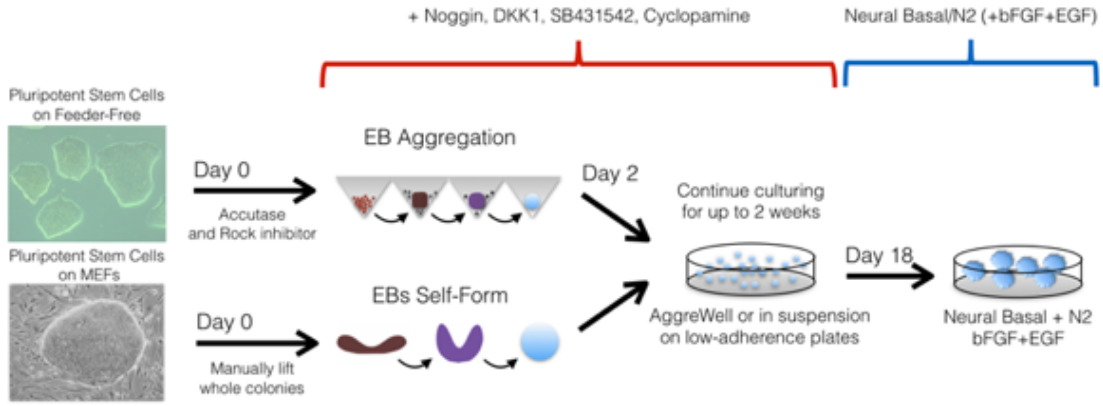

Figure 3.1: Cortical neural epithelium differentiation protocol. An outline of our dorsal neuron differentiation assay for both feeder grown and feeder-free cultures is provided. Human and rhesus embryonic stem cell cultures began on MEFs, were manually lifted, and allowed to self-form into embryoid bodies. Chimpanzee and orangutan cultures on feeder-free conditions were first dissociated into single cell suspensions and aggregated into embryoid bodies consisting of approximately 10,000 cells using AggreWell plates (Stem Cell Tech). Embryoid bodies prepared with both methods were incubated with the addition of Noggin, DKK1, SB431542, and cyclopamine for 18 days to induce cortical neural epithelium differentiation. The cultures were then switched to Neural Basal media supplemented with N2 with the addition of bFGF and EGF for chimpanzee and orangutan. Adjusted time points were used for rhesus to account for gestational time differences as explained in the methods section.

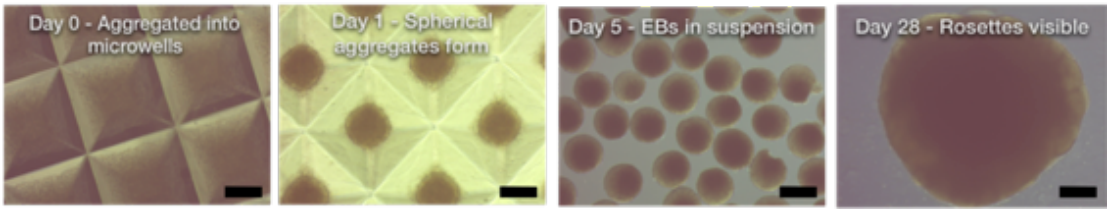

Figure 3.2: Embryoid body formation and differentiation. An example of chimpanzee aggregation and differentiation is illustrated. Single cell suspensions were aggregated on day 0 in AggreWell plates (Stem Cell Tech) at a density of 10,000 cells per embryoid body. These aggregates form into homogenous spheroids within a day. By day 28 neural rosettes are clear along the periphery of the neurospheres.

pluripotent stem cell lines for chimpanzee (Epi-8919-1A) and orangutan (Jos-3C1) from primary fibroblasts.

Performance of these stem cell lines in our cortical organoid differentiation assay was evaluated by immunofluorescence staining at day 35 (or day 28 in rhesus) showing efficient production of radial glia, intermediate progenitors, and early deep layer cortical neurons (Figure 3.3) in highly structured neural rosettes as described previously (Eiraku et al., 2008; Eiraku and Sasai, 2012; Lancaster et al., 2013; Camp et al., 2015; Qian et al., 2016). RNA samples were collected in at least duplicate from pluripotent stem cells and weekly time points over 5 weeks of neural differentiation in each species and used to create total-transcriptome strand-specific RNA sequencing libraries. Due to the shorter gestational period in rhesus macaque, those samples had adjusted time points with ~5.5 day weeks (see Methods). In all, over 2 billion paired-end RNA sequencing reads were uniquely mapped to their respective genomes from 49 libraries, averaging 41 million reads per library with a minimum of 46 million total reads across replicates per species time point. Cufflinks v2.0.2 (Trapnell et al., 2010; Trapnell et al., 2012) was used to assemble potential novel transcripts in each species and the Cuffmerge tool combined gene models across time points in each respective species using FANTOM5 lv3 (Hon et al., 2017) as a reference annotation. The resulting annotations from each species were projected through Cactus alignment (Stanke et al., 2008) to each of the other primate genomes (Figure 3.4). Guided by the Cufflinks annotation set in each genome, these

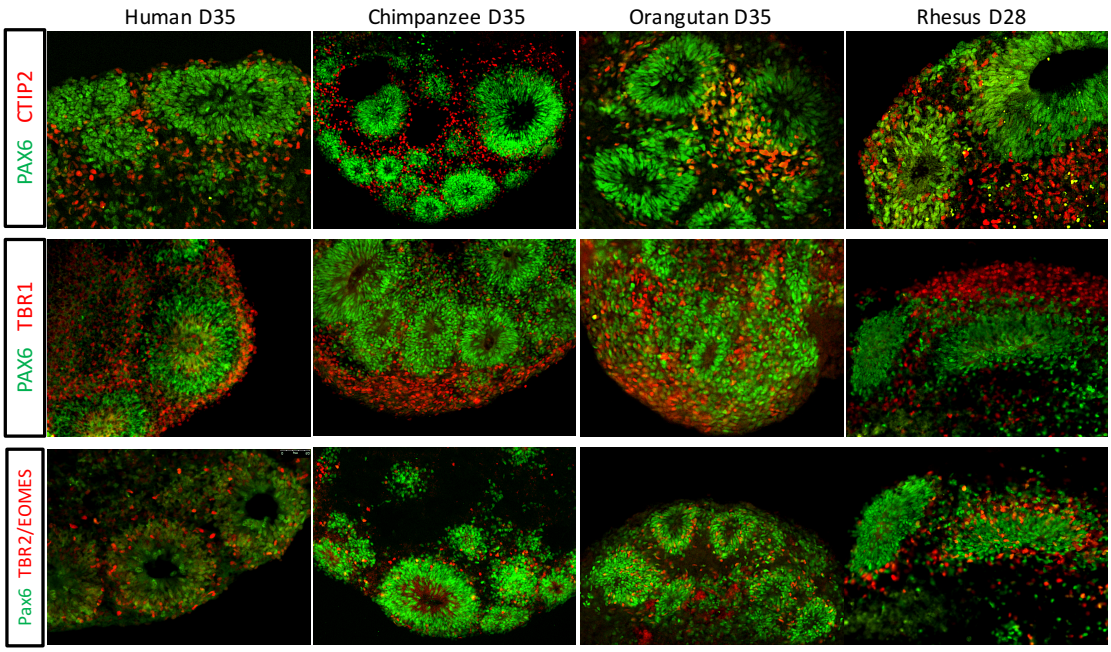

Figure 3.3: Immunofluorescence staining of week 5 neurospheres. Immunofluorescence was used to confirm on-target differentiation of our neurospheres at our endpoint of 5 weeks (28 days in rhesus). PAX6, marking neural progenitors, form the center of a rosette pattern forming around lumen like structures that are prominent throughout neurospheres in each species. Positive CTIP2 and TBR1 indicate early deep layer neurons projected radially out from neural progenitors. Finally, TBR2 is characteristic of intermediate progenitors or early migrating neurons as they transit radially outward from the lumen-like structures.

projections from the other genomes were assigned a putative gene locus. In cases where a projection overlapped multiple genes, the gene whose transcripts had the highest exonic Jaccard similarity were chosen. RSEM (v1.3.0, Li and Dewey, 2011) was used to calculate expression values of these new gene models (Figure 3.4).

On target differentiation of dorsal cortical tissue was confirmed by the transcriptional profile of marker genes in all species (Figure 3.5). Pluripotency markers such as OCT3/4 are downregulated by the week 1 time point in all species and early neural stem cell markers like PAX6 are upregulated. There was strong expression of deep layer neuron markers, such as TBR1, by week 5 in all species (Figure 3.6). Overall, there is strong induction of early neural differentiation and dorsal forebrain markers with little expression of ventral forebrain, midbrain, hindbrain, cerebellum, and spinal cord markers (Figure 3.5). Although I see expression of canonical interneuron markers including GAD1, MEIS2, and DLX1, these cultures exhibit low expression of NKX2-1 and LHX6 which would normally be associated with interneurons and their progenitors in the early ganglionic eminence. It has been suggested that there is a source of GABAergic interneuons that originate from the dorsal telencephalon in primate species which is in accordance with our results obtained here (Radonjić et al., 2014).

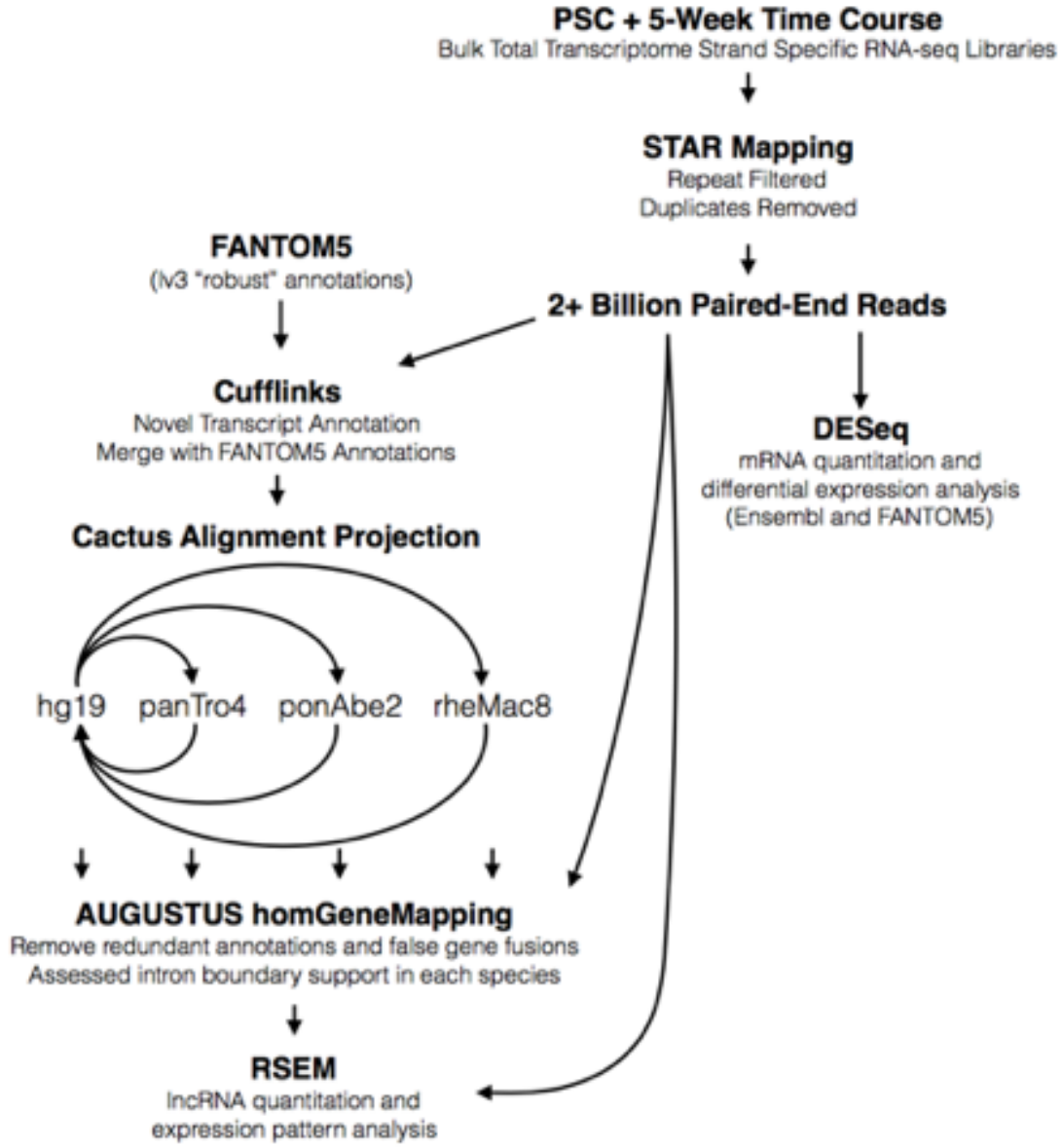

Figure 3.4: Transcriptomic analysis. A flowchart depicts the analysis pipeline used to process the neural differentiation time course RNA sequencing data and identify expressed lncRNAs.

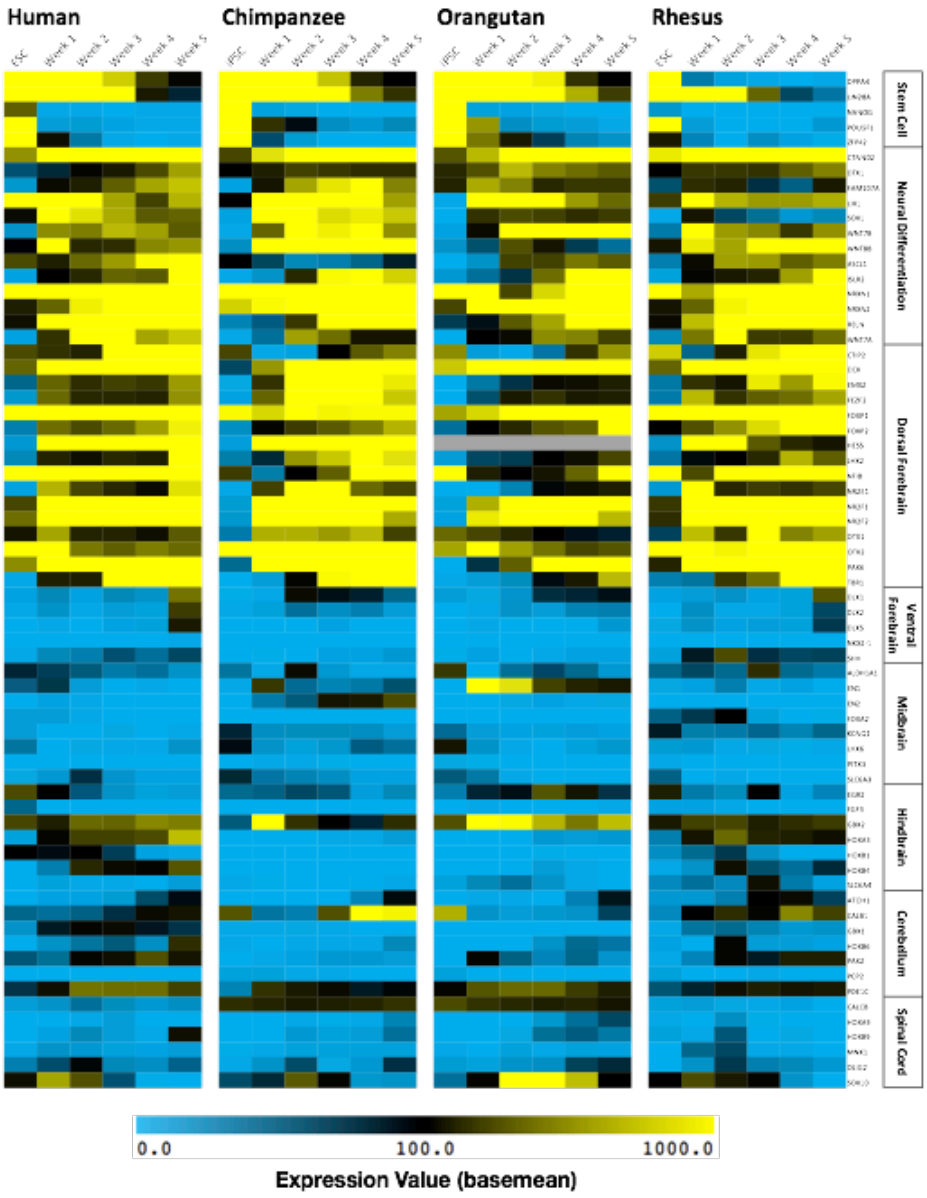

Figure 3.5: On-target differentiation by cell marker expression. A heatmap depicting gene expression colored by basemean values calculated with DESeq2. Canonical marker genes for pluripotent stem cells, neural differentiation, dorsal forebrain, ventral forebrain, midbrain, hindbrain, cerebellum, and spinal cord are displayed.

3C. Classification of lncRNAs and examples of TrEx lncRNAs

The expressed lncRNAs detected in our neural differentiation assay fell into two main categories: multi-exonic transcripts and long actively transcribed regions (Figure 3.7). Multi-exonic transcripts appeared similar to canonical messenger RNAs featuring small exons and large introns. In all, multi-exonic transcripts account for 920 of our primate conserved loci. Long actively transcribed regions, on the other hand, were long regions often stretching hundreds of kilobases with some approaching 1 megabase. They feature what seems to be continuous transcription without a dominant exonic structure and account for over a thousand gene loci in human. These regions included many that are known to be associated with imprinting during development such as PWRN1/2, so some clearly have functional significance. For this study, however, I will focus on multi-exonic transcripts that are more likely to exhibit their function through the transcripts themselves rather than just the act of transcription (Ulitsky, 2016).

Striking among multi-exonic transcripts are what I have coined transiently expressed (TrEx) lncRNAs. These TrEx lncRNAs are seen to be expressed primarily at one time point and are absent at all others. This signal is incredibly reproducible in replicates and, for the select transcripts seen here, they have appeared in all RNA sequencing replicates at their relevant time point. TrEx lncRNAs were assigned numbers arbitrarily by an algorithm used to

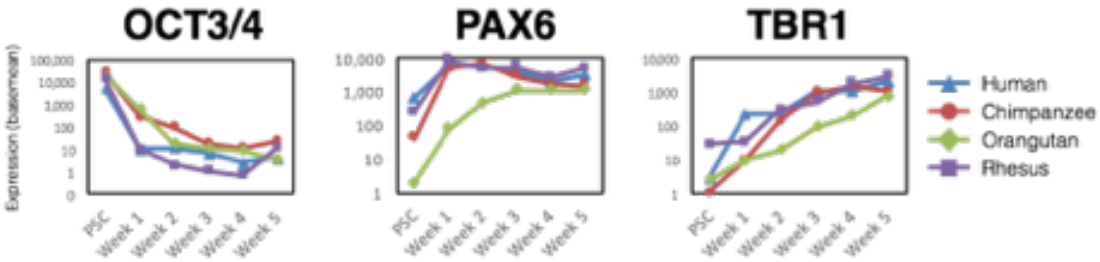

Figure 3.6: Expression of canonical cell markers over the time course in each species. Line graphs depicting the expression in basemean value of the genes OCT3/4, PAX6, and TBR1 in human (blue), chimpanzee (red), orangutan (green), and rhesus (purple) over the neural differentiation time course.

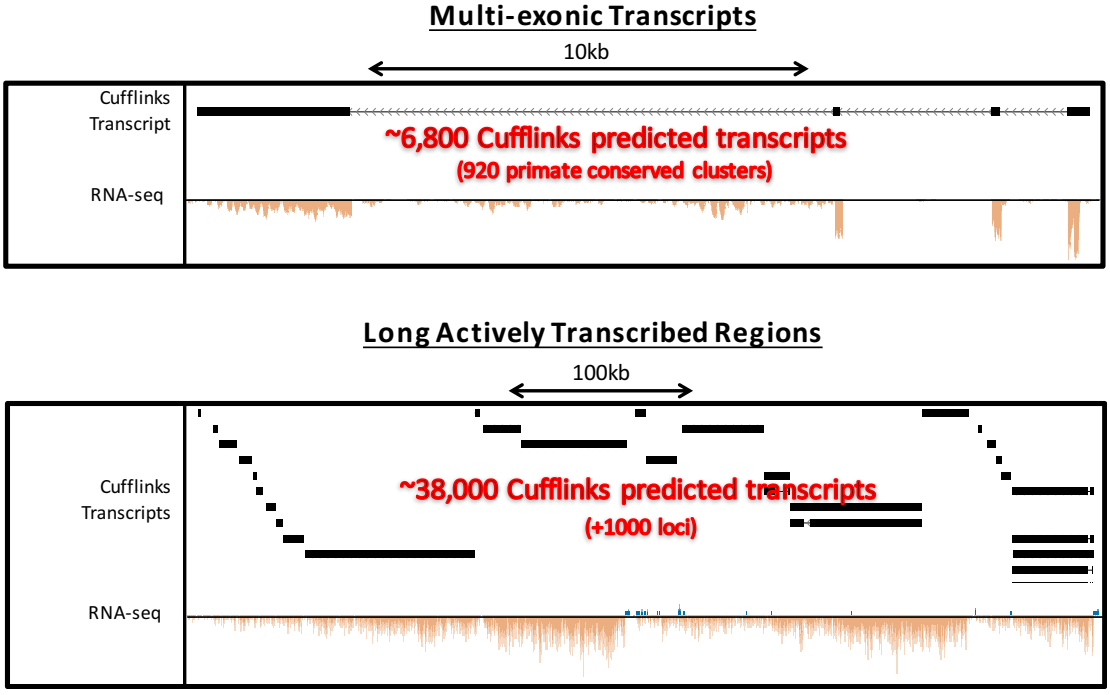

Figure 3.7: Classification of lncRNA transcript types. UCSC Genome Browser screenshots are shown of two examples of lncRNA loci identified by our block loci caller. The first depicts a typical multi-exonic transcript with well-defined short exons and long introns similar to the structure of protein-coding genes. The bottom screenshot depicts a lncRNA gene locus with no clear dominant exonic structure, very long exonic sequences, and the detectable splice junctions are short. Both examples show coverage from RNA sequencing in blue for positive strand and orange for negative.

identify new gene loci. Transcript models from Cufflinks were linked together into block loci if they were expressed from the same strand, within 10kb of each other, and uninterrupted by antisense transcripts. Here, I will describe a few TrEx lncRNA gene loci with heretofore undefined function.

TrEx5700 is a striking human-specific transcript I identified that met our transiently-expressed multi-exonic criteria (Figure 3.8). It is strongly expressed at week 2 of cortical organoid differentiation in human, yet is nearly off at all other time points. No evidence of its expression was seen in any chimpanzee, orangutan, or rhesus RNA sequencing samples. Curiously, we've noticed a stretch of 20 homopolymer adenines just upstream of the transcription start site. This site appears to consist of fewer adenines in other primate species with the fewest in rhesus lacking 14 of the 20 observed in human, possibly indicating DNA polymerase slippage as a mechanism for its expansion. It was also noted that Denisovan and Neanderthal carry the chimpanzee version of this region lacking 4 of the adenines seen in human. Such an expansion can potentially lead to novel accessibility to chromatin at the promoter of this gene allowing for transcription initiation truly specific to human.

Some examples of human TrEx lncRNAs were also observed in other species. TrEx2174 is also notable in its week 2 specific expression and the same pattern is observed in chimpanzee (Figure 3.9). This gene is not expressed in our orangutan or rhesus data. Similarly to TrEx5700, TrEx2174 also has a 19bp

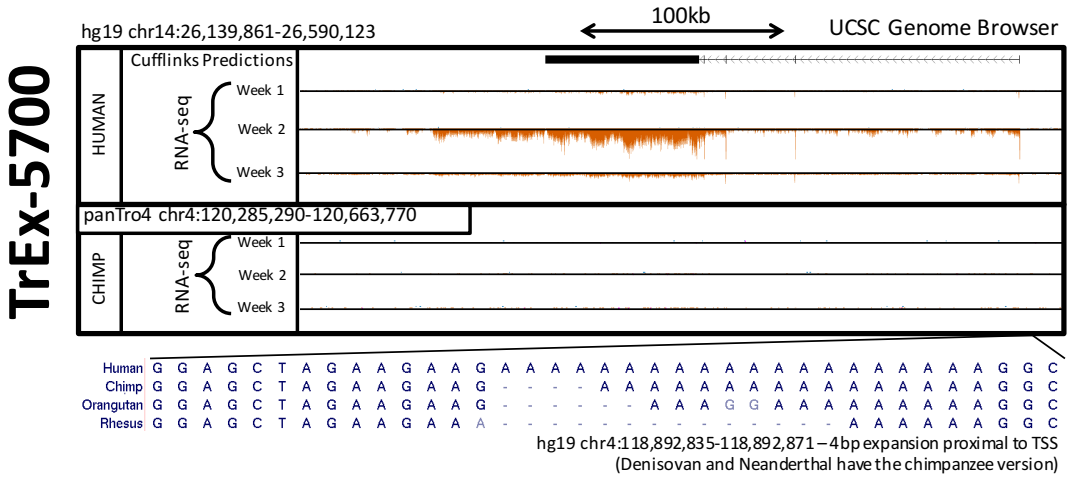

Figure 3.8: Human-specific TrEx5700. Bulk RNA sequencing coverage tracks are displayed from human and chimpanzee for weeks 1, 2, and 3 of neural differentiation over the TrEx5700 gene locus. Expression at this locus is only observed in human. The Multiz alignment just upstream of the transcription start site for TrEx5700 has a stretch of 20 homopolymer adenines that is specific to human among extant great apes and not seen in the current assemblies of Denisovan and Neanderthal.

insertion just upstream of its transcription start site that is specific to the species that express this transcript suggesting a possible mode of its birth.

Many of these lncRNAs were observed in all 4 of our species. TrEx4039 was one such transcript (Figure 3.10). Expression peaks at weeks 1 or 2 in all species and is extinguished by week 5. Chimpanzee appears to express a unique isoform of this transcript that is not shared with human or rhesus, but can be seen expressed early in orangutan. While chimpanzee ceases expression from this locus at week 2, orangutan appears to switch to the longer isoform observed in the other species at week 2. So, even among conserved transcripts, the entire gene body is not necessarily conserved.

3D. Expression and gene structure conservation of lncRNAs across primates

It has been suggested that conservation of exon boundaries within a lncRNA gene may be a feature indicative of functional transcripts (Ulitsky, 2016). Gene structure conservation of expressed transcripts among our primate species was assessed using the homGeneMapping tool from the AUGUSTUS toolkit (Konig et al., 2016). homGeneMapping makes use of cactus alignments to project annotation features in all pairwise directions, providing an accounting of

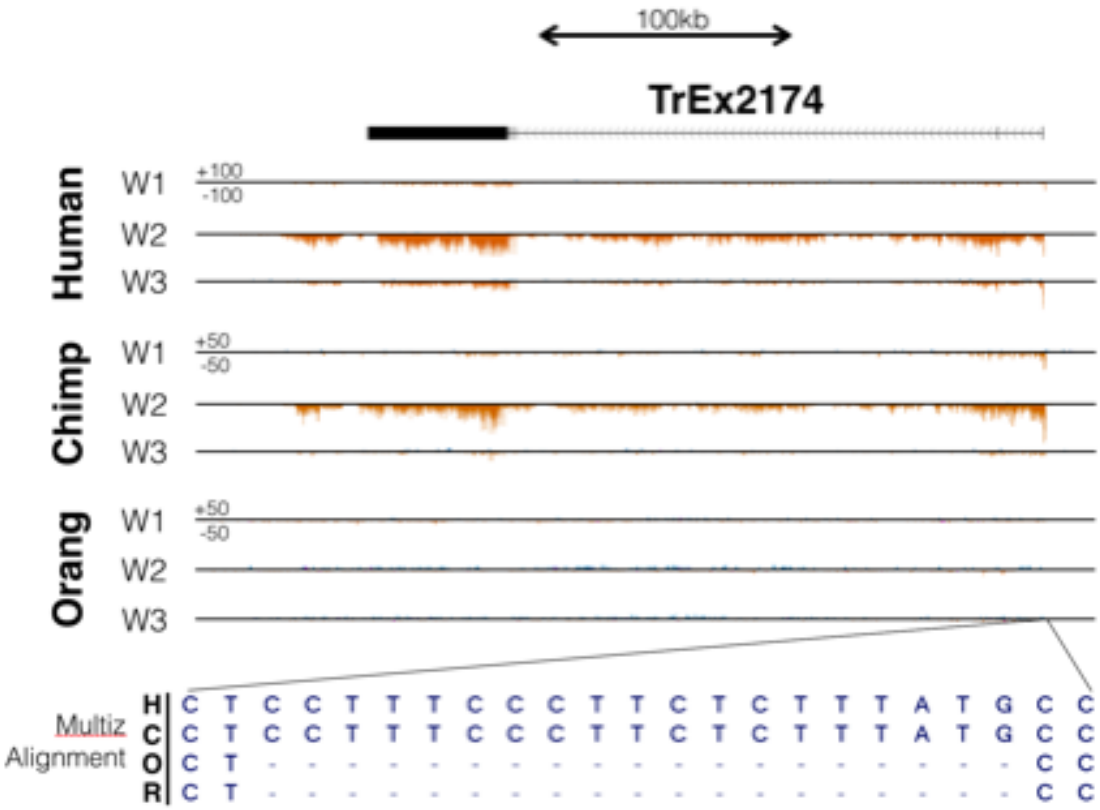

Figure 3.9: Human-chimpanzee specific TrEx2174. Bulk RNA sequencing coverage tracks are displayed from human, chimpanzee, and orangutan for weeks 1, 2, and 3 of neural differentiation over the TrEx2174 gene locus. Expression at this locus is conserved in human and chimpanzee but lacking in orangutan. The Multiz alignment just upstream of the transcription start site for TrEx2174 has a 19bp insertion that is specific to human and chimpanzee among extant great apes.

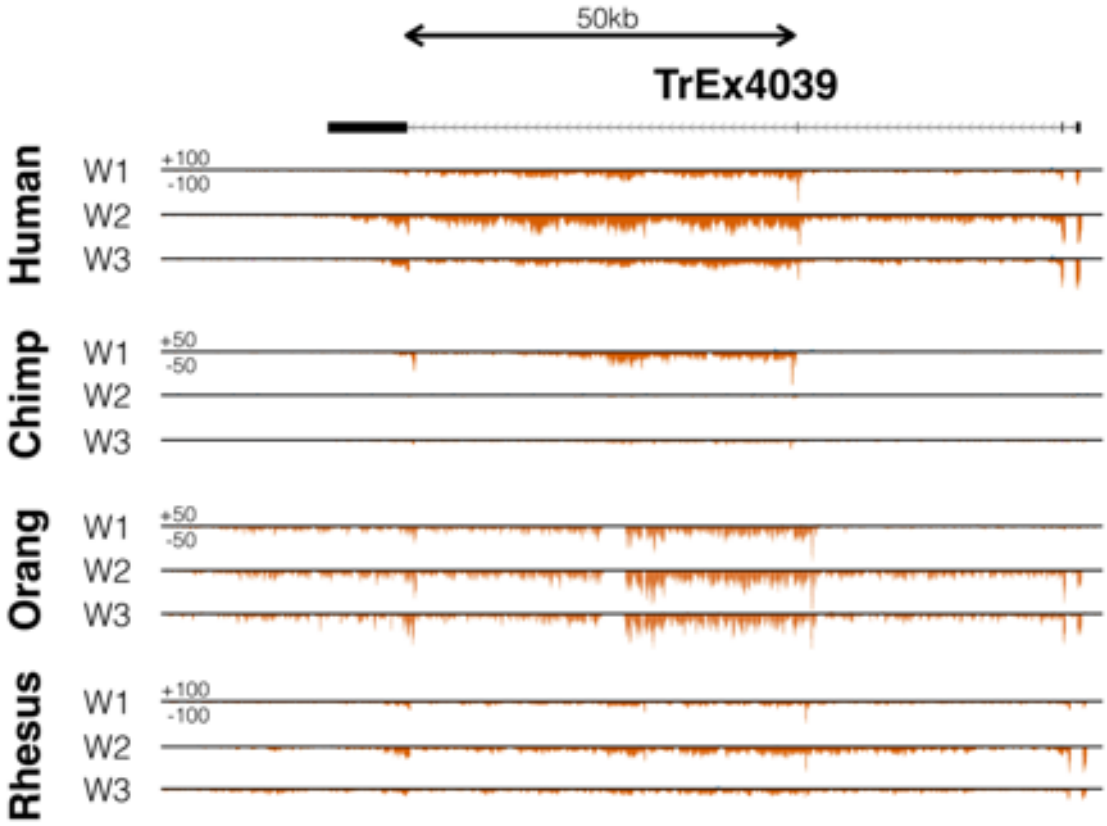

Figure 3.10: Primate conserved TrEx4039. Bulk RNA sequencing coverage tracks are displayed from human, chimpanzee, orangutan, and rhesus for weeks 1, 2, and 3 of neural differentiation over the TrEx4039 gene locus. Chimpanzee appears to express a unique isoform of this lncRNA at an earlier time point. Orangutan also exhibits this shorter isoform though appears to switch to the longer isoform observed in human and rhesus at later time points.

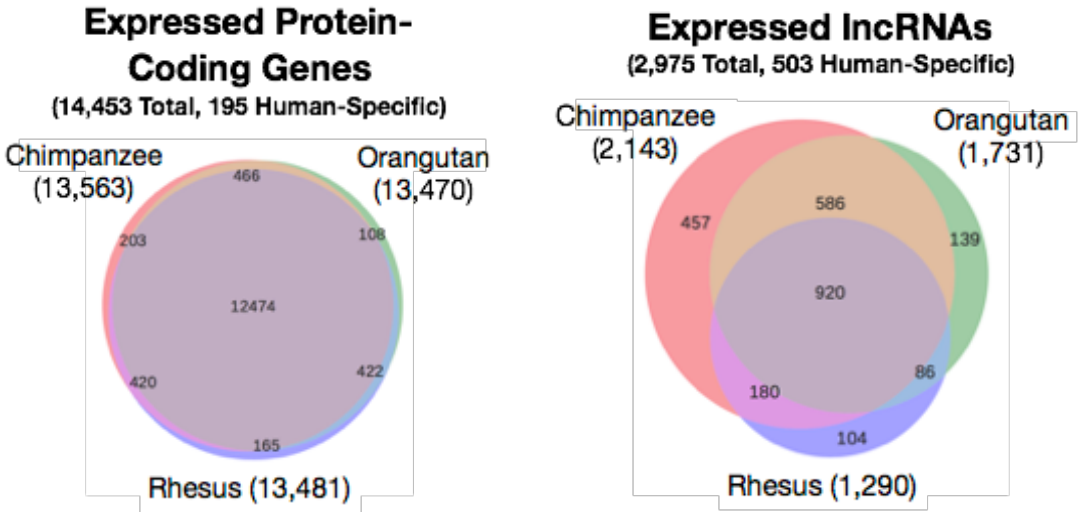

Figure 3.11: Transcript structure conservation. Venn diagrams depicting the support of intron boundaries within each genome show the degree of structural conservation of human protein-coding genes and lncRNAs in each of the other species. To be considered conserved, a gene must have a TPM greater than 0.1, 50% intron boundary support in human, and a non-zero intron boundary support in the target species.

features found in other genomes. homGeneMapping was provided both the Cufflinks transcript assemblies as well as expression estimates derived from the combination of the week 0 to week 5 RNA sequencing experiments in all four species. The results of this pipeline were combined with the transcript cactus projections to ascertain a set of gene loci that appear to have human specific expression, human-chimp specific expression, great-ape specific expression, and expression in all primates (Figure 3.11). Transcript models that had at least 50% intron junction support in human were considered conserved in a non-human primate genome if the target genome had RNA sequencing read support for any of its intron junctions and the gene cluster had a TPM value greater than 0.1. All single-exon transcripts were filtered out to reduce noise. To eliminate the possibility of the specificity results being skewed by assembly gaps or alignment error, loci who appear to have sub-tree specific expression were checked against the cactus alignment to ensure that there was a matching locus in each other genome. If a genome appeared to be missing sequence, then this locus was flagged as having incomplete information and excluded from further analysis, which accounted for about 4% of gene loci (see Methods).

In all, support for 2,975 human poly-exonic lncRNA gene clusters were found in our human RNA sequencing reads. 503 human-specific, 457 human-chimp specific, 586 great ape-specific, and 920 primate-conserved lncRNAs were found by intron boundary support and a maximum expression of at least 0.1 TPM (Figure 3.12) showing the expected higher overlap in species that are

separated by less evolutionary distance. Among the primate conserved category are the previously described mammalian conserved lncRNAs MALAT1, NEAT1, H19, PRWN1, and CRNDE. 347 novel gene clusters were also found by Cufflinks in the human RNA sequencing data points, 160 of which were human specific, 164 conserved in chimp, 105 in great apes, and 79 conserved across all of the primates (Figure 3.8), displaying a similar distribution to lncRNAs defined by the FANTOM5 consortium (Hon et al., 2017). 580 chimpanzee-specific, 1,709 orangutan-specific, and 593 rhesus-specific gene loci were also detected further supporting a relatively fast turn-over of transcribed intergenic loci, though I suspect that the orangutan estimates are inflated due to its relatively poor genome assembly. Comparing these figures to protein-coding genes, 14,453 genes were found to be expressed in human of which 12,474 shared intron boundaries among all species (Figure 3.7) confirming a much higher degree of gene structural conservation by these metrics.

LncRNAs have previously been reported to be time point specific in developing tissues (Cao et al., 2006; Amaral and Mattick, 2008; Chodroff et al., 2010; Qureshi et al., 2010; Pauli et al., 2012). For this reason, I looked for time point-specific expression of lncRNAs in our data sets. I defined human transiently expressed (TrEx) lncRNAs as those whose max expression was between weeks 1 and 4 and had less than 50% their maximal expression value at weeks 0 and 5. Using these metrics, I identified 386 human TrEx lncRNAs, most of which were expressed primarily at one time point (Fig. 3.13). I next assessed

if these transcripts were also transiently expressed in other species requiring that they also have max expression at weeks 1-4 and were below 80% max expression at weeks 0 and 5. Using these strict parameters, 176 were conserved in chimpanzee, 148 in orangutan, and 49 in rhesus (Fig 3.13). These figures do not account for the possibility that conserved lncRNAs may be transient when taken to later time points though it does suggest that the exact timing of lncRNA expression during these weekly time points is not well conserved through rhesus.

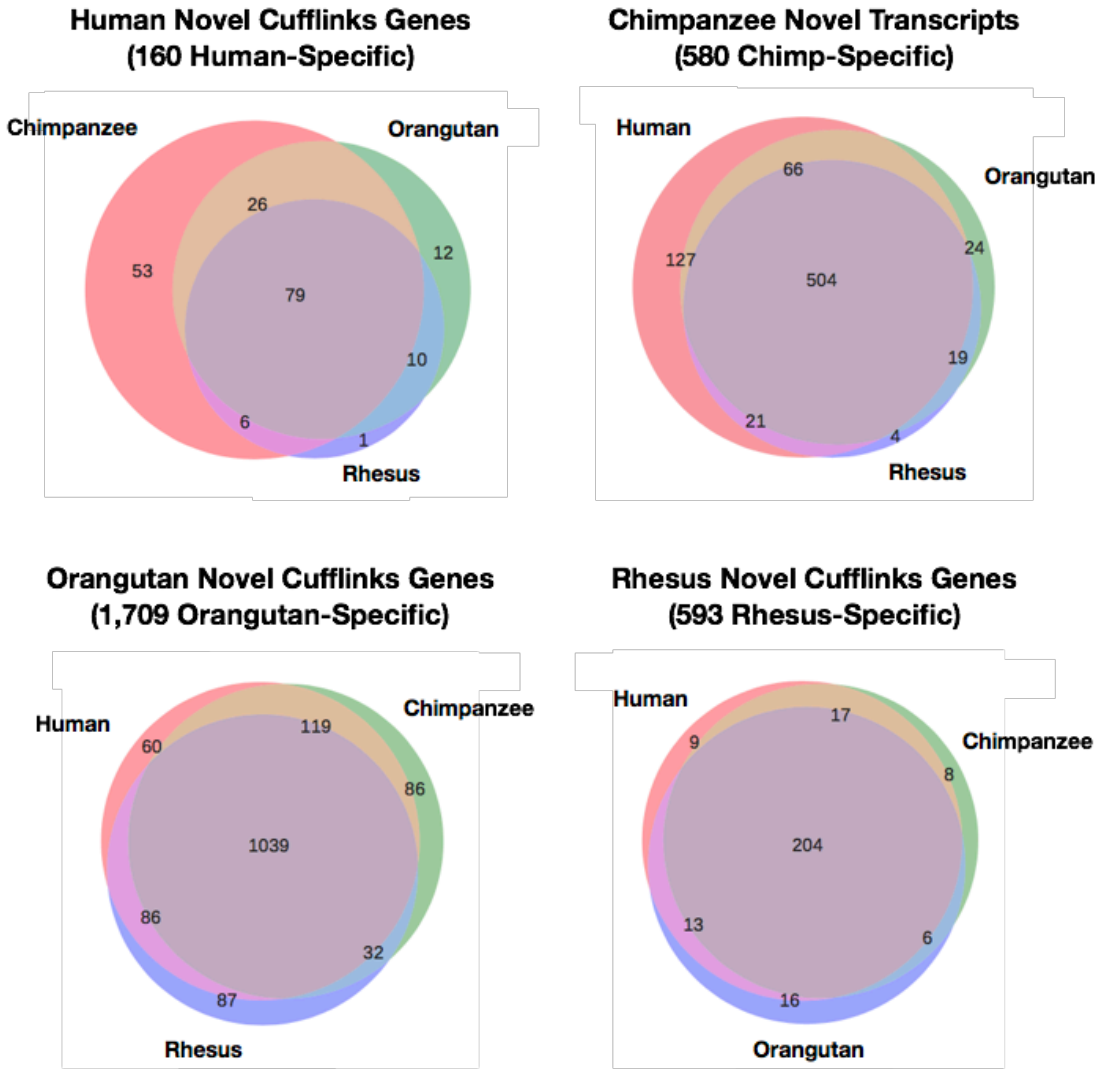

Figure 3.12: Novel detected Cufflinks transcripts. Novel multi-exonic gene loci with no overlap to genes annotated in Ensembl or FANTOM5 (Hon et al., 2017) were identified and assembled by Cufflinks (Trapnell et al., 2010; Trapnell et al., 2012). Novel gene loci identified in each species were evaluated for overlap with transcripts detected in other genomes by pairwise liftOver using Cactus alignments. Loci with transcripts that shared at least one intron junction, had a maximum expression greater than 0.1 TPM, and was align-able between genomes were considered "shared".

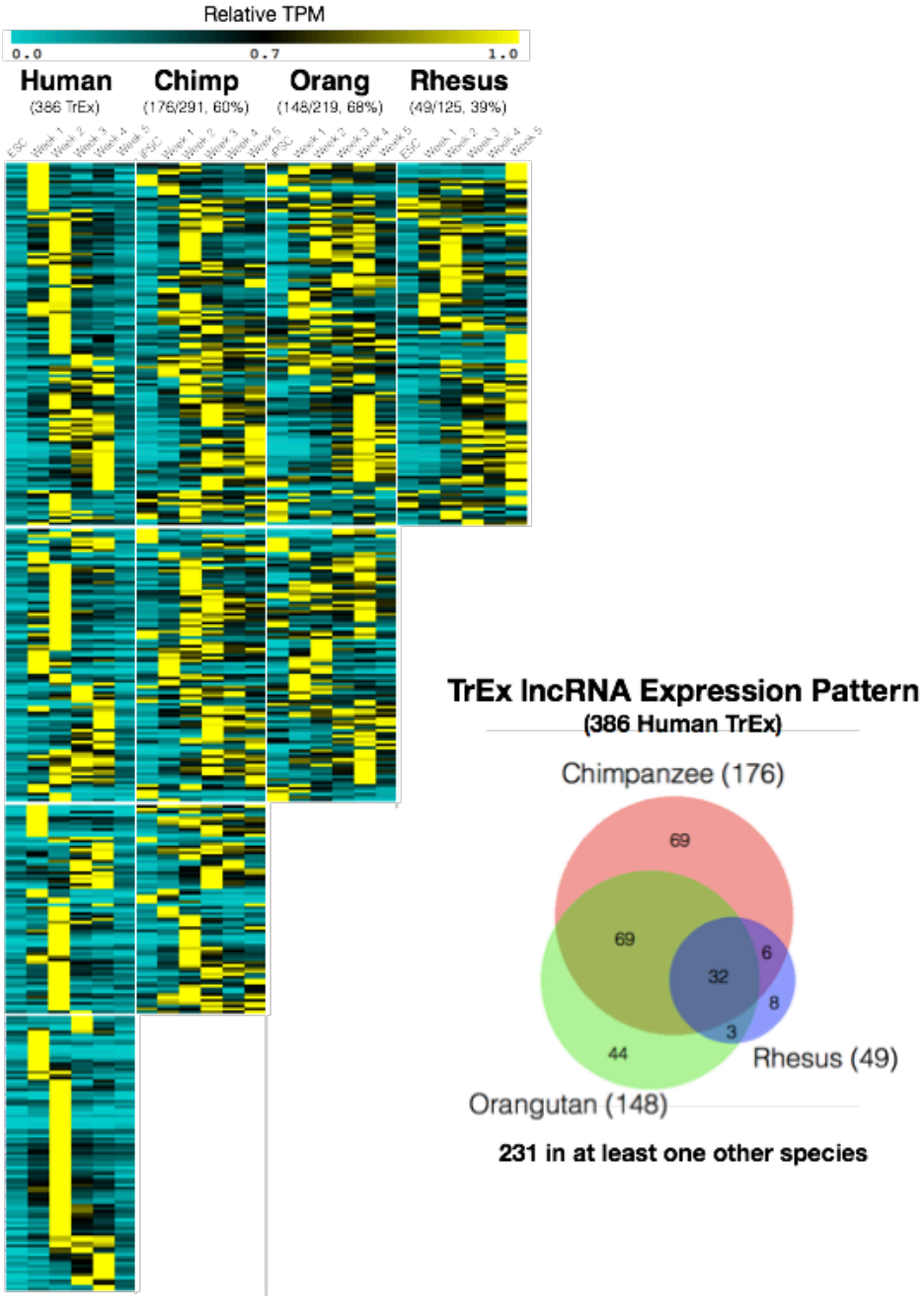

Figure 3.13: Conservation of the TrEx expression pattern. LncRNAs that had transient expression patterns in human samples were compared across species. A heatmap depicting expression in relative TPM normalized to the max value in each species shows expression patterns of human TrEx lncRNAs with conserved structure in the other species. To be considered conserved, a gene must have a TPM greater than 0.1, 50% intron boundary support in human, and a non-zero intron boundary support in the target species (Venn diagram).

Chapter 4

Verifying Gene Regulatory Network Correlation from Single Cell RNA-Sequencing with CRISPR-Activation

4. Verifying Gene Regulatory Network Correlation from Single Cell RNA-Sequencing with CRISPR-Activation

4A. Introduction

Bulk RNA-sequencing does not capture the complexity of the organoid differentiation system and is limited to displaying only the average of gene expression across all cells harvested at any given time point. I know this this be a gross over-simplification as these cortical organoids consist of at least 3 major cell types including radial glia, intermediate progenitors, and early neurons as evidenced by our own immunofluorescence data (Figure 3.3). Other studies using these systems have found other subtle cell subtypes whose signatures would be missed or drowned-out in bulk RNA-sequencing measurements including excitatory neurons and deep layer neurons (Eiraku et al., 2008; Eiraku and Sasai, 2012; Lancaster et al., 2013; Qian et al., 2016). Bulk RNA-sequencing gives us little insight into the biological context in which specific expression events occur.

It has been suggested that lncRNAs are not only tissue specific but can be used as specific cell type markers (Guttman 2009). It is still unknown if transiently expressed lncRNAs are associated with specific, short-lived cell states that may be important for tissue development. Here I have taken steps to address these questions generally and provide a method to test their putative functionality by CRISPR-activation.

4B. Cell type identification in single cell RNA-sequencing

I performed 10X Chromium 3' end single cell RNA-sequencing on human ESCs and cortical organoids at weeks 0, 1, 2, and 5. Reads were mapped with the CellRanger 2.0 pipeline using modified FANTOM5 gene models as described earlier for our bulk RNA sequencing analysis. In all, nearly 800 million reads were obtained from 14,086 cells averaging 56.6k reads per cell. The total number of genes detected per library ranged from 28k in week 5 neurospheres to 36k at week 1, averaging between 1702 and 4978 genes per cell.

tSNE plots generated with CellRanger v1.2 (10X Genomics) identified increasing cell heterogeneity as differentiation progressed (Figure 4.1). Using a combination of k-means clustering, graphical clustering (CellRanger v1.2), and visual inspection, I were able to manually curate clusters of cells with expression profiles matching neuroepithelium (NE), radial glia (RG), and Cajal-

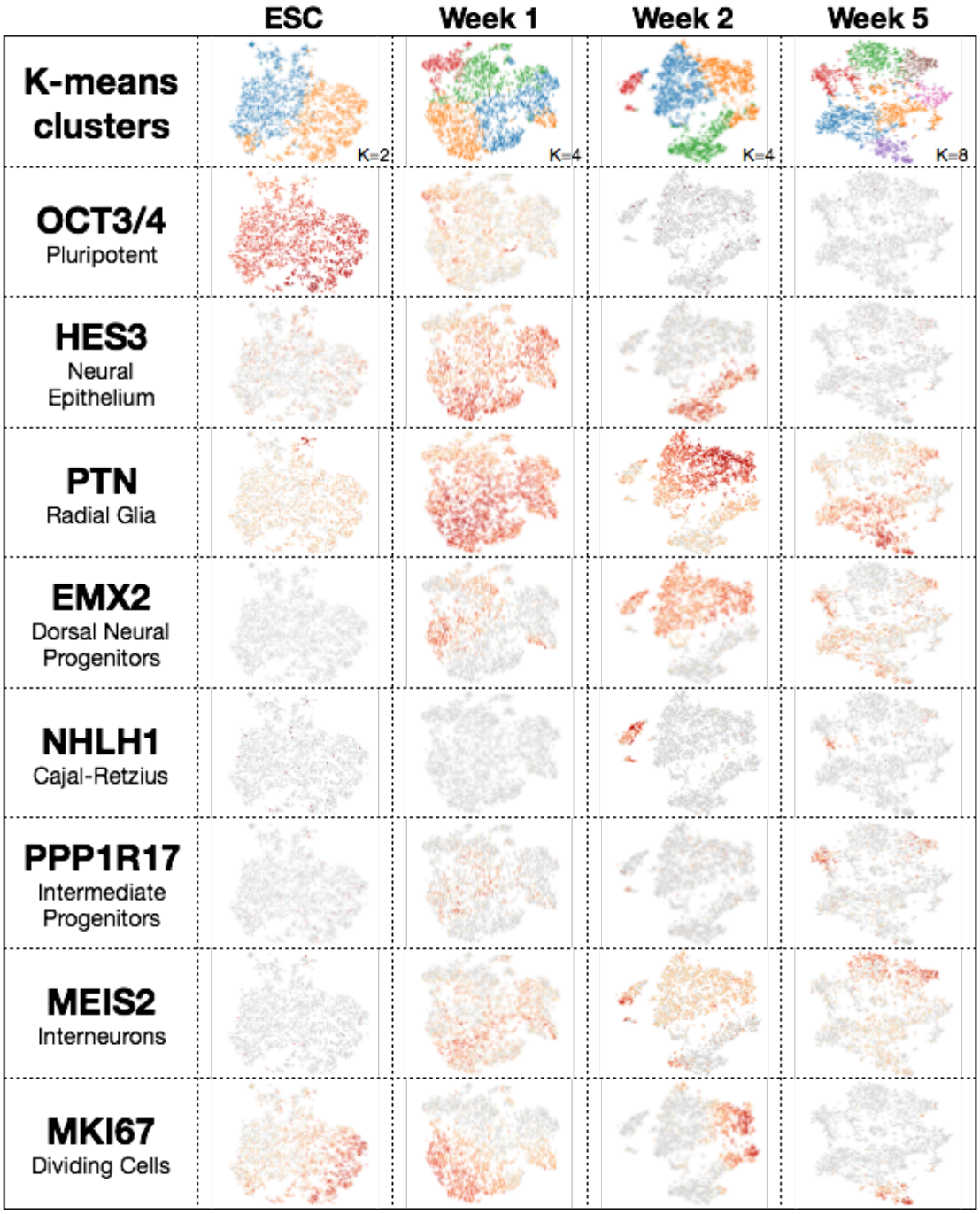

Figure 4.1: Human neurosphere single cell RNA-sequencing time course. t-SNE plots from single cell RNA sequencing displaying the expression of OCT3/4, HES3, PTN, EMX2, NHLH1, PPP1R17, MEIS2, and MKI67 show the increasing cell heterogeneity as time progresses in weeks 0, 1, 2, and 5 of neural differentiation in from human ESCs.

Retzius (CR) cells in our week 2 libraries (Figure 4.2). NE cells were identified by expression of HES3 and NR2F1 forming a cluster of 1261 cells (29%) (Figure 4.2C). CR cells exhibited a clear signature expressing TBR1, EOMES, LHX9, and NHLH1 comprising a cluster of 356 cells (8%) (Figure 4.2E). The largest proportion of cells showed strong expression of cortical RG markers SOX2, EMX2, NNAT, PTN, and TLE4 making up 2593 cells (59%) (Figure 4.2D). About 176 cells (4%) showed no strong association with these manually curated cell clusters and had no significant distinguishing genes, so I determined that they likely represented cell doublets and their prevalence is consistent with theoretical estimates from the number of cells I captured per library. At week 5, cells expressing NE markers were virtually absent and instead additional clusters expressing mature neuronal markers were present (Figure 4.1).

4C. Transiently expressed lncRNAs show cell type-specific expression patterns

LncRNAs have previously been shown to be associated with specific tissues and developmental time points (Cabili et al., 2011; Pontig et al., 2009; Derrien et al., 2012; Pauli et al., 2012) with spatial and temporal expression patterns within the brain (Cao et al., 2006; Amaral and Mattick, 2008; Chodroff

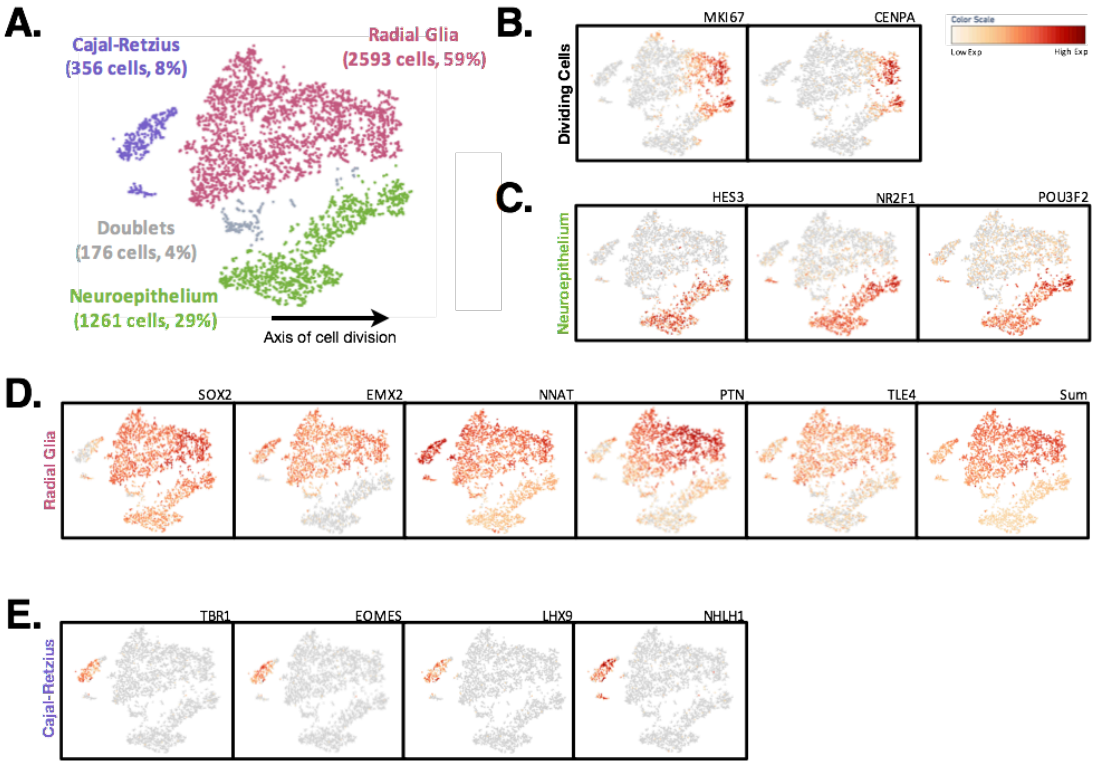

Figure 4.2: Cell type detection in week 2 neurosphere single cell RNA-seq. (A) Shown is a tSNE plot of cell types detected in week 2 single cell RNA-sequencing libraries. Putative cell types were manually curated by combination of K-means clustering, Louvain graphical clustering, and canonical cell type marker expression. Both the radial glia and neuroepithelium clusters showed expression of MKI67 and CENPA concentrated toward one end of the tSNE plot indicating dividing cells as would be expected for these cell populations (B). 1261 cells (29%) most closely fit undifferentiated neuroepithelial cells with strong expression of HES3 and NR2F1 (C). Though I also see POU3F2, which is commonly associated with midbrain development, since I do not see midbrain markers later in differentiation, it is possible that this cell state occurs prior to brain region specification. The largest fraction of cells at this time point, 2593 cells (59%), exhibited the radial glia markers SOX2, EMX2, NNAT, PTN, and TLE4 (D). 356 cells (8%) were found to have transcriptional profiles consistent with Cajal-Retzius cells predominantly expressing TBR1, EOMES, LHX9, and NHLH1 (E).

et al., 2010; Qureshi et al., 2010) and many may be indicative of specific cell states (Guttman, 2009). A recent single cell RNA-sequencing study in fetal brain has confirmed these results in more mature neural tissue and shown that lncRNAs have higher expression in specific cell clusters than would it appear from bulk RNA-sequencing (Liu et al., 2016). Indeed, I find this to be the case within our own data where TrEx lncRNAs make up some of the top distinguishing genes between cell type clusters in week 2 neurospheres (Figure 4.3 and 4.4). In particular, I identified 8 TrEx lncRNAs that are expressed in single cell data and are predominantly expressed in one cell type. TrEx108 (CATG00000005887), TrEx2174 (RP11-314P15), TrEx2578 (NR2F2-AS1), and TrEx8168 (MIR219-2) are all most highly expressed in neuroepithelium and absent in week 5 single cell data (Figure 4.4A). TrEx5008 (RP11-71N10) and TrEx6514 (CATG00000085368) are predominantly restricted to the large radial glia cluster at week 2 (Figure 4.4B). Interestingly, TrEx5008 appears to be transiently expressed in bulk RNA sequencing data but is still highly expressed in a subset of radial glia in week 5 single cell RNA sequencing data (Figure 4.5) indicating that some TrEx lncRNAs may be longer lived than what is implied by bulk RNA-sequencing and are in fact restricted to one cell subtype that may be overtaken by others as more cell types are generated. TrEx4039 (AC011306/MIR217HG locus) is lowly expressed in both neuroepithelium and radial glia, but most concentrated in a portion of Cajal-Retzius cells, indicating

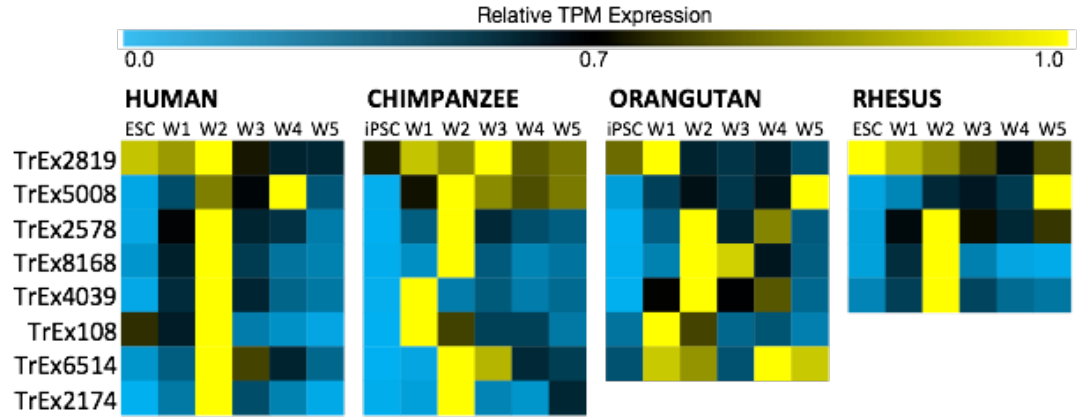

Figure 4.3: Selected TrEx lncRNAs. A selection of 8 TrEx lncRNAs which were analyzed by CRISPRa. lncRNAs were chosen by their transient expression patterns in human conservation in at least one other species. Heatmaps of their relative expression is displayed in this heatmap. Generally, the transient expression pattern is well conserved in species that preserve the transcript structure.

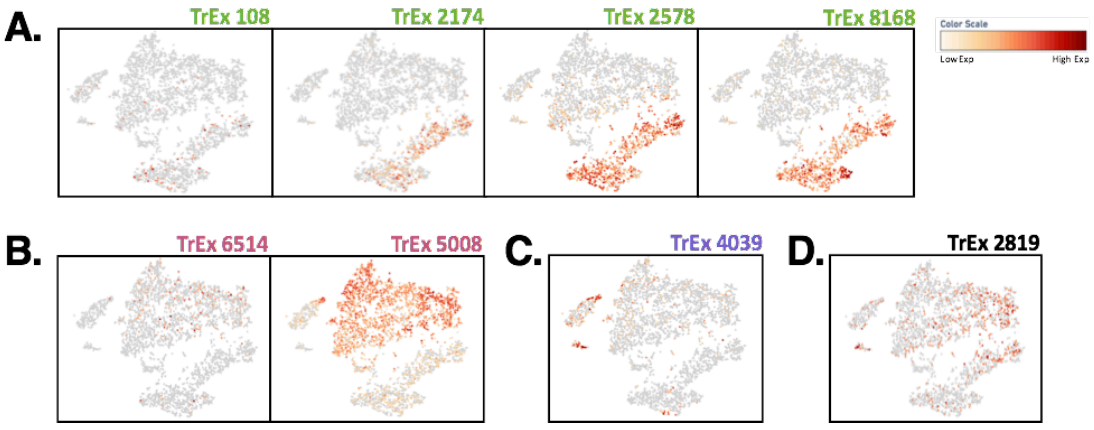

Figure 4.4: TrEx lncRNAs are cell cluster specific. TrEx lncRNAs segregated cleanly into major cell types (A, B, C). (A) TrEx108, TrEx2174, TrEx2578, and TrEx8168 were all expressed almost entirely within neuroepithelial cells. (B) Both TrEx6514 and TrEx5008 were restricted mainly to radial glia. (C) TrEx4039 is expressed primarily in a subpopulation of the Cajal-Retzius cluster. (D) TrEx2819 was expressed highly in both radial glia and neuroepithelium, but appeared most highly expressed in dividing cells.

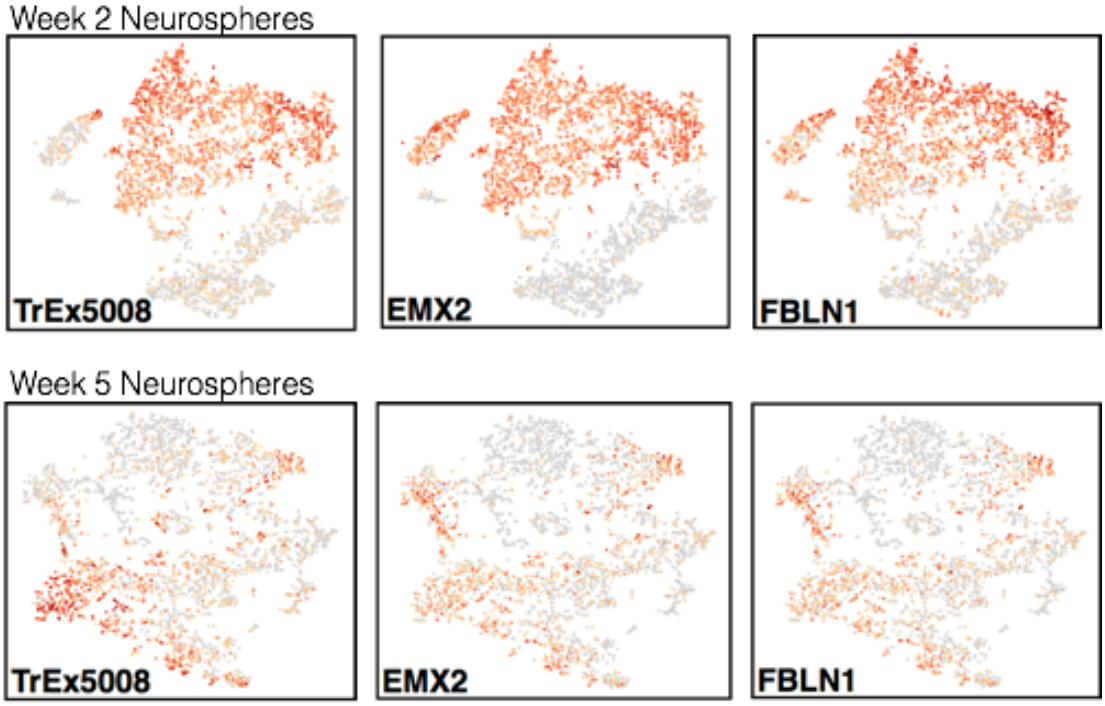

Figure 4.5: Persistence of TrEx5008 in immature radial glia cell. t-SNE plots of single cell RNA sequencing data is displayed from human week 2 and week 5 neurospheres. High expression of the indicated genes is shown in orange-red. Though TrEx5008 appears transiently expressed in bulk data, it seems to track with a cell subpopulation that is high in EMX2 at weeks 2 and 5.

that they may be implicated in differentiation, a specific cell subtype, or a cell state transition (Figure 4.4C). TrEx2819 (AC004158) perhaps had the most dispersed expression pattern of the TrEx lncRNAs I investigated, but it seems more highly expressed in dividing cells in both radial glia and neuroepithelial cell clusters (Figure 4.4D).

4D. TrEx lncRNA gene network correlations

Specific lncRNAs are known to have diverse gene regulatory functions including scaffolding of transcription factors (Rinn et al., 2007; Nagano et al., 2008; Pandey et al., 2008; Zhao et al., 2008; Khalil et al., 2009; Kozoil and Rinn, 2010; Zhao et al., 2010), microRNA sponges (Rani, 2016), and establishment of enhancer regions (DeSanta et al., 2010; Kim et al., 2010). In order to deduce potential regulatory networks and begin to identify potential mechanisms of action of functional TrEx lncRNAs, I used the single cell RNA sequencing data to find genes correlated and anti-correlated with TrEx lncRNA expression. This was approached in two ways: identifying active gene networks in cells that express the lncRNA and pairwise Pearson correlation of lncRNA expression to all other expressed genes.

Genes that were significantly correlated with cells expressing TrEx lncRNAs were identified in the Loupe Cell Browser v1.0.0 (10X Genomics)

(Figure 4.6). Cells expressing a TrEx were filtered and the "locally distinguishing genes" function was used to determine gene networks that were most associated with the lncRNA's native cell state. In Figure 4.6, the examples show genes associated with cells co-expressed with TrEx2174 and TrEx4039. The average expression of each gene across all the cells is compared to those that lack the lncRNA. Those with the most significant differences, referred to as either correlated with the cell type or anti-correlated, are represented in the heatmaps below the t-SNE plots. TrEx2174, being most associated with NE cells, correlates best with PDPN, MGST1, NR2F2, and HES3 (Figure 4.6B). TrEx4039 correlates with multiple cell types, but is strongest in Cajal-Retzius cells, exhibiting strong correlations with NEUROG1 and NHLH1 (Figure 4.6A).

Direct pairwise gene correlations with TrEx lncRNAs was also attempted in a single time point at week 2 of differentiation and over all time points (supplemental file: Table_S1.xlsx). Due to the inflated dispersion estimates on such a large set of data points, most pairwise comparisons appeared significant by 2-tailed t-test. For this reason, only the top and bottom 10 correlation values were considered for the functional assay described in the next section.

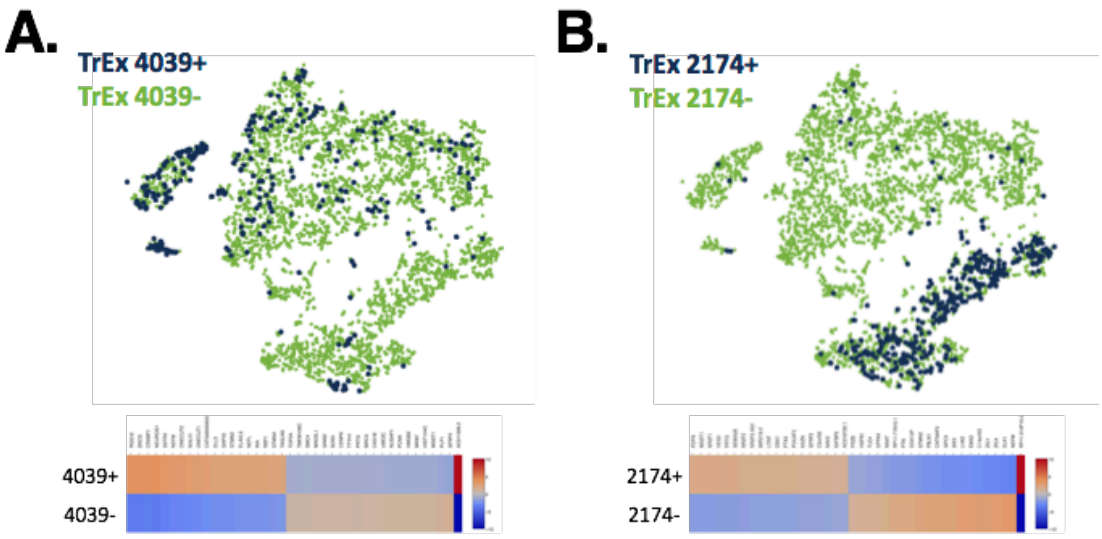

Figure 4.6: Single cell RNA sequencing gene correlations. Example t-SNE plots exhibiting cells expressing TrEx4039 (A) and TrEx2174 (B) are provided. Gene correlations and anti-correlations with these selected cells was determined by the Loupe Cell Browser locally distinguishing genes function. The resulting heatmaps of the average gene expression among all cells expressing each lncRNA for the top 20 and bottom 20 correlated genes are displayed below the t-SNE plots, respectively.

4E. Activation of lncRNAs in HEK293FT regulates gene expression

Since many lncRNAs have been implicated in gene regulatory function in either cis (Leighton et al., 1995; Penny et al., 1996; Pandey et al., 2008; Zhao et al., 2008; DeSanta et al., 2010; Kim et al., 2010; Orom et al., 2010; Wang et al., 2011) or trans (Rinn et al., 2007; Nagano et al., 2008; Pandey et al., 2008; Guttman et al., 2009; Khalil et al., 2009; Huarte et al., 2010; Kozoil and Rinn 2010; Loewer et al., 2010; Zhao et al., 2010; Guttman et al., 2011; Hung et al., 2011) I assessed the potential gene regulatory function of TrEx lncRNAs by CRISPR activation (CRISPRa) using dCas9-VP64 to drive transcription from the endogenous locus out of context in HEK293FT cells (Konermann et al., 2014). TrEx lncRNA candidates that were differentially expressed in bulk RNA sequencing data, detectable by single cell RNA sequencing, had little to no expression in HEK293FT cells, and appeared to be expressed from independent promoters in our bulk RNA sequencing coverage were selected for endogenous locus activation. Priority was given to lncRNAs that were max expressed at week 2 by bulk RNA sequencing or in the top 20 locally distinguishing genes expressed in our week 2 manually curated single cell RNA sequencing. 4 or 5 CRISPR single-guide RNAs (sgRNAs) were designed 50 to 450bp upstream from each candidate TrEx lncRNA and co-transfected into HEK293FT with dCas9-VP64 to drive transcription from the endogenous locus. A plasmid containing MS2, p65, and HSF1 was included to amplify PolII recruitment to the locus. Transfected

cells were selected by puromycin at 24 hours and harvested at 48 hours post-transfection (Figure 4.7). In all, I achieved activation of 8 TrEx lncRNAs ranging from 2.5-fold to 8600-fold activation over non-targeting scrambled sgRNA controls, four of which were activated to a similar or higher expression level compared to bulk week 2 human cortical organoid RNA.

To assess the regulatory potential of these activated TrEx lncRNAs, I used quantitative reverse transcriptase PCR (qPCR) to measure the expression of protein-coding genes that were correlated or anti-correlated with target TrEx lncRNA expression in single cell data as potential targets for lncRNA mediated regulation. A gene was considered correlated if either it was expressed in the same cell type in single cell RNA sequencing or was in the top 10 scores by Pearson correlation. Similarly, a gene was considered anti-correlated if it predominantly appeared in another cell type cluster or was in the bottom 10 scores by Pearson correlation. I tested expression changes of these genes upon CRISPRa of their respective TrEx lncRNAs compared to a set of scrambled sequence non-targeting guide controls (Figure 4.8, Figure 4.9, Figure 4.10, Figure 4.11, and Figure 4.13). If a TrEx lncRNA loci was located within 200kb of a protein coding gene or its nearest neighbor exhibited cluster-specific expression in single cell RNA-seq, I also tested adjacent gene expression to test for cis regulatory action of the lncRNA or potential off-target effects of dCas9-VP64. I found that all 8 activated TrEx lncRNAs had significant effects on gene expression of at least one of its potential targets with a p-value of less than 0.05

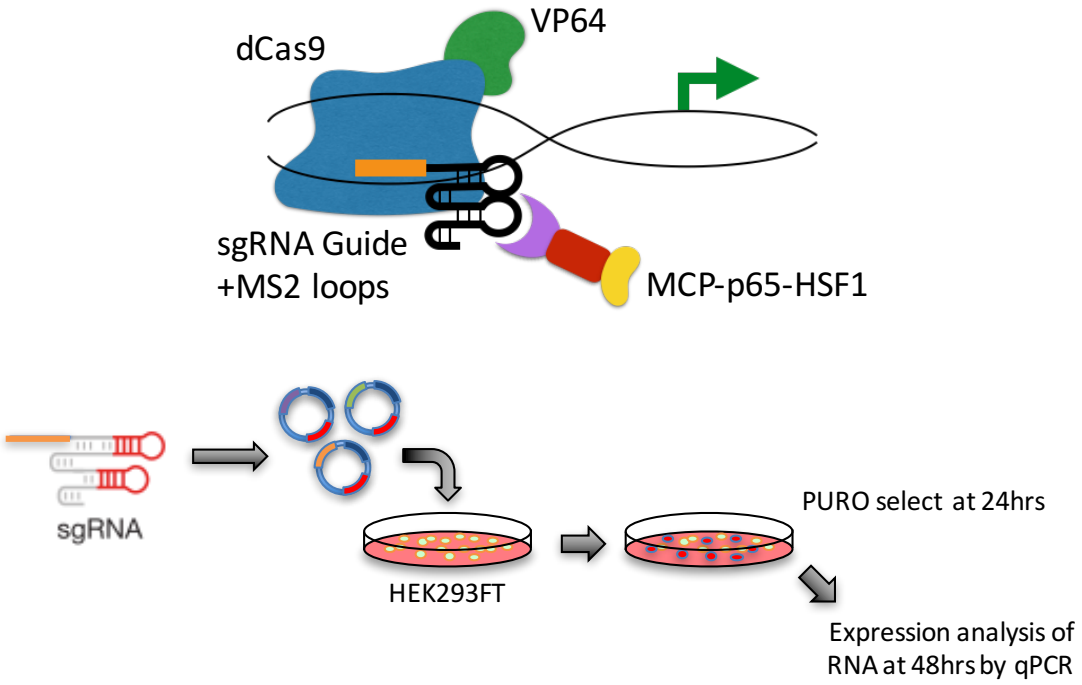

Figure 4.7: CRISPR activation assay. A cartoon depicting the experimental set up of the CRISPR activation assay. dCas9 lacking endonuclease activity fused to a VP64 activating domain is guided 50-450bp upstream of a target TrEx lncRNA. Small guide RNAs (sgRNAs) also included 2 MS2 loops for coordinating with supplied MCP-p65-HSF1 fusion protein to further boost activation of the target lncRNAs. A combination of 4 or 5 sgRNA guides was co-transfected into HEK293FT cells and selected with puromycin prior to harvest at 48 hours.

in a two-tailed t-test. Though most showed only modest effects on distal gene regulation, 3 induced a clear 2-fold change or greater on gene expression in *trans*.

The two TrEx lncRNAs associated with radial glia cells, TrEx5008 and TrEx6514, generally activated genes that correlated well with their expression in single cell RNA sequencing (Figure 4.8 and Figure 4.9A). TrEx5008 in particular exhibited significant activation of EMX2, increasing its expression nearly 2-fold over non-targeting controls and a more modest, but significant, effect on FBLN1 (1.5 fold) (Figure 4.8). Interestingly, TrEx5008 expression continues to track with EMX2 and FBLN1 in week 5 neurospheres within a subset of radial glia (Figure 4.5) further supporting an association with the same gene regulatory network. There is also a significant repression of the upstream gene ZIC1 (-1.6-fold) upon TrEx5008 activation (Figure 4.8), but since this gene is over 500kb upstream it is unlikely due to direct dCas9-VP64 interference, so it may instead imply a local chromatin structure change influenced by the expression of TrEx5008 itself.

The Cajal-Retzius associated lncRNA TrEx4039 was activated to a level 2.5-fold over bulk week 2 organoids and exhibited a 66% elevation in NEUROD1 along with a decrease of about 40% in MAB21L1 (Figure 4.10A, B). While these effects are modest, it correlates well with the heterogeneous distribution of MAB21L1 and NEUROD1 within the Cajal-Retzius cell cluster (Figure 4.10C). Indeed, expression of TrEx4039 is concurrent with NEUROD1 but segregates

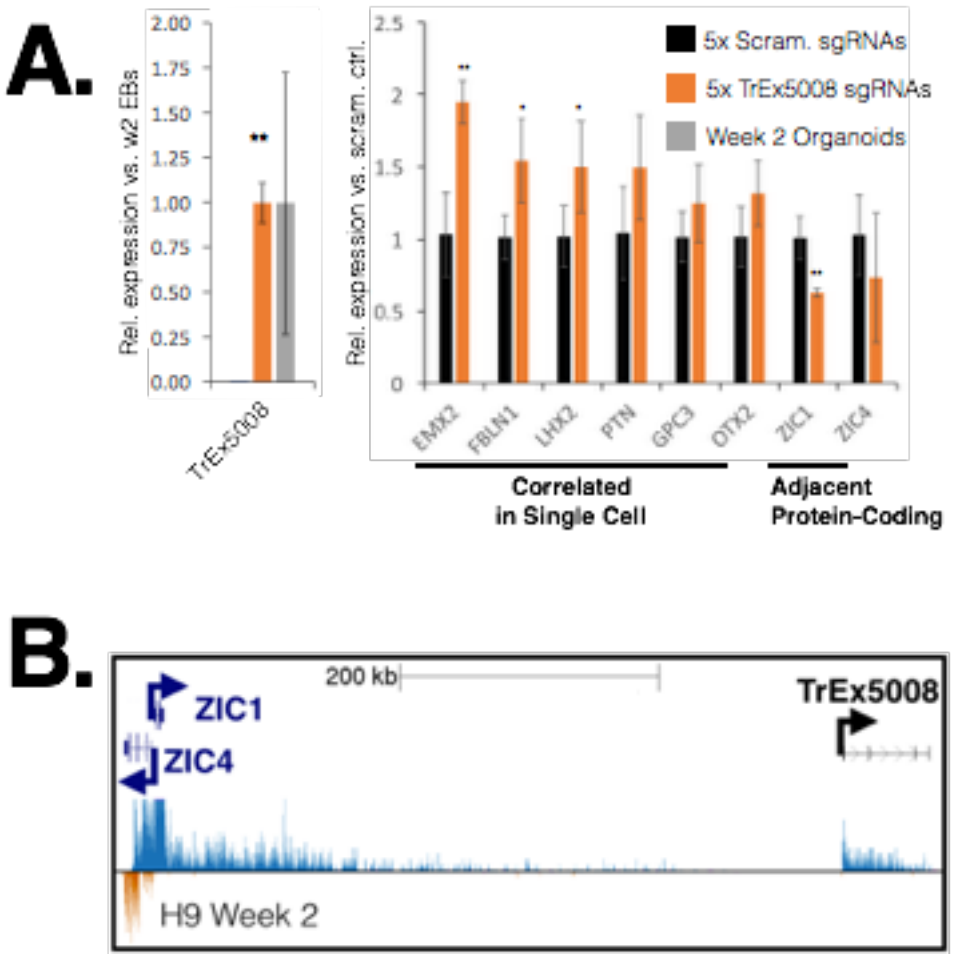

Figure 4.8: CRISPRa of TrEx5008. (A) qRT-PCRs of TrEx5008 lncRNAs and its correlated genes from single cell RNA-sequencing upon CRISPRa in HEK293FT cells. * indicates $p<0.05$ and ** $p<0.01$ versus scrambled non-targeting controls. Expression of TrEx5008 lncRNA in human week 2 cortical organoids is also depicted to indicate normal expression levels of the gene. (B) A UCSC Genome Browser screenshot depicting a representative Cufflinks gene model of TrEx5008 and its nearest protein-coding genes. Coverage tracks from human week 2 bulk RNA-seq show positive strand reads represented in blue and negative strand in orange.

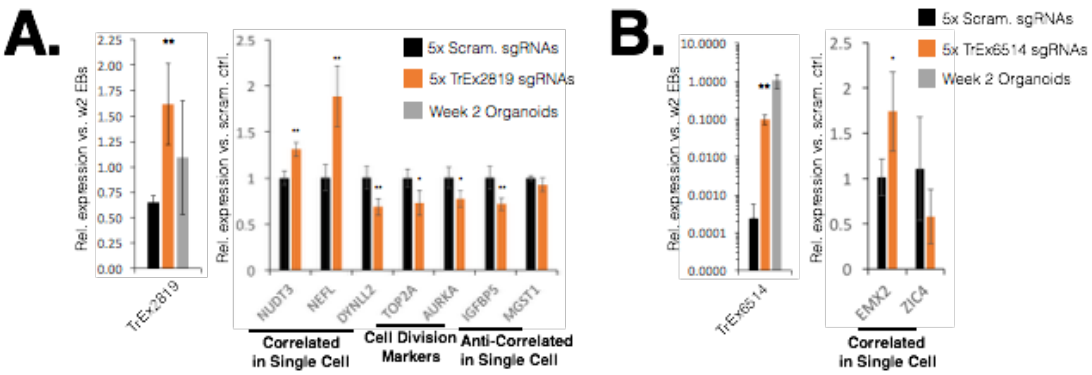

Figure 4.9: CRISPRa of TrEx2819 and TrEx6514. qRT-PCRs of target lncRNAs and their correlated genes from single cell RNA-sequencing upon CRISPRa of TrEx2819 (A) and TrEx6514 (B) in HEK293FT cells. * indicates $p<0.05$ and ** $p<0.01$ versus scrambled non-targeting controls. Expression of each TrEx lncRNA in human week 2 cortical organoids is also depicted to indicate normal expression levels of the gene.

from cells expressing MAB21L1, implicating that TrEx4039 may contribute to the establishment or maintenance of a specific Cajal-Retzius cell subtype or cell state.

TrEx2819 had the weakest change in expression versus the non-targeting control with about 2.48-fold activation, likely due to this transcript having low expression in unperturbed HEK293FT cells, though I achieved a similar level of expression to that observed in week 2 organoids (Figure 4.9B). I hypothesized a role in cell division by its expression pattern in the single cell RNA sequencing data, but I saw only modest down-regulation of TOP2A and AURKA upon CRISPRa of TrEx2819. I did see a significant increase of two genes found by Pearson correlation, NUDT3 (1.3-fold) and NEFL (1.9-fold), and down-regulation of the correlated DYNLL2 (-1.5-fold). TrEx2819 action on cytoskeletal proteins may indicate a role for it in cell structure reorganization pre- or post-cell division, but the data are not strong enough to support this at this point.

Though many of the activated TrEx lncRNAs had the effect of up-regulating positively correlated genes (i.e. TrEx5008, TrEx6514, TrEx4039, and TrEx2819) many had the opposite effect where genes positively correlated in single cell data were generally down-regulated upon activation (TrEx108, TrEx2174, TrEx8168). Curiously, these were all associated with NE and specifically expressed at week 2 as seen by both single cell and bulk RNA-seq. Specifically, activation of TrEx8168 significantly reduces HES3 expression by

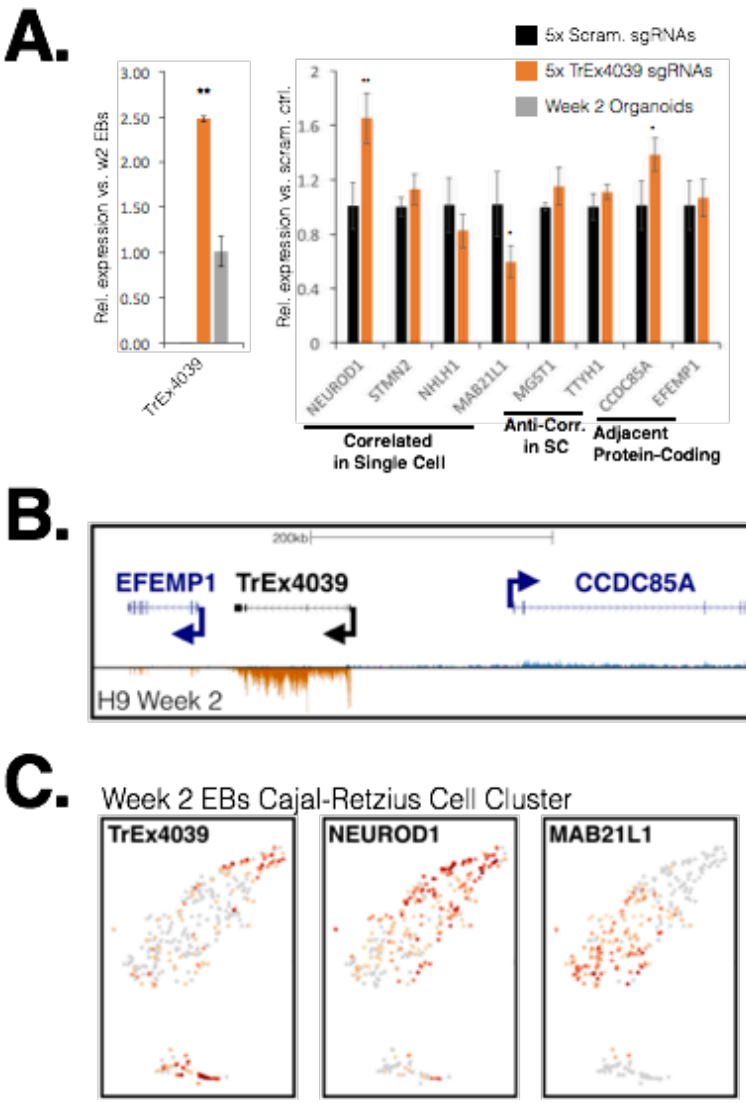

Figure 4.10: CRISPRa of TrEx4039. (A) qRT-PCRs of target TrEx4039 lncRNA and its correlated genes from single cell RNA-sequencing upon in HEK293FT cells. * indicates p<0.05 and ** p<0.01 versus scrambled non-targeting controls. Expression of TrEx4039 in human week 2 cortical organoids is also depicted to indicate normal maximum expression level of the gene. (B) AUCSC Genome Browser screenshots depicting a representative Cufflinks gene model of TrEx5008 and its nearest protein-coding genes. Coverage tracks from human week 2 bulk RNA-seq show positive strand reads represented in blue and negative strand in orange. (C) t-SNE plots focusing on the Cajal-Retzius cell cluster in week 2 neurosphere single cell RNA sequencing display the expression of TrEx4039, NEUROD1, and MAB21L1.

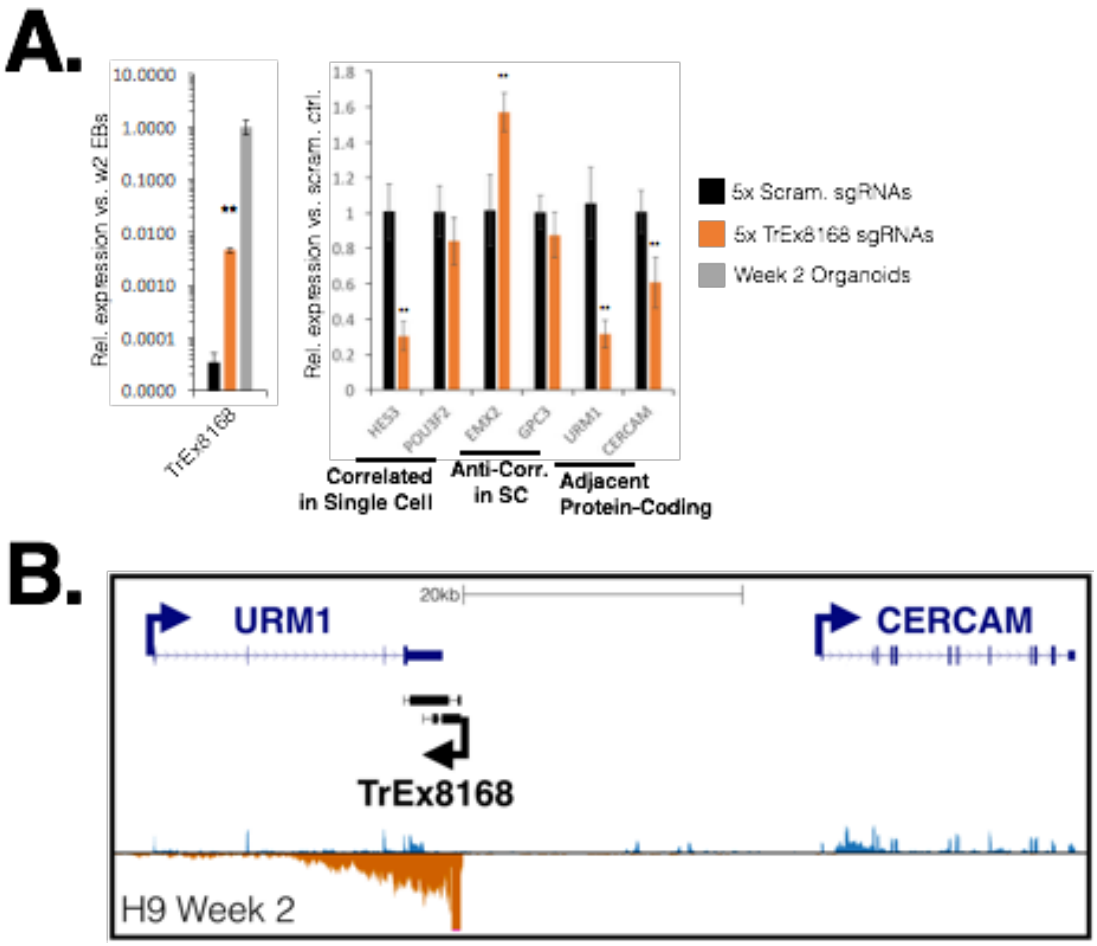

Figure 4.11: CRISPRa of TrEx8168. (A) RT-qPCRs of target TrEx8168 and its correlated genes from single cell RNA-sequencing upon CRISPR activation in HEK293FT cells. * indicates $p<0.05$ and ** $p<0.01$ versus scrambled non-targeting controls. Expression of TrEx8168 in human week 2 cortical organoids is also depicted to indicate normal maximal expression level of the gene. (B) A UCSC Genome Browser screenshot depicting a representative Cufflinks gene model of TrEx8168 and its nearest protein-coding genes, URM1 and CERCAM, is shown. Coverage tracks from human week 2 bulk RNA-seq were provided with positive strand reads represented in blue and negative strand in orange.

about 70% and up-regulates the anti-correlated EMX2 by almost 60% even with less than a tenth of the expression seen in week 2 organoids (Figure 4.11). TrEx8168 also seems to significantly disrupt the local transcriptional landscape, reducing the expression of its neighboring genes CERCAM by 40% and URM1 by 68% even with less than 0.5% the level of expression as week 2 organoids (Figure 4.11). URM1 appears to have a low dispersed expression pattern in single cell RNA sequencing but CERCAM does seem to be co-expressed though not significantly correlated by Pearson correlation with TrEx8168 (Figure 4.12).

TrEx108 exhibited largely repressive effects on genes that were both correlated and anti-correlated with its expression in single cell RNA-seq (Figure 4.13A). Its strongest effect was on PAPLN, an extracellular matrix protein involved in cell adhesion.

We achieved relatively low activation of TrEx2174 (Figure 4.13B) and TrEx2578 (Figure 4.13C) compared to the expression in week 2 organoids, yet still saw effects on associated genes. TrEx2174 reduced the correlated POU3F2 by about 25% suggesting a role in the exit from NE cells. TrEx2578, despite being antisense to NR2F2, increased its expression by over 1.4-fold. However, these effects are more modest than those seen at the other loci, and may not suggest function.

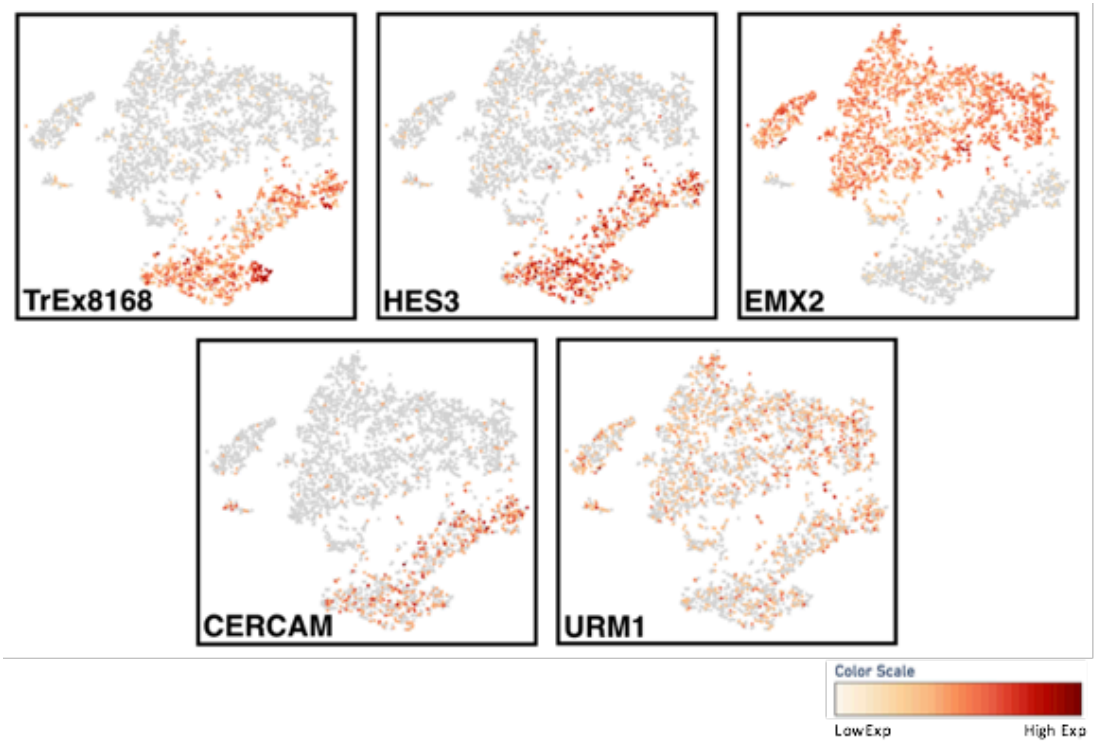

Figure 4.12: TrEx8168 in week 2 cortical organoids. t-SNE plots displaying expression of TrEx8168, HES3, EMX2, CERCAM and URM1 from single cell RNA-seq of week 2 human neurospheres illustrate cell cluster association.

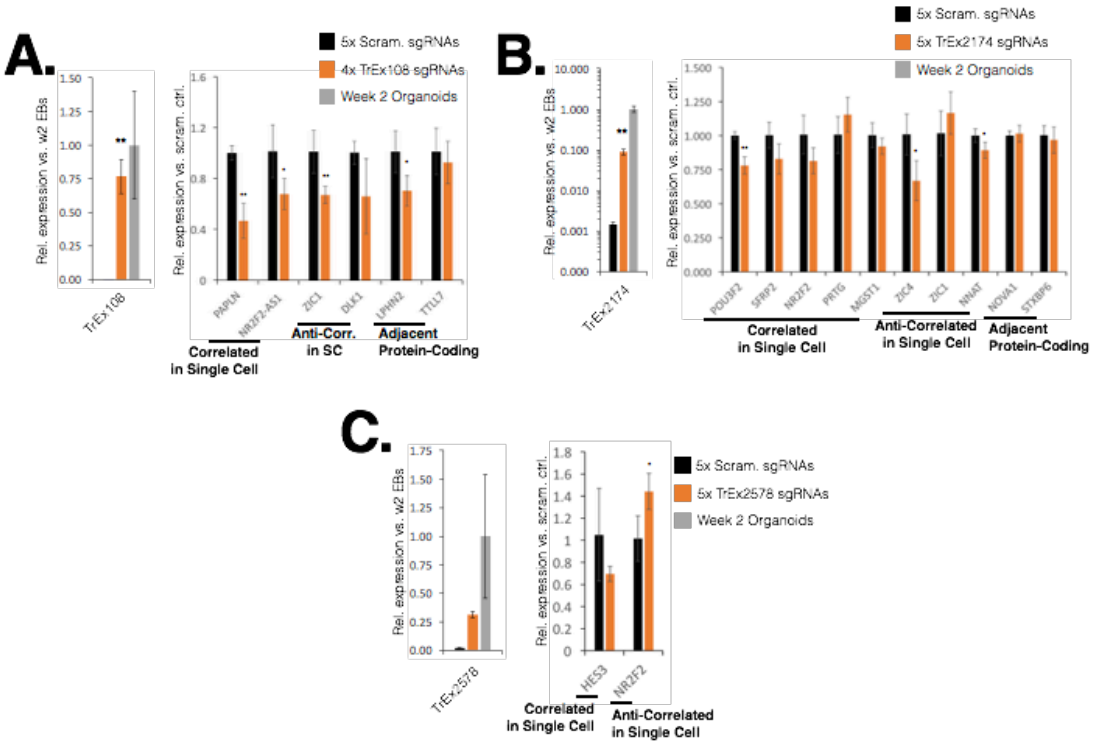

Figure 4.13: CRISPRa of TrEx108, 2174, and 2578. RT-qPCRs of target lncRNAs and their correlated genes from single cell RNA-sequencing upon CRISPRa of TrEx108 (A), TrEx2174 (B), and TrEx2578 (C) in HEK293FT cells. * indicates $p<0.05$ and ** $p<0.01$ versus scrambled non-targeting controls. Expression of each TrEx lncRNA in human week 2 cortical organoids is also depicted to indicate normal expression levels of the gene.

Chapter 5

Discussion and Conclusions

5. Discussion and Conclusions

The RNA-Seq data generated in this study provides a valuable resource for comparative studies aimed at understanding human, chimpanzee, orangutan, and rhesus cortical development. These tissues provide insight into early differentiation stages that are largely inaccessible *in vivo* and could shed light on what makes great apes and humans unique from each other and from other species. Further, while chimpanzee, human, and rhesus cortical neuron differentiation has been studied with neural organoids, to our knowledge, we provide the first look at orangutan early cortical neuron differentiation events. This system allows us to look at equivalent time points during this important stage of development in a protocol that is robust in all of our pluripotent stem cell lines and species. Pairing weekly bulk RNA-seq across species with analysis of the cell type composition of these heterogeneous cultures by single cell RNA-seq in human provides additional insight into the context of expression events during the formation of these early neural cell types. Here we have gone deeper into a small subset of the functional gene regulatory elements influencing primate cortical development that might be found in these data. Many more

could be discovered and studied by further analysis of this comprehensive RNA-seq data set.

The lncRNA field has been mired in controversy over the functional relevance of the tens of thousands of identified transcripts in human (Gingeras 2012; Kowalczyk et al., 2012; Hon et al., 2017), with claims that, despite a few notable exceptions, most represent non-functional transcription from enhancer elements or spurious transcriptional noise (Ponjavic et al., 2007; Struhl 2007; De Santa et al., 2010) due to their low sequence conservation across vertebrates (Wang et al., 2004; Babak et al., 2005; Pang et al., 2005; Ponjavic et al., 2007; Church et al., 2009; Kutter et al., 2012) or low levels of expression in bulk tissues (Cabili et al., 2011). It has also been suggested that tissue-specific lncRNAs are often less conserved than those expressed in multiple tissues (Ulitsky 2016). We have shown here, however, that many human lncRNA transcripts expressed during the early stages of cortical neuron differentiation have structural conservation over great apes or to old world monkeys. Of the 2,975 lncRNAs expressed over our time course in human, 72% had conserved structure through chimp, 58% through orangutan, and 43% through rhesus. 51% were conserved in both great ape species and 31% had evidence of conserved structure in all species, much greater than the vanishingly small estimates of sequence conservation even of functional lncRNAs through mouse (Babak et al., 2005; Pang et al., 2006; Ponjavic et al., 2007). Striking among these transcripts were those that were specifically expressed at single time points in human, which we

term TrEx lncRNAs. 386 of these TrEx lncRNAs were observed in human and had a remarkably conserved expression pattern in great ape species with at least 223 (58%) exhibiting a preserved TrEx pattern in chimpanzee or orangutan. While transient expression patterns seemed far less conserved than exonic structure, it is possible we are under-sampling relevant time points in each of our species for optimal detection of these lncRNAs, especially considering most TrEx lncRNAs are primarily expressed at a single time point.

I found that many of these conserved transiently expressed transcripts were associated with specific cell types by single cell RNA-seq, especially at our week 2 time point where there was a clear distinction between radial glia, neuroepithelium, and Cajal-Retzius cells. Using a strict definition of transcript conservation, requiring both gene intron boundary and expression pattern conservation between human and at least one other species, we set out to find transiently expressed lncRNAs associated with specific cell subtypes, reasoning that those have the highest likelihood of gene regulatory function. In all, 8 TrEx lncRNAs that appeared independently regulated in bulk RNA-seq with independent promoters from neighboring genes were selected for CRISPRa, allowing for detection of both cis and trans regulatory function. A recent study has shown that cellular context is vitally important for lncRNA function (Liu et al., 2017), so it is likely that the full functionality of these genes is significantly under-sampled in our HEK293FT-based CRISPRa assay. Even still, we see effects on distal genes upon endogenous activation of these loci indicating robust gene

regulatory function even out of its normal biological context, which warrants further study. All tested TrEx lncRNAs showed effects on distal gene regulation upon activation, but only 3 showed effects of 2-fold change or greater on genes correlated or anticorrelated with their expression in single cell RNA-seq data.

TrEx5008, which is highly expressed in most radial glia at week 2, induces a 2-fold increase in EMX2, an essential gene for dorsal telencephalon differentiation whose expression in neural precursors is proposed to lead to patterning of the neocortex into distinct functional areas (Hamasaki et al., 2004). Interestingly, though it appears transiently expressed in bulk RNA sequencing data, TrEx5008 continues to be expressed in a fraction of radial glia at week 5 tracking with EMX2 expression, further suggesting its association with a subpopulation of radial glia with high EMX2.

TrEx108 and TrEx8168, both associated with neuroepithelial cells, both showed max expression at week 2 and had incredibly conserved expression patterns in other species (Fig. 4B,C). Both showed repressive effects on genes correlated with their expression in single cell data when activated out of context in HEK293FT cells (Fig. 6A,D). This seemingly contradictory result may be justified by taking into account the timing of the expression of these elements and their max expression at a time when neuroepithelium gives rise to radial glia. In particular, TrEx8168, even at expression levels below 0.5% that of bulk neurospheres, significantly disrupted transcription of its two neighboring genes URM1 and CERCAM suggesting a drastic change to its local chromatin. TrEx8168

activation also resulted in a large reduction of the distal gene HES3 essential for the establishment and maintenance of neuroepithelial cells paired with a modest upregulation of the radial glia associated EMX2. Together, the sharp conserved expression pattern, repressive effects, and absence of neuroepithelial markers at later time points support the model of these lncRNAs marking the transition from neuroepithelium to radial glia.

Finally, TrEx4039, although showing only modest effects on distal genes, suggests that some of these transiently expressed lncRNAs may be involved in the establishment of specific cell types. HEK293FT cells expressing TrEx4039 had elevated NEUROG1 and depressed MAB21L1 which matches well with the expression of these two genes in relation to the lncRNA in week 2 single cell RNA sequencing. Week 5 neurospheres exhibit a much lower and more diffuse expression pattern of TrEx4039 (data not shown). Together, this would suggest that this lncRNA is either involved in a transient cell state on the way to making mature Cajal-Retzius cells or a short-lived cell subpopulation.

We have demonstrated that lncRNAs expressed during early cortical neuron development are well conserved over short evolutionary distances in exonic structure and, to a lesser extent, expression pattern. We found examples of transiently expressed lncRNAs with plausible roles in cell subtype specification and maintenance (TrEx5008), exiting progenitor states (TrEx108 and TrEx8168), and cell type establishment after differentiation (TrEx4039) using an assay that measures both cis and trans regulatory effects. The approach

outlined in this paper provides a means to identify promising candidates for in depth mechanistic studies to establish the complete functional roles for lncRNAs. The findings here only scratch the surface of our comparative bulk RNA and single cell RNA sequencing data sets. With thousands of expressed intergenic RNA elements, including hundreds of species-specific and cell-type specific elements, further interrogation of our matched time points from human, chimpanzee, orangutan, and rhesus bulk RNA sequencing is likely to yield many more functional elements involved in primate cortical development, including protein-coding genes as well as lncRNAs, facilitating exploration of gene networks and biological processes that may provide important insights into the molecular mechanisms underlying cortical development.

Appendix

Methods

A. Methods

A.1 iPSC generation

Primate primary fibroblasts were grown as adherent cultures in MEM Alpha (ThermoFisher) supplemented with 10% Gibco FBS (ThermoFisher) and 1% Pen-Strep (ThermoFisher). Integration-free chimpanzee induced pluripotent stem cells were produced at Applied StemCell from S008919 primary fibroblasts (Yerkes Primates, Coriell) by episomal reprogramming using the Y4 plasmid combination described in Okita et al., 2011. Integration-free Sumatran orangutan induced pluripotent stem cells were generated using the CytoTune 2.0 Sendai Reprogramming kit (ThermoFisher) from 11045-4593 primary fibroblasts obtained from the Frozen Zoo® (http://institute.sandiegozoo.org/resources/frozen-zoo%C2%AE). Both chimpanzee and orangutan iPSCs were initially established on mouse embryonic fibroblasts with KSR-15 (KO DMEM/F-12 + 20% KOSR, 1% NEAA, 1% GlutaMAX, 1% Pen-Strep, and 0.1mM 2-mercaptoethanol supplemented with 15 ng/mL bFGF) media and were transferred to feeder-free conditions on Matrigel (Corning) with mTeSR-1 (Stem Cell Tech) for chimpanzee or vitronectin

(ThermoFisher) with Essential-8 Flex (ThermoFisher) for orangutan. Pluripotency was confirmed by immunofluorescence staining of pluripotency markers, RT-PCR, teratoma assay, and karyotype.

A.2 Teratoma Assay

Mice were anesthetized by intraperitoneal injection with 100mg/kg ketamine. 2 subcutaneous injections of 1 to 5 million cells suspended in 30% Matrigel (Corning) were made in the dorsolateral or ventral lateral areas of NOD-SCID mice (NOD.CB17-Prkdc^{scid}/NCrCrl, Charles River) similar to Prokhorova et al., 2009. Mice were observed for up to 12 weeks for the appearance of tumors in the injected areas. The animals were euthanized by cervical dislocation and teratomas were harvested, fixed in 4% paraformaldehyde, saturated in 30% sucrose in PBS, embedded in Tissue Freezing Medium™ (Triangle Biomedical Sciences), and frozen for cryostat sectioning. Sections of the tumors were stained with hematoxylin (Mayer's Hematoxylin Solution, Sigma) & eosin (Eosin Y solution, Sigma) and analyzed for the generation of all three germ layers.

A.3 Karyotyping

Chimpanzee and orangutan iPSC lines were confirmed to have a stable wildtype 48/XX karyotype through at least passage 32 or 36, respectively. Karyotyping services were performed by Cell Line Genetics or the Coriell Institute for Medical Research.

A.4 Cortical organoid generation

The Eiraku et al., (2008) protocol was optimized for use with human ESCs, rhesus ESCs, chimpanzee iPSCs, and orangutan iPSCs. Human H9 and rhesus LyonESC1 embryonic stem cells were cultured on mouse embryonic fibroblasts with KSR-8 media (KO DMEM/F-12 + 20% KOSR, 1% NEAA, 1% GlutaMAX, 1% Pen-strep, and 0.1mM 2-mercaptoethanol supplemented with 8ng/mL bFGF). Embryonic stem cells were manually lifted from MEF feeders and allowed to self-form into embryoid bodies on low attachment plates (Corning) in KSR media. Chimpanzee and orangutan induced pluripotent stem cells were grown in feeder-free conditions on matrigel (Corning) with mTeSR-1 (Stem Cell Tech) or vitronectin (ThermoFisher) with Essential-8 Flex media (ThermoFisher), respectively, and 10,000 cells per EB were aggregated using AggreWell-800 plates (Stem Cell Technologies) in Aggrewell media (Stem Cell

Technologies) supplemented with 10 uM Y-27632 rock inhibitor (ATCC) and transferred to low attachment plates (Corning) on day 2. Both methods supplemented the respective media with 500ng/mL DKK1 (Peprotech), 500 ng/mL NOGGIN (R & D Systemes), 10 uM SB431542 (Sigma), and 1 uM Cyclopamine V. californicum (VWR) for the first 18 days of differentiation. The media was changed to Neuralbasal (Invitrogen) supplemented with N2 (Gibco) and 1 uM Cyclopamine on day 18. At this time, chimpanzee and orangutan neurospheres were also supplemented with 10ng/mL bFGF and 10ng/mL EGF to improve survivability in Neuralbasal media. After day 26, all cultures were grown in Neuralbasal/N2 media without any added factors. Total RNA was extracted at weekly time points for each species. This timeline was adjusted accordingly in rhesus, harvesting on days 6, 11, 17, 22, and 28, to account for differences in gestational timing.

A.5 Immunofluorescence Staining

Cortical organoids were fixed for 15 minutes in 4% paraformaldehyde and saturated in 30% sucrose prior to being embedded in Tissue Freezing Medium™ (Triangle Biomedical Sciences) and frozen for cryostat sectioning. Sections were adhered to glass slides and fixed a second time in 4% paraformaldehyde for 10 minutes. Cells grown in 2-dimensional culture were

grown on acid etched coverslips coated in Matrigel (Corning) and fixed for 10 minutes with 4% paraformaldehyde prior to staining. Samples were incubated at 4°C in blocking solution (3% BSA and 0.1% Triton X-100 in PBS) for 4 hours. Primary antibody incubation was performed overnight at 4°C in blocking solution. Secondary antibody incubation was for 1-4 hours at room temperature in blocking solution. A list of antibodies used is provided in the supplementary materials file: MaterialsList.xlsx.

A.6 Primate Genome Alignment and Annotation

A progressive Cactus (Paten et al 2011) whole genome alignment was generated between the human hg19 assembly, chimpanzee panTro4 assembly, orangutan ponAbe2 assembly, and rhesus macaque rheMac8 assembly. This alignment was used as input to the Comparative Annotation Toolkit (https://github.com/ComparativeGenomicsToolkit/Comparative-Annotation-Toolkit, citation pending) along with the FANTOM5 (Hon et al., 2017) lv3 annotation set. This process projects the annotations from human to the other primates in the alignment. Subsequent filtering and post-processing produces a high-quality comparative annotation set. RNA-seq obtained from SRA were used to help guide the annotation process. Annotations provided in supplementary

materials: hg19.cufflinks.gtf (human), panTro4.cufflinks.gtf (chimpanzee), ponAbe2.cufflinks.gtf (orangutan), rheMac8.cufflinks.gtf (rhesus macaque).

A.7 RNA-Sequencing Analysis

Paired-end Illumina reads were trimmed from the 3' end of read1 and read2 to 100x100bp for human and rhesus libraries and 80x80bp for chimpanzee and orangutan based on sequence quality. Bowtie2 v2.2.1 (Langmead et al., 2012) was used with the "--very-sensitive" parameter to filter reads against the repeatMasker library (Smit et al., 2015) for each respective species which were removed from further analysis. STAR (v2.5.1b, Dobin et al., 2012) was used to map RNA-seq reads to hg19 (human, Genome Reference Consortium GRCh37, 2009), panTro4 (chimpanzee, CGSC Build 2.1.4, 2011), ponAbe2 (orangutan, WUSTL Pongo_albelii-2.0.2, 2007), and rheMac8 (rhesus macaque, Baylor College of Medicine HGSC Mmul_8.0.1, 2015) respective to the origin species. STAR was run with the default parameters with the following exceptions: --outFilterMismatchNmax 999, --outFilterMismatchNoverLmax 0.04, --alignIntronMin 20, --alignIntronMax 1000000, and --alignMatesGapMax 1000000. STAR alignments were converted to genomic position coverage with the bedtools command genomeCoverageBed -split.

DESeq2 v1.14.1 (Love et al., 2014) was used to provide basemean expression values and differential expression analysis across the time course in each species. Total gene coverage for a gene was converted to read counts by dividing the coverage by N+N (100+100 for human and rhesus and 80+80 for chimpanzee and orangutan) since each paired-end NxN mapped read induces a total coverage of N+N across its genomic positions.

A.8 LncRNA annotation analysis, structure conservation, and expression estimates

Cufflinks v2.0.2 suite (Trapnell et al 2010; Trapnell et al 2012) was used to assemble transcript predictions of potentially unannotated lncRNAs in each species and the Cuffmerge tool was used to combine these annotations with FANTOM5 transcripts. The resulting cufflinks-assembled and merged transcript sets were then projected through the cactus alignment (Stanke et al 2008) to each of the other three genomes. Guided by the cufflinks annotation set in each genome, these projections from the other genomes were assigned a putative gene locus. In cases where a projection overlapped multiple genes, the gene whose transcripts had the highest exonic Jaccard similarity were chosen. RSEM (v1.3.0, Li and Dewey 2011) was used to provide TPM expression values for these newly generated transcripts.

Expressed lncRNAs were assessed using the homGeneMapping tool from the AUGUSTUS toolkit (Konig et al 2016). homGeneMapping makes use of cactus alignments to project annotation features in all pairwise directions, providing an accounting of features found in other genomes. homGeneMapping was provided both the Cufflinks transcript assemblies as well as expression estimates derived from the combination of the week 0 to week 5 RNA-seq experiments in all four species. The results of this pipeline were combined with the above transcript projections to ascertain a set of lncRNA loci that appear to have human specific expression, human-chimp specific expression, great-ape specific expression, and expressed in all primates. For this analysis, a locus was considered expressed in the current reference genome if one or more transcripts had RNA-seq support for every single one of its intron junctions, and considered expressed in another genome if the transcripts that mapped from that genome to the current reference had RNA-seq support for any of its intron junctions. All single-exon transcripts were filtered out to reduce noise.

To eliminate the possibility of the specificity results being skewed by assembly gaps or alignment error, loci which appeared to have sub-tree specific expression were checked against the cactus alignment to ensure that there was a matching locus in each other genome. If a genome appeared to be missing sequence, then this locus was flagged as having incomplete information.

The Illumina sequences were trimmed then filtered by mapping to RepeatMasker (Smit et al., 2015) using bowtie2 v2.2.1 (Langmead and Salzberg

2012). Remaining RNA-seq reads were mapped using STAR v2.5.1b (Dobin et al., 2012) to hg19 (human, Genome Reference Consortium GRCh37, 2009), panTro4 (chimpanzee, CGSC Build 2.1.4, 2011), ponAbe2 (orangutan, WUSTL Pongo_albelii-2.0.2, 2007), and rheMac8 (rhesus macaque, Baylor College of Medicine HGSC Mmul_8.0.1, 2015) respective to the source species. In all, over 2 billion paired-end RNA-seq reads were uniquely mapped to their respective genomes from 49 libraries, averaging 41 million reads per library with a minimum of 46 million total reads across replicates per species time point. Cufflinks v2.0.2 (Trapnell et al 2010; Trapnell et al 2012) was used to assemble potential novel transcripts in each species and the Cuffmerge tool combined gene models across time points in each respective species using FANTOM5 lv3 (Hon et al 2017) as a reference annotation. The resulting annotations from each species were projected through Cactus alignment (Stanke et al 2008) to each of the other primate genomes. Guided by the cufflinks annotation set in each genome, these projections from the other genomes were assigned a putative gene locus. In cases where a projection overlapped multiple genes, the gene whose transcripts had the highest exonic Jaccard similarity were chosen.

A.9 3' Single Cell RNA-sequencing

Human H9 embryonic stem cells were grown on vitronectin with E8-Flex media (ThermoFisher). Neurospheres were aggregated and as described above above for chimpanzee and orangutan induced pluripotent stem cells. Single cell suspensions for 10X Genomics Chromium single cell RNA-sequencing were dissociated with TrypLE (ThermoFisher) and handled according to the 10X protocol RevA (version 1 chemistry) for undifferentiated hESCss and week 5 cortical organoids and RevB (version 2 chemistry) for weeks 1 and 2 cortical organoids. Cell count, quality, and viability was assessed using Trypan Blue (ThermoFisher) on a TC20 automated cell counter (BioRad). Single cell suspensions were made aiming for 1500-3000 cells captured per library. The data was analyzed by CellRanger (v1.2) using a custom annotation set based on FANTOM5 lv3 (Hon et al., 2017; supplemental materials: hg19_fantom_lv3_allF2.gtf) and visualized using the Loupe Cell Browser (v1.0.0).

A.10 CRISPRa assay

The CRISPR-activation (CRISPRa) assay was modified from Konermann et al 2014. HEK293FT cells were cultured with DMEM+GlutaMAX (ThermoFisher) supplemented with 10%FBS without antibiotic. Each well of a 6-well plate was

seeded with 500k cells and co-transfected at 60-70% confluence the next day using Xfect reagent (Takara) with dCas9-VP64_Blast (Feng Zhang, addgene #61425), MS2-p65-HSF1_Hygro (Feng Zhang, addgene #61426), and a combination of 5 custom guide RNAs per target in the custom plasmid 783 (Figure A.1) for a total of 7.5ug DNA in a ratio of 1:1:2 respectively. Transfected cells were selected at 24hrs by incubation with 2ug/mL puromycin until harvest. RNA was harvested at 48 hours after transfection using TRIzol reagent (ThermoFisher) and RNA was extracted using Direct-zol columns (ZYMO). Quantitect SYBR® Green RT-PCR (Qiagen) was used with 50ng of total RNA per reaction, 4 replicates per condition. Relative expression was calculated by ddCt normalized to HEK293FT transfection with non-targeting scrambled control guides (supplemental materials: MaterialsList.xlsx).

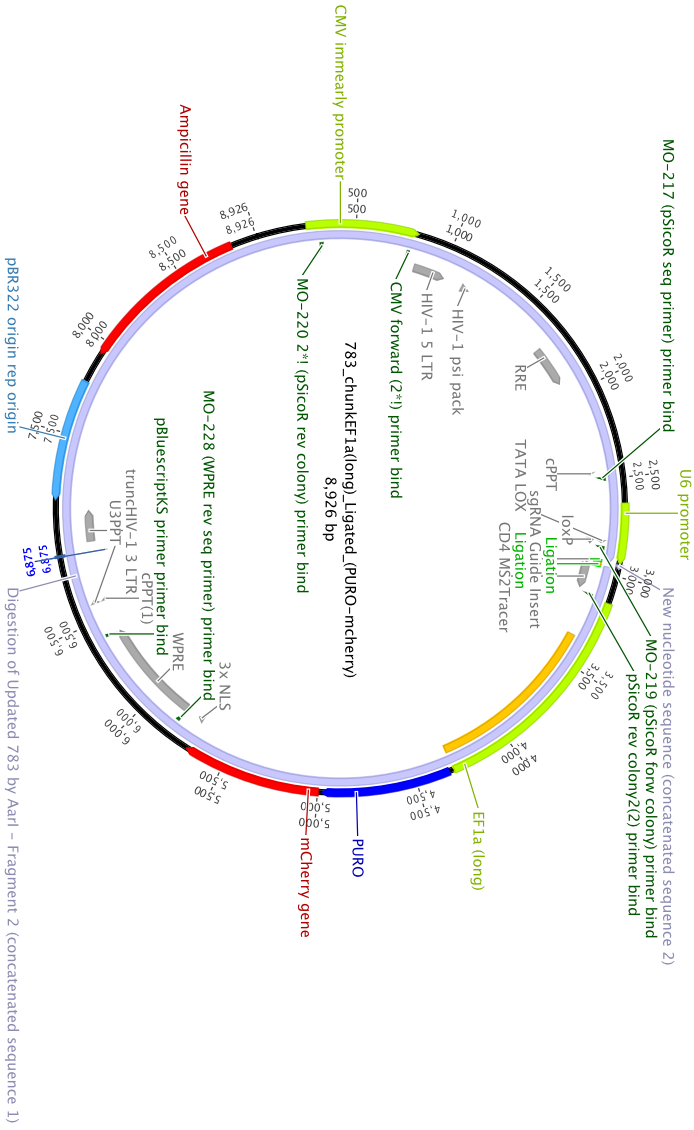

Figure A.1: 783 plasmid map. A map of the elements present in the 783 expression plasmid for CRISPR activation small guide RNAs.

Supplemental Files

MaterialsList.xlsx

Primer sequences, sgRNA guide sequences, primary and secondary antibodies list.

Table_S1.xlsx

Pairwise Peason correlations of lncRNAs to other genes in week 2 organoids. Each tab is one of our 8 target TrEx lncRNAs. Gene correlations with a p-value lower than 1E-09 are displayed.

hg19.cufflinks.gtf

Cufflinks based annotations used for human bulk RNA sequencing. Mapped in hg19.

panTro4.cufflinks.gtf

Cufflinks based annotations used for chimpanzee bulk RNA sequencing. Mapped in panTro4.

ponAbe2.cufflinks.gtf

Cufflinks based annotations used for orangutan bulk RNA sequencing. Mapped in ponAbe2.

rheMac8.cufflinks.gtf

Cufflinks based annotations used for rhesus macaque bulk RNA sequencing. Mapped in rheMac8.

hg19_fantom_lv3_allF2.gtf

Human FANTOM5 lv3 (Hon et al., 2017) based annotations used for bulk and single cell RNA sequencing analysis. Mapped in hg19.

panTro4_transmap_fantom_lv3_allF2.gtf

Human FANTOM5 lv3 (Hon et al., 2017) annotations lifted by transMap to panTro4 used for bulk and single cell RNA sequencing analysis.

ponAbe2_transmap_fantom_lv3_allF2.gtf

Human FANTOM5 lv3 (Hon et al., 2017) annotations lifted by transMap to ponAbe2 used for bulk and single cell RNA sequencing analysis.

rheMac8_transmap_fantom_lv3_allF2.gtf

Human FANTOM5 lv3 (Hon et al., 2017) annotations lifted by transMap to rheMac8 used for bulk and single cell RNA sequencing analysis.

Bibliography

Alexander, R., Fang, G. & Rozowsky, J. Annotating non-coding regions of the genome. Nat. Rev. 11, 559–571 (2010).

Amaral, P.P., and Mattick, J.S. Noncoding RNA in development. Mamm. Genome 19, 454–492 (2008).

Babak, T., Blencowe, B. J. & Hughes, T. R. A systematic search for new mammalian noncoding RNAs indicates little conserved intergenic transcription. BMC Genomics 6, 104 (2005).

Babbitt, C.C., Fedrigo, O., Pfefferle, A.D., Boyle, A.P., Horvath, J.E., Furey, T.S., and Wray, G.A. Both noncoding and protein-coding RNAs contribute to gene expression evolution in the primate brain. Genome Biol Evol 2, 67–79 (2010).

Ban, H, Nishishita, N, Fusaki, N, Tabata, T, Saeki, K, Shikamura, M, Takada, N, Inoue, M, Hasegawa, M, Kawamata, S, et al. Efficient generation of transgene-free human induced pluripotent stem cells (iPSCs) by temperature-sensitive sendai virus vectors. Proceedings of the National Academy of Sciences, 108(34):14234-14239 (2011).

Barker, N. et al. Lgr5+ve stem cells drive self-renewal in the stomach and build long-lived gastric units in vitro. Cell Stem Cell 6, 25–36 (2010).
Barnett, J.H., Smoller, J.W. The genetics of bipolar disorder. Neuroscience. 164(1), 331-343 (2009).

Ben-Nun, IF, Montague, SC, Houck, ML, Tran, HT, Garitaonandia, I, Leonardo, TR, Wang, Y, Charter, SJ, Laurent, LC, Ryder, OA. Induced pluripotent stem cells from highly endangered species. Nature methods, 8(10):829-831 (2011).

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S., et al. Global identification of human transcribed sequences with genome tiling arrays. Science 306, 2242–2246 (2004).

Brawand, D. et al. The evolution of gene expression levels in mammalian organs. Nature **478,** 343–348 (2011).

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., et al. The evolution of gene expression levels in mammalian organs. Nature 478, 343–348 (2011).

Britten, R.J. Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. Proc. Natl. Acad. Sci. U.S.A. 99, 13633–13635 (2002).

Buiting, K., Nazlican, H., Galetzka, D., Wawrzik, M., Groβ, S., and Horsthemke, B. C15orf2 and novel noncoding transcript from the Prader-Willi/Angelman syndrome region show monoallelic expression in fetal brain. Genomics 89, 588-595 (2007).

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 25, 1915–1927 (2011).

Caceres, M. et al. Elevated gene expression levels distinguish human from non-human primate brains. Proc. Natl Acad. Sci. USA 100, 13030–-13035 (2003).
Camp, J. G. et al. Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. Proc. Natl. Acad. Sci. U. S. A. 112, 15672–7 (2015).

Camp, J.G., Badsha, F., Florio, M., Kanton, S., Gerber, T., Wilsch-Bräuninger, M., Lewitus, E., Sykes, A., Hevers, W., Lancaster, M., Knoblich, J.A., Lachmann, R., Pääbo, S., Huttner, W.B., and Treutlein, B. Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. Proc. Natl. Acad. Sci. U.S.A. 112, 15672–7 (2015).

Camprubí, C. et al. Imprinting center analysis in Prader-Willi and Angelman syndrome patients with typical and atypical phenotypes. Eur. J. Med. Genet. 50, 11–20 (2007).

Cantalupo, C. & Hopkins, W. D. Asymmetric Broca's area in great apes. Nature 414, 505 (2001).

Cao, X., Yeo, G., Muotri, A.R., Kuwabara, T., and Gage, F.H. Noncoding RNAs in the mammalian central nervous system. Annu. Rev. Neurosci. 29, 77–103 (2006).

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, et al. The transcriptional landscape of the mammalian genome. Sci- ence 309: 1559–1563 (2005).

Carrieri, C., Cimatti, L., Biagioli, M., Beugnet, A., Zucchelli, S., Fedele, S., Pesce, E., Ferrer, I., Collavin, L., Santoro, C., et al. Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. Nature 491, 454–457 (2012).

Carroll, S.B. Genetics and the making of Homo sapiens. Nature 422, 849–857 (2003).

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science 308, 1149–1154 (2005).

Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437, 69–87 (2005).

Chodroff, R.A., Goodstadt, L., Sirey, T.M., Oliver, P.L., Davies, K.E., Green, E.D., Molnár, Z., and Ponting, C.P. Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. Genome Biol. 11, R72 (2010).

Church, D.M., Goodstadt, L., Hillier, L.W., Zody, M.C., Goldstein, S., She, X., Bult, C.J., Agarwala, R., Cherry, J.L., DiCuccio, M., et al.; Mouse Genome Sequencing Consortium. Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol. 7, e1000112 (2009).

De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B.K., Muller, H,, Ragoussis, J.,Wei, C.L., and Natoli, G. A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol 8: e1000384 (2010). doi: 10.1371/journal.pbio.1000384

DeFelipe, J. The evolution of the brain, the human nature of cortical circuits, and intellectual creativity. Front. Neuroanat. **5,** 29 (2011).

Deininger, PL, and Batzer, MA. Alu repeats and human disease. Molecular Genetics and Metabolism, 67(3):183-193 (1999). ISSN 1096-7192. doi: http://dx.doi.org/10.1006/mgme.1999.2864.

Dekkers, J. F. et al. A functional CFTR assay using primary cystic fibrosis intestinal organoids. Nat. Med. 19, 939–945 (2013).

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 22, 1775–1789 (2012).

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29(1), 15-21 (2013).

Eiraku, M. & Sasai, Y. Self-formation of layered neural structures in three-dimensional culture of ES cells. Curr. Opin. Neurobiol. 22, 768–777 (2012).

Eiraku, M. et al. Self-organized formation of polarized cortical tissues from ES cells and its active manipulation by extrinsic signals. Cell Stem Cell 3, 519-532 (2008).

Eiraku, M. et al. Self-organizing optic-cup morphogenesis in three-dimensional culture. Nature 472, 51–56 (2011).

Enard, W. et al. Intra- and interspecific variation in primate gene expression patterns. Science 296, 340–-343 (2002).

ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447, 799–816 (2007).

Fuentes, P., Cánovas, J., Berndt, F.A., Noctor, S.C., Kukuljan, M. CoREST/LSD1 control the development of pyramidal cortical neurons. Cereb Cortex [Epub ahead of print] (2011).

Fusaki, N., Ban, H., Nishiyama, A., Saeki, K. & Hasegawa, M. Efficient induction of transgene-free human pluripotent stem cells using a vector based on Sendai virus, an RNA virus that does not integrate into the host genome. Proc. Japan Acad. Ser. B **85,** 348–362 (2009).

Gilbert, L. A. et al. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. Cell 159, 647–661 (2014).

Gingeras, T.R. Patience is a virtue. Nature 482, 6–7 (2012).

Gong, C., and Maquat, L.E. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 30 UTRs via Alu elements. Nature 470, 284–288 (2011).

Gupta, R.A., Shah, N., Wang, K.C., Kim, J., Horlings, H.M., Wong, D.J., Tsai, M.-C., Hung, T., Argani, P., Rinn, J.L., et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature 464, 1071–1076 (2010).

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458: 223–227 (2009).

Guttman, M. et al. lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature 477, 295–300 (2011).

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458, 223–227 (2009).

Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L., et al. lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature 477, 295–300 (2011).

Hamasaki, T., Leinga, A., Ringstedt, T., O'Leary, D.D.M., and Jolla, L. EMX2 Regulates Sizes and Positioning of the Primary Sensory and Motor Areas in Neocortex by Direct Specification of Cortical Progenitors. Neuron 43, 359–372 (2004).

Hangauer MJ, Vaughn IW, McManus MT. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. PLoS Genet. 9:e1003569 (2013).

Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D, et al. GENCODE: producing a reference annotation for ENCODE. Genome Biol 7: S4 (2006). doi: 10.1186/gb-2006-7-s1-s4.

Heo JB, Sung S. Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. Science 331: 76–79 (2011).

Hill, R. S. & Walsh, C. A. Molecular insights into human brain evolution. Nature **3,** 64–67 (2005).

Hill, R.S., and Walsh, C.A. Molecular insights into human brain evolution. Nature 437, 64–67 (2005).

Hon, C. et al. An atlas of human long non-coding RNAs with accurate 5' ends. Nature 543, 199–204 (2017).

Hopkins, W. D. & Cantalupo, C. Handedness in chimpanzees (Pan troglodytes) is associated with asymmetries of the primary motor cortex but not with homologous language areas. Behav. Neurosci. 118, 1176–-1183 (2004).

Huarte, M., Guttman, M., Feldser, D., Garber, M., Koziol, M.J., Kenzelmann-Broz, D., Khalil, A.M., Zuk, O., Amit, I., Rabani, M., et al. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell 142, 409–419 (2010).

Huch, M. et al. In vitro expansion of single Lgr5+ liver stem cells induced by Wnt-driven regeneration. Nature 494, 247-250 (2013).

Hung T, Wang Y, Lin MF, Koegel AK, Kotake Y, Grant GD, Horlings HM, Shah N, Umbricht C, Wang P, et al. Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. Nat Genet 43: 621–629 (2011).

Hung, T., Wang, Y., Lin, M.F., Koegel, A.K., Kotake, Y., Grant, G.D., Horlings, H.M., Shah, N., Umbricht, C., Wang, P., et al. Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. Nat. Genet. 43, 621–629 (2011).

Hyslop, LA, Armstrong, L, Stojkovic, M, Lako, M. Human embryonic stem cells: biology and clinical implications. Expert Reviews in Molecular Medicine, 7(19):1-21 (2005).

Iannaccone, PM, Taborn, GU, Garton, RL, Caplice, MD, Brenin, DR. Pluripotent embryonic stem cells from the rat are capable of producing chimeras. Dev Biol 185(1):124-125 (1997).

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature 431, 931–945 (2004).

Itzhaki, I., Maizels, L., Huber, I., Zwi-Dantsis, L., Caspi, O., Winterstern, A., Feldman, O., Gepstein, A., Arbel, G., Hammerman, H., Boulos, M., Gepstein, L. (2010). Modelling the long QT syndrome with induced pluripotent stem cells. Nature 471(7337), 225-9 (2004).

Jacobs, FM, Greenberg, D, Nguyen, N, Haeussler, M, Ewing, AD, Katzman, S, Paten, B, Salama, SR, Haussler, D. An evolutionary arms race between krab zinc finger genes znf91/93 and sva/l1 retrotransposons. Nature (2014).

Jensen, L.R., Amende, M., Gurok, U., Moser, B., Gimmel, V., Tzschach, A., Janecke, A.R., Taruverdian, G., Chelly, J., Fryns, J., Esch, H., Kleefstra, T., Hamel, B., Moraine, C., Gécz, J., Turner, G., Reinhardt, R., Kalscheuer, V.M., Ropers, H., Lenzner, S.

Mutations in the JARID1C gene, which is involved in transcriptional regulation and chromatin remodeling, cause X-linked mental retardation. Cell 76(2), 227-236 (2005).

Ji, J, Ng, SH, Sharma, V, Neculai, D, Hussein, S, Sam, M, Trinh, Q, Church, GM, McPherson, JD, Nagy, A, et al. Elevated coding mutation rate during the reprogramming of human somatic cells into induced pluripotent stem cells. Stem Cells 30(3), 435-40 (2012).

Kang, H., Kawasawa, Y., Cheng, F., Zhu, Y., and Xu, X. Spatio-temporal transcriptome of the human brain. Nature 478 (7370), 489-9 (2011).

Kapranov, P., Cheng, J., Dike, S., Nix, D.A., Duttagupta, R., Willingham, A.T., Stadler, P.F., Hertel, J., Hackermüller, J., Hofacker, I.L., et al. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 316, 1484–1488 (2007).

Kelley, D.R., and Rinn, J.L. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol. 13, R107 (2012).

Kent, WJ, Sugnet, CW, Furey, TS, Roskin, KM, Pringle, TH, Zahler, AM, Haussler, DH. The Human Genome Browser at UCSC. Genome Research, 12(6):996-1006 (2002). doi: 10.1101/gr.229102.

Khalil, A.M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B.E., van Oudenaarden, A., et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc. Natl. Acad. Sci. U.S.A. 106, 11667–11672 (2009).

Kim, T-K, Hemberg, M, Gray, JM, Costa, AM, Bear, DM, Wu, J, Harmin, DA, Laptewicz, M, Barbara-Haley, K, Kuersten, S, et al. Widespread transcription at neuronal activity- regulated enhancers. Nature 465: 182–187 (2010).

King, M, Wilson, AC. Evolution at two levels in humans and chimpanzees. Essential Readings in Evolutionary Biology, 188(4184), 301 (1975).

King, M.C., and Wilson, A.C. Evolution at two levels in humans and chimpanzees. Science 188, 107–116 (1975).

Konermann, S., Brigham, M. D., Trevino, A. E., Joung, J., Abudayyeh, O. O., Barcena, C., … Zhang, F. Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature, 517(7536), 583–8 (2014).

Konig, S., Romoth, L.W., Gerischer, L., Stanke, M. Simultaneous gene finding in multiple genomes. Bioinformatics 32(22), 3388-3395 (2016).

Kornack, D. R. & Rakic, P. Changes in cell-cycle kinetics during the development and evolution of primate neocortex. Proc. Natl Acad. Sci. USA 95, 1242–1246 (1998).

Kouzarides, T. Chromatin modifications and their function. Cell, 128(4), 693-705 (2007).

Kowalczyk, M. S., and Higgs, D. R. RNA discrimination. Nature 482, 6–7 (2012).

Koziol MJ, Rinn JL. RNA traffic control of chromatin complexes. Curr Opin Genet Dev 20, 142–148 (2010).

Kutter, C., Watt, S., Stefflova, K., Wilson, M.D., Goncalves, A., Ponting, C.P., Odom, D.T., and Marques, A.C. Rapid turnover of long noncoding RNAs and the evolution of gene expression. PLoS Genet. 8, e1002841 (2012).

Kuzdzal-Fick, JJ, Fox, SA, Strassmann, JE, Queller, DC. High relatedness
Lagarde, J. et al. High-throughput annotation of full-length long noncoding RNAs with Capture Long-Read Sequencing (CLS). bioRxiv 1–26 (2017).

Lagarde, J., Uszczynska-Ratajczak, B., Carbonell, S., Davis, C., Gingeras, T.R., Frankish, A., Harrow, J., Guigo, R., and Johnson, R. High-throughput annotation of full-length long noncoding RNAs with Capture Long-Read Sequencing (CLS). bioRxiv 1–26 (2017). doi:10.1101/105064

Lancaster, M.A., et al. Cerebral organoids model human brain development and microcephaly. Nature 501, 373–9 (2013).

Langmead, B., Salzberg, S. Fast gapped-read alignment with Bowtie 2. Nature Methods 9, 357-359 (2012).

Leighton PA, IngramRS, Eggenschwiler J, Efstratiadis A, Tilghman SM. Disruption of imprinting caused by deletion of the H19 gene region in mice. Nature 375: 34–39 (1995).

Levin, J.Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D.A., Friedman, N., Gnirke, A., and Regev, A. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat. Methods 7, 709–715 (2010).

Li, B., Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323 (2011).

Lin, M., Pedrosa, E., Shah, A., Hrabovsky, A., Maqbool, S., Zheng, D., and Lachman, H.M. RNA-Seq of human neurons derived from iPS cells reveals candidate long non-coding RNAs involved in neurogenesis and neuropsychiatric disorders. PLoS ONE 6, e23356 (2011).

Liu, S.J., Nowakowski, T.J., Pollen, A.A., Lui, J.H., Horlbeck, M.A., Attenello, F.J., He, D., Weissman, J.S., Kriegstein, A.R., Diaz, A.A., Lim, D.A. Single-cell analysis of long non-coding RNAs in the developing human neocortex. Genome Biol. 17, 67 (2016).

Liu, S. J., Horlbeck, M.A., Cho, S.W., Birk, H.S., Malatesta, M., He, D., Attenello, F.J., Villalta, J.E., Cho, M.Y., Chen, Y., Mandegar, M.A., Olvera, M.P., Gilbert, L.A., Conklin, B.R., Chang, H.Y., Weissman, J.S., and Lim, D.A. (2016). CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. Science. doi:10.1126/science.aah7111

Locke, DP, Hillier, LW, Warren, WC, Worley, KC, Nazareth, LV, Muzny, DM, Yang, S, Wang, Z, Chinwalla, AT, Minx, P, et al. Comparative and demographic analysis of orang-utan genomes. Nature, 469(7331), 529-533 (2011).

Loewer, S., Cabili, M.N., Guttman, M., Loh, Y.-H., Thomas, K., Park, I.H., Garber, M., Curran, M., Onder, T., Agarwal, S., et al. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. Nat. Genet. 42, 1113–1117 (2010).

Love MI, Huber W and Anders S. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome Biology, 15, 550 (2014). doi: 10.1186/s13059-014-0550-8.

Mariani, J. et al. Modeling human cortical development in vitro using induced pluripotent stem cells. Proc. Natl Acad. Sci. USA 109, 12770–12775 (2012).

Marin-Padilla, M. Ontogenesis of the pyramidal cell of the mammalian neocortex and developmental cytoarchitectonics: a unifying theory. J. Comp. Neurol. 321, 223–240 (1992).

Marques, A.C., and Ponting, C.P. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. Genome Biol. 10, R124 (2009).

Martins-Taylor, K. et al. Imprinted expression of UBE3A in non-neuronal cells from a Prader-willi syndrome patient with an atypical deletion. Hum. Mol. Genet. 23, 2364–2373 (2014).

Mattick, J.S. Non-coding RNAs: the architects of eukaryotic complexity. EMBO Rep. 2, 986–991 (2001).

Mattick, J.S. RNA regulation: a new genetics? Nat. Rev. Genet. 5, 316–323 (2004).

McLean, C.Y., Reno, P.L., Pollen, A.A., Bassan, A.I., Capellini, T.D., Guenther, C., Indjeian, V.B., Lim, X., Menke, D.B., Schaar, B.T., et al. Human-specific loss of regulatory DNA and the evolution of human-specific traits. Nature 471, 216–219 (2011).

Mercer, T.R., Dinger, M.E., Sunkin, S.M., Mehler, M.F., and Mattick, J.S. Specific expression of long noncoding RNAs in the mouse brain. Proc. Natl. Acad. Sci. USA 105, 716–721 (2008).

Molyneaux, B., and Arlotta, P. Neuronal subtype specification in the cerebral cortex. Nat. Rev. **8,** 427–437 (2007).

Mora-Bermudez, F., et al. Differences and similarities between human and chimpanzee neural progenitors during cerebral cortex development. Elife 5, 1–24 (2016).

Moretti, A., Bellin, M., Welling, A., Jung, C.B., Lam, J.T., Bott-Flügel, L., Dorn, T., Goedel, A., Höhnke, C., Hofmann, F., Seyfarth, M., Sinnecker, D., Schömig, A., Laugwitz, K.L. Patient-specific induced pluripotent stem-cell models for long-QT syndrome. N Engl J Med 363(15), 1397-409 (2010).

Mouse Genome Sequencing Consortium, Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., et al. Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562 (2002).

Mu, X.J., Lu, Z.J., Kong, Y., Lam, H.Y.K., and Gerstein, M.B. Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. Nucleic Acids Res. 39, 7058–7076 (2011).

Nagano, T., Mitchell, J.A., Sanz, L.A., Pauler, F.M., Ferguson-Smith, A.C., Feil, R., and Fraser, P. The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. Science 322, 1717–1720 (2008).

Noctor, SC, Flint, AC, Weissman, TA, Dammerman, RS, Kriegstein, AR. Neurons derived from radial glial cells establish radial units in neocortex. Nature 409(6821), 714-720 (2001).

O'Donnovan, M.C., Craddock, N.J., Owen, M.J. Genetics of psychosis; insights from views across the genome. Hum Genet. 126(1), 3-12 (2009).

Okita, K., Matsumura, Y., Sato, Y., Okada, A., Morizane, A., Okamoto, S., … Yamanaka, S. A more efficient method to generate integration-free human iPS cells. Nature Methods, 8(5), 409–12 (2011).

Orban, G. A., Van Essen, D. & Vanduffel, W. Comparative mapping of higher visual areas in monkeys and humans. Trends Cogn. Sci. 8, 315–324 (2004).

Ørom, U.A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q., et al. Long noncoding RNAs with enhancer-like function in human cells. Cell 143, 46–58 (2010a).

Ørom, U.A., Derrien, T., Guigo, R., and Shiekhattar, R. Long noncoding RNAs as enhancers of gene expression. Cold Spring Harb. Symp. Quant. Biol. 75, 325–331 (2010b).

Otani, T., Marchetto, M. C., Gage, F. H., Simons, B. D. & Livesey, F. J. 2D and 3D Stem Cell Models of Primate Cortical Development Identify Species-Specific Differences in Progenitor Behavior Contributing to Brain Size. Cell Stem Cell 18, 467–480 (2016).

Pandey, R.R., Mondal, T., Mohammad, F., Enroth, S., Redrup, L., Komorowski, J., Nagano, T., Mancini-Dinardo, D., and Kanduri, C. Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. Mol. Cell 32, 232–246 (2008).

Pang, K.C., Frith, M.C., and Mattick, J.S. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. Trends Genet. 22, 1-5 (2006).

Paten, B., Earl, D., Nguyen, N., Diekhans, M., Zerbino, D., Haussler, D. Cactus: algorithms for genome multiple sequence alignment. Genome Res 21, 1512-1528 (2011).

Pauli, A., Valen, E., Lin, M.F., Garber, M., Vastenhouw, N.L., Levin, J.Z., Fan, L., Sandelin, A., Rinn, J.L., Regev, A., and Schier, A.F. Systematic identi- fication of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res. 22, 577–591 (2012).

Penny, G.D., Kay, G.F., Sheardown, S.A., Rastan, S., and Brockdorff, N. Requirement for Xist in X chromosome inactivation. Nature 379, 131–137 (1996).

Pereira, J.D., Stephen, N.S., Smith, J., Dobenecker, M., Tarakhovsky, A., Livesey, F.J. Ezh2, the histone methyltransferase of PRC2, regulates the balance between self-renewal and differentiation in the cerebral cortex. PNAS 107(36), 15957-62 (2010).

Pollard, K.S., Salama, S.R., King,B., Kern, A.D., Dreszer, T., Katzman, S., Siepel, A., Pedersen, J.S., Bejerano, G., Baertsch, R., et al. Forces shaping the fastest evolving regions in the human genome. PLoS Genet. 2, e168 (2006).

Pollen, A., et al. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat. Biotechnol. 32, (2014).

Pollen, A. et al. Molecular Identity of Human Outer Radial Glia during Article Molecular Identity of Human Outer Radial Glia during Cortical Development. Cell 163, 55–67 (2015).

Ponjavic, J., Ponting, C.P., and Lunter, G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res. 17, 556–565 (2007).

Ponting, C.P., Oliver, P.L., and Reik, W. Evolution and functions of long noncoding RNAs. Cell 136, 629–641 (2009).

Prasad, A, Kumar, SS, Dessimoz, C, Jaquet, V, Bleuler, S, Laule, O, Hruz, T, Gruissem, W, Zimmermann, P. Global regulatory architecture of human, mouse and rat tissue transcriptomes. BMC genomics, 14(1), 716 (2013).

Prokhorova, T.A., Harkness, L.M., Frandsen, U., Ditzel, N., Schrøder, H.D., Burns, J.S., and Kassem, M. Teratoma Formation by Human Embryonic Stem Cells Is Site Dependent and Enhanced by the Presence of Matrigel. Stem Cells Dev. 18, 47–54 (2009).

Qian, X. et al. Brain-Region-Specific Organoids Using Mini-bioreactors for Modeling ZIKV Exposure. Cell 165, 1238–1254 (2016).

Qureshi, I.A., Mattick, J.S., and Mehler, M.F. Long non-coding RNAs in nervous system function and disease. Brain Res. 1338, 20–35 (2010).

Radonjić, N.V., Ayoub, A.E., Memi, F., Yu, X., Maroof, A., Jakovcevski, I., Anderson, S.A., Rakic, P., and Zecevic, N. Diversity of Cortical Interneurons in Primates: The Role of the Dorsal Proliferative Niche. Cell Rep. 9**,** 2139–2151 (2014).

Ramos, A.D., Diaz, A., Nellore, A., Delgado, R.N., Park, K., Gonzales-Roybal, G., Oldham, M.C., Song, J.S., Lim, D.A. Integration of genome-wide approaches identifies lncRNAs of adult neural stem cells and their progeny in vivo. Cell Stem Cell. 12(5), 616-628 (2013).

Rani et al., A Primate lncRNA Mediates Notch Signaling during Neuronal Development by Sequestering miRNA, Neuron (2016).

Ravasi, T., Suzuki, H., Pang, K.C., Katayama, S., Furuno, M., Okunishi, R., Fukuda, S., Ru, K., Frith, M.C., Gongora, M.M., et al. Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. Genome Res. 16, 11–19 (2006).

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 129, 1311–1323 (2007).

Rippon, HJ, Bishop, AE. Embryonic stem cells. Cell Proliferation, 37(1), 23-34 (2004).

Ruvolo, M. Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data sets. Mol. Biol. Evol. 14, 248–265 (1997).

Sasai, Y., Eiraku, M. & Suga, H. In vitro organogenesis in three dimensions: self-organising stem cells. Development 139, 4111–4121 (2012).

Sato, T. et al. Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium. Gastroenterology 141, 1762–1772 (2011).

Sato, T. et al. Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche. Nature 459, 262–265 (2009).

Semendeferi, K., Lu, A., Schenker, N. & Damasio, H. Humans and great apes share a large frontal cortex. Nature Neurosci. 5, 272–276 (2002).

Shi, Y, Kirwan, P, Livesey, FJ. Directed differentiation of human pluripotent stem cells to cerebral cortex neurons and neural networks. Nature protocols, 7(10), 1836-1846 (2012).

Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013-2015 http://www.repeatmasker.org.

Stange, D. E. et al. Differentiated Troy+ chief cells act as reserve stem cells to generate all lineages of the stomach epithelium. Cell 155, 357–368 (2013).

Stanke, M., Diekhans, M., Baertsch, R., Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics, 24(5), 637-644 (2008).

Struhl, K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat. Struct. Mol. Biol. 14, 103–105 (2007).

Sun, T. et al. Early asymmetry of gene transcription in embryonic human left and right cerebral cortex. Science 308, 1794–1798 (2005).

Takahashi, K, Tanabe, K, Ohnuki, M, Narita, M, Ichisaka, T, Tomoda, K, Yamanaka, S. Induction of pluripotent stem cells from adult human broblasts by defined factors. Cell, 131(5), 861-872 (2007).

Takahashi, T., Nowakowski, R. S. & Caviness, V. S. Jr. The cell cycle of the pseudostratified ventricular epithelium of the embryonic murine cerebral wall. J. Neurosci. 15, 6046–6057 (1995).

Trapnell, C., et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28, 511–515 (2010).

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., … Pachter, L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols, 7(3), 562–578 (2012).

Tsai, M.-C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J.K., Lan, F., Shi, Y., Segal, E., and Chang, H.Y. Long noncoding RNA as modular scaffold of histone modification complexes. Science 329, 689–693 (2010).

Tuch, BE, et al. Stem cells: A clinical update. Australian Family Physician, 35(9), 719 (2006).

Uddin, M. et al. Sister grouping of chimpanzees and humans as revealed by genome-wide phylogenetic analysis of brain gene expression profiles. Proc. Natl Acad. Sci. USA 101, 2957–2962 (2004).

Ulitsky, I. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. Nat. Rev. Genet. 17, 601–614 (2016).

Ulitsky, I., and Bartel, D.P. lincRNAs: genomics, evolution, and mechanisms. Cell 154, 26–46 (2013).

Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H., and Bartel, D.P. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell 147, 1537–1550 (2011).

Walker, JA, Konkel, MK, Ullmer, B, Monceaux, CP, Ryder, OA, Hubley, R, Smit, AF, Batzer, MA. Orangutan alu quiescence reveals possible source element: support for ancient backseat drivers. Mob. DNA 3(8), (2012).

Wang KC, Yang YW, Liu B, Sanyal A, Corces-Zimmerman R, Chen Y, Lajoie BR, Protacio A, Flynn RA, Gupta RA, et al. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature 472, 120–124 (2011).

Wang, J., Zhang, J., Zheng, H., Li, J., Liu, D., Li, H., Samudrala, R., Yu, J., and Wong, G.K. Mouse transcriptome: neutral evolution of 'non-coding' complementary DNAs. Nature 431 (2004).

Wang, K.C., Yang, Y.W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B.R., Protacio, A., Flynn, R.A., Gupta, R.A., et al. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature 472, 120–124 (2011).

Wildman, D.E., Uddin, M., Liu, G., Grossman, L.I., and Goodman, M. Implications of natural selection in shaping 99.4% nonsynonymous DNA identity between humans and chimpanzees: enlarging genus Homo. Proc. Natl. Acad. Sci. U.S.A. 100, 7181–7188 (2003).

Williams, H.J., Owen, M.J., O'Donnovan, M.C. New findings from genetic association studies of schizophrenia. J Hum Genet. 54(1), 9-14 (2009).

Yu, J, Vodyanik, MA, Smuga-Otto, K, Antosiewicz-Bourget, J. Frane, JL, Tian, S, Nie, J, Jonsdottir, GA, Ruotti, V, Stewart, R, Slukvin, II, Thomson, JA. Induced pluripotent stem cell lines derived from human somatic cells. Science 318(5858), 1917-1920 (2007).

Yu, J., Hu, K., Smuga-Otto, K., Tian, S., Stewart, R., Slukvin, I. I., Thomson, J. A. Human induced pluripotent stem cells free of vector and transgene sequences. Science 324(5928), 797–801 (2009).

Zhao, J, Ohsumi, TK, Kung, JT, Ogawa, Y, Grau, DJ, Sarma, K, Song, JJ, Kingston, RE, Borowsky, M, Lee, JT. Genome-wide identification of polycomb-associated RNAs by RIP-seq. Mol Cell 40, 939–953 (2010).

Zhao, J, Sun, BK, Erwin, JA, Song, JJ, Lee, JT. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science 322: 750–756 (2008).