Identification and functional characterization of long non-coding RNAs in development and disease

by

Siyuan John Liu

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biomedical Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

To my wife, Caterina

Acknowledgements

I have had the great privilege of working in the laboratory of Professor Daniel Lim. Prof. Lim is an exemplary scientist, physician, and mentor. Prof. Lim is a role model in every sense, not just because he is a successful physician/scientist but because of his commitment to his students, his work ethic, his ability to motivate and invigorate, and the stimulating and welcoming environment he has fostered in the lab. He has guided and supported me during critical times in my PhD. It was in his lab where I learned to be a mature, critical scientist, and for that I am for forever indebted to Prof. Lim. I owe most of what I know about computational biology to Professor Aaron Diaz, who mentored and taught me since the early days of my PhD and made sure at every corner that I was using rigorous statistics. In addition to bestowing knowledge, Prof. Diaz gave me the computational resources to carry out my analysis with the power of a high performance compute cluster. Prof. Jonathan Weissman has been the best collaborator anyone could have asked for. The decades of experience he brought to the table was unprecedented, as was his enthusiasm for working together. Without a doubt, Prof. Weissman elevated the profile of our work to levels I could have only dreamed of. And along with Prof. Weissman was Max Horlbeck, my friend and MSTP colleague whom I worked with side by side on the CRISPRi project. Working for over two years on the project together was not only the highlight of my entire PhD, but also marked the first time I felt I was contributing to a project that was truly remarkable and beyond anything either one of us could have accomplished alone.

I am indebted to Prof. Arnold Kriegstein for allowing me the opportunity to study early human brain development, a privilege given to only the brightest and most morally sound researchers in the nation. Driving the enthusiasm in the Kriegstein lab were postdoctoral fellows Alex Pollen and Tom Nowakowski, who were tremendously diligent and devoted scientists who

laid the intellectual and physical groundwork for the single cell RNA-seq efforts. Other Kriegstein lab members I am indebted to are Jan Lui, Lizzie di Lullo, Carmen Sandoval, and William Walantus.

The Lim lab was truly a collaborative environment. Alex Ramos performed foundational work in the lab, without which the lab would not be the lncRNA lab that it is today. Ryan Delgado fulfilled a big sibling role who influenced my decision to join the lab and was an outstanding colleague. Becky Andersen never ceased to inspire me with her work on *in vivo* studies of lncRNAs. Martina Malatesta, Frank Attenello, Harjus Birk, Daniel He, and David Wu worked tirelessly with me to better understand lncRNAs. Miao Cui and Sajad Hamid brought fresh insights into the lab and provided invaluable conversations. Sung Hong rescued me many times, by aiding me with various tasks when no one else was left in the laboratory. And without Melanie Bedolli, our superstar lab manager, my and the rest of the lab's research would have come to a halt on a regular basis.

Outside of the Lim lab, I am indebted to Sören Müller, who led the computational analysis of the single cell glioblastoma work. Prof. Gary Kohanbash introduced me to the world of cancer immunology and its potential for translation to the clinic. Dr. Steve Braunstein was the first clinical mentor I've ever had, and I thank him for his willingness to take me under his wing despite my novice abilities. Prof. Barbara Panning and Joe Costello served on my thesis and preliminary exam committees and provided invaluable guidance and critical eyes. Prof. Howard Chang of Stanford University was the closest I've had to a "outside" mentor, providing unique insights and also contributed greatly to our collaborative CRISPRi work. Seung Woo Cho in the Chang lab also strengthened our lncRNA work, in addition to providing multi-institutional credibility to our results.

I thank my mother Sherry Zhang for instilling in me a love of science from infancy, teaching me how to read, and letting me follow my passions (academic and otherwise). Her military discipline, high standards, and tirelessness are truly inspirational. I thank my father Jay Liu for showing me the value of hard work; for he grew up a peasant in rural China and worked his way to become the first PhD scholar in the Liu family. It's an honor to follow in Dr. Liu's footsteps. And finally, I thank my wife Caterina Liu, whose perpetual companionship, love, and support ultimately gave me the energy, motivation, and happiness that got me to where I am.

Chapter specific acknowledgements

Chapter 1 is an unpublished survey of lncRNA biology with a historical primer, insights into lncRNA mechanisms, and techniques used to study lncRNA biology, several of which are used in subsequent chapters.

Chapter 2 is reprinted largely as it appears in Liu SJ, Nowakowski TJ, Pollen AA, Lui JH, Horlbeck MA, Attenello FJ, He D, Weissman JS, Kriegstein AR, Diaz AA, Lim DA. Single-cell analysis of long non-coding RNAs in the developing human neocortex. *Genome Biol*. 2016;17:67 PMID 27081004

S.J.L and D.A.L. conceived the project. D.A.L., A.A.D., and A.R.K. supervised the project. S.J.L and A.A.D performed bioinformatic and statistical analyses, with input from D.A.L., A.A.P, and T.N.J. S.J.L and J.H.L performed bulk tissue experiments. T.J.N., A.A.P., and J.H.L. performed single cell RNA-seq experiments. T.N.J., S.J.L, and D.H. performed in situ hybridizations. S.J.L, M.A.H., J.S.W., and F.J.A. designed and performed CRISPRi experiments. Manuscript was prepared by S.J.L. and D.A.L. with input from all authors.

Chapter 3 is reprinted largely as it appears in Müller S, Liu SJ, Di Lullo E, Malatesta M, Pollen AA, Nowakowski TJ, Kohanbash G, Aghi M, Kriegstein AR, Lim DA, Diaz A. Single-cell sequencing maps gene expression to mutational phylogenies in PDGF and EGF driven gliomas. *Mol Syst Biol*. 2016;12(11) PMID 27888226.

M.S. and S.J.L. contributed equally to this work. A.D. conceived of and designed the study. S.M. developed the algorithms for inferring tumor evolution. E.D., G.K., S.J.L., M.M., T.N. and A.P. performed the experiments under the supervision of A.D, A.K., and D.L. M.A. provided the surgical specimens via the UCSF Tissue Core. A.D. and S.M. performed the bioinformatics analysis of single cell and exome-seq data. A.D. and S.J.L. performed the bioinformatics analysis of the *in vitro* RNA-seq data. A.D. , S.M., and S.J.L. wrote the manuscript with input from all authors.

Chapter 4 includes reprinted material from Liu SJ, Horlbeck MA, Cho SW, Birk HS, Malatesta M, He D, Attenello FJ, Villalta JE, Cho MY, Chen Y, Mandegar MA, Olvera MP, Gilbert LA, Conklin BR, Chang HY, Weissman JS, Lim DA. Science. 2017 Jan 6;355(6320). pii: aah7111. doi: 10.1126/science.aah7111, **in addition to unpublished sections on the relationship between chromatin contacts and lncRNA biology, mechanisms of a potentially *cis* acting lncRNA, and an unusual lncRNA-oncogene fusion critical for the survival of HeLa cells.**

S.J.L and M.A.H. contributed equally to this work. M.A.H., S.J.L., J.S.W., D.A.L., and H.Y.C. conceived the project, interpreted the data, and wrote the manuscript. M.A.H. and S.J.L designed sgRNA library, performed RNA-seq, qPCR, growth assays, and analyzed data. S.J.L. performed screens in U87 and HeLa cells, 4C-seq, and machine learning. M.A.H. performed

screens in K562, HEK293T, and iPS cells. S.W.C. performed screens and validation experiments in MCF7 and MDA-MB-231 cells. H.S.B., M.M., and F.J.A. performed qPCR, growth assays, and ASO experiments. D.H. performed ChIP-seq. J.E.V. cloned sgRNA libraries. M.Y.C. and Y.C. performed validation experiments. M.A.M., M.P.O., and B.R.C. contributed to iPSC experiments and performed iPSC protein-coding screens. L.A.G. contributed to data interpretation and project conception.

Chapter 5 is entirely unpublished work extending the results from the CRISPRi lncRNA screen toward applications in glioma therapy, by performing radiation modifier screens to discover lncRNA loci that, when inactivated, sensitize cells to radiation therapy. John Liu lead this work and performed experiments and data analysis, with Max Horlbeck contributing to sgRNA design and data analysis. Martina Malatesta performed one of the replicates of the screen, and Harjus Birk helped with RNA-seq library prep and proliferation assays. Daniel Lim and Jonathan Weissman surpervised.

Supplementary tables listed throughout may be accessed through the corresponding publication citations listed above. Further acknowledgements are described in each chapter.

Identification and functional characterization of long non-coding RNAs in development and disease

S. John Liu

Abstract

The human genome produces tens of thousands of long non-coding RNA (lncRNAs), transcripts larger than 200 nt that do not make proteins. Select lncRNAs are established regulators of gene expression in developmental processes and in the etiology of diseases. However, the functions of the vast majority of lncRNAs are unknown. Here we utilize high throughput descriptive and functional genomics to characterize and understand the roles of lncRNAs in development of the human brain and in cancer biology. Starting with deep RNA sequencing (RNA-seq) of developing brain tissues, we comprehensively annotated lncRNA transcript structures in the cerebral cortex, revealing thousands of novel lncRNAs, many of which were antisense or non-polyadenylated. We then applied single cell RNA-seq to quantitatively measure the transcriptomes of individual cells, revealing cell type-specific expression of lncRNAs and surprisingly high abundance of certain lncRNAs (e.g. *DLX6-AS1*) that appear to be low abundance at the population level, due to cellular heterogeneity in the brain. Combination of single cell RNA-seq and exome-seq were then applied to not only measure RNA expression heterogeneity, but also to map the mutational phylogenies of glioblastoma tumors, revealing a novel deletion in *PDGFRA* that confers survival advantage. We then developed a CRISPR interference (CRISPRi) platform to systematically repress 16,401 lncRNA loci across seven cancer and non-cancer cell lines, identifying 499 lncRNAs necessary for robust growth and proliferation. The functions of these lncRNAs were exquisitely cell type-specific and correlated with complex transcriptional networks and chromatin architecture. Finally, we applied CRISPRi

screening of lncRNAs in combination with ionizing radiation therapy and identified over 200 lncRNAs whose knockdown sensitizes glioblastoma cells to radiation induced cell death. As much as these efforts advance our knowledge of lncRNA biology in development and in disease, they also raise exciting new questions for future inquiry.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Long non-coding RNAs: initial discoveries

The central dogma of molecular biology postulates that ribonucleic acid (RNA), in particular messenger RNA, serves in the flow of genetic information as an intermediate between DNA and proteins (Crick, 1970), which were thought to be responsible for performing essential functions in the cell. Therefore, it was surprising when completion of the human genome sequence revealed far fewer protein coding genes than expected, in the range of 20,000, which is roughly the same number as in invertebrates such as worms and flies (Lander et al., 2001; Venter et al., 2001). Microarray and massively parallel RNA sequencing (RNA-seq) technologies, which were enabled by the drafting of the human genome, then revealed that the human genome harbored at least as many unique RNA transcripts that do not encode proteins (non-coding RNA) as protein coding transcripts (Cabili et al., 2011; Djebali et al., 2012; FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al., 2014; Hon et al., 2017; Kapranov et al., 2002). While many of these non-coding RNAs were known, such as ribosomal RNA, transfer RNA, small nuclear RNAs, and micro RNAs, another class of non-coding RNA was rapidly unearthed: long non-coding RNAs (lncRNAs).

LncRNAs are operationally defined as RNA transcripts greater than 200 nt long, to biophysically distinguish them from small regulatory RNAs, and have no evidence for protein coding potential. Assessment of protein coding potential has undergone debate, but useful approaches include sequence detection of open reading frames, comparative genomics methods such as codon substitution frequency (Lin et al., 2011), *in vitro* translation experiments, proteomics data analysis (Banfai et al., 2012), ribosome profiling analysis (Guttman et al., 2013;

Ingolia et al., 2014), and machine learning models trained on established coding and non-coding transcripts(Sun et al., 2013; Wang et al., 2013). Nonetheless, a small number of transcripts originally annotated as lncRNAs by these and other methods are subsequently found to encode functional peptides (Anderson et al., 2015; Nelson et al., 2016). Unlike small regulatory RNAs such as micro RNAs, lncRNAs encompass a broad range of RNA species (Ulitsky and Bartel, 2013). Both polyadenylated and nonpolyadenylated transcripts have fallen under the term, and lncRNAs can be oriented as intergenic, intronic, antisense, or overlapping with respect to coding genes. LncRNAs can also be unspliced single exons or spliced transcripts. The lncRNA class known as long intergenic/intervening non-coding RNA (lincRNA) is understood to be 5' methylguanine capped, 3' polyadenylated, and have prominent trimethylation of histone 3 lysine residue 4 (H3K4me3) at their promoters and broad trimethylation of histone 3 lysine 36 (H3K36me3) along their gene bodies (Guttman et al., 2009). Recent large scale sequencing studies in humans have identified tens of thousands of lncRNAs (Cabili et al., 2011; Derrien et al., 2012) – in one study as high as 50,000 (Iyer et al., 2015). The NONCODE database now lists a staggering 337,880 annotated lncRNA genes, 101,700 of which are found in human (Zhao et al., 2016). Ideally, as sets of lncRNAs become better characterized, more informative functional designations will be established to make sense of these highly heterogenous RNAs (Mattick and Rinn, 2015). There is now little doubt of pervasive transcription from the human genome and the diversity of RNA molecules that are produced. There is, however, much debate over which lncRNAs are functional, what their roles are, and how they perform these functions.

lncRNAs regulate the flow of genetic information

One of the earliest long non-coding RNAs to be recognized is the X inactive specific transcript – *XIST* (BARR and BERTRAM, 1949; Borsani et al., 1991; Brown et al., 1991). *XIST* is an RNA produced in female mammalian cells and is required for the initiation of X chromosome inactivation (XCI), a process of dosage compensation that randomly silences transcription from one of two female X chromosomes (Wutz, 2011). Decades of research have begun to elucidate the molecular mechanisms of *XIST* in XCI. The transcript establishes enrichment of heterochromatin at the inactive X by tethering the chromosome to the nuclear periphery via interaction between the *XIST* Rep A sequence and the lamin B receptor (Chen et al., 2016), allowing local spreading of the *XIST* RNA (Engreitz et al., 2016b), leading to histone deacetylation and finally H3K27 trimethylation through recruitment of the polycomb repressive complex 2 (PRC2) (Chu et al., 2015; McHugh et al., 2015).

In the wake of the *XIST* discovery and its ongoing characterizing has been an explosion of reports on functional lncRNAs, primarily thought to regulate the flow of genetic information in diverse ways. The early discovered lncRNA *H19* is transcribed from an imprinted locus and is thought to negatively regulate cellular growth through suppression of *IGF2* (Bartolomei et al., 1991; Brannan et al., 1990; Ripoche et al., 1997). LncRNA *MALAT1* has been shown to regulate RNA alternative splicing through interactions with SR proteins (Tripathi et al., 2010). The *HOXC* locus lncRNA *HOTAIR* regulates both *HOXD* and more global gene expression through interaction with PRC2 and LSD1, leading to transcriptional repression via heterochromatinization (Gupta et al., 2011; Rinn et al., 2007; Tsai et al., 2010). The concept of lncRNAs as molecular scaffolds that "guide" chromatin modifying enzymes has been adopted by many reports on other lncRNAs, including *ANRIL* (Yap et al., 2010), *MEG3*(Zhao et al., 2010a),

KCNQ1OT1 (Pandey et al., 2008), *Evf2/DLX6AS* (Berghoff et al., 2013; Bond et al., 2009), and *MORRBID* (Kotzin et al., 2016). LncRNAs are not restricted to functions in the nucleus. Cytoplasmic lncRNAs such as *NRON* (Willingham et al., 2005)and *UCHL1* (Carrieri et al., 2012)can regulate protein trafficking and translation, respectively (Mercer and Mattick, 2013).

However, recent observations suggest that many lncRNAs act not as free floating molecular agents but as local regulators of chromatin structure, such as through formation of chromatin loops between distal promoters and enhancers (e.g. *LUNAR1, CCAT1)* (Ma et al., 2015; Trimarchi et al., 2014; Xiang et al., 2014). The function of some lncRNAs are also thought to result from the act of transcription itself (Engreitz et al., 2016b; Kornienko et al., 2013; Latos et al., 2012; Li et al., 2013; Mele and Rinn, 2016), consistent with the relatively lower sequence conservation of lncRNA gene bodies (Hon et al., 2017; Ponting et al., 2009). Intriguingly, studies have also shown that the DNA loci of certain lncRNAs are responsible for the cellular phenotypes originally attributed to the lncRNA transcripts (Groff et al., 2016; Paralkar et al., 2016; Sauvageau et al., 2013). To add even more complexity to the study of lncRNA function, some lncRNA transcripts have different functions from their host DNA locus, as in the case of *HAUNT* (Yin et al., 2015). Therefore, while there is great excitement in the many possible flavors of lncRNA roles and mechanisms, caution must be exercised before arriving at definitive answers.

lncRNAs in development and disease

One of the striking features of lncRNAs is dynamic and tissue specific expression patterns, supporting the importance of lncRNAs in mammalian development (Cabili et al., 2011; Carninci et al., 2005; Derrien et al., 2012; Liu et al., 2016; Ramos et al., 2013; Tsoi et al., 2015).

LncRNAs are critical to the maintenance of pluripotency in embryonic stem cells (Guttman et al., 2011) and also to the formation and physiology of specialized cell types such as neural stem cells (Lin et al., 2014; Liu et al., 2016; Ramos et al., 2015), GABAergic interneurons (Bond et al., 2009; Briggs et al., 2015), adipocytes (Hacisuleyman et al., 2014), lung epithelia (Herriges et al., 2014), and cardiomyocytes (Grote et al., 2013; Klattenhoff et al., 2013). One organ system in particular, the nervous system, has implicated lncRNAs in a host of different functions (Briggs et al., 2015), from fate specification (Bond et al., 2009; Chalei et al., 2014; Lin et al., 2014; Ramos et al., 2015), to neurite outgrowth (Modarresi et al., 2012), and synaptogenesis (Bernard et al., 2010). Intriguingly, the lncRNA *HAR1* is transcribed in the brain from a genomic region under rapid evolutionary acceleration from chimpanzee to human, supporting a role of lncRNAs in the complexity of the human cerebral cortex (Pollard et al., 2006).

LncRNAs are also involved in the pathogenesis of disease, or at least correlated with the disease state (Huarte, 2015; Prensner and Chinnaiyan, 2011; Schmitt and Chang, 2016). Because of prominent examples of lncRNA function in various cancers and their restricted expression patterns, lncRNAs may be ideal biomarkers for diagnosis or disease progression, as well as therapeutic targets (Mouraviev et al., 2016). Some of the earliest studied lncRNAs are known to be aberrantly expressed in cancer. *H19* overexpression is characteristic of several cancers, including Wilms' tumor (Rainier et al., 1993), and *MALAT1* is important for metastasis of lung and breast cancer (Gutschner et al., 2013; Ji et al., 2003). The prostate cancer specific lncRNA *PCA3* is even used clinically as a biomarker for prostate cancer progression (Hessels and Schalken, 2009; Lee et al., 2011). *ANRIL* is overexpressed in prostate and gastric cancers, leading to silencing of its neighboring tumor suppressor gene, *CDKN2B* (also known as *INK4a/ARF*) (Kotake et al., 2011; Yap et al., 2010; Zhang et al., 2014). Breast cancer metastasis

also involves *HOTAIR*-mediated reprograming of the epigenetic state (Gupta et al., 2011). Colorectal cancer implicates *CCAT1* in stabilizing distal *MYC* chromatin looping (McCleland et al., 2016; Xiang et al., 2014), and also *PVT1* is frequently amplified in many tumors (Tseng et al., 2014). Not all cancer-associated lncRNAs act through overexpression, however. The glucocorticoid receptor antagonist lncRNA *GAS5* is downregulated in breast cancer, which presumably allows abnormally high hormone receptor signaling that increases fitness of the cancer cells (Mourtada-Maarabouni et al., 2009). One intriguing area of lncRNA cancer research is in brain tumors. *MEG3* and *NBAT1* have been shown to contribute to the etiology of meningioma and neuroblastoma, respectively (Pandey et al., 2014; Zhang et al., 2010), but relatively little is known about the role of lncRNAs in glioblastoma, the most fatal primary brain tumor.

Sequencing approaches to studying lncRNAs

Prerequisites to functional studies of lncRNAs involve accurate annotations of their transcript structure, e.g. what is the primary sequence, where is the promoter, where are the splice sites, and are there any splice variances. The explosion in the number of diverse lncRNAs annotated in the past decade is due mostly to advancements in not only RNA sequencing technology, but to computational algorithms that reconstruct RNA transcript structure from short sequencing reads (Garber et al., 2011; Guttman et al., 2010; Steijger et al., 2013). As the cost of high throughput sequencing decreased, the sequencing depth used in studies increased and so did the number of tissues and cell types examined, all leading to the mapping of tens of thousands of lncRNAs (Hon et al., 2017; Iyer et al., 2015). Microaray data, which were intended to probe protein coding gene expression, have also been retrospectively re-annotated to analyze lncRNA expression

across hundreds of clinical samples (Du et al., 2013). Long read sequencing and targeted resequencing of captured transcripts (CaptureSeq) have been useful for revealing novel isoforms and quantitative expression levels of lncRNAs (Clark et al., 2015; Ramos et al., 2013).

One technology in particular, single cell RNA-seq, has transformed our approach to transcriptome analysis (Linnarsson and Teichmann, 2016). While standard RNA sequencing approaches ("bulk" RNA-seq) average the expression profiles of genes across thousands or millions of cells, single cell RNA-seq allows quantitative measurement of global RNA expression in individual cells, even from complex tissues such as the cerebral cortex (Liu et al., 2016; Pollen et al., 2015; 2014; Zeisel et al., 2015) and tumors such as glioblastoma (Müller et al., 2016; Patel et al., 2014). These and other studies revealed many insights, including the vast heterogeneity of different cell types formed during development and disease, and also the bimodal or "bursting" nature of RNA transcription (Shalek et al., 2013; Tirosh et al., 2016). A pressing question in lncRNA biology has been whether the apparent low expression levels of lncRNAs in RNA sequencing data has been due to uniform low abundance across all cells, or small numbers of high lncRNA expressing cells being averaged out by bulk methods. Although its quantitative accuracy is still improving, single cell RNA-seq is an ideal method for addressing this question. Recent work demonstrated that exquisite cell type-specific expression of lncRNAs in heterogeneous tissues such as the brain can explain their low abundance at the population level. For instance, lncRNA *DLX6-AS1* is very highly expressed within inhibitory interneurons, but the relative dearth of these cells in early fetal brain development results in muted expression levels in tissue RNA-seq (Liu et al., 2016). However, single molecule fluorescence *in situ* hybridization (FISH) measuring the abundance of 61 lncRNAs in cell culture demonstrated relatively uniform expression (Cabili et al., 2015). It remains to be determined whether

discrepancies between these approaches are related to differences in lncRNA expression patterns in cell cultures versus heterogeneous tissues.

Looking beyond features of the lncRNA primary sequence, various methods have been developed to understand the secondary and tertiary structure of lncRNAs, which may provide clues about their function and biophysical properties (Wan et al., 2011). SHAPE-seq and PARIS utilize chemical crosslinking of base paired RNA and selective deep sequencing of paired and unpaired RNA, in order to map RNA secondary structure (Ding et al., 2014; Lu et al., 2016; Lucks et al., 2011). Since lncRNAs typically act together with proteins as ribonucleoprotein complexes, there is great interest in approaches to accurately identify protein binding partners of lncRNAs. Earlier efforts focused on surveying RNAs that associated with a given protein using methods such as RNA immunoprecipitation sequencing (RIP-seq) (Zhao et al., 2010b). Other approaches involve incubation of labeled RNA with nuclear or cell lysate, followed by RNA pull down and mass spectrometry of interacting proteins (Kretz et al., 2013; Ramos et al., 2015; Tacheny et al., 2013). Most recently, there has been a push to elucidate *in vivo* lncRNA-protein interactions, using techniques such as aptamer tagging of lncRNAs through gene editing (Butter et al., 2009), and high resolution cross linking of RNA and protein while they are still inside cells (e.g. ChIRP-MS, RAP-MS) (Chu et al., 2015; McHugh et al., 2015). While these descriptive genomics and other methods have been instrumental to our understanding of lncRNA biology, more direct functional genomics studies are needed to determine the roles of lncRNAs.

Functional genomics approaches to studying lncRNAs

Direct perturbation-based approaches such as deletion or knockdown experiments are instrumental for understanding lncRNA function (Bassett et al., 2014). In particular, systematic

approaches are especially suitable, due to the large number of lncRNAs annotated whose functions are not known. RNA interference (RNAi) using small interfering RNAs or short hairpin RNAs have been extensively applied to knockdown lncRNAs, even in pooled genetic screens, (Guttman et al., 2011; Lin et al., 2014; Ramos et al., 2015). However, many lncRNAs are enriched in the nucleus, where RNAi exhibits variable knockdown efficiency (Zeng and Cullen, 2002). Antisense oligonucleotides (ASOs), which act through RNase H mediated degradation of complementary RNA transcripts, exhibit long half-lives, efficient depletion of nuclear lncRNAs, and are also promising for therapeutic use in the clinic (Gutschner et al., 2013; Meng et al., 2015; Monteleone et al., 2015). Gene knockouts conducted in model organisms and human cells have also yielded important biological insights about lncRNAs, but this approach is difficult to scale (Aparicio-Prat et al., 2015; Ho et al., 2014; Meller and Rattner, 2002; Sauvageau et al., 2013). Introduction of insertion/deletion mutations via CRISPR/Cas9 nuclease is useful for targeted loss of function studies of protein coding genes by altering the coding frame, but they are not suitable for the study of lncRNA gene function, since small deletions do not generally disrupt lncRNA biological activity (Shalem et al., 2014; Shi et al., 2015; Wang et al., 2014). Nonetheless, larger Cas9-mediated genetic deletions can be effective at eliminating lncRNA genes (Bassett et al., 2014; Groff et al., 2016; Paralkar et al., 2016; Yin et al., 2015; Zhu et al., 2016).

Engineered CRISPR/Cas9 systems utilizing nuclease-dead dCas9 fused to either a repressor or activator domain has proved effective at controlling lncRNA transcription (Gilbert et al., 2014; 2013; Liu et al., 2017; 2016; Mandegar et al., 2016; Qi et al., 2013). In particular, CRISPR interference (CRISPRi), which targets the dCas9-KRAB repressor to the transcription start site (TSS) of any gene by a single guide RNA (sgRNA), has been used to screen over

16,000 lncRNA loci across 7 different cancer and non-cancer cell lines (Liu et al., 2017). Since CRISPRi has a small window of activity (1kb) around the targeted TSS (Gilbert et al., 2014), and as dCas9 occupies only 23bp of the targeted DNA strand (Nishimasu et al., 2015), CRISPRi allows for precise repression of any lncRNA gene. By placing repressive chromatin modifications around the TSS and serving as a transcriptional roadblock, CRISPRi tests a broad range of potential lncRNA functions: the production of *cis*- and *trans*-acting RNA (Rinn and Chang, 2012), *cis* gene regulation from transcription itself (Engreitz et al., 2016a; Kornienko et al., 2013; Li et al., 2013; Ørom et al., 2010), and enhancer function of the actual DNA loci (Fulco et al., 2016; Paralkar et al., 2016; Yin et al., 2015). Therefore, dissecting the contributions of the lncRNA transcript versus the DNA locus from which it's transcribed will require additional experiments, such as massively parallel reporter assays and genetic knock-ins of poly adenylation sites (Engreitz et al., 2016b; Kheradpour et al., 2013). The synthesis of these and other approaches with a modicum of creativity will undoubtedly reveal new surprises in lncRNA biology.

References

Anderson, D.M., Anderson, K.M., Chang, C.-L., Makarewich, C.A., Nelson, B.R., McAnally, J.R., Kasaragod, P., Shelton, J.M., Liou, J., Bassel-Duby, R., et al. (2015). A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. Cell *160*, 595–606.

Aparicio-Prat, E., Arnan, C., Sala, I., Bosch, N., Guigó, R., and Johnson, R. (2015). DECKO: Single-oligo, dual-CRISPR deletion of genomic elements including long non-coding RNAs. BMC Genomics *16*, 846.

Banfai, B., Jia, H., Khatun, J., Wood, E., Risk, B., Gundling, W.E., Kundaje, A., Gunawardena, H.P., Yu, Y., Xie, L., et al. (2012). Long noncoding RNAs are rarely translated in two human cell lines. Genome Research *22*, 1646–1657.

BARR, M.L., and BERTRAM, E.G. (1949). A Morphological Distinction between Neurones of the Male and Female, and the Behaviour of the Nucleolar Satellite during Accelerated Nucleoprotein Synthesis. Nature *163*, 676–677.

Bartolomei, M.S., Zemel, S., and Tilghman, S.M. (1991). Parental imprinting of the mouse H19 gene. Nature *351*, 153–155.

Bassett, A.R., Akhtar, A., Barlow, D.P., Bird, A.P., Brockdorff, N., Duboule, D., Ephrussi, A., Ferguson-Smith, A.C., Gingeras, T.R., Haerty, W., et al. (2014). Considerations when investigating lncRNA function in vivo. eLife *3*, e03058.

Berghoff, E.G., Clark, M.F., Chen, S., Cajigas, I., Leib, D.E., and Kohtz, J.D. (2013). Evf2 (Dlx6as) lncRNA regulates ultraconserved enhancer methylation and the differential transcriptional control of adjacent genes. Development *140*, 4407–4416.

Bernard, D., Prasanth, K.V., Tripathi, V., Colasse, S., Nakamura, T., Xuan, Z., Zhang, M.Q., Sedel, F., Jourdren, L., Coulpier, F., et al. (2010). A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression. The EMBO Journal *29*, 3082–3093.

Bond, A.M., VanGompel, M.J.W., Sametsky, E.A., Clark, M.F., Savage, J.C., Disterhoft, J.F., and Kohtz, J.D. (2009). Balanced gene regulation by an embryonic brain ncRNA is critical for adult hippocampal GABA circuitry. Nat Neurosci *12*, 1020–1027.

Borsani, G., Tonlorenzi, R., Simmler, M.C., Dandolo, L., Arnaud, D., Capra, V., Grompe, M., Pizzuti, A., Muzny, D., Lawrence, C., et al. (1991). Characterization of a murine gene expressed from the inactive X chromosome. Nature *351*, 325–329.

Brannan, C.I., Dees, E.C., and Ingram, R.S. (1990). The product of the H19 gene may function as an RNA. Molecular and Cellular ….

Briggs, J.A., Wolvetang, E.J., Mattick, J.S., Rinn, J.L., and Barry, G. (2015). Mechanisms of Long Non-coding RNAs in Mammalian Nervous System Development, Plasticity, Disease, and Evolution. Neuron *88*, 861–877.

Brown, C.J., Ballabio, A., Rupert, J.L., Lafreniere, R.G., Grompe, M., Tonlorenzi, R., and Willard, H.F. (1991). A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. Nature *349*, 38–44.

Butter, F., Scheibe, M., Mörl, M., and Mann, M. (2009). Unbiased RNA-protein interaction screen by quantitative proteomics. Proceedings of the National Academy of Sciences *106*, 10626–10631.

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

Cabili, M.N., Dunagin, M.C., McClanahan, P.D., Biaesch, A., Padovan-Merhar, O., Regev, A., Rinn, J.L., and Raj, A. (2015). Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. Genome Biol *16*, 20.

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

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

Chalei, V., Sansom, S.N., Kong, L., Lee, S., Montiel, J.F., Vance, K.W., and Ponting, C.P. (2014). The long non-coding RNA Dali is an epigenetic regulator of neural differentiation. eLife *3*, e04530.

Chen, C.-K., Blanco, M., Jackson, C., Aznauryan, E., Ollikainen, N., Surka, C., Chow, A., Cerase, A., McDonel, P., and Guttman, M. (2016). Xist recruits the X chromosome to the nuclear lamina to enable chromosome-wide silencing. Science *354*, 468–472.

Chu, C., Zhang, Q.C., da Rocha, S.T., Flynn, R.A., Bharadwaj, M., Calabrese, J.M., Magnuson, T., Heard, E., and Chang, H.Y. (2015). Systematic discovery of Xist RNA binding proteins. Cell *161*, 404–416.

Clark, M.B., Mercer, T.R., Bussotti, G., Leonardi, T., Haynes, K.R., Crawford, J., Brunck, M.E., Cao, K.-A.L., Thomas, G.P., Chen, W.Y., et al. (2015). Quantitative gene profiling of long noncoding RNAs with targeted RNA sequencing. Nat Meth *12*, 339–342.

Crick, F. (1970). Central dogma of molecular biology. Nature *227*, 561–563.

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

Ding, Y., Tang, Y., Kwok, C.K., Zhang, Y., Bevilacqua, P.C., and Assmann, S.M. (2014). In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature *505*, 696–700.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. Nature *489*, 101–108.

Du, Z., Fei, T., Verhaak, R.G.W., Su, Z., Zhang, Y., Brown, M., Chen, Y., and Liu, X.S. (2013). Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. Nature Structural & Molecular Biology *20*, 908–913.

Engreitz, J.M., Haines, J.E., Munson, G., Chen, J., and Perez, E.M. (2016a). Neighborhood regulation by lncRNA promoters, transcription, and splicing. bioRxiv.

Engreitz, J.M., Haines, J.E., Perez, E.M., Munson, G., Chen, J., Kane, M., McDonel, P.E.,

Guttman, M., and Lander, E.S. (2016b). Local regulation of gene expression by lncRNA promoters, transcription and splicing. Nature *539*, 452–455.

FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest, A.R.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J.L., Haberle, V., Lassmann, T., Kulakovskiy, I.V., Lizio, M., et al. (2014). A promoter-level mammalian expression atlas. Nature *507*, 462–470.

Fulco, C.P., Munschauer, M., Anyoha, R., Munson, G., Grossman, S.R., Perez, E.M., Kane, M., Cleary, B., Lander, E.S., and Engreitz, J.M. (2016). Systematic mapping of functional enhancer-promoter connections with CRISPR interference. Science 1–13.

Garber, M., Grabherr, M.G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Meth *8*, 469–477.

Gilbert, L.A., Horlbeck, M.A., Adamson, B., Villalta, J.E., Chen, Y., Whitehead, E.H., Guimaraes, C., Panning, B., Ploegh, H.L., Bassik, M.C., et al. (2014). Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. Cell *159*, 647–661.

Gilbert, L.A., Larson, M.H., Morsut, L., Liu, Z., Brar, G.A., Torres, S.E., Stern-Ginossar, N., Brandman, O., Whitehead, E.H., Doudna, J.A., et al. (2013). CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell *154*, 442–451.

Groff, A.F., Sanchez-Gomez, D.B., Soruco, M.M.L., Gerhardinger, C., Barutcu, A.R., Li, E., Elcavage, L., Plana, O., Sanchez, L.V., Lee, J.C., et al. (2016). In Vivo Characterization of Linc-p21 Reveals Functional cis-Regulatory DNA Elements. Cell Reports *16*, 2178–2186.

Grote, P., Wittler, L., Hendrix, D., Koch, F., Währisch, S., Beisaw, A., Macura, K., Bläss, G.,

Kellis, M., Werber, M., et al. (2013). The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. Developmental Cell *24*, 206–214.

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. (2011). Long non-coding RNA HOTAIR reprograms chromatinstate to promote cancer metastasis. Nature *464*, 1071–1076.

Gutschner, T., Hämmerle, M., Eissmann, M., Hsu, J., Kim, Y., Hung, G., Revenko, A., Arun, G., Stentrup, M., Gross, M., et al. (2013). The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. Cancer Research *73*, 1180–1189.

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. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223–227.

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. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature *477*, 295–300.

Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C., et al. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol *28*, 503–510.

Guttman, M., Russell, P., Ingolia, N.T., Weissman, J.S., and Lander, E.S. (2013). Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. Cell *154*, 240–251.

Hacisuleyman, E., Goff, L.A., Trapnell, C., Williams, A., Henao-Mejia, J., Sun, L., McClanahan, P., Hendrickson, D.G., Sauvageau, M., Kelley, D.R., et al. (2014). Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. Nature Structural & Molecular Biology *21*, 198–206.

Herriges, M.J., Swarr, D.T., Morley, M.P., Rathi, K.S., Peng, T., Stewart, K.M., and Morrisey, E.E. (2014). Long noncoding RNAs are spatially correlated with transcription factors and regulate lung development. Genes & Development *28*, 1363–1379.

Hessels, D., and Schalken, J.A. (2009). The use of PCA3 in the diagnosis of prostate cancer. Nat Rev Urol *6*, 255–261.

Ho, T.-T., Zhou, N., Huang, J., Koirala, P., Xu, M., Fung, R., Wu, F., and Mo, Y.-Y. (2014). Targeting non-coding RNAs with the CRISPR/Cas9 system in human cell lines. Nucleic Acids Res. *43*, gku1198–e17.

Hon, C.-C., Ramilowski, J.A., Harshbarger, J., Bertin, N., Rackham, O.J.L., Gough, J., Denisenko, E., Schmeier, S., Poulsen, T.M., Severin, J., et al. (2017). An atlas of human long non-coding RNAs with accurate 5' ends. Nature.

Huarte, M. (2015). The emerging role of lncRNAs in cancer. Nat Med *21*, 1253–1261.

Ingolia, N.T., Brar, G.A., Stern-Ginossar, N., Harris, M.S., Talhouarne, G.J.S., Jackson, S.E., Wills, M.R., and Weissman, J.S. (2014). Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. Cell Reports *8*, 1365–1379.

Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner,

J.R., Evans, J.R., Zhao, S., et al. (2015). The landscape of long noncoding RNAs in the human transcriptome. Nature Genetics.

Ji, P., Diederichs, S., Wang, W., Böing, S., Metzger, R., Schneider, P.M., Tidow, N., Brandt, B., Buerger, H., Bulk, E., et al. (2003). MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. Oncogene *22*, 8031–8041.

Kapranov, P., Cawley, S.E., Drenkow, J., Bekiranov, S., Strausberg, R.L., Fodor, S.P.A., and Gingeras, T.R. (2002). Large-scale transcriptional activity in chromosomes 21 and 22. Science *296*, 916–919.

Kheradpour, P., Ernst, J., Melnikov, A., Rogov, P., Wang, L., Zhang, X., Alston, J., Mikkelsen, T.S., and Kellis, M. (2013). Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. Genome Research *23*, 800–811.

Klattenhoff, C.A., Scheuermann, J.C., Surface, L.E., Bradley, R.K., Fields, P.A., Steinhauser, M.L., Ding, H., Butty, V.L., Torrey, L., Haas, S., et al. (2013). Braveheart, a Long Noncoding RNA Required for Cardiovascular Lineage Commitment. Cell *152*, 570–583.

Kornienko, A.E., Guenzl, P.M., Barlow, D.P., and Pauler, F.M. (2013). Gene regulation by the act of long non-coding RNA transcription. BMC Biol. *11*, 59.

Kotake, Y., Nakagawa, T., Kitagawa, K., Suzuki, S., Liu, N., Kitagawa, M., and Xiong, Y. (2011). Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15(INK4B) tumor suppressor gene. Oncogene *30*, 1956–1962.

Kotzin, J.J., Spencer, S.P., McCright, S.J., Kumar, D.B.U., Collet, M.A., Mowel, W.K., Elliott, E.N., Uyar, A., Makiya, M.A., Dunagin, M.C., et al. (2016). The long non-coding RNA Morrbid regulates Bim and short-lived myeloid cell lifespan. Nature *537*, 239–243.

Kretz, M., Siprashvili, Z., Chu, C., Webster, D.E., Zehnder, A., Qu, K., Lee, C.S., Flockhart, R.J., Groff, A.F., Chow, J., et al. (2013). Control of somatic tissue differentiation by the long non-coding RNA TINCR. Nature *493*, 231–235.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Raymond, C., and Szustakowki, J. (2001). Initial sequencing and analysis of the human genome. Nature.

Latos, P.A., Pauler, F.M., Koerner, M.V., Şenergin, H.B., Hudson, Q.J., Stocsits, R.R., Allhoff, W., Stricker, S.H., Klement, R.M., Warczok, K.E., et al. (2012). Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. Science *338*, 1469–1472.

Lee, G.L., Dobi, A., and Srivastava, S. (2011). Prostate cancer: diagnostic performance of the PCA3 urine test. Nat Rev Urol *8*, 123–124.

Li, W., Notani, D., Ma, Q., Tanasa, B., Nunez, E., Chen, A.Y., Merkurjev, D., Zhang, J., Ohgi, K., Song, X., et al. (2013). Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature *498*, 516–520.

Lin, M.F., Jungreis, I., and Kellis, M. (2011). PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics *27*, i275–i282.

Lin, N., Chang, K.-Y., Li, Z., Gates, K., Rana, Z.A., Dang, J., Zhang, D., Han, T., Yang, C.-S., Cunningham, T.J., et al. (2014). An evolutionarily conserved long noncoding RNA TUNA

controls pluripotency and neural lineage commitment. Molecular Cell *53*, 1005–1019.

Linnarsson, S., and Teichmann, S.A. (2016). Single-cell genomics: coming of age. Genome Biol *17*, 97.

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., et al. (2017). CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. Science *355*, eaah7111.

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., et al. (2016). Single-cell analysis of long non-coding RNAs in the developing human neocortex. Genome Biol *17*, 67.

Lu, Z., Zhang, Q.C., Lee, B., Flynn, R.A., Smith, M.A., Robinson, J.T., Davidovich, C., Gooding, A.R., Goodrich, K.J., Mattick, J.S., et al. (2016). RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. Cell *165*, 1267–1279.

Lucks, J.B., Mortimer, S.A., Trapnell, C., Luo, S., Aviran, S., Schroth, G.P., Pachter, L., Doudna, J.A., and Arkin, A.P. (2011). Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). Proceedings of the National Academy of Sciences *108*, 11063–11068.

Ma, W., Ay, F., Lee, C., Gulsoy, G., Deng, X., Cook, S., Hesson, J., Cavanaugh, C., Ware, C.B., Krumm, A., et al. (2015). Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. Nat Meth *12*, 71–78.

Mandegar, M.A., Huebsch, N., Frolov, E.B., Shin, E., Truong, A., Olvera, M.P., Chan, A.H.,

Miyaoka, Y., Holmes, K., Spencer, C.I., et al. (2016). CRISPR Interference Efficiently Induces Specific and Reversible Gene Silencing in Human iPSCs. Cell Stem Cell *18*, 541–553.

Mattick, J.S., and Rinn, J.L. (2015). Discovery and annotation of long noncoding RNAs. Nature Structural & Molecular Biology *22*, 5–7.

McCleland, M.L., Mesh, K., Lorenzana, E., Chopra, V.S., Segal, E., Watanabe, C., Haley, B., Mayba, O., Yaylaoglu, M., Gnad, F., et al. (2016). CCAT1 is an enhancer-templated RNA that predicts BET sensitivity in colorectal cancer. J. Clin. Invest. *126*, 639–652.

McHugh, C.A., Chen, C.-K., Chow, A., Surka, C.F., Tran, C., McDonel, P., Pandya-Jones, A., Blanco, M., Burghard, C., Moradian, A., et al. (2015). The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. Nature *521*, 232–236.

Mele, M., and Rinn, J.L. (2016). "Cat's Cradling" the 3D Genome by the Act of LncRNA Transcription. Molecular Cell *62*, 657–664.

Meller, V.H., and Rattner, B.P. (2002). The roX genes encode redundant male-specific lethal transcripts required for targeting of the MSL complex. The EMBO Journal *21*, 1084–1091.

Meng, L., Ward, A.J., Chun, S., Bennett, C.F., Beaudet, A.L., and Rigo, F. (2015). Towards a therapy for Angelman syndrome by targeting a long non-coding RNA. Nature *518*, 409–412.

Mercer, T.R., and Mattick, J.S. (2013). Structure and function of long noncoding RNAs in epigenetic regulation. Nature Structural & Molecular Biology *20*, 300–307.

Modarresi, F., Faghihi, M.A., Lopez-Toledano, M.A., Fatemi, R.P., Magistri, M., Brothers, S.P., van der Brug, M.P., and Wahlestedt, C. (2012). Inhibition of natural antisense transcripts in vivo

results in gene-specific transcriptional upregulation. Nat Biotechnol *30*, 453–459.

Monteleone, G., Neurath, M.F., Ardizzone, S., Di Sabatino, A., Fantini, M.C., Castiglione, F., Scribano, M.L., Armuzzi, A., Caprioli, F., Sturniolo, G.C., et al. (2015). Mongersen, an oral SMAD7 antisense oligonucleotide, and Crohn's disease. N Engl J Med *372*, 1104–1113.

Mouraviev, V., Lee, B., Patel, V., Albala, D., Johansen, T.E.B., Partin, A., Ross, A., and Perera, R.J. (2016). Clinical prospects of long noncoding RNAs as novel biomarkers and therapeutic targets in prostate cancer. Prostate Cancer Prostatic Dis. *19*, 14–20.

Mourtada-Maarabouni, M., Pickard, M.R., Hedge, V.L., Farzaneh, F., and Williams, G.T. (2009). GAS5, a non-protein-coding RNA, controls apoptosis and is downregulated in breast cancer. Oncogene *28*, 195–208.

Müller, S., Liu, S.J., Di Lullo, E., Malatesta, M., Pollen, A.A., Nowakowski, T.J., Kohanbash, G., Aghi, M., Kriegstein, A.R., Lim, D.A., et al. (2016). Single-cell sequencing maps gene expression to mutational phylogenies in PDGF- and EGF-driven gliomas. Molecular Systems Biology *12*, 889.

Nelson, B.R., Makarewich, C.A., Anderson, D.M., Winders, B.R., Troupes, C.D., Wu, F., Reese, A.L., McAnally, J.R., Chen, X., Kavalali, E.T., et al. (2016). A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. Science *351*, 271–275.

Nishimasu, H., Cong, L., Yan, W.X., Ran, F.A., Zetsche, B., Li, Y., Kurabayashi, A., Ishitani, R., Zhang, F., and Nureki, O. (2015). Crystal Structure of Staphylococcus aureus Cas9. Cell *162*, 1113–1126.

Pandey, G.K., Mitra, S., Subhash, S., Hertwig, F., Kanduri, M., Mishra, K., Fransson, S., Ganeshram, A., Mondal, T., Bandaru, S., et al. (2014). The Risk-Associated Long Noncoding RNA NBAT-1 Controls Neuroblastoma Progression by Regulating Cell Proliferation and Neuronal Differentiation. Cancer Cell *26*, 722–737.

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

Paralkar, V.R., Taborda, C.C., Huang, P., Yao, Y., Kossenkov, A.V., Prasad, R., Luan, J., Davies, J.O.J., Hughes, J.R., Hardison, R.C., et al. (2016). Unlinking an lncRNA from Its Associated cis Element. Molecular Cell *62*, 104–110.

Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B.V., Curry, W.T., Martuza, R.L., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science *344*, 1396–1401.

Pollard, K.S., Salama, S.R., Lambert, N., Lambot, M.-A., Coppens, S., Pedersen, J.S., Katzman, S., King, B., Onodera, C., Siepel, A., et al. (2006). An RNA gene expressed during cortical development evolved rapidly in humans. Nature *443*, 167–172.

Pollen, A.A., Nowakowski, T.J., Chen, J., Retallack, H., Sandoval-Espinosa, C., Nicholas, C.R., Shuga, J., Liu, S.J., Oldham, M.C., Diaz, A., et al. (2015). Molecular identity of human outer radial glia during cortical development. Cell *163*, 55–67.

Pollen, A.A., Nowakowski, T.J., Shuga, J., Wang, X., Leyrat, A.A., Lui, J.H., Li, N.,

Szpankowski, L., Fowler, B., Chen, P., et al. (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat Biotechnol *32*, 1053–1058.

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

Prensner, J.R., and Chinnaiyan, A.M. (2011). The emergence of lncRNAs in cancer biology. Cancer Discov *1*, 391–407.

Qi, L.S., Larson, M.H., Gilbert, L.A., Doudna, J.A., Weissman, J.S., Arkin, A.P., and Lim, W.A. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell *152*, 1173–1183.

Rainier, S., Johnson, L.A., Dobry, C.J., Ping, A.J., Grundy, P.E., and Feinberg, A.P. (1993). Relaxation of imprinted genes in human cancer. Nature *362*, 747–749.

Ramos, A.D., Andersen, R.E., Liu, S.J., Nowakowski, T.J., Hong, S.J., Gertz, C.C., Salinas, R.D., Zarabi, H., Kriegstein, A.R., and Lim, D.A. (2015). The long noncoding RNA Pnky regulates neuronal differentiation of embryonic and postnatal neural stem cells. Cell Stem Cell *16*, 439–447.

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

Rinn, J.L., and Chang, H.Y. (2012). Genome regulation by long noncoding RNAs. Annu. Rev.

Biochem. *81*, 145–166.

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. (2007). Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs. Cell *129*, 1311–1323.

Ripoche, M.A., Kress, C., and Poirier, F. (1997). Deletion of the H19 transcription unit reveals the existence of a putative imprinting control element. Genes & ….

Sauvageau, M., Goff, L.A., Lodato, S., Bonev, B., Groff, A.F., Gerhardinger, C., Sanchez-Gomez, D.B., Hacisuleyman, E., Li, E., Spence, M., et al. (2013). Multiple knockout mouse models reveal lincRNAs are required for life and brain development. eLife *2*, e01749.

Schmitt, A.M., and Chang, H.Y. (2016). Long Noncoding RNAs in Cancer Pathways. Cancer Cell *29*, 452–463.

Shalek, A.K., Satija, R., Adiconis, X., Gertner, R.S., Gaublomme, J.T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., et al. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature *498*, 236–240.

Shalem, O., Sanjana, N.E., Hartenian, E., Shi, X., Scott, D.A., Mikkelsen, T.S., Heckl, D., Ebert, B.L., Root, D.E., Doench, J.G., et al. (2014). Genome-scale CRISPR-Cas9 knockout screening in human cells. Science *343*, 84–87.

Shi, J., Wang, E., Milazzo, J.P., Wang, Z., Kinney, J.B., and Vakoc, C.R. (2015). Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. Nat Biotechnol *33*, 661–667.

Steijger, T., Abril, J.F., Engström, P.G., and Kokocinski, F. (2013). Assessment of transcript

reconstruction methods for RNA-seq. Methods.

Sun, L., Luo, H., Bu, D., Zhao, G., Yu, K., Zhang, C., Liu, Y., Chen, R., and Zhao, Y. (2013). Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. Nucleic Acids Res. *41*, e166–e166.

Tacheny, A., Dieu, M., Arnould, T., and Renard, P. (2013). Mass spectrometry-based identification of proteins interacting with nucleic acids. J Proteomics *94*, 89–109.

Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science *352*, 189–196.

Trimarchi, T., Bilal, E., Ntziachristos, P., Fabbri, G., Dalla-Favera, R., Tsirigos, A., and Aifantis, I. (2014). Genome-wide Mapping and Characterization of Notch-Regulated Long Noncoding RNAs in Acute Leukemia. Cell *158*, 593–606.

Tripathi, V., Ellis, J.D., Shen, Z., Song, D.Y., Pan, Q., Watt, A.T., Freier, S.M., Bennett, C.F., Sharma, A., Bubulya, P.A., et al. (2010). The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. Molecular Cell *39*, 925–938.

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

Tseng, Y.-Y., Moriarity, B.S., Gong, W., Akiyama, R., Tiwari, A., Kawakami, H., Ronning, P.,

Reuland, B., Guenther, K., Beadnell, T.C., et al. (2014). PVT1 dependence in cancer with MYC copy-number increase. Nature *512*, 82–86.

Tsoi, L.C., Iyer, M.K., Stuart, P.E., Swindell, W.R., Gudjonsson, J.E., Tejasvi, T., Sarkar, M.K., Li, B., Ding, J., Voorhees, J.J., et al. (2015). Analysis of long non-coding RNAs highlights tissue-specific expression patterns and epigenetic profiles in normal and psoriatic skin. Genome Biol *16*, 24.

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

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science *291*, 1304–1351.

Wan, Y., Kertesz, M., Spitale, R.C., Segal, E., and Chang, H.Y. (2011). Understanding the transcriptome through RNA structure. Nat Rev Genet *12*, 641–655.

Wang, L., Park, H.J., Dasari, S., Wang, S., Kocher, J.-P., and Li, W. (2013). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res. *41*, e74–e74.

Wang, T., Wei, J.J., Sabatini, D.M., and Lander, E.S. (2014). Genetic screens in human cells using the CRISPR-Cas9 system. Science *343*, 80–84.

Willingham, A.T., Orth, A.P., Batalov, S., Peters, E.C., Wen, B.G., Aza-Blanc, P., Hogenesch, J.B., and Schultz, P.G. (2005). A strategy for probing the function of noncoding RNAs finds a

repressor of NFAT. Science *309*, 1570–1573.

Wutz, A. (2011). Gene silencing in X-chromosome inactivation: advances in understanding facultative heterochromatin formation. Nat Rev Genet *12*, 542–553.

Xiang, J.-F., Yin, Q.-F., Chen, T., Zhang, Y., Zhang, X.-O., Wu, Z., Zhang, S., Wang, H.-B., Ge, J., Lu, X., et al. (2014). Human colorectal cancer-specific CCAT1-L lncRNA regulates long-range chromatin interactions at the MYC locus. Cell Research *24*, 513–531.

Yap, K.L., Li, S., Muñoz-Cabello, A.M., Raguz, S., Zeng, L., Mujtaba, S., Gil, J., Walsh, M.J., and Zhou, M.-M. (2010). Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. Molecular Cell *38*, 662–674.

Yin, Y., Yan, P., Lu, J., Song, G., Zhu, Y., Li, Z., Zhao, Y., Shen, B., Huang, X., Zhu, H., et al. (2015). Opposing Roles for the lncRNA Haunt and Its Genomic Locus in Regulating HOXA Gene Activation during Embryonic Stem Cell Differentiation. Cell Stem Cell *16*, 504–516.

Zeisel, A., Muñoz-Manchado, A.B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science *347*, 1138–1142.

Zeng, Y., and Cullen, B.R. (2002). RNA interference in human cells is restricted to the cytoplasm. Rna *8*, 855–860.

Zhang, E.-B., Kong, R., Yin, D.-D., You, L.-H., Sun, M., Han, L., Xu, T.-P., Xia, R., Yang, J.-S., De, W., et al. (2014). Long noncoding RNA ANRIL indicates a poor prognosis of gastric cancer

and promotes tumor growth by epigenetically silencing of miR-99a/miR-449a. Oncotarget *5*, 2276–2292.

Zhang, X., Gejman, R., Mahta, A., Zhong, Y., Rice, K.A., Zhou, Y., Cheunsuchon, P., Louis, D.N., and Klibanski, A. (2010). Maternally expressed gene 3, an imprinted noncoding RNA gene, is associated with meningioma pathogenesis and progression. Cancer Research *70*, 2350–2358.

Zhao, J., Ohsumi, T.K., Kung, J.T., Ogawa, Y., and Grau, D.J. (2010a). Genome-wide identification of polycomb-associated RNAs by RIP-seq. Molecular Cell.

Zhao, J., Ohsumi, T.K., Kung, J.T., Ogawa, Y., Grau, D.J., Sarma, K., Song, J.J., Kingston, R.E., Borowsky, M., and Lee, J.T. (2010b). Genome-wide Identificationof Polycomb-Associated RNAs by RIP-seq. Molecular Cell *40*, 939–953.

Zhao, Y., Li, H., Fang, S., Kang, Y., Wu, W., Hao, Y., Li, Z., Bu, D., Sun, N., Zhang, M.Q., et al. (2016). NONCODE 2016: an informative and valuable data source of long non-coding RNAs. Nucleic Acids Res. *44*, D203–D208.

Zhu, S., Li, W., Liu, J., Chen, C.-H., Liao, Q., Xu, P., Xu, H., Xiao, T., Cao, Z., Peng, J., et al. (2016). Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR-Cas9 library. Nat Biotechnol.

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

Chapter 2

Single-cell analysis of long non-coding RNAs in the developing human neocortex

Summary

Long non-coding RNAs (lncRNAs) comprise a diverse class of transcripts that can regulate molecular and cellular processes in brain development and disease. LncRNAs exhibit cell type- and tissue-specific expression, but little is known about the expression and function of lncRNAs in the developing human brain. Furthermore, it has been unclear whether lncRNAs are highly expressed in subsets of cells within tissues, despite appearing lowly expressed in bulk populations. We use strand-specific RNA-seq to deeply profile lncRNAs from polyadenylated and total RNA obtained from human neocortex at different stages of development, and we apply this reference to analyze the transcriptomes of single cells. While lncRNAs are generally detected at low levels in bulk tissues, single cell transcriptomics of hundreds of neocortex cells reveal that many lncRNAs are abundantly expressed in individual cells and are cell type-specific. Notably, *LOC646329* is a lncRNA enriched in single radial glia cells but is detected at low abundance in tissues. CRISPRi knockdown of *LOC646329* indicates that this lncRNA regulates cell proliferation. The discrete and abundant expression of lncRNAs among individual cells has important implications for both their biological function and utility for distinguishing neural cell types.

Introduction

Long non-coding RNAs (lncRNAs), transcripts longer than 200 nt without protein coding potential, comprise upwards of 58,000 genes in the human genome and have important roles in neural development, function, and disease (Bond et al., 2009; Cabili et al., 2011; Goff et al., 2015; Hangauer et al., 2013; Iyer et al., 2015; Qureshi et al., 2010; Ramos et al., 2013; Sauvageau et al., 2013; Ulitsky and Bartel, 2013). LncRNAs exhibit tissue specific expression, with the brain producing an extraordinary number and diversity of lncRNAs (Cabili et al., 2011; Mercer et al., 2010; Ponting et al., 2009; Ramos et al., 2013). Furthermore, in the brain, lncRNAs have regionally segregated expression patterns (Mercer et al., 2008; Ramos et al., 2013), and many lncRNAs are enriched in specific sub-populations of the mouse(Molyneaux et al., 2015) and human(Johnson et al., 2015) cortex. However, little is known about lncRNA expression and function in the developing human brain.

Current annotations of lncRNAs expressed in the human brain are incomplete, partly due to the use of polyadenylated (polyA) transcript selection and RNA-seq libraries that do not preserve strand information (Darmanis et al., 2015; Miller et al., 2014). As a result, non-polyadenylated lncRNAs and antisense lncRNAs – several of which have known biological functions (Carrieri et al., 2012; Kotake et al., 2011; Pandey et al., 2008) – remain poorly described. Furthermore, the expression of lncRNAs in the human brain has not been systematically analyzed at the single cell level, limiting our understanding of temporal- and cell type-specific lncRNAs.

Bulk tissue studies have suggested that lncRNAs are expressed, on average, at lower levels than mRNAs. It has been unclear whether this is due to uniformly low levels of lncRNAs in all cells, or due to high levels of lncRNAs in subpopulations of cells. In the brain, the latter

explanation would suggest that lncRNAs previously thought to be transcriptional noise might have highly specialized roles in the differentiation or function of specific cell types. While studies of cultured cells provide evidence for both possibilities (Cabili et al., 2015; Shalek et al., 2013), whether these observations are consistent with the *in vivo* expression of lncRNAs in cells within heterogeneous tissues – such as that of the developing human brain – has been not been determined.

Here, we combined bulk tissue RNA-seq and single cell RNA-seq to deeply profile lncRNA expression during neocortical development. Both polyA selected and total RNA were sequenced using strand-specific methods to comprehensively annotate and quantify lncRNAs in tissues. By applying this reference transcriptome to single cell RNA-seq, we found that many lncRNAs are specific to distinct cell types and are abundantly expressed in individual cells. Furthermore, we found that the cell type-specific expression of lncRNAs contributes to the low levels of lncRNAs observed in tissues. Finally**,** using CRISPRi (clustered regularly interspaced short palindromic repeats interference) knockdown, we demonstrated that *LOC646329*, a lncRNA that appears low in neocortical tissues but high in the radial glia subpopulation, regulates cell proliferation.

Results

Catalogue of long non-coding RNAs in human neocortex development

To identify lncRNAs expressed during human neocortical development, we microdissected radial sections of the tissue at gestational weeks (GW) 13/14.5, 16, 21 and 23. For each time point, we obtained biological duplicates and performed strand-specific RNA-seq of both polyA selected RNA and total RNA that had been rRNA depleted, generating over 200 million 100 bp paired-end mapped reads from each tissue specimen (Figure 1A, Table S1). After *de novo* transcriptome assembly of the polyA selected RNA-seq reads, previously annotated genes and short transcripts (<200nt) were filtered. Transcript models that did not pass an optimized read coverage threshold in both biological replicates were removed (Methods, Figure S1A). We then analyzed the protein coding potential of the remaining transcripts using three computational tools: CPC, CPAT, and Pfam (Finn et al., 2014; Kong et al., 2007; Wang et al., 2013), and any transcripts assigned a protein coding status by any of the 3 methods were classified as transcripts of uncertain coding potential (TUCP) (Cabili et al., 2011) (Figure S1B). These newly annotated transcripts were classified as intergenic, antisense, or intronic according to previously proposed nomenclature standards (Mattick and Rinn, 2015) and merged with the Ensembl build 75 transcriptome, resulting in the Full transcriptome reference. A Stringent transcriptome reference, in which novel single-exon transcripts were removed, was also generated (all deposited in GSE71315).

In our polyA selected reference transcriptomes, we identified 11642 lncRNAs (4124 multi-exonic) and 2571 TUCPs expressed in developing human neocortex (Figure 1B; Table S2). The majority of lncRNAs were intergenic, though strand specific RNA-seq enabled identification of 3047 antisense lncRNAs (Figure 1C). 8180 lncRNAs were novel to Ensembl 75/GENCODE

v19 (Table S3). 7492, 7892, and 2105 were novel to the annotations of Cabili et. al. 2011, Hangauer et. al. 2013, and Mitranscriptome (Iyer et al., 2015), respectively (Figure S1C). On average, lncRNAs were detected at 13.6 fold lower levels than mRNAs (Figure 1D). Novel polyA transcripts annotated from human brain tissues had genomic characteristics and conservation scores similar to previously annotated lncRNAs (Figure S2).

We next performed pairwise whole-transcriptome comparisons of all time points using DESeq2 (Anders and Huber, 2010) (FDR < 0.01). 1088 mRNAs and 424 polyA lncRNAs/TUCPs were differentially expressed across these time points (Figure 2A, Table S4). Among differentially expressed mRNAs, *PAX6* and *CENPA* were elevated in GW13-16, suggesting the increased presence of radial glia stem cells (Heins et al., 2002). Conversely, *CUX2* and *ADCY1* were elevated in GW21-23, consistent with increased neurogenesis at these time points (Molyneaux et al., 2007). Among differentially expressed lncRNAs, *MEG3* and *DLX6-AS1* (a lncRNA antisense to the interneuron transcription factor *DLX6),* increased with developmental progression (Figure 2B). Furthermore, gene ontology (GO) analysis of lncRNA gene neighbors suggested their function in neuronal differentiation (Figure S1F).

Some lncRNAs, such as *MALAT1*, have been described to be non-polyadenylated (Wilusz et al., 2008). To expand our catalogue to include non-polyA lncRNAs, we performed *de novo* transcriptome assembly with sequencing data from the total RNA (rRNA depleted) from each tissue sample. Full and Stringent lncRNA/TUCP references were generated with the same pipeline used for polyA selected transcripts (Figure 1A). 26241 lncRNAs (4477 multi-exonic) and 4606 TUCPs were annotated from the total RNA-seq libraries (Figure S1E). To identify transcripts that are likely to be non-polyA, we analyzed genes that were consistently > 10 fold enriched in the total RNA libraries versus the polyA libraries across all samples (Figure 2C,

Figure S3, Table S5). mRNAs that encode specific histone subunits are known to be non-polyA (Yang et al., 2011), and 52 out of the 58 mRNAs enriched in the total RNA-seq transcriptomes were for histone subunits, including *HIST1H2BK* and *HIST2H2AB*. By these methods, 85 lncRNAs were identified as non-polyA, with 65 being novel to Ensembl. Among the previously annotated lncRNAs in this set, known non-polyadenylated lncRNAs such as MALAT1, RMRP, and TERC, the RNA component of telomerase, were identified (Livyatan et al., 2013). Thus, our transcriptome references allow broad profiling of lncRNAs during brain development, regardless of genomic orientation or polyadenylation status.

Single cell RNA-seq analysis of lncRNA expression

RNA-seq of whole tissues averages gene expression signatures of many different cell types (Shalek et al., 2013). To study lncRNA expression at single cell resolution, we captured single cells from radial sections of GW19.5, GW20.5, and GW23.5 neocortex (Figure 3A). To mitigate the effects of technical noise, we added equal amounts of ERCC Spike-In Control RNA to each single cell lysis reaction. PolyA libraries were generated, and a median of 1 million mapped read pairs were obtained for each reaction (Table S1). We utilized our polyA Stringent transcriptome reference to perform transcriptome-guided genomic alignment and gene-level quantification of single cell RNA-seq reads (Methods). Cells in which we detected > 1000 genes and > 40 ERCC species were retained, resulting in 226 single cells for consideration. We also included 50 single cell libraries from GW16 and GW21 that we previously sequenced (Pollen et al., 2014) (Figure S4A-C). Although these libraries did not contain ERCC Spike-In Controls, they expanded the developmental range of our analyses.

To determine the sensitivity of our single cell sequencing, we calculated the detection rate of each ERCC species (Figure S4E). With the exception of ERCC-00116, which is inefficiently sampled by polyA selection (Qing et al., 2013), at more than 8 copies, ERCCs were detected in 99-100% of cells. We then fit a linear regression model relating normalized read counts (ncounts) to ERCC molecules across all cells (Figure S4F). For our analyses, we included genes whose non-zero mean levels were above 20.6 ncounts, corresponding to 2 copies per cell. Furthermore, we omitted genes that were detected in fewer than 3 cells unless there was evidence of expression in our bulk RNA-seq data. Using these methods, we detected 10929 mRNAs and 1400 lncRNAs expressed across the 276 cells (Table S6).

Abundant lncRNA expression in subpopulations of single cells

In tissue specimens, lncRNAs were detected at 13.6 fold lower levels than mRNAs on average, consistent with previous reports (Cabili et al., 2011; Hangauer et al., 2013; Ramos et al., 2013) (Figure 1D). To determine whether lncRNAs may be expressed at high levels by subpopulations of cells, we analyzed the abundance of lncRNAs by comparing the median expression of lncRNAs to the median expression of mRNAs in each single cell (lncRNA:mRNA median ratios). In single cells, the median lncRNA:mRNA ratio was 0.85, with 32.2% of cells exceeding 1.0 (Figure 3B; Figure S5A,F). To investigate whether the lncRNAs analyzed in the single cells exhibit lower expression in whole tissues, we analyzed the same set of lncRNAs and mRNAs in whole neocortical samples. In bulk tissue samples, the lncRNA:mRNA ratios were significantly lower compared to single cell samples (median 0.31, $p=1.9x10^{-6}$, Mann-Whitney U; Figure 3B, Figure S5B). Furthermore, we merged the single cells *in silico* and found that this "reconstituted" sample had a lncRNA:mRNA ratio as low as the bulk tissues' (ratio = 0.14, Figure 3B; Figure

S5C). Analyses of cultured non-neural cells processed using the same single cell isolation, library preparation, and computational pipeline provided additional evidence that the higher lncRNA:mRNA ratios observed in single neocortex cells is not primarily driven by the methods used (Figure S6). Specifically, the more homogenous K562 cell line exhibited single cell lncRNA:mRNA ratios (median 0.46) that were much lower than those from neocortex cells (Figure S6A,D). Thus, lncRNAs that are detected at low levels in whole tissues can be abundantly expressed in individual cells.

To account for the observation that lncRNAs can be abundant in individual cells but detected at low levels in the whole tissues, we hypothesized that specific lncRNAs are expressed in subpopulations of cells. Housekeeping genes were detected in the vast majority of the single neocortex cells (median 91%), as expected (Figure 3C). In contrast, across all quantiles of gene expression, lncRNAs were detected in smaller proportions of cells than mRNAs ($p=2.7x10^{-288}$, Mann-Whitney U; Figure 3C,D). Consistent with this observation, lncRNAs that were lowly detected in bulk tissues were also expressed in fewer single cells (Figure S5D), indicating that in heterogeneous tissues lncRNAs are expressed in more discrete populations of cells than mRNAs.

Cell type-specific expression of lncRNAs

To determine whether the discrete and abundant expression of lncRNAs could relate to their expression in molecularly distinct cell types of the developing human brain, we performed hierarchical clustering of single cells with genes that exhibited significant variability (Methods, Figure 4A, Figure S4D,S7). We identified 7 clusters, each comprising cells derived from at least 3 different brains (Figure S7). To infer the identity of these cell clusters, we determined the most specific mRNAs in each cluster (Methods, Figure 4B, Table S7) and compared these genes to

known cell type-specific markers. With these methods, we identified the cell clusters to be endothelial cells (*FLT1*), radial glia (*VIM*, *GFAP*), dividing radial glia (*MKI67*, *TOP2A*), intermediate progenitors (*EOMES*), newborn neurons (*SEMA3C*, *DCC*), maturing excitatory neurons (*SATB2*, *ADCY1*), and inhibitory interneurons (*DLX2*, *GAD1*) (Hansen et al., 2013a; Kriegstein and Alvarez-Buylla, 2009; Molyneaux et al., 2007; Zhang et al., 2014).

To identify cell type-specific lncRNAs, we ranked the most specific lnRNAs of each cluster (Figure 4C). Overall, lncRNAs exhibited specificity scores comparable to those of mRNAs, with lower abundance lncRNAs having slightly greater specificity than abundance-matched mRNAs ($p=0.01$; Mann-Whitney U; Figure S8C). Notably, lncRNAs that were detected at lower abundances in bulk tissues were more cell type-specific in single cells than higher abundance lncRNAs ($p=1.1x10^{-47}$, Mann-Whitney U; Figure S8D). Of the top 105 specific lncRNAs (15 in each of 7 clusters), 10 were not annotated in Ensembl. *DLX6-AS1,* whose mean expression was 6123 fold higher in interneurons than in all other cell types, exhibited the highest cell type-specific enrichment of any gene (Figure 4C). Its mouse ortholog *Evf2* has been shown to function in interneurons (Berghoff et al., 2013; Bond et al., 2009). While *MEG3* and *SOX2-OT* have been shown as brain and even neuron specific (Zhang et al., 2014), our clustering revealed these lncRNAs to be more specific to interneurons than to newborn or maturing excitatory neurons (Figure 4C).

Gene co-expression analyses have previously been used to infer biological functions for novel lncRNAs (Guttman et al., 2009; Ramos et al., 2013). We therefore constructed co-expression networks between the top specific lncRNAs and all mRNAs expressed in the single cells (Figure S9A). Isolating the top 10% most correlated or anticorrelated mRNAs to these lncRNAs revealed gene clusters with cell type-specific function, such as “angiogenesis” for the endothelial

lncRNAs and "GABA synthesis, release, reuptake and degradation" for the interneuron lncRNAs (Figure S9B).

To validate our cell type-specific lncRNA expression patterns, we performed *in situ* hybridizations for three lncRNAs: *LOC646329* (radial glia), *LINC00599* (maturing neuron), and *DLX6-AS1* (interneuron) (Figure 5A). *LOC646329* was enriched in the ventricular zone (VZ), where most radial glia reside. *LINC00599* was enriched in the cortical plate (CP), which harbors maturing neurons. *DLX6-AS1* was enriched in the subpial granular layer and also exhibited a gradient of punctate expression spanning from the VZ to the intermediate zone (IZ), consistent with the migration patterns of cortical interneurons (Hansen et al., 2013b; Rakic and Zecevic, 2003). Imaging of the radial glial marker *PAX6*, the neuron marker *RTN1*, the maturing neuron marker ADRA2A, and the control marker *NNAT*, which is expressed broadly across progenitor and differentiated cells (Pollen et al., 2014), further validated the regional expression patterns of the cell type-specific lncRNAs (Figure 5B, S10).

To ask whether cell type-specific expression contributes to genes being detected at low levels in tissues, we analyzed the expression levels of the top 105 cell-type specific mRNAs and lncRNAs. As expected, in bulk tissues, cell type-specific mRNAs were detected at lower levels as compared to housekeeping genes (Figure 6B). Cell type-specific lncRNAs were detected at even lower abundances than housekeeping genes (0.069 fold). In contrast, in single cells, these lncRNAs were detected at levels close to those of housekeeping genes (0.436 fold, Figure 6A, Figure S8A,B). Thus, these cell type-specific lncRNAs appeared 6.31 times less abundant in bulk tissues as compared to their expression in single cells.

We then reasoned that the tissue level expression of a given gene could be modeled as a weighted average of cell type-specific expression, where the weights are proportional to the

relative abundance of each cell type. We used multiple linear regression to estimate the expected fraction of each cell type, using the mean expression in each cell type as predictor variables and bulk expression as the response. For consistency, only single cells derived from all cortical layers between GW19.5-23.5 were considered, and only bulk tissues from GW21-23 were used. Remarkably, the expected fractions of cell types showed strong agreement with the observed cell types identified in this study (Figure 6C). This was true for models using lncRNAs only ($r = 0.81$) and mRNAs only ($r = 0.87$), showing that the degree of decreased expression in bulk tissues can be explained largely by the relative abundances of cell types. Nonetheless, the discreteness of lncRNA expression (Figure 3D), and the relatively lower explanatory power of the lncRNA linear model ($R^2 = 0.336$; mRNA $R^2 = 0.422$) suggest that lncRNAs exhibit additional expression variation even within cell types.

Radial glia-enriched lncRNA *LOC646329* regulates cell proliferation

LOC646329 was among the most radial glia-enriched lncRNAs, and it was also detected at very low levels in bulk tissues (TPM < 0.5, Figure 4C,7A). Radial glia – the neural stem cell population of the developing brain – share biological and transcriptional characteristics with glioblastoma multiforme (GBM), a malignant glial tumor (Figure 7B) (Vescovi et al., 2006). *LOC646329* is expressed in human GBM, including the U87 GBM cell line (Figure 7A,B). To investigate the biological function of *LOC646329*, we performed CRISPRi knockdown of the lncRNA in U87 cells. U87 cells stably expressing dCas9-KRAB were infected with one of two lentiviruses harboring distinct sgRNAs targeting the transcription start site of *LOC646329* (Figure 7A). After confirming knockdown of *LOC646329* by qPCR, we performed internally controlled growth assays by measuring the percentage of cells infected with sgRNA-expressing

lentivirus over time (Figure 7C,D). Knockdown of *LOC646329* with either sgRNA reduced the propagation of U87 cells, indicating an important role for this lncRNA in cell proliferation.

Discussion

LncRNAs are remarkably tissue specific (Cabili et al., 2011; Ramos et al., 2013; Sauvageau et al., 2013), and the mammalian brain expresses a tremendous diversity and number of this class of noncoding transcripts (Cabili et al., 2011; Qureshi et al., 2010). Furthermore, some neural lncRNAs are primate and/or human specific, suggesting that lncRNAs play a role in the evolutionary expansion of the human neocortex (Lewitus and Huttner, 2015; Pollard et al., 2006). To lay groundwork for the study of lncRNAs in human brain development and disease, we generated a reference catalogue of lncRNAs expressed at different stages of human neocortex development.

Our human neocortex lncRNA reference catalogue greatly improves upon previous annotations of human neural lncRNAs in several ways (Lewitus and Huttner, 2015; Miller et al., 2014). First, we studied several developmental stages (8 samples spanning GW13 to GW23). Second, we analyzed both polyA selected and total RNA from each of the tissue specimens, which enabled the identification of novel lncRNAs that are potentially non-polyadenylated. Third, all of our cDNA libraries retained strand-of-origin information, allowing more accurate annotation and quantification of antisense transcripts, several of which have been shown to have important functions (Carrieri et al., 2012; Kotake et al., 2011; Pandey et al., 2008). Finally, the inclusion of biological replicate samples at matched (or near matched) developmental stages also allowed us to identify lncRNAs that were differentially expressed over time. We anticipate that this reference will facilitate future studies focusing on the roles of lncRNAs in cortical development and evolution, including cell type-specific roles of individual lncRNAs.

Our single cell transcriptome analyses indicate that lncRNAs can be highly expressed in individual cells of the developing neocortex. In cultured cells, studies using single molecule fluorescence *in situ* hybridization (FISH) demonstrate relatively uniform expression of 61 different lncRNAs across all cells, and that cells expressing high levels of specific lncRNAs are uncommon (Cabili et al., 2015). Consistent with these observations in cell lines, we found that lncRNA abundance is uniformly low to moderate in single cells of the relatively homogenous K562 leukemia line (Figure S6A,D). In contrast, single cells from developing human neocortex exhibited a wide range of lncRNA abundances, with some lncRNAs such as *SOX2-OT* and *DLX6-AS1* reaching levels higher than those of housekeeping genes (Figure 6A, Figure S8B). Given that the cultured K562 and human neocortex cells were all processed using the same Fluidigm C_1 platform and processed with the same computational pipeline, the discrete and abundant expression of lncRNAs in neocortical cells likely relates to differences in their cellular identity.

Consistent with this hypothesis of cellular identity explaining lncRNA expression, our results indicate preferential "dilution" of lncRNAs in bulk samples (Figure 6A), suggesting that cell type-specific expression of lncRNAs contributes to the apparent low expression of lncRNAs in heterogenous tissues. Furthermore, we were able to predict the relative abundances of the cell types identified in this study by regressing bulk expression onto cell type-specific expression (Figure 6C). However, two caveats are that we have not captured all the cell types in the developing human cortex, and we have not ruled out the possibility that the processing of single cells may enrich for certain cell types. A much larger scale study, such as those performed on the adult mouse brain (Zeisel et al., 2015) and retina (Macosko et al., 2015), would be required to better understand relative abundances of cell types in the developing human brain.

Quantitative measurements in single cell RNA-seq have been refined by the use of unique molecular identifiers (UMIs) that deconvolute non-uniform amplification of cDNA (Islam et al., 2014; Klein et al., 2015; Macosko et al., 2015; Zeisel et al., 2015). However, these methods are not yet compatible with full-length transcript coverage, which is advantageous for the study of lncRNAs. We retained full-length transcript coverage using the SMARTer protocol and instead used ERCC Spike-In Control RNA to control for technical noise. Our use of these synthetic RNA spikes to determine an expression threshold and to identify variable genes greatly diminished technical noise as the primary driver of lncRNA abundance. However, the use of spike-ins set a lower bound of detection, and many truly low abundance lncRNAs likely fell below this threshold. Nonetheless, the same threshold was also used for mRNAs, and our internal comparisons of lncRNAs to mRNAs in the same cell also controls for technical noise, especially since lncRNAs are not necessarily less stable than mRNAs (Clark et al., 2012). Therefore, our observation of discrete and abundant expression of lncRNAs in individual cells is not primarily driven by technical noise.

Several lncRNAs have previously been described as brain specific (Cabili et al., 2011; Lewitus and Huttner, 2015; Mercer et al., 2008). In this study, we found that these as well as novel lncRNAs can be further attributed to distinct cell types within the neocortex. For instance, *MEG3* is highly expressed in the brain (Cabili et al., 2011; Miyoshi et al., 2000), and we found this lncRNA to be especially enriched in interneurons, with moderate expression in maturing excitatory neurons (Figure 4C). We also identified cell type-specific and cortical layer-specific expression of lncRNAs such as *LINC00599 and LOC646329* (Figure 5), which have not been studied at this resolution in the human brain. In addition, our study of cell type-specific transcripts revealed a role for *LOC646329* in regulating cell proliferation. Determining whether

LOC646329 acts as a host to the microRNAs *MIR29A/B1* (Figure 7A) or through a distinct and separable role (Lauressergues et al., 2015) will require further investigation. Nonetheless, this result illustrates that lncRNAs that appear to be lowly expressed in tissues can have important functions, and motivates the study of lncRNAs at the single cell level.

Conclusions

Whole transcriptome analysis of single cells, in combination with deep RNA-seq of tissues, allowed us to identify abundant cell type-specific lncRNAs in the developing neocortex. Importantly, many lncRNAs that appear lowly expressed at the population level are abundant in discrete cell types. Analysis of these lncRNAs in the brain both reveals useful markers of cell types during lineage progression from precursor to differentiated cell types, and enables the discovery of novel cellular functions of lncRNAs. As such, lncRNAs that are lowly expressed in a population may still regulate essential functions and should not be discounted solely based on apparent abundance. These data and workflow should facilitate future studies aimed at determining the function of lncRNAs in the brain during development and disease.

Experimental Procedures

Prenatal tissue collection

De-identified human prenatal brain tissue samples were collected from elective pregnancy termination specimens at San Francisco General Hospital, usually within 2 h of the procedure. Donated specimens were examined only from patients who had previously given informed consent and in strict observance of state and institutional legal and ethical requirements. Gestational age was determined by measuring foot length, and tissues were transported to the laboratory on ice in Leibowitz-15 medium for immediate tissue processing and RNA extraction.

Ethics

Research protocols were approved by the Human Gamete, Embryo, and Stem Cell Research (GESCR) Committee (Institutional Review Board) at University of California, San Francisco. All experimental methods comply with the Declaration of Helsinki.

Bulk tissue RNA-seq library preparation

Fresh neocortex tissues were dissected along the radial axis, and 50-100 mg sections spanning the ventricular zone to the marginal zone of the neocortex were harvested in TRIzol. Mixtures were homogenized by triturating 10 times through a 21G needle. Ethanol extracted RNA was loaded onto RNeasy columns (QIAGEN), and on-column DNase treatment was performed as previously described(Ramos et al., 2013). Purified RNA samples produced RIN scores between 8.6 and 9.5, measured by 2100 Bioanalyzer (Agilent). RNA-seq libraries were generated using both TruSeq Stranded mRNA and TruSeq Stranded Total RNA with Ribo-Zero Gold kits (Illumina) according to the manufacturer's protocols. cDNA validation and normalization were

performed using RT-PCR and Quant-iT PicoGreen (Invitrogen). Cluster generation and high-throughput sequencing were performed on a HiSeq 2500 (Illumina), using the paired-end 100 bp protocol.

Single cell capture and RNA-seq library preparation

Tissues were microdissected by embedding in 3.5% low melt agarose (Fisher) and sectioned along the radial axis, perpendicular to the ventricles, using a Leica VT1200S vibratome in artificial cerebrospinal fluid (ACSF) media containing 125 mM NaCl, 2.5 mM KCl, 1 mM $MgCl_2$, 1 mM $CaCl_2$, 1.25 mM NaH_2PO_4. Cortex dissections were added to papain (Worthington Biochem. Corp) and 2000 units/mL of DNase I freshly diluted in EBSS and incubated at 37° C for 30 minutes and centrifuged for 5 minutes at 300g. Supernatants were removed and dissociated cells were resuspended in 0.5 mL of sterile DPBS containing 3% FBS (Sigma) and 1000 units of DNAse I. Suspensions were further dissociated by pipetting up and down 10 times and then passing through a 40 μm strainer cap (BD Falcon) to yield single cell suspensions. Single cell dissociations were performed on samples separate from those used for bulk tissue RNA-seq.

Single cell capture was performed using the Fluidigm C1 Single-Cell Auto Prep Integrated Fluidic Circuit (IFC) and SMARTer Ultra Low RNA Kit as previously described(Pollen et al., 2014). We minimized the capture of cell doublets by only using small (5-10 μm) 96-plex IFCs, which in combination with neural cells were demonstrated to have the lowest instance of doublets (median 6%, SD 4%) (Corporation, 2016). ERCC Spike-In Controls (Ambion) were added to each single cell lysis reaction at a final dilution of 1:20,000, in order to capture a wide range of molecule count per reaction. cDNA was quantified using High

Sensitivity DNA Kits (Agilent) and diluted to 0.15–0.30 ng/μL in C_1 Harvest Reagent. Dual indexing and amplification were performed using the Nextera XT DNA Sample Preparation Kit (illumina) with the following modifications: reactions were performed at one quarter of recommended volume, tagmentation proceeded for 10 minutes, and the PCR extension time was 60 seconds. Amplified cDNA was size selected twice using 0.9X volume of Agencourt AMPure XP beads (Beckman Coulter). Final cDNA libraries were quantified using High Sensitivity DNA Kits (Agilent) and sequenced on a HiSeq 2500 (Illumina), using the paired-end 100 bp protocol.

LncRNA identification and quantification pipeline – bulk whole tissues

alignment and *de novo* transcript assembly. Quality control of RNA-seq reads was performed using FastQC. No read trimming was performed on reads from bulk tissues, since all samples exhibited 25%ile quality scores above Q30 at all 100bp positions. Strand-specific reads were aligned to the human reference genome, Ensembl GRCh37/hg19 release 75, using TopHat v2.0.10 with the flags (--library-type fr-firststrand --microexon-search). Each sample generated between 105 and 148 million mapped read pairs from rRNA depleted total RNA-seq, and between 60 and 105 million mapped read pairs from polyA selection RNA-seq (Table S1). *De novo* transcriptome assembly was performed separately on rRNA depletion total RNA-seq alignments, and on polyA selection RNA-seq alignments, using Cufflinks v2.2.1 with the flags (-M ensembl_75_mtRNA_rRNA.gtf -b genome.fa -u --library-type fr-firststrand --max-multiread-fraction 0.25 --3-overhang-tolerance 2000) to mask potential rRNA and mtRNA reads, enable bias correction and multi-map correction, and also to reduce the identification of polymerase run-on fragments as novel transcripts(Trapnell et al., 2010). Transcriptome assemblies at all developmental stages and replicates were merged, separately for rRNA depletion total RNA-seq

and polyA selection RNA-seq, with the Ensembl 75/GENCODE 19 reference transcriptome, using Cuffmerge. To identify transcripts novel as compared with Ensembl, we utilized Cuffcompare class codes and extracted those assembled transcripts classified as: i – novel intronic, u – novel intergenic, x – novel antisense. All novel transcripts under 200 nt in length were removed. Of the remaining transcripts, we determined minimal read coverage thresholds based on whether Cufflinks classified previously annotated transcripts as having "full_read_support." By analyzing the true positive rate vs. false positive rate of classifying known genes as obtaining "full_read_support" at various coverage thresholds, we determined the minimum coverage to be 1.4 for polyA and 1.67 for total RNA-seq (at FDR = 0.05). Starting with just the polyA RNA-seq data, transcripts with read coverage above 1.4 in both biological replicates of at least one developmental stage were included in the reference and considered to be expressed in the neocortex. Due to limited availability of early fetal tissue, the GW14.5 sample was treated as the biological duplicate of the GW13 sample. Novel transcripts that were predicted to have protein coding capability by one or more of the following methods were classified as transcripts of uncertain coding potential (TUCP): CPAT (Wang et al., 2013), threshold = 0.364; CPC (Kong et al., 2007), threshold = 0; Pfam (Finn et al., 2014). For comparing to the Pfam database, the longest potential open reading frame (ORF) of each novel transcript was obtained, and any putative ORF that had a significant match for a protein domain annotated in Pfam A or Pfam B resulted in the parent transcript being classified as a TUCP. All remaining novel lncRNAs and TUCPs were then named according to recently proposed nomenclature standards (Mattick and Rinn, 2015), for instance LINC-[nearest mRNA] for intergenic lncRNAs and [nearest mRNA]-AS for antisense lncRNAs, and were then merged to the Ensembl 75 reference transcriptome, resulting in the polyA Full reference transcriptome. The

polyA Stringent reference transcriptome was produced by removing all novel single-exon lncRNAs and TUCPs. Known lncRNAs from Ensembl were obtained by identifying transcripts with one of the following biotype classifications: "3prime_overlapping_ncrna," "antisense," "lincRNA," "processed_transcript," "sense_intronic," and "sense_overlapping." The same pipeline, with the coverage threshold of 1.67, was performed for reads derived from the total RNA-seq.

LncRNA quantification. Gene-level fragment counts for each polyA and total RNA sample were quantified using featureCounts v1.4.6 (Liao et al., 2014), using the flags: -p -s 2 -B -C -t exon -g gene_id. Count tables were normalized to TPM (Transcripts per Million) (Li and Dewey, 2011) for internal comparisons and visualizations of bulk RNA-seq. To identify differentially expressed genes, we used DESeq2 (Anders and Huber, 2010) on gene-level fragment counts derived from the polyA samples and polyA Full reference transcriptome. Pairwise negative binomial significance tests were performed between developmental stages using biological duplicates, and the union of genes that were significant at FDR < 0.01 were classified as differentially expressed. Gene ontology (GO) analysis was performed using the DAVID web server(Huang et al., 2009). To identify lncRNAs enriched in total RNA versus polyA RNA, we first identified all transcript annotations in the total Full reference that did not overlap with transcripts in the polyA Full reference. We then merged these transcripts with the polyA Full reference. We quantified gene-level fragment counts annotated in this augmented reference using all samples (both polyA RNA and total RNA), as described above. For each of the 8 tissue samples, we then compared the TPM for each gene as observed from the polyA and total RNA-

seq. mRNAs and lncRNAs that were consistently > 10 fold enriched in one or the other fraction across all 8 samples were then considered enriched in total RNA or enriched in polyA RNA.

Quality control and analysis of single cell RNA-seq

Paired end 100 reads from single cell cDNA libraries were quality trimmed using Trim Galore with the flags: -q 20 --nextera --length 20. Trimmed reads were aligned to the human reference genome, Ensembl GRCh37/hg19 release 75, augmented with the 92 ERCC Spike-In Control sequences, using TopHat v2.0.10 with the flags: --transcriptome-index=polya_stringent_reference.gtf --prefilter-multihits. The polyA Stringent reference transcriptome, derived from bulk tissue RNA-seq as described above, was used as a transcriptome reference. A median of 1 million 100 bp paired-end reads were successfully aligned per cell. Gene-level fragment counts were quantified using featureCounts v1.4.6 with the flags: -p -B -C -t exon -g gene_id. Since the SMARTer cDNA prep does not retain strand-of-origin information, we did not count reads that overlapped multiple features, even if they were annotated on opposite strands. Outlier identification was performed by calculating the number of genes and ERCC species detected in each cell (defined as the number of genes with > 1 count). The distribution of genes and ERCCs detected decayed rapidly below 1000 genes and 40 spikes, with the remaining few samples centered at 0 genes and 0 spikes (Figure S4A-B). Therefore, we excluded from analyses cells in which we detected fewer than 1000 genes and 40 ERCC spikes. The > 40 ERCC requirement was not used for previously sequenced single cells from Pollen et. al. 2014, as these cells from GW16 and GW21 brains did not contain ERCC Spike-In Controls.

To determine the most accurate expression metric to use for our single cell transcriptome analyses, we correlated absolute ERCC spike abundances in each cell with their expression

readout using four different metrics (Figure S4J). Count-based metrics, which were counts and CPM (Counts per Million mapped reads), outperformed length-normalized metrics, which were FPKM (Fragments per Kilobase per Million mapped reads) and TPM (Transcripts per Million). Therefore, we used counts and normalized each single cell library by transcriptome size factors according to DESeq (Anders and Huber, 2010). Separate size factors were calculated and used for genes and ERCCs.

Detection rate analysis was performed by calculating for each gene and ERCC spike the number of cells out of the 276 neocortex cells that exhibited > 1 normalized counts for that gene. We determined that the sensitivity of detection was 74% for a ERCC species present at 2 copies per cell. At more than 8 copies per cell, ERCCs were detected in 99-100% of cells (Figure S4E). The exception was ERCC-00116, which is known to be inefficiently sampled by polyA selection (Qing et al., 2013). To find a reliable count threshold for further analyses, we then fit a linear regression model relating normalized read counts (ncounts) to ERCC molecules across all cells (Figure S4F). We then included genes whose non-zero mean levels across all cells were above 20.6 ncounts, corresponding to 2 copies per cell. Furthermore, we omitted genes that were detected in fewer than 3 single cells unless they were considered to be expressed by the whole-tissue RNA-seq experiments. For single cell comparisons of gene expression levels of lncRNAs to mRNAs, only genes that were expressed above 2 normalized counts in each cell were considered. The median normalized counts for lncRNAs was then compared to the median normalized counts for mRNAs in single cells. *In silico* merged cells in Figure 3B and S5C were generated by taking the sum of gene read counts across all single cells that passed QC.

Clustering and cell type identification of single cells

To identify genes for unsupervised clustering, we modeled technical noise using ERCC Spike-In Control RNA and identified genes with significantly greater expression variation than expected from noise according to Brennecke et. al. 2013 (Brennecke et al., 2013). Briefly, we fit a gamma family generalized linear model to the coefficient of variation squared of ERCCs as a function of mean normalized counts. We then determine the expected variance model for genes that exhibit greater than 50% biological coefficient of variation at FDR < 0.01 (Benjamini-Hochberg adjusted p value from Chi-square distribution). These 5243 remaining genes were then used for principal component analysis (PCA). PCA was performed using $\log_2$ size factor-normalized counts with a pseudocount of 1 and visualized in R. We then ranked all genes (mRNAs and lncRNAs) based on their highest absolute values of gene loading scores across the first four principal components in PCA. We then performed complete linkage hierarchical clustering of single cell expression ($\log_2$(Normalized Counts + 1)) using 1-(Pearson correlation coefficients) as the distance metric, using the top 500 genes ranked by PCA loading scores (Figure S7). Cluster dendrograms were cut statically between r=0.8 and r=0.9, and cell types were inferred by comparing mRNAs specific to each cluster to known markers.

Cell type specificity analysis

Cell type-specificity for each gene was calculated as the odds ratio of a cell expressing a given gene, above a given threshold, within a cluster compared to outside a cluster. Specifically, if p_{ij} is the fraction of cells expressing gene i above threshold in cluster j, and q_{ij} is the fraction of cells expressing gene i above threshold that are not in cluster j, then we measure cell type

specific expression odds by: $\theta_{ij} = log\frac{p_{ij}(1-q_{ij})}{q_{ij}(1-p_{ij})}$. To generate the set of mRNAs and lncRNAs specific to each cell type, we first assigned each gene to a cell type by determining which cell type yielded the maximum log odds specificity score for that gene. We used each gene's own 75%ile for expression thresholds. Then we calculated expression enrichment scores for each gene in each cell type by: $e_{ij} = u_{ij}/v_{ij}$, where u_{ij} is the mean expression level of gene i within cluster j, and v_{ij} is the mean expression level of gene i outside of cluster j. Genes were ranked within their cell type cluster assignments using expression enrichment scores, and the top 15 mRNAs and lncRNAs in each cell type were obtained.

Linear regression model

We reasoned that the relationship between bulk tissue RNA-seq expression and single cell RNA-seq expression could be approximated using multiple linear regression in the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

where *Y* is bulk expression, *n* is the number of cell types identified in the study, β_0 is a constant that normalizes global differences between single cell and bulk RNA-seq, $[\beta_1 \ldots \beta_n]$ is the vector of slope coefficients that are proportional to the relative fraction of each cell type within the bulk tissue, and *X* is the vector of mean expression of each cell type as measured by single cell RNA-seq. Expression values were represented as log2 transformed, size factor-normalized counts, with a pseudocount of 1. To improve matching between bulk tissues and single cells, we used the mean expression of GW21 and GW23 bulk cortex samples (n=4) as the bulk expression level, and we used single cells derived from all cortical layers between GW19.5-23.5 (n=226) and calculated the mean expression for each cell type, as was determined by hierarchical clustering in Figure 4. Only genes that passed the single cell expression threshold (described above), and were

detected above 5 counts in bulk tissues, were included. The data were then fit using the *lm* function in R, and the resulting slope coefficients were normalized such that $\sum_{i=1}^{n} \beta_i = 1$. These normalized coefficients, which represent the expected fractions of cell types, were then compared to the observed relative fractions of cell types, according to the cluster sizes in Figure 4 (but only counting the 226 included cells).

Gene co-expression analysis

A correlation matrix was assembled by calculating all pairwise Pearson correlation coefficients between the 105 cell type-specific lncRNAs and all expressed mRNAs across single cells. Expression values were represented as log2 transformed, size factor-normalized counts, with a pseudocount of 1. mRNAs whose maximum, absolute value correlation coefficients were in the top 10%ile were kept. The resulting sub-matrix was clustered using Euclidean distance and complete linkage, and gene clusters were analyzed for gene ontology terms using Enrichr(Chen et al., 2013). Clusters with significant gene ontology terms are represented in Figure S9.

***In situ* hybridization**

Probes for *in situ* hybridization were synthesized (Genscript) or cloned from GW16 human fetal neocortex cDNA, which were reverse-transcribed using SuperscriptII (Invitrogen) with random hexamer primers. Probe sequences were subcloned into pGEM T-Easy vector (Promega). T7 or SP6 RNA polymerase (Roche) was then used for *in vitro* transcription of the probes, in the presence of DIG-RNA Labeling Mix (Roche). *In situ* hybridization was performed blinded to the sense/antisense status for each probe and sense control probes gave no signal (data not shown).

The *in situ* hybridization protocol has been described before(Wallace and Raff, 1999). Images were collected with a Leica DMI 4000B microscope using a Leica DFC295 camera.

Immunohistochemistry

Immunohistochemistry was performed as described in (Pollen et al., 2014). Briefly, tissue samples were fixed in 4% paraformaldehyde, cryoprotected in 30% sucrose, and embedded in a 1:1 mixture of 30% sucrose and optimal cutting temperature (Thermo Scientific). 20 μm cryosections were collected using a Leica CM3050S cryostat. Primary antibody: ADRA2A (1:100, Thermo Scientific, PA1-048). Heat-induced antigen retrieval was performed in 10 mM sodium citrate buffer, pH 6. Binding was revealed using Alexa Fluor™ 488 fluorophore-conjugated secondary antibody (Life Technologies). Images were collected with a Leica TCS SP5 X Confocal microscope.

CRISPRi knockdown of lncRNAs

CRISPRi mediated repression of lncRNA transcription was performed as described previously (Gilbert et al., 2014)**.** First, we created a stable polyclonal cell line expressing dCas9-KRAB by transducing U87 glioblastoma cells with dCas9-KRAB-BFP lentivirus and sorting for the top 30% of BFP expressing cells. sgRNAs were designed as previously described (Gilbert et al., 2014). sgRNA protospacer sequences were: sgLOC646329-1, GCTTAGGAAATCACCAGCTCC; sgLOC646329-2, GGTCTGCCGTGACAGTTCAGT; sgCtrl, GAACGACTAGTTAGGCGTGTA. We then cloned the sgRNA sequences into puromycin-resistant lentiviral vectors and infected dCas9-KRAB U87 cells with the resulting lentivirus particles. To assess lncRNA knockdown, we treated cells with 1 μg/mL puromycin for

4 days following sgRNA infection and performed RT-qPCR as previously described (Gilbert et al., 2014). RT-qPCR primers were: LOC646329 Forward, CTTGGGGATCCTCTGTACGC; LOC646329 Reverse, CTTCGGTATCCTGATGTAGGTGT.

Internally controlled, relative growth assays were performed separately. Triplicates cultures of dCas9-KRAB U87 cells were transduced at 40-50% infection rate with lentiviruses harboring sgRNAs. The proportion of cells that were BFP^{hi} ,indicating sgRNA expression, in each population was measured at every passage (every other day) by an LSR II flow cytometer (BD). These proportions were internally normalized to values at 5d, when all infected cells reached full sgRNA expression, and then compared to cells infected with non-targeting control sgRNAs, which demonstrated stable expression of the sgRNA containing vector.

Availability of Supporting Data

All sequencing data, bulk RNA-seq alignment signal, reference transcriptome GTF files, and expression tables are deposited in GSE71315. Glioblastoma cell line U87 RNA-seq data were obtained from GSE29738. Single cell RNA-seq libraries from the 50 GW16 and GW21 samples and the 46 K562 cells were obtained from SRP041736.

Acknowledgements

We thank Joe Shuga for assistance with single cell RNA-seq, Richard Lao for sequencing assistance, and Shaohui Wang, Yingying Wang, and the staff at the San Francisco General Hospital for providing access to donated tissue samples. This project was supported by VA 5I01 BX000252-06, NIH 1R01NS091544-01A1, NIH SPORE DRP, the Shurl and Kay Curci Foundation, and the Hana Jabsheh Initiative (to D.A.L.), UCSF-CTSI UL1 TR000004 (to A.A.D.), and NIH U01 MH105989 to A.R.K. S.J.L. is supported by NIH F30 NS092319-01. A.A.P. is supported by a Damon Runyon Cancer Research Foundation postdoctoral fellowship (DRG-2166-13).

Figures

Figure 1

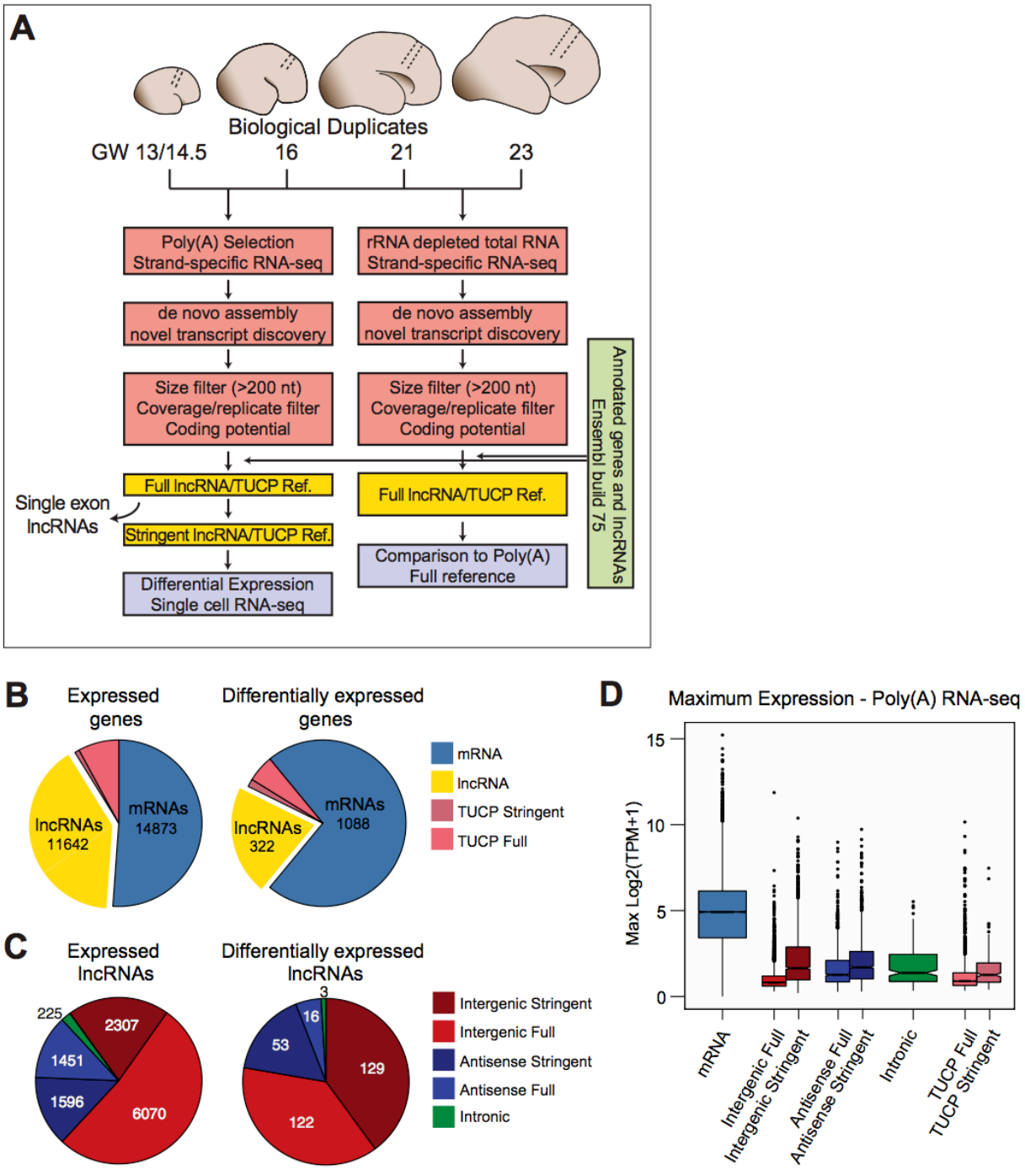


Figure 1. Catalogue of lncRNAs in human neocortex development

A) Schematic of neocortex tissue dissection, poly(A) and total RNA-seq library prep, and computational pipeline for lncRNA annotation and quantification. B) Numbers of expressed (left) and differentially expressed (right; DESEq2, FDR < 0.01) mRNAs, lncRNAs, and TUCPs during neocortex development in bulk tissues. Stringent references omit novel single exon transcripts. C) Breakdown of expressed (left) and differentially expressed (right) lncRNAs based on genomic orientation relative to mRNAs. D) Maximum expression levels of transcripts described in the Full and Stringent references derived from Poly(A) selection RNA-seq, across all samples. TPM, Transcripts per Million.

Figure 2

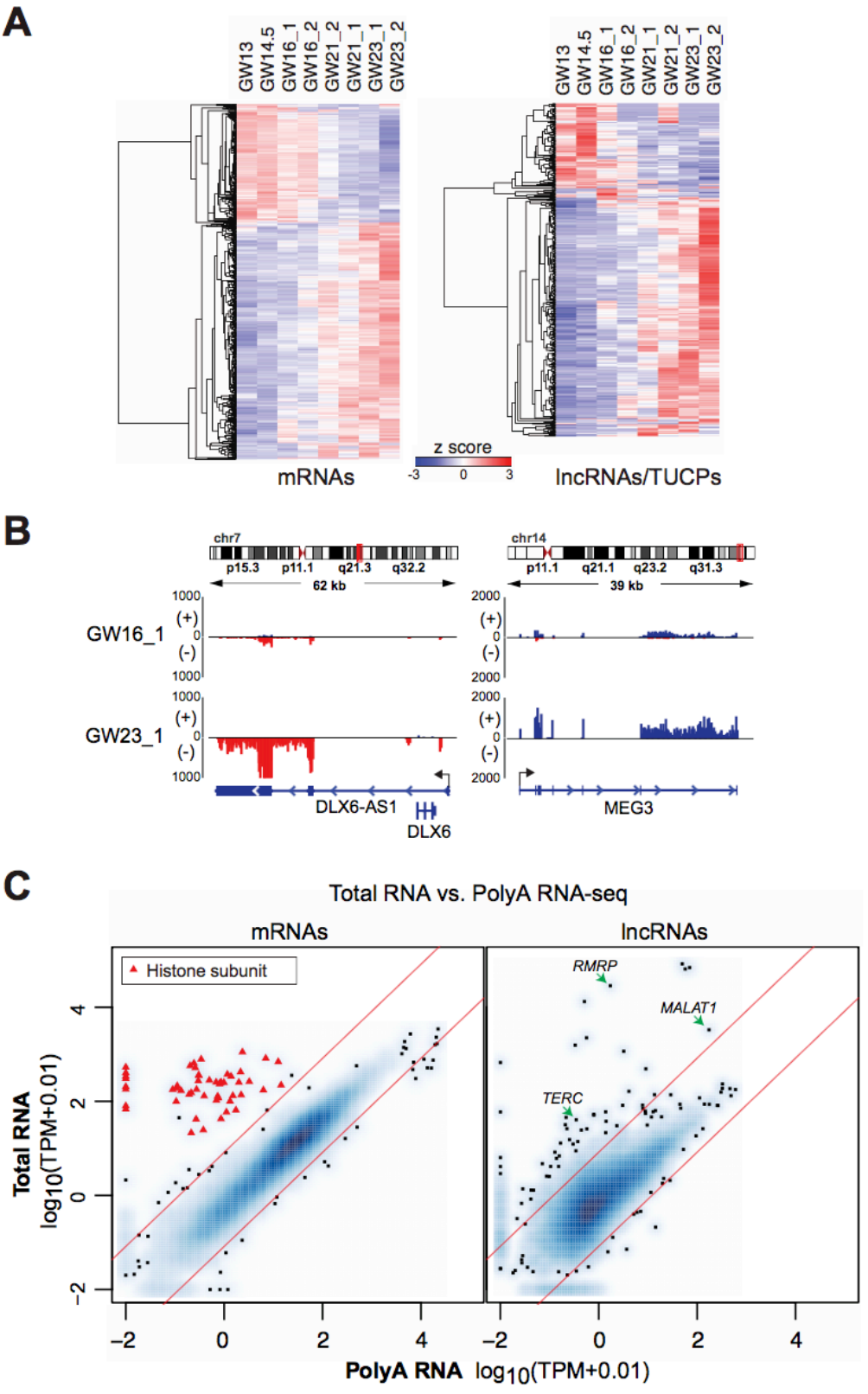

A
GW13
GW14.5
GW16_1
GW16_2
GW21_2
GW21_1
GW23_1
GW23_2
z score
-3
0
3
mRNAs
lncRNAs/TUCPs
B
chr7
p15.3
p11.1
q21.3
q32.2
62 kb
chr14
p11.1
q21.1
q23.2
q31.3
39 kb
GW16_1
GW23_1
(+)
(-)
DLX6-AS1
DLX6
MEG3
C
Total RNA vs. PolyA RNA-seq
mRNAs
lncRNAs
Histone subunit
RMRP
MALAT1
TERC
Total RNA
log10(TPM+0.01)
PolyA RNA log10(TPM+0.01)

Figure 2. Differential expression of mRNAs and lncRNAs/TUCPs during neocortex development

A) Heatmaps of differentially expressed mRNAs (left) and lncRNAs/TUCPs (right) throughout 8 samples of bulk neocortex tissues. B) Strand specific RNA-seq alignments at the *DLX6-AS1* and *MEG3* loci in GW16 and GW23 replicate 1 samples. Scale, number of aligned reads. C) Comparison of mRNA (left) and lncRNA (right) expression levels between poly(A) RNA-seq and total RNA-seq in GW16 sample 2. Red diagonals represent 10 fold enrichment in either total (upper) or polyA (lower) fractions. Red triangles, histone subunits enriched > 10 fold in total RNA. TPM, Transcripts per Million.

Figure 3

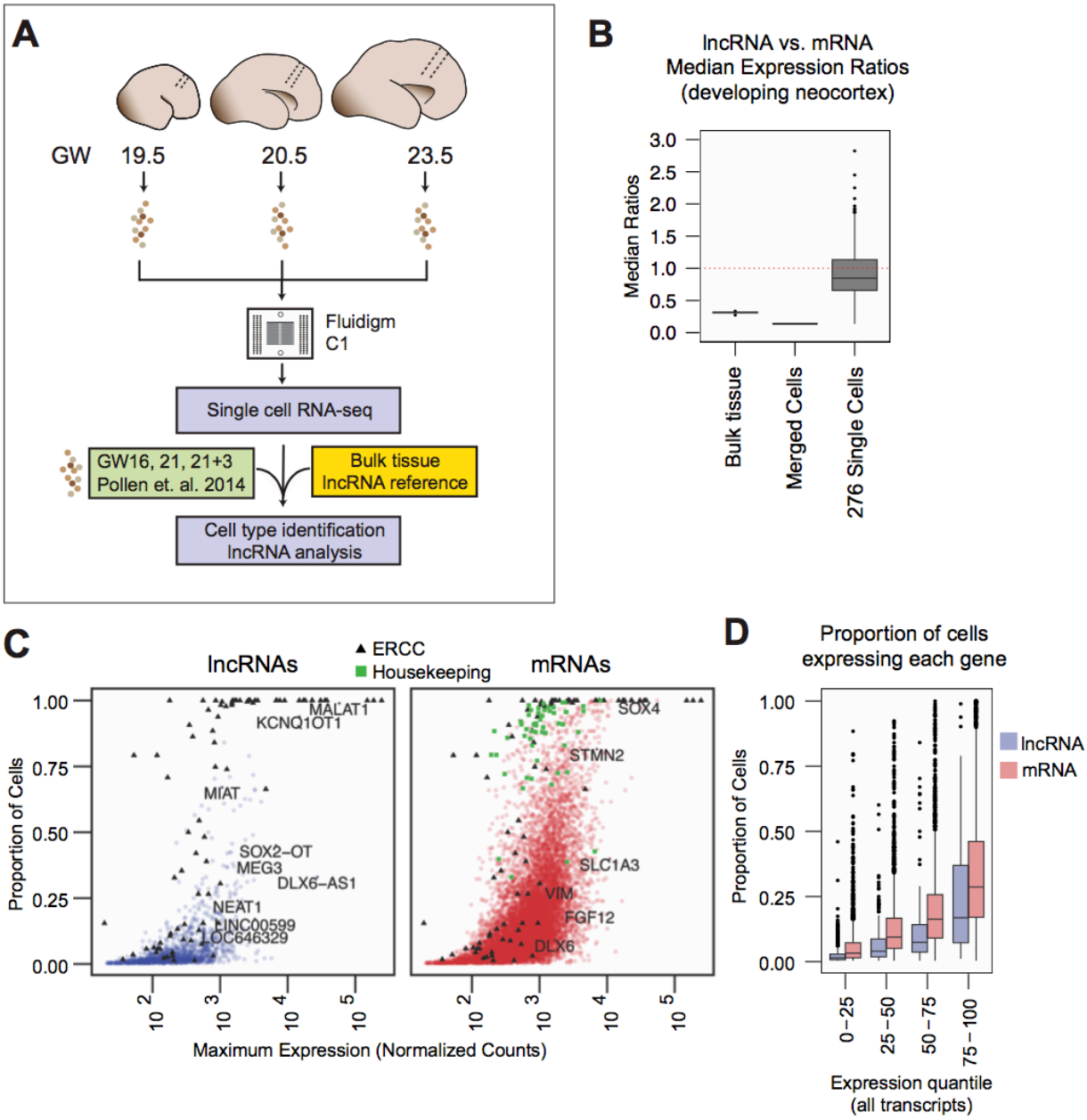

A
GW 19.5 20.5 23.5
Fluidigm C1
Single cell RNA-seq
GW16, 21, 21+3 Pollen et. al. 2014
Bulk tissue lncRNA reference
Cell type identification lncRNA analysis
B
lncRNA vs. mRNA Median Expression Ratios (developing neocortex)
Median Ratios
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Bulk tissue
Merged Cells
276 Single Cells
C
lncRNAs
ERCC
Housekeeping
mRNAs
Proportion of Cells
1.00
0.75
0.50
0.25
0.00
MALAT1
KCNQ1OT1
MIAT
SOX2-OT
MEG3
DLX6-AS1
NEAT1
LINC00599
LOC646329
SOX4
STMN2
SLC1A3
VIM
FGF12
DLX6
Maximum Expression (Normalized Counts)
D
Proportion of cells expressing each gene
lncRNA
mRNA
0-25
25-50
50-75
75-100
Expression quantile (all transcripts)

Figure 3. Single cell transcriptomics of lncRNA expression

A) Schematic of single cell microfluidic capture and integration of transcriptome reference generated from bulk tissue RNA-seq to conduct cell type identification and lncRNA analysis. Previously captured cells from Pollen et. al. 2014 were also included. B) Distributions of median lncRNA expression to median mRNA expression ratios (lncRNA:mRNA) in bulk tissues, *in silico* merged single cells, and single cells from the developing neocortex. C) Proportion of neocortex cells that expressed each lncRNA (blue) and mRNA (red), separated by maximum expression in single cells. D) Same as in C) but grouped by maximum expression quantile of the set of all transcripts (lncRNA and mRNA combined). Green squares, housekeeping genes. Black triangles, ERCC Spike-In Controls.

Figure 4

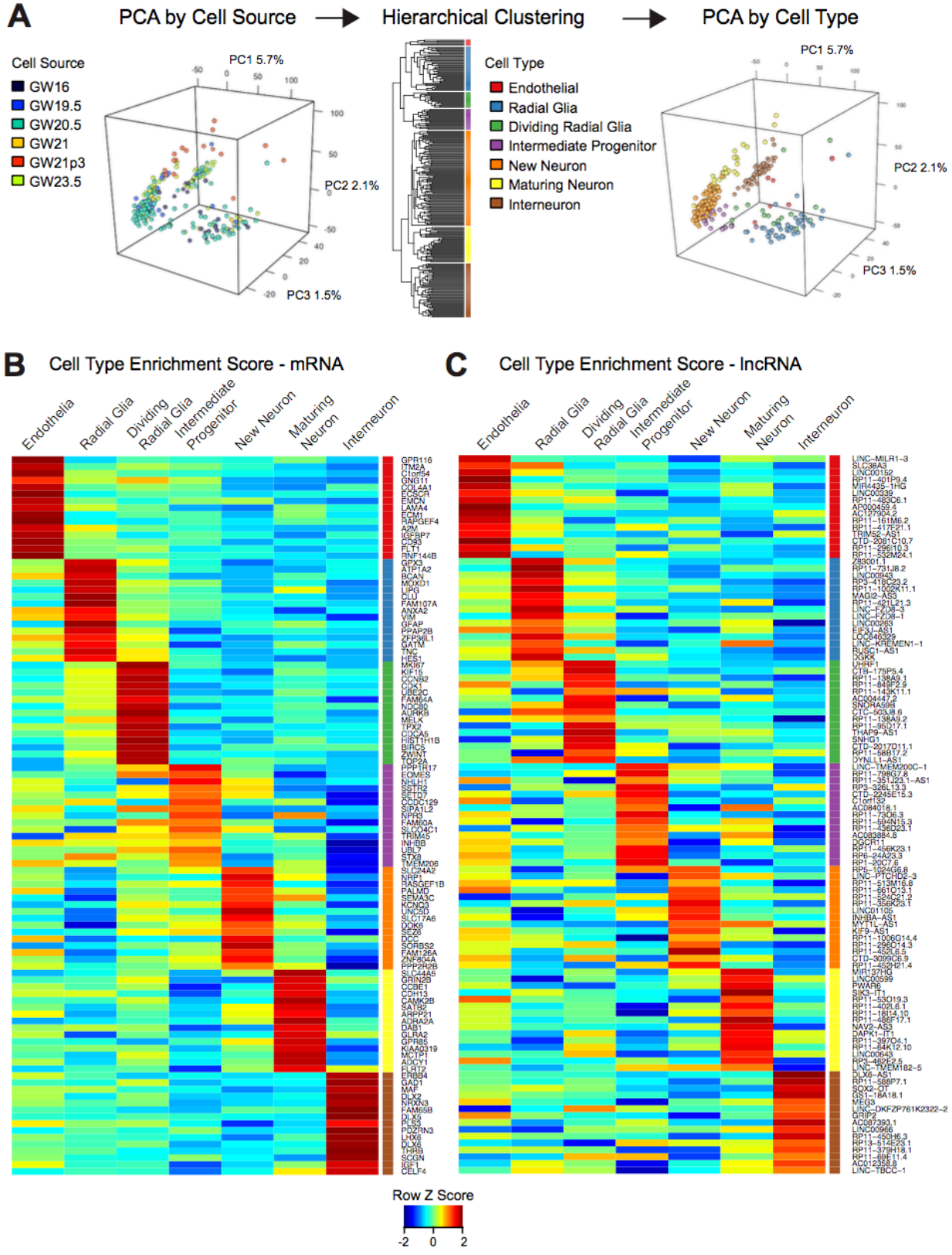

A
PCA by Cell Source
Hierarchical Clustering
PCA by Cell Type
Cell Source
GW16
GW19.5
GW20.5
GW21
GW21p3
GW23.5
PC1 5.7%
PC2 2.1%
PC3 1.5%
Cell Type
Endothelial
Radial Glia
Dividing Radial Glia
Intermediate Progenitor
New Neuron
Maturing Neuron
Interneuron
B
Cell Type Enrichment Score - mRNA
C
Cell Type Enrichment Score - lncRNA
Endothelia
Radial Glia
Dividing Radial Glia
Intermediate Progenitor
New Neuron
Maturing Neuron
Interneuron
Row Z Score
-2
0
2

Figure 4. Cell type-specific expression of lncRNAs

A) Identifying cell types using unsupervised clustering. Left - Principal component analysis (PCA) of single cells colored by developmental stage of source tissues. Middle - Complete linkage hierarchical clustering of single cells using genes exhibiting variance greater than expected than from technical noise. Right – PCA of single cells colored by cell types inferred from protein coding genes specific to each cluster. Axes labels indicate percent variation explained by each PC. B) Heatmaps of cell type enrichment scores for the 15 most specific mRNAs and (C) lncRNAs in each cluster. GW21p3, primary cells derived from GW21 brain that were cultured in differentiation media for 3 days.

Figure 5

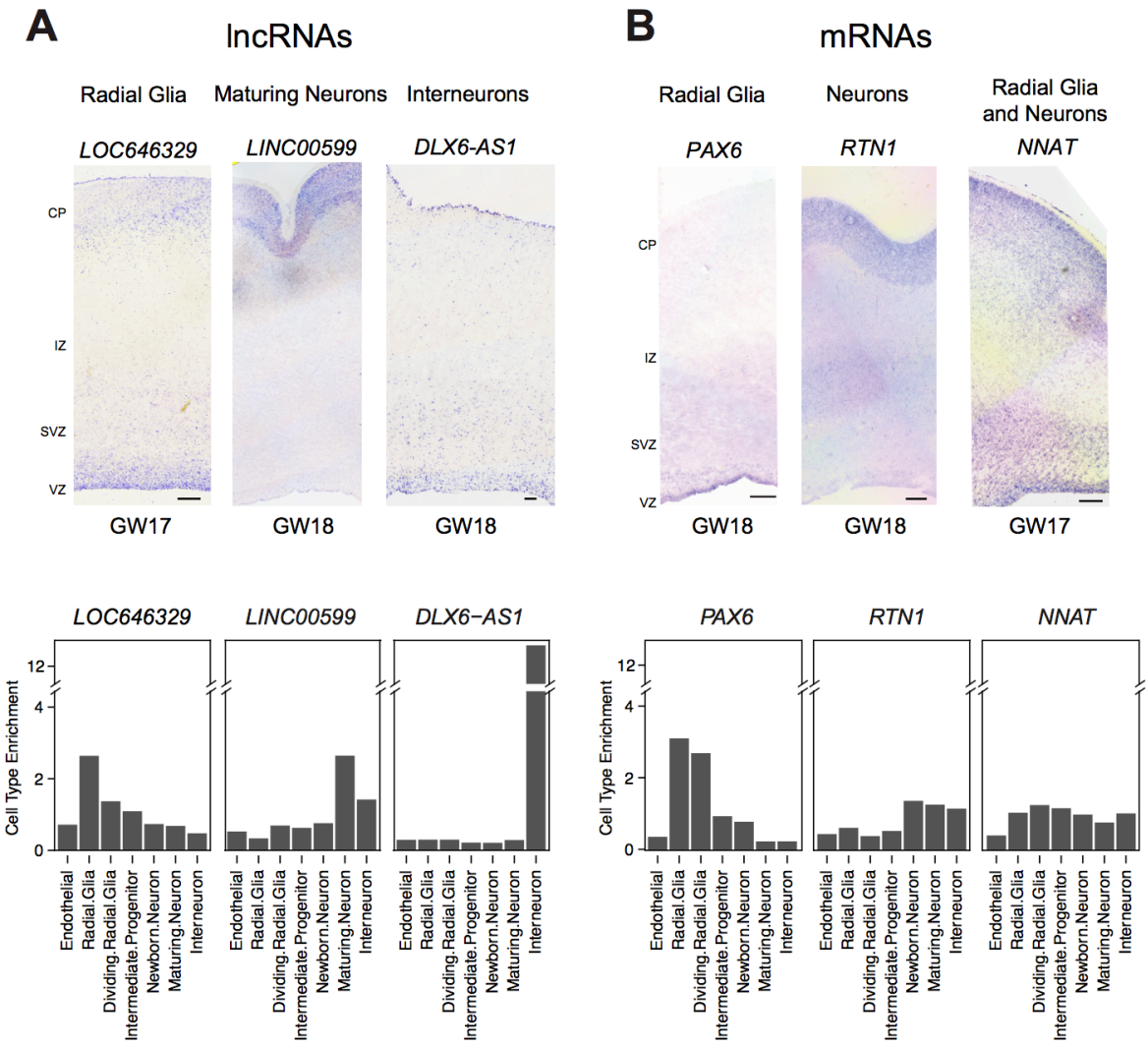

A
lncRNAs
Radial Glia
Maturing Neurons
Interneurons
LOC646329
LINC00599
DLX6-AS1
CP
IZ
SVZ
VZ
GW17
GW18
GW18
B
mRNAs
Radial Glia
Neurons
Radial Glia and Neurons
PAX6
RTN1
NNAT
CP
IZ
SVZ
VZ
GW18
GW18
GW17
LOC646329
LINC00599
DLX6−AS1
PAX6
RTN1
NNAT
Cell Type Enrichment
12
4
2
0
Endothelial
Radial.Glia
Dividing.Radial.Glia
Intermediate.Progenitor
Newborn.Neuron
Maturing.Neuron
Interneuron

Figure 5. *in situ* hybridization of cell type-specific lncRNAs and mRNAs in developing neocortex

A) *In situ* hybridizations and corresponding cell type enrichment values for radial glia-specific lncRNA *LOC646329* (left), maturing neuron-specific lncRNA *LINC00599* (middle), and interneuron-specific lncRNA *DLX6-AS1* (right). B) *In situ* hybridizations and corresponding cell type enrichment values for radial glia-specific mRNA *PAX6* (left), neuron mRNA marker *RTN1* (middle), and progenitor and differentiated cell-expressed mRNA *NNAT* (right). Scale bars, 250 μm. CP, cortical plate. IZ, intermediate zone. SVZ, subventricular zone. VZ, ventricular zone.

Figure 6

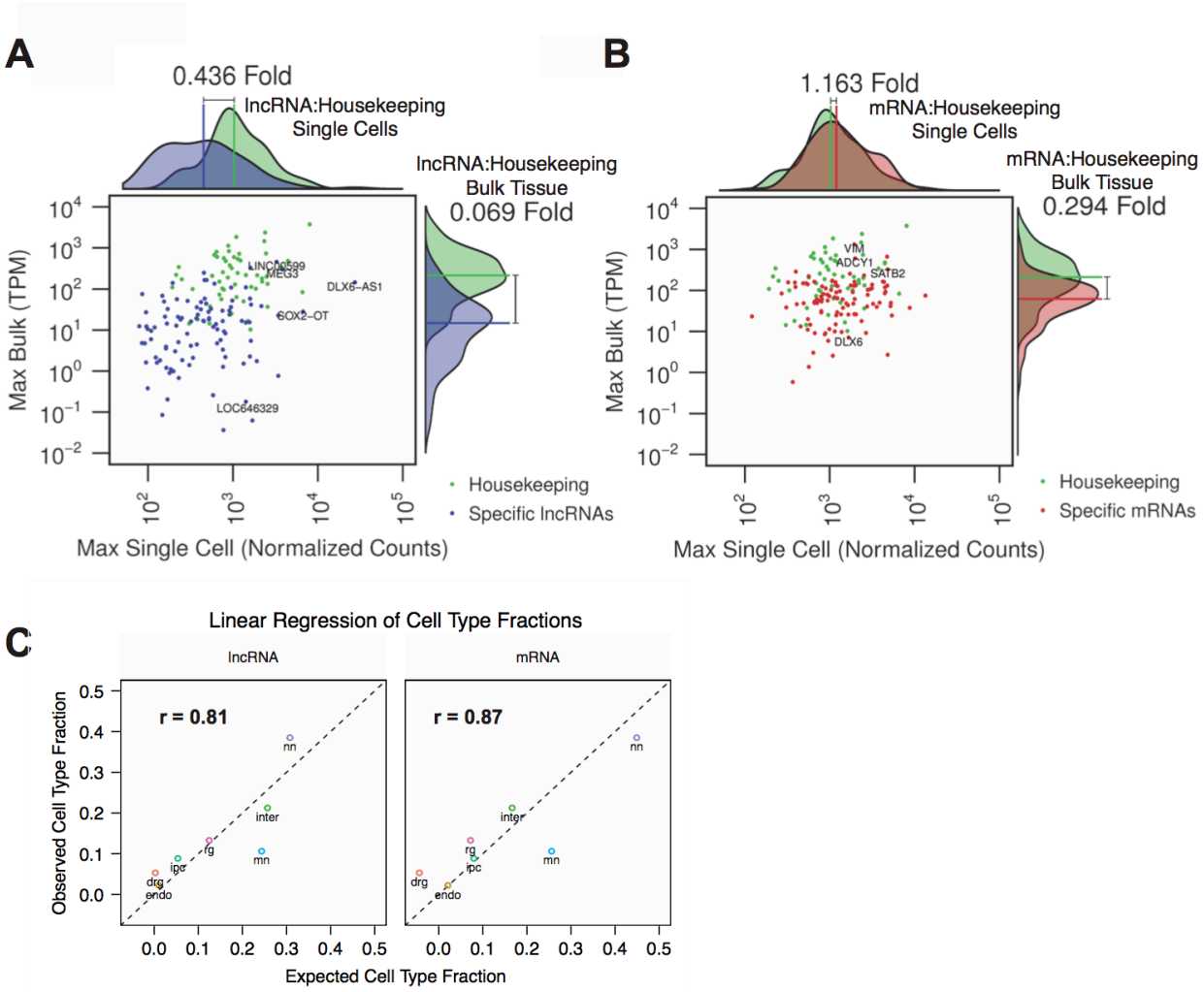

A
0.436 Fold
lncRNA:Housekeeping
Single Cells
lncRNA:Housekeeping
Bulk Tissue
0.069 Fold
Max Bulk (TPM)
LINC00599
MEG3
DLX6-AS1
SOX2-OT
LOC646329
Housekeeping
Specific lncRNAs
Max Single Cell (Normalized Counts)
B
1.163 Fold
mRNA:Housekeeping
Single Cells
mRNA:Housekeeping
Bulk Tissue
0.294 Fold
Max Bulk (TPM)
VIM
ADCY1
SATB2
DLX6
Housekeeping
Specific mRNAs
Max Single Cell (Normalized Counts)
C
Linear Regression of Cell Type Fractions
lncRNA
mRNA
r = 0.81
r = 0.87
Observed Cell Type Fraction
Expected Cell Type Fraction
nn
inter
rg
ipc
mn
drg
endo

Figure 6. Cell type-specific lncRNAs appear to be lowly expressed in bulk tissues.

A) Comparison of single cell and bulk tissue maximum expression levels of 105 cell type-specific lncRNAs and (B) 105 cell type-specific mRNAs. Green, housekeeping genes. Blue, cell type-specific lncRNAs. Red, cell type-specific mRNAs. Projected density plots summarize expression levels of scatterplots along the single cell (horizontal) and bulk tissue (vertical) axes. Fold changes noted alongside the projected density plots represent the ratio of the median expression of cell type-specific lncRNAs or mRNAs to the median expression of housekeeping genes in single cell or whole tissue RNA-seq. C) Comparison of expected cell type fractions as predicted by linear regression (x axis) and observed cell type fractions (y axis). TPM, Transcripts per million. Endo, endothelial. rg, radial glia. drg, dividing radial glia. ipc, intermediate progenitor cell. nn, newborn neurons. mn, maturing neurons. inter, interneurons.

Figure 7

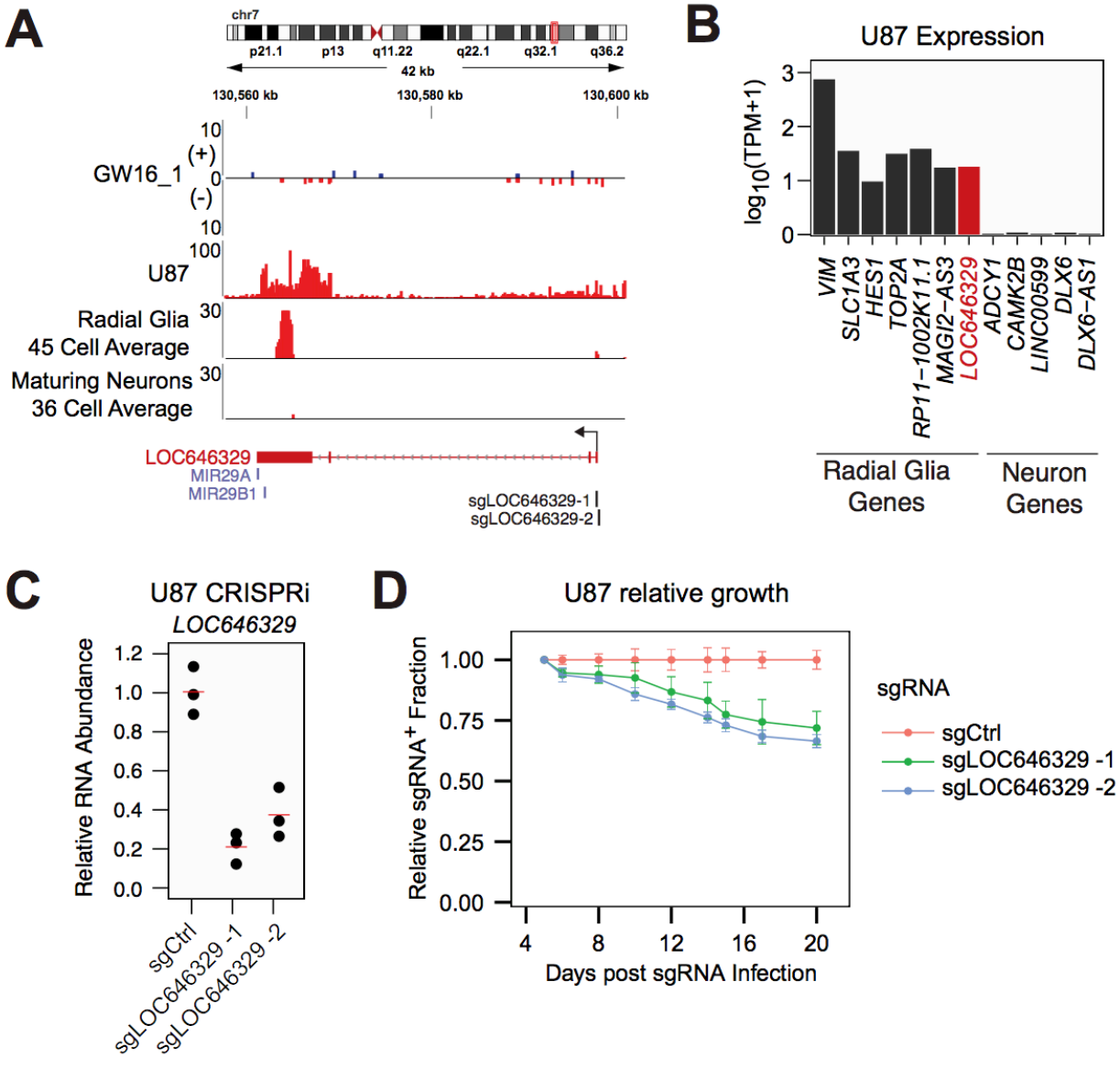
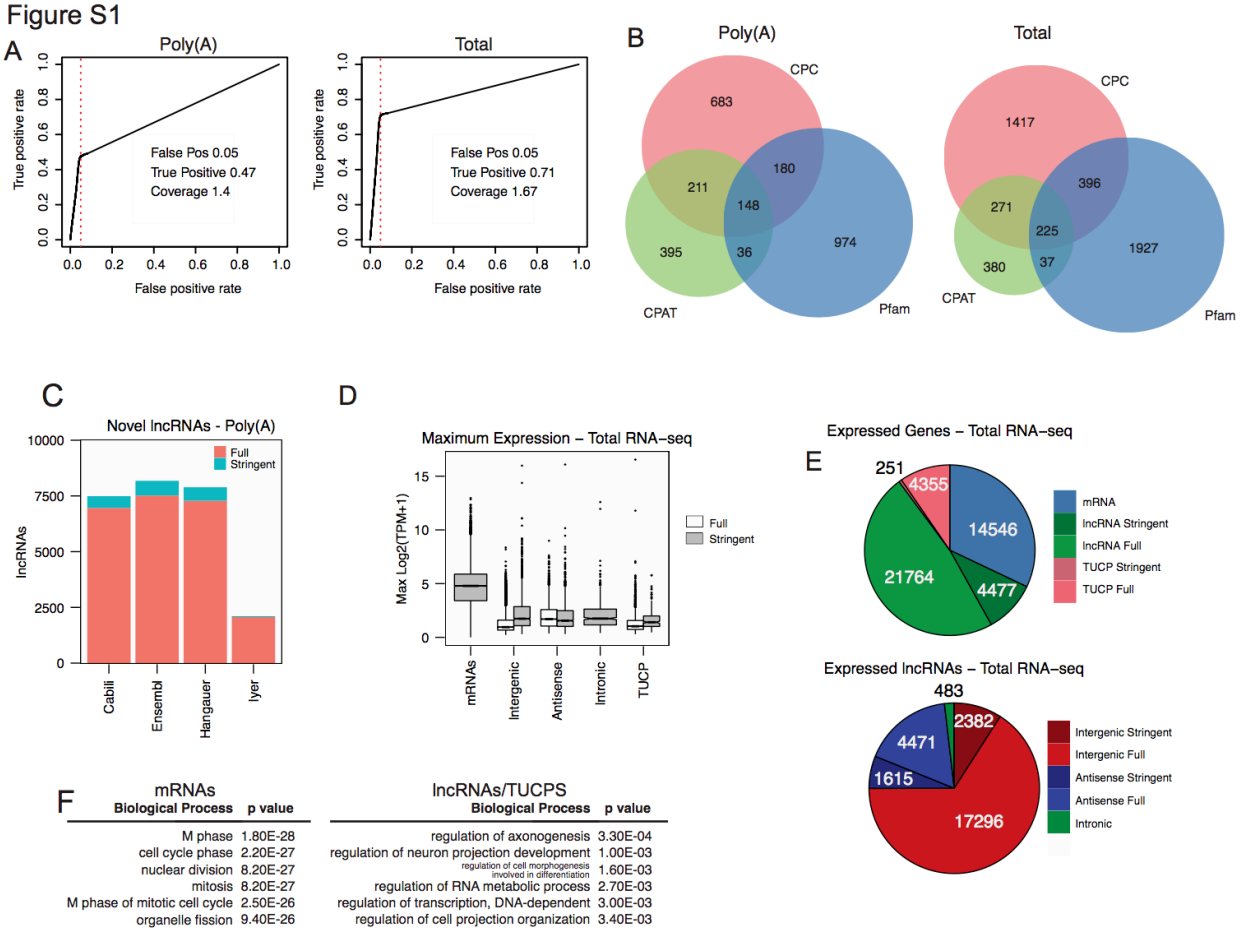

A
chr7
p21.1
p13
q11.22
q22.1
q32.1
q36.2
42 kb
130,560 kb
130,580 kb
130,600 kb
GW16_1
(+)
(-)
U87
Radial Glia
45 Cell Average
Maturing Neurons
36 Cell Average
LOC646329
MIR29A
MIR29B1
sgLOC646329-1
sgLOC646329-2
B
U87 Expression
log10(TPM+1)
VIM
SLC1A3
HES1
TOP2A
RP11-1002K11.1
MAGI2-AS3
LOC646329
ADCY1
CAMK2B
LINC00599
DLX6
DLX6-AS1
Radial Glia Genes
Neuron Genes
C
U87 CRISPRi
LOC646329
Relative RNA Abundance
sgCtrl
sgLOC646329 -1
sgLOC646329 -2
D
U87 relative growth
Relative sgRNA+ Fraction
Days post sgRNA Infection
sgRNA
sgCtrl
sgLOC646329 -1
sgLOC646329 -2

Figure 7. CRISPRi knockdown of radial glia-enriched lncRNA *LOC646329* inhibits proliferation

A) RNA-seq alignments at the *LOC646329* locus. GW16(+)/(-) replicate 1 and U87 alignments are number of reads. Radial glia and maturing neurons are merged alignments normalized by number of cells within each cell cluster. sgRNAs targeting the TSS of *LOC646329* are indicated. B) Expression of radial glia and neuron mRNAs and lncRNAs in U87 glioblastoma cells. C) qPCR of *LOC646329* following 4d of CRISPRi knockdown using 2 sgRNAs targeting the TSS of *LOC646329* relative to non-targeting control sgRNA. Biological triplicates (black circles) show 79.0% repression with sgLOC646329-1 (p=0.0014; Welch's t-test) and 62.6% repression in sgLOC646329-2 (p=0.0035; Welch's t-test). Red lines, mean. D) Relative growth assays of U87 cells following sgRNA infection. $sgRNA^+$ fraction was calculated relative to 5d post sgRNA infection and normalized to the $sgCtrl^+$ fraction at each timepoint. Biological triplicates show 28.1% depletion at 20d with sgLOC646329-1 (p=0.0073; Welch's t-test) and 33.5% depletion with sgLOC646329-2 (p=0.00048; Welch's t-test). Error bars, standard deviation of triplicate cultures.

Supplementary Figures

Figure S1



F

mRNAs

Biological Process	p value
M phase	1.80E-28
cell cycle phase	2.20E-27
nuclear division	8.20E-27
mitosis	8.20E-27
M phase of mitotic cell cycle	2.50E-26
organelle fission	9.40E-26

lncRNAs/TUCPS

Biological Process	p value
regulation of axonogenesis	3.30E-04
regulation of neuron projection development	1.00E-03
regulation of cell morphogenesis involved in differentiation	1.60E-03
regulation of RNA metabolic process	2.70E-03
regulation of transcription, DNA-dependent	3.00E-03
regulation of cell projection organization	3.40E-03

Figure S1. Computational pipeline of lncRNA identification

A) ROC analysis of coverage thresholds required to fully annotate known transcripts in poly(A) and total RNA-seq data. Coverage thresholds were selected for which false positive rates were 0.05, corresponding to the optimal balance of false positive and true positive rates. B) Venn diagrams of novel transcripts classified as protein coding by CPC, CPAT, and Pfam search. The union of the transcripts were annotated TUCPs. C) Number of novel poly(A) lncRNAs expressed in the Full (red) and Stringent (blue) references compared to Cabili et. al. 2013, Ensembl 75/GENCODE 19, Hangauer et. al. 2013, and Iyer et. al. 2015 (MiTranscriptome). D) Maximum expression levels of transcripts described in the Full (white) and Stringent (gray) references derived total RNA-seq across all samples. TPM, Transcripts per Million E) Numbers of expressed mRNAs, lncRNAs, and TUCPs during neocortex development in bulk tissues, according to total RNA-seq libraries. F) Top gene ontology terms for differentially expressed mRNAs (left) and mRNA neighbors of lncRNAs/TUCPs (right).

Figure S2

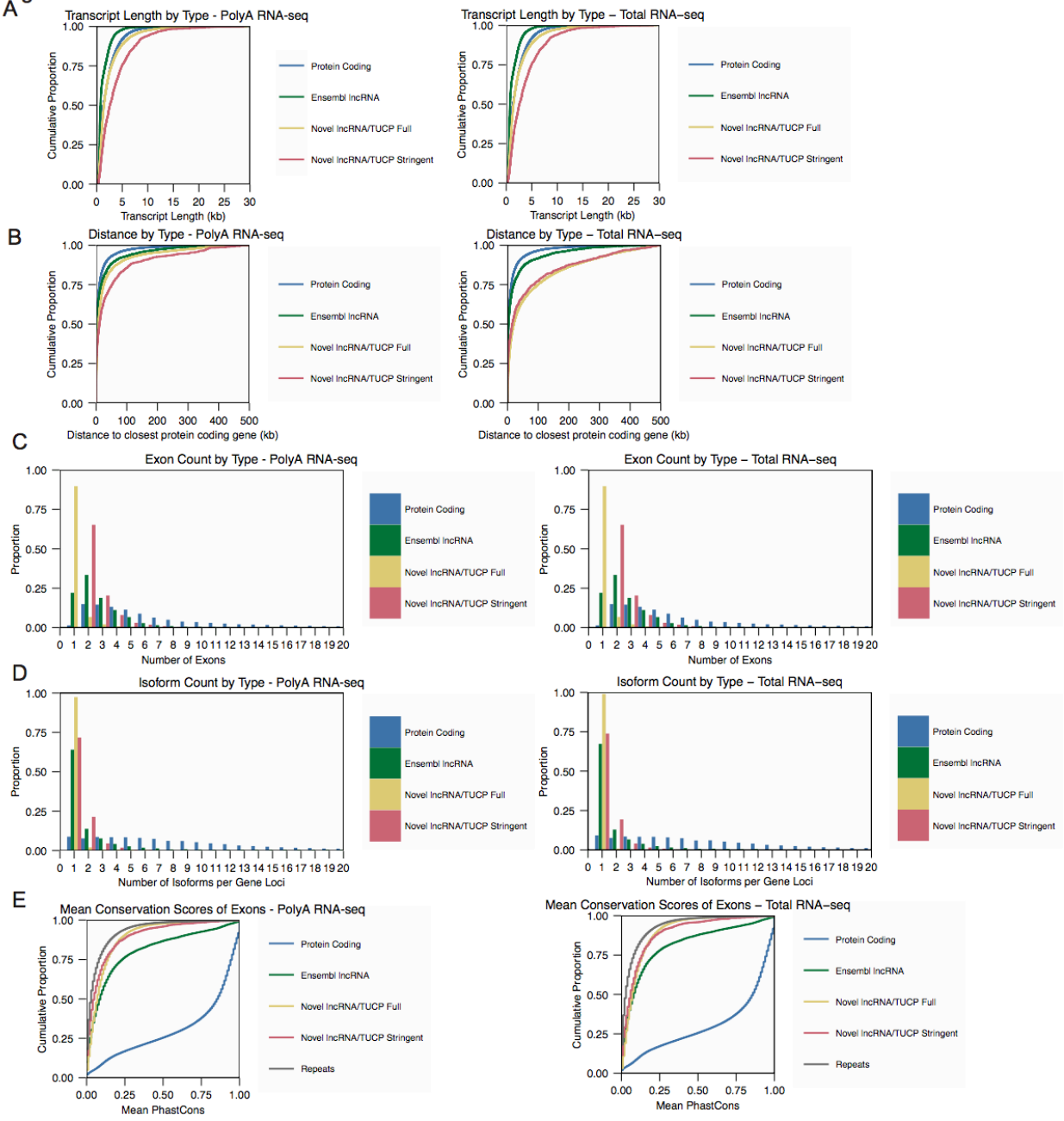

A
Transcript Length by Type - PolyA RNA-seq
Transcript Length by Type – Total RNA–seq
Cumulative Proportion
Transcript Length (kb)
B
Distance by Type - PolyA RNA-seq
Distance by Type – Total RNA–seq
Distance to closest protein coding gene (kb)
C
Exon Count by Type - PolyA RNA-seq
Exon Count by Type – Total RNA–seq
Proportion
Number of Exons
D
Isoform Count by Type - PolyA RNA-seq
Isoform Count by Type – Total RNA–seq
Number of Isoforms per Gene Loci
E
Mean Conservation Scores of Exons - PolyA RNA-seq
Mean Conservation Scores of Exons – Total RNA–seq
Mean PhastCons
Protein Coding
Ensembl lncRNA
Novel lncRNA/TUCP Full
Novel lncRNA/TUCP Stringent
Repeats

Figure S2. Genomic characteristics of lncRNAs expressed in human neocortex development

A) Cumulative distributions transcript length of mRNAs, Ensembl lncRNAs, and lncRNAs/TUCPs in the Full and Stringent references not annotated in Ensembl (novel). B) Cumulative distributions of distances between mRNAs or lncRNAs/TUCPs to their closest mRNA neighbors. Novel lncRNAs/TUCPs were farther from mRNAs than mRNAs were to their neighboring mRNAs. C) Histograms of exon counts for mRNA or lncRNA/TUCP transcripts. D) Histograms of number of unique isoforms for each mRNA or lncRNA/TUCP gene. E) Cumulative distributions of mean exonic PhastCons conservation scores for mRNAs or lncRNAs/TUCPs. All lncRNAs/TUCPs were less conserved than mRNAs, but were more conserved than repeat regions. Left panels represent results using polyA RNA data, right panels represent results using total RNA data.

Figure S3

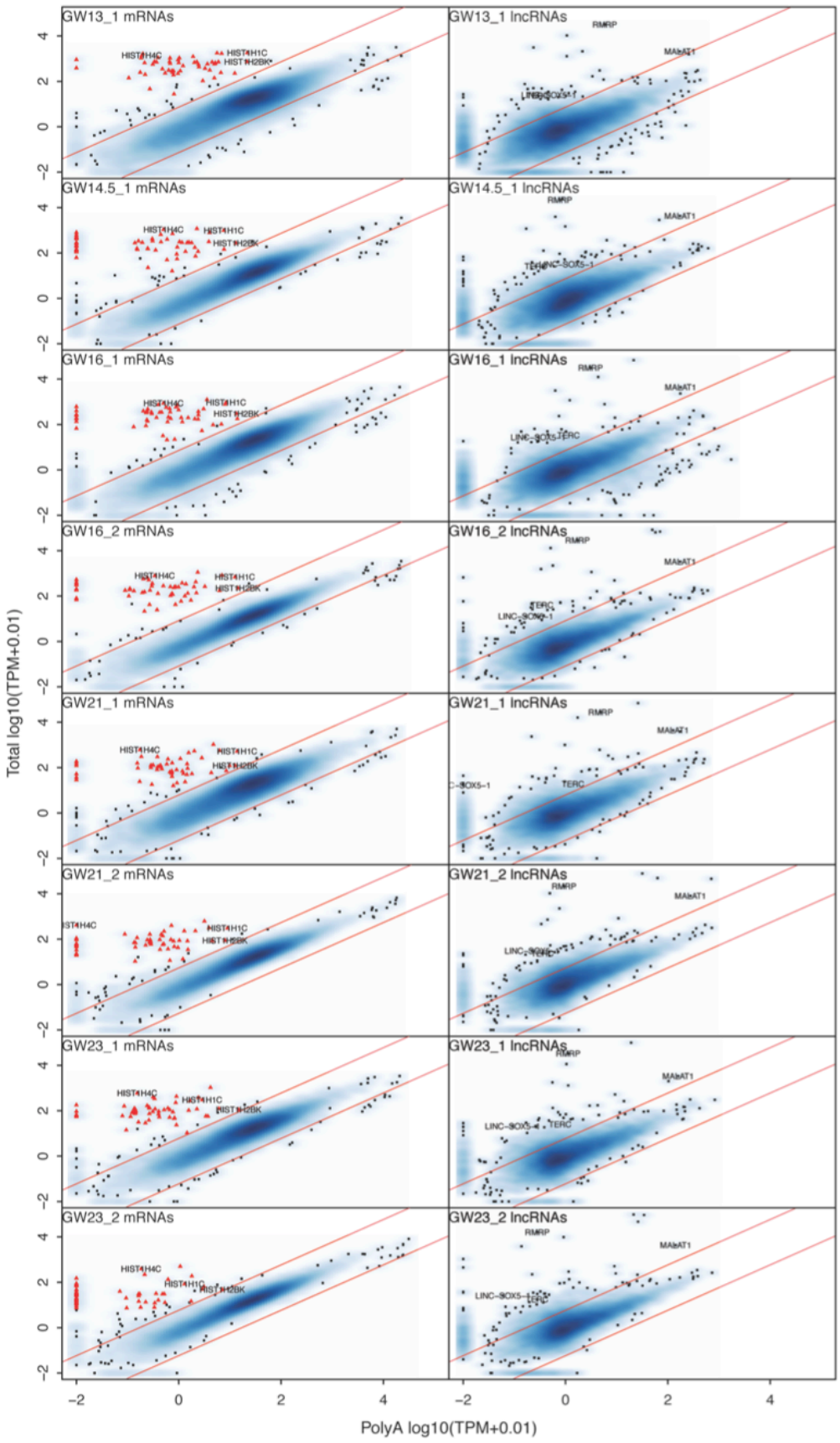

GW13_1 mRNAs
GW13_1 lncRNAs
GW14.5_1 mRNAs
GW14.5_1 lncRNAs
GW16_1 mRNAs
GW16_1 lncRNAs
GW16_2 mRNAs
GW16_2 lncRNAs
GW21_1 mRNAs
GW21_1 lncRNAs
GW21_2 mRNAs
GW21_2 lncRNAs
GW23_1 mRNAs
GW23_1 lncRNAs
GW23_2 mRNAs
GW23_2 lncRNAs
HIST4H4C
HIST1H1C
HIST1H2BK
RMRP
MALAT1
TERC
Total log10(TPM+0.01)
PolyA log10(TPM+0.01)
−2
0
2
4

Figure S3. Identification of lncRNAs enriched in total RNA-seq
Smooth scatter plot comparisons of mRNA (left) and lncRNA (right) expression levels between polyA RNA-seq and total RNA-seq in all 8 samples. Genes not expressed either polyA or total RNA-seq were omitted. 58 mRNAs and 85 lncRNAs were consistently > 10 fold enriched in total RNA-seq across all samples. Only one gene, NDUFC2-KCTD14, was identified as consistently enriched in the polyA selected RNA-seq. Red diagonals represent 10 fold enrichment in either total (upper) or polyA (lower) fractions. Red triangles, histone subunits enriched > 10 fold in total RNA. TPM, Transcripts per Million.

Figure S4

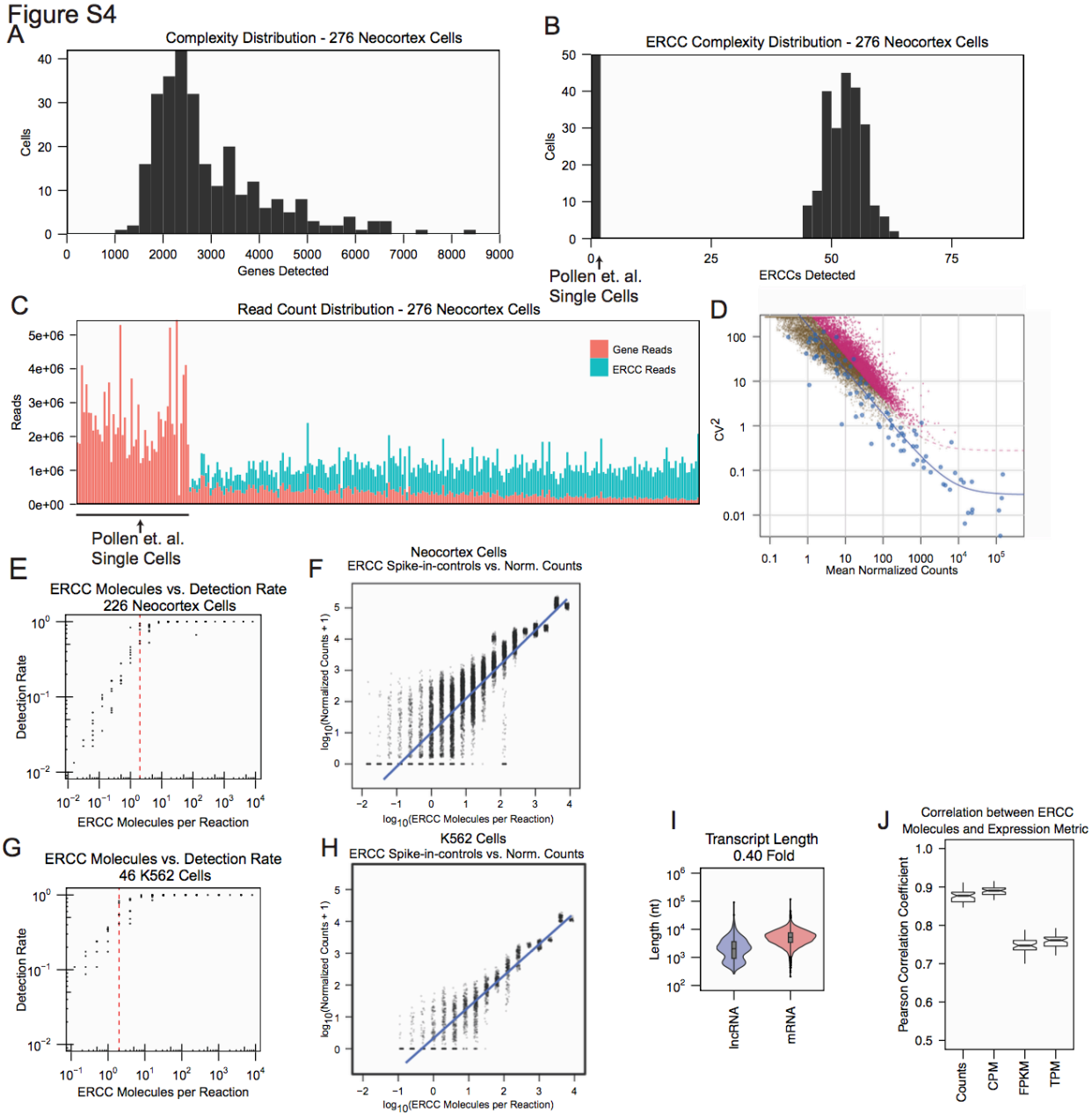

A
Complexity Distribution - 276 Neocortex Cells
Cells
Genes Detected
B
ERCC Complexity Distribution - 276 Neocortex Cells
Cells
ERCCs Detected
Pollen et. al.
Single Cells
C
Read Count Distribution - 276 Neocortex Cells
Gene Reads
ERCC Reads
Reads
Pollen et. al.
Single Cells
D
CV^2
Mean Normalized Counts
E
ERCC Molecules vs. Detection Rate
226 Neocortex Cells
Detection Rate
ERCC Molecules per Reaction
F
Neocortex Cells
ERCC Spike-in-controls vs. Norm. Counts
log10(Normalized Counts + 1)
log10(ERCC Molecules per Reaction)
G
ERCC Molecules vs. Detection Rate
46 K562 Cells
Detection Rate
ERCC Molecules per Reaction
H
K562 Cells
ERCC Spike-in-controls vs. Norm. Counts
log10(Normalized Counts + 1)
log10(ERCC Molecules per Reaction)
I
Transcript Length
0.40 Fold
Length (nt)
lncRNA
mRNA
J
Correlation between ERCC
Molecules and Expression Metric
Pearson Correlation Coefficient
Counts
CPM
FPKM
TPM

Figure S4. Summary statistics and quality control of single cell RNA-seq

A) Distribution of numbers of genes detected in each of the 276 neocortex single cells analyzed. The >1000 genes detected threshold was selected due to rapid decay of complexity below 1000 genes, potentially due to failed lysis of some cells. B) Distribution of numbers of ERCC species detected in each of the 276 neocortex single cells analyzed. The set of cells at 0 ERCCs represents single cell libraries previously sequenced and analyzed in Pollen et. al. 2014, which lack ERCCs. C) Number of reads mapping to genes (red) and ERCC species (blue) in each of the 276 neocortex single cells. D) Identification of highly variable genes (red; FDR < 0.05; Chi-squared distribution) used for cell type inference by modeling technical noise of ERCC Spike-In Controls (blue circles). Blue curve represents a gamma generalized linear model relating the coefficients of variation squared for ERCCs as a function of mean normalized counts. Dotted red curve represents 50% biological variance above technical noise. E) Detection rate of ERCC Spike-In Control RNA at predetermined quantities of spikes across 226 neocortex single cells. Red dotted line represents 74% mean detection rate at 2 copies per cell. F) Linear regression model relating size factor normalized counts to ERCC molecule quantity. G) Detection rate of ERCC Spike-In Control RNA at predetermined quantities of spikes across 46 K562 single cells. Red dotted line represents 62% mean detection rate at 2 copies per cell. H) Linear regression model relating normalized counts to ERCC molecule quantity in K562 cells. I) Transcript lengths of lncRNAs (blue) and mRNAs (red). Median lncRNA length was 0.40 fold that of median mRNA length. The omission of length normalization in single cell data therefore does not artificially overestimate lncRNA abundance. J) Pearson correlation coefficients between ERCC molecule quantity and various expression metrics in 46 K562 single cells. CPM, Counts per Million mapped reads. FPKM, Fragments per Kilobase per Million mapped reads. TPM,

Transcripts per Million. Count based metrics (Counts, CPM) outperformed length-normalized metrics (FPKM, TPM) in single cell RNA-seq data prepared using the SMARTer method.

Figure S5

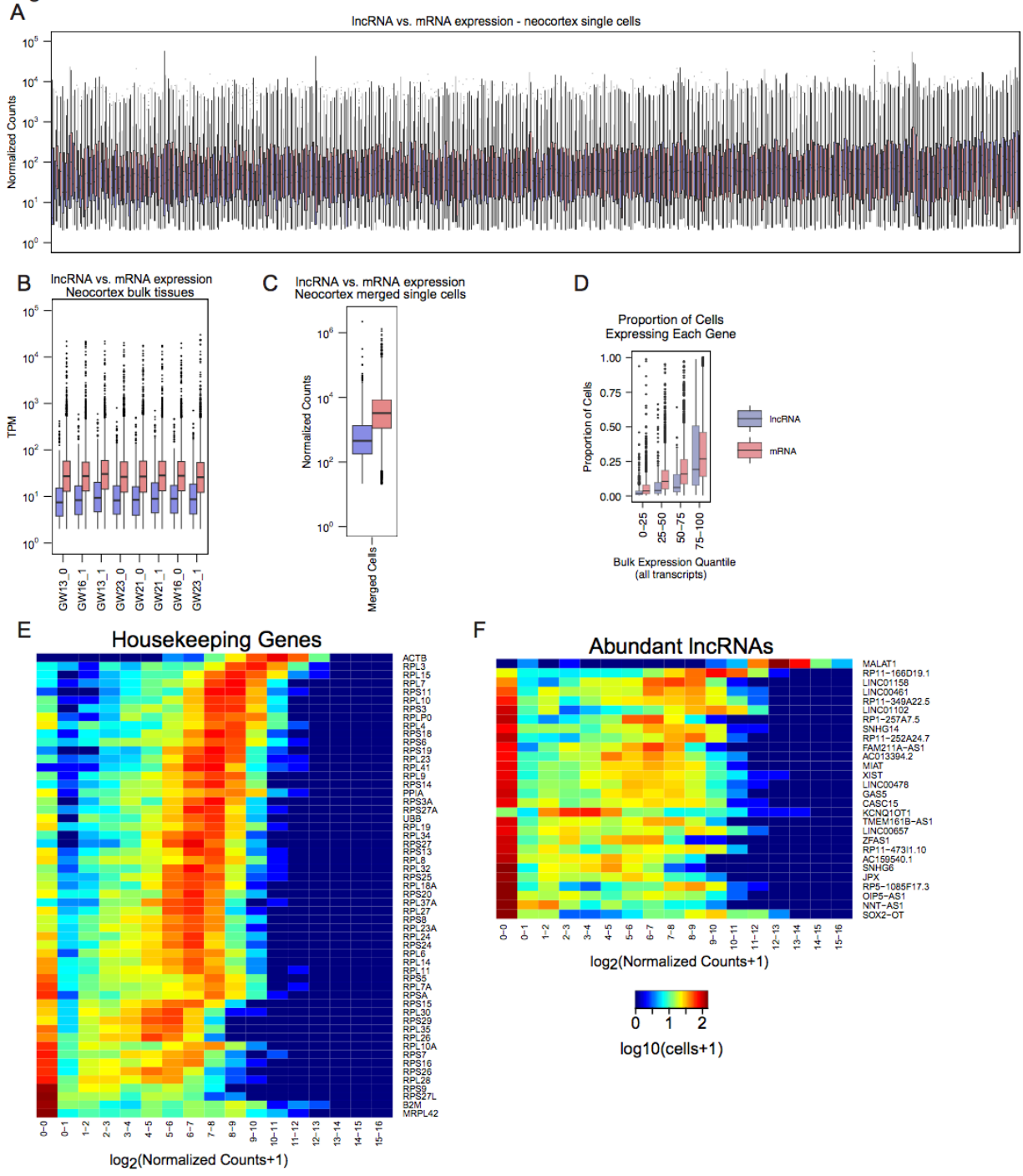

A
lncRNA vs. mRNA expression - neocortex single cells
Normalized Counts
B
lncRNA vs. mRNA expression
Neocortex bulk tissues
TPM
GW13_0
GW16_1
GW13_1
GW23_0
GW21_0
GW21_1
GW16_0
GW23_1
C
lncRNA vs. mRNA expression
Neocortex merged single cells
Normalized Counts
Merged Cells
D
Proportion of Cells
Expressing Each Gene
Proportion of Cells
lncRNA
mRNA
0–25
25–50
50–75
75–100
Bulk Expression Quantile
(all transcripts)
E
Housekeeping Genes
ACTB
RPL3
RPL15
RPL7
RPS11
RPL10
RPS3
RPLP0
RPL4
RPS18
RPS6
RPS19
RPL23
RPL41
RPL9
RPS14
PPIA
RPS3A
RPS27A
UBB
RPL19
RPL34
RPS27
RPS13
RPL8
RPL32
RPS25
RPL18A
RPS20
RPL37A
RPL27
RPS8
RPL23A
RPL24
RPS24
RPL6
RPL14
RPL11
RPS5
RPL7A
RPSA
RPS15
RPL30
RPS29
RPL35
RPL26
RPL10A
RPS7
RPS16
RPS26
RPL28
RPS9
RPS27L
B2M
MRPL42
0–0
0–1
1–2
2–3
3–4
4–5
5–6
6–7
7–8
8–9
9–10
10–11
11–12
12–13
13–14
14–15
15–16
log2(Normalized Counts+1)
F
Abundant lncRNAs
MALAT1
RP11–166D19.1
LINC01158
LINC00461
RP11–349A22.5
LINC01102
RP1–257A7.5
SNHG14
RP11–252A24.7
FAM211A–AS1
AC013394.2
MIAT
XIST
LINC00478
GAS5
CASC15
KCNQ1OT1
TMEM161B–AS1
LINC00657
ZFAS1
RP11–473I1.10
AC159540.1
SNHG6
JPX
RP5–1085F17.3
OIP5–AS1
NNT–AS1
SOX2–OT
0–0
0–1
1–2
2–3
3–4
4–5
5–6
6–7
7–8
8–9
9–10
10–11
11–12
12–13
13–14
14–15
15–16
log2(Normalized Counts+1)
0
1
2
log10(cells+1)

Figure S5. Expression of lncRNAs and mRNAs in single cells and whole tissues

A) Distributions of non-zero lncRNA (blue) and mRNA (red) expression in 276 single cells from neocortex. The median lncRNA expression and the median mRNA expression for each cell was compared as a ratio and summarized in Figure 3B. B) Distributions of non-zero lncRNA (blue) and mRNA (red) expression in 8 bulk tissue RNA-seq samples using the same set of 1400 lncRNAs and 10929 mRNAs as in A). C) Distributions of non-zero lncRNA (blue) and mRNA (red) expression in *in silico* merged neocortex single cells. D) Proportion of single neocortex cells that expressed each lncRNA (blue) and mRNA (red), binned by their expression levels in bulk tissues. E) Distributions of housekeeping gene and lncRNA (F) expression levels in neocortex single cells, binned by log2(Normalized Counts + 1). Colors represent number of cells in each bin. Abundant lncRNAs were ranked by their median expression levels across 276 neocortex single cells.

Figure S6

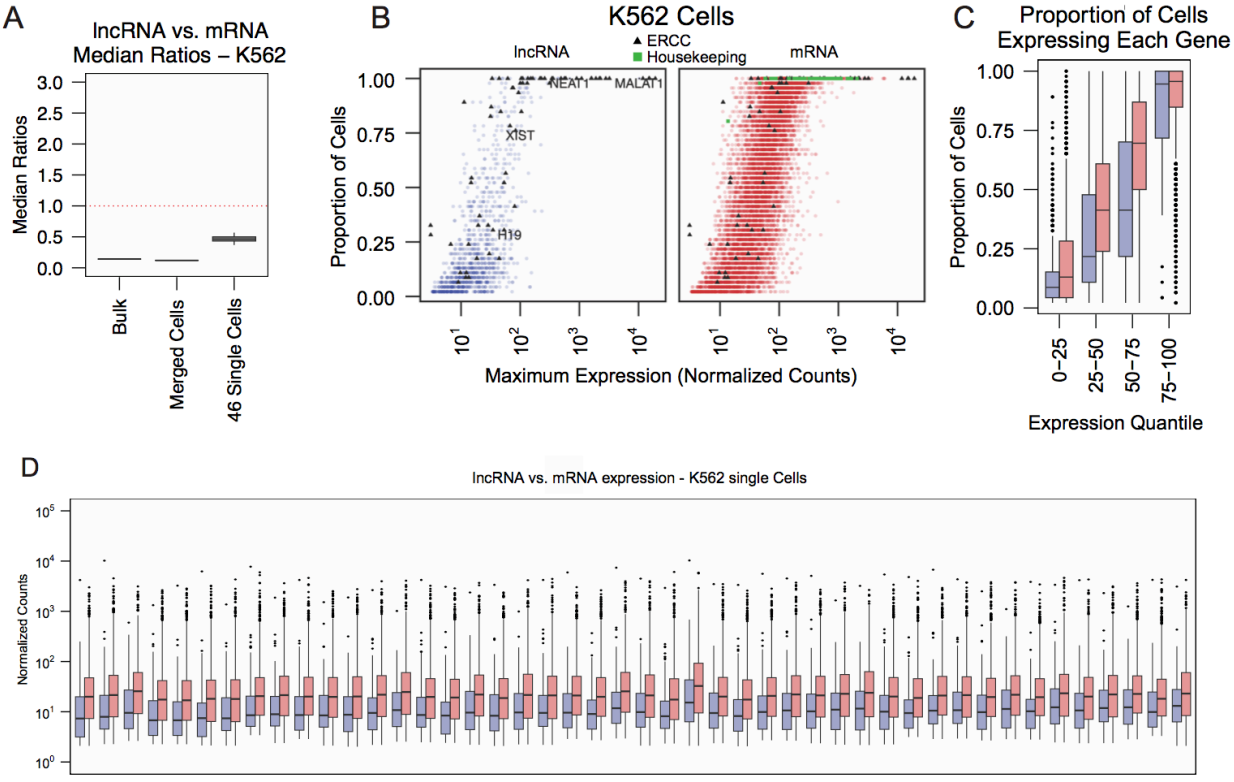

A
lncRNA vs. mRNA
Median Ratios – K562
Median Ratios
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Bulk
Merged Cells
46 Single Cells
B
K562 Cells
lncRNA
ERCC
Housekeeping
mRNA
Proportion of Cells
1.00
0.75
0.50
0.25
0.00
NEAT1
MALAT1
XIST
H19
Maximum Expression (Normalized Counts)
C
Proportion of Cells
Expressing Each Gene
Proportion of Cells
0–25
25–50
50–75
75–100
Expression Quantile
D
lncRNA vs. mRNA expression - K562 single Cells
Normalized Counts

Figure S6. Single cell transcriptomics of lncRNA expression in K562 cell cultures

A) Distributions of median lncRNA expression to median mRNA expression ratios (lncRNA:mRNA) in populations, *in silico* merged single cells, and single cells from K562 cultures. B) Proportion of K562 cells that expressed each lncRNA (blue) and mRNA (red), separated by maximum expression in single cells. C) Same as in B) but grouped by maximum expression quantile. D) Distributions of non-zero lncRNA (blue) and mRNA (red) expression in 46 single K562 cells. Green squares, housekeeping genes. Black triangles, ERCC Spike-In Controls.

Figure S7

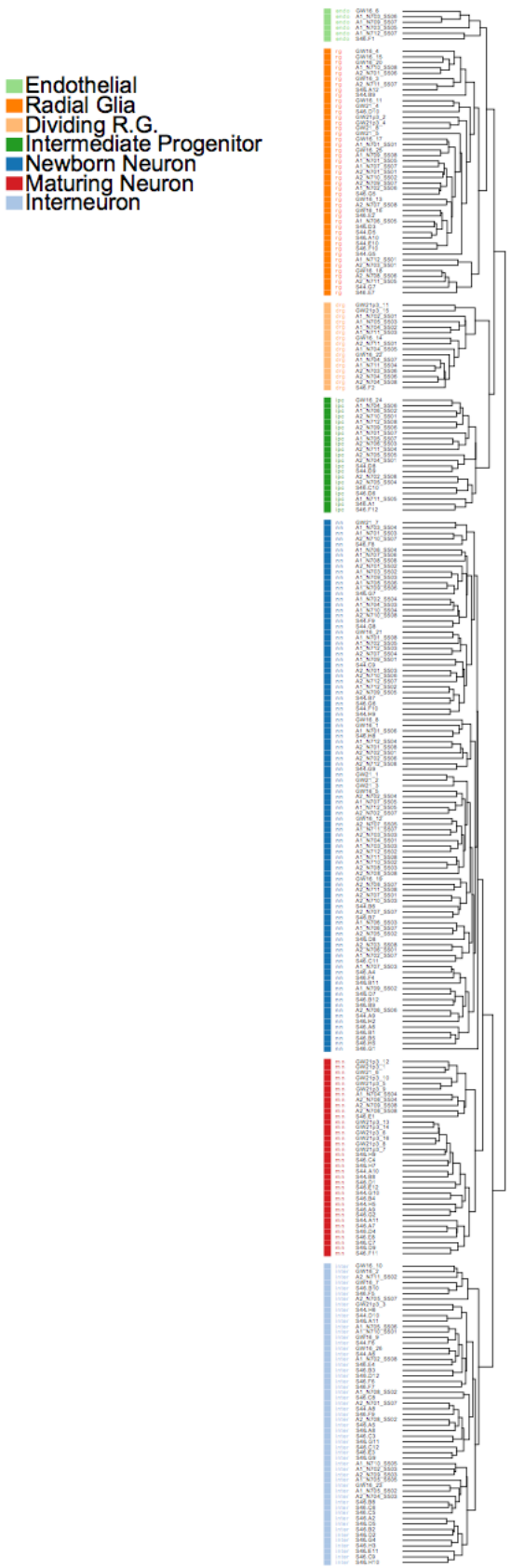

Endothelial
Radial Glia
Dividing R.G.
Intermediate Progenitor
Newborn Neuron
Maturing Neuron
Interneuron

Figure S7. Hierarchical Clustering of Neocortex Single Cells

Results of complete linkage hierarchical clustering of 276 neocortex single cells, using the top 500 genes ranked by the highest absolute values of gene loading scores across the first four principal components following PCA. Endo, endothelial. RG, radial glia. DRG, dividing radial glia. IPC, intermediate progenitor cell. NN, newborn neurons. MN, maturing neurons. INTER, interneurons.

Figure S8

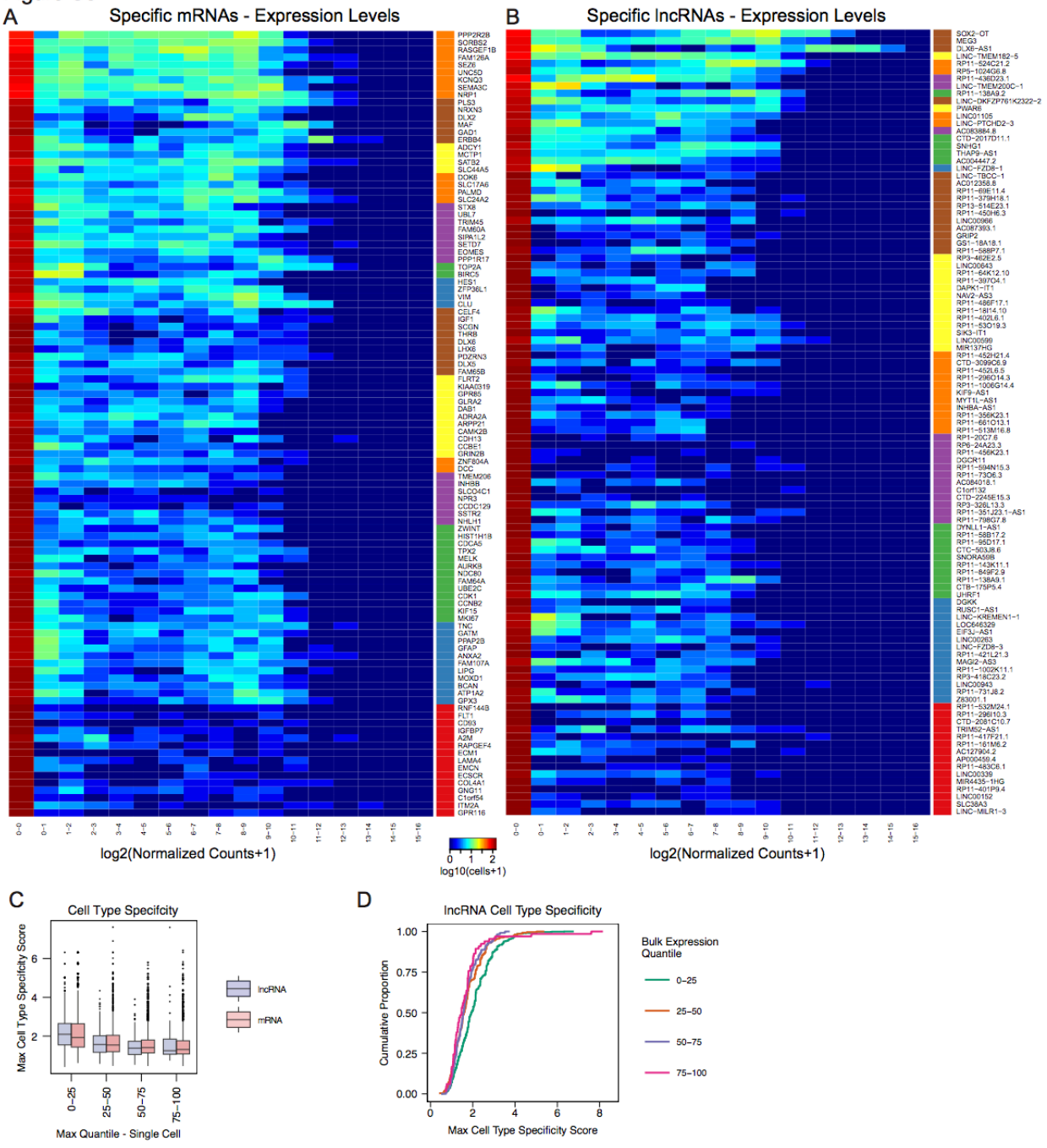
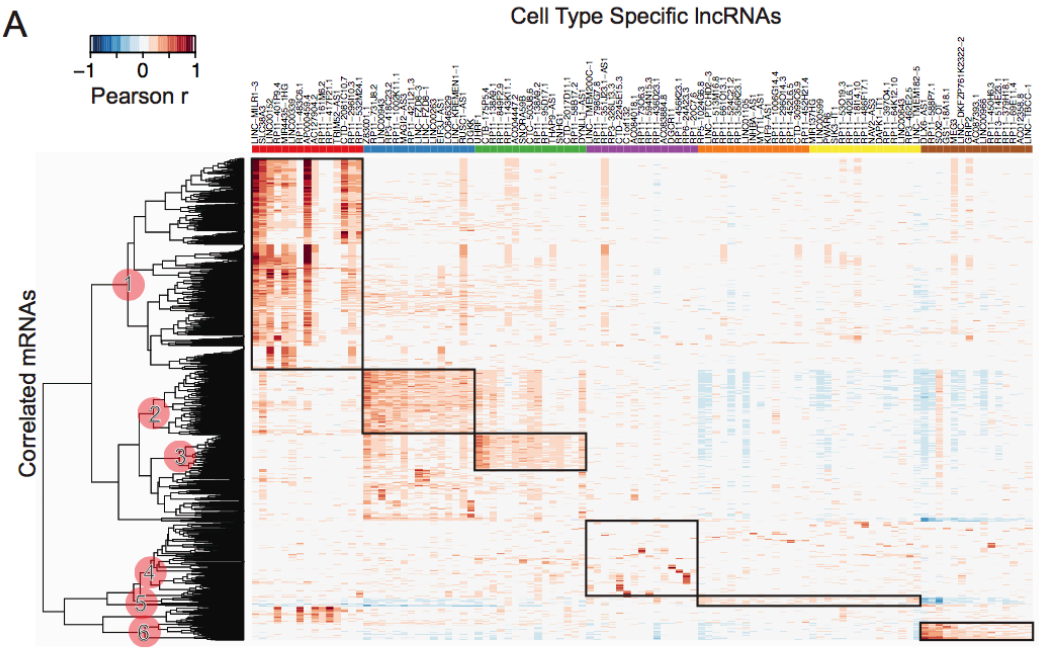

A
Specific mRNAs - Expression Levels
log2(Normalized Counts+1)
log10(cells+1)
B
Specific lncRNAs - Expression Levels
log2(Normalized Counts+1)
C
Cell Type Specifcity
Max Cell Type Specificity Score
lncRNA
mRNA
0–25
25–50
50–75
75–100
Max Quantile - Single Cell
D
lncRNA Cell Type Specificity
Cumulative Proportion
Max Cell Type Specificity Score
Bulk Expression Quantile
0–25
25–50
50–75
75–100

Figure S8. Abundant expression of cell type-specific lncRNAs in single cells

A) Distributions of expression levels of cell type-specific mRNAs and (B) lncRNAs in single neocortex cells, binned by log2(Normalized Counts + 1). Heatmap colors represent number of cells in each bin. Genes were ranked by their median expression levels across 276 neocortex single cells. Columns adjacent to gene names represent cell type assignments. C) Distributions of maximum log odds ratios for cell type-specific gene expression at 75%ile threshold for each lncRNA (blue) and mRNA (red) at different quantiles of expression. D) Cumulative distributions of cell type specificity scores for lncRNAs expressed at various levels in bulk tissues. Endo/red, endothelial. RG/blue, radial glia. DRG/green, dividing radial glia. IPC/purple, intermediate progenitor cell. NN/orange, newborn neurons. MN/yellow, maturing neurons. INTER/brown, interneurons.

Figure S9

A



B

1

Biological Process	p value
angiogenesis	7.39E-20
regulation of vasculature development	1.68E-11
hemostasis	7.47E-11
response to oxygen levels	3.01E-10
blood coagulation	1.72E-10
coagulation	1.72E-10
regulation of angiogenesis	5.93E-11
response to decreased oxygen levels	1.58E-09
response to hypoxia	4.60E-09
extracellular matrix organization	2.45E-09

2

Biological Process	p value
negative regulation of nervous system development	8.26E-12
negative regulation of neuron differentiation	4.38E-12
negative regulation of cell development	5.66E-11
negative regulation of neurogenesis	1.77E-10
glial cell differentiation	1.52E-09
response to mechanical stimulus	4.55E-09
regulation of neuron differentiation	1.46E-07
extracellular matrix organization	7.32E-07
extracellular structure organization	7.58E-07
positive regulation of neuroblast proliferation	2.96E-06

3

Biological Process	p value
mitotic cell cycle	8.94E-35
nuclear division	5.00E-30
organelle fission	5.40E-29
mitotic nuclear division	1.40E-29
chromosome segregation	7.81E-23
regulation of cell cycle process	6.82E-13
cell cycle checkpoint	1.89E-13
chromosome organization	3.60E-13
DNA repair	3.77E-12
microtubule-based process	1.31E-10

4

Biological Process	p value
dicarboxylic acid biosynthetic process	4.83E-03
glutamine family amino acid biosynthetic process	1.25E-02
GPI anchor metabolic process	1.64E-03
regulation of triglyceride biosynthetic process	9.22E-03
glutamate metabolic process	1.88E-02
neuroblast proliferation	9.22E-03
neuromuscular synaptic transmission	1.36E-02
GPI anchor biosynthetic process	1.74E-02
positive regulation of triglyceride metabolic process	1.03E-02
positive regulation of triglyceride biosynthetic process	4.83E-03

5

Biological Process	p value
limb bud formation	4.77E-07
cardiac ventricle morphogenesis	4.82E-06
cardiac chamber morphogenesis	6.68E-06
axon extension	8.96E-06
regulation of neuron differentiation	7.58E-06
neuron projection extension	1.38E-05
positive regulation of neuron differentiation	6.30E-06
positive regulation of neurogenesis	2.22E-05
glial cell development	2.01E-05
regulation of neuron projection development	1.94E-05

6

Reactome	p value
GABA synthesis, release, reuptake and degradation	3.25E-04
Transmission across Chemical Synapses	1.97E-03
Neurotransmitter Release Cycle	1.44E-03
Nuclear Receptor transcription pathway	2.09E-03
Signaling by ERBB2	1.90E-02
Signaling by FGFR	1.87E-02
Signaling by FGFR in disease	2.21E-02
Neuronal System	5.02E-03
Downstream signal transduction	1.85E-02
Downstream signaling of activated FGFR	1.60E-02

Figure S9. Gene co-expression analysis of cell type-specific lncRNAs in single cells

A) Matrix of Pearson correlation coefficients between cell type-specific lncRNAs and the top 10% most correlated or anticorrelated mRNAs across single cells. B) Gene ontology terms for gene clusters identified in the correlation matrix. Cluster nodes are labeled with red circles. Header bar colors for lncRNAs: red, endothelial. blue, radial glia. green, dividing radial glia. purple, intermediate progenitor cell. orange, newborn neurons. yellow, maturing neurons. brown, interneurons.

Figure S10

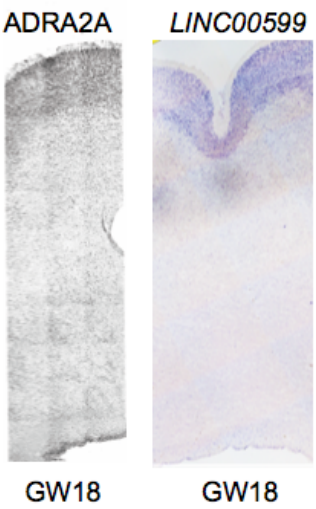


Figure S10. Immunohistochemistry of maturing neuron marker ADRA2A

Immunohistochemistry of maturing neuron marker protein ADRA2A (left) compared to *in situ* hybridization against maturing neuron lncRNA *LINC00599* (right, reproduced from Figure 5).

Supplemental Tables

Table S1

Sample and sequencing statistics. Developmental stages, number of reads sequenced and aligned for each bulk RNA-seq and single cell RNA-seq sample (separate tabs in xls file). Single cell RNA-seq statistics also include number of unique genes and ERCC Spike-In Control species detected, and how many reads were assigned to either.

Table S2

Expression table of all bulk tissue samples and genes (including known and novel lncRNAs) according to the polyA Full transcriptome reference. Expression values are in TPM (Transcripts per Million).

Table S3

Table of novel lncRNAs not annotated in Ensembl release 75. Gene names and coordinates are in BED format. These lncRNAs were the set of novel lncRNAs annotated in the polyA full reference, which includes both multi-exon and single exon lncRNAs.

Table S4

Expression table of all bulk tissue samples and differentially expressed genes, using polyA data. Expression values are in TPM (Transcripts per Million).

Table S5

Expression table of all bulk tissue samples showing genes enriched in either total RNA-seq or polyA RNA-seq. Both polyA and total RNA-seq data are represented for each tissue sample. Expression values are in TPM (Transcripts per Million).

Table S6

Single cell expression table. 276 single cells from human neocortex development and all genes (including known and novel lncRNAs) that pass the minimum expression threshold (Methods). Expression values are in size factor normalized counts, according to DESeq.

Table S7

Table of 105 cell type-specific mRNAs (1st tab) and lncRNAs (2nd tab) and their expression enrichment scores, which are log2 transformed, for each of the 7 cell type clusters.

References

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol *11*, R106.

Berghoff, E.G., Clark, M.F., Chen, S., Cajigas, I., Leib, D.E., and Kohtz, J.D. (2013). Evf2 (Dlx6as) lncRNA regulates ultraconserved enhancer methylation and the differential transcriptional control of adjacent genes. Development *140*, 4407–4416.

Bond, A.M., VanGompel, M.J.W., Sametsky, E.A., Clark, M.F., Savage, J.C., Disterhoft, J.F., and Kohtz, J.D. (2009). Balanced gene regulation by an embryonic brain ncRNA is critical for adult hippocampal GABA circuitry. Nat Neurosci *12*, 1020–1027.

Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., et al. (2013). Accounting for technical noise in single-cell RNA-seq experiments. Nat Meth *10*, 1093–1095.

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

Cabili, M.N., Dunagin, M.C., McClanahan, P.D., Biaesch, A., Padovan-Merhar, O., Regev, A., Rinn, J.L., and Raj, A. (2015). Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. Genome Biol *16*, 20.

Carrieri, C., Cimatti, L., Biagioli, M., Beugnet, A., and Zucchelli, S. (2012). Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. Nature.

Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics *14*, 128.

Clark, M.B., Johnston, R.L., Inostroza-Ponta, M., Fox, A.H., Fortini, E., Moscato, P., Dinger, M.E., and Mattick, J.S. (2012). Genome-wide analysis of long noncoding RNA stability. Genome Research *22*, 885–898.

Corporation, F. (2016). Doublet Rate and Detection on the C1 IFCs White Paper.

Darmanis, S., Sloan, S.A., Zhang, Y., Enge, M., Caneda, C., Shuer, L.M., Hayden Gephart, M.G., Barres, B.A., and Quake, S.R. (2015). A survey of human brain transcriptome diversity at the single cell level. Proceedings of the National Academy of Sciences *112*, 7285–7290.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., et al. (2014). Pfam: the protein families database. Nucleic Acids Res. *42*, D222–D230.

Gilbert, L.A., Horlbeck, M.A., Adamson, B., Villalta, J.E., Chen, Y., Whitehead, E.H., Guimaraes, C., Panning, B., Ploegh, H.L., Bassik, M.C., et al. (2014). Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. Cell *159*, 647–661.

Goff, L.A., Groff, A.F., Sauvageau, M., Trayes-Gibson, Z., Sanchez-Gomez, D.B., Morse, M., Martin, R.D., Elcavage, L.E., Liapis, S.C., Gonzalez-Celeiro, M., et al. (2015). Spatiotemporal expression and transcriptional perturbations by long noncoding RNAs in the mouse brain. Proceedings of the National Academy of Sciences *112*, 6855–6862.

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. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *457*, 223–227.

Hangauer, M.J., Vaughn, I.W., and McManus, M.T. (2013). Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. PLoS Genet *9*, e1003569.

Hansen, D.V., Lui, J.H., Flandin, P., Yoshikawa, K., Rubenstein, J.L., Alvarez-Buylla, A., and Kriegstein, A.R. (2013a). Non-epithelial stem cells and cortical interneuron production in the human ganglionic eminences. Nature Publishing Group *16*, 1576–1587.

Hansen, D.V., Lui, J.H., Flandin, P., Yoshikawa, K., Rubenstein, J.L., Alvarez-Buylla, A., and Kriegstein, A.R. (2013b). Non-epithelial stem cells and cortical interneuron production in the human ganglionic eminences. Nature Publishing Group *16*, 1576–1587.

Heins, N., Malatesta, P., Cecconi, F., Nakafuku, M., Tucker, K.L., Hack, M.A., Chapouton, P., Barde, Y.-A., and Götz, M. (2002). Glial cells generate neurons: the role of the transcription factor Pax6. Nat Neurosci *5*, 308–315.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. *37*, 1–13.

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Meth *11*, 163–166.

Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner, J.R., Evans, J.R., Zhao, S., et al. (2015). The landscape of long noncoding RNAs in the human transcriptome. Nature Genetics.

Johnson, M.B., Wang, P.P., Atabay, K.D., Murphy, E.A., Doan, R.N., Hecht, J.L., and Walsh, C.A. (2015). Single-cell analysis reveals transcriptional heterogeneity of neural progenitors in human cortex. Nat Neurosci *18*, 637–646.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. Cell *161*, 1187–1201.

Kong, L., Zhang, Y., Ye, Z.-Q., Liu, X.-Q., Zhao, S.-Q., Wei, L., and Gao, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res. *35*, W345–W349.

Kotake, Y., Nakagawa, T., Kitagawa, K., Suzuki, S., Liu, N., Kitagawa, M., and Xiong, Y. (2011). Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15(INK4B) tumor suppressor gene. Oncogene *30*, 1956–1962.

Kriegstein, A., and Alvarez-Buylla, A. (2009). The glial nature of embryonic and adult neural stem cells. Annu. Rev. Neurosci. *32*, 149–184.

Lauressergues, D., Couzigou, J.-M., Clemente, H.S., Martinez, Y., Dunand, C., Bécard, G., and Combier, J.-P. (2015). Primary transcripts of microRNAs encode regulatory peptides. Nature *520*, 90–93.

Lewitus, E., and Huttner, W.B. (2015). Neurodevelopmental LincRNA Microsyteny Conservation and Mammalian Brain Size Evolution. PLoS ONE *10*, e0131818.

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

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930.

Livyatan, I., Harikumar, A., Nissim-Rafinia, M., Duttagupta, R., Gingeras, T.R., and Meshorer, E. (2013). Non-polyadenylated transcription in embryonic stem cells reveals novel non-coding RNA related to pluripotency and differentiation. Nucleic Acids Res. *41*, 6300–6315.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202–1214.

Mattick, J.S., and Rinn, J.L. (2015). Discovery and annotation of long noncoding RNAs. Nature Structural & Molecular Biology *22*, 5–7.

Mercer, T.R., Dinger, M.E., Sunkin, S.M., Mehler, M.F., and Mattick, J.S. (2008). Specific expression of long noncoding RNAs in the mouse brain. Proceedings of the National Academy of Sciences *105*, 716–721.

Mercer, T.R., Qureshi, I.A., Gokhan, S., Dinger, M.E., Li, G., Mattick, J.S., and Mehler, M.F. (2010). Long noncoding RNAs in neuronal-glial fate specification and oligodendrocyte lineage maturation. BMC Neurosci *11*, 14.

Miller, J.A., Ding, S.-L., Sunkin, S.M., Smith, K.A., Ng, L., Szafer, A., Ebbert, A., Riley, Z.L., Royall, J.J., Aiona, K., et al. (2014). Transcriptional landscape of the prenatal human brain. Nature 1–19.

Miyoshi, N., Wagatsuma, H., Wakana, S., Shiroishi, T., Nomura, M., Aisaka, K., Kohda, T., Surani, M.A., Kaneko-Ishino, T., and Ishino, F. (2000). Identification of an imprinted gene, Meg3/Gtl2 and its human homologue MEG3, first mapped on mouse distal chromosome 12 and human chromosome 14q. Genes Cells *5*, 211–220.

Molyneaux, B.J., Arlotta, P., Menezes, J.R.L., and Macklis, J.D. (2007). Neuronal subtype specification in the cerebral cortex. Nat Rev Neurosci *8*, 427–437.

Molyneaux, B.J., Goff, L.A., Brettler, A.C., Chen, H.-H., Brown, J.R., Hrvatin, S., Rinn, J.L., and Arlotta, P. (2015). DeCoN: genome-wide analysis of in vivo transcriptional dynamics during pyramidal neuron fate selection in neocortex. Neuron *85*, 275–288.

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

Pollard, K.S., Salama, S.R., Lambert, N., Lambot, M.-A., Coppens, S., Pedersen, J.S., Katzman, S., King, B., Onodera, C., Siepel, A., et al. (2006). An RNA gene expressed during cortical development evolved rapidly in humans. Nature *443*, 167–172.

Pollen, A.A., Nowakowski, T.J., Shuga, J., Wang, X., Leyrat, A.A., Lui, J.H., Li, N., Szpankowski, L., Fowler, B., Chen, P., et al. (2014). Low-coverage single-cell mRNA

sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat Biotechnol.

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

Qing, T., Yu, Y., Du, T., and Shi, L. (2013). mRNA enrichment protocols determine the quantification characteristics of external RNA spike-in controls in RNA-Seq studies. Sci China Life Sci *56*, 134–142.

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

Rakic, S., and Zecevic, N. (2003). Emerging complexity of layer I in human cerebral cortex. Cerebral Cortex.

Ramos, A.D., Diaz, A., Nellore, A., Delgado, R.N., Park, K.-Y., Gonzales-Roybal, G., Oldham, M.C., Song, J.S., and Lim, D.A. (2013). Integration of Genome-wide Approaches Identifies lncRNAs of Adult Neural Stem Cells and Their Progeny In Vivo. Cell Stem Cell *12*, 616–628.

Sauvageau, M., Goff, L.A., Lodato, S., Bonev, B., Groff, A.F., Gerhardinger, C., Sanchez-Gomez, D.B., Hacisuleyman, E., Li, E., Spence, M., et al. (2013). Multiple knockout mouse models reveal lincRNAs are required for life and brain development. eLife *2*, e01749.

Shalek, A.K., Satija, R., Adiconis, X., Gertner, R.S., Gaublomme, J.T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., et al. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature *498*, 236–240.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol *28*, 511–515.

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

Vescovi, A.L., Galli, R., and Reynolds, B.A. (2006). Brain tumour stem cells. Nat Rev Cancer *6*, 425–436.

Wallace, V.A., and Raff, M.C. (1999). A role for Sonic hedgehog in axon-to-astrocyte signalling in the rodent optic nerve. Development *126*, 2901–2909.

Wang, L., Park, H.J., Dasari, S., Wang, S., Kocher, J.-P., and Li, W. (2013). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res. *41*, e74–e74.

Wilusz, J.E., Freier, S.M., and Spector, D.L. (2008). 3' end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. Cell *135*, 919–932.

Yang, L., Duff, M.O., Graveley, B.R., Carmichael, G.G., and Chen, L.-L. (2011). Genomewide characterization of non-polyadenylated RNAs. Genome Biol *12*, R16.

Zeisel, A., Muñoz-Manchado, A.B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science *347*, 1138–1142.

Zhang, Y., Chen, K., Sloan, S.A., Bennett, M.L., Scholze, A.R., O'Keeffe, S., Phatnani, H.P., Guarnieri, P., Caneda, C., Ruderisch, N., et al. (2014). An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex. J. Neurosci. *34*, 11929–11947.

Chapter 3

Single-cell sequencing maps gene expression to mutational phylogenies in PDGF and EGF driven gliomas

Abstract

Glioblastoma multiforme (GBM) is the most common and aggressive type of primary brain tumor. Epidermal growth-factor (EGF) and platelet-derived growth-factor (PDGF) receptors are frequently amplified and/or possess gain-of-function mutations in GBM. However, clinical trials of tyrosine-kinase inhibitors have shown disappointing efficacy, in part due to intra-tumor heterogeneity. To assess the effect of clonal heterogeneity on gene expression, we derived an approach to map single-cell expression profiles to sequentially-acquired mutations identified from exome sequencing. Using 288 single cells, we constructed high-resolution phylogenies of EGF-driven and PDGF-driven GBMs, modeling transcriptional kinetics during tumor evolution. Descending the phylogenetic tree of a *PDGF*-driven tumor corresponded to a progressive induction of an oligodendrocyte progenitor-like cell type, expressing pro-angiogenic factors. In contrast, phylogenetic analysis of an *EGFR*-amplified tumor showed an up-regulation of pro-invasive genes. An in-frame deletion in a specific dimerization domain of PDGF-receptor correlates with an up-regulation of growth pathways in a proneural GBM, and enhances proliferation when ectopically expressed in glioma cell lines. In-frame deletions in this domain are frequent in public GBM data.

Introduction

Glioblastoma multiforme (GBM) is an extremely aggressive type of brain tumor, characterized by a high degree of intra-tumor heterogeneity (Patel *et al*, 2014). Amplifications and gain-of-function mutations in receptor-tyrosine kinases (RTKs) are common in GBM. However, these mutations are typically regional and mosaic (Szerlip *et al*, 2012), and combinatorial application of RTK inhibitors is required to achieve a complete treatment *in vitro* (Stommel *et al*, 2007). Clinical trials of RTK inhibitors have shown only minimal efficacy, and these limitations may be in part due to intra-tumor heterogeneity (Prados *et al*, 2015). More broadly, developing treatments that circumvent specific, regional genotypic differences is challenging, since the number of biopsies per tumor is generally limited, and bulk-sequencing methods reduce such regional variation to population averages.

There is strong evidence that treatment itself can drive clonal evolution. Temozolomide chemotherapy is part of the current standard of care for newly diagnosed GBM. But, for some glioma patients, temozolomide treatment can also drive a hyper-mutated phenotype, as has been demonstrated by phylogenetic analysis of exome sequencing (exome-seq) data (Johnson *et al*, 2014). Phylogenetic analyses of bulk exome-seq and methylation array data in glioma cohorts have also been used to identify recurrent events in tumor evolution (Mazor *et al*, 2015; Johnson *et al*, 2014). Recent advances in single-cell sequencing have enabled fine mapping of *EGFR* variant heterogeneity (Francis *et al*, 2014), as well as studies of the evolutionary history of individual tumors at unprecedented resolution (Navin *et al*, 2011; Garvin *et al*, 2014). Furthermore, large-scale copy-number variations (CNVs) have been inferred from single-cell RNA sequencing (RNA-seq) in GBM (Patel *et al*, 2014). In this study, we called CNVs from bulk exome-seq and quantified them in single-cell RNA-seq from the same tumor sample. Based

on this, we produced a clonal ordering of individual cells that we used to infer transcriptional kinetics during tumor evolution and to perform inter-clone differential transcriptomics. We used this approach to contrast EGF-driven and PDGF-driven GBMs and identified pathways that show a dose-response to EGF and PDGF receptor copy-number.

Results

Primary GBMs contain heterogeneous mixtures of cell types with recurrent transcriptional signatures

We collected fresh tissue from 3 cases of primary, untreated GBM directly from the operating room (SF10282, SF10345, SF10360) and subjected these biopsies to both single-cell RNA-seq and bulk exome-seq (Fig 1A). We also performed bulk exome-seq on a separate blood sample from each patient. We characterized the landscape of genomic, somatic mutations for each patient using a robust exome-seq pipeline (Johnson *et al*, 2014), this analysis included identifying single-nucleotide variants (SNVs), small insertions/deletions (indels) and copy number variants (CNVs) (Materials and Methods).

All cases demonstrated an amplification of growth factor genes. We found *EGFR* to be highly amplified in SF10345 (122 copies), and *PDGFRA* to be amplified in SF10282 (12 copies). Deletion and putative loss-of-function mutations in tumor suppressors were also common events (Datasets EV 1-6). For example, all cases had non-synonymous point mutations in *PTEN* (with variant-allele frequencies (VAFs) from 41%-89%). A copy of chromosome 10 was lost in SF10345 and SF10360. Furthermore, these two cases harbored a deletion in chromosome 9, in a region encoding tumor suppressor genes *KLHL9* and *CDKN2A/B*. *KLHL9* deletions are correlated with the mesenchymal GBM subtype and poor prognosis (Chen *et al*, 2014). In our data, *KLHL9* is not expressed in either SF10345 or SF10360, and both samples classify as mesenchymal/classical. SF10360 and SF10282 share other mutations, such as a loss of 13q14 that contains the tumor suppressive micro-RNA cluster miR-15a/16 (Aqeilan *et al*, 2010; Afonso-Grunz & Müller, 2015). Between 5 to 35 small indels were detected per sample, e.g.

TP53 (SF10282, frame-shift deletion), *NF1* (SF10360, frame-shift deletion) and *PLAGL1* (SF10345, frame-shift deletion).

Prior to the analysis of single-cell RNA-seq libraries, low-complexity and low-coverage libraries were filtered (Fig EV 1A, B), and stromal/non-malignant cells were identified (Materials and Methods). This workflow left 61, 66 and 63 tumor cells from SF10282, SF10345, and SF10360 respectively. Consistent with previous reports (Patel *et al*, 2014), classification of single cells according to the Verhaak subtypes (Verhaak *et al*, 2010) identified heterogeneous mixtures of distinct subtypes within the same tumor (Fig 1C). SF10345 and SF10360 are classical/mesenchymal and predominantly EGFR driven. SF10282 is predominantly pro-neural, up-regulates PDGF-pathway genes, and markers of oligodendrocyte progenitor cells (OPCs) are broadly expressed. Yet SF10282 contains a subpopulation of cells with a neural stem cell (NSC) –like expression profile (Fig 1C). These cells classify as mesenchymal/classical in the Verhaak scheme. We sought to infer the relative ordering of the NSC and OPC-like cells in the tumor's phylogeny, and more generally to establish cellular phylogenies for all samples.

Phylogenies of copy-number alterations map gene expression to clonal structure

We chose to focus on large, somatic CNVs of 100 exons or more (Materials and Methods). The median size of CNVs exceeding this threshold was 18-21 mega base-pairs, comprising 300-400 genes. This size is much larger than the size of CNVs previously observed to occur frequently in the germline (Sudman*t et al*, 2015), which had a median size of 36 kilo base-pairs. We found that GBM to normal-brain control single-cell expression ratios correlated with CNV status (Appendix Fig S1), motivating us to quantify these CNVs in individual cells (Materials and Methods). Briefly, for each CNV identified in the exome-seq, the 5% significance level of the distribution

of normal brain read-counts covering that locus was used as a threshold to assign CNV presence/absence calls to individual cells (Fig 2A). This triage was unaffected by the application of a wavelet-smoothing filter to the single-cell data. Furthermore, exome-seq read histograms showed excellent agreement with single-cell trend-lines (Fig 2B,C). This indicated that our approach was robust to the stochastic expression of individual genes. We validated the error-rate of this classifier using 10-fold cross-validation, as well as empirical testing on a control dataset (Pollen *et al*, 2015a), (Appendix Fig S2, Materials and Methods). Using Jaccard distance between CNV genotypes to assess inter-cell similarity, we fit phylogenies to each of the tumor samples using the Fitch-Margoliash method (Materials and Methods). The amplifications of chromosomes 7 and 19p13.3, which were shared across cases in the exome-seq, occurred early in all 3 of our single-cell phylogenies. In SF10360, chromosome 7 gain was a founding event, occurring together with a loss of chromosome 10 (Fig 3A). Intriguingly, a loss of chromosome 13 arose independently in two distinct sub-clones of SF10360. Since 13q14 harbors the miR-15a/16 micro-RNA cluster, a known tumor-suppressor in prostate cancer (Bonci *et al*, 2008) and chronic lymphocytic leukemia (Pekarsky & Croce, 2014), this loss may convey a survival advantage here as well.

MiR-15a/16 mutant cells up-regulate downstream adhesion and Aurora B kinase pathway genes, enriched in the leading edge and infiltrating tumor

We compared gene expression in chromosome 13 wild-type cells to cells harboring the deletion in SF10360 (Dataset EV 7), and scanned the promoters of differentially expressed genes ($p<0.05$) for conserved transcription-factor recognition motifs (Materials and Methods). 16% of genes up-regulated upon chromosome 13 deletion were validated direct targets of miR-15a/16;

and an additional 78% were expressed from loci enriched for motifs of transcription factors targeted by miR-15a/16 (Fig 3B) (Chou *et al*, 2015). The most significant differentially-expressed genes (adjusted $p<0.1$) showed distinct patterning across tumor anatomical-structures, when cross-referenced with the Ivy Glioblastoma Atlas (glioblastoma.alleninstitute.org). Genes up-regulated in the wild-type cells were enriched in the peri-vascular region, and genes up-regulated upon chromosome 13 deletion were enriched in the leading edge and infiltrating tumor (Fig 3C, Fig EV 2). Consistent with an infiltrating phenotype, cell adhesion molecules were over-represented ($q=0.06$) (Materials and Methods). This included junction-adhesion molecules, integrins, disintegrins and cell-surface receptors implicated in invasion (Fig 3D) (Tenan *et al*, 2010; Reyes *et al*, 2013; Sloan, 2005; Venkatesh *et al*, 2015; Sarkar *et al*, 2015; Nath *et al*, 2000). 76% of these cell-adhesion pathway genes were enriched for NF- κB /REL recognition motifs (Fig 3B). Taken together with the EGFR-driven/mesenchymal classification of the case, we speculated that the loss of the miR-15a/16 cluster enhances growth-factor-stimulated cell invasion here, as has been described in other cancers (Bonci *et al*, 2008). Among the direct targets of miR-15a/16 that were up-regulated, Aurora B kinase, survivan and genes that complex with them were over-represented ($q=0.03$) (Fig 3E). Survivan is a well-studied inhibitor of apoptosis. Aurora B kinase overexpression increases genomic instability (Ota *et al*, 2002), resulting in multinuclearity and aneuploidy (Tatsuka *et al*, 1998).

Dose-response analysis correlates an in-frame deletion in *PDGFRA* to a pro-growth signature *in vivo*.

We found that the PDGF-receptor gene, *PDGFRA,* was amplified in SF10282 (12 copies as estimated by exome-seq). We also detected a small deletion in exon 7, that was broadly

expressed (Fig 4A). This mutant transcript, which we denote as *PDGFRAΔ*^{7}, is found in 98% of SF10282's cells that express *PDGFRA* (69% of cells overall). Since the deletion is in-frame, broadly expressed, and affects an immunoglobulin-like fold involved in receptor dimerization, we reasoned that *PDGFRAΔ*^{7} might enhance PDGF-receptor signaling. We sorted cells by *PDGFRAΔ*^{7} expression and identified genes that showed a strong rank-correlation with *PDGFRAΔ*^{7} levels (Materials and Methods). Positively correlated genes were enriched for the PDGF-receptor signaling network and cell cycle, when compared to the Pathway Commons and DAVID databases (Cerami *et al*, 2011; Huang *et al*, 2009). Negatively correlating genes were enriched for oxidative phosphorylation (Fig 4B). Genes correlating with increasing *PDGFRAΔ*^{7} were scanned for over-represented transcription factor binding motifs. Genetic and physical interaction databases were queried against significant transcription factors (Fig 4C), implicating STAT1 and NF-κB as downstream effectors of *PDGFRAΔ*^{7}. By comparison, an analogous dose-response analysis of *EGFR* in SF10345 identified an increasing gene-set of cell-cycle genes, as well as genes related to chromatin modification and cell motion. Inference of mediating transcription factors implicated STAT signaling, as in SF10282. Additionally, SOX2 (a pluripotency factor highly expressed in embryonic, neural and glioma stem cells (Suvà *et al*, 2014)) and c-Jun (an anti-apoptotic factor involved in glioma-genesis (Yoon *et al*, 2012)) targets correlated to *EGFR* dose (Fig 5).

PDGFRAΔ^7 confers a growth advantage and stimulates wild-type *PDGFRA* expression *in vitro*

We expressed *PDGFRAΔ*^{7}*,* wild-type *PDGFRA* and a GFP control from lentivirus, in two patient-derived cell lines that we had cultured as monolayers (Fig 6A). One we derived from

SF10360 (described here), and the second was from a primary GBM: SF10281. These cell-lines do not strongly express *PDGFRA* endogenously, but we detected robust expression of *PDGFRA* and *PDGFRAΔ*^{7} mRNA in the respective cultures where they were ectopically expressed (Fig 6B). To specifically quantify the expression of wild-type *PDGFRA,* we designed an RT-qPCR assay with a probe targeted to the deleted region. Intriguingly, we found that endogenous, wild-type *PDGFRA* was induced in both cell lines upon ectopic expression of *PDGFRAΔ*^{7} (Fig 6C). When we identified genes that were differentially, recurrently expressed in both *PDGFRAΔ*^{7} cultures compared to wild-type *PDGFRA* and GFP, we found that these genes enriched for gene-ontology molecular functions associated with PDGF binding and the binding of other growth factors (Fig 6D). In particular, we saw an up-regulation epiregulin mRNA (*EREG)*, an epidermal growth factor family member which ligates EGFR and most members of the v-erb-b2 oncogene homolog (ERBB*)* family. We saw an up-regulation of the notch-receptor ligand jagged 1 (*JAG1)* and the master-regulator of angiogenesis, *VEGFA*. Additionally, we saw an induction of regulators of inflammation *IL1B* and *IL6*, as well as *COX-2* and colony stimulating factor 3 (*CSF3). VEGFA*, *COX-2* and *CSF3* are all chemotactic factors for MDSC (Fujita *et al*, 2011; Lechner *et al*, 2010; Cao *et al*, 2014) (Fig 6E). Additionally, we performed an MTT colorimetric assay and found that ectopic *PDGFRAΔ*^{7} expression significantly enhanced cell growth *in vitro*, compared to over-expression of wild-type *PDGFRA* or GFP control (Fig 6F).

In-frame deletions in the PFGFRA dimerization domain are frequent events in The Cancer Genome Atlas's GBM data

We then processed exome-seq data from 389 GBM patients and corresponding blood controls available from The Cancer Genome Atlas (TCGA), and quantified the frequency of in-frame

deletions in *PDGFRA* (Materials and Methods). An in-frame deletion resulting in the loss of exons 8 and 9 (*PDGFRA*$\Delta^{8,9}$) had been previously cloned from a GBM biopsy (Kumabe *et al*, 1992). *PDGFRA*$\Delta^{8,9}$ affects the same dimerization domain as *PDGFRA*Δ^{7}, and has been shown to be transforming (Clarke & Dirks, 2003). A more recent TCGA study showed that *PDGFRA* mRNA lacking exons 8 and 9 was expressed in 17.8% of GBMs, however *PDGFRA*$\Delta^{8,9}$ prevalence was not interrogated at the DNA level (Brennan *et al*, 2014). We compared the distributions of reads mapping exons 8 and 9 in *PDGFRA* between tumor samples and the blood controls in TCGA data. The tumor distribution is clearly bimodal (Fig 7A), this second mode corresponds to a set of samples depleted of reads mapping exons 8 and 9. By thresholding at the 10% level of the blood distribution as a control, we estimate *PDGFRA*$\Delta^{8,9}$ occurs in 16% of cases in our dataset (n=389), after Benjamini-Hochberg correction for multiple hypothesis testing. This is remarkably close to the 17.8% of cases where *PDGFRA*$\Delta^{8,9}$-consistent mRNA was observed in the TCGA study (n=206). Additionally, we found a second family of deletions in exon 7, occurring in 1.8% of cases we analyzed (including SF10282). The frequency of *PDGFRA* amplification was 13.6%, but we did not find a strong correlation between *PDGFRA*$\Delta^{8,9}$ and *PDGFRA* amplification. On the other hand, all of the small deletions occurred in *PDGFRA* amplified cases (Fig 7B). Both *PDGFRA*Δ^{7}, *PDGFRA*$\Delta^{8,9}$, and the other small in-frame deletions target immunoglobulin I-set sub-domains of the extracellular domain of PDGFRA (Fig 7C). These domains are involved in receptor dimerization (Chen *et al*, 2012).

Accumulating mutations correlates with the acquisition of an OPC signature in a proneural GBM, and an invasion signature in a classical/mesenchymal case

We performed differential gene expression analysis between SF10282 and SF10345, and as expected, *EGFR* was up-regulated in SF10345 (Dataset EV 8). Additionally, there was an over-representation of cell-adhesion molecules and genes mediating motility in SF10345's differentially expressed genes (Fig 8A). For example *CD44*, encoding an adhesion molecule that mediates stem-cell homing (Pietras *et al*, 2014), was over 14-fold enriched in SF10345 (Fig 8B). Intriguingly, transcripts coding for chemotactic factors for myeloid-derived suppressor cells (MDSC) were differentially expressed in SF10345. C3 convertase is a core component of the complement cascade, mediating inflammation and the innate immune response. Complement pathway cytokines attract MDSC and induce their expression of reactive-oxygen species, contributing to a tumor-supportive microenvironment (Markiewski *et al*, 2008). Periostin (POSTN) is secreted by glioma stem cells, recruiting tumor-associated macrophages that enhance tumor growth (Zhou *et al*, 2015). Both *C3* and *POSTN* were up-regulated in SF10345 by several hundred fold (Fig 8B).

On the other hand, the neuron-differentiation pathway was significant in genes up-regulated in SF10282 (Fig 8A). Upon inspection however, these genes were factors predominantly expressed by OPCs during development: *PDGFRA*, *NKX2-2*, *SOX10*, *SEMA5A*, *LINGO1*, *S100B*, *MAP2* (Goldberg, 2004; Jepson *et al*, 2012; Deloulme *et al*, 2004; Shafit-Zagardo *et al*, 2000; Petryniak *et al*, 2007). Our initial inspection of SF10282 had revealed a NSC-like subpopulation (Fig 1C), which occurred early in our phylogeny of SF10282. Analysis of NSC and OPC gene expression in our phylogeny showed constitutively high expression of *PAX6*, *SOX2* and *TNC*, but a gradual increase in the OPC genes *OLIG2*, *ASCL1*, *NKX2-2* and

SOX10 along pseudo-time. PI3K/AKT pathway genes, and genes implicated in angiogenesis also increased concomitantly (Fig 8C). By comparison, SF10345 showed a progressive up-regulation of genes encoding extracellular matrix and transmembrane proteins associated with glioma motility and invasion, such as tenacin-C, neurocan and integrin (Cuddapah *et al*, 2014). AKT-pathway genes *AKT2* and *AKT3*, which contribute to glioma invasiveness and malignancy (Chautard *et al*, 2014), and class II myosins, required for glioma invasion and neural stem cell migration (Beadle *et al*, 2008; Ostrem *et al*, 2014), were also progressively up-regulated as one moves along the backbone of the SF10345 phylogeny (Fig 8D).

Discussion

Despite standard of care treatment, GBM has an extremely high recurrence rate, approximately 90% (Weller *et al*, 2013). There is an urgent need for combinatorial strategies to address residual disease (Prados *et al*, 2015). In particular, intra-tumor receptor heterogeneity is a confounder for tyrosine-kinase inhibitor therapy (De Witt Hamer, 2010). Recent advances in single-cell analysis have enabled high-resolution estimates of clonal heterogeneity (Meyer *et al*, 2015; Francis *et al*, 2014; Patel *et al*, 2014). In this study, we combined whole exome and single-cell mRNA sequencing to map transcriptional signatures to mutational phylogenies in EGF and PDGF-driven GBMs. These data implicate in-frame deletions in the dimerization domain of PDGFRA as potential therapeutic targets. And, they identify a cell-type hierarchy, which occurs in early brain development, as being recapitulated during tumor evolution.

In the developing forebrain, OPCs arise from neuroepithelial stem cells in sequential waves of oligodendrocyte production (Kessaris *et al*, 2006; Menn *et al*, 2006). ASCL1 and OLIG2 are interacting transcription factors required for oligodendrogenesis and highly expressed in OPCs (Nakatani *et al*, 2013; Zhou & Anderson, 2002; Petryniak *et al*, 2007). OLIG2 drives SOX10 transcription in OPCs (Kuspert *et al*, 2011), which in turn regulates myelin expression (Stolt *et al*, 2002), critical for oligodendrocyte function. Consistent with aberrant activation of this developmental program in GBM, SF10282 progressively up-regulated the above OPC-specific genes in its model of tumor evolution. Early cells expressed *PAX6*, *SOX2* and other markers of neural stem cells but produced daughters with a more OPC-like profile (Fig 8C). *PDGFRA* expression is characteristic of OPCs in non-malignant brain, and was amplified in SF10282.

However, 68% of tumor cells in SF10282 express an in-frame deletion mutant, *PDGFRAΔ*^{7}. We found that *PDGFRAΔ*^{7} is part of a family of in-frame deletions in *PDGFRA*'s dimerization domain that occur in GBM. The most common of these mutations, *PDGFRAΔ*^{8,9}, occurs with 16% frequency in the TCGA DNA sequencing data we analyzed (n=389). Our finding is consistent with a recent TCGA paper that found *PDGFRA* mRNA lacking exons 8 and 9 in 17.8% of their samples (n=206) (Brennan *et al*, 2014). *PDGFRAΔ*^{8,9} expression has been shown to induce constitutive signaling of the PDGF receptor, and is sufficient for malignant transformation (Clarke & Dirks, 2003). In-frame deletions in *PDGFRA*'s dimerization domain have also been observed in pediatric high grade gliomas, those studies implicate constitutive receptor signaling too (Paugh *et al*, 2013). Our single-cell data show that increasing *PDGFRAΔ*^{7} dosage correlates *in vivo* with an activation of genes downstream of the PDGF-receptor, in particular, targets of rapid-acting transcription factors (e.g. STAT, NF-κB). When ectopically over-expressed, in patient-derived glioma cell lines, *PDGFRAΔ*^{7} enhances proliferation compared to wild-type *PDGFRA* over-expression. *PDGFRAΔ*^{7} induces wild-type *PDGFRA* expression *in vitro*, along with tyrosine-kinase receptor ligands and pro-angiogenesis genes. The deletions we observed aggregate in the I-set domains of *PDGFRA*'s immunoglobulin-like folds. These highly conserved domains are involved with receptor dimerization, and in principle can interact with a variety of different domain types; however, these domains are typically glycosylated which protects against promiscuous receptor interaction in the absence of ligand (Barclay, 2003). All of the deletions we have observed are either on or near predicted glycosylation sites (Fig EV3).

Single-cell RNA-seq has enabled composition assessments and lineage reconstruction in the highly dynamic, highly heterogeneous context of the developing brain (Pollen *et al*, 2015b;

Darmanis *et al*, 2015; Diaz *et al*, 2016; Liu *et al*, 2016). Single-cell RNA-seq of tumor biopsies provide a similar snapshot of tumor evolution. By using a proven approach such as bulk exome-seq to identify mutations, single-cell RNA-seq libraries can then be triaged to assign a transcriptional signature to a mutational profile or phylogeny. By cross-referencing public atlases such as the Ivy Glioblastoma Atlas we further related this information to tumor anatomical structure. When we cross-referenced the Ivy Glioblastoma Atlas with the chromosome 13 deletion subclone of SF10360, we identified a spatial segregation of differentially-expressed genes, aggregating in the perivascular niche and the leading edge respectively. Cells harboring the chromosome 13 deletion were also characterized by an up-regulation of pro-invasion, adhesion genes. However, whether the deletion event preceded or succeeded any physical separation of these two clones could not be inferred from these data. Targeted resections that record relative biopsy position can resolve spatial evolution more accurately (Sottoriva *et al*, 2013).

Experimental Procedures

Sample Acquisition and processing

Fresh tumor tissue was acquired from patients undergoing resection for glioblastoma. De-identified samples were obtained from the Neurosurgery Tissue Bank at the University of California San Francisco (UCSF). Sample use was approved by the Committee on Human Research at UCSF. The experiments conformed to the principles set out in the WMA Declaration of Helsinki and the Department of Health and Human Services Belmont Report. All patients provided informed written consent. Tissues were minced in collection media (Leibovitz's L-15 medium, 4 mg/mL glucose, 100 u/mL Penicillin, 100 ug/mL Streptomycin) with a scalpel. Samples were further dissociated in a mixture of papain (Worthington Biochem. Corp) and 2000 units/mL of DNase I freshly diluted in EBSS and incubated at 37° C for 30 minutes. The suspension was centrifuged for 5 min at 300g and re-suspended in PBS. Suspensions were triturated by pipetting up and down 10 times and passed through a 70μm strainer cap (BD Falcon), followed by centrifugation for 5 min at 300g. Pellets were re-suspended in PBS and passed through a 40μm strainer cap (BD Falcon), followed by centrifugation for 5 min at 300g. Dissociated, single cells were then re-suspended in GNS (Neurocult NS-A (Stem Cell Tech.), 2mM L-Glutamine, 100 U/mL Penicillin, 100 ug/mL Streptomycin, N2/B27 supplement (Invitrogen), sodium pyruvate).

Single cell capture and cDNA generation was performed using the Fluidigm C1 Single-Cell Integrated Fluidic Circuit (IFC) and SMARTer Ultra Low RNA Kit. cDNA was quantified using Agilent High Sensitivity DNA Kits and diluted to 0.15–0.30 ng/μL. Dual indexing and amplification were performed using the Nextera XT DNA Library Prep Kit (Illumina) according to the Fluidigm C1 protocol. 96 single cell RNA-seq libraries were generated from each tumor

sample and were pooled for 96-plex sequencing. Amplified and pooled cDNA was purified and size selected twice using 0.9X volume of Agencourt AMPure XP beads (Beckman Coulter). Final cDNA libraries were quantified using High Sensitivity DNA Kits (Agilent) and sequenced on a HiSeq 2500 (Illumina), using the paired-end 100 base pair (bp) protocol.

Exome-Sequencing and genomic mutation identification

Exome capture was done using NimbleGen SeqCap EZ Human Exome Kit v3.0 (Roche) exome capture kits on a tumor sample and a blood control sample from each patient. Sequencing was carried out with an Illumina-HiSeq 2500 machine acquiring 100 bp paired-end reads. Reads were aligned to the human genome (hg19) using BWA (Li & Durbin, 2009), whereas only uniquely matched paired reads were retained. PicardTools (http://broadinstitute.github.io/picard/) and the GATK toolkit (McKenna *et al*, 2010) were used for quality score re-calibration, duplicate-removal and re-alignment around indels. The resulting BAM files were sorted by genomic coordinates. Subsequently, the percent contamination was assessed with ContEst (SF10345:0.1%, SF10360:0.1%, SF10282:0.2%). OxoG metrics were calculated with PicardTools' CollectOxoGMetrics (Dataset EV9). After these control steps, single nucleotide variants (SNVs), short indels (<50 bps) and large CNVs comprising more than three exons were detected. Somatic SNVs were inferred with MuTect (https://www.broadinstitute.org/cancer/cga/mutect) for each tumor/control pair and annotated with. SNVs with less than a 10% variant frequency in the tumor, with more than 5 variant reads in the patient matched normal, or greater than a 10% variant frequency in the patient-matched normal were excluded from further analysis. Small indels were detected with Pindel (Ye *et al*, 2009) and those with fewer than 6 supporting reads in the tumor, any supporting reads or less

than 14 total reads in the patient-matched normal, and replacements for which the deletion and non-template inserted sequence were of the same length were excluded. All indels and SNVs were annotated for their mutational context and effect using the Annovar software package (Wang *et al*, 2010). Only protein-coding or splice-site mutations were retained for further analysis. ADTex (Amarasinghe *et al*, 2014) was used for detection of large somatic CNVs. Only CNVs comprising more than 100 Exons were retained for downstream analysis. Proximal (<1 Mbp) somatic CNVs were merged in the output file to maximize CNV regions.

Single cell RNA-sequencing data pre-processing, quality control and GBM subtype classification

Reads were trimmed for quality and Nextera adapters removed with TrimGalore! (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), paired-end reads were mapped to the human genome (hg19) with tophat2 (Kim *et al*, 2013) using the –prefilter-multihits option and a GENCODE V19 transcriptome-guided alignment. Quantification of GENCODE genes was carried out with featureCounts (Liao *et al*, 2014). Only fragments corresponding to uniquely mapped, correctly paired reads were kept. Expression values were normalized to CPM in each cell. We filtered samples whose background fraction is significantly high, via a threshold on the (Benjamini-Hochberg corrected) q-value of a Lorenz statistic on the samples' cumulative densities, described previously (Diaz *et al*, 2016). In our tests, samples that have a small q-value have low complexity, as measured by Gini-Simpson index, and they have low coverage, as estimated by the Good-Turing statistic (Fig EV1).

To compare genotypes from the single-cell RNA-seq data across patients and to the exome-seq data, we identified 17, 21 and 15 single-nucleotide variants (SNVs) in the single-cell

RNA-seq that are patient specific in SF10282, SF10345 and SF10360 respectively. These SNVs are likewise detected in, and only in, their respective blood and tumor-derived exome-seq datasets. The median difference in RNA-seq to tumor exome-seq variant allele frequency (VAF) is 0.056. By comparison, the median difference in tumor exome-seq to blood exome-seq VAF is 0.044 (see Appendix Fig S3 and Table S1).

Infiltrating stromal/non-malignant cells were identified as those cells which contained none of the CNVs, none of the indels or point-mutations found in the exome-seq data, and clustered away from tumor cells in a hierarchical clustering. These putative stromal cells formed two clusters, one that is comprised of cells from all samples that expresses markers of immune cells (particularly those of macrophages/microglia) and a second cluster comprised of cells from SF10282 that express oligodendrocyte markers (Fig EV 2C). We classified all of the remaining cells according to the Verhaak (Verhaak *et al*, 2010) and Sun (Sun *et al*, 2014) molecular subtypes. To perform the Verhaak classification in individual cells we fit a linear regression model using the 4 centroids in the original Verhaak clustering as predictors, and the gene expression profile of the tumor cell to be classified as a response. We restricted expression profiles in individual cells to those genes used in the original Verhaak clustering, and represented expression by standardized, log-transformed CPMs. The cell to be classified is assigned to the subtype whose corresponding centroid has the largest regression coefficient in magnitude. Sun *et al.* determined their subtypes by identifying gene modules that co-expressed with *PDGFRA* or *EGFR* in a large cohort of adult diffuse gliomas. In each cell to be classified we averaged gene expression (measured by log-CPM) across both of these two gene modules. If either module's average was more than 2-fold higher than the other, then we assigned the cell to the corresponding subtype.

Single-cell CNV presence/absence calls

Based on the premise that copy-number changes are reflected in RNA-seq read-counts when averaged over large, adjacent genomic regions (Patel *et al*, 2014), we examined loci that were called for somatic CNV in our ADTex pipeline. For each CNV candidate region (CNVCR), we sum the library-size normalized read-counts across genes in that region, for each cell in our tumor sample. We do the same for each cell in a non-malignant, human brain control (Darmanis *et al*, 2015). We use the distribution across cells in the control for each CNVCR, to assess the sum in a given tumor cell. We use the 5% significance level of the control distribution as our threshold for making a CNV call (Fig 2A), and control for multiple-hypothesis testing using Benjamini-Hochberg correction. This results in a genotype assignment to each cell, determined by which CNVs called in the exome-seq data are present in that cell. To estimate the false discovery rate (FDR) of this classification procedure, we performed 10-fold cross-validation using the normal-brain control cells. For each patient, we randomly selected tranches of 10% as test and 90% as training data. We estimated the FDR as (# positive CNV calls)/(# total CNV calls), for each of the 10 folds. We found the FDR to be < 0.01 for all tests (Appendix Fig S2A). As a second estimator of the FDR, we classified the presence of CNVs on a dataset comprised of nonmalignant, fetal-brain cells (Pollen *et al*, 2015b), and estimated the FDR as above. We found these FDR estimates to all be < 0.06 (Appendix Fig S2B).

Phylogenetic trees

We measure pairwise distance between individual cells using Jaccard distance between CNV genotypes. This measures the number of shared CNV calls, as a fraction of the number of unique calls in either cell. To obtain a phylogenetic tree of tumor cells based on this distance metric, we

use the Fitch-Margoliash method (Fitch & Margoliash, 1967) as implemented in the phylip R package (Revell & Chamberlain, 2014), adding a "normal" genotype with no CNVs as an out-group to root the tree. We identified 5-6 cells per sample that did not harbor the CNV with the highest frequency (chromosome 7 gain in SF10345, chromosome 4q12 gain in SF10282 and chromosome 10 loss in SF10360). Since these mutations are nearly ubiquitous, they represent founding mutations for the dominant clones sampled in our cells. This is consistent with the fact that they affect cancer driver genes *EGFR*, *PDGFRA* and *PTEN* respectively. These rare cells lacking founding mutations may be technical outliers, or members of a lineage under-sampled in the biopsy.

PDGFR *in vitro* overexpression

Primary dissociated tumors were plated for cell culture in DMEM-F12 with 10% FBS. Expression vectors driving wild-type *PDGFR*, *PDGFRΔ*[7], or GFP (pLV[Exp]-EGFP/Bsd-EF1A) were generated and packaged into third generation lentivirus particles (VectorBuilder, Cyagen Biosciences). Triplicate cell cultures were infected with lentiviruses at equal titers at a MOI of ~1.0 and with 0.5 ug/mL polybrene. 2 days following infection, cells were selected with 33.3 ug/mL blasticidine for 3 days. For RNA collection, cells were grown for 1 additional day with no drug selection and then harvested. For proliferation assays, cells were grown continually in the presence of 33.3 ug/mL blasticidine. Vector expression was confirmed using fluorescence microscopy and RNA-seq.

Bulk RNA-seq sample processing

RNA was harvested using TRIzol reagent, followed by Direct-zol MiniPrep RNA purification kits (Zymo Research) with the on-column DNAse digestion step. RNA integrity was confirmed using the Agilent 2200 RNA ScreenTape (Agilent Technologies). RNA-seq libraries were generated using TruSeq Stranded mRNA kit according to manufacturer's protocol (illumina). cDNA was validated using the Agilent 2200 DNA 1000 ScreenTape, Qubit 2.0 Fluorometer (Life Technologies), and ddPCR (Bio-Rad). Cluster generation and sequencing was performed on a HiSeq 4000, using the single end 50 read protocol.

Reads were mapped to the human genome (hg19) with tophat2 (Kim et al, 2013). Quantification of GENCODE V19 genes was carried out with featureCounts (Liao et al, 2014) using only uniquely mapped reads. Differential expression analyses were performed via DESeq2, using the likelihood ratio test applied against the wild-type *PDGFR*, *PDGFRΔ*^{7}, and GFP triplicate samples as a 3 level factor.

Dose response analysis

In SF10282, cells were sorted by *PDGFRAΔ*^{7} expression, in SF10345 cells were sorted by EGFR expression. Genes which had a Spearman's rank correlation in the top 5% with *PDGFRAΔ*^{7} were considered for further analysis. Pathway analysis was done using Pathway Commons (Cera*mi et al*, 2011) via WEBGESTALT (Wang *et al*, 2013) and DAVID (Huang *et al*, 2009) transcription-factor motif enrichment analysis was done via OPOSSOM (Sui Ho *et al*, 2007). Network interactions were computed via GeneMANIA (Montojo *et al*, 2010).

In vitro proliferation and wild-type *PDGFRA* quantification assays

For the WST-1 proliferation assay, 1×10^4 cells were cultured in a 96 well plate for 24 hours in 100 µl of complete media. Then, 10 µl of WST-1 reagent (Roche) was added to each well. Cells were incubated at 37°C, 5% CO2 for 4 hours, and placed on a shaker for 1 min. The plates were then read on a microplate reader with a wavelength of 420 nm and a reference at 620 nm. For the Taqman gene expression assay, cDNA was synthesized from 50 ng of glioma cell line-derived RNA with qScript™ XLT cDNA SuperMix (Quanta Biosciences, Gaithersburg, MD). 2 µl of converted cDNA was then used in the qRT-PCR reaction with PerfeCta® FastMix® II (Quanta Biosciences), according to the manufacturer's protocol. A custom Taqman assay specific to wild-type, but not mutant, *PDGFRA* was designed by Life Technologies. The fluorescent probe was targeted to the region deleted in $PDGFRA\Delta^7$, but present in wild-type PDGFRA. Real-time PCR and data analysis were performed using the StepOnePlus Real-Time PCR system (Life Technologies, Carlsbad, CA). GAPDH (Assay ID: Hs02758991_g1) expression was used as the housekeeping gene and relative expression was determined using the $2^{\Delta\Delta CT}$ method.

Differential expression and time-series analysis

To identify genes differentially expressed between SF10345 and SF10282, and assess inter-cellular heterogeneity, we used the scde R package (Kharchenko *et al*, 2014). We used DESeq to compare chromosome 13 loss to wild-type cells in SF10360 and treated each cell as a replicate. Genes that were expressed in more than 80% of cells at < 1 CPM were filtered prior to each analysis. Transcription-factor motif enrichment analysis was done via OPOSSOM (Sui Ho *et al*, 2007). JEPETTO (Winterhalter *et al*, 2014), run via Cytoscape (Shannon *et al*, 2003), was used to compute pathways having significant overlap with genes up-regulated upon chromosome 13

loss, and their protein interactions. For the time-series analysis we chose paths in our phylogenetic trees corresponding to the most frequent mutation at each level. Our goal was to identify an ordering of the dominant clones in the sample. But, in principle expression along an arbitrary path can be measured. We grouped cells with identical copy number profiles into three intermediate points along each branch: early, mid and late. We subjected monotonically increasing genes to gene-ontology analysis via DAVID (Huang *et al*, 2009).

TCGA data analysis

Alignments for all GBM exome-seq with available paired blood controls from TCGA (http://cancergenome.nih.gov/) were retrieved from CGHub (https://cghub.ucsc.edu/) in BAM format**.** Reads in FASTQ format were extracted with bedtools from these alignments (Quinlan & Hall, 2010). We detected the quality encoding with a custom perl script and clipped adapters/low quality based with Trimmomatic (Bolger *et al*, 2014). Next, we mapped reads to the human genome (hg19) with HISAT2 (Kim *et al*, 2015). Small Indels were detected with pindel (Ye *et al*, 2009) for each tumor/control pair. If multiple sequencing libraries existed for one patient, the most recently published library was used. For the detection of loss of exons 8/9 we calculated the fraction of reads mapping to these exons from all reads aligned to the PDGFRA in each GBM and each blood sample. We used the distribution of all blood controls to assign significance values to each GBM sample. P-values were corrected for multiple testing with Benjamini-Hochberg (Benjamini & Hochberg, 1995) and an adjusted p-value of <0.25 was considered significant, which is in agreement with the significance threshold for CNVs used by the TCGA consortium (Brennan *et al*, 2013).

Data availability and algorithm parameters

The RNA-seq and Exome-Seq data from this publication has been deposited at the European Genome-phenome Archive (EGA, http://www.ebi.ac.uk/ega/) which is hosted at the EBI, under accession number EGAS00001001900. Parameters of all algorithms are available in the Appendix Supplementary Methods.

Acknowledgments

We would like to thank Joanna Phillips and Anny Shai of the UCSF Neurosurgery Tissue Core who facilitated tissue acquisition, and Pamela Paris (UCSF) for consultation on genomic arrays. This work has been supported by a Shurl and Kay Curci Foundation Research Grant, a UCSF Brain Tumor SPORE Career Development Award (P50-CA097257-13:7017), and a gift from the Dabbiere Family to A.D., a Damon Runyon Cancer Research Foundation postdoctoral fellowship (DRG-2166-13) to A.P., a VA award 5I01 BX000252 to D.A.L., and by NIH awards U01 MH105989 and R01NS075998 to A.R.K.

Figures

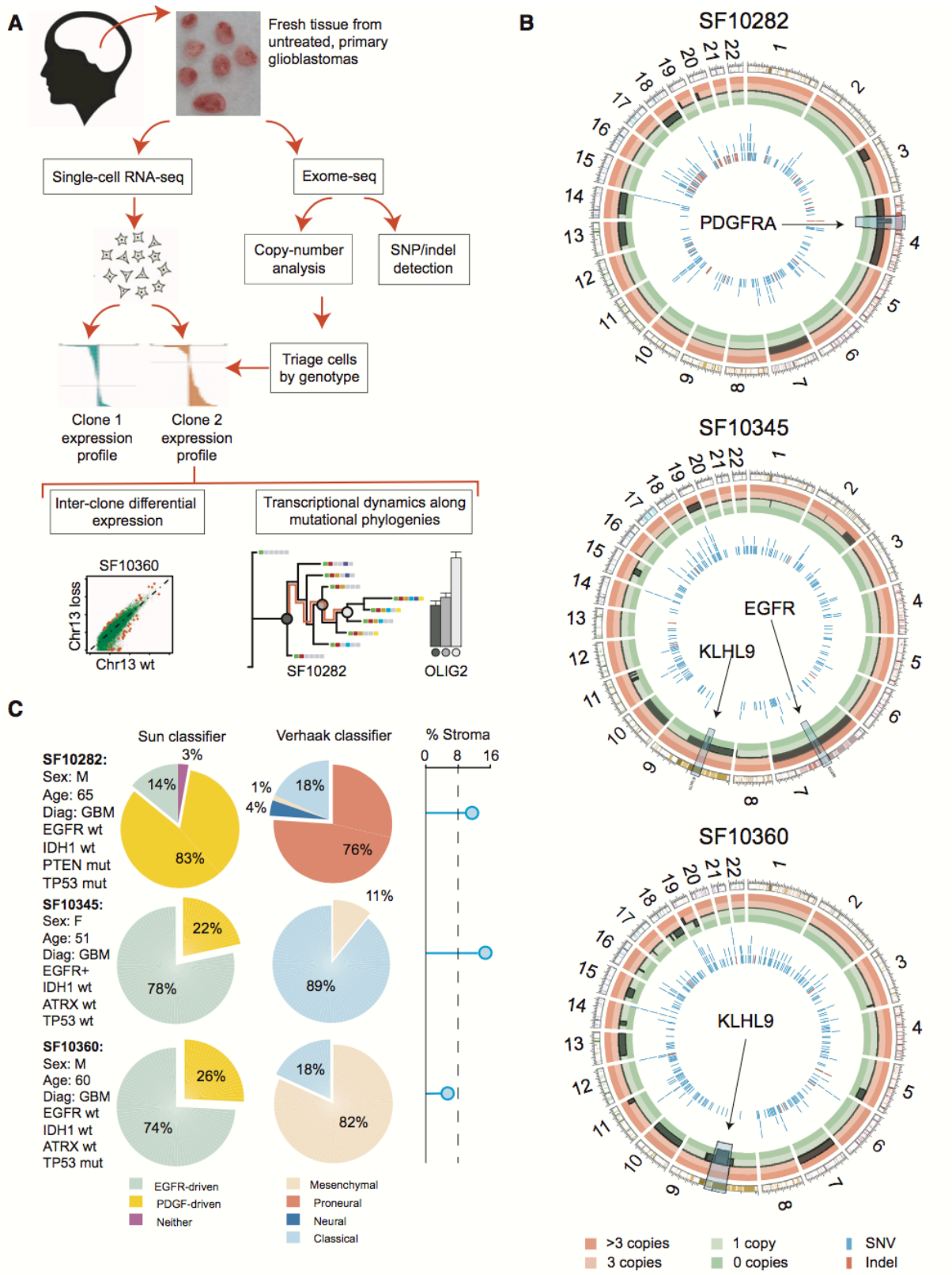

A
Fresh tissue from untreated, primary glioblastomas
Single-cell RNA-seq
Exome-seq
Copy-number analysis
SNP/indel detection
Triage cells by genotype
Clone 1 expression profile
Clone 2 expression profile
Inter-clone differential expression
Transcriptional dynamics along mutational phylogenies
SF10360
Chr13 loss
Chr13 wt
SF10282
OLIG2
B
SF10282
PDGFRA
SF10345
EGFR
KLHL9
SF10360
KLHL9
>3 copies
3 copies
1 copy
0 copies
SNV
Indel
C
Sun classifier
Verhaak classifier
% Stroma
0 8 16
SF10282:
Sex: M
Age: 65
Diag: GBM
EGFR wt
IDH1 wt
PTEN mut
TP53 mut
3%
14%
83%
1%
4%
18%
76%
SF10345:
Sex: F
Age: 51
Diag: GBM
EGFR+
IDH1 wt
ATRX wt
TP53 wt
22%
78%
11%
89%
SF10360:
Sex: M
Age: 60
Diag: GBM
EGFR wt
IDH1 wt
ATRX wt
TP53 mut
26%
74%
18%
82%
EGFR-driven
PDGF-driven
Neither
Mesenchymal
Proneural
Neural
Classical

Figure 1. Experimental design and pipeline, pathology, and genomic and transcriptomic signatures for three primary GBMs.

A) The sample acquisition and processing pipeline. B) Circos plot of somatic, genomic alterations detected in the bulk DNA of each patient using ADTEx. Copy-number alterations are highlighted in the outer circle by thick black bars, and SNVs (MuTect) and small indels (Pindel) by the vertical lines in the inner circle. Regions with strong amplifications/losses are highlighted. C) Summaries of the patient's sex, age, pathology report, sample's stromal infiltration, and molecular classification.

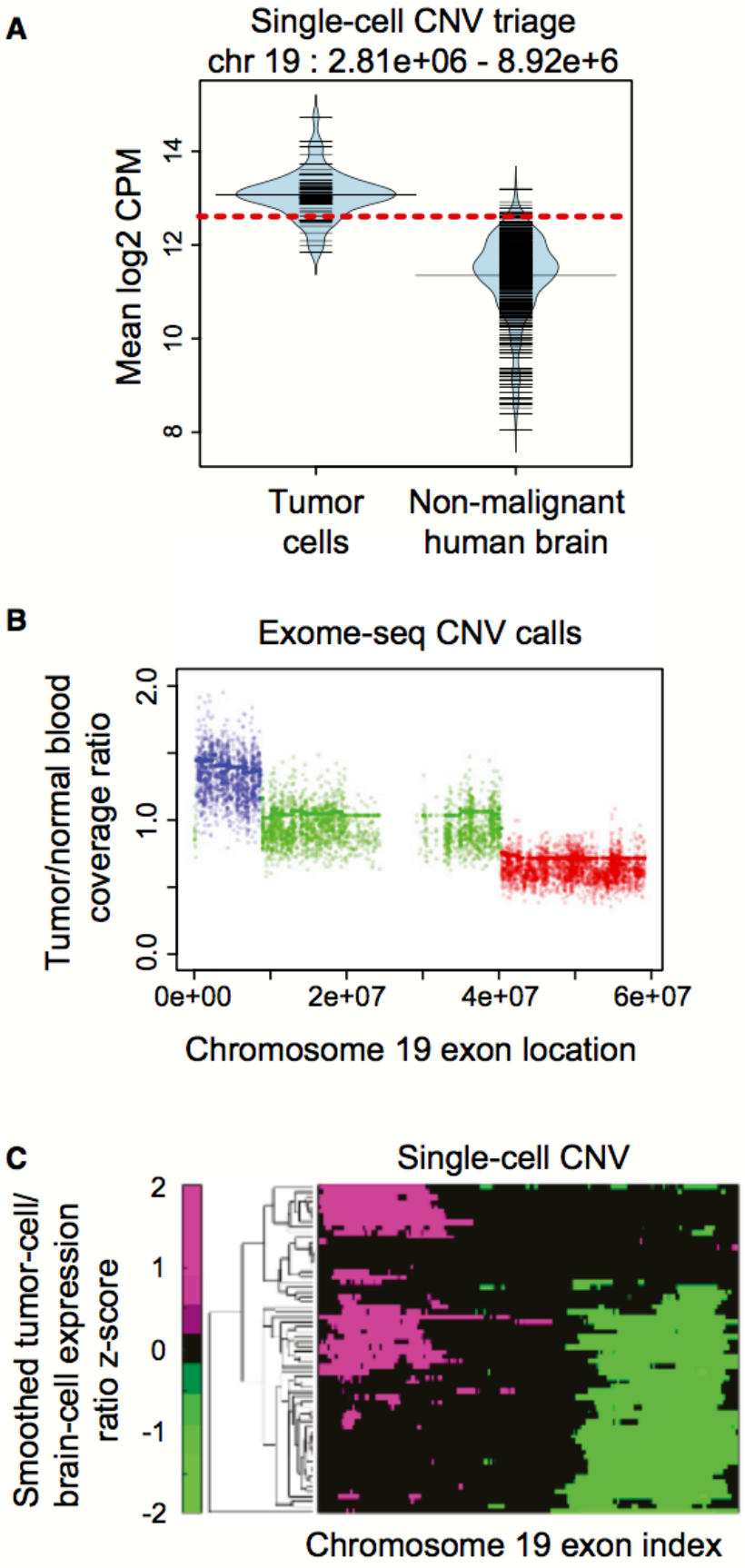

A
Single-cell CNV triage
chr 19 : 2.81e+06 - 8.92e+6
Mean log2 CPM
8
10
12
14
Tumor cells
Non-malignant human brain
B
Exome-seq CNV calls
Tumor/normal blood coverage ratio
0.0
1.0
2.0
0e+00
2e+07
4e+07
6e+07
Chromosome 19 exon location
C
Single-cell CNV
Smoothed tumor-cell/ brain-cell expression ratio z-score
2
1
0
-1
-2
Chromosome 19 exon index

Figure 2. Presence/absence assignments in individual cells, of CNVs called from bulk exome-seq

A) Read-count distributions in a locus of copy-number gain on 19p13.3, comparing cells in SF10360 to a human, normal-brain control. The 5^{th} percentile of the normal-cell distribution is indicated by a red line. B) Per-exon, normalized bulk exome-seq read-count computed by ADTEx, compared between SF10360 and a blood control. Regions with a read-count ratio > 1.3 are putative DNA copy-number gains (blue), and regions with fold-changes < 0.7 are putative losses (red). C) A heatmap visualizing the z-score of ratios, of wavelet-smoothed read counts, compared between single-cell RNA-seq expression in tumor cell (rows) and the median expression in single-cell RNA-seq of normal-brain tissue.

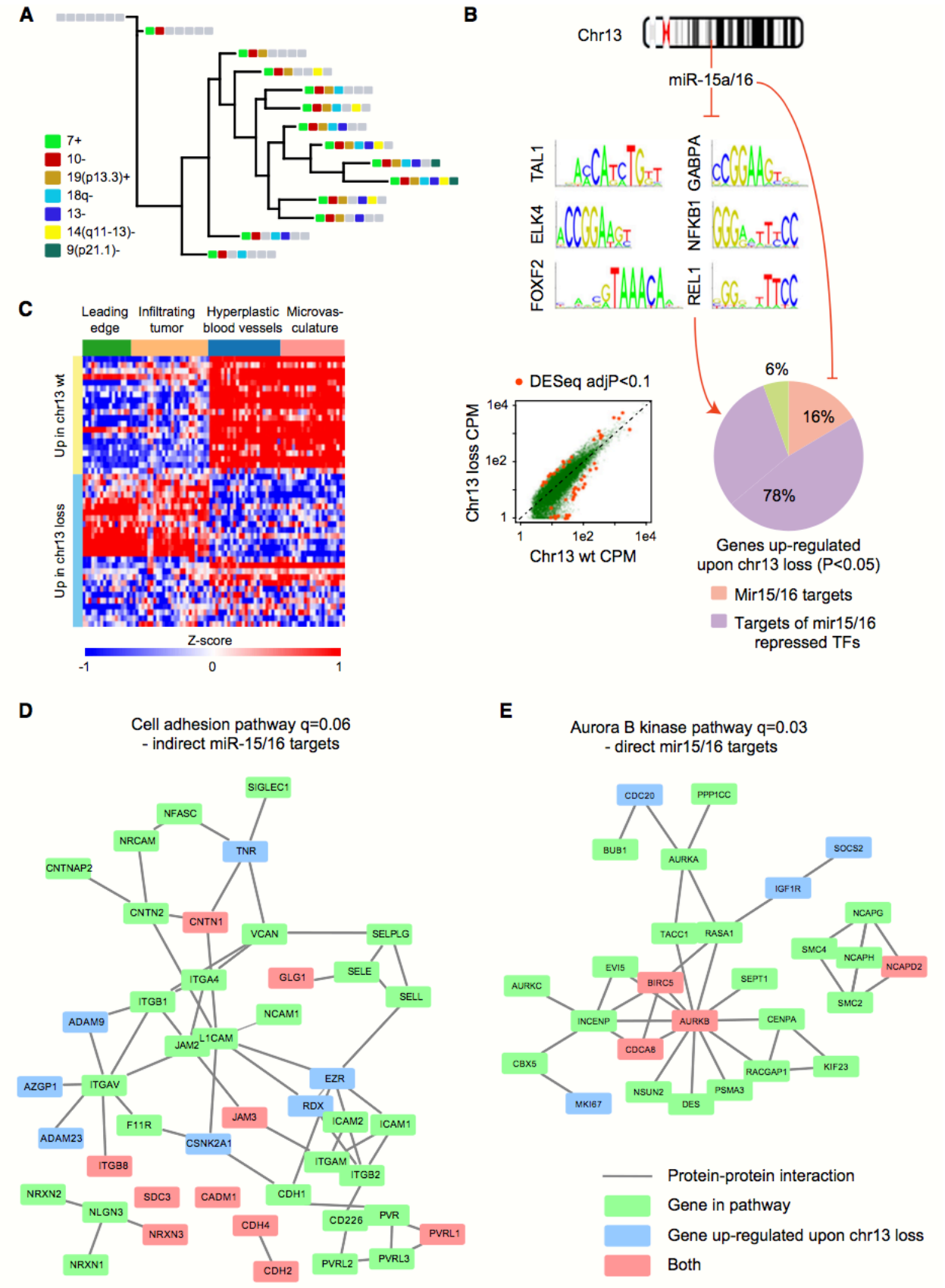

A
7+
10-
19(p13.3)+
18q-
13-
14(q11-13)-
9(p21.1)-
B
Chr13
miR-15a/16
TAL1
GABPA
ELK4
NFKB1
FOXF2
REL1
DESeq adjP<0.1
Chr13 loss CPM
Chr13 wt CPM
1e4
1e2
1
6%
16%
78%
Genes up-regulated upon chr13 loss (P<0.05)
Mir15/16 targets
Targets of mir15/16 repressed TFs
C
Leading edge
Infiltrating tumor
Hyperplastic blood vessels
Microvas-culature
Up in chr13 wt
Up in chr13 loss
Z-score
-1
0
1
D
Cell adhesion pathway q=0.06
- indirect miR-15/16 targets
SIGLEC1
NFASC
NRCAM
TNR
CNTNAP2
CNTN2
CNTN1
VCAN
SELPLG
GLG1
SELE
ITGA4
SELL
ITGB1
ADAM9
NCAM1
JAM2
L1CAM
AZGP1
ITGAV
EZR
RDX
JAM3
ICAM2
ICAM1
F11R
ADAM23
CSNK2A1
ITGB8
ITGAM
ITGB2
NRXN2
SDC3
CADM1
CDH1
NLGN3
CDH4
CD226
PVR
PVRL1
NRXN3
PVRL2
PVRL3
NRXN1
CDH2
E
Aurora B kinase pathway q=0.03
- direct mir15/16 targets
CDC20
PPP1CC
BUB1
AURKA
SOCS2
IGF1R
NCAPG
TACC1
RASA1
SMC4
NCAPH
NCAPD2
EVI5
BIRC5
SEPT1
AURKC
SMC2
INCENP
AURKB
CENPA
CDCA8
CBX5
KIF23
RACGAP1
NSUN2
PSMA3
MKI67
DES
Protein-protein interaction
Gene in pathway
Gene up-regulated upon chr13 loss
Both

Figure 3 Analysis of gene-sets, differentially expressed between subclones of SF10360

A) A phylogeny of CNV cellular genotypes identified in SF10360. Each leaf corresponds to a genotype, defined by a set of CNV presence/absence calls shared between a group of cells. B) Overrepresented transcription factor recognition motifs in genes up-regulated (DESeq, $P < 0.05$) upon chromosome 13 loss. 16% of up-regulated genes are direct, validated miR-15a/16 targets. Another 78% of up-regulated genes are targets of transcription factors that are repressed by miR-15a/16. C) Heatmap of the most significantly, differentially expressed genes (DESeq, FDR < 0.1) upon chromosome 13 loss in the Ivy Glioblastoma Atlas. Each row is a gene, each column is an RNA-seq from an anatomically defined tumor compartment, micro-dissected from an untreated GBM biopsy. D) Protein interaction network of genes differentially expressed upon chromosome 13 loss, which are targets of a miR-15a/16 regulated transcription factor, with an overlay of protein interactions in the cell-adhesion pathway. E) Protein interaction network of genes differentially expressed upon chromosome 13 loss, which are direct, validated targets of miR-15a/16, with an overlay of protein interactions in the Aurora B kinase. Data information: Significance of network overlaps in (D, E) was computed via JEPETTO ($q < 0.1$).

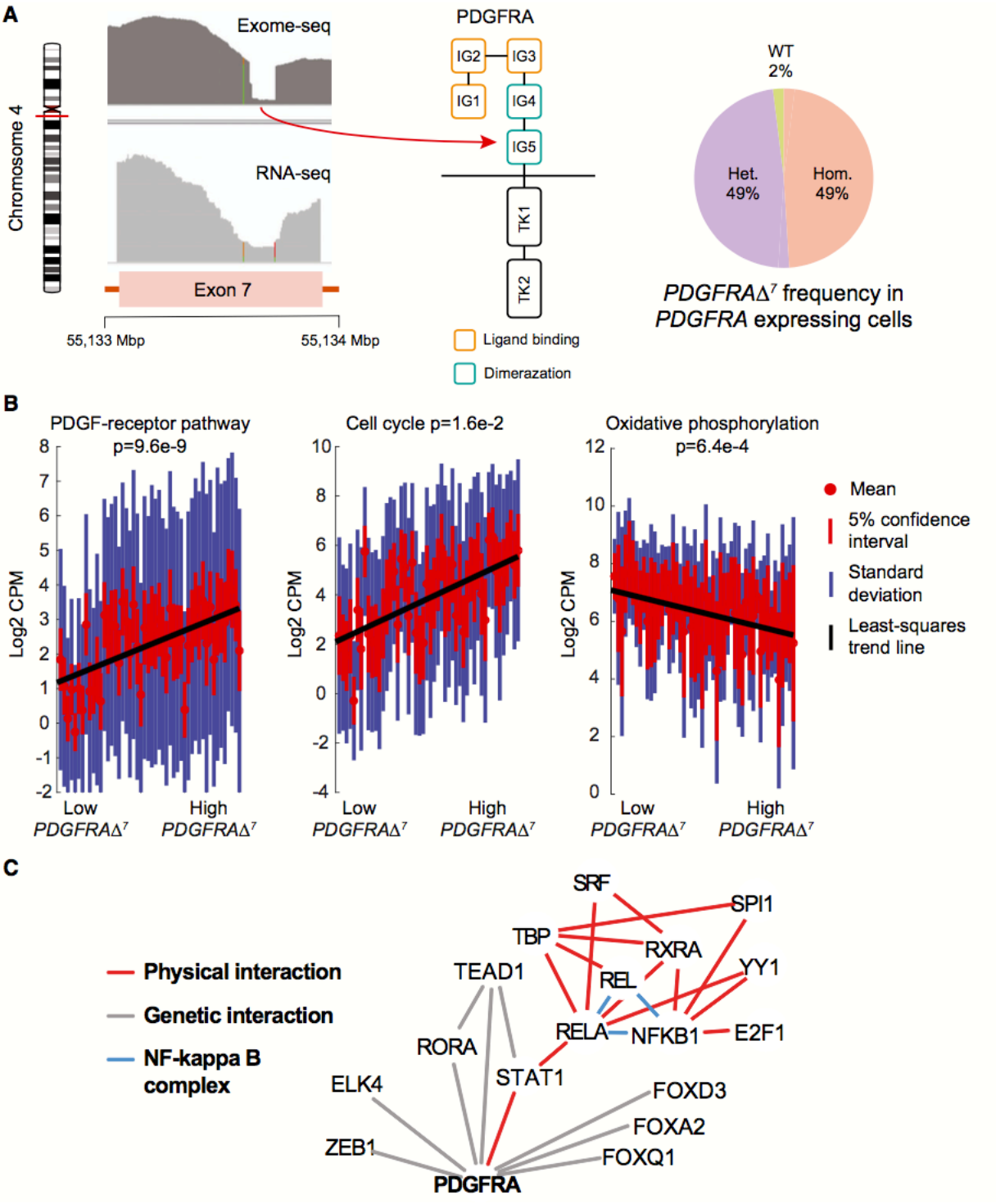

A
Chromosome 4
Exome-seq
RNA-seq
Exon 7
55,133 Mbp
55,134 Mbp
PDGFRA
IG2
IG3
IG1
IG4
IG5
TK1
TK2
Ligand binding
Dimerazation
WT
2%
Het.
49%
Hom.
49%
PDGFRAΔ7 frequency in
PDGFRA expressing cells
B
PDGF-receptor pathway
p=9.6e-9
Cell cycle p=1.6e-2
Oxidative phosphorylation
p=6.4e-4
Log2 CPM
Low
PDGFRAΔ7
High
PDGFRAΔ7
Mean
5% confidence
interval
Standard
deviation
Least-squares
trend line
C
Physical interaction
Genetic interaction
NF-kappa B
complex
SRF
SPI1
TBP
RXRA
TEAD1
REL
YY1
RELA
NFKB1
E2F1
RORA
STAT1
ELK4
FOXD3
FOXA2
ZEB1
FOXQ1
PDGFRA

Figure 4. Dose–response analysis of a PDGFR mutant

A) Coverage of exome-seq (top left) and RNA-seq reads (bottom left) in exon 7 of the *PDGFRA* gene. The deletion targets the immunoglobulin-like domain IG5 of the PDFG receptor (center). 49% of PDGFRA expressing cells express PDGFRAΔ^7 homozygously, another 49% express it heterozygously (right). B) Enriched gene sets (WEBGESTALT, DAVID, adj. P-value < 0.05) correlated to PDGFRAΔ^7. Distributions of in-pathway genes in individual cells, sorted from low PDGFRAΔ^7 to high PDGFRAΔ^7. C) An interaction network (generated via geneMANIA) of physical and genetic interactions of transcription factors, whose recognition motifs are overrepresented (OPOSSUM, z-score ≥ 10, Fisher score ≥ 7) in correlated genes. Physical interactions are interactions between the protein products, identified from proteomics experiments. Genetic interactions are changes in gene expression that occur when another gene is suppressed in a knockdown experiment.

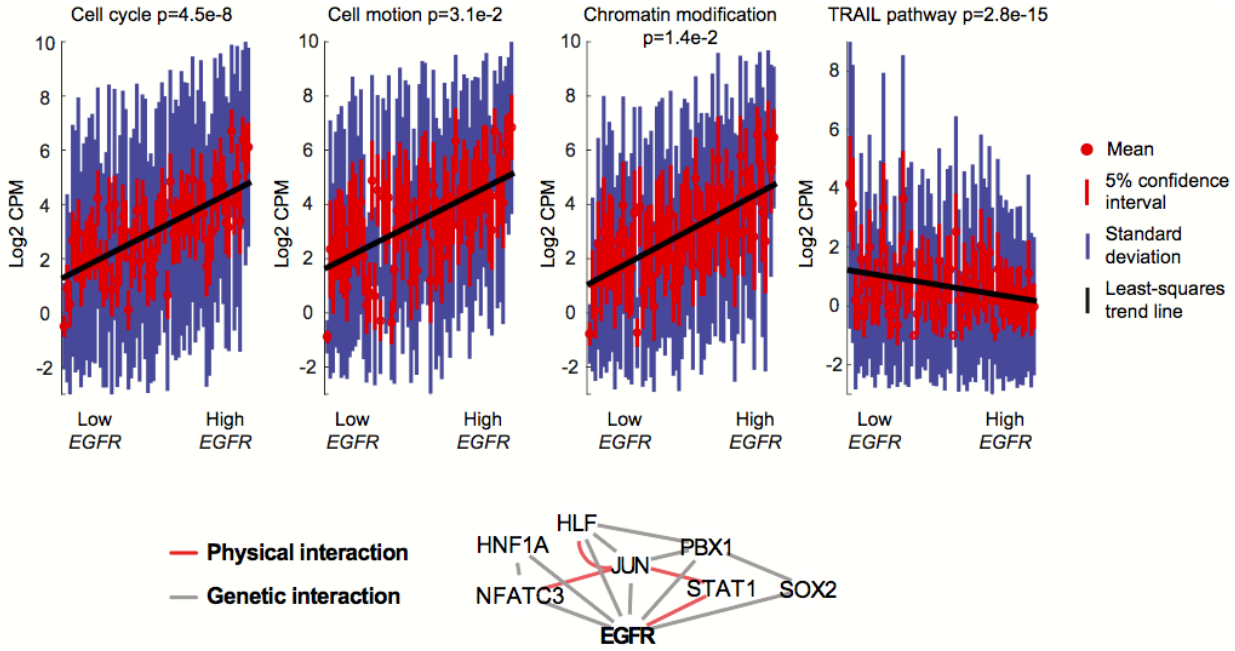

Cell cycle p=4.5e-8
Cell motion p=3.1e-2
Chromatin modification p=1.4e-2
TRAIL pathway p=2.8e-15
Log2 CPM
Low EGFR
High EGFR
Mean
5% confidence interval
Standard deviation
Least-squares trend line
Physical interaction
Genetic interaction
HLF
HNF1A
PBX1
JUN
NFATC3
STAT1
SOX2
EGFR

Figure 5. Dose response analysis of an EGFR-amplified GBM

Enriched gene sets (WEBGESTALT, DAVID, adj. p-value < 0.05) correlated to EGFR. Distributions of in-pathway genes in individual cells, sorted from low EGFR to high EGFR (top panel). An interaction network (generated via geneMANIA) of physical and genetic interactions of transcription factors, whose recognition motifs are overrepresented (OPOSSUM, z-score ≥ 10, Fisher score ≥ 7) in correlated genes. Physical interactions are interactions between the protein product, identified from proteomics experiments. Genetic interactions are changes in gene expression that occur when another gene is suppressed in a knockdown experiment.

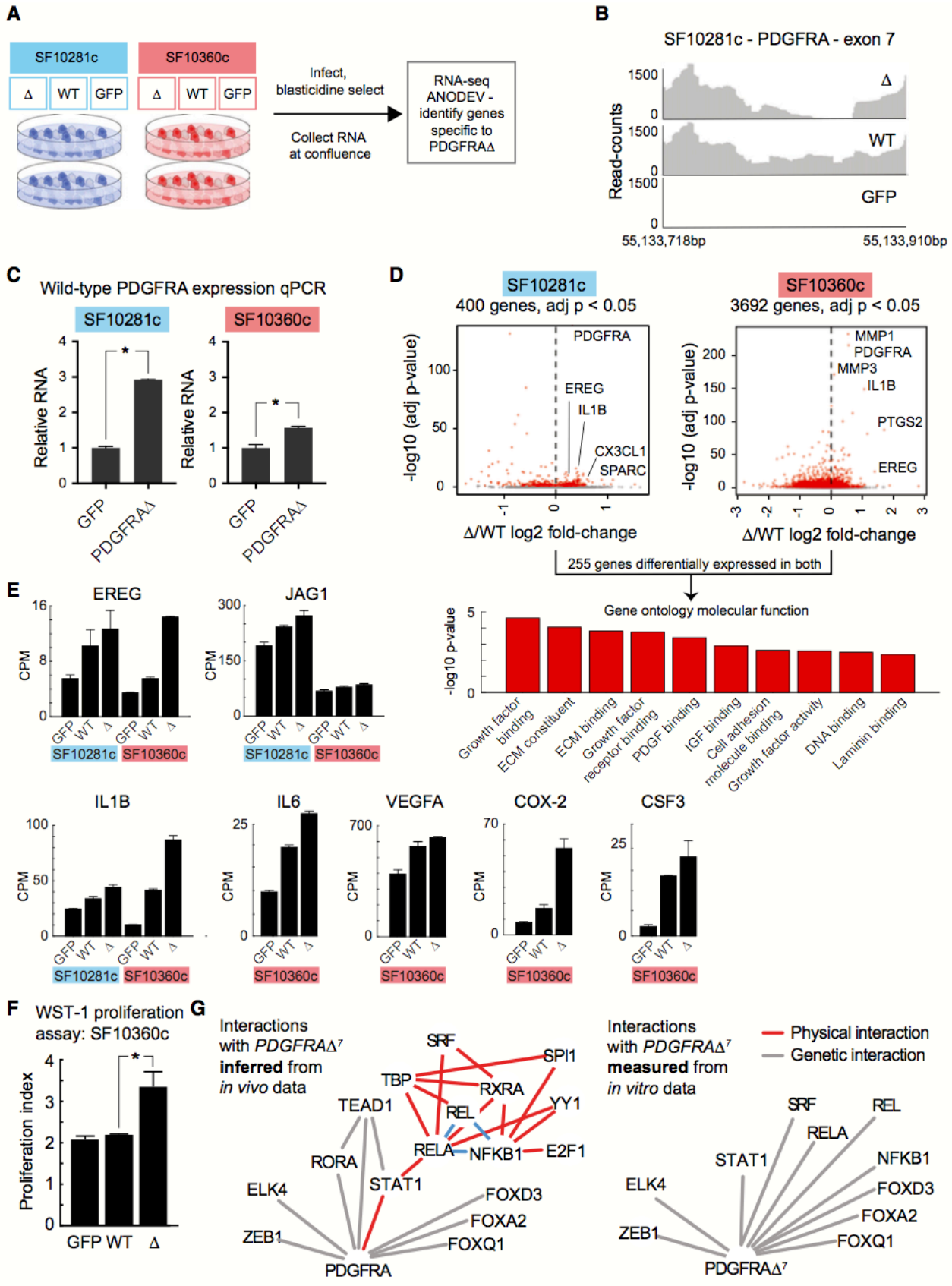

A
SF10281c
SF10360c
Δ
WT
GFP
Infect, blasticidine select
Collect RNA at confluence
RNA-seq ANODEV - identify genes specific to PDGFRAΔ
B
SF10281c - PDGFRA - exon 7
Read-counts
1500
0
Δ
WT
GFP
55,133,718bp
55,133,910bp
C
Wild-type PDGFRA expression qPCR
SF10281c
SF10360c
Relative RNA
GFP
PDGFRAΔ
D
SF10281c
400 genes, adj p < 0.05
SF10360c
3692 genes, adj p < 0.05
-log10 (adj p-value)
PDGFRA
EREG
IL1B
CX3CL1
SPARC
MMP1
MMP3
PTGS2
Δ/WT log2 fold-change
255 genes differentially expressed in both
Gene ontology molecular function
-log10 p-value
Growth factor binding
ECM constituent
ECM binding
Growth factor receptor binding
PDGF binding
IGF binding
Cell adhesion molecule binding
Growth factor activity
DNA binding
Laminin binding
E
EREG
JAG1
IL1B
IL6
VEGFA
COX-2
CSF3
CPM
GFP
WT
Δ
SF10281c
SF10360c
F
WST-1 proliferation assay: SF10360c
Proliferation index
G
Interactions with PDGFRAΔ7 inferred from in vivo data
Interactions with PDGFRAΔ7 measured from in vitro data
Physical interaction
Genetic interaction
SRF
SPI1
TBP
RXRA
TEAD1
REL
YY1
RELA
NFKB1
E2F1
RORA
STAT1
ELK4
FOXD3
FOXA2
ZEB1
FOXQ1
PDGFRA
PDGFRAΔ7

Figure 6. *In vitro* analysis of *PDGFRAΔ*^{7}

A) Schematic representation of *in vitro* experiment. B) Reads mapped to exon 7 of *PDGFRA* in *PDGFRAΔ*^{7} over-expressing, wild-type *PDGFRA* over-expressing, and GFP control cultures. C) Quantitative PCR with a probe targeted to the region deleted in *PDGFRAΔ*^{7}. Results (mean +/- SD) comparing wild-type *PDGFRA* expression between GFP control and *PDGFRAΔ*^{7} expressing cells from SF10360. The asterisk indicates $P < 0.05$ (t-test). D) Top: Volcano plot of gene expression between *PDGFRAΔ*^{7} and *PDGFRA* wild-type expressing cells from SF10281 (left) and SF10360 (right). Differentially expressed genes (adjusted P-value < 0.05, ANODEV test from DESeq2 package) are indicated in red. Bottom: Gene-ontology enrichment of genes differentially expressed between *PDGFRAΔ*^{7} and *PDGFRA* wild type in both cell lines (right). E) Bar plots of mean gene expression (+/- SD) across duplicates in GFP, wild-type *PDGFRA*, and expressing cells from SF10281 (left) and SF10360 (right). F) WST-1 assay (n = 3) comparing proliferation (mean +/- SD) of SF10360c cells expressing GFP, *PDGFRA WT* or *PDGFRAΔ*^{7}. The asterisk indicates $P < 0.05$ (t-test). G) Genetic interactions and physical interactions (via Genemania) of transcription factors (via OPOSSOM) whose motifs are enriched in the promoters of genes correlating with *PDGFRAΔ*^{7} *in vivo*. This is compared to the genetic interactions between upstream transcription factors of genes which are differentially expressed in the *PDGFRAΔ*^{7} over-expression experiment.

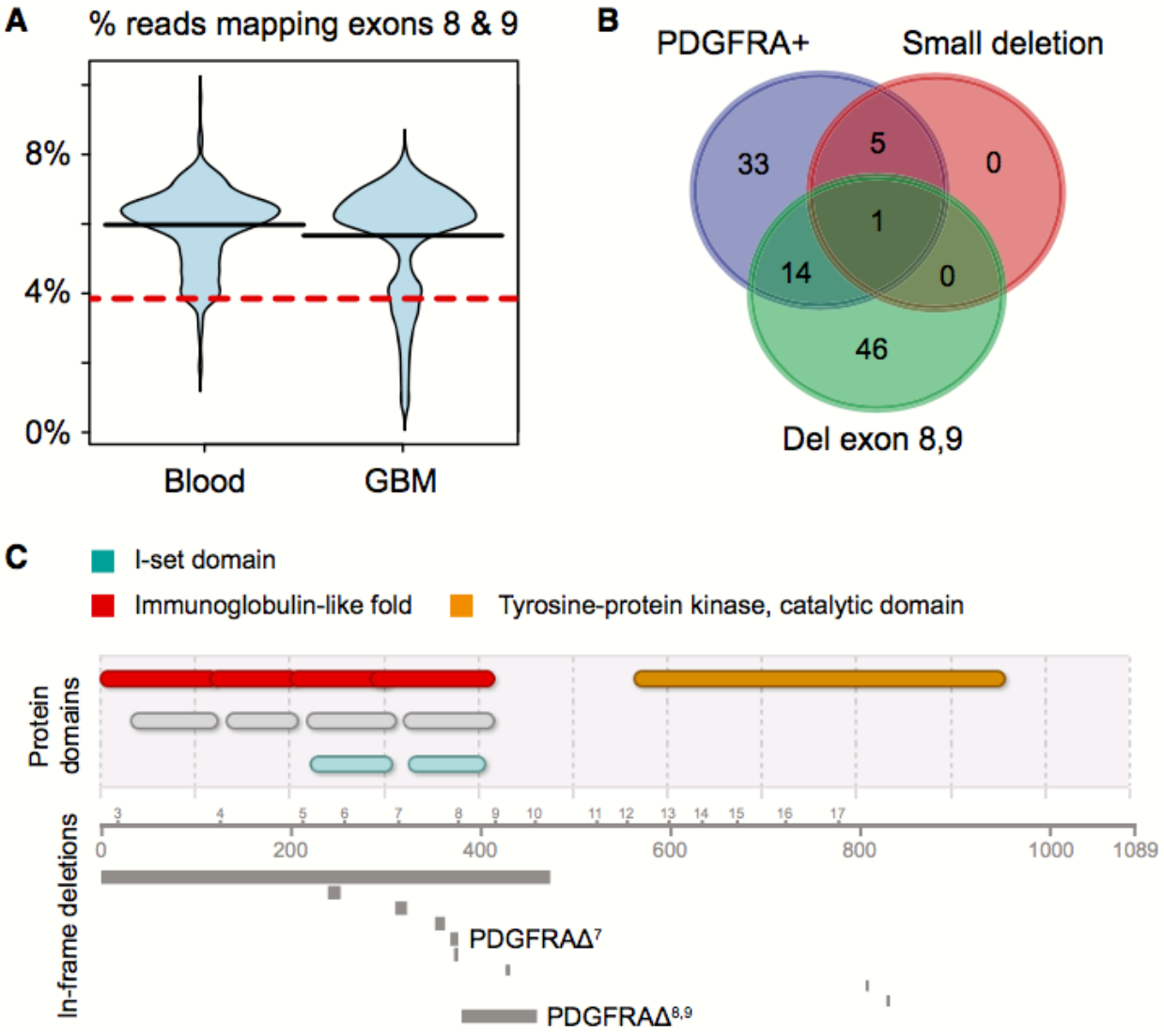

A
% reads mapping exons 8 & 9
8%
4%
0%
Blood
GBM
B
PDGFRA+
Small deletion
33
5
0
1
14
0
46
Del exon 8,9
C
I-set domain
Immunoglobulin-like fold
Tyrosine-protein kinase, catalytic domain
Protein domains
In-frame deletions
0
200
400
600
800
1000
1089
PDGFRAΔ7
PDGFRAΔ8,9

Figure 7. Analysis of TCGA data reveals a family of *PDGFRA* mutations affecting the dimerization domain

A) The fraction of reads assigned to exons 8 and 9 from all exome-seq reads mapping to *PDGFRA*, compared between blood control and GBM samples. The 10^{th} percentile of the blood control distribution is indicated by a red line. B) Venn diagram of GBM cases harboring an amplification, a deletion of exons 8 and 9, or a small deletion affecting the dimerization domain of *PDGFRA*. C) Visualization of deletions detected in *PDGFRA* from TCGA data. Domains are indicated by color (top), exons and protein residues by number (middle), and deleted regions by bars (bottom).

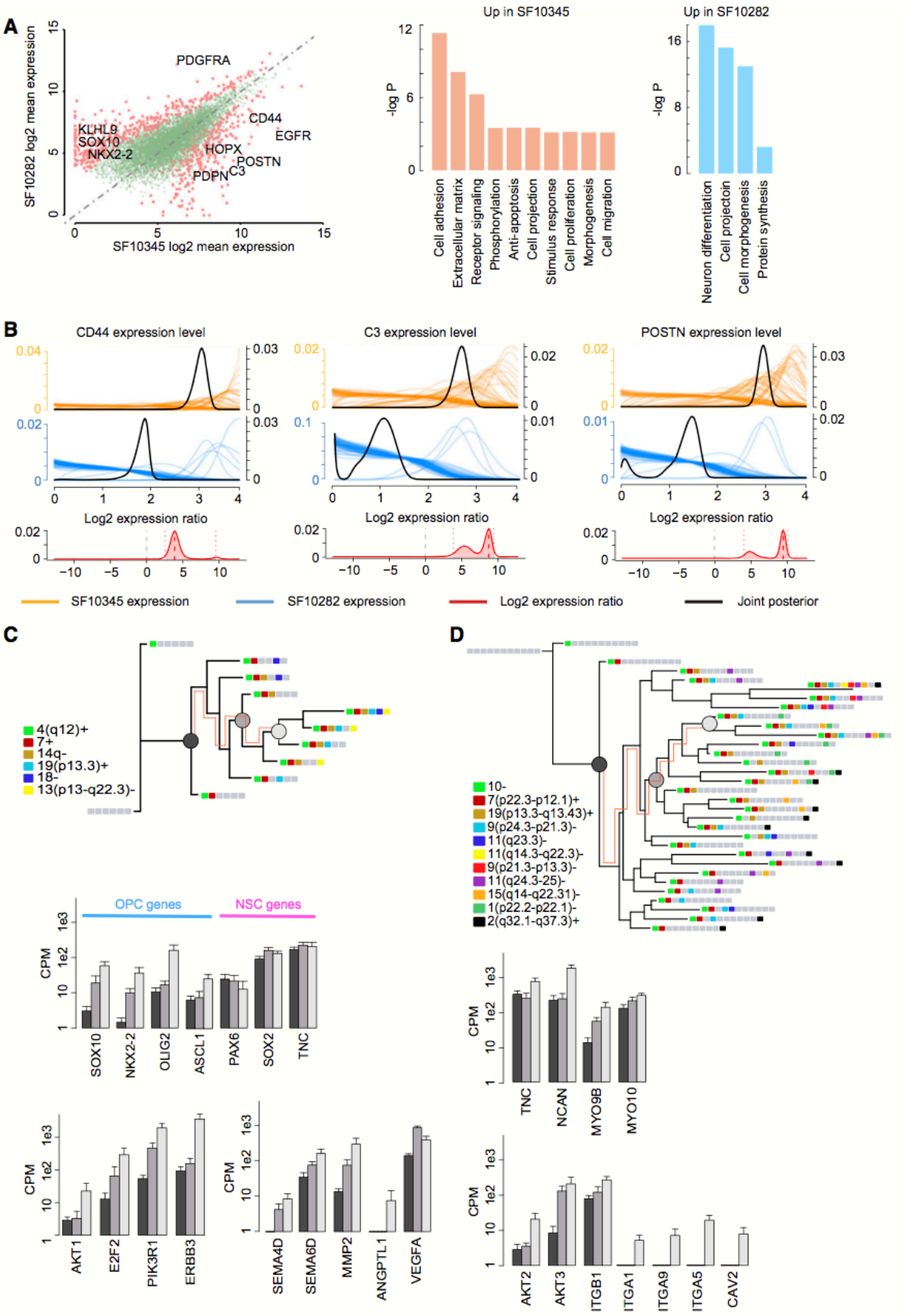

A
SF10282 log2 mean expression
SF10345 log2 mean expression
PDGFRA
CD44
EGFR
KLHL9
SOX10
NKX2-2
HOPX
POSTN
PDPN
C3
Up in SF10345
-log P
Cell adhesion
Extracellular matrix
Receptor signaling
Phosphorylation
Anti-apoptosis
Cell projection
Stimulus response
Cell proliferation
Morphogenesis
Cell migration
Up in SF10282
-log P
Neuron differentiation
Cell projectoin
Cell morphogenesis
Protein synthesis
B
CD44 expression level
C3 expression level
POSTN expression level
Log2 expression ratio
SF10345 expression
SF10282 expression
Log2 expression ratio
Joint posterior
C
4(q12)+
7+
14q-
19(p13.3)+
18-
13(p13-q22.3)-
OPC genes
NSC genes
CPM
SOX10
NKX2-2
OLIG2
ASCL1
PAX6
SOX2
TNC
AKT1
E2F2
PIK3R1
ERBB3
SEMA4D
SEMA6D
MMP2
ANGPTL1
VEGFA
D
10-
7(p22.3-p12.1)+
19(p13.3-q13.43)+
9(p24.3-p21.3)-
11(q23.3)-
11(q14.3-q22.3)-
9(p21.3-p13.3)-
11(q24.3-25)-
15(q14-q22.31)-
1(p22.2-p22.1)-
2(q32.1-q37.3)+
TNC
NCAN
MYO9B
MYO10
AKT2
AKT3
ITGB1
ITGA1
ITGA9
ITGA5
CAV2

Figure 8. Differences in gene expression between an EGF- and a PDGF-driven GBM

A) A scatterplot of log2 mean expression between SF10345 and SF10282. Adjusted P-values of biological process GO terms, enriched in differentially expressed genes (SCDE), with an adjusted P-value < 0.05 (on a log10 scale). B) Single-cell gene expression estimates (colored lines), their SF10345/SF10282 log2 fold-changes, and their joint posteriors for select genes. A 95% confidence interval is represented by dotted lines. C) Phylogenetic trees for SF10282 and SF10345 (D) . Each leaf corresponds to a unique set of cells with the same CNV genotype. Bar plots show mean (+/- SD) expression of select genes across sets of cells that progressively gain CNVs.

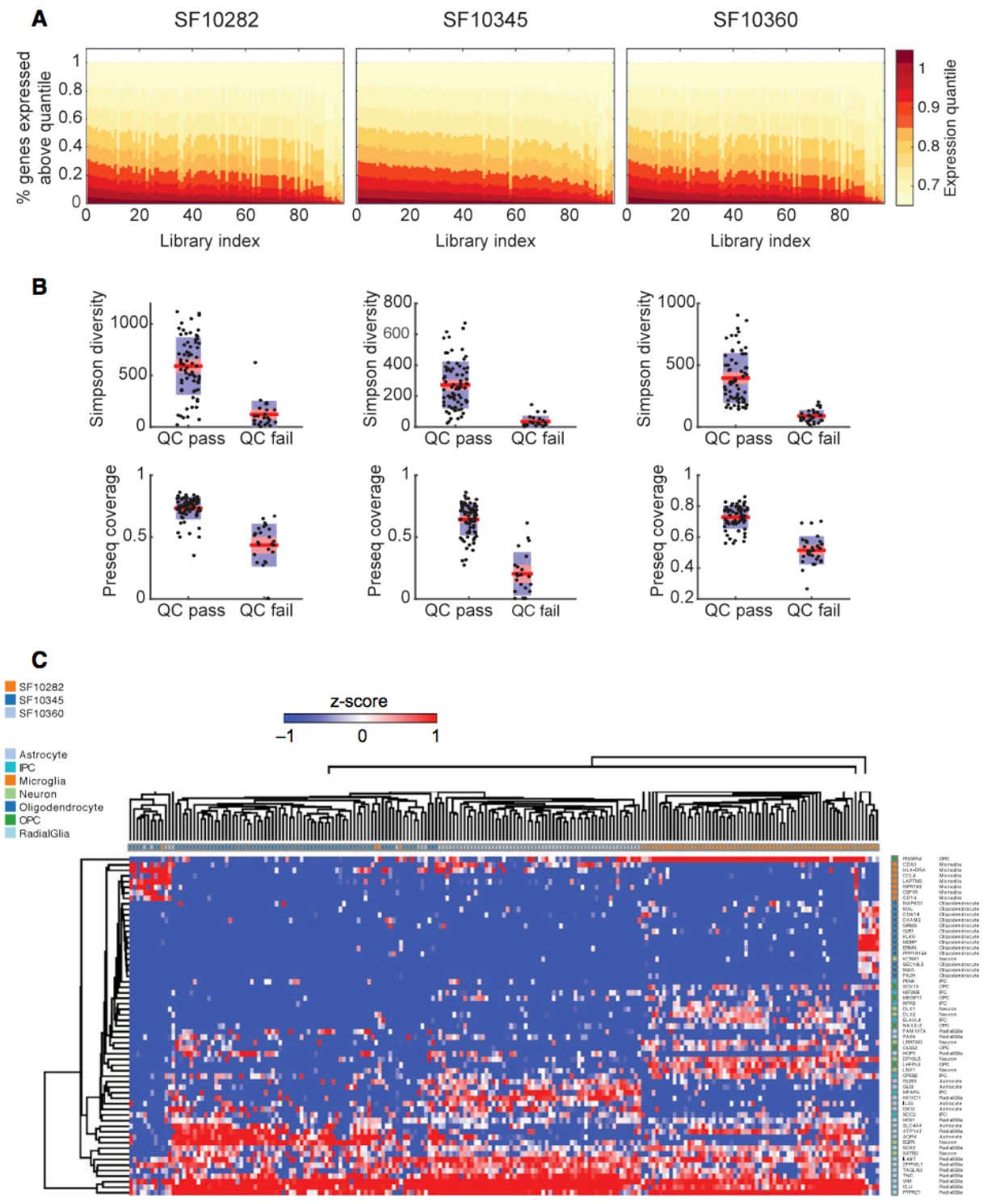

A
SF10282
SF10345
SF10360
% genes expressed above quantile
Library index
Expression quantile
B
Simpson diversity
Preseq coverage
QC pass
QC fail
C
SF10282
SF10345
SF10360
z-score
Astrocyte
IPC
Microglia
Neuron
Oligodendrocyte
OPC
RadialGlia

Figure EV1. Quality control and preprocessing

A) Percent of genes expressed above a given expression percentile for each case. B) Simpson diversity scores and Preseq coverage for each sample. C) Hierarchical clustering of all cells, using a biomarker panel of genes expressed by cell types found in the brain.

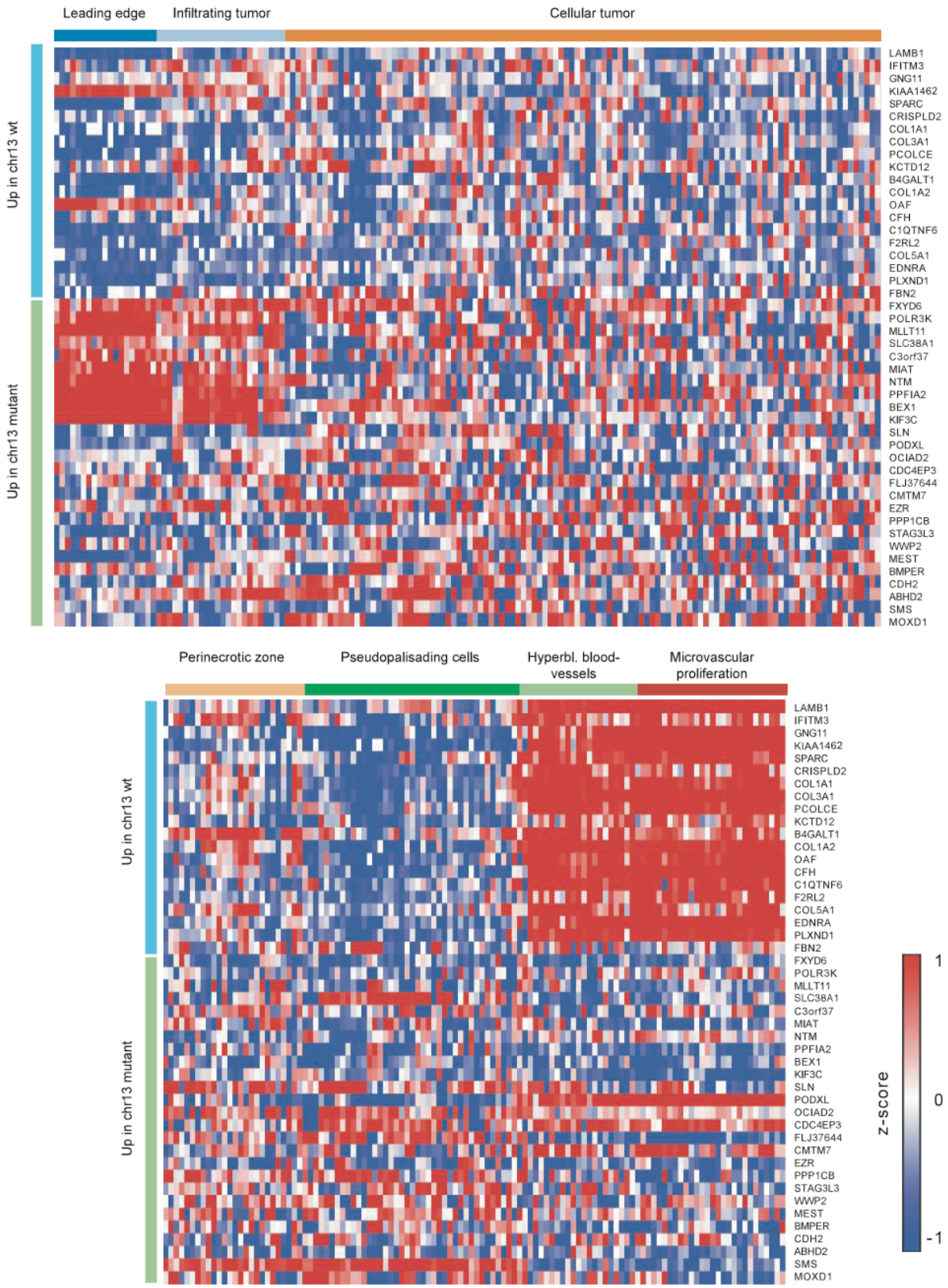

Leading edge
Infiltrating tumor
Cellular tumor
Up in chr13 wt
Up in chr13 mutant
LAMB1
IFITM3
GNG11
KIAA1462
SPARC
CRISPLD2
COL1A1
COL3A1
PCOLCE
KCTD12
B4GALT1
COL1A2
OAF
CFH
C1QTNF6
F2RL2
COL5A1
EDNRA
PLXND1
FBN2
FXYD6
POLR3K
MLLT11
SLC38A1
C3orf37
MIAT
NTM
PPFIA2
BEX1
KIF3C
SLN
PODXL
OCIAD2
CDC4EP3
FLJ37644
CMTM7
EZR
PPP1CB
STAG3L3
WWP2
MEST
BMPER
CDH2
ABHD2
SMS
MOXD1
Perinecrotic zone
Pseudopalisading cells
Hyperbl. blood-vessels
Microvascular proliferation
Up in chr13 wt
Up in chr13 mutant
LAMB1
IFITM3
GNG11
KIAA1462
SPARC
CRISPLD2
COL1A1
COL3A1
PCOLCE
KCTD12
B4GALT1
COL1A2
OAF
CFH
C1QTNF6
F2RL2
COL5A1
EDNRA
PLXND1
FBN2
FXYD6
POLR3K
MLLT11
SLC38A1
C3orf37
MIAT
NTM
PPFIA2
BEX1
KIF3C
SLN
PODXL
OCIAD2
CDC4EP3
FLJ37644
CMTM7
EZR
PPP1CB
STAG3L3
WWP2
MEST
BMPER
CDH2
ABHD2
SMS
MOXD1
z-score
1
0
-1

Figure EV2. Expression of genes differentially expressed between the chromosome 13 loss clone, and the wild-type clone in SF10360, in the Ivy Glioblastoma Atlas.

Gene expression across all samples in the Ivy Glioblastoma Atlas, for genes with and adjusted P-value < 0.05 in the differential analysis between cells harboring the chr13 loss and those free of CNV alterations on chr13.

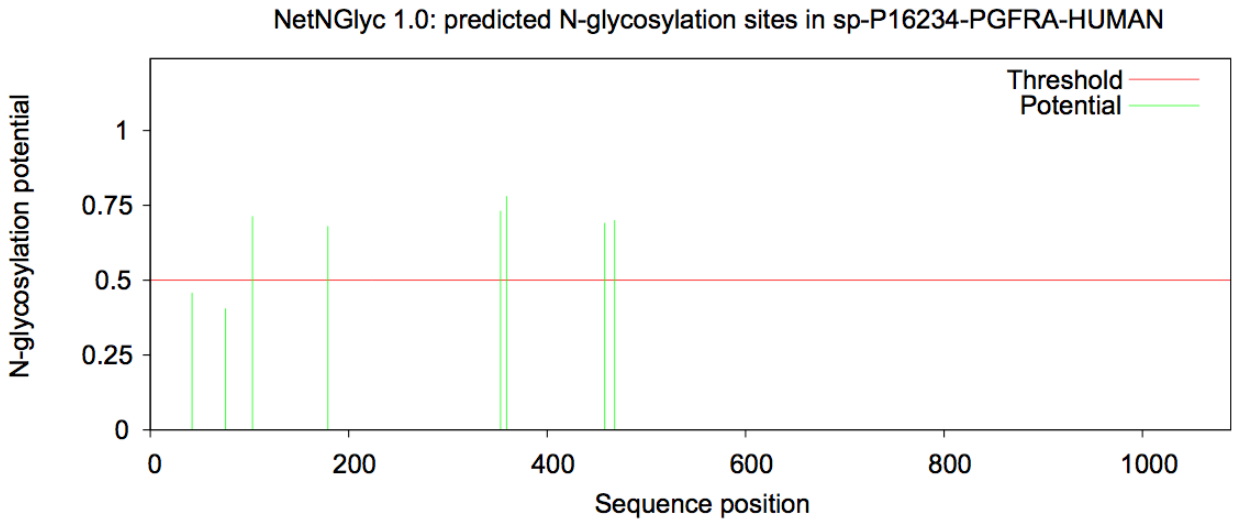

NetNGlyc 1.0: predicted N-glycosylation sites in sp-P16234-PGFRA-HUMAN
Threshold
Potential
N-glycosylation potential
1
0.75
0.5
0.25
0
0
200
400
600
800
1000
Sequence position

Figure EV3. Predicted glycosylation sites in PDGFRA.

Glycosylation sites in PDGFRA, predicted by NetNGly1.0.

References

Afonso-Grunz F & Müller S (2015) Principles of miRNA-mRNA interactions: Beyond sequence complementarity. *Cell. Mol. Life Sci.* **72:** 3127–3141

Amarasinghe KC, Li J, Hunter SM, Ryland GL, Cowin PA, Campbell IG & Halgamuge SK (2014) Inferring copy number and genotype in tumour exome data. *BMC Genomics* **15:** 732

Aqeilan RI, Calin GA & Croce CM (2010) miR-15a and miR-16-1 in cancer: discovery, function and future perspectives. *Cell Death Differ.* **17:** 215–220 Available at: http://www.nature.com/doifinder/10.1038/cdd.2009.69

Beadle C, Assanah MC, Monzo P, Vallee R, Rosenfeld SS & Canoll P (2008) The role of myosin II in glioma invasion of the brain. *Mol. Biol. Cell* **19:** 3357–68 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2488307&tool=pmcentrez&rendertype=abstract

Benjamini Y & Hochberg Y (1995) Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57:** 289–300

Bolger AM, Lohse M & Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30:** 2114–2120

Bonci D, Coppola V, Musumeci M, Addario A, Giuffrida R, Memeo L, D'Urso L, Pagliuca A, Biffoni M, Labbaye C, Bartucci M, Muto G, Peschle C & De Maria R (2008) The miR-15a–miR-16-1 cluster controls prostate cancer by targeting multiple oncogenic activities. *Nat. Med.* **14:** 1271–1277 Available at: http://www.nature.com/doifinder/10.1038/nm.1880

Brennan CW, Verhaak RGW, McKenna A, Campos B, Noushmehr H, Salama SR, Zheng S, Chakravarty D, Sanborn JZ, Berman SH, Beroukhim R, Bernard B, Wu CJ, Genovese G,

Shmulevich I, Barnholtz-Sloan J, Zou L, Vegesna R, Shukla SA, Ciriello G, et al (2013) The somatic genomic landscape of glioblastoma. *Cell* **155:**

Brennan CW, Verhaak RGW, McKenna A, Campos B, Noushmehr H, Salama SR, Zheng S, Chakravarty D, Sanborn Z, Berman SH, Beroukhim R, Bernard B & Chin L (2014) The Somatic Genomic Landscape of Glioblastoma. *Cell* **155:** 462–477

Cao Y, Slaney CY, Bidwell BN, Parker BS, Johnstone CN, Rautela J, Eckhardt BL & Anderson RL (2014) BMP4 inhibits breast cancer metastasis by blocking myeloid-derived suppressor cell activity. *Cancer Res.* **74:** 5091–5102

Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, Schultz N, Bader GD & Sander C (2011) Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* **39:** D685–D690 Available at: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkq1039

Chautard E, Ouédraogo ZG, Biau J & Verrelle P (2014) Role of Akt in human malignant glioma: From oncogenesis to tumor aggressiveness. *J. Neurooncol.* **117:** 205–215

Chen JCC, Alvarez MJJ, Talos F, Dhruv H, Rieckhof GEE, Iyer A, Diefes KLL, Aldape K, Berens M, Shen MMM & Califano A (2014) Identification of causal genetic drivers of human disease through systems-level analysis of regulatory networks. *Cell* **159:** 402–14 Available at: http://www.ncbi.nlm.nih.gov/pubmed/25303533\nhttp://linkinghub.elsevier.com/retrieve/pii/S0092867414011702

Chen P-HH, Chen X & He X (2012) Platelet-derived growth factors and their receptors: Structural and functional perspectives. *Biochim. Biophys. Acta (BBA)-Proteins …* **1834:**

2176–86 Available at: http://dx.doi.org/10.1016/j.bbapap.2012.10.015

Chou C-H, Chang N-W, Shrestha S, Hsu S-D, Lin Y-L, Lee W-H, Yang C-D, Hong H-C, Wei T-Y, Tu S-J, Tsai T-R, Ho S-Y, Jian T-Y, Wu H-Y, Chen P-R, Lin N-C, Huang H-T, Yang T-L, Pai C-Y, Tai C-S, et al (2015) miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res.* **5712121:** gkv1258 Available at: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv1258

Clarke ID & Dirks PB (2003) A human brain tumor-derived PDGFR-α deletion mutant is transforming. *Oncogene* **22:** 722–733 Available at: http://www.nature.com/doifinder/10.1038/sj.onc.1206160

Cuddapah VA, Robel S, Watkins S & Sontheimer H (2014) A neurocentric perspective on glioma invasion. *Nat. Rev. Neurosci.* **15:** 455–465 Available at: http://www.nature.com/doifinder/10.1038/nrn3765

Darmanis S, Sloan SA, Zhang Y, Enge M, Caneda C, Shuer LM, Hayden Gephart MG, Barres BA & Quake SR (2015) A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci.* **112:** 7285–7290 Available at: http://www.pnas.org/lookup/doi/10.1073/pnas.1507125112

Deloulme JC, Raponi E, Gentil BJ, Bertacchi N, Marks A, Labourdette G & Baudier J (2004) Nuclear expression of S100B in oligodendrocyte progenitor cells correlates with differentiation toward the oligodendroglial lineage and modulates oligodendrocytes maturation. *Mol. Cell. Neurosci.* **27:** 453–465 Available at: http://linkinghub.elsevier.com/retrieve/pii/S1044743104001721

Diaz A, Liu SJ, Sandoval C, Pollen A, Nowakowski TJ, Lim DA & Kriegstein A (2016) SCell:

integrated analysis of single-cell RNA-seq data. *Bioinformatics* **32:** 2219–2220 Available at: http://bioinformatics.oxfordjournals.org/lookup/doi/10.1093/bioinformatics/btw201

Fitch WM & Margoliash E (1967) Construction of phylogenetic trees. *Science (80-.).* **155:** 279–284

Francis JM, Zhang CZ, Maire CL, Jung J, Manzo VE, Adalsteinsson VA, Homer H, Haidar S, Blumenstiel B, Pedamallu CS, Ligon AH, Love JC, Meyerson M & Ligon KL (2014) EGFR variant heterogeneity in glioblastoma resolved through single-nucleus sequencing. *Cancer Discov.* **4:** 956–971

Fujita M, Kohanbash G, Fellows-Mayle W, Hamilton RL, Komohara Y, Decker SA, Ohlfest JR & Okada H (2011) COX-2 blockade suppresses gliomagenesis by inhibiting myeloid-derived suppressor cells. *Cancer Res.* **71:** 2664–2674

Garvin T, Aboukhalil R, Kendall J, Baslan T, Atwal GS, Hicks J, Wigler M & Schatz M (2014) Interactive analysis and quality assessment of single-cell copy-number variations. *bioRxiv* **12:** 11346 Available at: http://www.biorxiv.org/content/early/2014/11/12/011346.abstract

Goldberg JL (2004) An Oligodendrocyte Lineage-Specific Semaphorin, Sema5A, Inhibits Axon Growth by Retinal Ganglion Cells. *J. Neurosci.* **24:** 4989–4999 Available at: http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.4390-03.2004

Huang DW, Sherman BT & Lempicki R a (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4:** 44–57 Available at: http://www.ncbi.nlm.nih.gov/pubmed/19131956 [Accessed January 28, 2013]

Jepson S, Vought B, Gross CH, Gan L, Austen D, Frantz JD, Zwahlen J, Lowe D, Markland W & Krauss R (2012) LINGO-1, a Transmembrane Signaling Protein, Inhibits

Oligodendrocyte Differentiation and Myelination through Intercellular Self-interactions. *J. Biol. Chem.* **287:** 22184–22195 Available at: http://www.jbc.org/cgi/doi/10.1074/jbc.M112.366179

Johnson BE, Mazor T, Hong C, Barnes M, Aihara K, McLean CY, Fouse SD, Yamamoto S, Ueda H, Tatsuno K, Asthana S, Jalbert LE, Nelson SJ, Bollen AW, Gustafson WC, Charron E, Weiss WA, Smirnov I V, Song JS, Olshen AB, et al (2014) Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. *Science* **343:** 189–93 Available at: http://www.sciencemag.org/content/343/6167/189.full

Kessaris N, Fogarty M, Iannarelli P, Grist M, Wegner M & Richardson WD (2006) Competing waves of oligodendrocytes in the forebrain and postnatal elimination of an embryonic lineage. *Nat. Neurosci.* **9:** 173–179

Kharchenko P V, Silberstein L & Scadden DT (2014) Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11:** 740–2 Available at: http://www.ncbi.nlm.nih.gov/pubmed/24836921 [Accessed July 9, 2014]

Kim D, Langmead B & Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12:** 357–360 Available at: http://www.nature.com/nmeth/journal/v12/n4/full/nmeth.3317.html?WT.ec_id=NMETH-201504\nhttp://www.nature.com/nmeth/journal/v12/n4/extref/nmeth.3317-S1.pdf\nhttp://www.nature.com/nmeth/journal/v12/n4/pdf/nmeth.3317.pdf

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R & Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14:** R36 Available at: http://genomebiology.com/2013/14/4/R36

Kumabe T, Sohma Y, Kayama T, Yoshimoto T & Yamamoto T (1992) Amplification of alpha-platelet-derived growth factor receptor gene lacking an exon coding for a portion of the extracellular region in a primary brain tumor of glial origin. *Oncogene* **7:** 627–633 Available at: http://europepmc.org/abstract/MED/1314366

Kuspert M, Hammer A, Bosl MR & Wegner M (2011) Olig2 regulates Sox10 expression in oligodendrocyte precursors through an evolutionary conserved distal enhancer. *Nucleic Acids Res.* **39:** 1280–1293 Available at: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkq951

Lechner MG, Liebertz DJ & Epstein AL (2010) Characterization of Cytokine-Induced Myeloid-Derived Suppressor Cells from Normal Human Peripheral Blood Mononuclear Cells. *J. Immunol.* **185:** 2273–2284 Available at: http://www.jimmunol.org/cgi/content/abstract/185/4/2273

Li H & Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25:** 1754–60

Liao Y, Smyth GK & Shi W (2014) FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30:** 923–930

Liu SJ, Nowakowski TJ, Pollen AA, Lui JH, Horlbeck MA, Attenello FJ, He D, Weissman JS, Kriegstein AR, Diaz AA & Lim DA (2016) Single-cell analysis of long non-coding RNAs in the developing human neocortex. *Genome Biol.* **17:** 67 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4831157&tool=pmcentrez&rendertype=abstract

Markiewski MM, DeAngelis RA, Benencia F, Ricklin-Lichtsteiner SK, Koutoulaki A, Gerard C,

Coukos G & Lambris JD (2008) Modulation of the antitumor immune response by complement. *Nat. Immunol.* **9:** 1225–1235 Available at: http://www.nature.com/doifinder/10.1038/ni.1655

Mazor T, Pankov A, Brett E, Chang SM, Jun S & Costello JF (2015) DNA Methylation and Somatic Mutations Converge on the Cell Cycle and Define Similar Evolutionary Histories in Brain Tumors Article DNA Methylation and Somatic Mutations Converge on the Cell Cycle and Define Similar Evolutionary Histories in Brain. *Cancer Cell* **28:** 307–317 Available at: http://dx.doi.org/10.1016/j.ccell.2015.07.012

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M & DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20:** 1297–303

Menn B, Garcia-Verdugo JM, Yaschine C, Gonzalez-Perez O, Rowitch D & Alvarez-Buylla A (2006) Origin of oligodendrocytes in the subventricular zone of the adult brain. *J. Neurosci.* **26:** 7907–7918

Meyer M, Reimand J, Lan X, Head R, Zhu X, Kushida M, Bayani J, Pressey JC, Lionel AC, Clarke ID, Cusimano M, Squire JA, Scherer SW, Bernstein M, Woodin M a, Bader GD & Dirks PB (2015) Single cell-derived clonal analysis of human glioblastoma links functional and genomic heterogeneity. *Proc. Natl. Acad. Sci. U. S. A.* **112:** 851–6

Montojo J, Zuberi K, Rodriguez H, Kazi F, Wright G, Donaldson SL, Morris Q & Bader GD (2010) GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. *Bioinformatics* **26:** 2927–2928 Available at:

http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btq562

Nakatani H, Martin E, Hassani H, Clavairoly A, Maire CL, Viadieu A, Kerninon C, Delmasure A, Frah M, Weber M, Nakafuku M, Zalc B, Thomas J-L, Guillemot F, Nait-Oumesmar B & Parras C (2013) Ascl1/Mash1 Promotes Brain Oligodendrogenesis during Myelination and Remyelination. *J. Neurosci.* **33:** 9752–9768 Available at: http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.0805-13.2013

Nath D, Slocombe PM, Webster a, Stephens PE, Docherty a J & Murphy G (2000) Meltrin gamma(ADAM-9) mediates cellular adhesion through alpha(6)beta(1)integrin, leading to a marked induction of fibroblast cell motility. *J. Cell Sci.* **113 (Pt 1:** 2319–28 Available at: http://www.ncbi.nlm.nih.gov/pubmed/10825303

Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, Muthuswamy L, Krasnitz A, McCombie WR, Hicks J & Wigler M (2011) Tumour evolution inferred by single-cell sequencing. *Nature* **472:** 90–94 Available at: http://www.nature.com/doifinder/10.1038/nature09807 [Accessed March 14, 2011]

Ostrem BEL, Lui JH, Gertz CC & Kriegstein AR (2014) Control of Outer Radial Glial Stem Cell Mitosis in the Human Brain. *Cell Rep.* **8:** 656–664 Available at: http://dx.doi.org/10.1016/j.celrep.2014.06.058

Ota T, Suto S, Katayama H, Han Z, Suzuki F, Maeda M, Tanino M, Terada Y & Tatsuka M (2002) Increased mitotic phosphorylation of histone H3 attributable to AIM-1/Aurora-B overexpression contributes to chromosome number instability. *Cancer Res.* **62:** 5168–77 Available at: http://www.ncbi.nlm.nih.gov/pubmed/12234980

Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, Cahill DP, Nahed B

V, Curry WT, Martuza RL, Louis DN, Rozenblatt-Rosen O, Suvà ML, Regev A & Bernstein BE (2014) Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344:** 1396–401 Available at: http://www.ncbi.nlm.nih.gov/pubmed/24925914 [Accessed July 9, 2014]

Paugh BS, Zhu X, Qu C, Endersby R, Diaz AK, Zhang J, Bax DA, Carvalho D, Reis RM, Onar-Thomas A, Broniscer A, Wetmore C, Zhang J, Jones C, Ellison DW & Baker SJ (2013) Novel oncogenic PDGFRA mutations in pediatric high-grade gliomas. *Cancer Res.* **73:** 6219–6229

Pekarsky Y & Croce CM (2014) Role of miR-15/16 in CLL. *Cell Death Differ.***:** 1–6 Available at: http://www.ncbi.nlm.nih.gov/pubmed/24971479

Petryniak MA, Potter GB, Rowitch DH & Rubenstein JLR (2007) Dlx1 and Dlx2 control neuronal versus oligodendroglial cell fate acquisition in the developing forebrain. *Neuron* **55:** 417–33 Available at: http://www.sciencedirect.com/science/article/pii/S089662730700493X

Pietras A, Katz AM, Ekström EJ, Wee B, Halliday JJ, Pitter KL, Werbeck JL, Amankulor NM, Huse JT & Holland EC (2014) Osteopontin-CD44 signaling in the glioma perivascular niche enhances cancer stem cell phenotypes and promotes aggressive tumor growth. *Cell Stem Cell* **14:** 357–69 Available at: http://www.sciencedirect.com/science/article/pii/S193459091400006X

Pollen AA, Chen J, Leyrat AA, West JA, Kriegstein AR, Nowakowski TJ, Chen J, Retallack H & Sandoval-espinosa C (2015a) Molecular Identity of Human Outer Radial Glia during Article Molecular Identity of Human Outer Radial Glia during Cortical Development. *Cell*

163: 55–67 Available at: http://dx.doi.org/10.1016/j.cell.2015.09.004

Pollen AA, Nowakowski TJ, Chen J, Retallack H, Sandoval-Espinosa C, Nicholas CR, Shuga J, Liu SJ, Oldham MC, Diaz A, Lim DA, Leyrat AA, West JA & Kriegstein AR (2015b) Molecular Identity of Human Outer Radial Glia during Cortical Development. *Cell* **163:** 55–67 Available at: http://dx.doi.org/10.1016/j.cell.2015.09.004

Prados MD, Byron SA, Tran NL, Phillips JJ, Molinaro AM, Ligon KL, Wen PY, Kuhn JG, Mellinghoff IK, de Groot JF, Colman H, Cloughesy TF, Chang SM, Ryken TC, Tembe WD, Kiefer JA, Berens ME, Craig DW, Carpten JD & Trent JM (2015) Toward precision medicine in glioblastoma: the promise and the challenges. *Neuro. Oncol.* **17:** 1051–1063 Available at: http://neuro-oncology.oxfordjournals.org/lookup/doi/10.1093/neuonc/nov031

Quinlan AR & Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–2 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2832824&tool=pmcentrez&rendertype=abstract [Accessed May 25, 2014]

Revell LJ & Chamberlain SA (2014) Rphylip: An R interface for PHYLIP. *Methods Ecol. Evol.***:** n/a-n/a

Reyes SB, Narayanan AS, Lee HS, Tchaicha JH, Aldape KD, Lang FF, Tolias KF & McCarty JH (2013) avb8 integrin interacts with RhoGDI1 to regulate Rac1 and Cdc42 activation and drive glioblastoma cell invasion. *Mol. Biol. Cell* **24:** 474–482 Available at: http://www.molbiolcell.org/cgi/doi/10.1091/mbc.E12-07-0521

Sarkar S, Zemp FJ, Senger D, Robbins SM & Yong VW (2015) ADAM-9 is a novel mediator of tenascin-C-stimulated invasiveness of brain tumor-initiating cells. *Neuro. Oncol.* **17:** 1095–

1105 Available at: http://neuro-oncology.oxfordjournals.org/content/early/2015/02/01/neuonc.nou362.abstract

Shafit-Zagardo B, Davies P, Rockwood J, Kress Y & Lee SC (2000) Novel microtubule-associated protein-2 isoform is expressed early in human oligodendrocyte maturation. *Glia* **29:** 233–245

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B & Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13:** 2498–504 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=403769&tool=pmcentrez&rendertype=abstract [Accessed May 21, 2013]

Sloan KE (2005) CD155/PVR Enhances Glioma Cell Dispersal by Regulating Adhesion Signaling and Focal Adhesion Dynamics. *Cancer Res.* **65:** 10930–10937 Available at: http://cancerres.aacrjournals.org/cgi/doi/10.1158/0008-5472.CAN-05-1890

Sottoriva A, Spiteri I, Piccirillo SGM, Touloumis A, Collins VP, Marioni JC, Curtis C, Watts C & Tavaré S (2013) Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc. Natl. Acad. Sci. U. S. A.* **110:** 4009–14 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3593922&tool=pmcentrez&rendertype=abstract [Accessed July 16, 2014]

Stolt CC, Rehberg S, Ader M, Lommes P, Riethmacher D, Schachner M, Bartsch U & Wegner M (2002) Terminal differentiation of myelin-forming oligodendrocytes depends on the transcription factor Sox10. *Genes Dev.* **16:** 165–70 Available at: http://genesdev.cshlp.org/content/16/2/165.long

Stommel JM, Kimmelman AC, Ying H, Nabioullin R, Ponugoti AH, Wiedemeyer R, Stegh AH, Bradner JE, Ligon KL, Brennan C, Chin L & DePinho R a (2007) Coactivation of receptor tyrosine kinases affects the response of tumor cells to targeted therapies. *Science* **318:** 287–290

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, Konkel MK, Malhotra A, Stütz AM, Shi X, Paolo Casale F, Chen J, Hormozdiari F, Dayama G, Chen K, Malig M, et al (2015) An integrated map of structural variation in 2,504 human genomes. *Nature* **526:** 75–81 Available at: http://www.ncbi.nlm.nih.gov/pubmed/26432246

Sui Ho SJ, Fulton DL, Arenillas DJ, Kwon AT & Wasserman WW (2007) OPOSSUM: Integrated tools for analysis of regulatory motif over-representation. *Nucleic Acids Res.* **35:** 245–252

Sun Y, Zhang W, Chen D, Lv Y, Zheng J, Lilljebjörn H, Ran L, Bao Z, Soneson C, Sjögren HO, Salford LG, Ji J, French PJ, Fioretos T, Jiang T & Fan X (2014) A glioma classification scheme based on coexpression modules of EGFR and PDGFRA. *Proc. Natl. Acad. Sci.* **111:** 3538–3543 Available at: http://www.pnas.org/lookup/doi/10.1073/pnas.1313814111

Suvà ML, Rheinbay E, Gillespie SM, Patel AP, Wakimoto H, Rabkin SD, Riggi N, Chi AS, Cahill DP, Nahed B V, Curry WT, Martuza RL, Rivera MN, Rossetti N, Kasif S, Beik S, Kadri S, Tirosh I, Wortman I, Shalek AK, et al (2014) Reconstructing and reprogramming the tumor-propagating potential of glioblastoma stem-like cells. *Cell* **157:** 580–94 Available at: http://www.ncbi.nlm.nih.gov/pubmed/24726434 [Accessed July 10, 2014]

Szerlip NJ, Pedraza a., Chakravarty D, Azim M, McGuire J, Fang Y, Ozawa T, Holland EC,

Huse JT, Jhanwar S, Leversha M a., Mikkelsen T & Brennan CW (2012) Intratumoral heterogeneity of receptor tyrosine kinases EGFR and PDGFRA amplification in glioblastoma defines subpopulations with distinct growth factor response. *Proc. Natl. Acad. Sci.* **109:** 3041–3046

Tatsuka M, Katayama H, Ota T, Tanaka T, Odashima S, Suzuki F & Terada Y (1998) Multinuclearity and increased ploidy caused by overexpression of the Aurora- and Ipl1-like midbody-associated protein mitotic kinase in human cancer cells. *Cancer Res.* **58:** 4811–4816

Tenan M, Aurrand-Lions M, Widmer V, Alimenti A, Burkhardt K, Lazeyras F, Belkouch M-C, Hammel P, Walker PR, Duchosal M a, Imhof B a & Dietrich P-Y (2010) Cooperative expression of junctional adhesion molecule-C and -B supports growth and invasion of glioma. *Glia* **58:** 524–37 Available at: http://www.ncbi.nlm.nih.gov/pubmed/19795504

Venkatesh HS, Johung TB, Caretti V, Noll A, Tang Y, Nagaraja S, Gibson EM, Mount CW, Polepalli J, Mitra SS, Woo PJ, Malenka RC, Vogel H, Bredel M, Mallick P & Monje M (2015) Neuronal Activity Promotes Glioma Growth through Neuroligin-3 Secretion. *Cell* **161:** 803–816 Available at: http://linkinghub.elsevier.com/retrieve/pii/S0092867415004298

Verhaak RGW, Hoadley K a, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP, Alexe G, Lawrence M, O'Kelly M, Tamayo P, Weir B a, Gabriel S, Winckler W, Gupta S, Jakkula L, Feiler HS, et al (2010) Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17:** 98–110 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2818769&tool=pmcentrez&ren

dertype=abstract [Accessed July 11, 2014]

Wang J, Duncan D, Shi Z & Zhang B (2013) WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res.* **41:** W77-83

Wang K, Li M & Hakonarson H (2010) ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38:** 1–7

Weller M, Cloughesy T, Perry JR & Wick W (2013) Standards of care for treatment of recurrent glioblastoma--are we there yet? *Neuro. Oncol.* **15:** 4–27 Available at: http://neuro-oncology.oxfordjournals.org/cgi/doi/10.1093/neuonc/nos273

Winterhalter C, Widera P & Krasnogor N (2014) JEPETTO: a Cytoscape plugin for gene set enrichment and topological analysis based on interaction networks. *Bioinformatics* **30:** 1029–30 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3967109&tool=pmcentrez&rendertype=abstract\nhttp://www.ncbi.nlm.nih.gov/pubmed/24363376

De Witt Hamer PC (2010) Small molecule kinase inhibitors in glioblastoma: a systematic review of clinical studies. *Neuro. Oncol.* **12:** 304–16

Ye K, Schulz MH, Long Q, Apweiler R & Ning Z (2009) Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25:** 2865–2871

Yoon C-H, Kim M-J, Kim R-K, Lim E-J, Choi K-S, An S, Hwang S-G, Kang S-G, Suh Y, Park M-J & Lee S-J (2012) c-Jun N-terminal kinase has a pivotal role in the maintenance of self-renewal and tumorigenicity in glioma stem-like cells. *Oncogene***:** 4655–4666

Zhou Q & Anderson DJ (2002) The bHLH transcription factors OLIG2 and OLIG1 couple neuronal and glial subtype specification. *Cell* **109:** 61–73

Zhou W, Ke SQ, Huang Z, Flavahan W, Fang X, Paul J, Wu L, Sloan AE, McLendon RE, Li X, Rich JN & Bao S (2015) Periostin secreted by glioblastoma stem cells recruits M2 tumour-associated macrophages and promotes malignant growth. *Nat. Cell Biol.* **17:** 170–182 Available at: http://www.nature.com/doifinder/10.1038/ncb3090

Chapter 4

CRISPRi-based genome-scale identification of functional long non-coding RNA loci in human cells

Abstract

The human genome produces thousands of long non-coding RNAs (lncRNAs) – transcripts >200 nucleotides long that do not encode proteins. While critical roles in normal biology and disease have been revealed for a subset of lncRNAs, the function of the vast majority remains untested. Here, we developed a CRISPR interference (CRISPRi) platform targeting 16,401 lncRNA loci in 7 diverse cell lines including 6 transformed cell lines and human induced pluripotent stem cells (iPSCs). Large-scale screening identified 499 lncRNA loci required for robust cellular growth, of which 89% showed growth modifying function exclusively in one cell type. We further found that lncRNA knockdown can perturb complex transcriptional networks in a cell type-specific manner and that chromatin architecture may underlie some cell type-specific function. Detailed follow up revealed potential *cis* regulatory function for *LINC00263,* and also a lncRNA-oncogene fusion with exceptional negative growth phenotype. These data underscore the functional importance and cell type-specificity of many lncRNAs.

Introduction

Sequencing efforts have revealed that the human genome produces tens of thousands of long non-coding RNAs (lncRNAs), transcripts over 200 nucleotides that are often spliced and polyadenylated but have no apparent protein coding potential (Djebali et al., 2012; FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al., 2014; Ulitsky and Bartel, 2013). Certain lncRNAs play critical roles in cellular function, development, and disease (Ponting et al., 2009; Rinn and Chang, 2012). However, of the very large set of lncRNAs – many of which are differentially expressed in tissues and disease states – only a very small fraction have established biological functions, and even fewer are known to function in fundamental aspects of cell biology such as cell proliferation. Currently, it is not possible to predict which lncRNAs are functional, let alone what function they perform. Thus, a large-scale, systematic approach to evaluating the function of the vast population of lncRNAs is critical to understanding the roles that these non-coding transcripts play in cell biology.

A central limitation to systematic efforts to evaluate lncRNA function has been the lack of highly specific, scalable tools for inhibiting lncRNA gene activity (Bassett et al., 2014). Gene deletion studies conducted in mice, flies, and human cells have yielded important biological insights about lncRNAs, but this approach is difficult to scale up (Aparicio-Prat et al., 2015; Ho et al., 2014; Meller and Rattner, 2002; Sauvageau et al., 2013). CRISPR/Cas9 nuclease approaches based on introduction of indels – while both scalable and useful for targeted loss of function studies of protein coding genes by altering the coding frame – are not well suited for the study of lncRNA gene function, as small deletions do not generally disrupt their biological activity (Shalem et al., 2014; Shi et al., 2015; Wang et al., 2014). Nonetheless, larger Cas9-mediated genetic deletions can be effective at eliminating lncRNA genes (Bassett et al., 2014;

Groff et al., 2016; Paralkar et al., 2016; Yin et al., 2015; Zhu et al., 2016). Screens based on RNA interference (RNAi) have been valuable (Guttman et al., 2011; Lin et al., 2014) despite challenges with off-target effects (Adamson et al., 2012). However, many lncRNAs localize to the nucleus, where RNAi exhibits variable knockdown efficiency (Zeng and Cullen, 2002).

We previously developed CRISPRi, a technology which can repress transcription of any gene via the targeted recruitment of the nuclease-dead dCas9-KRAB repressor fusion protein to the transcriptional start site (TSS) by a single guide RNA (sgRNA) (Gilbert et al., 2014; 2013; Qi et al., 2013). As CRISPRi acts only within a small window (1kb) around the targeted TSS (Gilbert et al., 2014), and as dCas9 occludes only 23bp of the targeted DNA strand (Nishimasu et al., 2015), CRISPRi allows for precise perturbation of any lncRNA gene. By catalyzing repressive chromatin modifications around the TSS and serving as a transcriptional roadblock, CRISPRi tests a broad range of lncRNA gene functions including the production of *cis*- and *trans*-acting RNA transcripts (Rinn and Chang, 2012), *cis*-mediated regulation related to lncRNA transcription itself (Engreitz et al., 2016a; Kornienko et al., 2013; Li et al., 2013; Ørom et al., 2010), and enhancer-like function of some lncRNA loci (Fulco et al., 2016; Paralkar et al., 2016; Yin et al., 2015). The repressive chromatin modification H3K9me3 catalyzed by CRISPRi is highly specific, with little to no off-target effects due to either spurious dCas9 binding or unintended silencing of distal regulatory elements, as measured by ChIP-seq or RNA-seq (Amabile et al., 2016; Braun et al., 2016; Gilbert et al., 2013; Mandegar et al., 2016; Thakore et al., 2015) see also Figure 4C below). To enhance CRISPRi for large-scale screening, we have improved upon the design of CRISPRi sgRNA libraries to optimize on-target activity while further minimizing-off target effects, enabling highly sensitive detection of essential coding genes (Horlbeck et al., 2016).

Here, we developed CRISPRi libraries targeting 16,401 lncRNA loci (with 10 sgRNAs per TSS), and conducted screens for genes that are required for robust growth in 7 human cell types—6 transformed cell lines and induced pluripotent stem cells (iPSCs)(Takahashi et al., 2007). These large-scale screens, coupled with extensive validation studies, greatly increased the number of lncRNA genes known to have biological function and revealed lncRNA function to be highly cell type-specific. Our studies thus help elucidate the biology contained within the lncRNA genome, and provide a tool for both large-scale and targeted investigations of lncRNA function.

Results

CRISPRi screens identify lncRNA loci that modify cell growth

We first designed an sgRNA library to enable genome-scale CRISPRi screening of lncRNA gene function. We generated a comprehensive lncRNA gene set by merging three major non-coding transcriptome annotations (Cabili et al., 2011; Iyer et al., 2015; Yates et al., 2016), prioritized ~1/3 of these genes based on expression in any of a panel of cancer and non-transformed cell lines (Table S1), and designed 10 sgRNAs targeting each lncRNA transcription start site (TSS) using the hCRISPRi-v2.1 algorithm (Horlbeck et al., 2016) (Figure 1A and S1). The cell lines represent a broad range of cell types studied by the ENCODE project (Consortium, 2004), including a chronic myeloid leukemia cell line (K562), the cervical cancer line HeLa, a glioblastoma line (U87), and two mammary adenocarcinoma lines (MCF7 and MDA-MB-231). We also chose an iPSC line that inducibly expresses CRISPRi components (Kreitzer et al., 2013; Mandegar et al., 2016). The library, termed "CRiNCL" for CRISPRi Non-Coding Library, is available as pooled lentiviral plasmid libraries on Addgene and *in silico* as Table S2.

We used this library to conduct screens for lncRNA loci that increase or decrease cell growth in each of 7 cell lines. We infected the full lentiviral library or targeted sublibraries (Figure S2A) into each cell line engineered to express dCas9-KRAB (Gilbert et al., 2013; 2014; Liu et al., 2016; Mandegar et al., 2016), selected for infected cells by puromycin selection, and cultured for between 12 and 20 days, measuring sgRNA enrichment by Illumina sequencing (Figure 1B and Table S3). The fraction of cells infected with the sgRNA library remained stable over the course of the screen (Gilbert et al., 2014), indicating that CRISPRi targeting of lncRNA loci does not exhibit non-specific toxicity (Figure S2B). To facilitate comparisons between screens conducted for different durations and in cell lines with different growth rates, we

normalized sgRNA enrichment by total cell doublings to obtain the quantitative growth phenotype γ, which reflects the positive or negative impact on cell growth caused by knockdown of a given gene (Kampmann et al., 2013) (Figure 1B).

Analysis of biological replicates revealed that the γ for targeting sgRNAs showed strong and reproducible phenotypes (Pearson r = 0.34-0.90) while non-targeting control sgRNAs were tightly distributed around 0 (Figure 1C and S2C, Table S3). We averaged replicate sgRNA phenotypes and used these to score lncRNA genes (Gilbert et al., 2014; Horlbeck et al., 2016), calculating gene phenotypes from the mean of the top 3 sgRNAs targeting the gene and Mann-Whitney p-values from all 10 sgRNAs compared to non-targeting control sgRNAs (Figure 1D and S3A, Table S4). Within each screen, we also randomly sampled non-targeting sgRNA phenotypes to generate "negative control genes" and analyzed them as with lncRNA genes (see Methods), enabling us to estimate an empirical false discovery rate for each screen as well as the combined screen dataset (Figure S2D). LncRNA genes were considered to be hits if their combined phenotype effect size and p-value (referred to here as "screen score") exceeded a consistent threshold applied to each screen corresponding to an empirical false discovery rate of 5% (Figure S3C). Overall, we found between 28 and 438 lncRNA loci hits in each cell line (Figure 1E and S3A, Table S4).

We observed that for 169 of these lncRNA hits, the TSS of the non-coding gene was within 1kb of the TSS of a coding gene previously found to be essential in a CRISPRi screen (Gilbert et al., 2014), making it difficult to determine whether the observed phenotypes were due to knockdown of the target lncRNA or direct inhibition of the neighboring coding gene (Figure S3B). We thus removed these hits from the total set of hit genes for downstream analyses (Figure 1E, S3A, S3D), resulting in 169 "neighbor hits" and 499 "lncRNA hits," 299 of which are distal

from any protein coding gene (~90% of which would not measurably impact growth upon knockdown). The 1kb threshold was chosen based on the maximum distance at which CRISPRi is effective as revealed by analysis of dense sgRNA tiling and genome-scale screens (Figure S4) (Gilbert et al., 2014); increasing this threshold to 10kb classifies only an additional 19 genes as neighbor hits (Figure S3D).

A larger fraction of lncRNAs hits were observed in the iPSC screen, suggesting that this cell line is either more susceptible to growth perturbations or that iPSCs were differentiating to other cell types with lower growth rates. We therefore investigated iPSC differentiation in a secondary fluorescence activated cell sorting (FACS)-based screen by assessing loss of pluripotency as indicated by decreased *POU5F1/OCT4* expression. CRISPRi targeting of only 9 lncRNA loci reduced *POU5F1/OCT4* expression (Figure S5, Tables S5-6), suggesting that the majority of lncRNA hits identified in iPSCs primarily affect cell growth. To confirm that the increased fraction of lncRNA hits in iPSCs was not due to technical differences in CRISPRi function between cell lines, we performed a CRISPRi screen for protein-coding genes required for cell growth in iPSCs (Figure S6A, Table S7). These results corresponded well with our previously published K562 growth screen (Horlbeck et al., 2016) in both the number of genes found to have function and in the ability to specifically identify known essential genes (Figure S6B-C) (Hart et al., 2014). Taken together, our screens identified 499 lncRNA genes that modify cell growth and have no essential coding gene neighbors, representing a large set of unstudied non-protein-coding genes serving important functions in cell biology.

lncRNA CRISPRi phenotypes are reproducible with robust knockdown

Extensive validation studies argue for the low false-positive and -negative rates of our studies. First, we individually cloned the top two sgRNAs targeting 65 representative lncRNA hit loci, 41 of which were hits in only one cell line. We tested whether the observed phenotypes from the screens were reproducible using internally-controlled growth assays, in which the fraction of cells infected with an sgRNA were measured over time by flow cytometry. We monitored the growth effects of sgRNAs in the cell lines in which they exhibited a phenotype in the screen, as well as several sgRNAs in cell lines where they showed no effect, and found that the individual sgRNA growth phenotypes (γ) correlated well with the screen γ (Pearson r = 0.72, Figure 2A). This confirmed both that lncRNA knockdown phenotypes were reproducible and that the difference in lncRNA phenotype between cell lines was not due to technical differences between genome-scale screens. Analyzing these phenotypes over time further revealed distinct kinetics of cell depletion mediated by lncRNA knockdown (Figure 2B). For 12 lncRNA hits, we measured the levels of knockdown by qPCR and found over 70-95% knockdown for most of the targeted transcripts (14/14 sgRNAs in U87; 10/16 sgRNAs in MCF7) despite the effect of cellular depletion (Figure S7A).

In four cell lines, knockdown of lncRNA *PVT1* had a pro-growth phenotype. As *PVT1* had previously been characterized as a proto-oncogene (Tseng et al., 2014) and pro-growth phenotypes in cancer cell lines are uncommon (Gilbert et al., 2014; Wang et al., 2015), we validated the pro-growth phenotype (Figure 2C and S7A) and investigated this complex locus further by conducting a CRISPRi screen in U87 cells with an sgRNA library tiling every possible site along the locus (17,469 sgRNAs). We found that only sgRNAs within 1kb of the most upstream TSSs, which is distal to any mapped enhancers, caused a consistent pro-growth

phenotype (Figure 2D and S7B, Table S8). Within this TSS region, the majority of sgRNAs promoted cell growth, and knockdown of the major isoform was confirmed by qPCR (Figure S7A). sgRNAs outside of this 1 kb window around the TSS, which would not be expected to affect transcription of the major *PVT1* isoform (Gilbert et al., 2014), showed no consistent impact on growth, arguing that the observed pro-growth phenotype is mediated by transcriptional interference.

Repression of lncRNA loci elicits lncRNA-specific transcriptome responses

To better understand the consequences of lncRNA CRISPRi, we performed RNA-seq following CRISPRi knockdown of 42 lncRNA hits in 3 cell types. 32 of these lncRNA loci were hits in only one cell type. Selected lncRNA loci did not have essential coding gene neighbors, and 2 or more sgRNAs per gene were tested individually. Distinct sgRNAs targeting the same lncRNA TSS resulted in highly correlated transcriptome responses (mean Pearson r = 0.980; Figure 2E) that were generally proximal to each other in hierarchical clustering analysis (Figure S8A-D). By contrast, pairs of sgRNAs targeting different hit lncRNA loci with the same phenotype direction had transcriptome responses that were more dissimilar (mean Pearson r = 0.942, Mann-Whitney p-value compared to same-gene pairs = 6.4×10^{-08};), suggesting distinct molecular mechanisms of the lncRNAs despite having similar phenotypes (Figure 2E).

RNA-seq analysis of differential gene expression also revealed several clusters of co-expressed genes, suggesting that growth modifier lncRNA loci regulate critical pathways (Figure S8A-D and Table S9). For instance, 2 lncRNA knockdowns that caused increased growth in U87 cells clustered by upregulation of translation genes ($p = 3.2 \times 10^{-37}$), while other pro-growth sgRNAs showed correlated changes in expression of DNA replication ($p = 2.0 \times 10^{-10}$) and post-

transcriptional regulation ($p = 3.0x10^{-08}$). Clusters enriched for genes in the p53 pathway (e.g. *ATF3*) were upregulated by many anti-growth sgRNAs in both U87 and HeLa cells. Interestingly, K562 cells showed clusters of genes enriched for platelet degranulation ($p = 1.6x10^{-05}$) and response to decreasing oxygen levels ($p = 5.0x10^{-05}$). The median magnitude of $\log_2$ fold changes for differentially expressed genes in U87, HeLa, and K562 were 0.67, 0.86, and 1.17, respectively (Figure S8E), with several genes exhibiting > 2 fold up- or down-regulation consistently across many samples (Figure S8F). These results indicate that different lncRNAs can regulate distinct biological pathways that affect cell growth and proliferation.

Analysis of the chromosomal location of differentially expressed genes did not reveal a global trend toward transcriptional changes on the targeted chromosome (Figure S9). We did however find that knockdown of 14 lncRNA loci resulted in local transcriptional changes within a 20 gene window (Figure S10), suggesting that certain lncRNAs may preferentially act locally.

CRISPRi robustly inhibits lncRNA transcription

The fraction of growth modifier lncRNA loci identified in our screens (1-8% per cell line) was less than the fraction of essential protein-coding genes in previous reports (10-11%) (Horlbeck et al., 2016; Wang et al., 2015). We therefore wanted to assess whether lncRNA genes that did not appear as a hit in any screen were true negatives or simply a result of ineffective repression by CRISPRi. To this end, using all 10 sgRNAs per gene, we measured the knockdown of five arbitrarily selected lncRNA genes that had no observed phenotype in any cells and were expressed in both K562 and U87 cells (Figure 2F and S7C). Of these 100 knockdown measurements, 61 showed over 90% repression of the targeted lncRNA. Furthermore, with the exception of *LOC100506710* in U87 cells, all lncRNAs were repressed by at least 90% by at

least three different sgRNAs. For all sgRNAs, lncRNA knockdown efficiency correlated with their predicted CRISPRi activity, and the efficiency of knockdown was highly correlated between K562 and U87 cells (Pearson r = 0.78; Figure 2G). Based on these findings, with the exception of cases where a small amount of residual transcript is sufficient for lncRNA function, we infer that the majority of lncRNA loci that did not appear as a screen hit produce transcripts that are not essential for robust growth of the cell line screened.

Growth modifier lncRNA function is highly cell type-specific

We next determined the number of lncRNA hits that were unique to a specific cell type or common to any combination of two or more of the cell types screened. The vast majority (89.4%) of lncRNA hits were unique to only one cell type, with none being a hit in 5 or more cell types (Figure 3A-C). Even when we restricted this analysis to the 1,329 lncRNAs expressed in all 7 cell types, 82.6% of the lncRNA hits modified growth in only one cell type (Figure 3B). Analysis of cell type-specificity scores based on the Jensen-Shannon distance, which quantifies how closely a given distribution resembles "perfect" specificity (Cabili et al., 2011), revealed that the specificity of lncRNA screen scores was far greater than the specificity of lncRNA expression, for lncRNA hits (Figure 3D). Therefore, differential expression patterns alone are not sufficient to predict functional lncRNAs. Cross comparison of screen score distributions for lncRNAs that scored as hits in each cell type revealed that the threshold used for calling hits did not account for the cell type specificity (Figure 3E, S11D-E). Furthermore, cross-comparison of screen scores between replicates does not support technical variation as the source of the apparent cell type-specific function (Figure 3F and S11F).

In contrast to the sparse cell type overlap of lncRNA hits, analysis of published protein coding screens across similar numbers of cell types (Hart et al., 2015; Wang et al., 2015) revealed that the majority (54.8% in (Hart et al., 2015), 67.3% in (Wang et al., 2015)) of essential protein coding genes are hits in 2 or more cell types, with 20.4% and 30.8% being essential to all cell types screened in (Hart et al., 2015) and (Wang et al., 2015), respectively (Figure 3C, S11A-B). In addition, "neighbor hits" (lncRNA loci that are within 1kb of an essential protein coding gene), were more likely to modify growth in multiple cell types, suggesting that CRISPRi targeted to these loci represses the adjacent essential coding gene, at least in some cases (Figure 3C, S11C,E).

Cell type-specific lncRNAs elicit highly divergent phenotypes

We sought to better understand the cell type-specific function of specific lncRNAs. We focused on *LINC00263*, which despite being expressed in all 7 cell lines screened, had a much stronger negative growth phenotype in U87 than in any other cell line (Figure S12A). The abundance of *LINC00263* transcript in a given cell line was also poorly correlated with the corresponding screen phenotype (Pearson r = 0.266). Validating these screen results, in internally controlled growth assays, two distinct sgRNAs to the TSS of *LINC00263* reduced the propagation of only U87 cells and not K562, MCF7 or HeLa cells (Figure 4A). H3K9me3 is a chromatin modification that is a result of local dCas9-KRAB activity (Thakore et al., 2015), and in both U87 and HeLa cells with *LINC00263* CRISPRi targeting, ChIP-seq analysis demonstrated equal enrichment of H3K9me3 specifically at the *LINC00263* promoter for two independent sgRNAs (Figure 4B,C, S12B,C). However, despite such evidence of equivalent and specific CRISPRi targeting, U87 and HeLa cells had substantially different transcriptome changes after *LINC00263*

knockdown. While U87 cells upregulated genes related to ER stress (e.g. *ATF4*, *CHAC1*; GO term $p = 4.51x10^{-09}$) and apoptosis (e.g. *DDIT3*, *SOD2*; GO term $p = 3.39x10^{-08}$), only *LINC00263* itself was differentially expressed in HeLa cells (adj. $p < 0.05$; Figure 4D). In K562 cells, these same 2 sgRNAs also produced very little transcriptional change (Figure S12D). Of note, in all three cell lines, knockdown efficiency of *LINC00263* was equivalent (Figure 4D, S12D). Consistent with our observations for *LINC00263*, knockdown of *PVT1* and *LINC00909,* which were hits in U87 but not in HeLa, produced many more differentially expressed genes in U87 (Figure S12E). By contrast, depletion of *LINC00680,* which was a hit in both U87 and HeLa cells, resulted in comparable numbers of differentially expressed genes in U87 and HeLa cells (Figure S12E). Our results suggest that the specificity of lncRNA function is not due to differences in CRISPRi activity, but is related to differences in transcriptional networks across cell types.

We then targeted the *LINC00263* lncRNA transcript with antisense oligonucleotides (ASOs) that degrade RNA via an RNaseH-based mechanism. In both U87 and HeLa cells, ASOs reduced *LINC00263* transcript levels by 85-95% (Figure 4E). However, *LINC00263* ASOs decreased proliferation in U87 cells but not in HeLa cells (Figure 4F,G). The magnitude of proliferation decrease was also comparable to CRISPRi (Figure S12F,G), further supporting the cell type-specific function of this lncRNA. ASO knockdown of three other U87 lncRNA hits also reduced cell proliferation (Figure S12H,I), providing additional evidence for the functional contribution of the lncRNA molecule in these examples.

Insights to molecular mechanism of *LINC00263*

To begin to investigate possible mechanisms of *LINC00263*, we looked for evidence of *cis* gene interaction between *LINC00263* and its closest gene neighbor, *SCD*, whose locus is 10kb upstream of *LINC00263* and is transcribed in the same genomic orientation as *LINC00263* (Figure 4B tracks). *SCD*, or stearoylCoA desaturase-1, is the primary enzyme for catalyzing the transformation of saturated fatty acids to monounsaturated fatty acids, a type of lipid conducive to oncogenic processes such as proliferation and metastasis (Igal, 2011). Analysis of RNA-seq data from TCGA revealed positive correlations between *LINC00263* and *SCD* expression levels in glioblastoma (GBM pearson r = 0.429), low grade glioma (LGG pearson r = 0.814), and breast cancer (BRCA pearson r = 0.526) (Figure 4H). CRISPRi knockdown of *LINC00263* also resulted in decreased *SCD* expression in U87 cells (Figure 4D). Although the fold changes were not statistically significant and were therefore not evidence in differential expression analyses in DESeq2, *SCD* RNA levels did modestly decrease in HeLa and K562 cells upon *LINC00263* knockdown (Figure 4I). Conversely, CRISPRi knockdown of *SCD* resulted in modest depletion of *LINC00263* RNA levels (Figure 4I). This "crosstalk" between gene neighbors is consistent with previously reported local gene regulation, whereby transcriptional repression of one gene negatively affects the levels of adjacent genes, potentially through shared use of chromatin regulators and transcription factors (Engreitz et al., 2016b). Global transcriptome analysis of *LINC00263* knockdown compared to *SCD* knockdown revealed a correlated cell death and cell stress response in U87 cells, however this response was only observed in *SCD* knockdown in HeLa cells (Figure 4J).

To test whether *LINC00263* acts solely as a *cis* regulator of *SCD* expression, we performed relative growth assays following CRISPRi knockdown of *LINC00263* and *SCD*,

individually and in combination, in U87 and HeLa cells. While *LINC00263* knockdown exhibited a negative phenotype in U87 only, *SCD* knockdown resulted in negative growth in both U87 and HeLa, consistent with its essential function (Figure 4K). Interestingly, double knockdown of *LINC00263* and *SCD* resulted in a similar phenotype as *SCD* depletion alone in U87, but exhibited a synergistic effect in HeLa cells. These results suggest that *LINC00263* either has significant *cis* regulation of *SCD* specifically in the context of U87, or that the lncRNA has functions beyond that of a *cis* gene regulator.

Machine learning identifies features predictive of growth modifier lncRNAs

Using data from our genome-scale screens, we sought to identify properties of the lncRNA hits that can distinguish them from non-hit lncRNAs. 18 classes of genomic data such as enhancer maps, expression levels, chromosomal looping data, conservation, and copy number variation from ENCODE (Consortium, 2004), FANTOM (Andersson et al., 2014), Vista (Visel et al., 2007), and other sources (Chen et al., 2016; Hnisz et al., 2013; Yan et al., 2015) were compared with all lncRNA loci screened in this study. 8 of these properties (expression, Pol2/CTCF looping by ChIA-PET, enhancers and super enhancers from (Hnisz et al., 2013), copy number variation) were cell type-dependent. Generalized linear models were constructed to assess which genomic properties are predictive of lncRNA function (see Methods). Expression levels within each cell line, lncRNA gene body within 1kb of a mapped FANTOM Enhancer, lncRNA gene body within 5kb of a cancer-associated single nucleotide polymorphism (SNP) (Yan et al., 2015), and the number of exons were significant predictors of lncRNAs hits ($p < 0.01$) in repeated 10-fold cross validation (Figure 5, Table S10). 99.6% of lncRNA genes that were screened but not apparently expressed were not called as hits (Figure 5C). Whether the 11

growth modifier hits of such "non-expressed" lncRNA loci represent non-lncRNA mediated effects, inaccurate quantitation of the transcript levels, or effects mediated by lncRNAs acting at low expression remains to be determined. In support of the latter possibility, *HOTTIP* has been reported to function despite being expressed at ~0.3 copies per cell (Wang et al., 2011). Nonetheless, many highly expressed lncRNAs were not hits (*e.g.* 154 non-hit lncRNAs were detected at FPKM > 100), and the accuracy for predicting lncRNA hits was greater for a model using all variables as compared to a model that relied only on expression levels (Figure 5B).

Compared to non-hit lncRNAs, hit lncRNA gene bodies were 1.66 times more likely to be within 1kb of a mapped enhancer (Figure 5D). This represented 127 of the lncRNA hit loci identified in our screens. However, the FANTOM enhancer annotations used for our analyses were derived from hundreds of different cell types, and thus only a fraction of these enhancers are active in any given cell type in our screen (Andersson et al., 2014; Visel et al., 2007). Hit loci were also 1.4 times more likely to be within 5kb of a cancer-associated SNP (Figure 5E). That our hits were enriched for multi-exonic lncRNAs is consistent with the concept that lncRNA splicing can be an aspect of lncRNA function (Engreitz et al., 2016a) (Figure 5F). However, the explanatory power of exon number was relatively low, and our screen did identify several single exon hits such as *NEAT1*. However, no genomic property analyzed, alone or in aggregate, fully predicted growth modifier lncRNAs in a given cell type, underscoring the importance of performing loss-of-function screens for defining sets of functional genes.

Cell type specific chromatin structure identifies lncRNA-oncogene fusion in HeLa cells

Cell type-specific super enhancers, which are highly active loci of transcription that can act as master regulators of cell fate (Hnisz et al., 2013; Whyte et al., 2013), were enriched near the loci

of growth modifier lncRNAs (Figure 6A), albeit at a less significant level ($p = 0.0165$) as compared to FANTOM enhancers ($p = 6.44 \times 10^{-09}$) in our prediction model. LncRNA *CCAT1* is transcribed near enhancers defined by FANTOM and is also overlapped by a HeLa-specific super enhancer (Figure 6B). *CCAT1*, which is overexpressed in cancer (McCleland et al., 2016; Xiang et al., 2014), was a highly significant hit that was specific to HeLa cells (Figure 6C). Circularized chromosome conformation capture sequencing (4C-seq), revealed that the *CCAT1* locus forms a highly specific distal loop with the *MYC* locus in HeLa cells, and this looping interaction was absent in U87 cells (Figure 6B,D-E). Consistent with its negative growth phenotype, *CCAT1* CRISPRi knockdown also resulted in decreased *MYC* expression (Figure S10, upper left panel). Genome wide, the only other gene loci that exhibited significant differential contact with the *MYC* locus were *POU5F1B* and *CASC8*, a region containing a known *MYC* enhancer (Pomerantz et al., 2009), although interaction between *MYC* and *CCAT1* was much stronger. Mining of published Hi-C data also revealed that the interaction between *CCAT1* and *MYC* was highest in HeLa cells among 8 cell types previously analyzed (Ma et al., 2015; Rao et al., 2014) (Figure 6F). Previous studies indicate that ASO-mediated degradation of the *CCAT1* lncRNA transcript diminishes the looping interaction between the *CCAT1* locus and *MYC* (Xiang et al., 2014), suggesting *CCAT1*'s role in higher order chromatin contacts.

CCAT1 was also remarkable because its phenotype upon knockdown was more severe than that of any other gene targeted in this study, regardless of cell type, and we sought to determine why this was the case. Upon closer inspection of the haplotype-resolved HeLa genome sequence, the *CCAT1* locus spanned the site of integration for human papilloma virus (HPV)18, which is thought to be the event that initiated tumorigenesis in HeLa cells (Adey et al., 2013). Consistent with this observation, we detected RNA-seq reads supporting the presence of a fusion

RNA transcript joining *CCAT1* and the E6/E7 oncogenes of HPV18 at coordinates chr8 128,241,374 (human hg19) and 928 (HPV18 NC_001357.1) (Figure 6G). These reads mapping to fusion transcripts were virtually depleted upon CRISPRi targeting of the *CCAT1* TSS (Figure 6H). CRISPRi against *CCAT1* also resulted in ~80-90% depletion of total E6, E7, and E1 RNA (Figure 6I), a degree of knockdown equivalent to that of *CCAT1* itself in these experiments. Importantly, cells with complete *CCAT1* knockdown are likely not represented in these experiments, due to the anti-proliferative phenotype observed upon knockdown. Therefore, the vast majority of these integrated HPV oncogenes are likely transcribed as fusion transcripts with *CCAT1*. Broadly, our results underscore the importance of studying cellular context—whether chromatin contacts, fusion transcripts, or other undetermined properties—for determining the cell type-specific function of lncRNAs.

Discussion

By employing CRISPRi for systematic, large-scale screens for lncRNA function in multiple cell lines, we identified 499 lncRNA loci that are required for robust cell growth. This work increases considerably the number of known functional lncRNAs and revealed that the large majority (89%) of lncRNA genes identified modified growth in just one cell type. Studies of the protein-coding genome with similar large-scale screening efforts showed that an essential gene in one cell type is highly likely to be essential in the other cell types tested (Hart et al., 2015; Wang et al., 2015). In contrast to protein coding genes, of the 1,329 lncRNA genes expressed in all of the seven different cell lines tested, not one lncRNA gene was required for robust cell growth in all cell types, with the large majority of lncRNA gene hits being specific to just one cell line. Our results thus reveal a critical role of cellular context in determining lncRNA function.

Several clues to this specificity of lncRNA function emerge from our analyses. First, although cell type-specific expression of lncRNAs was the strongest predictor of lncRNA hits in our machine learning model (Figure 5A,C), it did not fully explain this functional specificity (Figure 3, 5B). For example, RNA-seq analysis points to *LINC00263* playing a role in a complex transcriptional network required for U87 cells, but that despite being expressed in other cell types, *LINC00263* appears dispensable for the normal expression of nearly all genes in these other cells (Figure 4D and S12D,E). Further dissection of the *LINC00263* locus through knockdown of its adjacent gene *SCD* revealed some evidence of *cis* regulatory function for this lncRNA, but this interplay might only occur in certain contexts, such as within the nucleus of U87 cells.

Taking advantage of the scale of our dataset, we have also begun to discover genomic features that predict growth modifying function. Our finding that enhancer proximity and

chromosome contacts correlate with lncRNA function suggests that higher-order chromatin structure can play a role in such specificity of lncRNA function (Li et al., 2013; Ørom et al., 2010) (Fulco et al., 2016). The extent to which cell type-specific function of enhancer-templated lncRNAs results from repression of the transcript itself or its genomic locus remains an important open question. In any case, the association of lncRNA function with higher order chromatin structure is consistent with the emerging view that chromosomal looping between lncRNA promoters and target genes differs between cell types (Ma et al., 2015) and is critical to lncRNA function (Engreitz et al., 2013). Finally, our finding that genomic regions containing growth modifier lncRNAs are enriched for cancer risk SNPs suggests that these lncRNAs may contribute to the pathogenesis of cancer.

The observation that *CCAT1*, the lncRNA with the most severe growth phenotype upon knockdown, was fused to the E6/E7 oncogenes of the HPV18 viral genome was surprising for several reasons. Because the HPV integration event is not evident in the human reference genome, the targeting of these oncogenes was purely serendipitous, and something that resulted only from systematic functional perturbation of lncRNAs. The existence of this fusion transcript between *CCAT1* and HPV18 is also supported by single cell RNA-seq experiments of HeLa cells (Wu et al., 2015). Although RNA interference-based strategies to target HPV oncogenes have been previously performed (Jiang and Milner, 2002), our results represent a novel approach to silencing viral oncogenes. Nonetheless, it is important to note that *CCAT1* has been shown to function in chromatin looping in cells that are not known to be initiated by viral oncogenes, such as colon cancer (Xiang et al., 2014), suggesting that this lncRNA can function independently of this unique fusion event. These remarkable results motivate further study of lncRNA loci that

harbor viral sequences, beginning with a comprehensive survey of putative lncRNA-oncogene fusions in cancer that typically go undetected due to the use of standard genome references.

Regardless of the mechanism(s) of the observed cell type-specificity of lncRNAs, these findings have implications for understanding the biological roles of lncRNAs. LncRNAs appear to have originated much later than protein coding genes, consistent with their not playing generic housekeeping roles (Necsulea et al., 2014; Ulitsky and Bartel, 2013). Our study, which focused on lncRNAs required for robust cell growth, underestimates the true number of functional lncRNAs in these cell types, as lncRNAs have been shown to regulate more evolutionarily complex cellular decisions such as cell fate (Kretz et al., 2013; Lin et al., 2014; Ramos et al., 2015; Sauvageau et al., 2013), cancer metastasis (Gupta et al., 2011; Gutschner et al., 2013), and perhaps neuronal function (Briggs et al., 2015). The CRISPRi tools developed here can now be applied to the study of such higher order cellular processes, where lncRNAs might exhibit even greater richness of function. Finally, the exquisite cell type-specificity of lncRNA gene function has clear implications for targeted therapy.

Materials and Methods

lncRNA CRISPRi library design

lncRNA target selection

LncRNA annotations were retrieved from Ensembl build 75 (using the biotypes lincRNA, antisense, 3 prime overlapping ncRNA, processed transcript, sense intronic, sense overlapping) (*39*), the Broad human lincRNA catalog (*37*), the MiTranscriptome (*38*) , and a set of human brain specific lncRNAs (*42*). Annotations were merged using the cuffmerge command in Cufflinks v2.2.1 (*62*).

LncRNAs that were transcribed in at least one of the 7 cell lines in this study were identified by quantifying the expression of lncRNAs using RNA-seq data. RNA-seq data were obtained from ENCODE and other sources: HEK293T (GSE56010), HeLa (GSE30567, GSE33480, GSE23316), K562 (GSE30567, GSE33480, GSE23316), MCF7 (GSE30567, GSE33480), MDAMB231 (GSE73526, GSE45732), iPS (clone PCBC15hsi2012040401 (*63*)), HFF (GSE69906). RNA-seq was performed in-house for U87 cells, using the illumina TruSeq Stranded mRNA kit. Reads were quality trimmed using seqtk v1.0 and aligned to the human genome (GRCh37) with tophat v2.0.10, using the merged transcriptome reference as the transcriptome index, the prefilter-multihits flag, and strand specific flag when appropriate. Transcript abundance estimation was performed using Cufflinks v2.2.1. For each gene, the median FPKM value of the replicate samples were obtained. For each cell line, a minimum expression threshold was set between 0.25-0.50 FPKM in order to screen as many genes as possible, given cell culture scale limitations. 21,578 lncRNAs were identified.

Generating lncRNA TSS annotations

From these 21,578 transcripts passing the expression filter, an initial set of 17,740 TSSs were obtained from transcripts belonging to the same gene and with 5' ends within 100bp of each other. These TSS annotations were further refined using the FANTOM cap analysis of gene expression (CAGE)-based TSS annotations as previously described (*35*) with adjustments. LncRNA TSS annotations could not be directly matched to FANTOM "p1@gene" CAGE peaks, and instead were matched to any same-stranded CAGE peak within 400bp labeled as "p1" or "p2," and annotation support was labeled as "CAGE, primary peaks." If no primary peaks were found, annotations could instead be refined by robust or permissive peaks within 200bp of the starting annotation, and were labeled as "CAGE, robust peak" or "CAGE, permissive peak," respectively. Where no CAGE peaks were found (due to the cell type-specific nature of lncRNA expression only 30% of TSS annotations were refined with CAGE peaks), the TSS as determined by the annotation sets above was used and labeled "Annotation." 66 of the original TSSs were assigned the same start site by this method, reducing the total number targeted to 17,674. This annotation is included as Table S1. As detailed below, a further 692 TSSs could not be uniquely targeted, reducing the total TSSs to 16,982. Finally, to avoid redundant information from different TSSs located in close proximity, TSSs within 100bp of each other were assigned to a single gene ID (designated LHnnn in Tables S1-6,8) for a total of 16,401 distinct lncRNA target loci.

sgRNA selection

All potential sgRNAs within 25bp upstream and 500bp downstream of the refined lncRNA TSS annotations were scored for predicted activity using the hCRISPRi-v2.1 algorithm, scored for

off-target sites near TSSs and in the genome using weighted Bowtie v1.0.0 (*64*), and filtered for required restriction sites (BstXI, BlpI, and SbfI) and overlap with higher-ranking sgRNAs as previously described (*35*). For 692 TSSs, 10 sgRNAs passing all filters could not be found and were discarded. 87.5% of sgRNAs accepted into the library passed the highest off-target stringency threshold, while 10.9% passed at the second-highest stringency. Non-targeting control sgRNAs were generated randomly weighted by the per-base nucleotide frequencies of the targeting sgRNAs in the library, and filtered for no target sites in the genome. sgRNAs targeting lncRNA genes specifically expressed in a subset of cell types were assigned to the appropriate hierarchical sublibrary (along with a proportional number of non-targeting controls) to enable screening only the desired gene sets (Figure S2A). Sublibraries were designed as the intersection of genes expressed in the cell lines indicated in Figure S2A, and then the full set of genes for a given cell line could be generated by combining sublibraries as follows:

iPSC = Common + (iPSC, HFF) + iPSC

HFF (not screened in this study) = Common + (iPSC, HFF) + iPSC

U87 = Common + Cancer common + (U87, HEK293T) + U87

HEK293T = Common + Cancer common + (U87, HEK293T) + HEK293T

K562* = Common + Cancer common + (K562, HeLa, MCF7) + (K562, HeLa) + K562

HeLa = Common + Cancer common + (K562, HeLa, MCF7) + (K562, HeLa) + HeLa

MCF7/MDA-MB-231 = Common + Cancer common + (K562, HeLa, MCF7) + MCF7

*all 13 sublibraries were screened in K562s to validate the cell line expression sublibrary strategy; see Figure 5C

Oligonucleotide pools were designed with flanking cloning and PCR sites as described, synthesized by Agilent Technologies (Santa Clara, CA), and cloned into the library sgRNA expression vector pCRISPRia-v2 (*23*, *35*).

For the *PVT1* tiling library, all possible sgRNAs from 25bp upstream of the *PVT1* locus to 25bp downstream that passed the second-highest off-target stringency filter and restriction site filters were included. Non-targeting sgRNAs matching this design in base composition and off-target stringency were included, and oligonucleotide pools were synthesized and cloned as above.

CRISPRi screens

Growth screens

Several cell lines expressing dCas9-KRAB were obtained from previous publications: HEK293T(*22*), HeLa(*22*), K562 (*23*), iPSCs (WTC-CRISPRi Gen IC)(*33*), and U87(*42*). MCF7 and MDA-MB-231 were generated for this study by infecting lentivirus expressing dCas9-KRAB-BFP (Addgene #46911; (*22*)) and sorting for single cell clones stably expressing high BFP. Replicates for these cell lines were performed in different clones. All cell lines except iPSCs were infected in duplicate with sgRNA the sublibraries described above or the *PVT1* tiling library, packaged with TransIT-LT1 (Mirus, Madison, WI) transfection in HEK293T cells (not expressing dCas9-KRAB), at an initial infection rate of 30-50% (300-500x coverage of the library). Cells were cultured for two days following infection, treated for two days with 0.75-1.00 μg/mL puromycin, allowed to recover for one day, and then cultured at a minimum coverage of 1000x for 12 days (K562, HEK293T, U87, HeLa) or 20 days (MCF7, MDA-MB-231) starting from this "T0." K562 cells were passaged daily, while adherent cells were split on

alternate days. iPSCs were infected at ~15% infection (with double the starting cell population to yield 300x coverage), grown for 3 days, selected with 1.5 μg/mL puromycin for 9 days, and allowed to recover for 3 days. iPSCs were then divided into two independent replicates and treated daily with 2μM doxycycline starting from this T0 and for the following 18 days (with primary endpoint at 12 days used here unless otherwise specified). Cells with a minimum of 1000x library coverage were harvested the day following puromycin recovery (T0) and at the endpoint, and processed for sequencing on Illumina HiSeq 2500 or 4000 as previously described(*23*, *35*).

Sequencing reads were aligned to the expected CRiNCL library sequences, counted, and quantified using the ScreenProcessing pipeline (https://github.com/mhorlbeck/ScreenProcessing; (*35*)) Negative control genes were generated by randomly sampling (with replacement) ~10 non-targeting sgRNAs per negative control gene to match the true gene TSSs targeted by the library, and then scoring the negative control genes for effect size and Mann-Whitney p-value as was done for the true genes. Genes (LHnnn) with multiple TSSs were collapsed to a single score by selecting the one with the lowest Mann-Whitney p-value.

In order to call hit genes from screens, we defined a "screen score" incorporating both the effect size and the p-values of genes in each screen. The screen score was calculated as | γ z-score from negative control gene distribution | x $-\log_{10}$ p-value, and for all screens a threshold of greater than or equal to 7 was applied to call hits.

Neighbor hits were classified by first calculating the distance between the each lncRNA TSS and the TSS of the closest protein coding gene (TSS-pc distance). LncRNAs whose TSS-pc distance was less than 1000 bp, and whose neighboring protein coding gene was 1) scored as essential in our previous screen in K562 cells (*23*), 2) expressed in the cell type in consideration, and 3) had the same phenotype direction as the lncRNA, were then classified as neighbor hits. Hits that did not meet these criteria were left as lncRNA hits.

OCT4 FACS-based screen

iPSCs were harvested 9 days post-doxycycline addition from the growth screen samples above and fixed with 4% paraformaldehyde for 10 minutes at room temperature followed by PBS wash. 9 days was chosen to balance the longer duration of continuous target gene knockdown with dropout of cells containing sgRNAs conferring a negative growth phenotype. Cells were permeablized with 0.5% saponin (Sigma) in PBS with 4% FBS and 2mM EDTA, stained with 1:100 mouse monoclonal α-*OCT3/4* antibody (sc-5279, Santa Cruz Biotechnology), washed with permeablization buffer, and stained with 1:200 Goat α-Mouse IgG-488 (A11029, Invitrogen) (*33*). Cells in the top and bottom 30% of *OCT4* signal as measured on the FITC-A channel were sorted for purity on a FACS AriaII custom. Sorted cells were then harvested for de-crosslinked genomic DNA using QIAamp DNA Formalin Fixed Paraffin Embedded Tissue kit (QIAGEN) following manufacturer's instructions but omitting paraffin removal steps, using one column per million sorted cells. Genomic DNA was directly amplified for Illumina sequencing using Q5 DNA polymerase (New England Biolabs) and sequenced on a HiSeq 4000. Sequencing data was analyzed as described above.

A screen for protein-coding genes required for robust growth in iPSCs was performed as with the iPSC lncRNA screen, with an endpoint at 14 days post-doxycycline addition. The screen was performed using the hCRISPRi-v2 H1 Drug Targets, Kinases, and Phosphatases sublibrary with 5 sgRNAs/gene (*35*), and was analyzed as above.

sgRNA Validation

sgRNAs for individual validation were cloned by annealing oligo pairs containing the sgRNA protospacer and flanking BstXI and BlpI cloning sites and ligating the resulting fragment into the sgRNA expression vector pU6-sgRNA EF1Alpha-puro-T2A-BFP (Addgene #60955). Internally controlled growth assays were performed by infecting cells with sgRNA lentiviruses at MOI < 1.0 and measuring the sgRNA+ fraction by BFP using flow cytometry on an LSRII (BD). Double knockdown experiments were performed by cloning sgRNAs onto BFP and GFP plasmids and tracking the double positive population over time. Experiments were performed in biological triplicates from the infection step.

RT-qPCR

Cells were puromycin selected for 4 d (1 μg/mL) and subjected to 1 d recovery. K562 cells were infected for 48 hr, followed by 2 d puromycin treatment (3 μg/mL) and 2 d recovery. RNA was harvested with TRIzol and purified using the Direct-zol MiniPrep RNA purification kits (Zymo Research) with the on-column DNAse digestion step. cDNA were prepared with Transcriptor First Strand cDNA Synthesis Kit (Roche) using the oligo-dT protocol, and RT-qPCR was performed using LightCycler 480 SYBR Green I Master Mix (Roche) on a LightCycler 480

instrument (Roche). Experiments were performed in biological triplicates from the infection step. sgRNA protospacer sequences and RT-qPCR primers are listed in Table S11.

RNA-seq sample preparation and data analysis following CRISPRi

Cells were infected with sgRNA lentiviruses for 48 hr, followed by 4 d puromycin treatment (1 μg/mL) and 1 d recovery. K562 cells were infected for 48 hr, followed by 2 d puromycin treatment (3 μg/mL) and 2 d recovery. RNA was harvested using TRIzol and purified using the Direct-zol MiniPrep RNA purification kits (Zymo Research) with the on-column DNAse digestion step. RNA integrity was confirmed using the Agilent 2200 RNA ScreenTape. RNA-seq libraries were generated using TruSeq HT Stranded mRNA kit according to manufacturer's protocol (illumina). cDNA was validated using the Agilent 2200 DNA 1000 ScreenTape, Qubit 2.0 Fluorometer (Life Technologies), and ddPCR (Bio-Rad). Cluster generation and sequencing was performed on a HiSeq 4000, using the single end 50 read protocol. Reads were aligned to the human genome (GRCh37) using the spliced read aligner HISAT2 v2.0.3 (*65*) against an index containing SNP and transcript information (*genome_snp_tran*). Quantification of Ensembl build 75 genes was carried out with featureCounts (*66*) using only uniquely mapped reads.

Library complexity was calculated by counting the number of genes with greater than 2 reads, and knockdown efficiency was calculated by normalizing gene Transcripts per Million (TPM) for the experimental samples with the mean TPM of the control knockdown samples. Samples with fewer than 11,000 genes detected and weaker than 40% lncRNA knockdown were filtered. Pairwise Pearson correlations between RNA-seq samples were obtained using the sets of genes exhibiting significant variation within each cell type using the likelihood ratio test in DESeq2

(*67*) with an adjusted p value threshold of 0.001. Differential expression analysis for individual lncRNA knockdowns was performed using the Wald test in DESeq2 with an adjusted p value threshold of 0.05, using unique sgRNAs against the same lncRNA TSS as biological replicates. For hierarchical clustering of co-expressed genes across multiple samples, we first grouped cells by cell type and then by the direction of phenotype of the sgRNA. Within each subgroup, we obtained the set of variable genes using the likelihood ratio test in DESeq2 (*67*) with an adjusted p value threshold of 0.001. These genes were then used for complete linkage hierarchical clustering using Pearson correlation coefficients as the distance metric. Sequencing data are deposited in GSE85011.

Analysis of lncRNA-oncogene fusion transcripts was performed by aligning RNA-seq reads to a custom genome annotation that includes all chromosomes of hg19 and also the HPV18 genome (NC_001357.1). tophat-fusion v 2.1.0 was used with the flags (--bowtie1 --library-type fr-firststrand --fusion-search --no-coverage-search). Reads spanning the fusion junction were reported, and fold changes were calculated by normalizing by library size and then by the mean of the negative control samples.

TCGA data analysis was performed on the TANRIC web server (Li et al., 2015).

ChIP-seq sample preparation and data analysis

Cells were infected with sgRNA lentiviruses for 48 hr, followed by 4 d puromycin treatment (1 ug/mL) and 1 d recovery. Genome-wide histone modifications were determined by ChIP against H3K9me3 (Abcam ab8898) on 5 million cells as described in (*68*). Cells were cross-linked by

adding 37% formaldehyde to a final concentration of 1% into culture medium and gently shaking for 10 min at room temperature. Reaction was quenched with glycine, and cells were then washed twice with ice-cold PBS containing protease inhibitors (1mM PMSF, 1X Roche cOmplete EDTA-free cocktail). Cells were scraped off of the plate using a cell lifter and pelleted for 5 min at 2,000 rpm at 4°C. Pellet was snap-frozen in liquid nitrogen and stored at −80°C. Pellet was then thawed and resuspended in Cell Lysis Buffer (5 mM PIPES pH 8, 85 mM KCl, freshly added 1% IGEPAL) with protease inhibitors (Pierce Halt Protease Inhibitor Cocktail). Cells were then homogenized using a type B glass dounce homogenizer, pelleted, and resuspended in Nuclei Lysis Buffer (50 mM Tris-HCl pH 8, 10 mM EDTA, 1% SDS). Chromatin was sonicated in Diagenode TPX tubes using the Diagenode Bioruptor for 20 cycles and DNA was ranged from 150–700 bps as determined by gel electrophoresis. Debris was pelleted and discarded, and an aliquot was removed for Input DNA sequencing from the sonicated chromatin within the supernatant. Sonicated chromatin was then diluted 5-fold in IP Dilution Buffer (50 mM Tris–HCl pH 7.4, 150 mM NaCl, 1% IGEPAL, 0.25% deoxycholic acid, 1 mM EDTA pH 8) with protease inhibitors and pre-cleared with Life Technologies Protein G Dynabeads for 2 hr at 4°C. 5 µg of antibody was added per million cells, and samples were incubated overnight at 4°C. Antibody-bound chromatin was then collected using Life Technologies Protein G Dynabeads and washed twice with IP Dilution Buffer, twice with IP Wash Buffer 2 (100 mM Tris–HCl pH 9, 500 mM LiCl, 1% IGEPAL, 1% deoxycholic acid), and once with IP Wash Buffer 3 (100 mM Tris–HCl pH 9, 500 mM LiCl, 150 mM NaCl, 1% IGEPAL, 1% deoxycholic acid). Precipitated chromatin was then eluted for 30 min at 65°C with Elution Buffer (1% SDS, 50 mM $NaHCO_3$). ChIP and Input DNA crosslinks were reversed by adding 5 M NaCl and heating at 65°C overnight. The following day, 10 mg/ml RNase A was

added to precipitated chromatin, and chromatin was incubated for 30 min at 37°C. DNA was then recovered using Agencourt AMPure XP Beads and quantified using the Life Technologies Qubit Fluorometer.

ChIP DNA was then used for library preparation using the Kapa HyperPlus library preparation kit. 100ng of ChIP DNA was used for end repair and A-tailing. Illumina adapters were then ligated to the end-repair products. The library was amplified for 6 cycles before post-amplification cleanup using SPRI beads. Libraries were then quantified with the Life Technologies Qubit Fluorometer, and library size was confirmed using Agilent TapeStation 2200. ChIP-seq libraries were sequenced on a HiSeq 4000, 50 read single end.

Reads were aligned to the human genome (GRCh37) using bowtie v2.2.8 (*69*). Enrichment at promoter regions, which were defined as +/- 1kb of each TSS and generated from Ensembl GRCh37 build 75, were quantified using featureCounts v1.5.0-p2 (*66*). Signal was visualized using deepTools2 bamCoverage (*70*), normalizing reads to 1x sequencing depth. Differential H3K9me3 enrichment was analyzed using DESeq2 (*67*) , treating distinct sgRNAs against the same lncRNA TSS as replicate samples. Sequencing data are deposited in GSE85011.

4C-seq sample preparation and data analysis

Chromosome conformation capture combined with high-throughput sequencing (4C-seq) was performed according to (Stadhouders et al., 2013) with modifications. Single cell suspensions consisting of 5M cells per sample were prepared by dissociating cells with Trypsin/EDTA, centrifuging at 300g for 5 min, and resuspending in 10% (vol/vol) FBS/PBS. Cells were fixed by

adding 37% formaldehyde to a final concentration of 1% and incubating for 10 min at room temperature while tumbling, followed by addition of 1M glycine to a final concentration of 0.125M to quench. Samples were centrifuged at 300g for 8 min at 4°C, and supernatant was removed. Samples were washed with cold PBS and resuspended in 5 mL ice-cold lysis buffer (10 mM Tris-HCl (pH 8.0), 10 mM NaCl, 0.2% NP-40, 1× complete protease inhibitor (Roche)), followed by 10 min incubation on ice. Cytoplasmic lysis was then confirmed by light microscopy. Samples were centrifuged at 300g for 5 min at 4°C and resuspended in 500 uL 1.2x CutSmart restriction digest buffer (NEB). 20% SDS was added to a final concentration of 0.3%, and samples were incubated for 1 hr at 37°C while shaking at 900 rpm on a thermomixer. 20% Triton X-100 was added to a final concentration of 2%, followed by incubation for 1 hr at 37°C while shaking at 900 rpm on a thermomixer. 600U of NlaIII (NEB) was added to each sample and incubated 20 hr at 37°C while shaking at 900 rpm on a thermomixer. To inactivate enzyme, 20% SDS was added to a final concentration of 1.6% SDS, followed by incubation for 25 min at 65°C while shaking at 900 rpm on a thermomixer. Samples were added to 6.125 mL of 1.15x T4 DNA Ligase buffer (NEB) and 10,000 U of T4 DNA Ligase (NEB), followed by incubation for 18 hr at 16°C. De-crosslinking was performed by adding 300 ug proteinase K (Machery Nagel) and incubating for 16 hr at 65°C. 300 ug RNase (Sigma) was added and incubated for 30 min at 37°C. 7 mL phenol:chloroform was added to each sample and centrifuged at 3,200g for 15 min. Aqueous phase was transferred to 7mL H2O, 1.5 mL 2M NaOAc (pH 5.6), 7 uL glycogen, and 35 mL 100% ethanol. Mixtures were frozen at -80°C for 2 hrs, until solid, and immediately centrifuged at 3,200g for 45 min at 4°C. Pellets were washed with 70% ethanol and centrifuged at 3,200g for 15 min at 4°C. Pellets were air dried and resuspended in 150 uL 10 mM Tris-HCl (pH 7.5). The second 4C-seq digest was then performed by adding 50 U DpnII (NEB), 50 uL 10x

NEBuffer DpnII, and H2O to 500 uL, followed by incubation for 16 hr at 37°C. Digests were then purified using phenol:chloroform and resuspended in 100 uL H2O. Second ligation was performed by transferring samples to 12.5 mL H2O, 1.4 mL 10x T4 DNA Ligase buffer (NEB), and 20,000 U of T4 DNA Ligase (NEB), followed by incubation for 18 hr at 16°C. Circularized fragments were purified by phenol:chloroform and resuspended in 75 uL 10 mM Tris-HCl (pH 7.5), and then purified again using NucleoSpin PCR Clean-up (Machery Nagel), distributing each sample across 3 columns. 1.2 ug of each template was amplified using the Expand Long Template PCR kit (Roche) for 30 cycles. PCR primers used were selected from a genome-wide database generated by (van de Werken et al., 2012): *MYC* viewpoint Forward – TTGCACTTTTCTTGTCCATG; *MYC* viewpoint Reverse – GGGGAAGGAATAAAAGAAAA . PCR amplicons were purified using Agencourt Ampure XP beads. Sequencing libraries were generated using the KAPA HyperPlus kit according to manufacturer's recommendations, with the following modifications: fragmentation was carried out for 20 min, and adapter ligation was performed using 10 uL of 15 uM illumina adapters, incubated for 75 min at 20°C. No further amplification was performed following ligation. 4C-seq libraries were validated using the Agilent 2200 High Sensitivity DNA ScreenTape, Qubit 2.0 Fluorometer (Life Technologies), and ddPCR (Bio-Rad). Cluster generation and sequencing was performed on a HiSeq 4000, using the single end 50 read protocol.

Reads were aligned to the human genome (GRCh37) using the spliced read aligner HISAT2 v2.0.3 (Kim et al., 2015) . 4C-seq signal at promoter regions, which were defined as +/- 1kb of each TSS and generated from Ensembl GRCh37 build 75, were quantified using featureCounts v1.5.0-p2 (Liao et al., 2014). Alignments were visualized using deepTools2 bamCoverage

(Ramírez et al., 2016), normalizing reads to 1x sequencing depth. Genome-wide differential 4C signal between cell types was analyzed using DESeq2 (Anders and Huber, 2010) , treating distinct sgRNAs against the same lncRNA TSS as replicate samples. To filter signals dominated by noise, differential looping calls required a minimum baseline threshold of 1000 reads. Sequencing data are deposited in GSE85011.

Hi-C data mining

Hi-C data at 5kb resolution was obtained from (Rao et al., 2014) (HeLa, HUVEC, NHEK, GM12878, HMEC, IMR90, KMB7, K562) and (Ma et al., 2015) (H1). Visualization was performed using Juicebox (Rao et al., 2014) using the Balanced – Knight-Ruiz algorithm for data normalization. Values for observed vs. expected contacts were determined at chr8:128,230,001 - chr8:128,745,001, corresponding to the *CCAT1* and *MYC* loci, respectively.

Antisense oligonucleotide knockdown and proliferation assay

Antisense locked nucleic acid gapmers were designed against *LINC00263* using the Exiqon web server. Cells were transfected with the specified ASOs including negative control “A” at a final concentration of 50nM using Lipofectamine RNAimax Reagent (Invitrogen) according to the manufacturer's instructions. After 48 hours of transfection cells were seeded in duplicate. In order to maintain gene depletion, cells were transfected for a second time 7 days after the first transfection. Cell counting was performed every 2 days using Countess Automated Cell Counter (Invitrogen).

Flow cytometry for cell cycle analysis

Cells were transfected with the specified ASOs as described above. After 72 hours of transfection cells were pulsed with 33 μM bromodeoxyuridine (BrdU) for 20 min, and afterwards fixed in 70% ethanol. Subsequently cells were stained with primary anti-BrdU antibody (Clone B44; BD Biosciences) for 1 h, followed by 1 h incubation with Alexa Fluor 488 anti-mouse IgG (Invitrogen). DNA was counterstained by 0.1 mg/ml propidium iodide supplemented with RNase for 1 h at 37°C. Analysis was performed on a FACSCalibur using CellQuest software (BD). Quantification and analysis of cell-cycle profiles were obtained using FlowJo (Tree Star, Inc).

Machine learning of lncRNA properties

Genomic features were obtained from multiple sources. RNA-seq data were the same as used above for sgRNA library generation. Enhancer maps were obtained from the Fantom 5 Transcribed Enhancer Atlas (*48*), and VISTA Enhancer Browser set of experimentally confirmed human enhancers (*49*). Cell type-specific enhancer and super-enhancer maps for HeLa, U87, K562, and MCF7 cells were obtained from (*51*). LncRNA loci were considered near a (super)enhancer if it overlapped with or was within 1kb of a mapped enhancer. Cancer associated SNPs from the NHGRI GWAS Catalog were obtained from (*50*) and noted if any were within 5 kb of a lncRNA locus. Cell type-specific copy number variation data for HEK293T, HeLa, K562, U87, and MCF7 cells were obtained from ENCODE (GSE40698) and intersected with lncRNA loci. ChIA-pet data for HeLa, K562, and MCF7 cell lines were obtained from ENCODE (GSE39495). LncRNA loci that were overlapped completely by a Pol2 or CTCF loop with a score of at least 400 were noted. LncRNAs with mouse orthologs were identified using Slncky (*52*).

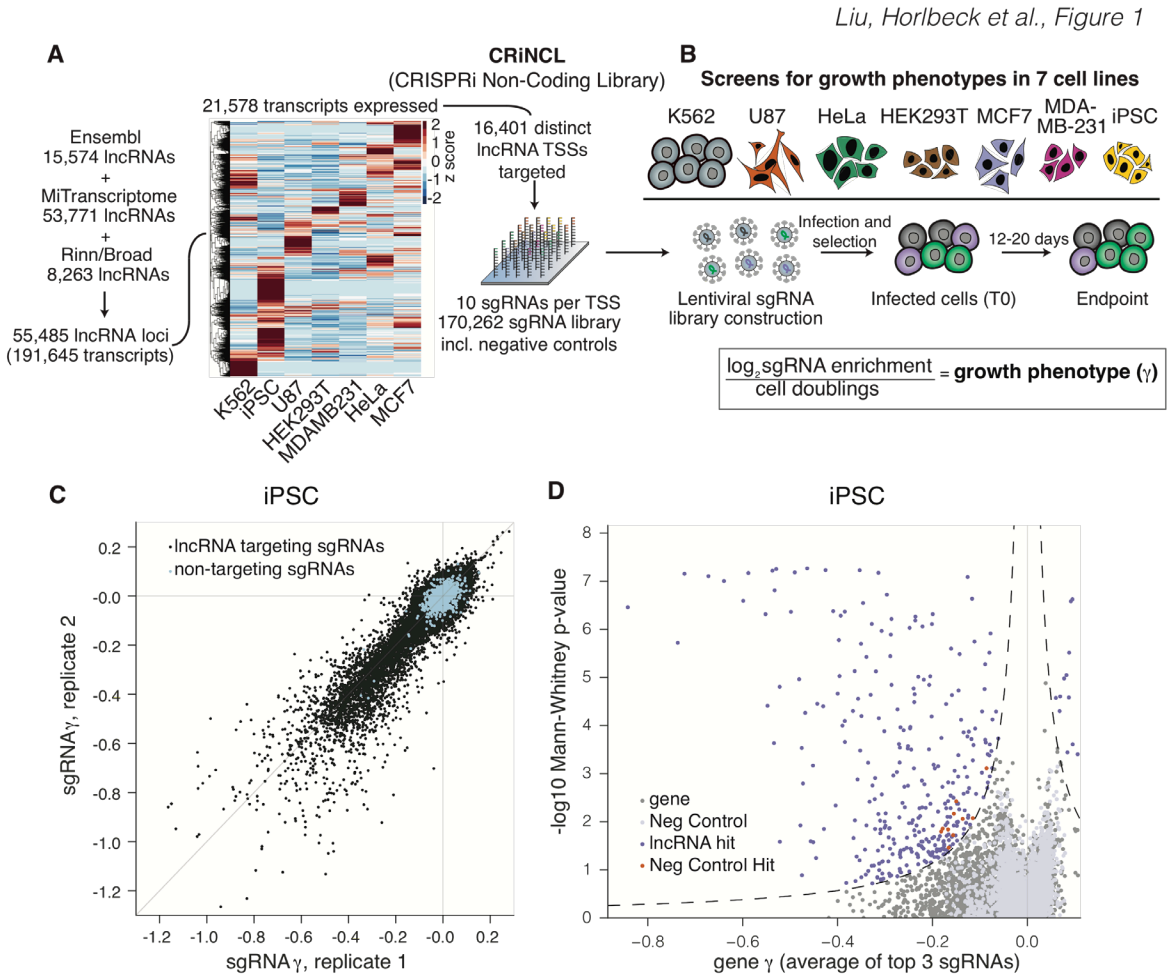
To generate machine learning models, lncRNAs phenotypes were binarized as hit (1) or non-hit (0) and used as the response variable. Categorical variables were assigned as either 1 or 0. Missing data, e.g. super-enhancer or CNV information for cell lines for which data were not available, were assigned the value of 0. Predictor variables were then centered to the mean and z standardized. Expression levels were log transformed. To avoid confounding by nearby protein coding genes, only lncRNAs whose TSS were > 1kb from a protein coding TSS were considered. Several classes of models were generated and tested, using the R package caret on randomly sampled training (75% of data) and testing (25% of data) sets from our screen results. Logistic regression outperformed both support vector machines (least squares, polynomial kernel, radial kernel) and random forests in accurately classifying test sets of lncRNAs as hits or non-hit. Therefore, we used logistic regression to identify significant predictors of lncRNA hits. 100 iterations of ten-fold cross validation was performed by randomly withholding 10% of the dataset and training logistic regression models using the remaining data. Those predictors that repeatedly scored as significant ($p < 0.01$) were noted as reliable.

Acknowledgments

We thank the members of the Lim and Weissman labs, particularly Alex Fields, Joshua Dunn, Manny DeVera, Miao Cui, and David Wu for helpful discussions and assistance. We thank Annie Truong for assistance with iPS cell culturing and Nathan Salomonis for iPSC RNA-seq data. We also thank Eric Chow and Derek Bogdanoff of the UCSF Center for Advanced Technology for sequencing assistance, and Laurakay Bruhn, Daniel Ryan, Luke Fairbairn, and Peter Tsang of Agilent Technologies for their assistance on the design and synthesis of oligonucleotide pools.

This project was supported by NIH 1R01NS091544-01A1, VA 5I01 BX000252-07, NIH SPORE DRP, the Shurl and Kay Curci Foundation, the LoGlio Foundation, and the Hana Jabsheh Initiative (to D.A.L.). S.J.L. is supported by NIH F30 NS092319-01. M.A.H., J.E.V., M.Y.C., Y.C., L.A.G., and J.S.W. were supported by the Howard Hughes Medical Institutes and the National Institutes of Health (P50 GM102706, U01 CA168370, R01 DA036858). S.W.C. and H.Y.C. are supported by National Institutes of Health (R35-CA209919, P50-HG007735). B.R.C and M.A.M are supported by the Gladstone Institutes and National Institutes of Health (U01HL100406, P01HL089707, R01HL130533). L.A.G. is supported by the NIH/NCI Pathway to Independence Award (K99CA204602). Oligonucleotide pools were provided courtesy of the Innovative Genomics Initiative. Sequencing data are deposited in GSE85011.

Figures



E

	Genes targeted in screen	Total hits	lncRNA genes neighboring essential coding genes	**lncRNA hits**
iPSC	5,534	438	112	**326**
MCF7	5,725	117	60	**57**
U87	5,689	88	23	**65**
K562	16,401	144	79	**65**
MDA-MB-231	5,725	44	14	**30**
HeLa	6,158	52	28	**24**
HEK293T	5,785	28	11	**17**

Figure 1. CRISPRi screens identify lncRNA genes that modify cell growth

A) Schematic of CRISPRi library design strategy. Three lncRNA annotation sets were merged, prioritized by expression in the indicated cell lines, and targeted by 10 sgRNAs per TSS using the hCRISPRi-v2.1 algorithm. Heatmap represents expression as z-score of fragments per kilobase million (FPKM) within each cell line (see Figure S1 for TPM values). B) Schematic of growth screens performed in 7 different cell lines, and formula for calculation of the growth phenotype (γ). C) Scatter plot of sgRNA phenotypes from two independent replicates of a CRISPRi screen performed in iPSCs. D) Volcano plot of gene γ and p-value. Screen replicates were averaged, and sgRNAs targeting the same gene were collapsed into a growth phenotype for each gene by the average of the 3 top scoring sgRNAs by absolute value, and assigned a p-value by the Mann-Whitney test of all 10 sgRNAs compared to the non-targeting controls. Negative control genes were randomly generated from the set of non-targeting sgRNAs, and dashed lines represents a threshold for calling hits by screen score (see Methods). Neighbor hits are not displayed for clarity (see Figure S3A,B). E) Summary table of all CRISPRi growth screens performed.

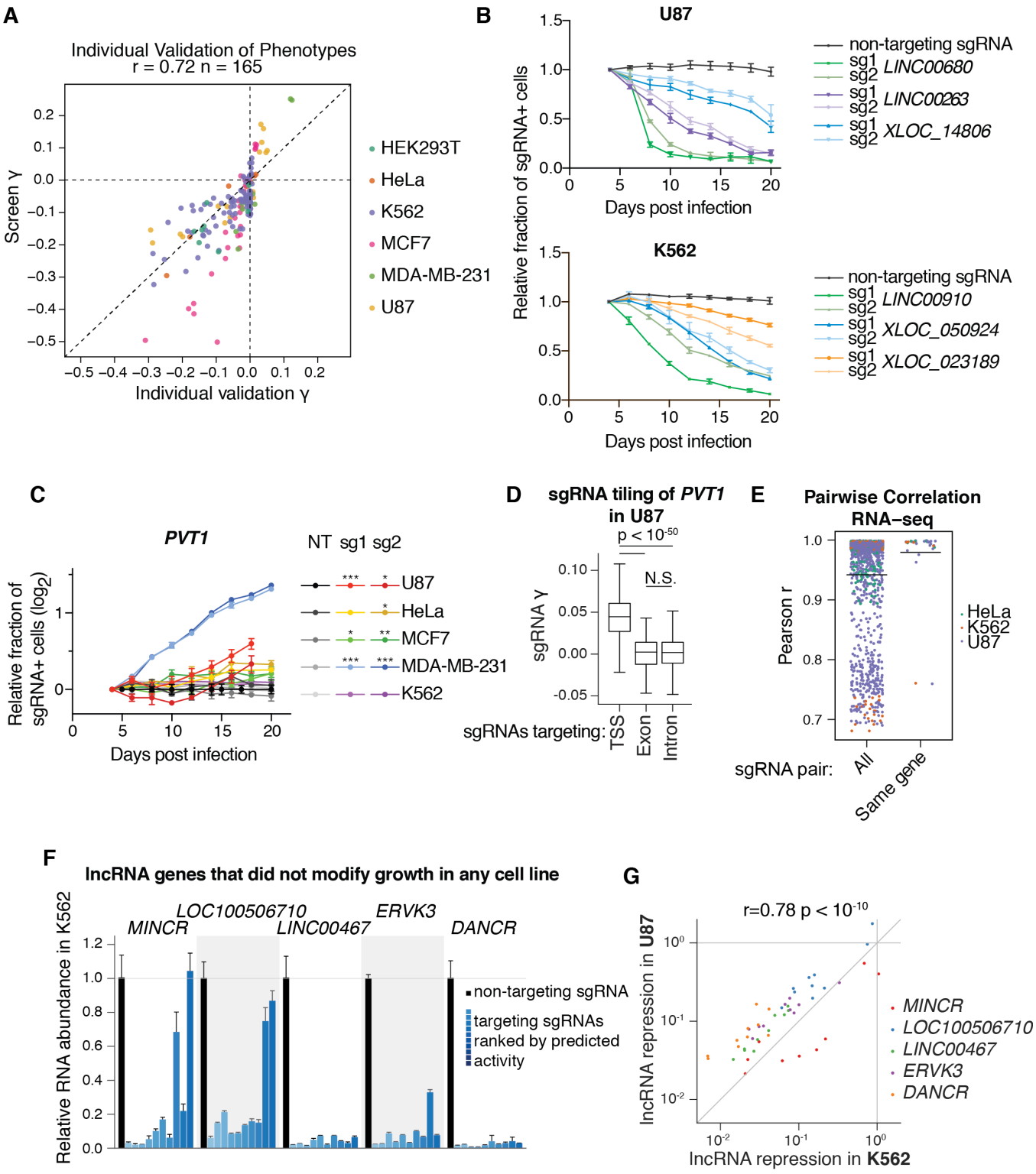

A
Individual Validation of Phenotypes
r = 0.72 n = 165
Screen γ
Individual validation γ
HEK293T
HeLa
K562
MCF7
MDA-MB-231
U87
B
U87
Relative fraction of sgRNA+ cells
Days post infection
non-targeting sgRNA
sg1 LINC00680
sg2 LINC00680
sg1 LINC00263
sg2 LINC00263
sg1 XLOC_14806
sg2 XLOC_14806
K562
sg1 LINC00910
sg2 LINC00910
sg1 XLOC_050924
sg2 XLOC_050924
sg1 XLOC_023189
sg2 XLOC_023189
C
PVT1
NT sg1 sg2
Relative fraction of sgRNA+ cells (log2)
Days post infection
U87
HeLa
MCF7
MDA-MB-231
K562
D
sgRNA tiling of PVT1 in U87
p < 10-50
N.S.
sgRNA γ
sgRNAs targeting:
TSS
Exon
Intron
E
Pairwise Correlation RNA–seq
Pearson r
sgRNA pair:
All
Same gene
HeLa
K562
U87
F
lncRNA genes that did not modify growth in any cell line
Relative RNA abundance in K562
MINCR
LOC100506710
LINC00467
ERVK3
DANCR
non-targeting sgRNA
targeting sgRNAs ranked by predicted activity
G
r=0.78 p < 10-10
lncRNA repression in U87
lncRNA repression in K562
MINCR
LOC100506710
LINC00467
ERVK3
DANCR

Figure 2. Validation of screen results shows reproducible phenotypes, correlated transcriptome responses, and robust knockdown of target transcripts

A) Individual sgRNA phenotypes from internally-controlled growth assays (B,C) compared to sgRNA phenotypes from screens. Individual growth phenotypes were calculated from relative fraction of sgRNA-containing cells at the endpoint, divided by the number of doublings from 4 days post-infection. Screen growth phenotypes represent the replicate average phenotype from the indicated cell line. B) Internally-controlled growth assays performed with sgRNAs targeting lncRNA hit genes in U87 and K562. Cells were infected with lentivirus of the sgRNA expression vector (including a BFP marker gene) and passaged for 20 days. The fraction of sgRNA-containing cells was measured as the fraction of high-BFP-expressing cells by flow cytometry, and expressed relative to the fraction at 4 days post infection. Points represent the mean and standard deviation of 3 biological replicates. C) Internally-controlled growth assays of *PVT1*-targeting sgRNAs in 5 cell lines. Assays were performed as in (B). Asterisks represent t-test p-values compared to the non-targeting (NT) sgRNA at the assay endpoint (* < 0.05, ** < 0.01, *** < 0.001). D) Boxplot of sgRNA growth phenotypes from tiling screen of *PVT1* in U87 cells. TSS represents all sgRNAs within 1kb of the *PVT1* "p1" and "p2" TSSs as annotated by FANTOM, exon represents sgRNAs targeting any *PVT1* exon annotated by Ensembl, and intron represents all other sgRNAs (see Figure S7B). sgRNA γs are the average of two replicates. E) Pairwise correlation of gene expression profiles for independent sgRNAs. Expression profiles were measured by RNA-seq and correlations were calculated from transcripts per million (TPM) of genes with significant variation of expression (see Methods). "All" represents every sgRNA pair from the same cell line with the same phenotype direction, except same-sgRNA and same-gene pairs. F) Relative RNA abundance in K562 of lncRNA genes that were not hits in any cell

line. RNA abundance for all 10 sgRNAs targeting the indicated genes in the CRiNCL library was measured by qPCR. Each bar represents the mean and standard deviation of 3 biological replicates, and is ordered by decreasing activity as predicted by the hCRISPRi-v2.1 algorithm. G) Correlation of lncRNA repression in K562 and U87. Points represent mean values from (F) and Figure S7C.

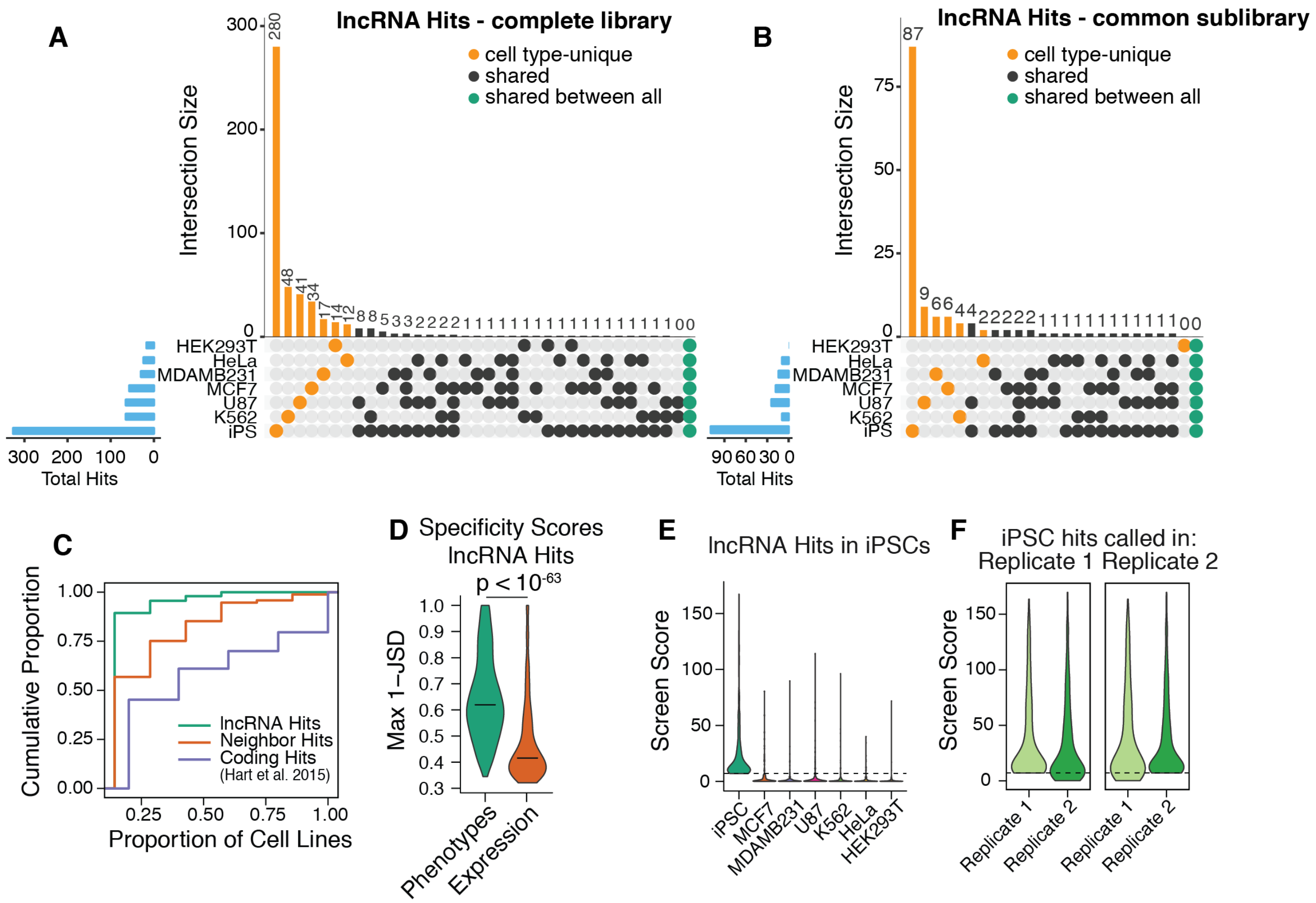

A
lncRNA Hits - complete library
cell type-unique
shared
shared between all
Intersection Size
Total Hits
HEK293T
HeLa
MDAMB231
MCF7
U87
K562
iPS
B
lncRNA Hits - common sublibrary
C
Cumulative Proportion
Proportion of Cell Lines
lncRNA Hits
Neighbor Hits
Coding Hits
(Hart et al. 2015)
D
Specificity Scores
lncRNA Hits
p < 10-63
Max 1–JSD
Phenotypes
Expression
E
lncRNA Hits in iPSCs
Screen Score
F
iPSC hits called in:
Replicate 1 Replicate 2
Replicate 1
Replicate 2

Figure 3. Growth modifier lncRNA function is highly cell type-specific

A) Numbers of lncRNA hits for each set of cell types in the complete library and (B) common sublibrary (lncRNAs that were expressed and screened in all cell types). Blue bars indicate total number of lncRNA hits in each cell type. C) Cumulative distribution function for the proportion of cell types in which each gene is a hit. Protein coding hits were obtained from Hart et al. 2015 using the authors' 5% FDR Bayes Factor threshold. D) Distributions of the maximum 1 - Jenson Shannon distance (JSD) metric of cell type-specificity for lncRNA hit screen scores and expression values. Horizontal lines – median. E) Distributions of screen scores across all cell types for lncRNAs that were hits in iPSCs. Dashed line represents screen score threshold for calling hit genes. F) Distributions of screen scores across both replicates of iPS cells, for lncRNAs that would be called as hits in replicate 1 (left) and in replicate 2 (right).

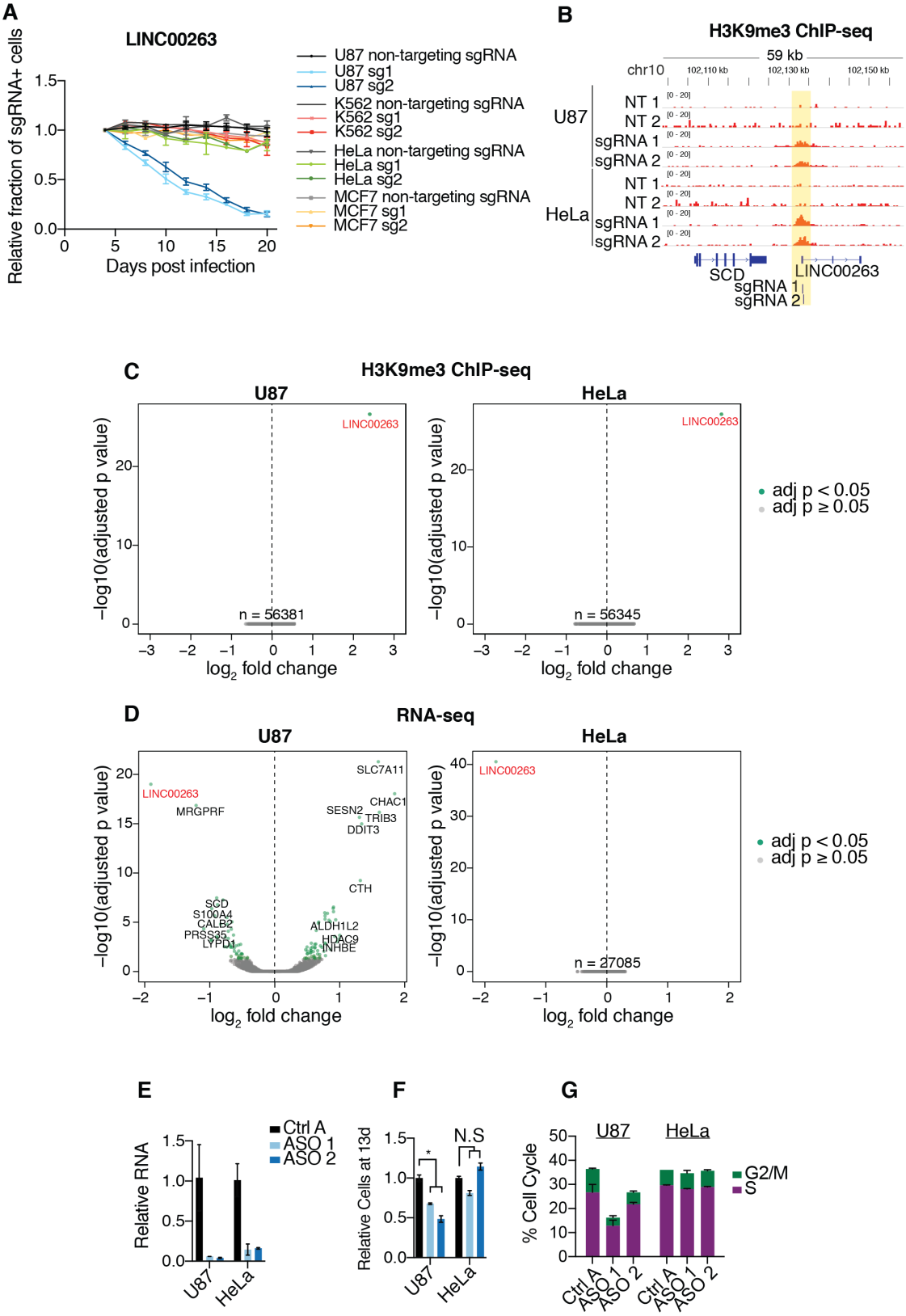
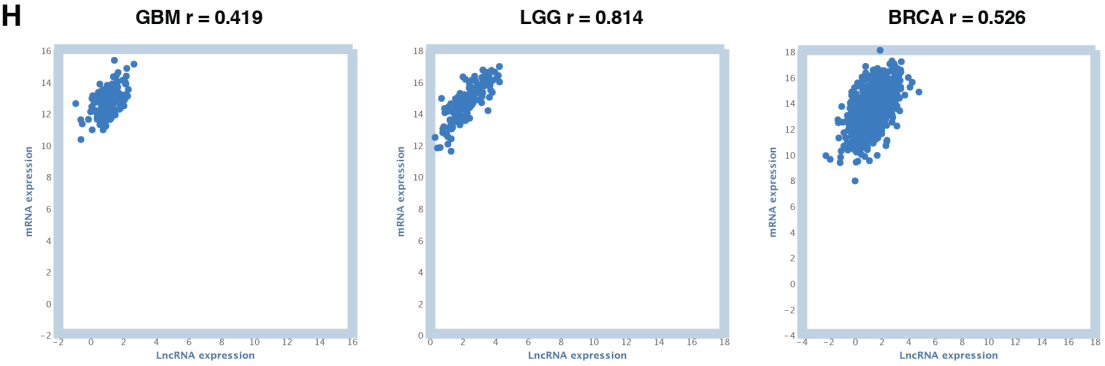
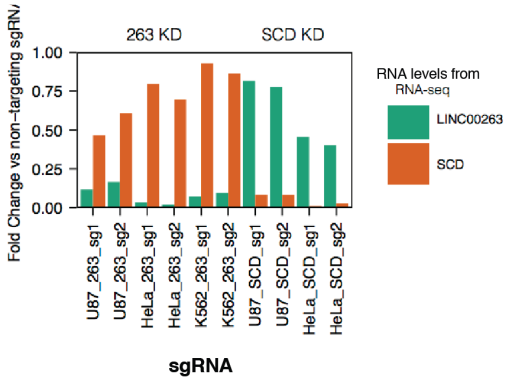

A
LINC00263
Relative fraction of sgRNA+ cells
Days post infection
U87 non-targeting sgRNA
U87 sg1
U87 sg2
K562 non-targeting sgRNA
K562 sg1
K562 sg2
HeLa non-targeting sgRNA
HeLa sg1
HeLa sg2
MCF7 non-targeting sgRNA
MCF7 sg1
MCF7 sg2
B
H3K9me3 ChIP-seq
59 kb
chr10
102,110 kb
102,130 kb
102,150 kb
U87
HeLa
NT 1
NT 2
sgRNA 1
sgRNA 2
SCD
LINC00263
C
H3K9me3 ChIP-seq
U87
HeLa
LINC00263
−log10(adjusted p value)
log2 fold change
n = 56381
n = 56345
adj p < 0.05
adj p ≥ 0.05
D
RNA-seq
U87
HeLa
LINC00263
MRGPRF
SLC7A11
CHAC1
SESN2
TRIB3
DDIT3
CTH
SCD
S100A4
CALB2
PRSS35
LYPD1
ALDH1L2
HDAC9
INHBE
n = 27085
E
Relative RNA
Ctrl A
ASO 1
ASO 2
F
Relative Cells at 13d
N.S
G
% Cell Cycle
G2/M
S

Figure 4 contd

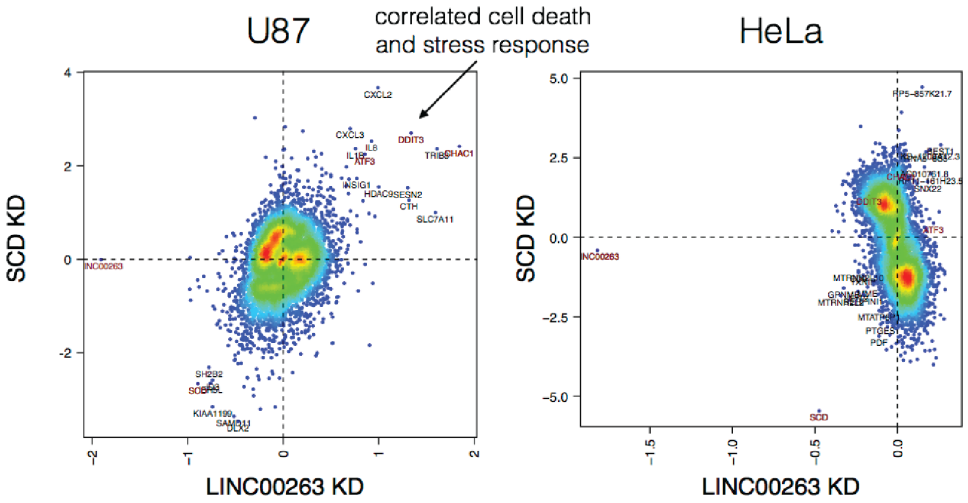
H



I



J



K

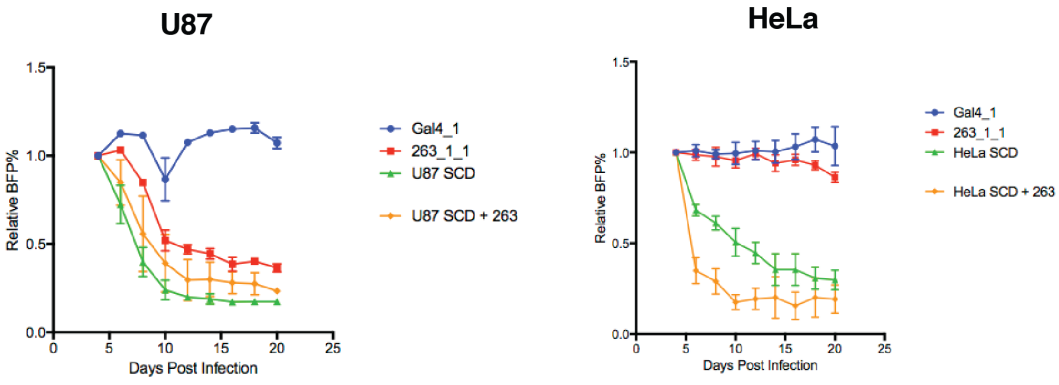


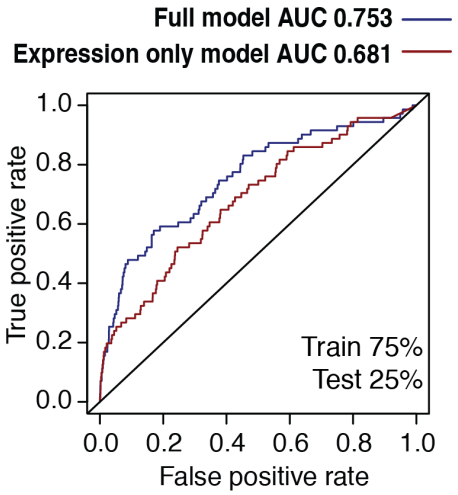
Figure 4. Dissection of cell type-specific growth modifier lncRNA *LINC00263*

A) Internally-controlled growth assays for 2 independent sgRNAs targeting the TSS of *LINC00263* and non-targeting sgRNA in U87, K562, HeLa, and MCF7 cells. B) ChIP-seq against H3K9me3 in replicates of U87 and HeLa cells infected with non-targeting sgRNAs or *LINC00263* sgRNAs. Values represent normalized reads. C) Volcano plots for ChIP-seq samples in (B), representing genome-wide differential enrichment of H3K9me3 at promoter regions. Fold changes are between *LINC00263* sgRNAs over non-targeting sgRNAs. D) Volcano plots for RNA-seq differential expression following infection of *LINC00263* sgRNAs compared to infection of non-targeting sgRNAs. E) qPCR of ASO knockdown of *LINC00263* in U87 and HeLa cells. F) Proportion of cells at 13 days post ASO transfection, relative to control ASO. G) Percentage of cells in S or G2/M phases following ASO knockdown of *LINC00263*. * indicates p = 0.0029. H) RNA-seq expression levels of *LINC00263* (x axis) and *SCD* (y axis) from TCGA data obtained from the TANRIC web server. I) RNA-seq fold changes of the genes *LINC00263* and *SCD* following infection of sgRNAs targeting *LINC00263* or *SCD,* relative to non-targeting control sgRNAs. J) Global RNA-seq analysis of gene fold changes following infection of sgRNAs targeting *LINC00263* (x axis) or *SCD* (y axis) in either U87 (left) or HeLa (right) cells. Arrow points to correlated upregulated cell death/stress genes in U87. K) Internally-controlled growth assays for sgRNAs targeting the TSS of *LINC00263*, *SCD,* or both, in addition to non-targeting sgRNA, in U87 and HeLa. Gal4 = non-targeting control sgRNA.

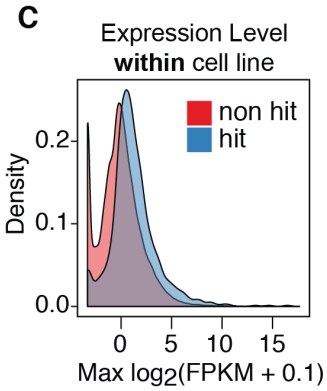
A

Variable	Cell Type Specific	Continuous or Discrete	odds	p value	% of CV where p < 0.01
Expression level	●	**C**	**1.62**	**4.52E-28**	**100**
Near FANTOM Enhancer		**D**	**1.34**	**6.44E-09**	**100**
Near cancer associated SNP		**D**	**1.10**	**1.84E-03**	**81**
Number of exons		**C**	**1.11**	**5.74E-03**	**66**
Within Pol2 ChIA-PET loop	●	D	0.83	1.27E-02	23
Near Super Enhancer	●	D	1.14	1.65E-02	23
Within CTCF ChIA-PET loop	●	D	0.83	1.81E-02	14
Near Traditional Enhancer	●	D	0.84	2.20E-02	5
Has mouse ortholog		D	1.08	4.44E-02	0
Locus is heterozygous deleted	●	D	0.92	1.98E-01	0
Is intergenic		D	1.12	2.57E-01	0
Transcript length		C	0.95	4.45E-01	0
Locus is amplified	●	D	0.96	5.12E-01	0
Is antisense		D	1.04	6.86E-01	0
Locus-Nearest coding gene dist.		C	0.98	8.12E-01	0
TSS-Nearest coding gene dist.		C	1.02	8.63E-01	0
Near Vista Enhancer		D	1.01	8.68E-01	0
Locus is homozygous deleted	●	D	0.79	9.69E-01	0

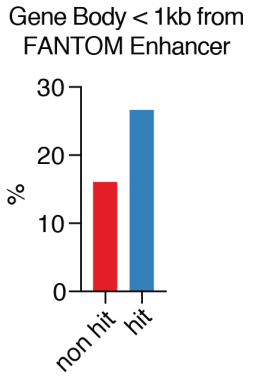
B



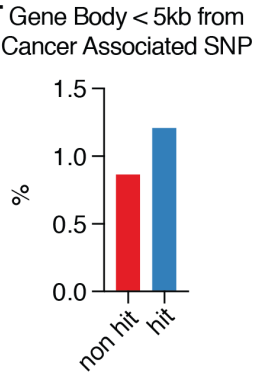
C



D



E



F

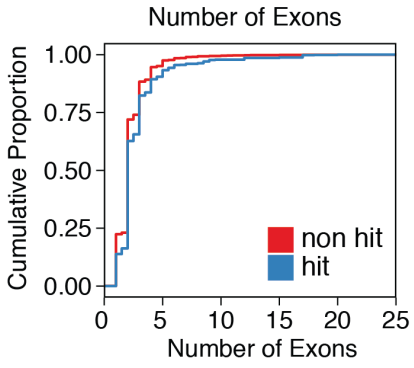


Figure 5. Machine learning identifies genomic features of growth modifier lncRNAs

A) Results from logistic regression model using 18 classes of genomic data as possible predictors of growth modifier lncRNAs. Cell type dependent variables are marked. Odds ratios represent relative impact of 1 standard deviation increase of given variable. Significant variables ($p < 0.01$) are bolded. Results of 10-fold cross validation are represented as the % of cross validation iterations where the given variable is significant. B) ROC curves for full model compared to model using only expression data. C) Density plot of expression levels for lncRNAs that scored as hits and non-hits, aggregated across all cell types. D) Percentage of non-hit (red) and hit (blue) lncRNAs whose gene bodies resided < 1 kb from an annotated FANTOM enhancer. E) Percentage of non-hit (red) and hit (blue) lncRNAs whose gene bodies resided < 5 kb from a cancer associated SNP. F) Cumulative distribution function of number of exons for non-hit (red) and hit (blue) lncRNAs transcripts.

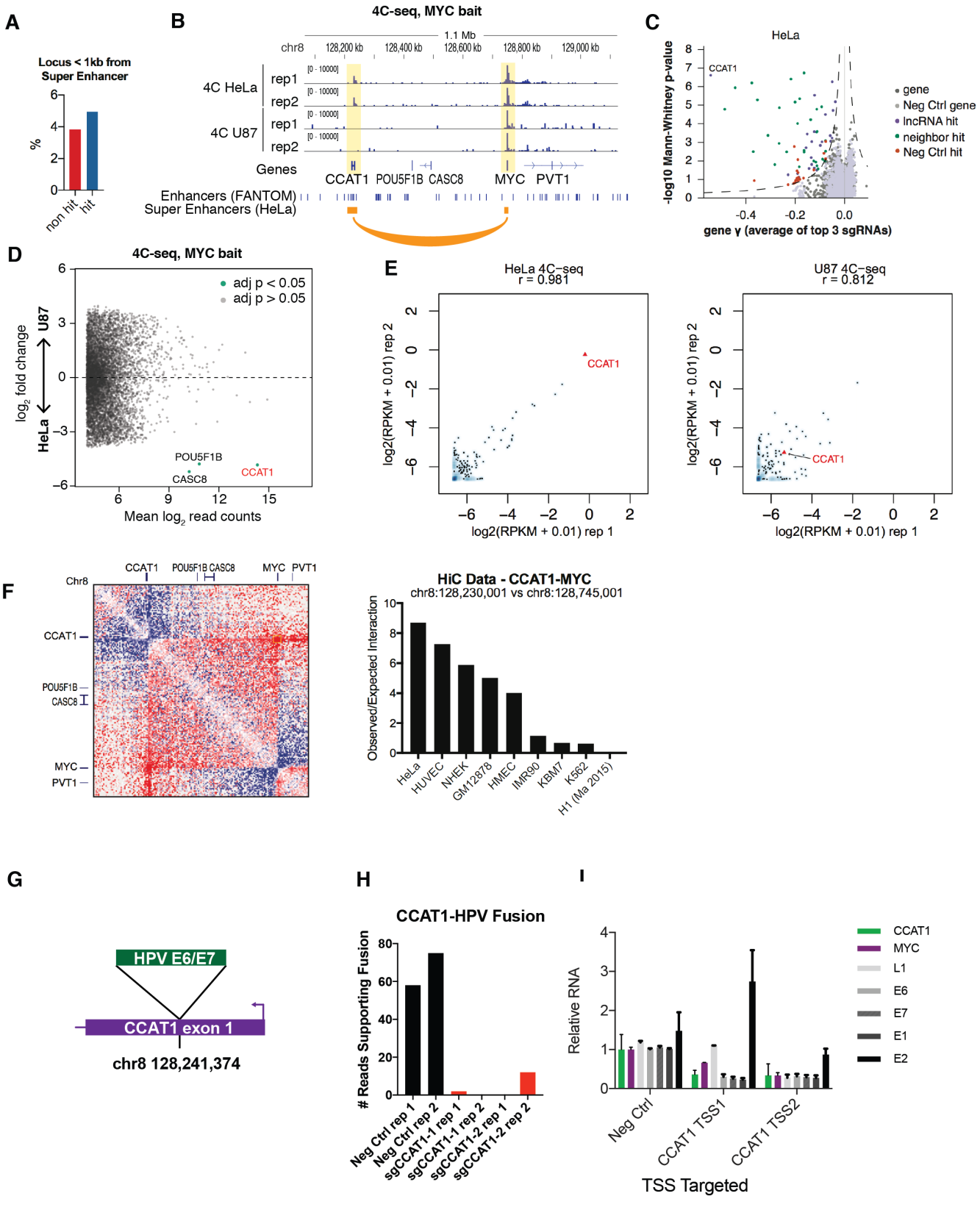

A
Locus < 1kb from Super Enhancer
%
non hit
hit
B
4C-seq, MYC bait
1.1 Mb
chr8
128,200 kb
128,400 kb
128,600 kb
128,800 kb
129,000 kb
4C HeLa
rep1
rep2
4C U87
rep1
rep2
Genes
CCAT1 POU5F1B CASC8
MYC PVT1
Enhancers (FANTOM)
Super Enhancers (HeLa)
C
HeLa
-log10 Mann-Whitney p-value
CCAT1
gene
Neg Ctrl gene
lncRNA hit
neighbor hit
Neg Ctrl hit
gene γ (average of top 3 sgRNAs)
D
4C-seq, MYC bait
adj p < 0.05
adj p > 0.05
log2 fold change
U87
HeLa
POU5F1B
CASC8
CCAT1
Mean log2 read counts
E
HeLa 4C-seq
r = 0.981
U87 4C-seq
r = 0.812
log2(RPKM + 0.01) rep 2
log2(RPKM + 0.01) rep 1
CCAT1
F
CCAT1
POU5F1B CASC8
MYC
PVT1
Chr8
HiC Data - CCAT1-MYC
chr8:128,230,001 vs chr8:128,745,001
Observed/Expected Interaction
HeLa
HUVEC
NHEK
GM12878
HMEC
IMR90
KBM7
K562
H1 (Ma 2015)
G
HPV E6/E7
CCAT1 exon 1
chr8 128,241,374
H
CCAT1-HPV Fusion
Reads Supporting Fusion
Neg Ctrl rep 1
Neg Ctrl rep 2
sgCCAT1-1 rep 1
sgCCAT1-1 rep 2
sgCCAT1-2 rep 1
sgCCAT1-2 rep 2
I
Relative RNA
CCAT1
MYC
L1
E6
E7
E1
E2
Neg Ctrl
CCAT1 TSS1
CCAT1 TSS2
TSS Targeted

Figure 6. Characterization of a lncRNA-oncogene fusion transcript *CCAT1*

A) Percentage of non-hit (red) and hit (blue) lncRNAs whose locus resided < 1 kb from a cell type-specific super enhancer. B) 4C-seq in replicates of HeLa and U87 cells, using the *MYC* locus as bait. Orange curve schematizes observed interaction between *CCAT1* and *MYC*. C) Volcano plot of screen results in HeLa cells, reproduced from Figure S1, but labeling *CCAT1*. D) Genome-wide MA plot of 4C-seq differential contact from the *MYC* bait region between HeLa and U87 cells. E) Reproducibility of 4C-seq between replicate 1 and 2 of HeLa (left) and U87 (right) cells. F) Normalized ratios of observed/expected interaction between *CCAT1* and *MYC* in published Hi-C data (Rao et al. 2014; Ma et al. 2015). Left shows Hi-C contact matrix in HeLa cells at the *CCAT1/MYC* region. Yellow box represents CCAT1-MYC contact. Right shows observed/expected interaction for CCAT1-MYC across cell types. G) Schematic of the HPV18 integration at the *CCAT1* locus. H) Number of RNA-seq reads spanning the fusion junction between HPV18 and *CCAT1* following infection of negative control sgRNAs and sgRNAs targeting either TSS of *CCAT1* in HeLa. I) RNA-seq quantification of RNA levels for *CCAT1*, *MYC*, and the viral oncogenes of HPV18 following infection of negative control sgRNAs and sgRNAs targeting either TSS of *CCAT1* in HeLa.

Supplementary Figures

Liu, Horlbeck et al., Figure S1

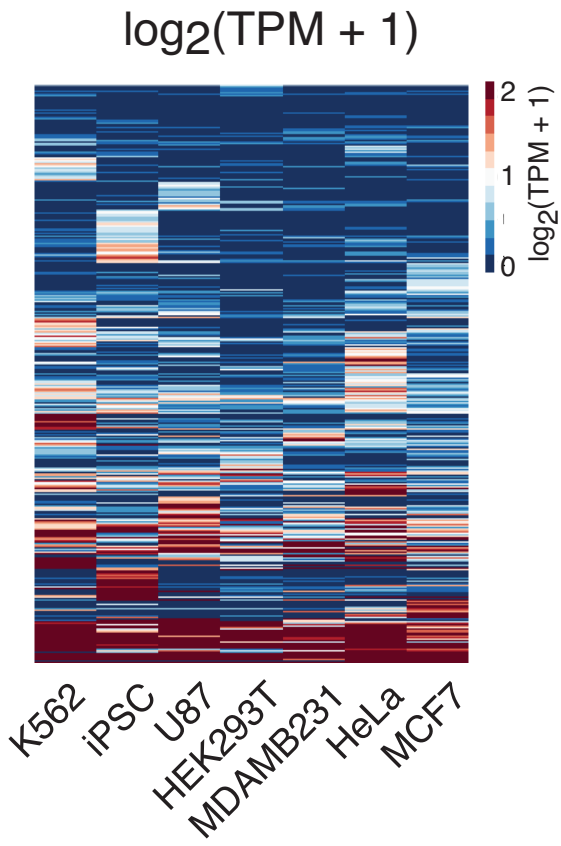


Figure S1. Expression levels of lncRNAs targeted in the CRiNCL library.

Rows correspond to those used in Figure 1A. TPM, transcripts per million.

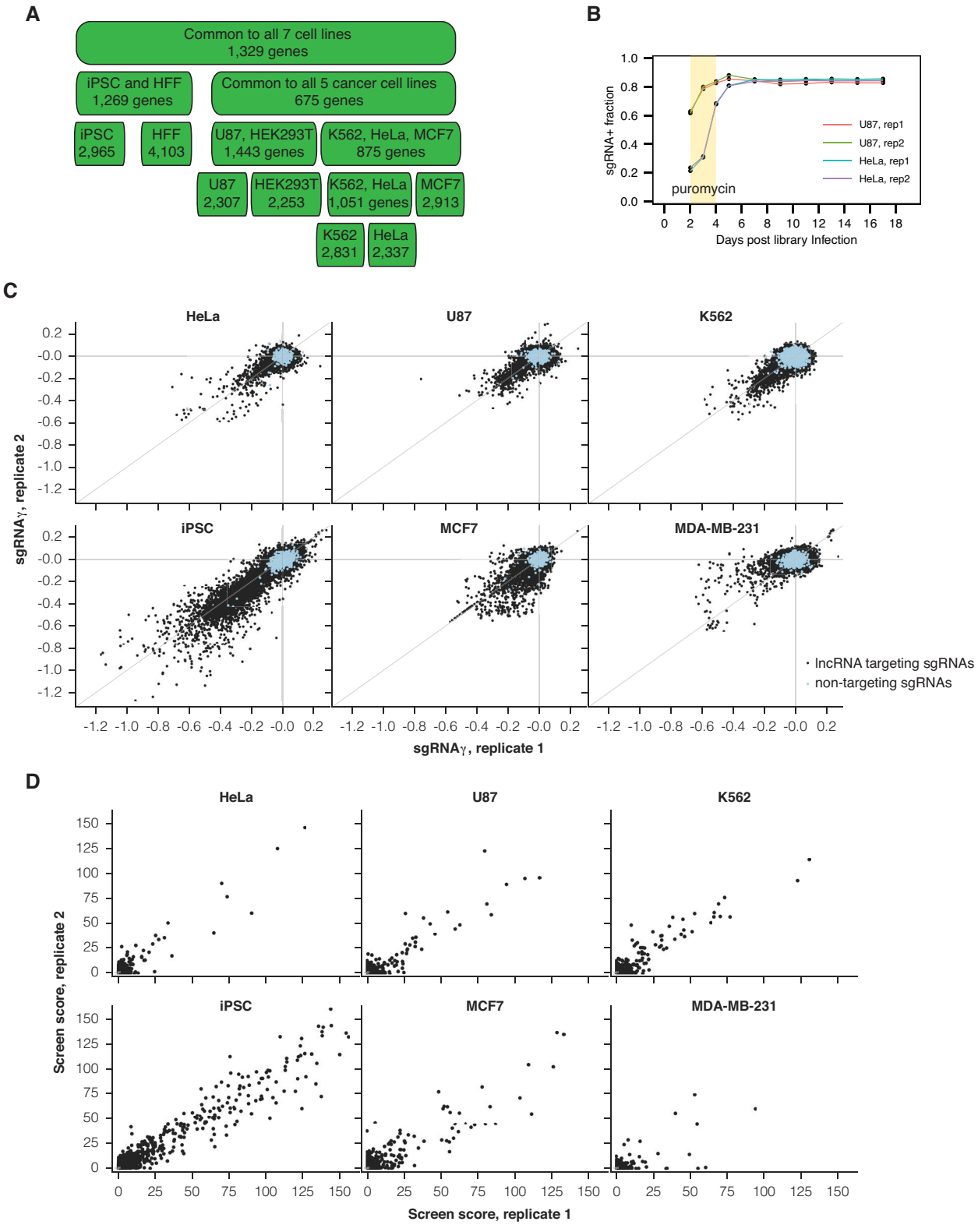
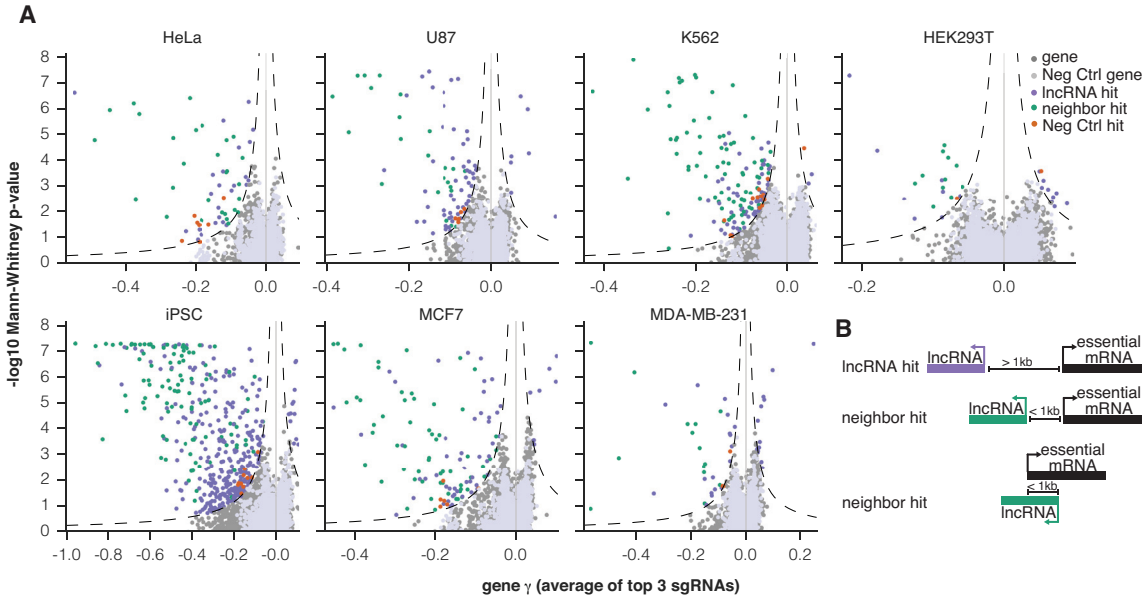
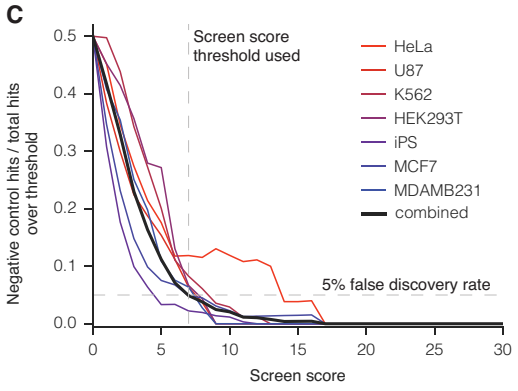

A
Common to all 7 cell lines
1,329 genes
iPSC and HFF
1,269 genes
Common to all 5 cancer cell lines
675 genes
iPSC
2,965
HFF
4,103
U87, HEK293T
1,443 genes
K562, HeLa, MCF7
875 genes
U87
2,307
HEK293T
2,253
K562, HeLa
1,051 genes
MCF7
2,913
K562
2,831
HeLa
2,337
B
sgRNA+ fraction
puromycin
U87, rep1
U87, rep2
HeLa, rep1
HeLa, rep2
Days post library Infection
C
HeLa
U87
K562
iPSC
MCF7
MDA-MB-231
sgRNAγ, replicate 2
sgRNAγ, replicate 1
lncRNA targeting sgRNAs
non-targeting sgRNAs
D
HeLa
U87
K562
iPSC
MCF7
MDA-MB-231
Screen score, replicate 2
Screen score, replicate 1

Figure S2. CRISPRi growth screens performed in seven cell lines.

A) Schematic of sublibrary divisions of the CRiNCL library. The library was divided into 13 sublibraries based on expression in 7 cell lines to facilitate library cloning and allow for targeted screens. Combinations of sublibraries were selected for screening in each cell line studied as described in Methods. B) Fraction of cells containing the sgRNA library over the course of the U87 and HeLa screens. sgRNA-containing fraction measured as the fraction of high-BFP-expressing cells by flow cytometry. C) sgRNA γ for replicate screens performed in 6 cell lines, as in Figure 1C. Only one replicate was performed for HEK293T. D) Screen scores from replicate screens for individual lncRNA loci. Screen scores were calculated as described in Methods.





D

Distance to nearest essential protein-coding gene TSS	lncRNA hits passing threshold
1.0kb	499
2.5kb	490
5.0kb	485
10.0kb	480

Figure S3. CRISPRi growth screen results and validation of thresholds used in screen analysis.

A) Volcano plots for screens performed in 7 cell lines, as in Figure 1D. Hits that were considered neighbor hits (see B) are labeled. B) Schematic of definitions of "lncRNA hit" and "neighbor hit." C) Fraction of negative control genes called as hits out of total number of hits above the indicated screen score threshold, calculated for each cell line and the combined dataset. D) Number of lncRNA hits classified after eliminating neighbor hits within the indicated distance from any essential protein-coding gene. A 1.0kb threshold was used for all analysis.

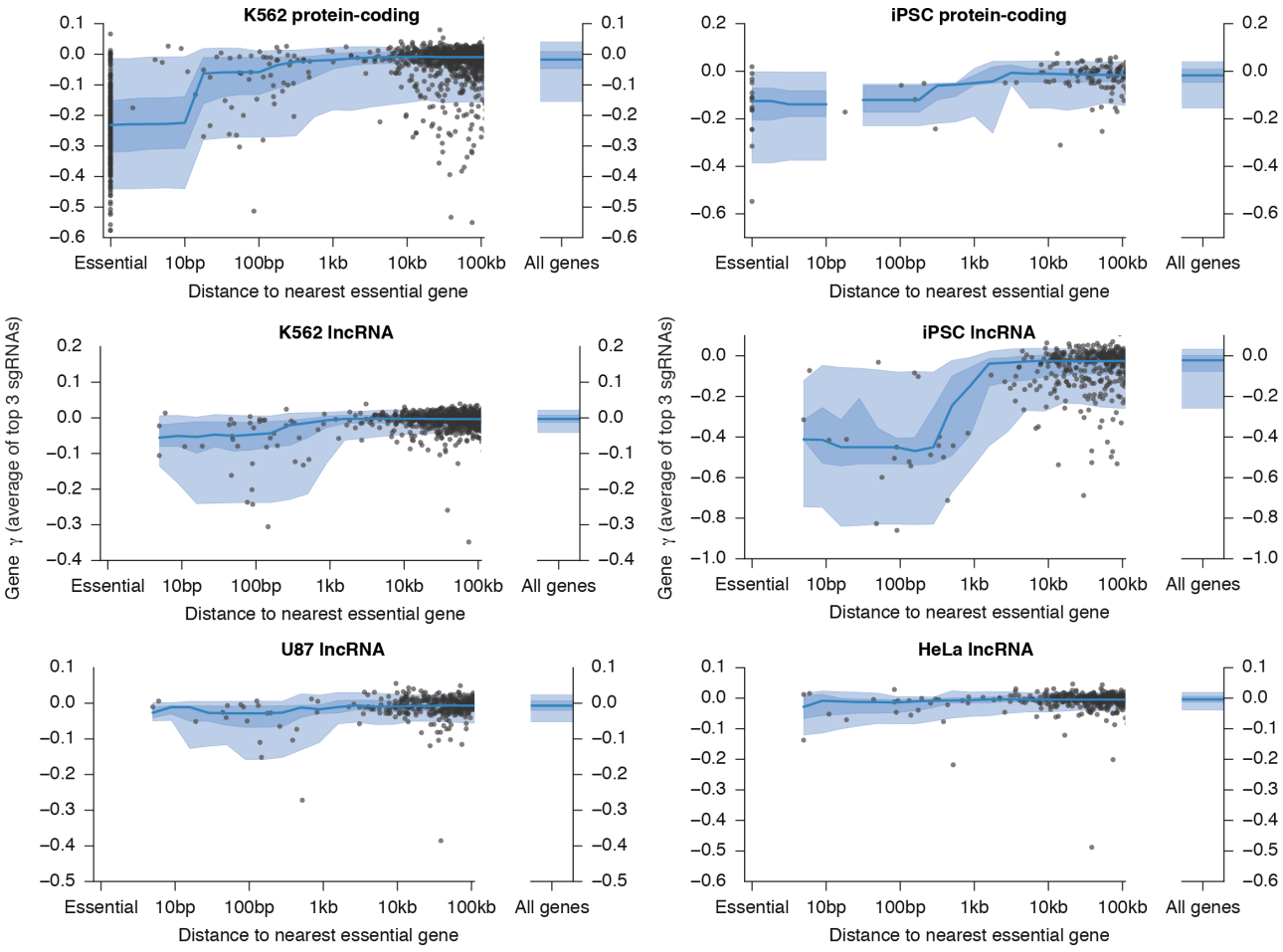

K562 protein-coding
iPSC protein-coding
K562 lncRNA
iPSC lncRNA
U87 lncRNA
HeLa lncRNA
Gene γ (average of top 3 sgRNAs)
Essential
10bp
100bp
1kb
10kb
100kb
All genes
Distance to nearest essential gene

Figure S4. CRISPRi growth phenotypes relative to gold standard essential genes.

Distribution of gene γ relative to the nearest gold standard essential protein-coding gene (*44*). Points indicate individual gene γ scores and blue shaded regions represent 5th, 25th, 50th, 75th, and 95th percentiles of all genes within 10-fold of the position. Screen data plotted are from the indicated protein-coding screens ((*35*) and Figure S4) or lncRNA screens.

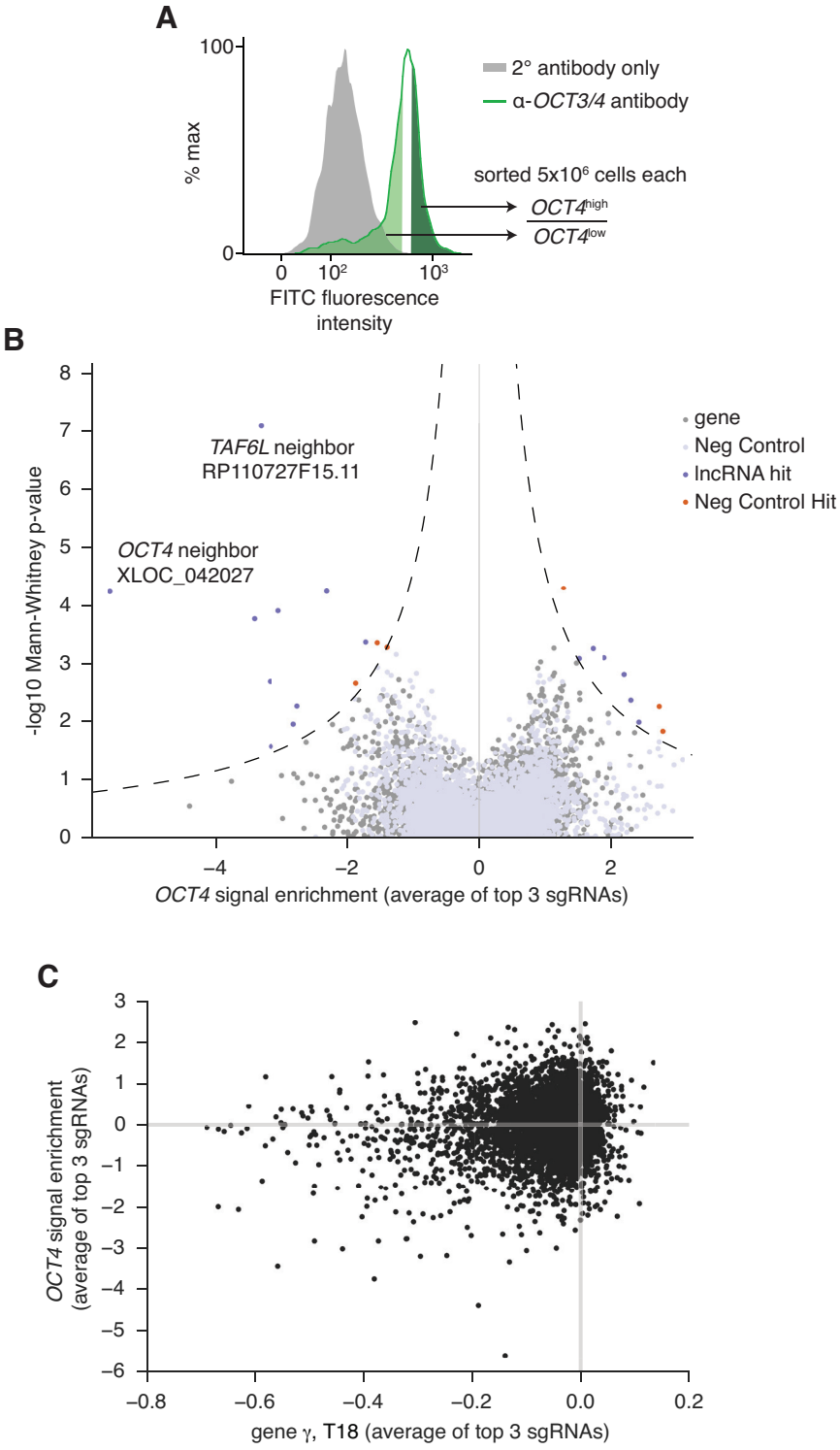

A
100
0
% max
0
10²
10³
FITC fluorescence intensity
2° antibody only
α-OCT3/4 antibody
sorted 5x10⁶ cells each
OCT4high
OCT4low
B
-log10 Mann-Whitney p-value
TAF6L neighbor
RP110727F15.11
OCT4 neighbor
XLOC_042027
gene
Neg Control
lncRNA hit
Neg Control Hit
OCT4 signal enrichment (average of top 3 sgRNAs)
C
OCT4 signal enrichment (average of top 3 sgRNAs)
gene γ, T18 (average of top 3 sgRNAs)

Figure S5. A FACS-based screen for *OCT4* expression identifies genes that modify iPSC differentiation.

A) Representative FACS histogram of *OCT4* staining of iPSCs, with high and low 30% fractions highlighted. $5x10^6$ cells from each fraction were sorted and processed for Illumina sequencing. *OCT4* signal enrichment was calculated as the fraction of each sgRNA present in the high sample versus the low sample. B) Volcano plot of screen results, as in Figure 1D. C) Gene growth phenotypes from iPSC screen compared to *OCT4* signal enrichment. The results suggests that all *OCT4* screen hits also modify cell growth rates, but that most growth screen hits are not accompanied by changes in *OCT4* expression.

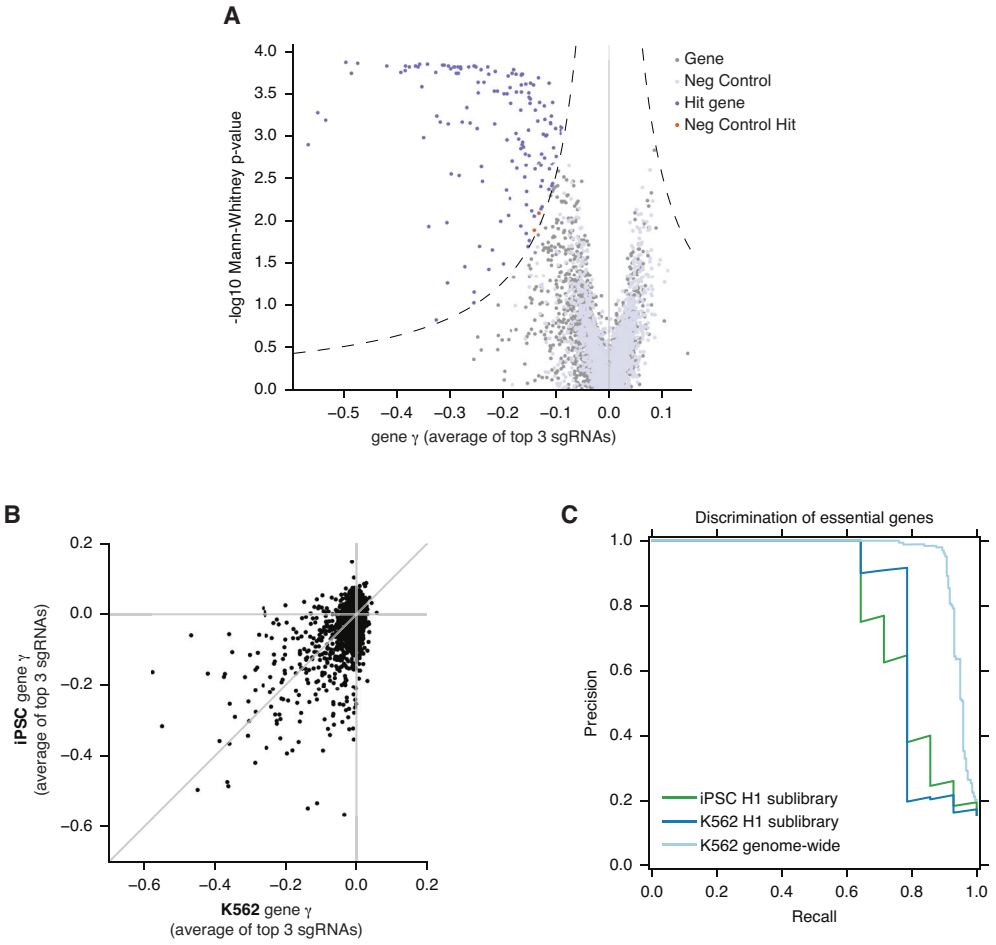

A
-log10 Mann-Whitney p-value
gene γ (average of top 3 sgRNAs)
Gene
Neg Control
Hit gene
Neg Control Hit
B
iPSC gene γ
(average of top 3 sgRNAs)
K562 gene γ
(average of top 3 sgRNAs)
C
Discrimination of essential genes
Precision
Recall
iPSC H1 sublibrary
K562 H1 sublibrary
K562 genome-wide

Figure S6. A CRISPRi screen for protein-coding genes that modify growth in iPSC.

A) Volcano plot for iPSC screen performed with the hCRISPRi-v2 H1 sublibrary, 5 sgRNAs/gene, displayed as in Figure 1D. Data are the average of two replicate screens. B) Scatter plot of gene γ from growth screens performed in K562 (*35*) and iPSC. K562 screen was performed with the genome-wide hCRISPRi-v2 library and reanalyzed here using only the H1 5 sgRNAs/gene sublibrary. C) Discrimination of gold-standard essential genes from non-essential genes (*44*) for K562 and iPSC screens, ranked by gene γ. K562 genome-wide and H1 data were analyzed using the 5 sgRNAs/gene sublibraries.

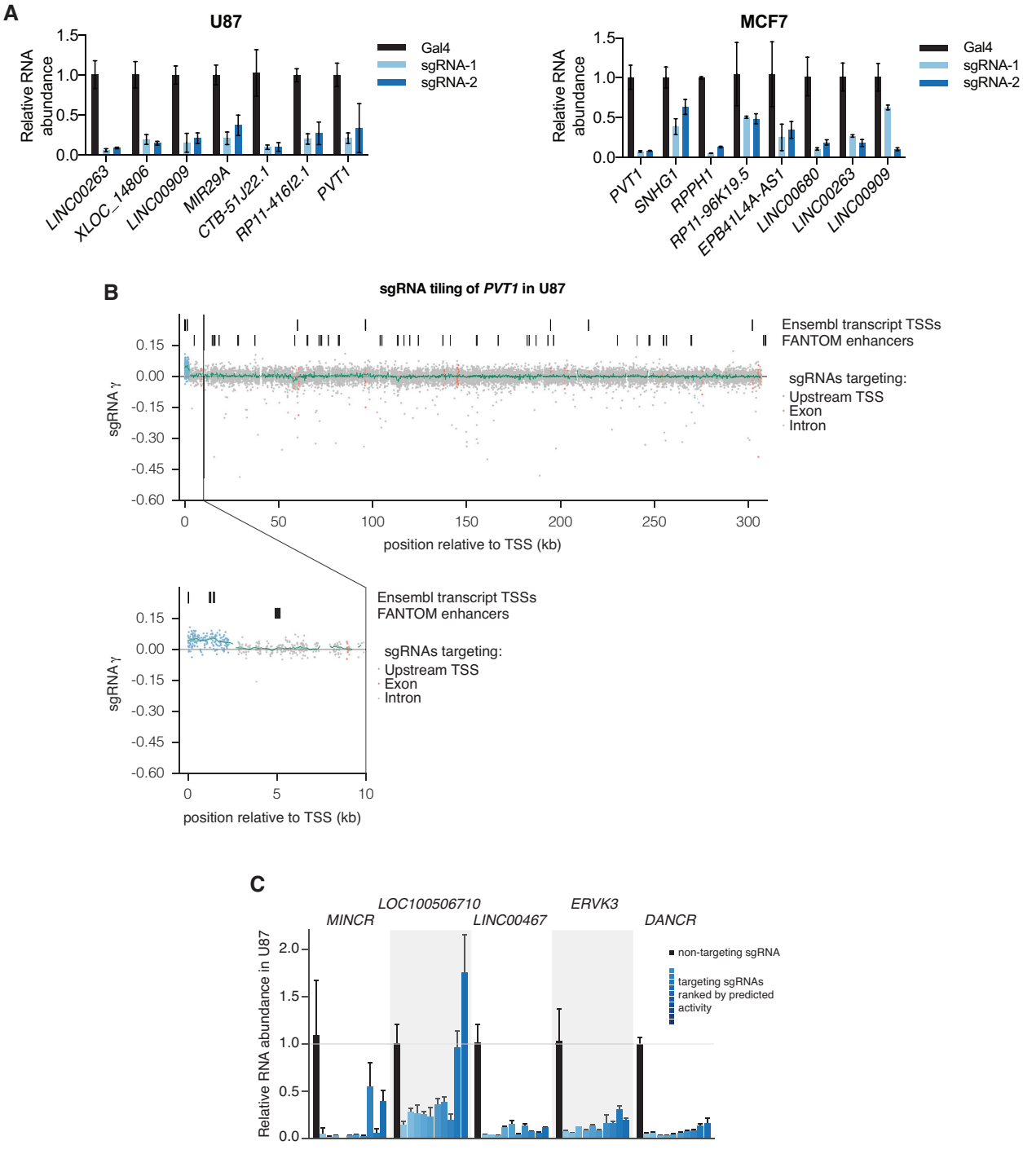

A
U87
MCF7
Relative RNA abundance
Gal4
sgRNA-1
sgRNA-2
LINC00263
XLOC_14806
LINC00909
MIR29A
CTB-51J22.1
RP11-416I2.1
PVT1
SNHG1
RPPH1
RP11-96K19.5
EPB41L4A-AS1
LINC00680
B
sgRNA tiling of PVT1 in U87
sgRNA γ
position relative to TSS (kb)
Ensembl transcript TSSs
FANTOM enhancers
sgRNAs targeting:
Upstream TSS
Exon
Intron
C
LOC100506710
ERVK3
MINCR
LINC00467
DANCR
Relative RNA abundance in U87
non-targeting sgRNA
targeting sgRNAs ranked by predicted activity

Figure S7. lncRNA CRISPRi produces robust knockdown and is specific to the TSS.

A) Relative RNA abundance of lncRNA hits upon knockdown with CRIPSRi. Bars represent mean and standard deviation of 3 biological replicates. B) sgRNA growth phenotypes from tiling screen of *PVT1* in U87 cells by position. sgRNA position was calculated as the genomic coordinate of the protospacer adjacent motif (PAM) relative to the *PVT1* p1 FANTOM TSS. sgRNA γ is the average of two replicates. TSS, exon, and intron are defined as in Figure 2D. Green line represents median phenotype of all sgRNAs within 250bp. C) Relative RNA abundance in U87 of lncRNA genes that were not hits in any cell line, as with Figure F,G.

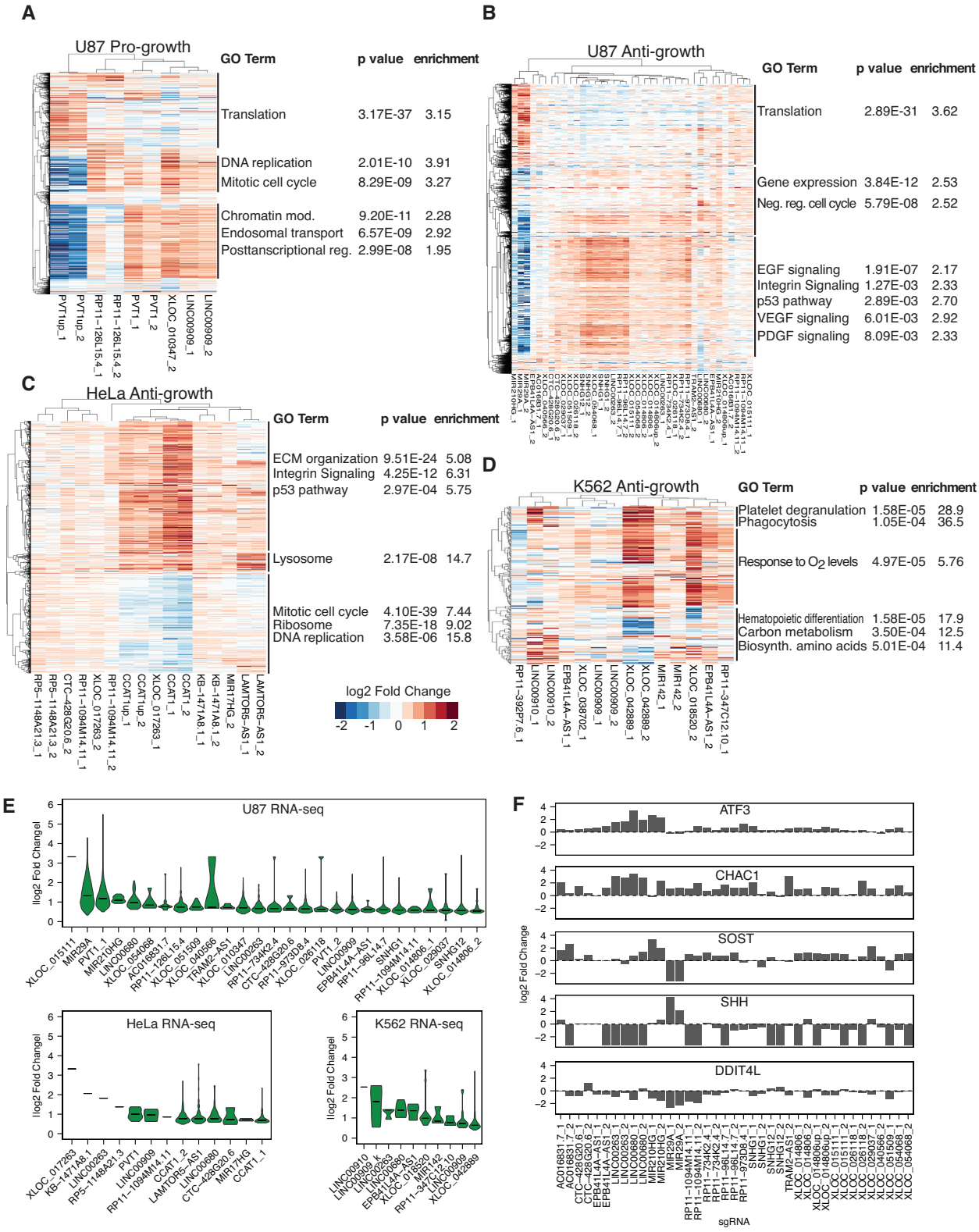

A
U87 Pro-growth
GO Term p value enrichment
Translation 3.17E-37 3.15
DNA replication 2.01E-10 3.91
Mitotic cell cycle 8.29E-09 3.27
Chromatin mod. 9.20E-11 2.28
Endosomal transport 6.57E-09 2.92
Posttanscriptional reg. 2.99E-08 1.95
PVT1up_1
PVT1up_2
RP11−126L15.4_1
RP11−126L15.4_2
PVT1_1
PVT1_2
XLOC_010347_2
LINC00909_1
LINC00909_2
B
U87 Anti-growth
GO Term p value enrichment
Translation 2.89E-31 3.62
Gene expression 3.84E-12 2.53
Neg. reg. cell cycle 5.79E-08 2.52
EGF signaling 1.91E-07 2.17
Integrin Signaling 1.27E-03 2.33
p53 pathway 2.89E-03 2.70
VEGF signaling 6.01E-03 2.92
PDGF signaling 8.09E-03 2.33
C
HeLa Anti-growth
GO Term p value enrichment
ECM organization 9.51E-24 5.08
Integrin Signaling 4.25E-12 6.31
p53 pathway 2.97E-04 5.75
Lysosome 2.17E-08 14.7
Mitotic cell cycle 4.10E-39 7.44
Ribosome 7.35E-18 9.02
DNA replication 3.58E-06 15.8
log2 Fold Change
-2 -1 0 1 2
D
K562 Anti-growth
GO Term p value enrichment
Platelet degranulation 1.58E-05 28.9
Phagocytosis 1.05E-04 36.5
Response to O2 levels 4.97E-05 5.76
Hematopoietic differentiation 1.58E-05 17.9
Carbon metabolism 3.50E-04 12.5
Biosynth. amino acids 5.01E-04 11.4
RP11−392P7.6_1
LINC00910_1
LINC00910_2
EPB41L4A−AS1_1
XLOC_038702_1
LINC00909_1
LINC00909_2
XLOC_042889_1
XLOC_042889_2
MIR142_1
MIR142_2
XLOC_018520_2
EPB41L4A−AS1_2
RP11−347C12.10_1
E
U87 RNA-seq
log2 Fold Change
XLOC_015111
MIR29A
PVT1_1
MIR210HG
LINC00680
XLOC_054068
AC016831.7
RP11−126L15.4
XLOC_051509
XLOC_040566
TRAM2−AS1
XLOC_010347
LINC00263
RP11−734K2.4
CTC−428G20.6
RP11−973D8.4
XLOC_026118
PVT1_2
LINC00909
EPB41L4A−AS1
RP11−96L14.7
SNHG1_1
RP11−1094M14.11
XLOC_014806_1
XLOC_029037
SNHG12_1
XLOC_014806_2
HeLa RNA-seq
log2 Fold Change
XLOC_017263
KB−1471A8.1
LINC00263
RP5−1148A21.3
PVT1
LINC00909
RP11−1094M14.11
CCAT1_2
LAMTOR5−AS1
LINC00680
CTC−428G20.6
MIR17HG
CCAT1_1
K562 RNA-seq
log2 Fold Change
LINC00910
LINC00909
LINC00263
LINC00680
EPB41L4A−AS1
XLOC_018520
MIR142_1
RP11−347C12.10
LINC00909
XLOC_042889
F
ATF3
CHAC1
SOST
SHH
DDIT4L
log2 Fold Change
AC016831.7_1
AC016831.7_2
CTC−428G20.6_1
CTC−428G20.6_2
EPB41L4A−AS1_1
EPB41L4A−AS1_2
LINC00263_1
LINC00263_2
LINC00680_1
LINC00680_2
MIR210HG_1
MIR210HG_2
MIR29A_1
MIR29A_2
RP11−1094M14.11_1
RP11−1094M14.11_2
RP11−734K2.4_1
RP11−734K2.4_2
RP11−96L14.7_1
RP11−96L14.7_2
RP11−973D8.4_1
RP11−973D8.4_2
SNHG1_1
SNHG1_2
SNHG12_1
SNHG12_2
TRAM2−AS1_1
TRAM2−AS1_2
XLOC_014806_1
XLOC_014806_2
XLOC_014806up_1
XLOC_014806up_2
XLOC_015111_1
XLOC_015111_2
XLOC_026118_1
XLOC_026118_2
XLOC_029037_1
XLOC_029037_2
XLOC_040566_1
XLOC_040566_2
XLOC_054068_1
XLOC_054068_2
sgRNA

Figure S8. lncRNA CRISPRi produces co-expressed transcriptome responses.

Heatmaps of differentially expressed genes across samples segregated by (A) U87 pro-growth, (B) U87 anti-growth, (C) HeLa anti-growth, (D) K562 anti-growth. Significant gene ontology terms are indicated next to the clusters in which they are enriched, annotated with p value and enrichment. Fold changes are relative to non-targeting controls within the same cell type. E) Distributions of absolute value $\log_2$ fold changes for differentially expressed genes (adj. $p < 0.05$) for each lncRNA knockdown. F) Panel of genes consistently upregulated or downregulated across multiple CRISPRi samples in U87.

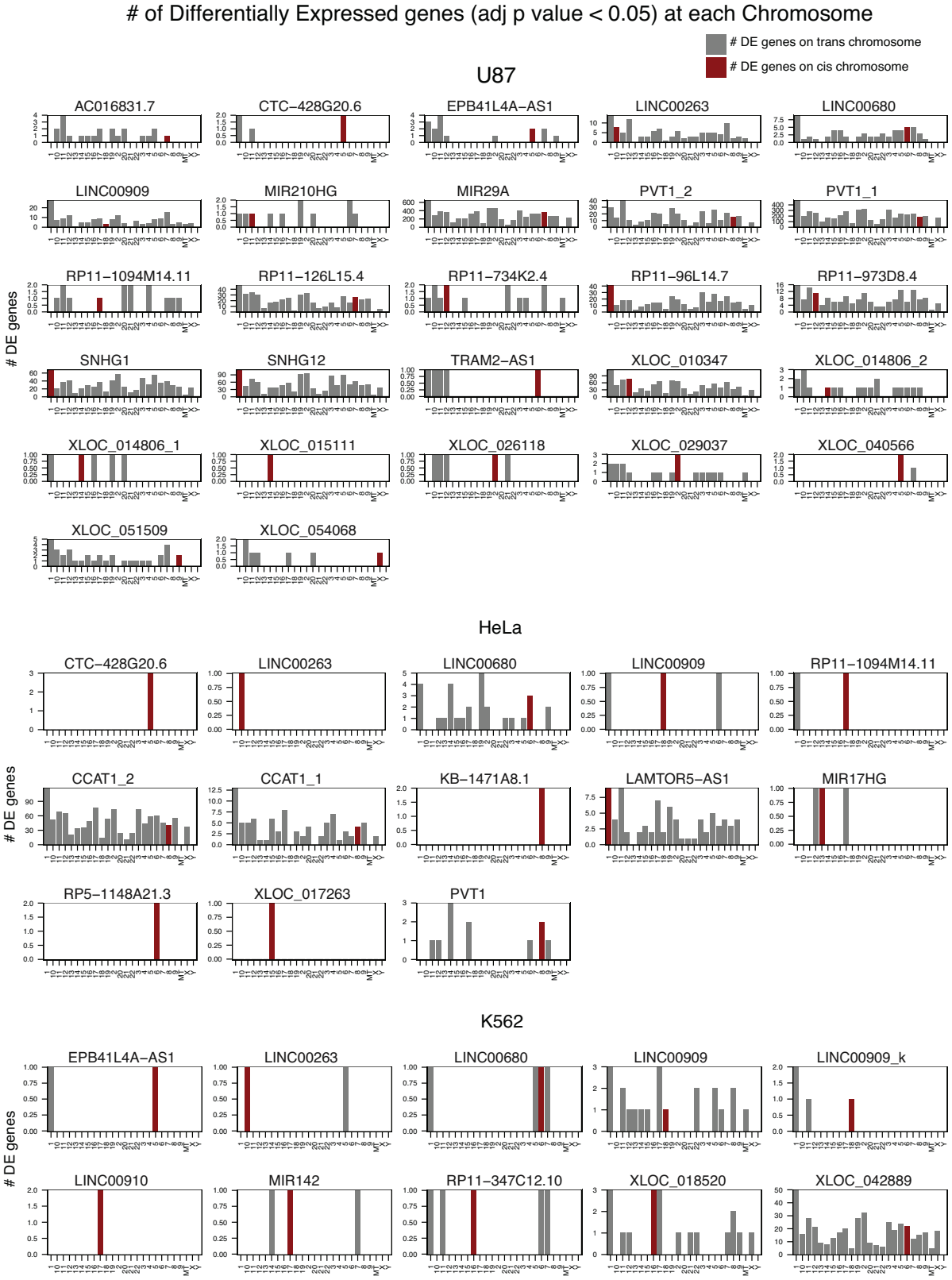

# of Differentially Expressed genes (adj p value < 0.05) at each Chromosome
DE genes on trans chromosome
DE genes on cis chromosome
U87
AC016831.7
CTC−428G20.6
EPB41L4A−AS1
LINC00263
LINC00680
LINC00909
MIR210HG
MIR29A
PVT1_2
PVT1_1
RP11−1094M14.11
RP11−126L15.4
RP11−734K2.4
RP11−96L14.7
RP11−973D8.4
SNHG1
SNHG12
TRAM2−AS1
XLOC_010347
XLOC_014806_2
XLOC_014806_1
XLOC_015111
XLOC_026118
XLOC_029037
XLOC_040566
XLOC_051509
XLOC_054068
DE genes
HeLa
CTC−428G20.6
LINC00263
LINC00680
LINC00909
RP11−1094M14.11
CCAT1_2
CCAT1_1
KB−1471A8.1
LAMTOR5−AS1
MIR17HG
RP5−1148A21.3
XLOC_017263
PVT1
DE genes
K562
EPB41L4A−AS1
LINC00263
LINC00680
LINC00909
LINC00909_k
LINC00910
MIR142
RP11−347C12.10
XLOC_018520
XLOC_042889
DE genes

Figure S9. Chromosome distribution of differentially expressed genes after each lncRNA knockdown.

Red bars indicate chromosomes harboring the lncRNAs of interest (*cis*). Gray bars represent chromosomes that do not contain the lncRNAs of interest (*trans*).

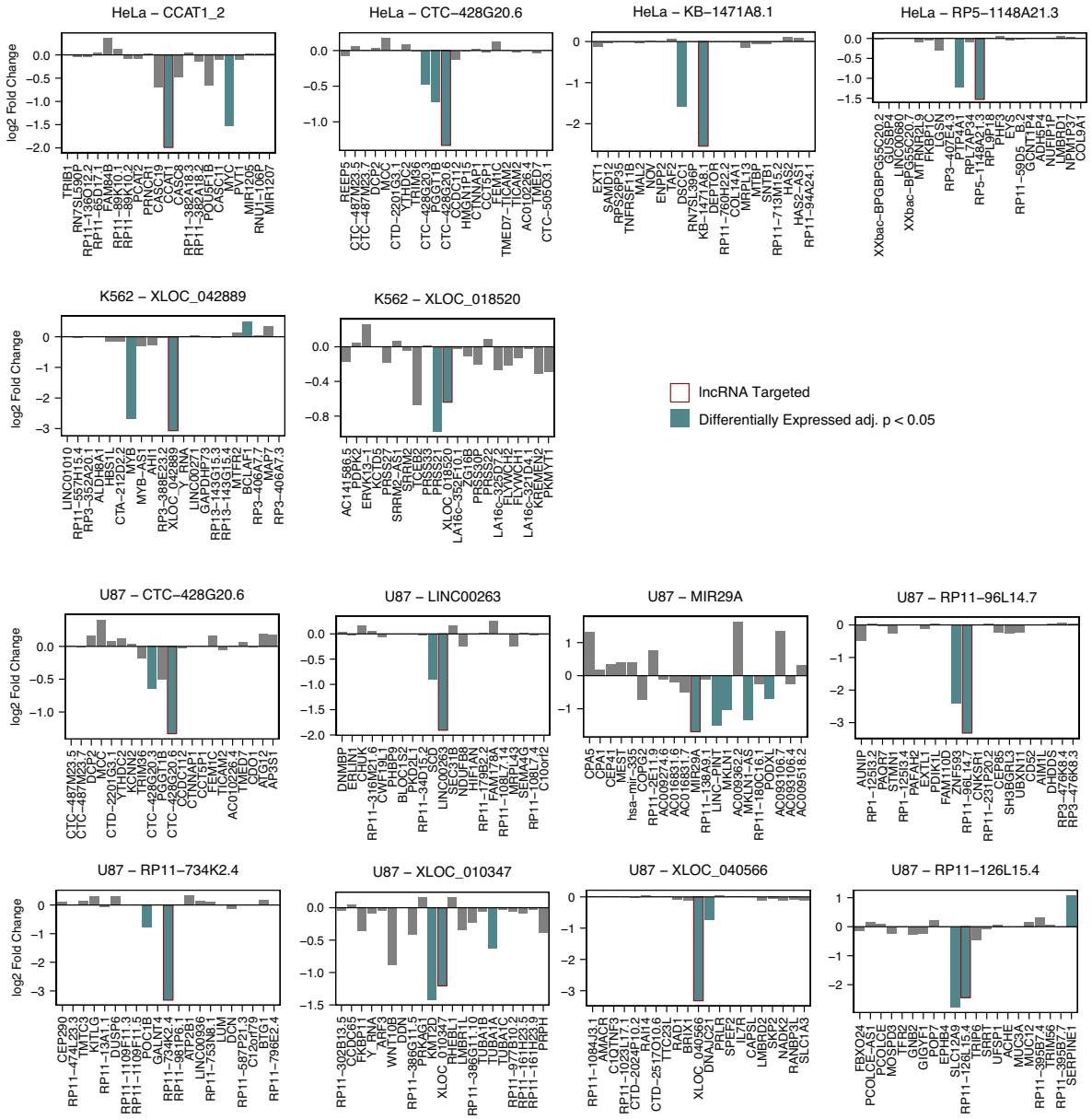

HeLa – CCAT1_2
HeLa – CTC–428G20.6
HeLa – KB–1471A8.1
HeLa – RP5–1148A21.3
K562 – XLOC_042889
K562 – XLOC_018520
lncRNA Targeted
Differentially Expressed adj. p < 0.05
U87 – CTC–428G20.6
U87 – LINC00263
U87 – MIR29A
U87 – RP11–96L14.7
U87 – RP11–734K2.4
U87 – XLOC_010347
U87 – XLOC_040566
U87 – RP11–126L15.4
log2 Fold Change

Figure S10. Local transcriptional changes within 20 gene-windows surrounding each lncRNA of interest following CRISPRi knockdown.

Red outline indicates targeted lncRNA. Blue bars indicate differentially expressed genes, among the broader set of differentially expressed genes genome-wide (DESeq2 adj. $p < 0.05$) upon knockdown.

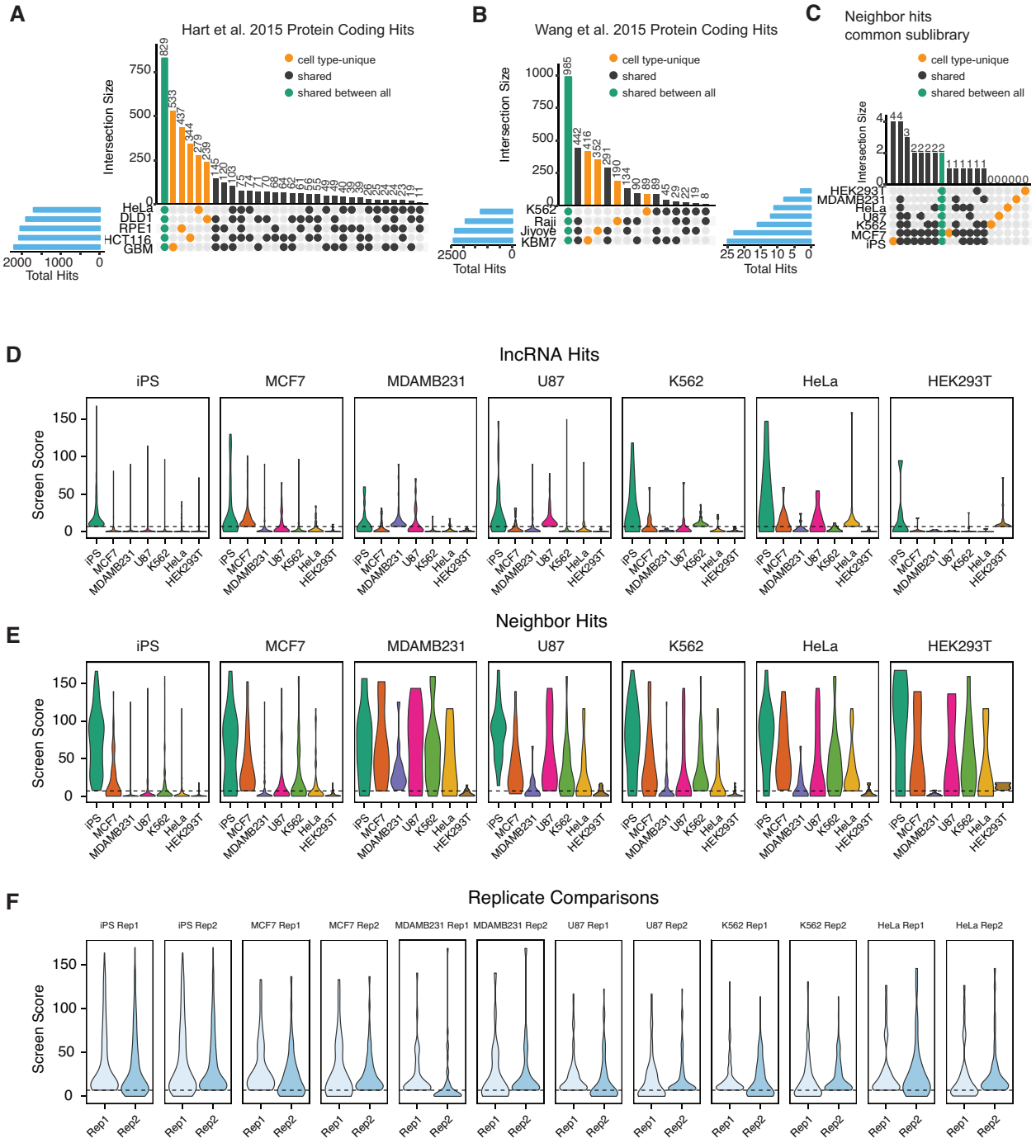

A
Hart et al. 2015 Protein Coding Hits
cell type-unique
shared
shared between all
Intersection Size
HeLa
DLD1
RPE1
HCT116
GBM
Total Hits
B
Wang et al. 2015 Protein Coding Hits
K562
Raji
Jiyoye
KBM7
C
Neighbor hits
common sublibrary
HEK293T
MDAMB231
HeLa
U87
K562
MCF7
iPS
D
lncRNA Hits
E
Neighbor Hits
F
Replicate Comparisons
Screen Score
iPS Rep1
iPS Rep2
MCF7 Rep1
MCF7 Rep2
MDAMB231 Rep1
MDAMB231 Rep2
U87 Rep1
U87 Rep2
K562 Rep1
K562 Rep2
HeLa Rep1
HeLa Rep2
Rep1
Rep2

Figure S11. lncRNA hit specificity is greater than essential protein coding gene specificity. A) Numbers of protein coding gene hits for each set of cell types screened in Hart et al. 2015 and (B) Wang et al. 2015. Wang et al. genes were considered hits if they passed a 5% false discovery threshold set by precision-recall analysis (*44*). C) Numbers of hits in our study that share promoters with essential protein coding genes (neighbor hits). Blue bars indicate total number of hits in each cell type. D) Distributions of screen scores across all cell types, for lncRNAs and (E) "neighbors" that were called hits in each given cell type. F) Distributions of screen scores across both replicates of each cell type, for lncRNAs that would be called as hits in replicate 1 (left) and in replicate 2 (right).

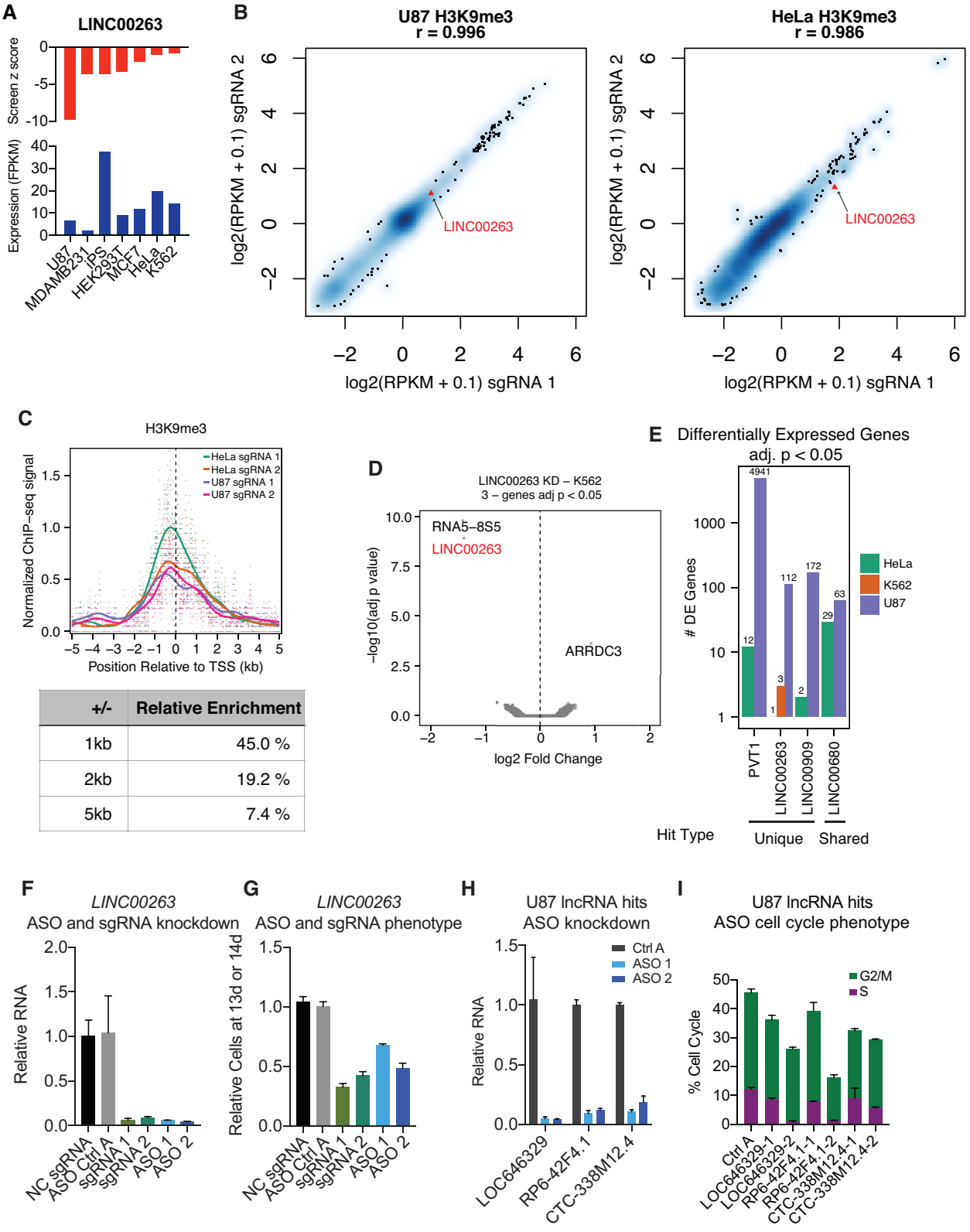

A
LINC00263
Screen z score
Expression (FPKM)
U87
MDAMB231
iPS
HEK293T
MCF7
HeLa
K562
B
U87 H3K9me3
r = 0.996
HeLa H3K9me3
r = 0.986
log2(RPKM + 0.1) sgRNA 1
log2(RPKM + 0.1) sgRNA 2
LINC00263
C
H3K9me3
HeLa sgRNA 1
HeLa sgRNA 2
U87 sgRNA 1
U87 sgRNA 2
Normalized ChIP−seq signal
Position Relative to TSS (kb)
+/- | Relative Enrichment
1kb | 45.0 %
2kb | 19.2 %
5kb | 7.4 %
D
LINC00263 KD − K562
3 − genes adj p < 0.05
RNA5−8S5
LINC00263
ARRDC3
−log10(adj p value)
log2 Fold Change
E
Differentially Expressed Genes
adj. p < 0.05
DE Genes
HeLa
K562
U87
PVT1
LINC00263
LINC00909
LINC00680
Hit Type
Unique
Shared
F
LINC00263
ASO and sgRNA knockdown
Relative RNA
NC sgRNA
ASO Ctrl A
sgRNA 1
sgRNA 2
ASO 1
ASO 2
G
LINC00263
ASO and sgRNA phenotype
Relative Cells at 13d or 14d
H
U87 lncRNA hits
ASO knockdown
Ctrl A
ASO 1
ASO 2
Relative RNA
LOC646329
RP6-42F4.1
CTC-338M12.4
I
U87 lncRNA hits
ASO cell cycle phenotype
G2/M
S
% Cell Cycle
Ctrl A
LOC646329-1
LOC646329-2
RP6-42F4.1-1
RP6-42F4.1-2
CTC-338M12.4-1
CTC-338M12.4-2

Figure S12. Cell type-specificity of *LINC00263*.

A) Screen phenotype z scores (red) and expression values (blue) for *LINC00263* across the 7 cell types. B) Reproducibility of H3K9me3 ChIP-seq between sgRNA 1 and sgRNA 2 targeting the TSS of *LINC00263* in U87 (left) and HeLa (right) cells. C) ChIP-seq enrichment of H3K9me3 surrounding the TSS of LINC00263, comparing 2 independent sgRNAs in U87 and HeLa cells. Smoothed lines were obtained by applying a Gaussian kernel smoother against ChIP-seq coverage that had been background subtracted with coverage of H3K9me3 in cells infected with non-targeting control sgRNAs. Signal was then normalized to the peak of the highest smoothed line. Table summarizes relative enrichment of H3K9me3 at various distances beyond the TSS, obtained from the median value of the smoothed lines at each distance. D) Volcano plots for RNA-seq differential expression following infection of *LINC00263* sgRNAs compared to infection of non-targeting sgRNAs in K562 cells. E) Numbers of differentially expressed genes (DESeq2 adj $p < 0.05$) following knockdown of lncRNAs in HeLa, K562, and U87 cells. For each gene, the same sgRNAs were used across the cell types. F) qPCR comparing *LINC00263* knockdown using CRISPRi and ASO. G) Proportion of cells at 14 days post sgRNA infection, or 13 days post ASO transfection against *LINC00263*, relative to control sgRNA or control ASO, respectively. H) qPCR of ASO knockdown of additional lncRNA hits in U87 cells. I) Percentage of cells in S and G2/M phases following ASO knockdown of additional lncRNA hits in U87.

Supplementary Tables

Table S1. TSS Annotations

Table S2. CRiNCL library sgRNAs

Table S3. Growth screen sgRNA read counts and phenotypes

Table S4. Growth screen gene phenotypes and p-values

Table S5. OCT4 screen sgRNA read counts and phenotypes

Table S6. OCT4 screen gene phenotypes and p-values

Table S7. iPSC protein-coding screen sgRNA and gene phenotypes

Table S8. PVT1 tiling library sgRNAs and phenotypes

Table S9. Differential expression of genes following lncRNA CRISPRi

Table S10. Genomic Properties of lncRNAs

Table S11. Individually cloned sgRNAs and primer pairs used in this study

References

Adamson, B., Smogorzewska, A., Sigoillot, F.D., King, R.W., and Elledge, S.J. (2012). A genome-wide homologous recombination screen identifies the RNA-binding protein RBMX as a component of the DNA-damage response. Nat Cell Biol *14*, 318–328.

Adey, A., Burton, J.N., Kitzman, J.O., Hiatt, J.B., Lewis, A.P., Martin, B.K., Qiu, R., Lee, C., and Shendure, J. (2013). The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. Nature *500*, 207–211.

Amabile, A., Migliara, A., Capasso, P., Biffi, M., Cittaro, D., Naldini, L., and Lombardo, A. (2016). Inheritable Silencing of Endogenous Genes by Hit-and-Run Targeted Epigenetic Editing. Cell *167*, 219–232.e14.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol *11*, R106.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461.

Aparicio-Prat, E., Arnan, C., Sala, I., Bosch, N., Guigó, R., and Johnson, R. (2015). DECKO: Single-oligo, dual-CRISPR deletion of genomic elements including long non-coding RNAs. BMC Genomics *16*, 846.

Bassett, A.R., Akhtar, A., Barlow, D.P., Bird, A.P., Brockdorff, N., Duboule, D., Ephrussi, A., Ferguson-Smith, A.C., Gingeras, T.R., Haerty, W., et al. (2014). Considerations when investigating lncRNA function in vivo. eLife *3*, e03058.

Braun, C.J., Bruno, P.M., Horlbeck, M.A., Gilbert, L.A., Weissman, J.S., and Hemann, M.T. (2016). Versatile in vivo regulation of tumor phenotypes by dCas9-mediated transcriptional perturbation. Proceedings of the National Academy of Sciences *113*, E3892–E3900.

Briggs, J.A., Wolvetang, E.J., Mattick, J.S., Rinn, J.L., and Barry, G. (2015). Mechanisms of Long Non-coding RNAs in Mammalian Nervous System Development, Plasticity, Disease, and Evolution. Neuron *88*, 861–877.

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

Chen, J., Shishkin, A.A., Zhu, X., Kadri, S., Maza, I., Guttman, M., Hanna, J.H., Regev, A., and Garber, M. (2016). Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. Genome Biol *17*, 19.

Consortium, T.E.P. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. Science *306*, 636–640.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. Nature *489*, 101–108.

Engreitz, J.M., Haines, J.E., Munson, G., Chen, J., and Perez, E.M. (2016a). Neighborhood regulation by lncRNA promoters, transcription, and splicing. bioRxiv.

Engreitz, J.M., Haines, J.E., Perez, E.M., Munson, G., Chen, J., Kane, M., McDonel, P.E.,

Guttman, M., and Lander, E.S. (2016b). Local regulation of gene expression by lncRNA promoters, transcription and splicing. Nature *539*, 452–455.

Engreitz, J.M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., Kadri, S., Xing, J., Goren, A., Lander, E.S., et al. (2013). The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. Science *341*, 1237973–1237973.

FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest, A.R.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J.L., Haberle, V., Lassmann, T., Kulakovskiy, I.V., Lizio, M., et al. (2014). A promoter-level mammalian expression atlas. Nature *507*, 462–470.

Fulco, C.P., Munschauer, M., Anyoha, R., Munson, G., Grossman, S.R., Perez, E.M., Kane, M., Cleary, B., Lander, E.S., and Engreitz, J.M. (2016). Systematic mapping of functional enhancer-promoter connections with CRISPR interference. Science 1–13.

Gilbert, L.A., Horlbeck, M.A., Adamson, B., Villalta, J.E., Chen, Y., Whitehead, E.H., Guimaraes, C., Panning, B., Ploegh, H.L., Bassik, M.C., et al. (2014). Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. Cell *159*, 647–661.

Gilbert, L.A., Larson, M.H., Morsut, L., Liu, Z., Brar, G.A., Torres, S.E., Stern-Ginossar, N., Brandman, O., Whitehead, E.H., Doudna, J.A., et al. (2013). CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell *154*, 442–451.

Groff, A.F., Sanchez-Gomez, D.B., Soruco, M.M.L., Gerhardinger, C., Barutcu, A.R., Li, E., Elcavage, L., Plana, O., Sanchez, L.V., Lee, J.C., et al. (2016). In Vivo Characterization of Linc-p21 Reveals Functional cis-Regulatory DNA Elements. Cell Reports *16*, 2178–2186.

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. (2011). Long non-coding RNA HOTAIR reprograms chromatinstate to promote cancer metastasis. Nature *464*, 1071–1076.

Gutschner, T., Hämmerle, M., Eissmann, M., Hsu, J., Kim, Y., Hung, G., Revenko, A., Arun, G., Stentrup, M., Gross, M., et al. (2013). The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. Cancer Research *73*, 1180–1189.

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. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature *477*, 295–300.

Hart, T., Brown, K.R., Sircoulomb, F., Rottapel, R., and Moffat, J. (2014). Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. Molecular Systems Biology *10*, 733–733.

Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K.R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., et al. (2015). High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell *163*, 1515–1526.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-enhancers in the control of cell identity and disease. Cell *155*, 934–947.

Ho, T.-T., Zhou, N., Huang, J., Koirala, P., Xu, M., Fung, R., Wu, F., and Mo, Y.-Y. (2014). Targeting non-coding RNAs with the CRISPR/Cas9 system in human cell lines. Nucleic Acids Res. *43*, gku1198–e17.

Horlbeck, M.A., Gilbert, L.A., Villalta, J.E., Adamson, B., Pak, R.A., Chen, Y., Fields, A.P., Park, C.Y., Corn, J.E., Kampmann, M., et al. (2016). Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. eLife *5*.

Igal, R.A. (2011). Roles of StearoylCoA Desaturase-1 in the Regulation of Cancer Cell Growth, Survival and Tumorigenesis. Cancers 2011, Vol. 3, Pages 2462-2477 *3*, 2462–2477.

Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner, J.R., Evans, J.R., Zhao, S., et al. (2015). The landscape of long noncoding RNAs in the human transcriptome. Nature Genetics.

Jiang, M., and Milner, J. (2002). Selective silencing of viral gene expression in HPV-positive human cervical carcinoma cells treated with siRNA, a primer of RNA interference. Oncogene *21*, 6041–6048.

Kampmann, M., Bassik, M.C., and Weissman, J.S. (2013). Integrated platform for genome-wide screening and construction of high-density genetic interaction maps in mammalian cells. Proceedings of the National Academy of Sciences *110*, E2317–E2326.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat Meth *12*, 357–360.

Kornienko, A.E., Guenzl, P.M., Barlow, D.P., and Pauler, F.M. (2013). Gene regulation by the act of long non-coding RNA transcription. BMC Biol. *11*, 59.

Kreitzer, F.R., Salomonis, N., Sheehan, A., Huang, M., Park, J.S., Spindler, M.J., Lizarraga, P., Weiss, W.A., So, P.-L., and Conklin, B.R. (2013). A robust method to derive functional neural

crest cells from human pluripotent stem cells. Am J Stem Cells *2*, 119–131.

Kretz, M., Siprashvili, Z., Chu, C., Webster, D.E., Zehnder, A., Qu, K., Lee, C.S., Flockhart, R.J., Groff, A.F., Chow, J., et al. (2013). Control of somatic tissue differentiation by the long non-coding RNA TINCR. Nature *493*, 231–235.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Meth *9*, 357–359.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

Li, J., Han, L., Roebuck, P., Diao, L., Liu, L., Yuan, Y., Weinstein, J.N., and Liang, H. (2015). TANRIC: An Interactive Open Platform to Explore the Function of lncRNAs in Cancer. Cancer Research *75*, 3728–3737.

Li, W., Notani, D., Ma, Q., Tanasa, B., Nunez, E., Chen, A.Y., Merkurjev, D., Zhang, J., Ohgi, K., Song, X., et al. (2013). Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature *498*, 516–520.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930.

Lin, N., Chang, K.-Y., Li, Z., Gates, K., Rana, Z.A., Dang, J., Zhang, D., Han, T., Yang, C.-S., Cunningham, T.J., et al. (2014). An evolutionarily conserved long noncoding RNA TUNA controls pluripotency and neural lineage commitment. Molecular Cell *53*, 1005–1019.

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., et al. (2016). Single-cell analysis of long non-coding RNAs in the developing human neocortex. Genome Biol *17*, 67.

Ma, W., Ay, F., Lee, C., Gulsoy, G., Deng, X., Cook, S., Hesson, J., Cavanaugh, C., Ware, C.B., Krumm, A., et al. (2015). Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. Nat Meth *12*, 71–78.

Mandegar, M.A., Huebsch, N., Frolov, E.B., Shin, E., Truong, A., Olvera, M.P., Chan, A.H., Miyaoka, Y., Holmes, K., Spencer, C.I., et al. (2016). CRISPR Interference Efficiently Induces Specific and Reversible Gene Silencing in Human iPSCs. Cell Stem Cell *18*, 541–553.

McCleland, M.L., Mesh, K., Lorenzana, E., Chopra, V.S., Segal, E., Watanabe, C., Haley, B., Mayba, O., Yaylaoglu, M., Gnad, F., et al. (2016). CCAT1 is an enhancer-templated RNA that predicts BET sensitivity in colorectal cancer. J. Clin. Invest. *126*, 639–652.

Meller, V.H., and Rattner, B.P. (2002). The roX genes encode redundant male-specific lethal transcripts required for targeting of the MSL complex. The EMBO Journal *21*, 1084–1091.

Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker, J.C., Grützner, F., and Kaessmann, H. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature *505*, 635–640.

Nishimasu, H., Cong, L., Yan, W.X., Ran, F.A., Zetsche, B., Li, Y., Kurabayashi, A., Ishitani, R., Zhang, F., and Nureki, O. (2015). Crystal Structure of Staphylococcus aureus Cas9. Cell *162*, 1113–1126.

O'Geen, H., Echipare, L., and Farnham, P.J. (2011). Using ChIP-seq technology to generate

high-resolution profiles of histone modifications. Methods Mol. Biol. *791*, 265–286.

Paralkar, V.R., Taborda, C.C., Huang, P., Yao, Y., Kossenkov, A.V., Prasad, R., Luan, J., Davies, J.O.J., Hughes, J.R., Hardison, R.C., et al. (2016). Unlinking an lncRNA from Its Associated cis Element. Molecular Cell *62*, 104–110.

Pomerantz, M.M., Ahmadiyeh, N., Jia, L., Herman, P., Verzi, M.P., Doddapaneni, H., Beckwith, C.A., Chan, J.A., Hills, A., Davis, M., et al. (2009). The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nature Genetics *41*, 882–884.

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

Qi, L.S., Larson, M.H., Gilbert, L.A., Doudna, J.A., Weissman, J.S., Arkin, A.P., and Lim, W.A. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell *152*, 1173–1183.

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. *44*, W160–W165.

Ramos, A.D., Andersen, R.E., Liu, S.J., Nowakowski, T.J., Hong, S.J., Gertz, C.C., Salinas, R.D., Zarabi, H., Kriegstein, A.R., and Lim, D.A. (2015). The long noncoding RNA Pnky regulates neuronal differentiation of embryonic and postnatal neural stem cells. Cell Stem Cell *16*, 439–447.

Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T.,

Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680.

Rinn, J.L., and Chang, H.Y. (2012). Genome regulation by long noncoding RNAs. Annu. Rev. Biochem. *81*, 145–166.

Salomonis, N., Dexheimer, P.J., Omberg, L., Schroll, R., Bush, S., Huo, J., Schriml, L., Ho Sui, S., Keddache, M., Mayhew, C., et al. (2016). Integrated Genomic Analysis of Diverse Induced Pluripotent Stem Cells from the Progenitor Cell Biology Consortium. Stem Cell Reports *7*, 110–125.

Sauvageau, M., Goff, L.A., Lodato, S., Bonev, B., Groff, A.F., Gerhardinger, C., Sanchez-Gomez, D.B., Hacisuleyman, E., Li, E., Spence, M., et al. (2013). Multiple knockout mouse models reveal lincRNAs are required for life and brain development. eLife *2*, e01749.

Shalem, O., Sanjana, N.E., Hartenian, E., Shi, X., Scott, D.A., Mikkelsen, T.S., Heckl, D., Ebert, B.L., Root, D.E., Doench, J.G., et al. (2014). Genome-scale CRISPR-Cas9 knockout screening in human cells. Science *343*, 84–87.

Shi, J., Wang, E., Milazzo, J.P., Wang, Z., Kinney, J.B., and Vakoc, C.R. (2015). Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. Nat Biotechnol *33*, 661–667.

Stadhouders, R., Kolovos, P., Brouwer, R., Zuin, J., van den Heuvel, A., Kockx, C., Palstra, R.-J., Wendt, K.S., Grosveld, F., van Ijcken, W., et al. (2013). Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. Nat Protoc *8*, 509–524.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., and Tomoda, K. (2007). Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. Stem Cells *131*, 1–12.

Thakore, P.I., D'Ippolito, A.M., Song, L., Safi, A., Shivakumar, N.K., Kabadi, A.M., Reddy, T.E., Crawford, G.E., and Gersbach, C.A. (2015). Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. Nat Meth *12*, 1143–1149.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol *28*, 511–515.

Tseng, Y.-Y., Moriarity, B.S., Gong, W., Akiyama, R., Tiwari, A., Kawakami, H., Ronning, P., Reuland, B., Guenther, K., Beadnell, T.C., et al. (2014). PVT1 dependence in cancer with MYC copy-number increase. Nature *512*, 82–86.

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

van de Werken, H.J.G., Landan, G., Holwerda, S.J.B., Hoichman, M., Klous, P., Chachik, R., Splinter, E., Valdes-Quezada, C., Oz, Y., Bouwman, B.A.M., et al. (2012). Robust 4C-seq data analysis to screen for regulatory DNA interactions. Nat Meth *9*, 969–972.

Visel, A., Minovitsky, S., Dubchak, I., and Pennacchio, L.A. (2007). VISTA Enhancer Browser--a database of tissue-specific human enhancers. Nucleic Acids Res. *35*, D88–D92.

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. (2011). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature *472*, 120–124.

Wang, T., Birsoy, K., Hughes, N.W., Krupczak, K.M., Post, Y., Wei, J.J., Lander, E.S., and Sabatini, D.M. (2015). Identification and characterization of essential genes in the human genome. Science *350*, 1096–1101.

Wang, T., Wei, J.J., Sabatini, D.M., and Lander, E.S. (2014). Genetic screens in human cells using the CRISPR-Cas9 system. Science *343*, 80–84.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307–319.

Wu, L., Zhang, X., Zhao, Z., Wang, L., Li, B., Li, G., Dean, M., Yu, Q., Wang, Y., Lin, X., et al. (2015). Full-length single-cell RNA-seq applied to a viral human cancer: applications to HPV expression and splicing analysis in HeLa S3 cells. Gigascience *4*, 51.

Xiang, J.-F., Yin, Q.-F., Chen, T., Zhang, Y., Zhang, X.-O., Wu, Z., Zhang, S., Wang, H.-B., Ge, J., Lu, X., et al. (2014). Human colorectal cancer-specific CCAT1-L lncRNA regulates long-range chromatin interactions at the MYC locus. Cell Research *24*, 513–531.

Yan, X., Hu, Z., Feng, Y., Hu, X., Yuan, J., Zhao, S.D., Zhang, Y., Yang, L., Shan, W., He, Q., et al. (2015). Comprehensive Genomic Characterization of Long Non-coding RNAs across Human Cancers. Cancer Cell *28*, 529–540.

Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., et al. (2016). Ensembl 2016. Nucleic Acids Res. *44*, D710–D716.

Yin, Y., Yan, P., Lu, J., Song, G., Zhu, Y., Li, Z., Zhao, Y., Shen, B., Huang, X., Zhu, H., et al. (2015). Opposing Roles for the lncRNA Haunt and Its Genomic Locus in Regulating HOXA Gene Activation during Embryonic Stem Cell Differentiation. Cell Stem Cell *16*, 504–516.

Zeng, Y., and Cullen, B.R. (2002). RNA interference in human cells is restricted to the cytoplasm. Rna *8*, 855–860.

Zhu, S., Li, W., Liu, J., Chen, C.-H., Liao, Q., Xu, P., Xu, H., Xiao, T., Cao, Z., Peng, J., et al. (2016). Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR-Cas9 library. Nat Biotechnol.

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

Chapter 5

Genome-scale screening of lncRNAs reveals radiation sensitizers for glioma therapy

Summary

Glioblastoma (GBM) is a uniformly fatal cancer with a median survival of 14 months, despite advancements in surgery, chemotherapy, and radiation therapy. Therefore, there is urgency to not only develop novel therapy for GBM, but also discover ways to improve the efficacy of current therapeutics. Long non-coding RNAs (lncRNAs), which are a diverse class of non-protein producing transcripts transcribed throughout the genome, have been demonstrated to regulate gene expression in various cancers. To discover lncRNAs that could potentially be targeted for novel GBM therapy, we performed systematic genetic screens to identify lncRNAs whose knockdown sensitizes GBM cells to radiation induced cell death. After evaluating the model GBM cell line U87 for radiation sensitivity and transcriptional changes following clinically relevant doses of ionizing radiation, we performed CRISPRi-based screens targeting 5689 lncRNAs and found 269 that synergize with radiation therapy. We further identified lncRNAs whose knockdown increased radiation induced cell death but had little to no phenotype with no radiation treatment. We then validated *LINC00339* as a potential lncRNA target for radiation sensitization. Although further validation in patient derived cell lines and xenograft models are required to further understand these novel lncRNAs, our study demonstrates how genetic vulnerabilities to existing therapeutics can be identified at large scale.

Introduction

Glioblastoma multiforme (GBM) is the most common and the most fatal primary tumor in the brain(Thomas et al., 2014). While advances in next generation sequencing and 'omics profiling have improved our understanding of the molecular pathways and genetic aberrations in GBM (Brennan et al., 2013; Verhaak et al., 2010), current treatments – which includes surgical resection, radiation therapy, and the alkylating agent temozolomide – can only prolong survival to a median of 14-16 months (Raizer and Parsa, 2014; Thomas et al., 2014). Therefore, there is a pressing need to develop more efficacious therapeutics, or agents that can increase the potency of current therapies while minimizing toxicity to normal tissues.

High dose fractionated ionizing radiation (IR) is a critical part of the standard of care for GBM, due to the ability of IR to control microscopic disease not resected by surgery (Raizer and Parsa, 2014). However, disease response to IR is variable (Fouse et al., 2014), and radiation resistance is a major issue. Increasing the sensitivity of tumors to IR therapy and overcoming radiation resistance have been approaches previously taken to improve upon the current standard of care, often through perturbation of relatively well understood cellular pathways such as DNA damage response, survival signaling, and immune responses (Badie et al., 1999; De Bacco et al., 2016; Higgins et al., 2010; Naidu et al., 2010; Reichert et al., 2016; Zheng et al., 2008). However, there has not been genome scale, systematic efforts to identify genes that could be therapeutically targeted for radiation sensitization.

Long non-coding RNAs (lncRNAs), are transcripts > 200 nt long that are pervasively transcribed throughout the genome (Cabili et al., 2011; Derrien et al., 2012; Iyer et al., 2015). Functional studies have implicated lncRNAs in the pathophysiology of several diseases, including cancer (Batista and Chang, 2013; Gupta et al., 2011). In particular, several lncRNAs

such as *DINO* (Schmitt et al., 2016), *NORAD* (Lee et al., 2015), *PARTICLE* (O'Leary et al., 2015), and *DDSR1* (Sharma et al., 2015) have been shown to function in DNA damage response and genome stability, which are pathways that could be targeted for radiation sensitization. A common theme that has emerged from lncRNA research is that both lncRNA expression patterns and functionality are exquisitely cell type- and disease type-specific (Cabili et al., 2011; Liu et al., 2017; 2016). These features make lncRNAs attractive targets for GBM therapy, because targeting lncRNAs might preferentially treat the disease while sparing normal tissue.

CRISPR interference (CRISPRi) is an engineered CRISPR/Cas9 system that represses transcription through recruitment of catalytically-inactive Cas9 fused to a KRAB repressor domain (dCas9-KRAB) to the transcription start site (TSS) of any gene by a single guide RNA (sgRNA). We previously developed a CRISPRi platform specifically targeting 16,401 lncRNA loci across 7 different cancer and non-cancer cell types (Liu et al., 2017). 499 lncRNA loci were identified as required for robust growth and proliferation, in the absence of additional perturbations such as drugs. Here, we sought to identify and characterize lncRNAs that could be targeted to sensitize cells to radiation therapy, and therefore potentially be candidates for improved GBM treatment. After evaluating a model *in vitro* cell line, U87, for use in radiation studies, we analyzed transcriptome-wide lncRNA expression changes in response to radiation. We then used CRISPRi to screen all lncRNAs expressed in U87 to identify lncRNAs whose knockdown increases radiation-induced cell death and senescence. *LINC00339* reproducibly sensitized U87 cells to radiation when validated by individual knockdown assays. Our study represents a key step toward developing novel GBM therapeutics and also demonstrates the use of CRISPRi to study complex cellular processes on a massive scale.

Results

U87 Glioblastoma cells are responsive to ionizing radiation

To begin to study genes that control cellular responses to radiation therapy, we first evaluated the effects of clinically relevant doses of ionization radiation on the proliferation and survival of U87 glioblastoma cells. The hypotriploid U87 glioblastoma line was chosen because we have previously used this as an *in vitro* platform for discovering novel lncRNA therapeutic targets for glioma using CRISPRi. U87 cells exhibited decreased growth rates when treated with single doses of 4 Gy and 8 Gy of ionization radiation, but not with 2 Gy (Figure 1). Interestingly, treatment with a single 4 Gy dose lead to the recovery of U87 cells to pre-treatment proliferation rates at 10 days post irradiation, consistent with an intact p53 response and the ability to repair DNA damage in U87 cells(Badie et al., 1999). Because clinical treatment for gliomas involves fractionated radiation (Barani and Larson, 2014), we also subjected U87 cultures to 4 doses of 2 Gy ionizing radiation every other day (4x 2Gy). The fractionated treatment decreased proliferation rates similarly to the single 4 Gy dose up to 8 days post irradiation, but unlike the single 4 Gy dose, the growth rate continued to decrease under the fractionated regimen. Therefore, U87 glioblastoma cells are a viable system for modeling response to radiation therapy.

Transcriptional profile of U87 cells to ionizing radiation

To understand potential mechanisms of radiation resistance and to identify lncRNAs that may play a role in the response to radiation, we performed transcriptome profiling using RNA-seq following treatment of U87 cultures to varying levels of ionizing radiation. Increasing amounts of radiation (2, 4, 8 Gy) resulted in dose-dependent differential gene expression (Figure 2A),

with 507 genes upregulated and 603 downregulated (adj p. value < 0.05) when all doses were considered (Figure 2B). Gene Ontology (GO) analysis identified upregulated genes belonging to the p53 signaling (adj p. value 3.34E-08) and TNF signaling pathways (adj p. value 8.46E-04). Downregulated genes were enriched for spliceosome components (adj p. value 7.85E-13) and DNA replication (adj p. value 9.97E-11), consistent with the decreased viability of U87 cells in response to radiation.

Differentially Expressed lncRNAs following radiation

Among the 1110 differentially expressed genes identified from RNA-seq following radiation, 30 were lncRNAs (Figure 2C). Similar to differentially expressed protein coding genes, these lncRNAs also altered in expression levels in a radiation dose-dependent manner. To ask whether any of these differentially expressed lncRNAs might be regulated by p53, we cross referenced the loci of these 30 lncRNAs (including their upstream 1kb promoter regions) with p53 ChIP-seq data from lymphoblastoid cells (Zeron-Medina et al., 2013). One lncRNA, *LINC01021* (also known as *LOC643401*), exhibited a significant p53 binding site at its promoter (Figure 2D), suggesting *LINC01021* might be a radiation responsive lncRNA that is induced through p53 activation.

CRISPRi screens to identify lncRNAs whose knockdown sensitize cells to radiation

CRISPRi screens have been previously performed to identify lncRNAs that are necessary for robust cellular growth and proliferation of U87 glioblastoma cells (Liu et al., 2017). Extending upon these results, we sought out to discover lncRNAs whose knockdown enhanced the efficacy of radiation therapy-induced cell death, or in other words lncRNAs that normally play a role in

radiation resistance. To do this, we generated lentiviral sgRNA libraries derived from the CRISPRi Non-Coding Library (CRiNCL) (Liu et al., 2017) specifically targeting lncRNAs expressed in U87 cells. This library, which targets 5689 lncRNA TSS's with 10 independent sgRNAs each and also includes negative control sgRNAs, was infected into duplicate cultures of U87 cells stably expressing dCas9-KRAB (Liu et al., 2016) at an MOI of ~1.0. Infected cells were selected by puromycin selection, and four doses of 2 Gy ionizing radiation was applied to the entire cultures every other day, for a total of 8 Gy radiation fractionated across 8 days. The cells were cultured for another 4 days without radiation, corresponding to a total of ~6 cell doublings over the course of 12 days under these conditions (Figure 3A). Enrichment or depletion of sgRNAs were measured by illumina sequencing sgRNA tags at the beginning and end of the 12 day period. sgRNA enrichment was then normalized by total number of cell doublings to obtain the quantitative phenotype τ, which reflects the negative or positive growth phenotype caused by knockdown of a gene in combination with radiation, compared to radiation with negative control knockdown (Kampmann et al., 2013).

As was observed in growth screens without radiation (Liu et al., 2017), the proportion of cells infected with the sgRNA library remained stable throughout the duration of the radiation modifier screen, indicating that the sgRNA library is not toxic to U87 cells in general, even while being irradiated (Figure 3B). To identify lncRNAs with cellular phenotypes in the presence of radiation, we first averaged the biological duplicate τ values for each sgRNA and calculated the mean phenotype of the top 3 sgRNAs targeting each gene along with the Mann-Whitney p-value from all 10 sgRNAs, which was obtained by comparing the beginning (T0) and end (T12) timepoints (Gilbert et al., 2014; Horlbeck et al., 2016; Liu et al., 2017). In addition, we randomly sampled from the set of non-targeting control sgRNAs to score "negative control genes" and

analyzed them as with lncRNA genes, allowing us to estimate the empirical false discovery rate based on what proportion of negative control genes score as hits (Liu et al., 2017). LncRNA genes were scored as hits if their combined phenotype effect size and p-value (referred to here as "screen score") surpassed a threshold corresponding to an empirical false discovery rate of 5%. We identified a total of 404 genes with significant radiation phenotypes, almost all of which caused decreased growth upon knockdown compared to radiation alone (Figure 3C, left panel). Because we cannot rule out the possibility of silencing adjacent neighbor genes, we separately categorized 135 of these hits as "neighbor hits" because their TSS's were within 1kb of *any* protein coding TSS (Figure 3D). Nonetheless, the 269 lncRNA hits identified in this radiation modifier screen far surpassed the 50 lncRNA hits identified in the growth-only screen performed previously (Liu et al., 2017) and analyzed here in an identical fashion as the radiation screen (Figure 3C, right panel).

As internal positive controls, we identified 2 protein coding genes with established DNA repair function – *RAD21* and *ERCC6L2* – whose transcription start sites were either overlapping or within 200bp of a lncRNA TSS that scored as a hit in the radiation modifier screen but not in the growth only screen (Figure 3E). Because the inter-TSS distances between these intended lncRNA targets and neighbor genes were within the window of activity of CRISPRi (Gilbert et al., 2014), we interpret these results as direct inhibition of both lncRNA and coding gene. *RAD21* and *ERCC6L2* are proteins that play roles in DNA double strand break repair and excision repair (Birkenbihl and Subramani, 1992; Tummala et al., 2014), respectively. The observation that knockdown of these lncRNAs results in enhanced radiation-induced cell death but have little phenotype when knocked down in the absence of radiation supports the ability of our screen to detect lncRNA gene knockdowns that synergize with radiation.

Identification and validation radiation modifier lncRNA *LINC00339*

We then searched for lncRNAs without protein coding gene neighbors that synergized with ionization radiation upon knockdown. Five of these lncRNAs had highly significant phenotypes with radiation but insignificant phenotypes without radiation (Figure 4A). One of these lncRNAs, *LINC00339*, was selected for further validation. *LINC00339* is an intergenic lncRNA approximately 13 kb downstream of *CELA3A* and 21 kb upstream of *CDC42* (Figure 4B). Internally controlled relative growth assays were performed by individually cloning two sgRNAs targeting each of the two TSS's of *LINC00339.* Infection of all four sgRNAs resulted in decreased cellular proliferation when combined with four doses of 2Gy radiation (Figure 4C), consistent with the screen results.

Discussion

By utilizing both an expression profiling approach (RNA-seq) and a functional genomics approach (CRISPRi screen), we have begun to understand the role of lncRNAs in the cellular response to ionizing radiation. 269 lncRNA loci exhibited significant growth phenotypes in U87 glioblastoma cells when knocked down with CRISPRi in combination with clinically relevant doses of ionizing radiation, far greater than the number of lncRNAs that affected normal growth and proliferation withour radiation. Interestingly, all but one of these lncRNA radiation hits resulted in decreased growth upon knockdown, compared to cells that received radiation only. These results suggest that many lncRNAs normally function to keep U87 cells resistant to radiation-induced cell death, hence promoting the fitness of the tumor. It also suggests that lncRNAs more frequently play roles in higher order processes such as DNA damage response, compared to essential growth and proliferation. The additional validation of intergenic lncRNA *LINC00339* further demonstrates that radiation hit lncRNAs can be individually targeted for potential therapeutic use, although antisense oligonucleotide (ASO)-based knockdown of lncRNAs may be better primed for therapeutic use in the clinical setting (Meng et al., 2015; Monteleone et al., 2015).

The lncRNAs discovered in this study represent a large step toward developing radiation sensitizer targets in GBM therapy. However, future genome-scale screens should be performed with both radiation and temozolomide treatment, to more accurately mimic the standard of care for patients with GBM(Omuro and DeAngelis, 2013). Similar radiation modifier screens could also be performed against the protein coding genome, as has been done to study drug resistance and sensitivity (Gilbert et al., 2014; Kampmann et al., 2013; Konermann et al., 2015). The interplay between tumor and microenvironment must also not be discounted in future studies, as

immune modulating therapy may synergize with radiation(Kalbasi et al., 2013). To extend the clinical relevance of our results, candidate lncRNAs that were identified in the CRISPRi radiation-modifier screen should also be tested in recently derived tumor cultures (Müller et al., 2016), in addition to xenograft models of brain tumors (Hashizume et al., 2014). Nonetheless, the rapid identification of genetic vulnerabilities to established treatments may enable the next generation of cancer therapy.

Experimental Procedures

Cell culture and radiation treatment

U87 cells were grown in DMEM with 10% FBS and antibiotics/antimycotics. Radiation was applied using a Cesium-137 irradiator with rotating platform, with cells in suspension for the screen and adherent on plates for validation. Proliferation was measured using a Countess automated hemocytometer (Agilent).

RNA-seq sample preparation and data analysis

RNA was harvested using TRIzol 48 hours following radiation treatment and purified using the Direct-zol MiniPrep RNA purification kits (Zymo Research) with the on-column DNAse digestion step. RNA integrity was confirmed using the Agilent Bioanalyzer. RNA-seq libraries were generated using TruSeq Stranded mRNA kit according to manufacturer's protocol (illumina). cDNA was validated using the Agilent Bioanalyzer, Qubit 2.0 Fluorometer (Life Technologies), and ddPCR (Bio-Rad). Cluster generation and sequencing was performed on a HiSeq 2500, using the paired end 100 read protocol.

Reads were aligned to the human genome (GRCh37) using the spliced read aligner HISAT2 v2.0.3 (Kim et al., 2015) against an index containing SNP and transcript information (*genome_snp_tran*). Quantification of Ensembl build 75 genes was carried out with featureCounts (Liao et al., 2014) using only uniquely mapped reads. Differential expression analysis was performed using DESeq2 (Anders and Huber, 2010) using the Wald test with an adjusted p value threshold of 0.05 as threshold for differential expression. Complete linkage hierarchical clustering was performed using 1 – Pearson correlation coefficients as the distance matrix, using only differentially expressed genes or differentially expressed lncRNAs. Gene

ontology terms were obtained using Enrichr (Chen et al., 2013), taking gene names from the clusters of all upregulated or downregulated genes.

ChIP-seq data

P53 genomic occupancy data were obtained by analyzing ChIP-seq peaks from GEO accession GSE46993, and intersecting all lncRNA loci +/- 1kb with the p53 peaks.

CRISPRi screens

sgRNA library was derived from the CRISPRi Non-Coding Library (CRiNCL) (Liu et al., 2017), selecting sub-libraries that targets all expressed lncRNAs in U87: Common + Cancer common + (U87, HEK293T) + U87 unique. sgRNAs were cloned into the library expression vector pCRISPRia-v2 (Gilbert et al., 2014; Horlbeck et al., 2016) and lentivirus pools were generated as previously described (Liu et al., 2017). U87-dCas9-KRAB cells were generated previously in (Liu et al., 2016). Lentivirus libraries were infected in duplicate cultures, cultured for 2 days following infection, puromycin (1 mg/mL) selected for two days, and recovered for one day without puromycin. Cells were then cultured for 12 days at a minimum coverage of 1000x, starting at this "T0." For the radiation modifier screen, doses of 2 Gy radiation were given at the following days: T0, T2, T4, T6, for a total of 8 Gy fractionated ionizing radiation. Genomic DNA was harvested from aliquots of ~60M cells each from T0 and T12 and processed for sequencing as previously described (Gilbert et al., 2014; Horlbeck et al., 2016). Data processing and hit analysis was performed as described in (Liu et al., 2017), with the exception that neighbor hits were considered to be hits whose TSS's were within 1kb of ANY protein coding gene TSS. Growth-only screen data for U87 was obtained from (Liu et al., 2017) and compared

to the radiation screen data obtained in this study. Growth and radiation z scores were calculated as log2 enrichment normalized by the standard deviation of negative control genes. sgRNA validation was performed as described in (Liu et al., 2017), with the addition of 4x 2 Gy radiation starting two days following sgRNA infection, fractionated every other day.

Acknowledgements

We thank Melanie Bedolli for radiation therapy support, Joe Costello, Howard Chang, and Steve Braunstein for helpful discussions, Eric Chow and Derek Bogdanoff of the UCSF Center for Advanced Technology for sequencing assistance, and Laurakay Bruhn, Daniel Ryan, Luke Fairbairn, and Peter Tsang of Agilent Technologies for their assistance on the design and synthesis of oligonucleotide pools.

Figures

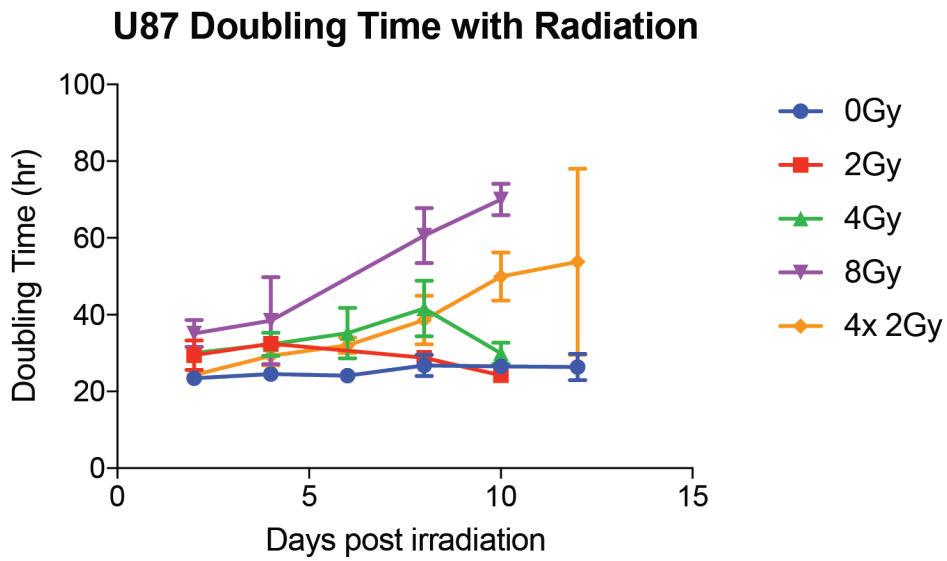
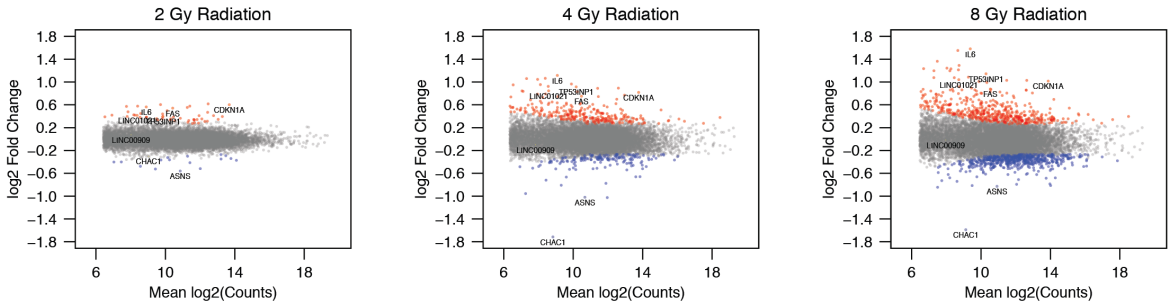

U87 Doubling Time with Radiation
Doubling Time (hr)
100
80
60
40
20
0
0
5
10
15
Days post irradiation
0Gy
2Gy
4Gy
8Gy
4x 2Gy

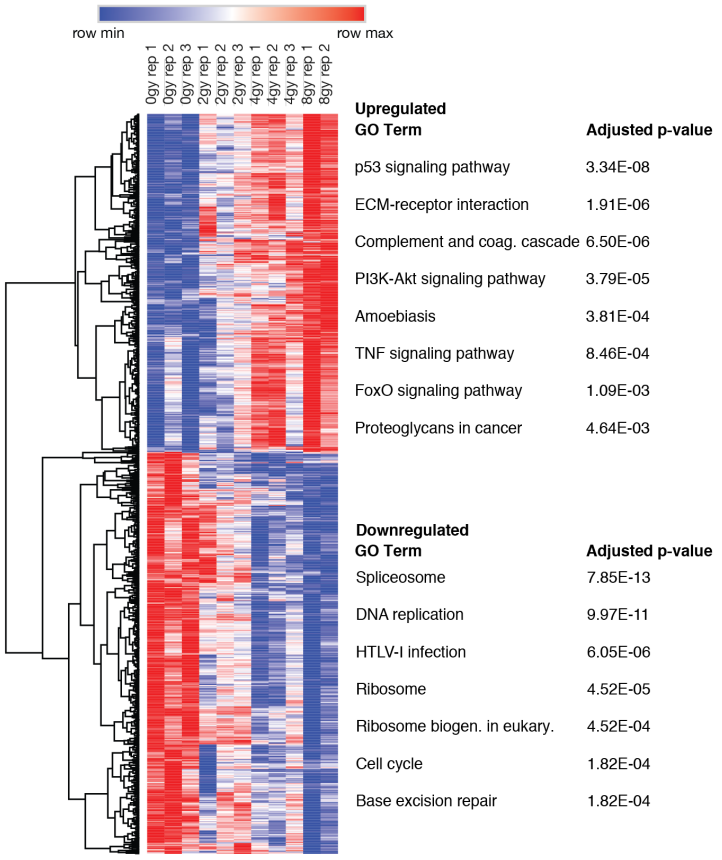
Figure 1. Clinically relevant doses of ionizing radiation slow down proliferation of U87 cells

Doubling times of U87 cells growth *in vitro* following treatment with varying doses of ionizing radiation. 4x 2Gy indicates 4 separate doses of 2Gy each, given every other day starting at day 0.

A



B



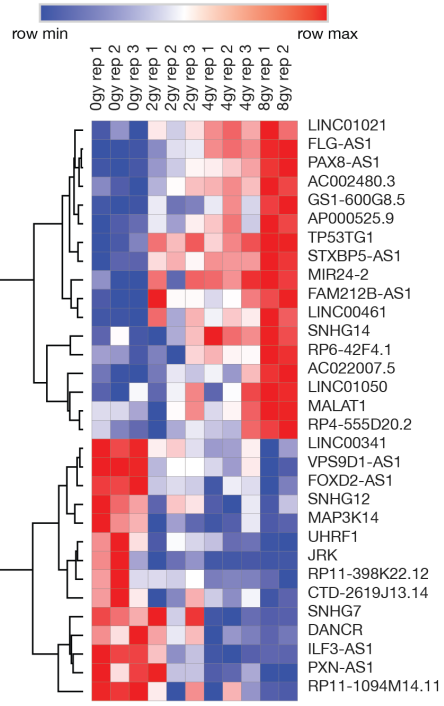
Upregulated

GO Term	Adjusted p-value
p53 signaling pathway	3.34E-08
ECM-receptor interaction	1.91E-06
Complement and coag. cascade	6.50E-06
PI3K-Akt signaling pathway	3.79E-05
Amoebiasis	3.81E-04
TNF signaling pathway	8.46E-04
FoxO signaling pathway	1.09E-03
Proteoglycans in cancer	4.64E-03

Downregulated

GO Term	Adjusted p-value
Spliceosome	7.85E-13
DNA replication	9.97E-11
HTLV-I infection	6.05E-06
Ribosome	4.52E-05
Ribosome biogen. in eukary.	4.52E-04
Cell cycle	1.82E-04
Base excision repair	1.82E-04

C



D

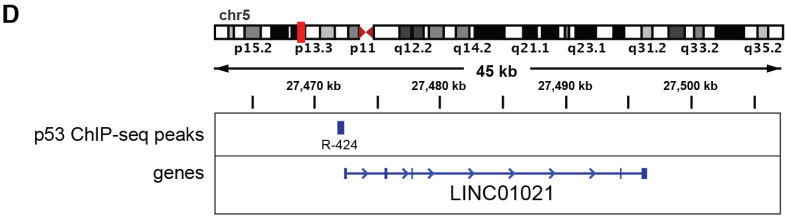
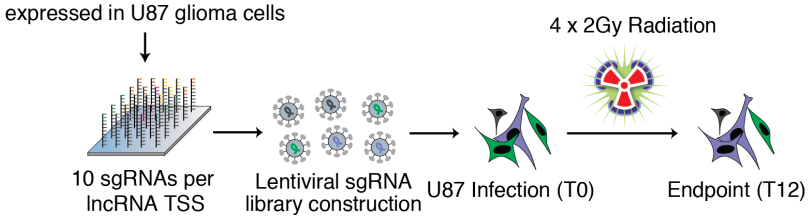


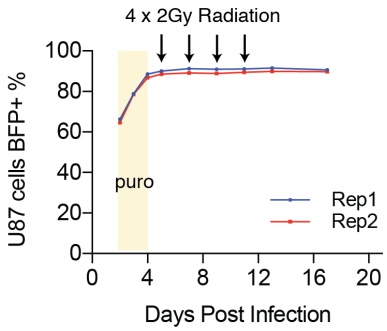
Figure 2. Transcriptome profiling following ionizing radiation in U87 cells

A) MA plots of RNA-seq analysis 48 hr after treatment of U87 cells *in vitro* with 2, 4, and 8 Gy of ionizing radiation. Upregulated genes (adj p value < 0.05) are labeled in red, and downregulated genes (adj p value < 0.05) are labeled in blue. Labeled genes demonstrate dose-dependent fold changes across increases levels of radiation. B) Heatmap of differentially expressed genes (adj p value < 0.05) for replicate samples of irradiated U87 cells, along with gene ontology terms of upregulated and downregulated gene clusters. C) Heatmap of differentially express lncRNAs (adj p value < 0.05) for replicate samples of irradiated U87 cells. D) Transcript model and locus of *LINC01021* with p53 ChIP-seq peaks.

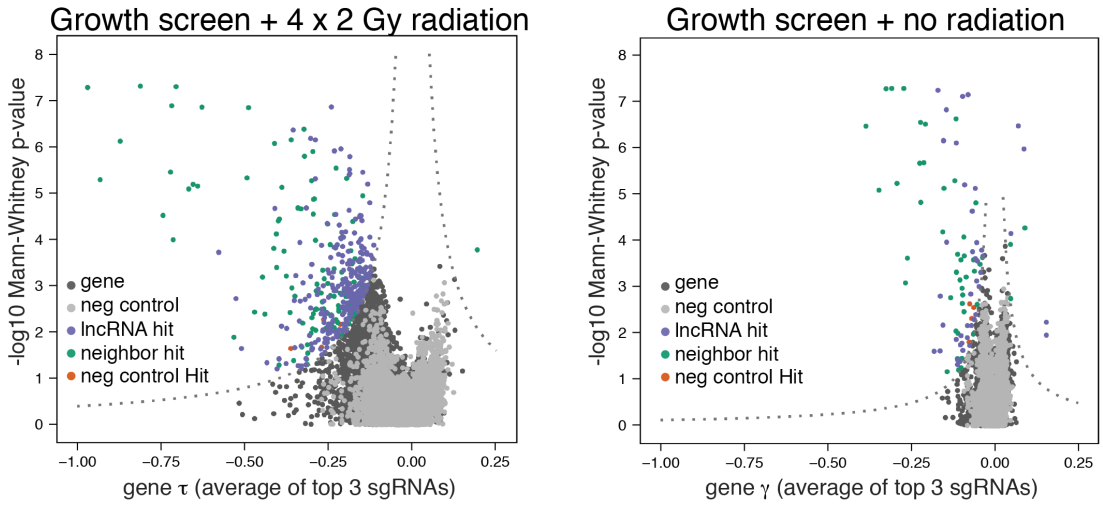
A



B



C



D

	Genes targeted in screen	Total hits	Hits neighboring **ANY** coding genes	**lncRNA hits**
U87 - Radiation	5,689	404	135	**269**
U87 - No Rad	5,689	100	50	**50**

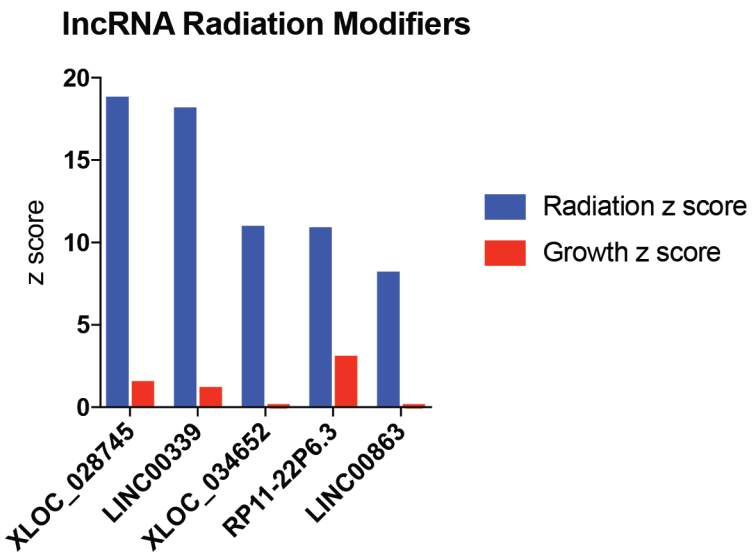
E

Hit	Neighbor	Distance	Rad z score	Growth z score	Neighbor role
RAD21-AS1	RAD21	166	32.7	0.2	Double Strand Break repair
LINC00476	ERCC6L2	0	30.8	1.2	Excision Repair

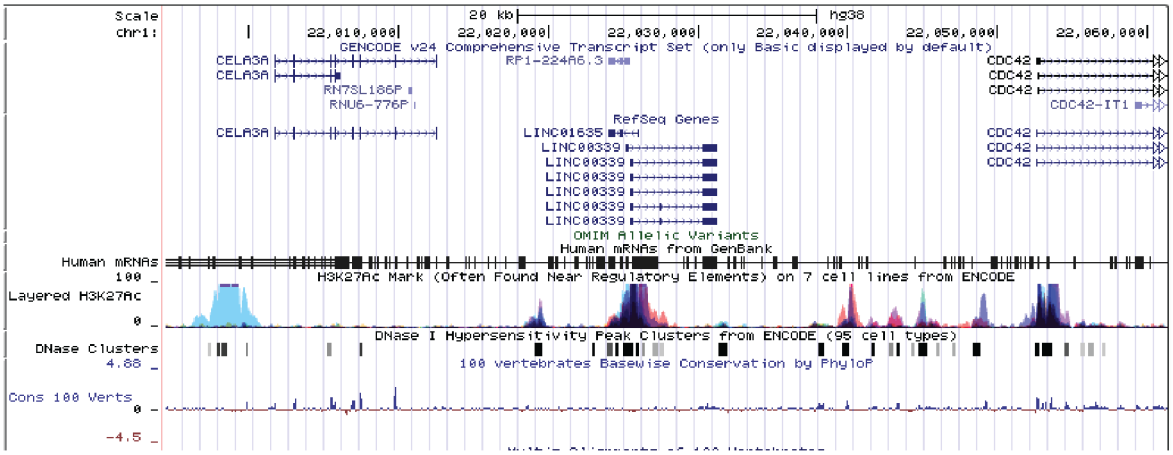
Figure 3. CRISPRi screens identify lncRNAs whose knockdown sensitize U87 cells to radiation induced cell death

A) Schematic of CRISPRi screen targeting 5689 lncRNA TSS's in U87 cells, with addition of 4 doses of 2 Gy ionizing radiation during growth period. B) Percentage of cells during replicate screen experiments showing stable expression of integrated sgRNA vectors over the course of the screen. Radiation time points are indicated with arrows. C) Volcano plots of screen results comparing gene phenotype (τ) of the average of the top 3 sgRNAs targeting each lncRNA TSS, and the significance of each gene. Neighbor hits represent lncRNA hits whose TSS is within 1kb of ANY protein coding gene TSS. Dotted line is threshold calculated from the combined phenotype and significance corresponding to 5% empirical FDR. D) Summary table of genes targeted and hits in the radiation modifier screen and the growth screen. E) Table of TSS-TSS distances (in bp), and radiation and growth screen z scores, for internal positive control neighbor hits.

A


lncRNA Radiation Modifiers
z score
20
15
10
5
0
XLOC_028745
LINC00339
XLOC_034652
RP11-22P6.3
LINC00863
Radiation z score
Growth z score

B


Scale
chr1:
20 kb
hg38
22,010,000
22,020,000
22,030,000
22,040,000
22,050,000
22,060,000
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)
CELA3A
RP1-224A6.3
CDC42
RN7SL186P
RNU6-776P
CDC42-IT1
RefSeq Genes
LINC01635
LINC00339
OMIM Allelic Variants
Human mRNAs from GenBank
Human mRNAs
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types)
DNase Clusters
100 vertebrates Basewise Conservation by PhyloP
Cons 100 Verts

C

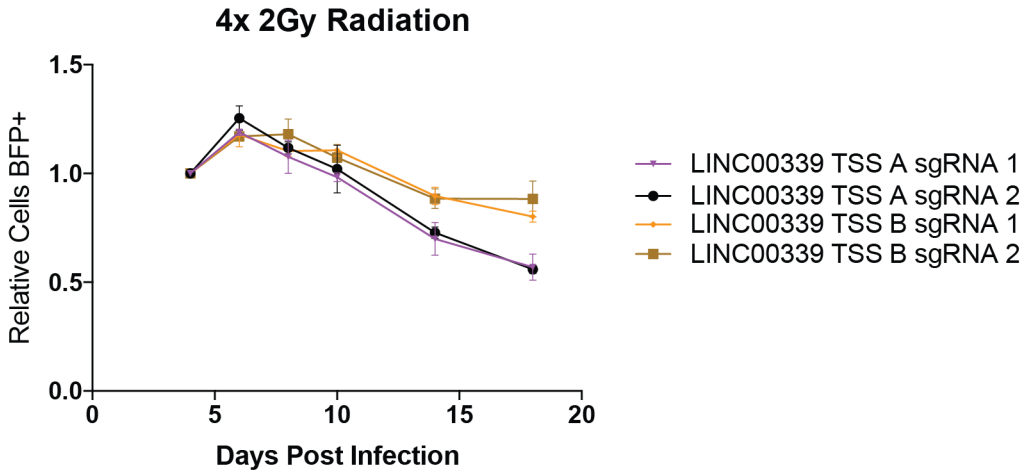

4x 2Gy Radiation
Relative Cells BFP+
1.5
1.0
0.5
0.0
0
5
10
15
20
Days Post Infection
LINC00339 TSS A sgRNA 1
LINC00339 TSS A sgRNA 2
LINC00339 TSS B sgRNA 1
LINC00339 TSS B sgRNA 2

Figure 4. Characterization of radiation sensitizer lncRNA hit *LINC00339*

A) Screen phenotype z scores for representative lncRNAs with strong radiation phenotypes but little to no growth phenotypes. B) Genome locus of *LINC00339.* C) Internally controlled growth assays for individual knockdown of two separate *LINC00339* TSS's with two independent sgRNAs for each TSS, in the presence of 4 doses of 2 Gy ionizing radiation at the beginning of the growth assay.

References

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol *11*, R106.

Badie, B., Goh, C.S., Klaver, J., Herweijer, H., and Boothman, D.A. (1999). Combined radiation and p53 gene therapy of malignant glioma cells. Cancer Gene Ther. *6*, 155–162.

Barani, I.J., and Larson, D.A. (2014). Radiation Therapy of Glioblastoma. In Current Understanding and Treatment of Gliomas, (Cham: Springer International Publishing), pp. 49–73.

Batista, P.J., and Chang, H.Y. (2013). Long noncoding RNAs: cellular address codes in development and disease. *152*, 1298–1307.

Birkenbihl, R.P., and Subramani, S. (1992). Cloning and characterization of rad21 an essential gene of Schizosaccharomyces pombe involved in DNA double-strand-break repair. Nucleic Acids Res.

Brennan, C.W., Verhaak, R.G.W., McKenna, A., Campos, B., Noushmehr, H., Salama, S.R., Zheng, S., Chakravarty, D., Sanborn, J.Z., Berman, S.H., et al. (2013). The somatic genomic landscape of glioblastoma. *155*, 462–477.

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

Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool.

BMC Bioinformatics *14*, 128.

De Bacco, F., D'Ambrosio, A., Casanova, E., Orzan, F., Neggia, R., Albano, R., Verginelli, F., Cominelli, M., Poliani, P.L., Luraghi, P., et al. (2016). MET inhibition overcomes radiation resistance of glioblastoma stem-like cells. EMBO Molecular Medicine *8*, 550–568.

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

Fouse, S.D., Nakamura, J.L., James, C.D., Chang, S., and Costello, J.F. (2014). Response of primary glioblastoma cells to therapy is patient specific and independent of cancer stem cell phenotype. Neuro-Oncology *16*, 361–371.

Gilbert, L.A., Horlbeck, M.A., Adamson, B., Villalta, J.E., Chen, Y., Whitehead, E.H., Guimaraes, C., Panning, B., Ploegh, H.L., Bassik, M.C., et al. (2014). Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. Cell *159*, 647–661.

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. (2011). Long non-coding RNA HOTAIR reprograms chromatinstate to promote cancer metastasis. Nature *464*, 1071–1076.

Hashizume, R., Andor, N., Ihara, Y., Lerner, R., Gan, H., Chen, X., Fang, D., Huang, X., Tom, M.W., Ngo, V., et al. (2014). Pharmacologic inhibition of histone demethylation as a therapy for pediatric brainstem glioma. Nat Med *20*, 1394–1396.

Higgins, G.S., Prevo, R., Lee, Y.-F., Helleday, T., Muschel, R.J., Taylor, S., Yoshimura, M., Hickson, I.D., Bernhard, E.J., and McKenna, W.G. (2010). A small interfering RNA screen of genes involved in DNA repair identifies tumor-specific radiosensitization by POLQ knockdown. Cancer Research *70*, 2984–2993.

Horlbeck, M.A., Gilbert, L.A., Villalta, J.E., Adamson, B., Pak, R.A., Chen, Y., Fields, A.P., Park, C.Y., Corn, J.E., Kampmann, M., et al. (2016). Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. eLife *5*.

Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner, J.R., Evans, J.R., Zhao, S., et al. (2015). The landscape of long noncoding RNAs in the human transcriptome. Nature Genetics.

Kalbasi, A., June, C.H., Haas, N., and Vapiwala, N. (2013). Radiation and immunotherapy: a synergistic combination. J. Clin. Invest. *123*, 2756–2763.

Kampmann, M., Bassik, M.C., and Weissman, J.S. (2013). Integrated platform for genome-wide screening and construction of high-density genetic interaction maps in mammalian cells. Proceedings of the National Academy of Sciences *110*, E2317–E2326.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat Meth *12*, 357–360.

Konermann, S., Brigham, M.D., Trevino, A.E., Joung, J., Abudayyeh, O.O., Barcena, C., Hsu, P.D., Habib, N., Gootenberg, J.S., Nishimasu, H., et al. (2015). Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature *517*, 583–588.

Lee, S., Kopp, F., Chang, T.-C., Sataluri, A., Chen, B., Sivakumar, S., Yu, H., Xie, Y., and Mendell, J.T. (2015). Noncoding RNA NORAD Regulates Genomic Stability by Sequestering PUMILIO Proteins. Cell 1–33.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930.

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., et al. (2017). CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. Science *355*, eaah7111.

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., et al. (2016). Single-cell analysis of long non-coding RNAs in the developing human neocortex. Genome Biol *17*, 67.

Meng, L., Ward, A.J., Chun, S., Bennett, C.F., Beaudet, A.L., and Rigo, F. (2015). Towards a therapy for Angelman syndrome by targeting a long non-coding RNA. Nature *518*, 409–412.

Monteleone, G., Neurath, M.F., Ardizzone, S., Di Sabatino, A., Fantini, M.C., Castiglione, F., Scribano, M.L., Armuzzi, A., Caprioli, F., Sturniolo, G.C., et al. (2015). Mongersen, an oral SMAD7 antisense oligonucleotide, and Crohn's disease. N Engl J Med *372*, 1104–1113.

Müller, S., Liu, S.J., Di Lullo, E., Malatesta, M., Pollen, A.A., Nowakowski, T.J., Kohanbash, G., Aghi, M., Kriegstein, A.R., Lim, D.A., et al. (2016). Single-cell sequencing maps gene expression to mutational phylogenies in PDGF- and EGF-driven gliomas. Molecular Systems Biology *12*, 889.

Naidu, M.D., Mason, J.M., Pica, R.V., Fung, H., and Peña, L.A. (2010). Radiation resistance in glioma cells determined by DNA damage repair activity of Ape1/Ref-1. J. Radiat. Res. *51*, 393–404.

Omuro, A., and DeAngelis, L.M. (2013). Glioblastoma and Other Malignant Gliomas: A Clinical Review. Jama *310*, 1842–1850.

O'Leary, V.B., Ovsepian, S.V., Carrascosa, L.G., Buske, F.A., Radulovic, V., Niyazi, M., Moertl, S., Trau, M., Atkinson, M.J., and Anastasov, N. (2015). PARTICLE, a Triplex-Forming Long ncRNA, Regulates Locus-Specific Methylation in Response to Low-Dose Irradiation. Cell Reports *11*, 474–485.

Raizer, J., and Parsa, A. (2014). Current Understanding and Treatment of Gliomas.

Reichert, Z.R., Wahl, D.R., and Morgan, M.A. (2016). Translation of Targeted Radiation Sensitizers into Clinical Trials. Semin Radiat Oncol *26*, 261–270.

Schmitt, A.M., Garcia, J.T., Hung, T., Flynn, R.A., Shen, Y., Qu, K., Payumo, A.Y., Peres-da-Silva, A., Broz, D.K., Baum, R., et al. (2016). An inducible long noncoding RNA amplifies DNA damage signaling. Nature Genetics *48*, 1370–1376.

Sharma, V., Khurana, S., Kubben, N., Abdelmohsen, K., Oberdoerffer, P., Gorospe, M., and Misteli, T. (2015). A BRCA1-interacting lncRNA regulates homologous recombination. EMBO Reports *16*, e201540437–e201541534.

Thomas, A.A., Brennan, C.W., DeAngelis, L.M., and Omuro, A.M. (2014). Emerging Therapies for Glioblastoma. JAMA Neurol *71*, 1437–1444.

Tummala, H., Kirwan, M., Walne, A.J., Hossain, U., Jackson, N., Pondarre, C., Plagnol, V., Vulliamy, T., and Dokal, I. (2014). ERCC6L2 mutations link a distinct bone-marrow-failure syndrome to DNA repair and mitochondrial function. Am. J. Hum. Genet. *94*, 246–256.

Verhaak, R.G.W., Hoadley, K.A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M.D., Miller, C.R., Ding, L., Golub, T., Mesirov, J.P., et al. (2010). Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell *17*, 98–110.

Zeron-Medina, J., Wang, X., Repapi, E., Campbell, M.R., Su, D., Castro-Giner, F., Davies, B., Peterse, E.F.P., Sacilotto, N., Walker, G.J., et al. (2013). A polymorphic p53 response element in KIT ligand influences cancer risk and has undergone natural selection. Cell *155*, 410–422.

Zheng, M., Morgan-Lappe, S.E., Yang, J., Bockbrader, K.M., Pamarthy, D., Thomas, D., Fesik, S.W., and Sun, Y. (2008). Growth Inhibition and Radiosensitization of Glioblastoma and Lung Cancer Cells by Small Interfering RNA Silencing of Tumor Necrosis Factor Receptor-Associated Factor 2. Cancer Research *68*, 7570–7578.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature

3/29/2017

Date