CONTROL OF POLYCOMB BY CIS-REPRESSIVE LONG NON-CODING RNAS

Megan Danielle Schertzer

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Curriculum in Genetics and Molecular Biology.

Chapel Hill
2019

Approved by:

J. Mauro Calabrese

Folami Ideraabdullah

Jonathan Berg

Fernando Pardo-Manuel de Villena

Brian Strahl

ABSTRACT

Megan Danielle Schertzer: Control of polycomb by cis-repressive long non-coding RNAs
(Under the direction of J. Mauro Calabrese)

Cis-repressive long non-coding RNAs (lncRNAs) spread Polycomb Repressive Complexes (PRCs) within specific genomic regions to achieve chromatin compaction and stable gene silencing. *Xist* is the best characterized lncRNA; it is required to spread PRCs and silence genes across the entire 165 Mb X chromosome. Despite decades of research using the *Xist* lncRNA as a model, the relationship between lncRNAs and PRCs remains unclear. *Airn* and *Kcnq1ot1* lncRNAs function similarly to *Xist*, but in smaller genomic regions. In this dissertation, we gained novel insights into lncRNA and PRC mechanism by comparing and contrasting lncRNA features and the genomic environments of *Xist*, *Airn*, and *Kcnq1ot1*.

First, we found that *Airn* and *Kcnq1ot1* spread PRCs and silence genes across multi-megabase domains in mouse trophoblast stem cells (TSCs). Similar to the X chromosome, *Airn* and *Kcnq1ot1* targeted regions contained non-uniform patterns of PRCs. We showed that PRC density in the 13 Mb *Airn* target region correlated with *Airn* abundance and was dependent on multiple aspects of genome architecture: linear distance to the *Airn* locus, pre-existing structure, TAD boundaries, and high-affinity chromatin sites of *Airn*. In *Airn* overexpression TSCs, eight PRC-bound CpG islands (CGIs) appeared to nucleate the spread of Polycomb. Deletion of one 2kb CGI caused loss of Polycomb across 4.5 Mb. *Xist* and *Kcnq1ot1* targeted regions showed similar patterns of Polycomb at PRC-bound CGIs. This suggests a common mechanism where lncRNAs depend on pre-bound CGIs to specifically target and spread Polycomb *in cis*.

ACKNOWLEDGEMENTS

During these six years of graduate school, there are many people that have contributed to my success. My advisor, Mauro Calabrese, has had the greatest scientific impact. I am so grateful to have been trained by him directly at the bench. He always took the time to explain the why behind steps in an experiment. He let me take time to teach myself to code, so that I could analyze my own sequencing data. Finally, he was always optimistic and encouraging through every experiment. If I got down about a result, he maintained a positive attitude, and encouraged me to continue moving forward.

I would also like to thank the members of my thesis committee: Folami Ideraabdullah, Jonathan Berg, Fernando Pardo-Manuel de Villena, and Brian Strahl for their enthusiasm, support, and suggestions throughout the years.

Day to day, David Lee and Jessime Kirk had the greatest impact on my development as a scientist. We were a small lab for the first four years, and we talked almost everyday about science and life. They gave me honest feedback on my work and challenged my ideas to make me better. On days, weeks, and sometimes months where I was down about research, they always listened and encouraged. They helped make my graduate school experience the best that it could be. Over the past year, we have gained new lab members: Keean Braceros, Rachel Cherney, Jackson Trotman, and Dan Sprague, all who have provided further advice and encouragement. I am especially grateful to Keean who will be taking over my project. We have had the chance to work side by side for the past year, and he has contributed to both of my first author papers. I look forward to seeing where the project goes!

In addition to people in my lab, many others at UNC contributed to my main first author paper in Chapter 2. I am grateful to Josh Starmer for help with the statistics throughout the manuscript. He has a natural gift for explaining statistics to biologists. I would like to thank Pablo Ariel for his enthusiasm about my project and for going above and beyond his role as the microscopy core director. He worked closely with me to help develop the methods necessary to answer our biological questions, and he taught me a lot about microscopy in the process. I thank Megan Justice and Jill Dowen for help with the 3D structure part of my project. Finally, I thank Gabby Salazar and Mark Zylka for their collaboration to supply us with F1 hybrid mouse neurons. For my other first other paper in Chapter 3, I would like to thank the rotation students that worked with me on this project: Eliza Thulson, Ryan Murphy, Emma Hinkle, and Eva Vitucci.

I would also like to acknowledge undergraduate mentors that made a significant contribution to my early science education. Lori West and Sherry Kasper encouraged me to apply for summer research during my second year of college. If it wasn't for them, I would probably be graduating from medical school instead, which was definitely not the right career path for me.

I am also grateful to all of the friends that I made over the past six years. Sam Miller, Lyndsay Wylie, and Kate Lansu were fellow graduate students that were also some of my closest friends. Outside or work, I am grateful to Mimi Huang and Amanda Hunsucker for being my best roommates and my adventure buddies. Thanks to all of the friends I made through Grace Church, especially Allison Everson and Kelley Roundtree. In the evenings, I always looked forward to playing sports. I am so thankful to all the members of the Bread and Butter softball team, that I have been apart of for five years. Thanks to all of my volleyball friends, especially

PK and Marie Vogel. All of the fun spent outside of lab gave me the happiness and motivation to keep working.

I would like to thank my family, especially my Mom and Dad, who have always supported me. They never put pressure on me, and always told me they were proud of me and loved me no matter what. They taught me to be responsible, independent, and hard working from the very beginning.

Finally, I would like to thank Daniel, who has been beside me for the best and worst moments of graduate school. Thank you for listening to the same things over and over again, and thank you for your honest advice and support that you also gave over and over again. And thanks for celebrating the big moments with me!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

3D	Three-dimensional
Bp	Basepair
B/C	Black 6 mother crossed with Castaneous father
C57BL/6J	Black 6 (mouse strain)
CAGE	Cap-analysis gene expression
Cas9	CRISPR-associate protein 9
CAST/EiJ	Castaneous (mouse strain)
C/B	Castaneous mother crossed with Black 6 father
CGI	CpG island
ChIA-PET	Chromatin interaction analysis by paired end tag sequencing
ChIP	Chromatin immunoprecipitation
CPM	Counts per million
CRISPR	Clustered regularly interspaced short palindromic repeat
CT	Chromosome territory
ESC	Embryonic stem cell
FISH	Fluorescence in situ hybridization
H2AK119ub	Mono-ubiquitylation of lysine 119 on histone H2A
H3K27me3	Tri-methylation of lysine 27 on histone H3
IF	Immunofluorescence
irMEFs	Irradiated mouse embryonic fibroblasts
Kb	Kilobase
KD	Knockdown

KO	Knockout
lncRNA	Long non-coding RNA
Mb	Megabase
mRNA	Messenger RNA
OE	Overexpression
PRC1	Polycomb Repressive Complex 1
PRC2	Polycomb Repressive Complex 2
PRCs	Polycomb Repressive Complexes
qPCR	Quantitative polymerase chain reaction
RBP	RNA binding protein
RE	Regulatory element
RNA-IP	RNA immunoprecipitation
RPKM	Reads per kilobase per million
RPM	Reads per million
rtTA	Reverse-tetracycline transactivator
sgRNA	Single guide RNA
TAD	Topologically associated domain
TSC	Trophoblast stem cell
WT	Wildtype
Xa	Active X
XCI	X chromosome inactivation
Xi	Inactive X

CHAPTER 1: General Introduction

1.1 Long non-coding RNAs

RNA is classically viewed as an intermediate that transfers information from DNA to make protein. However, the first examples of non-coding RNAs were identified as early as the 1950s with the discovery of ribosomal RNAs and transfer RNAs (Cech and Steitz, 2014). By the early 1990's, diverse types of non-coding RNAs were found, including large regulatory RNAs, *Xist* and *H19,* and smaller RNAs termed microRNAs (Rinn and Chang, 2012). The emergence of functional non-coding RNAs challenged the central dogma that RNA serves as a template for protein synthesis.

With the advent of high throughput sequencing technologies, it was discovered that 60-70% of genomic DNA is transcribed into RNA, but less than 2% is translated into protein (Fatica and Bozzoni, 2014; Geisler and Coller, 2013; Yan et al., 2015). Long non-coding RNAs, or lncRNAs, quickly emerged as a major portion of these non-coding RNAs. They were classified based on a length greater than 200bp with little to no protein coding potential (open reading frames that translate polypeptides of <50 amino acids; Fatica and Bozzoni, 2014; Rinn and Chang, 2012). Identification and annotation of lncRNAs has continued to evolve, primarily using RNA-sequencing methods and mapping transcription start sites via cap-analysis gene expression sequencing (CAGE-seq) and chromatin-immunoprecipiation sequencing (ChIP-seq) methods in numerous tissue types (Rinn and Chang, 2012). Estimates of human lncRNAs range from 15,000 to 60,000 transcripts (Iyer et al., 2015; Kornienko et al., 2016; Pertea et al., 2018).

Similar to messenger RNAs (mRNAs), lncRNAs are transcribed by RNA Polymerase II, capped, and frequently both polyadenylated and spliced (Ulitsky and Bartel, 2013). However, lncRNAs harbor many features that distinguish them from mRNAs. Generally, lncRNAs have lower sequence conservation, fewer exons, poor splicing efficiency, lower expression levels, and tissue-specific expression patterns (Fatica and Bozzoni, 2014; Kornienko et al., 2016; Ulitsky and Bartel, 2013). While an mRNA's job is to act as a template for protein synthesis, lncRNAs have more diverse functions that center on their ability to bind specific proteins. LncRNAs can act as (1) decoys, to pull proteins from other cellular processes, (2) scaffolds, to bring proteins together that would otherwise not interact, and (3) guides, to recruit proteins to DNA with their ability to specifically interact with both macro-molecules (Geisler and Coller, 2013; Rinn and Chang, 2012; Yan et al., 2015). While lncRNAs as a class seem to be well-characterized, only a few individual lncRNAs have been studied beyond their annotation.

1.2 Cis-repressive lncRNAs in X-inactivation and genomic imprinting

Cis-repressive lncRNAs represent a major class of lncRNAs that can act as both scaffolds and guides, as described above. They often recruit chromatin modifying complexes *in cis* to the same chromosome from which the lncRNA is transcribed for repression of nearby genes. *Xist*, *Airn*, and *Kcnq1ot1* were some of the earliest identified cis-repressive lncRNAs due to their requirement for X-chromosome inactivation and genomic imprinting, two important processes during development. Misregulation of these lncRNAs and their gene targets results in developmental disorders such as Rett Syndrome, Beckwith-Wiedemann Syndrome, and Silver Russels Syndrome (Lee and Bartolomei, 2013).

X-chromosome inactivation (XCI) is the mammalian mechanism to equalize gene dosage between males (XY) and females (XX). It results in whole chromosome gene silencing of one X chromosome in mammalian females (Lyon, 1961). In mice, there are two types of XCI: imprinted X-inactivation and random X-inactivation. Imprinted XCI is initiated at the 4- cell stage of development, where *Xist* is expressed solely from the paternal allele and therefore silences the paternal X allele (Kalantry et al., 2009; Tada et al., 2000; Takagi and Sasaki, 1975). This expression pattern is maintained in the extraembryonic tissues but is reversed in the inner cell mass. After reversal of imprinted XCI, random XCI is initiated in the developing embryo and maintained in all adult somatic tissues (Calabrese and Magnuson). In relation to these types of XCI, two main models are used to study *Xist* and X-inactivation: mouse trophoblast stem cells (TSCs; imprinted XCI) and mouse embryonic stem cells (ESCs; random XCI). Importantly, humans only undergo random XCI.

In the early 1990's, it was discovered that XCI in placental mammals was orchestrated by the *Xist* lncRNA (Brockdorff et al., 1992; Brown et al., 1992; Penny et al., 1996). *Xist* is an 18kb, spliced and polyadenylated lncRNA that remains exclusively in the nucleus. It is required for the repression of ~1000 genes across the entire 165 Mb X chromosome. To date, *Xist* is the most well characterized lncRNA and is a stand-alone example of a lncRNA that can silence an entire chromosome.

Other examples of multi-gene silencing by cis-acting lncRNAs are in genomic imprinted regions, where clusters of genes are expressed exclusively from either the maternal or paternal allele. Due to this process, both the paternal and maternal genomes are required for survival of the offspring (McGrath and Solter, 1984; Surani and Barton, 1983). In 1991, *Igf2r*, *Igf2*, and *H19* were the first imprinted genes to be discovered (Barlow et al., 1991; Bartolomei et al., 1991;

DeChiara et al., 1991; Ferguson-Smith et al., 1991). To date, up to 150 imprinted genes have been identified, with the highest number in the extra-embryonic tissue, where XCI is also imprinted (Andergassen et al., 2017; Barlow and Bartolomei, 2014; Calabrese et al., 2015). Defects in genomic imprinting can occur when a region harboring imprinted genes have deletions or epigenetic errors on one of the two alleles, affecting embryonic growth, placental development, and behavior (Lee and Bartolomei, 2013).

Two clusters of imprinted genes on mouse chromosome 17 and 7 are controlled by the lncRNAs *Airn* and *Kcnq1ot1*, respectively (Mancini-Dinardo et al., 2006; Sleutels et al., 2002). *Airn* and *Kcnq1ot1* are imprinted themselves through methylation of their promoters during oocyte development, which results in paternal-specific expression of the lncRNAs upon fertilization (Lewis et al., 2004; Stöger et al., 1993). Both lncRNAs function in the nucleus as predominantly unspliced, polyadenylated transcripts that are greater than 80kb in length (Lyle et al., 2000; Pandey et al., 2008; Redrup et al., 2009; Seidl et al., 2006). Similar to *Xist*, they repress genes *in cis*, exclusively on the paternal allele, but in smaller genomic regions. *Airn* and *Kcnq1ot1* target the largest genomic regions in the mouse placenta, where *Airn* represses 10 genes across 10 Mb and *Kcnq1ot1* represses seven genes across 800 kb (Andergassen et al., 2017; Lee and Bartolomei, 2013).

Interestingly, in all three regions targeted by *Xist*, *Airn*, and *Kcnq1ot1*, gene repression is non-uniform, in which repressed genes are interspersed with actively transcribed genes. On the inactive X, between 3% and 15% of genes escape silencing by *Xist* (Calabrese et al., 2012; Carrel and Willard, 2005). On chromosome 17 and 7, only 10 out of more than 100 genes are repressed by the *Airn* lncRNA and only 7 are repressed by the *Kcnq1ot1* lncRNA, respectively. This suggests that cis-repressive lncRNAs specifically target genes and chromatin for silencing

versus imposing a random, but uniform targeting with well-defined boundaries. It remains unclear how this specificity is achieved.

1.3 LncRNAs and Polycomb Repressive Complexes

Xist, *Airn*, and *Kcnq1ot1* lncRNAs are not catalytic and, therefore, function through the proteins that they bind. All three are known to interact with the two Polycomb Repressive Complexes (PRCs) in mammals: PRC2 and PRC1 (Pandey et al., 2008; Zhao et al., 2008, 2010). Much of the initial research focused on PRC2, which deposits the mono-, di-, and tri-methylation marks on lysine 27 of histone H3 (H3K27me3; (Chittock et al., 2017)). The core complex consists of EED, SUZ12, RbAp48, and catalytic component, EZH1/EZH2, although RbAp48 is not required for methyltransferase activity of the complex (Cao and Zhang, 2004; Simon and Kingston, 2013). PRC1 deposits a ubiquitin mark on lysine 119 of histone H2A (H2AK119ub). The core complex consists of a PCGF protein and the catalytic component, RING1B/RING1A, where RING1B is most common in mammals (Simon and Kingston, 2013). In addition to the core components, there are numerous accessory proteins that affect PRC1 recruitment. The canonical complex contains a CBX protein that can bind H3K27me3, while variant complexes have RYBP instead and can be recruited independently of H3K27me3 (Gao et al., 2012; Tavares et al., 2012). Similarly, EED within the PRC2 complex can bind H3K27me3 and can interact with PRC1, leading to significant overlap of the two complexes and their deposited histone marks throughout the genome (Blackledge et al., 2014; Cooper et al., 2014; Margueron et al., 2009).

Early experiments to decipher the relationship between PRCs and lncRNAs, mainly *Xist*, involved immunofluorescence (IF). These experiments showed that PRC2 and PRC1

components and H3K27me3 localized to the inactive X similar to the *Xist* cloud seen via RNA florescence in situ hybridization (FISH) (Mak et al., 2002; Plath et al., 2003, 2004; Silva et al., 2003). To investigate a specific role for the *Xist* lncRNA in this colocalization, different groups used RNA immunoprecipitation (RNA-IP) and in vitro binding assays. EZH2 RNA-IP methods concluded that PRC2 interacts with a specific region of the *Xist* lncRNA (Zhao et al., 2008, 2010), whereas more quantitative in vitro binding assays concluded that EZH2 has a high affinity for all RNA, rather than for a subset that included *Xist* (Davidovich et al., 2013, 2015). In support of the latter conclusion, *Xist* pull-down followed by mass spectrometry did not show interaction between *Xist* and PRC2 components, although RING1B, RYBP, and PCGF5 of PRC1 were detected (Chu et al., 2015). Additionally, high resolution microscopy showed spatial separation of *Xist* and PRC2 that supported an indirect interaction (Cerase et al., 2014).

Similar, but less extensive, experiments have investigated the relationship between the *Kcnq1ot1* and *Airn* lncRNAs and PRCs. One group performed immuno/RNA FISH and measured significant overlap between both lncRNAs and EZH2, RING1B, H3K27me3, and H2AK119ub (Mi Terranova et al., 2008). Additional EZH2 and SUZ12 RNA-IP experiments in mouse placenta pulled down the *Kcnq1ot1* lncRNA (Pandey et al., 2008). Further characterization of *Kcnq1ot1* and *Airn* has been complicated by their size (>80 kb) and their low expression levels. Collectively, these data for the *Xist*, *Airn*, and *Kcnq1ot1* lncRNAs support indirect recruitment of PRCs to the inactive X and the imprinted domains on chromosome 17 and 7, respectively.

Given this indirect model for lncRNA binding, recent work has investigated what PRCs can bind directly and how specific recruitment to chromatin can be achieved. From an RNA centric view, the RNA-binding protein (RBP), HNRNPK, mediates the interaction between the

Xist lncRNA and the PCGF3/5 core components of PRC1 (Pintacuda et al., 2017). Interestingly, HNRNPK was one of the most abundant proteins, amongst many other RBPs, pulled down by the *Xist* lncRNA (Chu et al., 2015). This suggests that specificity of lncRNA-PRC interactions may have more to do with RBPs. From a chromatin centric view, CpG islands (CGIs) can recruit both PRC1 and PRC2, although not all of them do (Ku et al., 2008; Mendenhall et al., 2010). Accessory proteins such as KDM2B for PRC1 and JARID2, PHF1, MTF2, and PHF19 for PRC2, have been shown to bind CGI DNA and mediate PRC recruitment (Farcas et al., 2012; Hunkapiller et al., 2012; Li et al., 2010; Oksuz et al., 2018; Wu et al., 2013).

The importance of PRC bound CGIs in lncRNA-targeted regions is unclear. Additionally, as *Xist* is primarily used to study lncRNA biology, it is not known which aspects of *Xist* function differ from *Airn, Kcnq1ot1*, and other less studied cis-repressive lncRNAs.

1.4 LncRNAs, PRCs, and three-dimensional genome structure

LncRNAs and PRCs have both been shown to contribute to the three-dimensional (3D) structure of DNA in the nucleus. 3D genome structure consists of multiple layers: (1) DNA loops, (2) topologically associated domains (TADs), and (3) chromosome territories (CTs; (Dixon et al., 2016)).

At the most basic layer, individual DNA loops are formed when two distant loci are brought into close spatial proximity. Loops commonly serve to bring together enhancers and gene promoters to regulate gene expression. At the anchors of these loops are architectural proteins, most commonly CTCF and cohesion (Rao et al., 2014). The next layer of 3D genome structure involves TADs, which are large compartments that harbor high intra-chromosomal interactions with distinct chromatin and gene expression profiles. TADs are on average 1 Mb in

size, allowing them to encompass many loops within their boundaries. Importantly, TADs are conserved between cell types, while individual loops vary (Dixon et al., 2012; Dowen et al., 2014; Nora et al., 2012; Rao et al., 2014). At the highest level of structure are CTs, where DNA of each chromosome occupies a defined space within the nucleus (Cremer and Cremer, 2010).

Previous work has shown that 3D structure and the *Xist* lncRNA work together to achieve XCI. At the initiation of XCI, *Xist* first transfers to distant genomic sites based on pre-existing 3D contacts and specific chromatin features before spreading across the remainder of the X chromosome (Engreitz et al., 2013; Simon et al., 2013). Not surprisingly, the inactive X (Xi) and the active X (Xa) differ in structure and position within the nucleus. The Xa forms many TADs similar to autosomal chromosomes, but the Xi has less defined TADs and a high frequency of long range contacts, or super-loops (Deng et al., 2015; Nora et al., 2012; Rao et al., 2014; Splinter et al., 2011). The unique structure of Xi is dependent on the *Xist* lncRNA, as deletion of *Xist* partially restores the structure on Xa (Splinter et al., 2011). While 3D structures also differ between the repressed and active alleles in imprinted regions, it is not known if these differences are dependent on the corresponding lncRNA (Deng et al., 2015; Rao et al., 2014).

There is also evidence that PRCs can dictate 3D structure. First, canonical PRC1 has been shown to compact chromatin (Francis et al., 2004; Grau et al., 2011). The catalytic component of PRC1, RING1B, is required for compaction, but this function is independent of its ubiquitinating activity (Eskeland et al., 2010; Kundu et al., 2017; Schoenfelder et al., 2015). Second, PRC1-bound promoters in mouse ESCs form three dimensional contacts that require RING1B, and a subset of these loops occur across >10Mb of chromatin (Schoenfelder et al., 2015). Finally, PRC1 forms visible foci within the nucleus in both *Drosophila* and mammalian cells (Buchenau et al., 1998; Isono et al., 2013). These aggregates require the polymerization activity of the sterile

alpha motifs (SAM) in Polyhomeotic proteins, Ph in *Drosophila* and Phc2 in mice (Isono et al., 2013; Wani et al., 2016). Given these data, it seems that PRCs can control genomic structure independent of CTCF and can function beyond defined TAD boundaries. Taken together, the relatedness between lncRNAs and PRCs, PRCs and structure, and lncRNAs and structure suggests that all three likely work together to regulate chromatin and gene expression in lncRNA-targeted regions of the genome.

1.5 Model to study *Xist, Airn,* and *Kcnq1ot1* lncRNA function

Xist, *Airn*, and *Kcnq1ot1* are great models to study epigenetic regulation by cis-repressive lncRNAs, for multiple reasons. First, each is expressed from a single parental allele, meaning the lncRNA-targeted allele can be compared to the non-targeted allele within the same cell. Second, data suggests that these three lncRNAs function through a similar mechanism. However, *Xist*, *Airn*, and *Kcnq1ot1* harbor different repressive capacity, targeting genes across 165, 10, and 0.8 Mb, respectively. Additionally, when inserted onto an autosome, *Xist* can silence genes across the entire chromosome, although not as effectively (Lee and Jaenisch, 1997; Tang et al., 2010; Wutz and Jaenisch, 2000). These differences confirm that features of the lncRNAs themselves dictate repressive potency, but highlight that genomic environment plays an additional role. Therefore, study of *Xist*, *Airn*, and *Kcnq1ot1* features and function within the same cell type would provide a unique opportunity to define lncRNA features and environmental factors important for cis-repressive lncRNA function.

To date, most studies of *Xist* have used ESC models, in which *Xist* is not normally expressed. Approaches have included differentiating ESCs to naturally induce *Xist* expression, inducing expression from an *Xist* transgene on the single male X chromosome, or using

transgenes inserted onto various autosomes (Engreitz et al., 2013; Lee and Jaenisch, 1997; Loda et al., 2017; Simon et al., 2013). Additionally, in ESCs, *Airn* is lowly expressed and *Kcnq1ot1* represses only 4 genes across 0.4 Mb, less than in trophoblast stem cells (TSC; Kanduri, 2011; Latos et al., 2009). Instead of ESCs, our lab previously derived reciprocal F1-hybrid mouse TSC lines (Calabrese et al., 2015). We reasoned that TSCs would be the ideal system to study *Xist*, *Airn*, and *Kcnq1ot1* simultaneously, as they are all naturally expressed and deposit PRCs across megabases of DNA.

1.6 Creating tools for studies in TSCs

There are several challenges in working with TSCs. They require feeder cells and additional growth factors, grow slowly, and have low transfection efficiency (Quinn et al., 2006). To study lncRNA, PRCs, and 3D structure, we needed a system to reliably and efficiently knockdown proteins, make genomic deletions, and activate and repress lncRNA transcription in TSCs. To achieve these goals, we created a system that combines CRISPR-Cas9 genome editing with the piggyBac transposon system (Cadiñanos and Bradley, 2007; Cong et al., 2013; Ding et al., 2005a; Wang et al., 2008; Wilson et al., 2007). Using this system, we can reproducibly create stable TSC lines in 12-14 days.

CHAPTER 2: LncRNA-induced spread of Polycomb controlled by genome architecture, RNA abundance, and CpG island DNA[1]

2.1 Introduction

Long noncoding RNAs (lncRNAs) play essential roles in development by directing Polycomb Repressive Complexes (PRCs) to broad genomic regions. In the most extreme example, expression of the lncRNA *Xist* causes PRCs to modify chromatin over the entire inactive X chromosome. Several other lncRNAs also cause PRCs to engage with smaller genomic regions. In the mouse placenta, the lncRNA *Airn* silences 10 genes in 10 megabases (Mbs) on chr17 and the lncRNA *Kcnq1ot1* silences seven genes in ~800 kilobases (kb) on chr7. Like *Xist*, these lncRNAs act *in cis*, meaning that they only target regions located on the same chromosome from which they were transcribed (Andergassen et al., 2017; Lee and Bartolomei, 2013).

The two major PRCs, PRC1 and PRC2, catalyze the mono-ubiquitylation of lysine 119 on histone H2A (H2AK119ub) and the trimethylation of lysine 27 on histone H3 (H3K27me3), respectively. These modifications repress gene expression through parallel mechanisms that compact chromatin and antagonize transcriptional activators (Schwartz and Pirrotta, 2013; Simon and Kingston, 2013). The two PRCs are also interdependent throughout the genome. Regions of chromatin modified by one PRC are usually modified by the other, PRC1 can be recruited by PRC2 and vice versa, and loss of either PRC destabilizes most, or all, PRC1- and

[1] This chapter previously appeared as an article in the journal, *Molecular Cell*. The original citation is as follows: Schertzer MD, Braceros KCA, Starmer J, Cherney RE, Lee DM, Salazar G, Justice M, Bischoff SR, Cowley DO, Ariel P, Zylka MJ, Dowen JM, Magnuson T, Calabrese JM (2019). lncRNA-induced spread of Polycomb controlled by genome architecture, RNA abundance, and CpG island DNA. *Molecular Cell*.

PRC2-silenced regions (Blackledge et al., 2014; Kalb et al., 2014; Schwartz and Pirrotta, 2013; Simon and Kingston, 2013). Indeed, *Xist, Airn,* and *Kcnq1ot1* require both PRC1 and 2 for full repressive activity (Almeida et al., 2017; Kalantry et al., 2006; Mi Terranova et al., 2008).

Nevertheless, the mechanisms through which *Xist* and related lncRNAs induce the spread of PRCs over chromatin remain unclear. While local features of the genome correlate with levels of *Xist*-induced, PRC-dependent modification, the mechanisms that underlie the correlations are unclear (Calabrese et al., 2012; Cotton et al., 2014; Kelsey et al., 2015; Loda et al., 2017; Pinter et al., 2012). The RNA-binding protein HNRNPK bridges PRC1 with *Xist*, and this interaction is required to spread both PRC1 and 2 over the inactive X (Almeida et al., 2017; Pintacuda et al., 2017). However, whether *Xist* directly travels with PRCs, or if *Xist* causes PRCs to spread via secondary interactions, remains debated (Cerase et al., 2014; Smeets et al., 2014; Sunwoo et al., 2015). Moreover, while *Airn* and *Kcnq1ot1* also direct PRCs to chromatin (Pandey et al., 2008; Regha et al., 2007), it is unclear if they do so through mechanisms shared with *Xist*. Indeed, transcription over the *Airn* and *Kcnq1ot1* genes, and not necessarily the lncRNAs themselves, plays a role in local silencing by both lncRNAs; whether the lncRNA products are required for distal silencing or spread of PRCs remains unclear (Andergassen et al., 2017; Korostowski et al., 2012; Latos et al., 2012). Lastly, the mechanisms that give rise to specific patterns of PRC-dependent modifications within the *Airn* and *Kcnq1ot1* target domains remain unclear (Andergassen et al., 2017; Pandey et al., 2008; Regha et al., 2007).

Considering these unknowns, we set out to study molecular phenotypes associated with *Airn* and *Kcnq1ot1* in contexts that allowed direct comparison to *Xist*, using female, F1-hybrid, mouse trophoblast stem cells (TSCs) that naturally express all three lncRNAs (Calabrese et al.,

2015). We found that genome architecture, lncRNA abundance, and CpG island (CGI) DNA all play roles in coordinating the spread of PRCs induced by *Xist*, *Airn*, and *Kcnq1ot1*.

2.2 Results

2.2.1 Megabase-sized domains of H3K27me3 require continued expression of *Airn* and *Kcnq1ot1*

In TSCs, *Xist*, *Airn*, and *Kcnq1ot1* are monoallelically expressed from paternally inherited chromosomes (Lee and Bartolomei, 2013). Their monoallelism, coupled with their cis-acting nature, necessitates the use of F1-hybrid cells to study their effects on chromatin (Figure S1A). To this end, we previously derived F1-hybrid TSCs from reciprocal crosses between C57BL/6J and CAST/EiJ mice, and demonstrated that the TSCs can be used to study all three lncRNAs (Calabrese et al., 2012, 2015).

While *Xist* is known to recruit PRCs to gene-dense regions of the X (Calabrese et al., 2012; Chadwick, 2007; Marks et al., 2009), PRC targeting had not been examined around *Airn* and *Kcnq1ot1* using sequencing-based approaches. Thus, we used ChIP-Seq to measure the density of H2AK119ub and H3K27me3, covalent modifications catalyzed by PRC1 and PRC2, respectively, in reciprocal F1-hybrid TSC lines. As expected, the two modifications were highly correlated throughout the genome ($r = 0.746$; Figure S1B). We used H3K27me3 as a surrogate for H2AK119ub in most of our work below owing to its higher signal in ChIP-Seq.

We observed that H3K27me3 density was high around known *Airn* and *Kcnq1ot1* target genes, and dropped sharply near non-targets, supporting previous views that chromatin-associated factors work in concert with lncRNAs to control the spread of PRCs in lncRNA domains (Figure 1A, B, upper panels; (Calabrese et al., 2012; Cotton et al., 2014; Kelsey et al.,

2015; Loda et al., 2017; Pinter et al., 2012)). Unexpectedly, both *Airn* and *Kcnq1ot1* were centered in H3K27me3-enriched regions that extended for megabases beyond their originally defined target genes (Figure 1A, B, lower panels area between dotted lines). We used hiddenDomains to call peaks of H3K27me3 in TSCs (Starmer and Magnuson, 2016), and then used allelic data within those peaks to identify sites of significant parent-of-origin bias (Table S1). Strikingly, 83% of all paternally-biased autosomal peaks of H3K27me3 in TSCs were found in the regions surrounding *Airn* and *Kcnq1ot1* (Figure S1C; Table S1), where patterns of H2AK119ub mirrored those of H3K27me3 (Figure S1D, E; Table S1). Of the 118 and 91 paternally-biased H3K27me3 peaks surrounding *Airn* and *Kcnq1ot1*, 43 and 32 overlapped genes (starting 2kb upstream of transcription starts and extending to transcription ends) and 76 and 3 peaks were intergenic, respectively. H3K27me3 signal in these peaks rivaled or surpassed signal in H3K27me3 peaks on the X (Figure S1F). These data suggest that in TSCs, *Airn* and *Kcnq1ot1* each direct PRCs to regions that span megabases, potentially in a manner analogous to *Xist*.

To determine whether PRC-induced chromatin modifications around *Airn* and *Kcnq1ot1* were lncRNA-dependent, we truncated each lncRNA in a way that phenocopies lncRNA knockout in embryos (Figure S2A; (Mancini-Dinardo et al., 2006; Sleutels et al., 2002)). We derived clonal truncation lines for each lncRNA, and in two of them we profiled H3K27me3 via ChIP-Seq (Figure S2B-D). We then compared paternal H3K27me3 between lncRNA-truncated and wild-type TSCs. Upon truncation, we observed loss of H3K27me3 in the regions surrounding both lncRNAs (Figure 1C, D; Table S1). Consistent with activity *in cis*, *Airn*-truncated TSCs had wild-type H3K27me3 levels around *Kcnq1ot1*, and vice versa (Figure S2D). These data show that in TSCs, *Airn* and *Kcnq1ot1* direct PRCs to 13 and 2.3 Mb regions, respectively.

We performed RNA-Seq in three truncation clones to determine whether lncRNA loss coincided with gene derepression. Of 61 genes in the *Airn* domain whose allelic expression met our threshold for consideration (avg. of ≥10 allelic reads per dataset), 14 were derepressed upon *Airn* truncation, 6 of which are known *Airn* targets in the placenta (Figure 1E; Table S2; (Andergassen et al., 2017)). In the *Kcnq1ot1* domain, 27 genes met our threshold, and 6 were derepressed upon *Kcnq1ot1* truncation; 4 were known *Kcnq1ot1* targets (Figure 1F; Table S2). Non-impacted genes also trended towards higher expression upon lncRNA truncation (Figure 1E, F). Thus, lncRNA truncation leads to derepression of genes *in cis*. Locations of derepressed and non-impacted genes relative to H3K27me3 levels are shown in Figure S2E. 62 and 16 genes in the *Airn* and *Kcnq1ot1* domains, respectively, did not meet our threshold for allelic analysis.

2.2.2 Spread of H3K27me3 in the *Airn* domain is influenced by pre-existing genomic architecture and additional features on the paternal allele

On the X, H3K27me3 levels correlate with 3D contacts in place prior to induction of *Xist*, and with specific regulatory elements (Calabrese et al., 2012; Cotton et al., 2014; Engreitz et al., 2013; Kelsey et al., 2015; Loda et al., 2017; Pinter et al., 2012). Whether similar trends are true in regions silenced by *Airn* and *Kcnq1ot1* is unknown.

Owing to its large size, we first focused on the region targeted by *Airn*. Based on the studies above, and data demonstrating DNA loops restrict signals that control gene expression (Dowen et al., 2014; Rao et al., 2014), we hypothesized that variation in H3K27me3 around *Airn* was due to three factors: first, a pre-existing genomic architecture that might make certain regions more susceptible to targeting due to their ability to contact *Airn*; second, within

susceptible regions, a greater affinity of *Airn* for specific sites over others; and third, DNA loops that restrict PRC spread around sites of *Airn* contact.

We tested the first two hypotheses using fluorescence in situ hybridization (FISH). We designed FISH probes to 9 regions surrounding *Airn*, each harboring different extents of *Airn*-dependent H3K27me3, including a probe in a region whose H3K27me3 was unaffected by *Airn* ("Neg control", purple bar; Figure 2A). We used RNA/DNA FISH to measure spatial distance between each region of interest and the *Airn* locus, distinguishing paternal from maternal alleles by co-localization of *Airn* RNA and DNA FISH probes (Figure 2B).

For 8 of 9 loci, average spatial distance to *Airn* was less on the paternal allele (Figure 2C). We also examined two loci in *Airn*-truncation TSCs, *Park2* and *Arid1b*, and found that differences between maternal and paternal distributions were reduced (Figure 2D). The average difference in distance between the paternal and maternal alleles, i.e. the extent of genomic compaction at a locus, was a strong predictor of H3K27me3 (Figure 2E, R2=0.738). These data indicate that compaction in the *Airn* region depends on continued expression of *Airn* and correlates with underlying levels of H3K27me3. Distance in base pairs to the *Airn* locus was also a strong predictor of paternal H3K27me3 (Figure 2F panel (i); R2 = 0.399).

Next, we tested the hypothesis that, in addition to distance in base pairs from *Airn*, H3K27me3 levels were influenced by chromosomal conformations in place prior to the onset of *Airn* expression, which rendered certain regions more likely than others to come into proximity to *Airn*. In our experiment, the maternal allele served as a surrogate to approximate the conformation that the paternal allele would be in if *Airn* were not expressed. If extent of *Airn*-induced H3K27me3 was influenced by pre-existing chromosomal conformations in place prior to the onset of *Airn* expression, we would expect distance to *Airn* on the maternal allele, which

provides a readout for those conformations, to be a better predictor of paternal H3K27me3 than the distance to the *Airn* locus in base pairs. Consistent with this notion, our expectation held true (Figure 2F panel (i) vs (ii); p= 0.003; empirically derived by bootstrapping).

Intriguingly, distance to *Airn* on the paternal allele was a better predictor of paternal H3K27me3 than distance on the maternal allele (Figure 2F panel (ii) vs (iii); p=0.037; empirically derived by bootstrapping). This increase in predictive power supports the view that, within broader domains capable of contacting *Airn*, additional factors on the paternal allele cause the lncRNA to associate with certain sites more than others. Thus, in addition to distance in base pairs from the lncRNA expressing locus (Figure 2Fi) and pre-existing genomic architecture (Figure 2Fii), local features of chromatin (Figure 2Fiii) likely contribute to the control of PRCs by *Airn*.

2.2.3 Intensity of H3K27me3 in lncRNA target domains correlates with TADs, DNA loops, and SMC1 and CTCF binding

DNA loops and topologically associated domains (TADs) divide the genome into compartments with distinct chromatin and gene expression patterns that may influence targeting by *Xist* (Darrow et al., 2016; Dixon et al., 2012; Dowen et al., 2014; Engreitz et al., 2013; Giorgetti et al., 2016; Rao et al., 2014). To determine whether DNA loops and TADs might also influence *Airn* and *Kcnq1ot1*, we examined Hi-C, ChIA-PET, and ChIP-Seq data in mouse embryonic stem cells (ESCs). TAD boundaries are often conserved between cell types (Dixon et al., 2012), and we reasoned that, as a first pass, inferring the location of DNA loops and TADs in TSCs using ESC data was a viable approach. Independently, we profiled, via ChIP-Seq, Cohesin (SMC1) and CTCF binding in F1-hybrid TSCs. Genome-wide and in the *Airn* and *Kcnq1ot1*

domains, SMC1 and CTCF peak locations were concordant between ESCs and TSCs (Figure S3A). Moreover, SMC1 and CTCF binding in TSCs was detected at DNA loops anchored by SMC1 and CTCF in ESCs (Figure S3B, C), consistent with the notion that the two cell types harbor many of the same DNA loops.

Throughout the *Airn* and *Kcnq1ot1* target domains, inflections in TSC H3K27me3 density coincided with ESC TAD boundaries and SMC1-bound DNA loops (Figure 3A, B), supporting the notions that DNA loops influence spread of H3K27me3 in lncRNA target domains, and that *Airn* and *Kcnq1ot1* direct PRCs over multiple TADs. Moreover, SMC1 and CTCF showed reduced binding in lncRNA-silenced domains on paternal relative to maternal alleles. This reduction was stronger for SMC1 than CTCF and correlated with the range over which *Xist*, *Airn*, and *Kcnq1ot1* direct PRCs to chromatin (Figure 3C, D). Thus, the more potent the lncRNA, the more likely that its targeted regions lack DNA loops anchored by SMC1 and CTCF.

2.2.4 LncRNA repressive potency correlates with abundance, stability, and underlying features of the genome

Our data show that in TSCs, *Airn* and *Kcnq1ot1* direct PRCs to megabase-sized regions in which H3K27me3 levels are influenced by genome architecture and underlying features of chromatin. We also found that *Airn* and *Kcnq1ot1* each control PRCs to different extents, and both to lesser extents than *Xist*.

We examined if differences in lncRNA abundance could account for differences in repressive potency, which we define here as the ability of a lncRNA to induce PRC-dependent chromatin modifications. We estimated copy number of *Xist*, *Airn*, and *Kcnq1ot1* in TSCs using

RNA-Seq and found that the lncRNAs are expressed at an average of 232, 8.7, and 7.6 copies per TSC, respectively (Figure 4A, S4A). Using Actinomycin D, we found the half-life of *Xist* was ~6.2 hours, while the half-lives of *Airn* and *Kcnq1ot1* were each ~1.7 hours (Figure 4B). Thus, in TSCs, abundance and stability correlate with lncRNA potency, but they do not account for differences in potency between *Airn* and *Kcnq1ot1*.

Based on these data, we hypothesized that changes in *Airn* abundance would affect its potency. We created TSCs in which we could recruit a transcriptional activator (dCas9-VP160) or a repressor (dCas9-KRAB) to the endogenous *Airn* promoter in a doxycycline-inducible manner (Schertzer et al., 2018). RNA-Seq showed *Airn* increased to ~27.2 copies per cell by recruiting the activator and decreased to ~0.6 copies per cell by recruiting the repressor. Almost all of the boost in *Airn* expression occurred on the paternal allele, presumably because DNA methylation prevented the activator from accessing the maternal allele (Figure S4B, C). Overexpression increased the size of *Airn* foci detected by FISH but did not change *Airn* subcellular distribution (Figure S4D-F).

We observed a striking correlation between RNA abundance and repressive potency in the *Airn* domain (Figure 4C-E). Expression-induced changes in H3K27me3 were variable throughout the domain and were inversely proportional to changes in gene expression (Figure 4C-E; Figure S5A). The largest changes in H3K27me3 occurred on the centromeric side of *Airn*, centered around three regions that appeared to be sites from which H3K27me3 spread outwards, owing to their high levels of H3K27me3 in *Airn*-overexpression cells that dropped rapidly with increasing distance on either or both sides (Fig 4E, arrows). We draw two major conclusions: (1) that RNA abundance can affect lncRNA control over PRCs, but it is not the only factor to do so, and (2) that genomic features – likely a mix of 3D architecture, chromatin-bound factors, and the

sequence of DNA itself – play important roles in the lncRNA-induced spread of PRCs on chromatin.

2.2.5 CGIs bind PRCs autonomously and can nucleate spread of H3K27me3 by *Airn*

We examined regions in the *Airn* domain whose patterns of H3K27me3 suggested the presence of H3K27me3 nucleation sites (Figure 4E, arrows). Strikingly, all 6 regions coincided with CGIs (Figure 4E, blue ticks; S5B), which are known to recruit PRCs in mammals (Farcas et al., 2012; Li et al., 2017a; Lynch et al., 2012; Mendenhall et al., 2010; Oksuz et al., 2018; Riising et al., 2014; Woo et al., 2010). The 8 CGIs were all found at lowly to moderately expressed genes and had the highest levels of H3K27me3 in *Airn*-overexpressing TSCs relative to all other CGIs in the *Airn* domain (Table S3; Figure S5B). Using non-allelic data as a search feature, 7 of the 8 CGIs co-localized with MACS-defined peaks of RING1B (catalytic subunit of PRC1), and 6 also co-localized with peaks of EZH2 (catalytic subunit of PRC2). In contrast, only 27 of the 83 remaining CGIs in the *Airn* domain co-localized with RING1B peaks, and none with EZH2 peaks (Table S3; Figure S5B). Moreover, consistent with trends elsewhere (Figure S3D, E), 5 of the 8 CGIs in question co-localized with SMC1 peaks and none with CTCF (Table S3; Figure S5B). Assuming that *Airn* targets PRCs de novo to chromatin, we hypothesized that the 8 CGIs should harbor higher RING1B and EZH2 signal relative to surrounding regions, and that CGIs on the *Airn*-targeted paternal allele would harbor more signal than CGIs on the untargeted maternal allele.

To test these hypotheses, we profiled RING1B in wild-type, *Airn*-overexpression, *Airn*-truncation, and *Kcnq1ot1*-truncation TSCs, and EZH2 in wild-type TSCs. We observed enrichment of RING1B and EZH2 at CGIs relative to surrounding DNA; however, near-equal

levels of RING1B and EZH2 were found at CGIs on maternal and paternal alleles (Figure 5A, red lines overlap with blue lines). Outside of CGIs, we observed broad enrichment of RING1B and EZH2 on the paternal allele. This enrichment was responsive to *Airn* expression and mirrored enrichment of H3K27me3 and H2AK119-ub (Figures 5B, C vs. Figures 1A, S1D). Analogous patterns of RING1B and EZH2 surrounded *Kcnq1ot1* (Figure S6A-C).

To test our hypothesis that specific CGIs nucleate the spread of PRCs in the *Airn* target domain, we used CRISPR to delete the CGI at *Slc22a3*, which is located ~234 kb upstream of *Airn* yet harbors some of the highest density of H3K27me3 in the *Airn* domain and shows evidence of RING1B binding on both alleles (Figure S6D; Table S3). As a control, we deleted a size-matched region ~1,383 kb upstream of *Airn* that occurs within an H3K27me3 peak in the *Park2* intron but does not overlap a CGI (Figure S6E). We profiled H3K27me3 in two independent clones of each deletion. Strikingly, deletion of the *Slc22a3* CGI, but not the size-matched control, caused a ~4.6 Mb reduction in H3K27me3 in the *Airn* domain (Figures 5D, 5E, S6F). This loss could not be ascribed to reduced *Airn* RNA abundance upon *Slc22a3* CGI deletion (Figure S6G). Thus, specific CGIs can play outsized roles in nucleating the spread of H3K27me3 in the *Airn* target domain.

2.2.6 *Xist*-induced H3K27me3 density is highest around CGIs that bind PRCs autonomously

The inactive X displayed patterns of H3K27me3 similar to those seen upon over-expression of *Airn*, where H3K27me3 levels culminated at single points, then decreased in intensity until inflecting or crossing into a H3K27me3-depleted region (Figure S7A vs. 4E; (Calabrese et al., 2012)). In light of these similarities, we examined levels of H3K27me3,

RING1B, and EZH2 at X-linked CGIs in TSCs. Analogous to the *Airn* region, the highest levels of H3K27me3 on the inactive X were found at CGIs over which we could detect peaks of RING1B and EZH2; the more PRC binding that could be detected, the greater the levels of surrounding H3K27me3, and the greater the difference in H3K27me3 between the inactive and active X (Figure 6A; Table S4). However, at CGIs, the binding of both RING1B and EZH2 was substantially higher on the active X relative to the inactive X, despite H3K27me3 levels showing the opposite enrichment in the majority of cases (Figure 6A). Most CGIs co-bound by RING1B and EZH2 on the active X coincided with lowly expressed genes, consistent with their PRC-mediated repression (Table S4). Outside of CGIs, RING1B and EZH2 were broadly enriched on the inactive X and their enrichment in H3K27me3 peaks was strongly co-correlated (Figure 6B; r=0.98). Thus, in TSCs, regions on the inactive X that harbor the most H3K27me3 coincide with CGIs that bind the highest levels of RING1B and EZH2. Unexpectedly, at these CGIs, more RING1B and EZH2 signal is found on the active X than on the inactive X.

We next examined *Xist*-induced spread of PRCs in ESCs, a commonly used cell-based model in *Xist* research. As part of a separate study, we inserted a doxycycline-inducible *Xist* gene into the Rosa26 locus on chr6 in mouse ESCs (D.M.L. and J.M.C., in press). We profiled RING1B and H3K27me3 in these ESCs before, 12 hours after, and 72 hours after induction of *Xist* (Figure S7B, C). Similar to what we observed in TSCs, the highest levels of *Xist*-induced H3K27me3 were found around CGIs on chr6 that bound PRCs prior to induction of *Xist* (Figure 6C). However, consistent with recent studies performed in ESC models (Fursova et al., 2019; Żylicz et al., 2019), there was little relative difference in the change in *Xist*-induced H3K27me3 levels at PRC-bound versus unbound CGIs (Δ1.08 vs Δ0.87, respectively; upper right corner of lower two panels in Figure 6C), whereas in TSCs, the difference was dramatic (Δ1.29 vs Δ0.26,

respectively; upper right corner of first and last panels in lowest row in Figure 6A). Moreover, unlike what was observed on the TSC inactive X, we observed no *Xist*-dependent depletion of PRCs at CGIs on ESC chr6 (Figure 6A vs 6C), and overall levels of *Xist*-induced H3K27me3 were lower on ESC chr6 than they were on the TSC inactive X, or even in the TSC *Airn* domain (Figure 6D vs S7A, 4E). Similarly, H3K27me3 levels around *Kcnq1ot1* were lower in ESCs than in TSCs and were lower in a third cell type, cortical neurons (Figure S7D, E). Lastly, we note that relative to the TSC X, there were ~10-fold more RING1B/EZH2-bound CGIs on ESC chr6 (Figure 6A vs 6C). These data highlight potential differences in interactions between PRCs and CGIs in ESCs versus TSCs, and suggest, along with data from (Andergassen et al., 2017; Lewis et al., 2006; Umlauf et al., 2004), that relative to other mouse cell types, TSCs are primed to respond to PRC-controlling lncRNAs.

2.2.7 *Xist, Airn*, and *Kcnq1ot1* require HNRNPK to spread H3K27me3

Xist requires the RNA-binding protein HNRNPK to induce PRC spread (Pintacuda et al., 2017). Considering the similarities between *Xist*, *Airn*, and *Kcnq1ot1*, we examined whether *Airn* and *Kcnq1ot1* also required HNRNPK to induce PRC spread.

We first examined whether *Airn* and *Kcnq1ot1* showed evidence of HNRNPK association. We used a formaldehyde-based RNA immunoprecipitation (RNA IP) protocol followed by RNA-Seq in *Xist*-expressing SM33 ESCs (which also express *Kcnq1ot1* and low levels of *Airn*) and in TSCs (Raab et al., 2019). IP revealed strong enrichment of HNRNPK over all three lncRNAs relative to IgG in both cell types, and peaks of HNRNPK enrichment were identified by MACS (Figure 7A, Table S5). In contrast, IP of CTCF, a protein that binds RNA with high affinity in a sequence non-specific manner (Kung et al., 2015), yielded little

enrichment over the lncRNAs (Figure 7A, Table S5). Moreover, in TSCs, relative to the set of 23366 UCSC Known Genes, *Xist*, *Airn*, and *Kcnq1ot1* harbored the 4th, 6th, and 7th most HNRNPK signal, respectively (Table S5, 'ts.hk.norm' column). In contrast, the lncRNAs harbored the 296th, 3048th, and 4244th most CTCF signal, and in terms of length-normalized expression, they were the 171st, 6550th, and 6798th most highly expressed transcripts, respectively (Table S5, 'ts.ctcf.norm' and 'TSC input' columns). HNRNPK enrichment was also observed over Repeat B and C in *Xist*, which are known HNRNPK-interacting regions (Pintacuda et al., 2017). Taken together, these data show that *Xist*, *Airn*, and *Kcnq1ot1* associate with HNRNPK, likely at levels above most other genes.

Next, we used CRISPR to knock-down HNRNPK in TSCs and profiled H3K27me3 and RNA expression after four days of Cas9 induction (Schertzer et al., 2018). By two days of Cas9 induction, HNRNPK knockdown caused TSC death (not shown). Nevertheless, four days after Cas9 induction, relative to non-targeting sgRNA controls, HNRNPK levels were substantially reduced in surviving TSCs (Figure S7F, G). Loss of HNRNPK coincided with a significant loss of H3K27me3 in *Xist*, *Airn*, and *Kcnq1ot1* target domains (Figure 7B), but did not coincide with changes in gene silencing, presumably because TSCs that lost the most HNRNPK died (Figure S7H). Moreover, upon HNRNPK knockdown, H3K27me3 was reduced around CGIs that bound PRCs even in the absence of lncRNA expression (Figure 7C). Thus, in lncRNA domains, H3K27me3 levels at PRC-bound CGIs are more dependent on HNRNPK than H3K27me3 levels at PRC-unbound CGIs. Also, similar to *Xist*, *Airn* and *Kcnq1ot1* rely on HNRNPK to spread PRCs.

2.3 Discussion

Via orthogonal assays, we compared the genomic properties in the domains targeted by *Airn* and *Kcnq1ot1* to those on the *Xist*-targeted X chromosome. We gained several insights into mechanism which we enumerate below. While *Xist*, *Airn*, and *Kcnq1ot1* are all monoallelically expressed due to X-inactivation and imprinting (Lee and Bartolomei, 2013), we posit that these transcriptional regulatory mechanisms are unlikely to impact function of the lncRNA after its production. Thus, principles defined by our study are likely to be relevant to other lncRNAs, as well.

We found that variation in intensity of H3K27me3 in *Airn* and *Kcnq1ot1* domains mirrored variation on the inactive X, where H3K27me3-enriched regions are separated by regions that partially or fully escape *Xist*-induced silencing (Calabrese et al., 2012; Chadwick, 2007; Marks et al., 2009; Pinter et al., 2012). Within the *Airn* domain, variation in H3K27me3 could be partly explained by large-scale, pre-existing conformations of chromatin that rendered strongly silenced regions more likely to come in proximity to the *Airn* locus than weakly-silenced ones. Our results support the view that long-distance contacts in place prior to the onset of lncRNA expression play roles in dictating the intensity of PRC-induced modification in lncRNA target domains (Engreitz et al., 2013; Kelsey et al., 2015; Marks et al., 2015).

However, more than *Xist*, *Airn* and *Kcnq1ot1* appeared to be influenced by genome architecture. Inflections in H3K27me3 density in the *Airn* and *Kcnq1ot1* domains tended to co-localize with DNA loops and TADs (Dixon et al., 2012; Dowen et al., 2014), whereas such structures are largely absent on the inactive X (Rao et al., 2014). Accordingly, we found that the more potent the lncRNA, the more likely it was to disrupt binding of SMC1 and, to a lesser

extent, CTCF. Thus, while DNA loops may influence the initial spread of H3K27me3 on the inactive X, they are more likely to be overridden, ultimately, by the repressive effect of *Xist*.

We observed a strong correlation between expression, stability, and potency of *Xist*, *Airn*, and *Kcnq1ot1*. In TSCs, *Xist*, the most potent of the three, was expressed most highly, at ~232 molecules per cell, and had the longest half-life, at ~6.2 hours. *Airn* and *Kcnq1ot1* were expressed at ~8 molecules per cell and had ~1.7-hour half-lives. Increasing or decreasing expression of *Airn* changed its potency over a 13 Mb domain. Thus, factors that control the balance between expression and stability likely play major roles in controlling the potency of *Airn* and other lncRNAs as well.

Within the *Airn* domain, specific CGIs appeared to nucleate the spread of H3K27me3 upon lncRNA exposure, owing to their high levels of H3K27me3 and the nearby decrease in H3K27me3 as distance from the CGIs increased. These CGIs bound RING1B and EZH2 at near equal levels on paternal and maternal alleles, but were centered in broad regions of H3K27me3 enrichment only on the lncRNA-expressing (paternal) allele. Similar relationships between CGIs, PRC binding, and H3K27me3 density were found surrounding *Kcnq1ot1* and on the TSC inactive X, although at X-linked CGIs, RING1B and EZH2 binding were higher on the active X than on the inactive one. Deletion of a lynchpin CGI at the *Slc22a3* promoter caused a multi-megabase loss of H3K27me3 in the *Airn* domain, whereas a control deletion did not. *Xist*, *Airn*, and *Kcnq1ot1* all required HNRNPK to induce the spread of PRCs, and HNRNPK loss reduced H3K27me3 at CGIs pre-loaded with PRCs in all three lncRNA target domains.

Our data suggest that lncRNAs preferentially induce the spread of PRCs from CGIs that autonomously bind PRCs. In our model (Figure 7D), individual lncRNA foci associate with high levels of PRCs owing to RNA-binding proteins such as HNRNPK, which bind both lncRNAs

and PRCs and may aggregate with themselves and other proteins (Hentze et al., 2018; Pintacuda et al., 2017). Relative to sites on chromatin that lack bound PRCs, these same RNA-binding proteins may stabilize lncRNA foci at PRC-bound CGIs. The stabilization of a lncRNA carrying a payload of PRCs at a CGI would initiate PRC spread in a domain of contact, beyond the spread that was nucleated by the CGI prior to lncRNA exposure. Network-like interactions between PRC-bound CGIs, which may e*Xist* even in the absence of lncRNA expression (Isono et al., 2013), could explain how lowly expressed lncRNAs like *Airn* induce multi-megabase effects (one lncRNA focus could contact multiple CGIs simultaneously), and why deletion of the *Slc22a3* CGI caused such a strong loss of H3K27me3 (disrupting a key CGI in a network might disrupt lncRNA access to other CGIs, as well). The PRCs and HNRNPK are now known to tether *Xist* to the X both during and after X-inactivation is complete (Colognori et al., 2019), providing precedent for the notion that interactions between lncRNAs, HNRNPK, and possibly other proteins tether lncRNAs to PRC-bound CGIs in domains including the X.

Importantly, our proposed CGI-mediated nucleation model appears to be more relevant for *Airn* and *Kcnq1ot1* than it is for *Xist*, owing to the greater non-uniformity of H3K27me3 centered around CGIs in the target domains of the former two lncRNAs, and the fact that deletion of a single PRC-bound CGI in the *Airn* domain caused a multi-megabase loss of H3K27me3. Relative to *Xist*, which is stable and can diffuse away from its site of transcription to form hundreds of nuclear foci (Cerase et al., 2014; Smeets et al., 2014; Sunwoo et al., 2015), lncRNAs such as *Airn* or *Kcnq1ot1* are unstable, may be less diffusible, and likely access far fewer regions on chromatin before being degraded or otherwise turned over. These differences rationalize how *Airn* or *Kcnq1ot1* might exhibit a greater reliance on PRC-bound CGIs to spread PRCs within their target domains. The stability of *Xist*, its affinity for actively transcribed

regions of chromatin, and the large number of PRC-bound CGIs that *Xist* has access to on the X almost certainly lessen its dependence on any one CGI, particularly during the early stages of X-inactivation in the embryo or in ESCs, where many regions on the X are still transcribed and many CGIs (likely many more than in TSCs) are PRC-bound (Engreitz et al., 2013; Fursova et al., 2019; Loda et al., 2017; Pinter et al., 2012; Zylicz et al., 2019). However, the sum of PRCs bound to chromatin at the onset of X-inactivation may still play important roles in tethering *Xist* to chromatin; indeed, recent work by Colognori, Sunwoo, and colleagues suggest this to be true (Colognori et al., 2019).

Moreover, in TSCs, a subset of CGIs appear to seed high levels of H3K27me3 in broad windows on the inactive X well after their initial exposure to *Xist* (see Figure 6A). In *Drosophila*, deposition of H3K27me3 on newly incorporated nucleosomes requires the presence of DNA elements that recruit PRCs (Laprell et al., 2017). The TSC lines used in our study have maintained their H3K27me3 levels for months in culture after their initial exposure to *Xist* in the blastocyst. Therefore, rather than being related to an event occurring at the onset of X-inactivation, it seems likely that on the TSC inactive X, the increased levels of H3K27me3 surrounding PRC-bound CGIs are due to ongoing synergy between CGI- and *Xist*-dependent PRC recruitment. Thus, in certain cell types, subsets of CGIs on the X may control the intensity of PRC-induced chromatin modifications locally, long after initial exposure to *Xist*.

CGIs are known to nucleate the spread of PRCs throughout the genome (Farcas et al., 2012; Li et al., 2017; Lynch et al., 2012; Mendenhall et al., 2010; Oksuz et al., 2018; Riising et al., 2014; Woo et al., 2010), and, in prior studies of *Xist*, it has been noted that the presence of PRCs bound to chromatin correlates with the intensity of PRC-induced modifications precipitated by expression of *Xist* (Cotton et al., 2014; Kelsey et al., 2015; Loda et al., 2017;

Pinter et al., 2012). However, to our knowledge, our work is the first to directly demonstrate that a single CGI is required to maintain wild-type levels of H3K27me3 in a lncRNA target domain. In our view, this result and others we describe above imply that the proteins that cause PRCs to engage with CGIs elsewhere in the genome likely nucleate the spread of PRCs even within lncRNA target domains.

A key question that remains is, how do lncRNAs induce the spread of PRCs from CGIs? It is possible that the act of transcription, and not the lncRNA per se, plays a role (Kornienko et al., 2013). Indeed, transcription of *Airn* over the *Igf2r* promoter silences the latter gene, and transcription of *Kcnq1ot1* blocks access to enhancers in its gene body (Korostowski et al., 2012; Latos et al., 2012). Still, the hundreds of *Xist* RNA foci that surround the inactive X must harbor function after being transcribed (Cerase et al., 2014; Smeets et al., 2014; Sunwoo et al., 2015). Considering this in relation to data we describe above, we posit that like *Xist*, the *Airn* and *Kcnq1ot1* lncRNAs encode function by recruiting RNA-binding proteins such as HNRNPK, which may nucleate super-stoichiometric interactions with themselves, other RNA-binding proteins, and the PRCs around a given lncRNA focus (Hentze et al., 2018; Pintacuda et al., 2017).

In concert with lncRNA-induced effects, transcription may alter nuclear architecture in a way that facilitates PRC spread over short and long genomic spans (Engreitz et al., 2013; Mele and Rinn, 2016; Nozawa et al., 2017). Within TADs, transcription may promote PRC spread in a process related to DNA loop extrusion (Fudenberg et al., 2016). Between TADs, affinity between transcribed regions may help PRC-bound CGIs co-localize with lncRNA foci in 3D space.

Within lncRNA target domains and between cell types, altered activity of factors that lncRNAs require to interface with PRCs and CGIs may cause lncRNAs to vary in potency. Similar alterations, induced pharmacologically, may offer new avenues to exogenously control lncRNA silencing function.

2.4 Methods

2.4.1 TSC derivation and culture

The C/B and B/C TSC lines used in this work correspond to the C/B and B/C TSCs used in (Calabrese et al., 2012), and are referred to as CB.1 and BC.1 TSCs in (Calabrese et al., 2015). TSCs were cultured at 37°C on pre-plated irradiated MEF feeder cells (irMEFs) in TSC media [RPMI (Invitrogen), 20% Qualified FBS (Invitrogen), 1mM sodium pyruvate (Invitrogen), 100μM β-mercaptoethanol (Sigma), and 2mM L-glutamine] supplemented with Fgf4 (25ng/ml; Invitrogen) and Heparin (1μg/ml; Sigma) just before use. At passage, TSCs were trypsinized with 0.125% Trypsin (Invitrogen) for 3 minutes at room temperature and gently dislodged from their plate with a sterile, cotton-plugged Pasteur pipette (Thermofisher). To deplete MEF feeder cells from TSCs prior to RNA isolation or crosslinking, TSCs were pre-plated for 40 minutes and cultured for three days in 70% MEF-conditioned TSC media supplemented with Fgf4 (25ng/ml; Invitrogen) and Heparin (1μg/ml; Sigma). This was done once for harvesting chromatin and twice before RNA isolation.

2.4.2 Cortical neuron derivation and culture

Reciprocal F1-hybrid cortical neurons were derived from crosses between C57BL/6J and CAST/EiJ mice, cultured, and fixed with 1% formaldehyde. Embryonic day 13.5 (E13.5) to

E16.5 mouse cortices were dissected and trypsinized with TrypLE express at 37 °C for 10 min. Dissociated neurons were cultured with Neurobasal medium with 5% fetal bovine serum, GlutaMAX (Invitrogen), B27 (Invitrogen) and Antibiotic-Antimycotic (Invitrogen) and changed into Neurobasal medium supplemented with 4.84 μg ml−1 uridine 5′-triphosphate (Sigma), 2.46 μg ml−1 5-fluoro-2′-deoxyuridine (Sigma,), GlutaMAX (Invitrogen), B27 (Invitrogen), and Antibiotic-Antimycotic (Invitrogen) at DIV1 and DIV3. Neurons were fixed in 10cm plates at DIV5.

2.4.3 ESC culture

ESCs (both the Rosa26 RMCE and SM33 lines) were maintained at 37°C in a humidified incubator at 5% CO2. Media was changed daily and consisted of DMEM high glucose plus sodium pyruvate, 0.1 mM non-essential AA, 100μM β-mercaptoethanol, 2 mM L-glutamine, 1000 U/ml LIF (ESG1107, Millipore Sigma, St. Louis, MO) and 15% ES-qualified FBS. RMCE cells were maintained on approximately 1.5 million irMEFs per 10-cm plate. At the passage prior to harvesting cells for ChIP, ESCs were pre-plated for 40 minutes and cultured for two days in 70% MEF-conditioned media supplemented as above.

2.4.4 Generation of stable cell lines

Generation of Airn and Kcnq1ot1 truncation TSCs

To create targeting vectors to truncate the *Airn* and *Kcnq1ot1* lncRNAs, a triple-polyA sequence (from (Meng et al., 2013); kind gift of L. Meng and A. Beaudet) was cloned into the NotI and XhoI restriction sites of the PGK-neo vector (Addgene #51422; (Luo et al., 2014)). 5’ prime and 3’ homology arms of about 800bp each were amplified from the RP23-81B3 and

RP23-101N20 BACs (BACPAC Resources), to target the *Airn* and *Kcnq1ot1* loci, respectively, and subsequently cloned upstream and downstream of the triple-polyA-neo cassette. These arms flanked sgRNA recognition sites that were designed to cut at approximately the same genomic coordinates of previous triple-polyA-mediated truncations of *Airn* and *Kcnq1ot1* in the mouse (Mancini-Dinardo et al., 2006; Sleutels et al., 2002). sgRNAs were cloned into pX330 (Addgene #42230; (Cong et al., 2013)). Oligonucleotides used to amplify homology arms and to clone sgRNAs are listed in Table S7.

Targeting vectors were linearized with HindIII and co-electroporated into C/B TSCs with pX330 at a 1:1 ratio using the Neon® Instrument (electroporation program: 950V, 30 ms, 2 pulses; Invitrogen). G418 selection (200μg/ml; Gibco) was started two days after electroporation. Individual colonies were picked on day 8 of G418 selection, and selection was continued for 7 additional days before cells were harvested for RNA expression analysis.

Generation of Airn OE and Airn KD TSCs

(Schertzer et al., 2018) describes our rationale and construction of the piggyBac-based vectors. To create doxycycline-inducible Cas9, dCas9-VP160, and dCas9-KRAB vectors, a parent vector was created in which a bGH-polyA signal and an EF1α promoter driving expression of a hygromycin resistance gene was ligated into the cumate-inducible piggyBac transposon vector from System Biosciences after its digestion with HpaI and SpeI, which cut just downstream of each chicken β-globin insulator sequence and removed all other internal components of the original vector. The TRE promoter from pTRE-Tight (Clontech) was then cloned upstream of the bGH-polyA site, and Cas9, dCas9-VP160, and dCas9-KRAB were then each cloned behind the TRE promoter by digestion with AgeI and SalI (NEB) followed by

Gibson Assembly (NEB), to generate piggyBac cargo vectors capable of inducibly expressing Cas9, dCas9-VP160, and dCas9-KRAB, respectively, upon addition of doxycycline (Addgene #126029, #126031, #126030; (Schertzer et al., 2018)).

In parallel, an sgRNA targeting the *Airn* promoter region was cloned into pX330 (Table S7). Subsequently, the entire U6-sgRNA expression cassette, as well as a U6-sgRNA expression cassette that lacked an sgRNA targeting sequence, was cloned into the PacI site upstream of the rtTA3-IRES-Neo cassette in the rtTA-piggyBac-Cargo vector described in ((Kirk et al., 2018); Addgene #126028; (Schertzer et al., 2018)). The *Airn*-targeting-sgRNA-rtTA-Cargo vector and the non-targeting-sgRNA-rtTA-Cargo vector were each co-electroporated into wild-type C/B TSCs along with the dCas9-VP160-Cargo and dCas9-KRAB-Cargo vectors, respectively, and with the pUC19-piggyBac transposase from (Kirk et al., 2018), using the Neon® Instrument (electroporation program: 1300V, 40 ms, 1 pulse; Invitrogen). C/B TSCs were then selected on G418 (200µg/ml; Gibco) and Hygromycin B (150µg/ml; Gibco) for 9 days. Electroporation of all four vector combinations was performed a second time, and piggyBac-expressing TSCs from both series of electroporation were expanded and treated with 1µg/ml of doxycycline (Sigma) for four days prior to crosslinking for H3K27me3 ChIP and RNA preparation for RNA-seq, as described above.

Generation of CGI and non-CGI deletion TSCs

To delete the *Slc22a3* CGI and the non-CGI control region, 4 unique sgRNAs were designed that flanked each region to be excised (Table S7; Figure S6D, E). Each sgRNA was cloned into the rtTA-BsmbI piggyBac vector from (Schertzer et al., 2018), and starter cultures for each sgRNA were pooled together in equal amounts prior to liquid culture expansion and

plasmid preparation using the Invitrogen HiPure Midiprep kit. The pooled vectors were co-electroporated into 500,000 TSCs with Cas9-Cargo and piggyBac transposase vectors at an 8:2:1 ratio using the Neon® Instrument (electroporation program: 1300V, 40 ms, 1 pulse; Invitrogen). Two days after electroporation, cells were selected with G418 (200ug/ml; Gibco) and Hygromycin B (150ug/ml; Gibco) for 13 days, followed by 4 days of dox treatment (1ug/ml). 2,000 dox-induced cells were then plated on a 10cm plate with pre-plated irMEFs. After 7 days, individual colonies were picked and plated on irMEFs for expansion. Clonal lines were passaged once off of irMEFs prior to harvests for genotyping and *Airn* RNA expression analysis.

Genotyping PCR reactions were performed with genomic DNA using Apex Taq DNA Polymerase (Genesee Scientific) and custom primers. The first set of primers flanked the deletion and identified clonal lines with at least one allele deleted. The second set only amplified a wildtype allele, with one primer sitting outside the deletion and the other inside. Primers used are listed in Table S7 and their locations relative to the sgRNAs are shown in Figure S6D, E.

Generation of ESCs with Xist inserted into the Rosa26 locus

A recombinase-mediated cassette exchange (RMCE) approach was used to insert a doxycycline inducible *Xist* gene into the Rosa26 locus in an ESC line. Briefly, a male F1-hybrid mouse ESC line derived from a cross between C57BL/6J (B6) and CAST/EiJ (Cast) mice was made competent for RMCE by insertion of a custom homing cassette into the Rosa26 locus via homologous recombination. *Xist* transgenes were cloned via PCR or recombineering into a custom RMCE-cargo vector and then electroporated along with a plasmid expressing Cre-recombinase into RMCE-competent cells using a Neon® Transfection System (Invitrogen). After selection on hygromycin (150µg/mL) and ganciclovir (3µM), individual colonies were picked

and genotyped, then selected on G418 (200 µg/mL) after transfection with a pUC19-piggyBAC transposase and a piggyBac-based cargo vector containing an rtTA-expression. Creation of the ESC line is described in its entirety in a manuscript currently under revision by D.M.L., S.R.B, D.O.C, and J.M.C.

Generation of HNRNPK knockdown TSCs

Two sgRNAs targeting HNRNPK were designed using Desktop Genetics and cloned into the rtTA-BsmbI piggyBac vector from (Schertzer et al., 2018). Starter cultures for each sgRNA were maxi prepped using the Qiagen kit. Both sgRNA vectors were co-electroporated into 500,000 TSCs with Cas9-Cargo and (piggyBac) transposase vectors at a 1:1:1 ratio using the Neon® Instrument (electroporation program: 1300V, 40 ms, 1 pulse; Invitrogen). Two days after electroporation, cells were selected with G418 (200ug/ml; Gibco) and Hygromycin B (150ug/ml; Gibco) for 10 and 8 days (first and second experimental replicate, respectively), followed by 4 days of dox treatment (1ug/ml) prior to crosslinking for H3K27me3 ChIP and RNA preparation for RNA-seq, as described above.

2.4.5 RNA Isolation, qPCR, and RNA-Seq

TSCs were passaged twice off of irMEFs with 40 minutes of pre-plating prior to RNA preparation. RNA was isolated using Trizol (Invitrogen). For RT-qPCR assays in Figure S2B, 2µg of RNA was reverse transcribed using Superscript III (Invitrogen). For assays in Figures 4B and S7B, MultiScribe RT (Applied Biosystems) was used with 2.5 µg RNA. qPCR was performed using iTaq Universal SYBR Green (Biorad) and custom primers (Table S7). RNA-

Seq libraries were prepared from 1µg of RNA using Stranded mRNA-Seq Kits (Kapa Biosciences) and RNA HyperPrep Kits with RiboErase (Kapa Biosciences; Table S6).

2.4.6 ChIP-Seq

Prior to crosslinking for ChIP, TSCs were passaged one time off of irMEFs with 40 minutes of pre-plating. For all ChIPs except EZH2, cells were crosslinked in RPMI media and 10% FBS with 0.6% formaldehyde for 10 minutes at room temperature. 125mM glycine was used to quench for 5 minutes at room temperature. For Ezh2 ChIPs, cells were crosslinked in 1.5mM EGS (ethylene glycol bis(succinimidyl succinate); ThermoFisher Pierce) in PBS for 30 minutes and then in 0.6% formaldehyde for 10 minutes at room temperature, and 50mM glycine was used to quench for 10 minutes at room temperature. ChIPs were performed using 10 to 20 million cells, 5 to 10µl of antibody, and 25µl of Protein A/G agarose beads (Santa Cruz). Sonication was performed on a Vibracell VCX130 (Sonics) with cycles of 30% intensity for 30 seconds with 1 minute of rest on ice between cycles. TSCs crosslinked in 0.6% formaldehyde and EGS/formaldehyde were sonicated for 10 and 12 cycles, respectively. Crosslinked cortical neurons and ESCs were sonicated for 10 cycles. Antibodies used were: H3K27me3 (Abcam ab6002), H2AK119ub (Cell Signaling #8240), CTCF (kind gift from V. Lobanenkov), EZH2 (Cell Signaling #5246), RING1B (Cell Signaling #5694), and SMC1 (Bethyl A300-055A).
For H3K27me3 and SMC1 ChIPs, 10 million crosslinked TSCs were re-suspended in lysis buffer 1 (50mM HEPES pH 7.3, 140mM NaCl, 1mM EDTA, 10% glycerol, 0.5% NP-40, 0.25% Triton X-100, and 1x PIC (protease inhibitor cocktail; Sigma) and incubated for 10 min at 4C, and then incubated with lysis buffer 2 (10mM Tris-HCl pH 8.0, 200mM NaCl, 1mM EDTA pH 8.0, 0.5 mM EGTA pH 8.0, and 1x PIC) for 10 min at RT. For H3K27me3 ChIPs, cells were re-

suspended in lysis buffer 3 (10 mM Tris-HCl pH 8.0, 100 mM NaCl, 1 mM EDTA pH8.0, 0.5 mM EGTA pH 8.0, 0.1% Na-deoxycholate, 0.5% N- lauroylsarcosine, and 1x PIC) and then sonicated. For SMC1 ChIPs, cells were resuspended in a sonication buffer (20mM Tris-HCl pH 8.0, 150 mM NaCl, 2 mM EDTA pH 8.0, 0.1% SDS, 1% Triton X-100, and 1x PIC) and then sonicated. ChIPs were performed by incubating sonicated cell lysates at a concentration of 20 million cells/1ml of lysis buffer 3 containing 1% Triton X-100 with pre-conjugated antibody/agarose beads overnight at 4°C. After overnight H3K27me3 ChIP, beads were washed 5x in RIPA buffer (50 mM HEPES pH 7.3, 500 mM LiCl, 1 mM EDTA, 1% NP-40 and 0.7% Na-Deoxycholate) for 5 minutes each and then once in TE. After overnight SMC1 ChIP, beads were washed once with lysis buffer 3 (20mM Tris-HCl pH 8.0, 150 mM NaCl, 2 mM EDTA pH 8.0, 0.1% SDS, 1% Triton X-100, and 1x PIC), once with High Salt Buffer C (20 mM Tris-HCl pH 8.0, 500 nM NaCl, 2 mM EDTA pH 8.0, 0.1% SDS, 1% Triton X-100, and 1x PIC), once with Buffer D (10mM Tris-HCl pH 8.0, 250 mM LiCl, 1mM EDTA pH 8.0, 1% NP-40, and 1xPIC), and once with TE + 50 mM NaCl. To elute the DNA, beads were re-suspended in Elution buffer (50mM Tris pH 8.0, 10mM EDTA, and 1% SDS) and placed on a 65°C heat block for 17 minutes to 1 hour with frequent vortexing. Crosslinks were reversed overnight at 65°C, eluates were incubated with Proteinase K and RNase A, and DNA was extracted with phenol/chloroform and precipitated with ethanol.

For CTCF, EZH2, and RING1B ChIPs, 10 million crosslinked cells were resuspended in buffer 4 (50 mM Tris-HCl pH 7.5, 140 mM NaCl, 1 mM EDTA, 1 mM EGTA, 0.1% Na-deoxycholate, 0.1% SDS), sonicated to generate 200-500 bp DNA fragments, cleared via centrifugation, and diluted to 20 million cells equivalents per ml of buffer 4 containing 1% Triton-X100. Post-ChIP, beads were washed 3x with buffer 4 containing 1% Triton-X100, once

with buffer 4 containing 1% Triton-X100 and 500mM NaCl, once with buffer 5 (20 mM Tris pH 8.0, 1 mM EDTA, 250 mM LiCl, 0.5% NP-40, 0.5% Na-deoxycholate), and once with TE before eluting and reversing crosslinks as above.

DNA was prepared for sequencing on the Illumina platform using Next Reagents (NEB) and Agencourt AMPure XP beads (Beckman Coulter).

2.4.7 DNA/RNA FISH

BACs and fosmids (Key Resources Table) were ordered from the BACPAC resource center and fingerprinted with restriction digestion prior to use to verify inserted DNA. Fluorescent labeling was performed using BioPrime (Invitrogen). For RNA/DNA FISH in wild-type TSCs, BAC RP23-309H2O was used to mark the *Airn* DNA locus, and fosmid Wl1-2156F18 was used to detect *Airn* RNA. For RNA/DNA FISH in *Airn* truncation TSCs, fosmid WI1-662A5 was used to detect expression of *Igf2r*, which marks the maternal allele owing to its monoallelic expression. *Igf2r* remained monoallelic even in *Airn* truncation cells, because the polyA signal for the G418 resistance gene that we used to select for *Airn* truncations is in the same orientation as *Igf2r* transcription. The G418 polyA signal therefore causes early termination of *Igf2r* on the paternal allele, effectively suppressing its transcription even if paternal *Igf2r* might have been reactivated by truncation of *Airn*.

RNA/DNA FISH was performed essentially as in (Byron et al., 2013). TSCs were fixed on coverslips for 10 minutes in 4% paraformaldehyde/PBS, followed by a 10-minute permeabilization on ice in 0.5% TritonX-100 in PBS and 1:200 Ribonucleoside Vanadyl Complex (NEB). Coverslips were stored at -20C in 70% ethanol until use. To initiate the RNA/DNA FISH protocol, coverslips were dehydrated by serial 3-minute incubations with 75%,

85%, 95%, and 100% ethanol, and air-dried for 5 minutes. Biotinylated RNA FISH probe was added and coverslips were placed cell-side down in a chamber humidified with 50% formamide/2xSSC overnight at 37°C. After overnight incubation, coverslips were washed 3x with 50% formamide/2xSSC at 42C, 3x with 1xSSC at 50C, 1x with 1xSSC at room temperature, and 1x with 4xSSC. Each wash was 5 minutes long. After the final wash, streptavidin AlexaFluor 555 (Invitrogen) was diluted 1 to 2000 in 4xSSC 1ug/ml BSA, and added dropwise to each coverslip. Coverslips were incubated cell-side down at 37°C for one hour in a chamber humidified with 4xSSC. Coverslips were then washed for 10 minutes each with 4xSSC, 4xSSC plus 0.1% TritonX-100 (Fisher Scientific), and then 4xSSC again in a chamber humidified with 4xSSC at 37C. Coverslips were rinsed with 1xSSC then 1xPBS, then fixed with 2% paraformaldehyde in PBS for 5 minutes at room temperature. Coverslips were rinsed twice with PBS, then incubated for 5 minutes at room temperature in 200mM NaOH to degrade RNA. Cells were then rinsed with 70% ethanol and denatured at 80°C for 20 minutes in preheated 70% formamide/2xSSC. Coverslips were then washed with ice cold 70% ethanol, and with 100% ethanol, then allowed to air dry. DNA FISH probes were then added and coverslips were placed face down in a chamber humidified with 50% formamide/2xSSC overnight at 37°C. The next day, coverslips were washed 3x with 50% formamide/2xSSC at 42°C, and 3x with 1xSCC at 55°C. Each wash was 5 minutes long. Coverslips were then rinsed 1x with PBS before a 2-minute incubation in DAPI stock diluted 1:1000 in water. Coverslips were rinsed twice more and affixed to glass slides using Vectashield (VectorLabs), then sealed with nail polish.

Four dimensional datasets were acquired by taking multi-channel Z-stacks on an Olympus BX61 widefield fluorescence microscope using a Plan-Aprochromat 60X/1.4 oil objective and a Hamamatsu ORCA R2 camera, controlled by Volocity 6.3 software. Excitation was provided by

a mercury lamp and the following filters were used for the four fluorescent channels that were imaged: 377/25 ex, 447/30 em for DAPI (DAPI-5060B Semrock filter); 482/17 ex, 536/20 em for AlexaFluor488, (Semrock FITC-3540B filter); 562/20 ex, 642/20 em for AlexaFluor 555 (Semrock TXRED-4040B filter); 628/20 ex, 692/20 em for Cy5 (Semrock Cy5 4040A filter). Pixel size was 0.103 µm, Z spacing was 0.2 µm, and images had 1344x1024 pixels.

Approximately 40 Z slices were acquired for each Z-stack. Z-stacks were deconvolved using the iterative-constrained algorithm (Mediacy AutoQuantX3) with default algorithm settings. Sample settings for the deconvolution were: peak emissions for dyes (670 nm, 565 nm, 519 nm, 461 nm for Cy5, AlexaFluor 555, AlexaFluor 488 and DAPI respectively), widefield microscopy mode, NA = 1.4, RI of oil = 1.518, and RI of sample = 1.45. After deconvolution, DNA and RNA FISH signals were located using the "Spots" function in Imaris software (version 8.3.1, Bitplane) and marked with equal sized spheres. To initially call spots on all images, spot detection values were set at 0.5µm for xy and 1.5µm for z, and background subtraction and auto quality settings were used. We manually optimized the quality/sensitivity setting to match the expected 1 RNA spot or 2 DNA spots per cell. Only nuclei in which we could observe the expected number of DNA and RNA FISH signals (four and one, respectively), were counted. Images are shown as maximum intensity projections that were made using ImageJ.

To correct for chromatic aberrations that distort the relative positions of spots (in XYZ) labeled with different fluorophores, we systematically corrected all spot positions based on calibrations performed with 0.2µm diameter Tetraspeck beads (Thermofisher). These were diluted in ethanol, dried on coverslips, mounted with Vectashield, and imaged with the same filters as the sample slides, and deconvolved and analyzed with the same settings as the samples. Because each Tetraspeck bead is labeled with four fluorophores, the spots in each fluorescent channel for a

given bead should localize to the same position in XYZ. Thus, by analyzing the deviations of the detected spot positions from the actual spot positions (where all channels should be colocalized) across the field of view we were able to determine the corrections that had to be applied to compensate for the system's chromatic aberrations. These corrections were a function of the particular pair of fluorophores that needed to be compared, as well as the position in X and Y in the field of view. We found that the Z position required a shift dependent on the fluorophore pair, independent of XY position in the field (mean shift = -0.462 μm). In contrast, the XY coordinates of detected spots required a shift dependent on the position in the field of view (the difference between actual and detected spot positions between channels varied linearly across X and Y). Once we had constructed the chromatic correction model for the entire field of view in XYZ and the necessary fluorescent channel pairs, we applied the corrections to the beads and found that the distance between a detected spot in the Cy5 and AlexaFluor488 channels was smaller than 136 nm in 95% of cases. This gives an upper bound on the precision of our measurement and analysis scheme, i.e. we could confidently make statements about distances larger than 136 nm, but not smaller. We then applied the chromatic correction model to each imaged spot, prior to measuring distances between DNA FISH probes.

2.6.8 ERCC spike-ins to measure lncRNA copy # per cell

We took the following approach to calculate transcript copy number per TSC. First, we purified and quantified total RNA from known numbers of TSCs, in triplicate, to determine that the average TSC contains 30pg of RNA (not shown). Second, prior to preparing RNA-seq libraries, we added 2μl of a 1:100 dilution of ERCC RNA Spike-In Mix #1 (Invitrogen, 4456740) to 1μg of the TSC RNA sample of interest according to Invitrogen's recommendation.

We then proceeded with RNA-seq library preparation using RNA HyperPrep Kits with RiboErase (Kapa Biosciences). For lncRNA copy number per cell measurements (Figure 4), RNA-seq libraries were sequenced from the following RNA preparations: single preparations of RNA from replicate derivations of dCas9-VP160/non-targeting-sgRNA-rtTA-expressing C/B TSCs, single preparations of RNA from replicate derivations of dCas9-VP160/*Airn*-targeting-sgRNA-rtTA-expressing C/B TSCs, single preparations of RNA from replicate derivations of dCas9-KRAB/non-targeting-sgRNA-rtTA-expressing C/B TSCs, and single preparations of RNA from replicate derivations of dCas9-KRAB/*Airn*-targeting-sgRNA-rtTA-expressing C/B TSCs. Counts from wild-type TSCs and dCas9-VP160/ and dCas9-KRAB/non-targeting sgRNA TSCs were collectively considered "wild-type".

To calculate transcript copy number per cell, reads were aligned to a version of the mm9 genome with ERCC.fa sequences doped in, and a standard curve was created to link RPKM values for each of the ERCC-Spike-In RNAs to their molecular abundance (see Figure S4A for an example). These RPKM-to-abundance ratios were used to calculate molecular abundance of our lncRNAs of interest in 1µg of RNA, and this abundance was divided by 33,333 (the approximate number of TSCs that would give rise to 1µg of RNA) to determine the lncRNA copy number per TSC reported in Figure 4A.

2.4.9 Measurement of lncRNA half-life

TSCs were treated with a final concentration of 5µg/ml Actinomycin D (Sigma) for 10 minutes, 20 minutes, and 30 minutes, and 1, 2, 4, and 8 hours prior to lysis with Trizol (Invitrogen) and RT-qPCR to measure expression of *Xist*, *Airn*, and *Kcnq1ot1* relative to

GAPDH (Table S7). Actinomycin D treatment and RNA preparation was performed twice in total, once each on separate days.

To model lncRNA half-life, lncRNA levels were measured by qRT-PCR relative to Gapdh mRNA at each time point. Levels were then represented as a percentage relative to 0 hr and transformed by the natural log. Linear models using time as a predictor of RNA percentage were fit to the data for each biological replicate and then used to find the time at which the percent of RNA remaining relative to the 0 hr time point was 50%. This value was reported as the half-life.

2.4.10 RNA fractionation

To isolate RNA from cytosolic, soluble nuclear, and chromatin-bound fractions, cells were passaged once off of irMEFs with 40 minutes of pre-plating and cultured in conditioned media on a 10cm plate. *Airn*-overexpression TSCs were induced with doxycycline at 1µg/mL for 3 days. For RNA harvest, cells were washed twice with 1mL cold PBS and scraped in 1mL PBS + 1mM PMSF + 1:100 protease inhibitor cocktail (PIC, Sigma P8340). 200uL was removed at this step and cells were resuspended in 1mL Trizol (total RNA). The remaining cells were centrifuged at 1500xrcf for 5min, and resuspended in 250uL low salt solution (10mM KCl, 1.5mM MgCl2, 20mM Tris-HCl pH 7.5) supplemented with 1mM PMSF, 1mM DTT, and 1x PIC. Triton X-100 was added to a final concentration of 0.1% and cells were rotated for 10min at 4°C, then centrifuged for 5min at 1500xrcf. 200uL of supernatant was added to 1mL Trizol (cytosolic fraction). The remaining supernatant was discarded and the nuclear pellet was washed by rotating for 2min at 4°C in low salt solution without Triton X-100 and centrifuged at 1300rpm for 10min. Nuclei were resuspended in 100uL Buffer B (3mM EDTA, 0.2mM EGTA)

supplemented with 1mM DTT, 1mM PMSF, and 1x PIC, rotated for 30min at 4°C, and centrifuged at 1700xrcf for 10min. 80uL of supernatant was added to 1mL Trizol (soluble nuclear fraction). The chromatin pellet was washed by rotating for 2-5min in Buffer B and centrifuged at 1700xrcf for 10min. The pellet was resuspended in 1mL Trizol (chromatin-bound fraction). Isolation of RNA from Trizol was performed according to manufacturer protocol. Equal amounts of RNA (1µg) were reverse transcribed using the High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems) with random primers. qPCR was performed using iTaq Universal SYBR Green (Bio-Rad) and custom primers (Table S7). For a given RNA species, log-transformed Cq values for each fraction were added together, and the percentage of total signal coming from each fraction was plotted in R.

2.4.11 Stellaris RNA FISH

Custom Stellaris® FISH probes were designed against the first 40kb of *Airn* and the first 2kb of *Xist* by utilizing the Stellaris® RNA FISH Probe Designer (Biosearch Technologies, Inc., Petaluma, CA) available online at www.biosearchtech.com/stellarisdesigner and labeled with Quasar 570 and 670 dyes, respectively. The ShipReady Quasar 670 probe set was used for Gapdh (Cat# SMF-3140-1). Cells were grown on MEFs on glass coverslips for 2 days in the presence or absence of 1ug/mL doxycycline before being washed once with 1x PBS, fixed for 10min at room temperature with 4% formaldehyde in 1x PBS, washed twice with 1x PBS, and permeabilized overnight with cold 75% ethanol at 4°C. 1uL of 2.5µM probe was added to 100uL of hybridization solution (10% dextran sulfate, 2x SSC, 10% formamide) and pre-warmed to 37°C. Coverslips were washed at 37°C for 2-5min in pre-warmed wash buffer (2x SSC, 10% formamide). Coverslips were incubated with diluted probes overnight at 37°C in a humidified

chamber, then washed twice with wash buffer at 37°C for 30min, adding DAPI to 5ng/mL for the second wash. Coverslips were rinsed with 2x SSC, mounted using Prolong Gold and allowed to cure overnight at room temperature. Images were acquired and deconvolved similarly to DNA/RNA FISH images and maximum intensity projections were made using ImageJ (Schindelin et al., 2012).

2.4.12 RNA immunoprecipitation

RNA IP experiments were performed using the protocol outlined in (Raab et al., 2019). Prior to fixation, TSCs were passaged once off of irMEFs with 40 minutes of pre-plating and cultured in conditioned media. Prior to fixation of SM33 ESCs, *Xist* expression was induced with 1µg/mL doxycycline for 24 hours. TSCs and SM33 ESCs were trypsinized, washed twice with PBS, then fixed in 0.3% methanol-free formaldehyde for 30 min at 4 °C. Formaldehyde was quenched with 125 mM glycine for 5 min at room temperature. Cells were snap frozen in liquid nitrogen and stored at −80 °C.

For each IP, 10 million cells were resuspended in 0.5 ml RIPA buffer (50 mM Tris-HCl pH 8, 1% Triton X-100, 0.5% sodium deoxycholate, 0.1% SDS, 5 mM EDTA, 150 mM KCl) supplemented with 0.5 mM DTT, 1x protease inhibitor cocktail (Sigma), and 2.5 µl RNAsin (Ambion), and incubated on ice for 10 min prior to lysing using a Vibracell VCX130 (Sonics) with two cycles of 30% intensity for 30 seconds with 1 minute of rest on ice between cycles, followed by centrifugation at 4 °C for 20 min at maximum speed. Subsequently, extracts were diluted with 0.5ml fRIP buffer (25 mM Trix-HCl pH 7.5, 5 mM EDTA, 0.5% Igepal CA-630, 150 mM KCl) supplemented with 0.5 mM DTT, 1x protease inhibitor cocktail (Sigma), and 2.5 µl RNAsin (Ambion). In parallel, per IP, 25ul of protein A/G agarose beads (Santa Cruz)

were pre-conjugated with antibody overnight in PBS and 0.5% BSA at 4 °C. 10 uL of HNRNPK (Abcam, ab39975) antibody was used for HNRNPK IP in SM33 cells, and 10 ul of HNRNPK (Santa Cruz, sc28380) antibody was used for HNRNPK IP in TSCs. 10ul of CTCF antibody (a kind gift from V. Lobanenkov) was used for CTCF IP in TSCs. 10ug of mouse IgG (Invitrogen, 02-6502) was used as the "IgG control". After sonication, clarification, and dilution in fRIP buffer, extracts were united with antibody/bead mixtures and incubated overnight at 4 °C with end-over-end rotation. Beads were washed consecutively with fRIP buffer (25 mM Trix-HCl pH 7.5, 5 mM EDTA, 0.5% Ipegal CA-630, 150 mM KCl), three times in ChIP buffer (50 mM Tris-HCl pH 7.5, 140 mM NaCl, 1 mM EDTA, 1 mM EGTA, 1% Triton X-100, 0.1% sodium deoxycholate, 0.1% SDS), once in high salt buffer (ChIP buffer, but with 500 mM NaCl) and once in (20 mM Tris pH 8.0, 1 mM EDTA, 250 mM LiCl, 0.5% NP-40, 0.5% Na-deoxycholate). All washes were performed for 5 min at 4 °C. After the first and final wash, solutions were transferred to clean tubes. After the final wash, beads were resuspended in 100ul of 1x reverse crosslinking buffer (1x PBS, 2% N-lauroyl sarcosine, 10 mM EDTA) supplemented with 5 mM DTT. Per IP, 20 µl proteinase K and 1 µl RNAsin were added and samples were incubated 1 h at 42 °C, 1 h at 55 °C, and 30 min at 65 °C and were mixed by pipetting every 15 minutes. Following, samples were mixed with 1ml of Trizol. 200ul of chloroform was added, the aqueous phase was extracted and mixed with 1 volume of ethanol, vortexed, then purified using a Zymo-spin IC column, using the on-column DNase digestion as per the manufacturer's instruction. RNA was eluted in 15 µl of deionized water. RNA-seq libraries were prepared using 9ul of immunoprecipitated RNA from each condition mixed with 1 µl of 1:250 µl dilution of ERCC spike-in mix 1 (Invitrogen, 4456740). The SM33 input library was prepared from 200ng of RNA and 1ul of a 1:100 dilution of ERCC spike-ins, and the TSC input library was prepared from

100ng of RNA and 1ul of a 1:250 dilution of ERCC spike-ins. Libraries were prepared using the Kapa RiboErase kit following the manufacturer's instructions, pooled, and sequenced using single-end 75-bp reads on an Illumina Nextseq 500.

2.4.13 HNRNPK and H3K27me3 Immunofluorescence

TSCs were fixed on coverslips as described above in preparation for DNA/RNA FISH. To initiate the IF protocol, coverslips were washed twice in PBS and blocked for 30 minutes at room temperature in blocking solution (1x PBS with 0.2% Triton X-100, 1% goat serum, and 6 mg/mL IgG-free BSA). Then, coverslips were washed in 0.2% triton/1x PBS and incubated with HNRNPK antibody (Santa Cruz 28380) and H3K27me3 antibody (Cell Signaling #9733) diluted 1:200 in block solution for 1 hour at RT. Coverslips were washed 3x in 0.2% triton/1x PBS for 4 minutes each and incubated with secondary antibody (AlexaFluor 488 goat anti-mouse, A-11029 and AlexaFluor 594 goat anti-rabbit, A-11037) diluted 1:400 in block solution for 30 minutes at RT. After incubation, coverslips were washed 3x in 0.2% triton/1x PBS for 4 minutes each and rinsed 1x with PBS before a 2-minute incubation in DAPI stock diluted to 1ug/ml in water. Coverslips were rinsed twice more and mounted to glass slides using Prolong Gold (Thermo Fisher Scientific P10144). Images were acquired and deconvolved similarly to DNA/RNA FISH images and maximum intensity projections were made using Image J.

2.4.14 Protein isolation and western blotting

To isolate protein for western blotting, cells were washed with PBS, and then lysed with RIPA buffer (10 mM Tris-Cl (pH 7.5), 1 mM EDTA, 0.5 mM EGTA, 1% NP40, 0.1% sodium deoxycholate, 0.1% SDS, 140 mM NaCl) supplemented with 1 mM PMSF (Fisher Scientific)

and 1x protease inhibitor cocktail (Sigma) for 15 minutes at 4°C. Prior to western blotting, protein levels were quantified using the DC assay from Biorad. For western blotting, primary and secondary antibody incubations were done for 1hr at room temperature. Antibodies used were HNRNPK (Santa Cruz sc-28380, 1:5000 dilution), TBP (Abcam ab818, 1:2000 dilution), donkey anti-mouse IgG-HRP secondary (Santa Cruz; sc-2314; 1:2500), and donkey anti-rabbit IgG-HRP secondary (Santa Cruz; sc-2313; 1:2500).

2.4.15 Sequence alignment and processing

RNA sequence reads were aligned to genomic sequence using Star (version 2.6.0a; (Dobin et al., 2013)) and ChIP sequencing reads were aligned using bowtie2 (Langmead et al., 2009), using default parameters. All mm9 genome annotations were obtained from the UCSC genome browser. Variant sequence data was obtained from the Sanger Institute (http://www.sanger.ac.uk/resources/mouse/genomes/). Samtools was used to filter for reads that had a mapping quality greater than or equal to 30 (Li et al., 2009). CAST/EiJ pseudogenome creation and allele-specific read retention was performed as in (Calabrese et al., 2012, 2015). All genome-related plots were generated using R.

2.4.16 ChIP-Seq Peak Calling

ChIP-Seq data between wild-type C/B and B/C TSCs were pooled to identify peaks in H3K27me3, CTCF, SMC1, and EZH2 datasets. Peaks of RING1B in TSCs were identified by pooling ChIP-Seq data from *Airn* OE, *Airn* WT, *Airn* KO, and *Kcnq1ot1* KO TSCs. H3K27me3 peaks in ESCs were called separately for the 0hr and 72hr datasets. RING1B peaks in ESCs were identified by pooling 0hr, 12hr, and 72hr datasets. hiddenDomains 3.0 was used to call peaks

from H3K27me3 ChIP-Seq data, using input DNA sequenced from formaldehyde crosslinked TSCs or ESCs as a control (Starmer and Magnuson, 2016). Default parameters plus max.read.count=10 were used in the main hiddenDomains.R script, and neighboring enriched bins were merged to generate the final set of H3K27me3 peaks. MACS2 was used to call peaks from all other ChIP-Seq datasets, with the same input used in hiddenDomains and parameters –broad –broad-cutoff 0.01 (Zhang et al., 2008). All peaks were called using allele-nonspecific data, owing to its significantly higher coverage relative to allele-specific data.

2.4.17 Parent-of-origin bias in H3K27me3 peaks

Allele-specific reads falling within peaks of H3K27me3 were counted in the C/B and B/C TSC datasets, and counts were imported into edgeR and normalized using edgeR's counts per million (CPM) metric (Robinson et al., 2010). H3K27me3 ChIP-seq data in C/B TSCs was from (Calabrese et al., 2012), and B/C data was generated as part of this study. Allelic data from individual H3K27me3 replicates within C/B and B/C TSC datasets were merged before importing into edgeR. Autosomal peaks with >1 allelic cpm in each dataset were analyzed. X-linked peaks were excluded. Differential enrichment between Cast and B6 alleles within each individual peak in each F1-hybrid TSC line was tested via edgeR's generalized linear model likelihood ratio test, and p-values from both tests were adjusted for false discovery using the Benjamini-Hochberg correction method. Peaks exhibiting PO biases with false discovery rates scores of ≤ 0.1 in both C/B and B/C TSCs were considered to be significantly biased (Table S1).

2.4.18 Allelic changes upon lncRNA truncation

Allelic reads were retained that fell within exons as well as introns for all UCSC Known Genes. Counts were then used to calculate the proportion of paternal expression for each gene in each sample, and this proportion was then arcsine transformed. To detect differentially expressed genes, two tailed t-tests were performed on the transformed data comparing three knockout TSC lines to five wild-type TSC lines. For genes in the *Airn* region, the three *Kcnq1ot1* knockout TSC lines plus the two wildtype C/B TSC lines were used as *Airn* wild-type. Similarly, for genes in the *Kcnq1ot1* region, the three *Airn* knockouts plus the two wildtype C/B TSC lines were used as *Kcnq1ot1* wild-type. The arcsine transformation was used to eliminate the bounds of '0' and '1' in proportions and to spread the data out at the extremes and only the extremes, thereby validating the assumptions inherent in a two-sided t-test. See Table S2.

2.4.19 Genome alignability

The proportion of mm9 that could be uniquely mapped using 45 or 75bp sequence tags, depending on the length of read from the dataset in question, was defined as genome alignability.

2.4.20 Genome-wide correlations in ChIP-Seq datasets

To derive Pearson's r values between H3K27me3 and H2AK119ub pooled datasets, reads were counted in 10kb bins genome wide using bedtools coverage (Quinlan and Hall, 2010). Counts were normalized for dataset size. For H3K27me3 and H2AK119ub comparisons, only bins with total H3 rpm >1 were used (H3 data from (Calabrese et al., 2012)). To derive Pearson's r values between RING1B and EZH2 density with H3K27me3 peaks on the X chromosome,

paternal reads were counted per H3K27me3 peak. Only peaks with an average of >10 paternal reads per dataset were used.

2.4.21 Bootstrap approach for FISH measurements

We used a bootstrap approach to determine if the correlations between paternal H3K27me3 density in the nine regions probed for DNA FISH and distance in base pairs to *Airn*, maternal distance to *Airn*, and paternal distance to *Airn* (Figure 2F) were significantly different. For each iteration of the bootstrap, we calculated the mean values for the bootstrap-dataset on the maternal and paternal allele, and used a linear regression fit to determine the R-squared value between the maternal and paternal distance measurements and paternal H3K27me3. We repeated this process 10,000 times and used the distributions of R-squared values under each condition and either a one-sample (Figure 2Fii) or two-sample (Figure 2Fiii) test to calculate empirical p-values that assess whether the differences in R-squared values between our three comparisons were significantly different.

2.4.22 Hi-C and ChIA-PET data

Hi-C data was downloaded from and processed using Juicebox (Durand et al., 2016). ESC SMC1 ChIP-PET contact calls were from (Dowen et al., 2014), and ESC SMC1 and CTCF ChIP-Seq data were from (Kagey et al., 2010; Stadler et al., 2011).

2.4.23 Chromosome tiling plots using bedtools

Chromosome-scale H3K27me3 tiling density plots in Figures 4, 5, 6, 7, S6, and S7 were created by summing total H3K27me3 counts in 40kb bins across each chromosome, moving in 4kb increments (using bedtools coverage on sorted bam files). All counts per bin were

normalized for alignability, where total reads per bin were divided by the proportion of alignable bases per bin. Bins with alignability of less than 0.5 (i.e. less than 50% alignable at 75bp (Figures 4, 5, 6, 7, and S6) and 45bp resolution (Figure S7) were excluded from tiling density plots to avoid potential uncertainty that would be introduced by normalizing highly non-unique regions.

In Figures 4, 5, 6, S6, and S7, read counts were normalized between datasets by multiplying by 1 million then dividing by the total number of reads per dataset. Read normalization for HNRNPK knockdown experiments (Figure 7) is described in a separate subheaded section below.

2.4.24 H3K27me3 ChIP normalization using *Drosophila* DNA

It was possible that HNRNPK knockdown would cause a global reduction in H3K27me3 that if left unaccounted for during library preparations, would have obscured a reduction in H3K27me3 in lncRNA target domains. To circumvent this possibility, after H3K27me3 ChIP in non-targeting and HNRNPK knockdown TSCs and prior to the preparation of sequencing libraries, an amount of sonicated *Drosophila* melanogaster DNA (kind gift of D. McKay) equal to 1% of the total amount of DNA in the lowest yielding ChIP sample (74.2 picograms) was added to 10ul of each ChIP (one third of the total volume of eluted ChIP'd DNA for each sample). After sequencing, reads were aligned to the mm9 and dm6 genomes. Normalization factors were created by dividing the total number of aligned *Drosophila* reads in each sample by the lowest *Drosophila* read count amongst samples, giving a factor of 1 for the lowest sample and values greater than 1 for all other samples. Binned read counts (from mm9) were divided by these normalization factors and then divided by the input DNA amount in ng for each IP. To be

able to directly compare Y-axes displayed in Figure 7B and C, read counts were then divided by bin size in kb, which was 40kb for Figure 7B and 4kb for Figure 7C.

2.4.25 Determining feature overlap using bedtools

To determine feature overlap for Venn diagrams in Figure S3 and for CGI classification in Figures 5, 6, 7, and S6, bedtools intersect was used on MACS2 "broadpeak" files and UCSC-annotated CGIs.

2.4.26 Metagenes

Allele specific metagene plots in Figures 5, 6, and S6 were constructed using the following approach. For each dataset, counts of sequence read starts were recorded in 500bp bins surrounding the annotated TSS or center of the feature for the gene/feature class in question. In addition to normalization for gene number, allelic counts were normalized for the number of uniquely alignable SNPs present in each bin for the specific gene/feature set being analyzed. Non-allelic metagenes in Figures 6 and S3 were generated with HOMER (Heinz et al., 2010). To create tag directories of aligned reads, "makeTagDirectory" was used. Then, "annotatePeaks.pl" was used to generate metagenes with 500bp bins in a 100kb window for Figure 6 and 50bp bins in a 4kb window for Figure S3. The "Coverage" column was used for plotting.

2.4.27 Measurement of Signal over IgG in HNRNPK RNA IP data

RNA-IP reads were aligned to a version of the mm9 genome with ERCC.fa sequences doped in. Samtools was used to filter aligned reads for q>30. (Li et al., 2009). Reads were overlaid with UCSC known gene annotations using featureCounts to determine the read count

per transcript (Liao et al., 2014). For normalization, counts per ERCC spike-in transcript were generated for each dataset using featureCounts on the ERCC92.gtf file. The upper quartile values from the set of ERCC spike-in transcripts quantified for each dataset were used to normalize all datasets relative to their respective total RNA input dataset (either SM33 or TSC). Wiggle tracks in Figure 7A and transcript read counts in Table S5 were scaled using these factors. HNRNPK and CTCF normalized counts were divided by IgG normalized counts to give the signal relative to IgG values that are reported in Figure 7A.

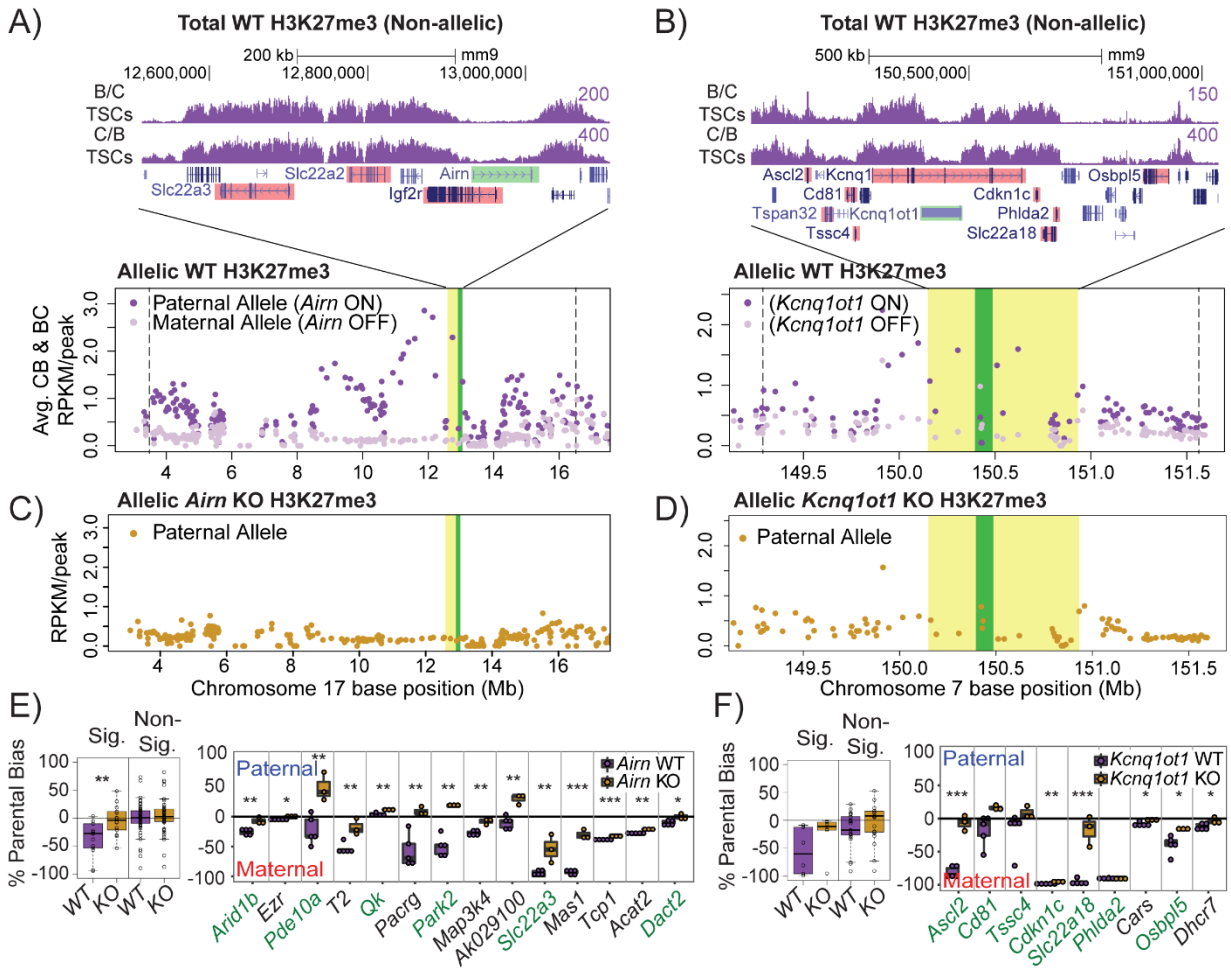


Figure 2. 1. Megabase-sized domains of H3K27me3 require continued expression of *Airn* and Kcnq1ot1.

(A, B) Paternally-biased H3K27me3 surrounds *Airn* and *Kcnq1ot1* in TSCs. Dark/light purple dots, paternal/maternal signal in H3K27me3 peaks avgd. from C/B + B/C TSCs. Green bars, *Airn*/*Kcnq1ot1* loci. Dotted lines, first/last biased H3K27me3 peak used to define domain size (13 Mb for *Airn* and 2.3 Mb for *Kcnq1ot1*). Yellow shading/upper insets, non-allelic H3K27me3 data in previously-defined *Airn*/*Kcnq1ot1* domains relative to maternally/paternally-biased genes (pink/green, respectively). (C, D) Paternal signal in H3K27me3 peaks in truncation C/B TSCs (gold). Peaks are as in A, B. (E, F) Parent-of-origin expression in wild-type (purple) and truncation (gold) C/B TSCs. Paternal/maternal biases represented from 0 to 100; maternal values multiplied by -1. Value of 0, equal expression from both alleles. Value of 100, 100% expression from one allele. Green name, known lncRNA target. Asterisks, genes de-repressed in lncRNA truncation (***, p<0.001; **, p<0.01; *, p<0.05, two-tailed t-test). Sig./Non-Sig, avg. bias of de-repressed/non-target gene.

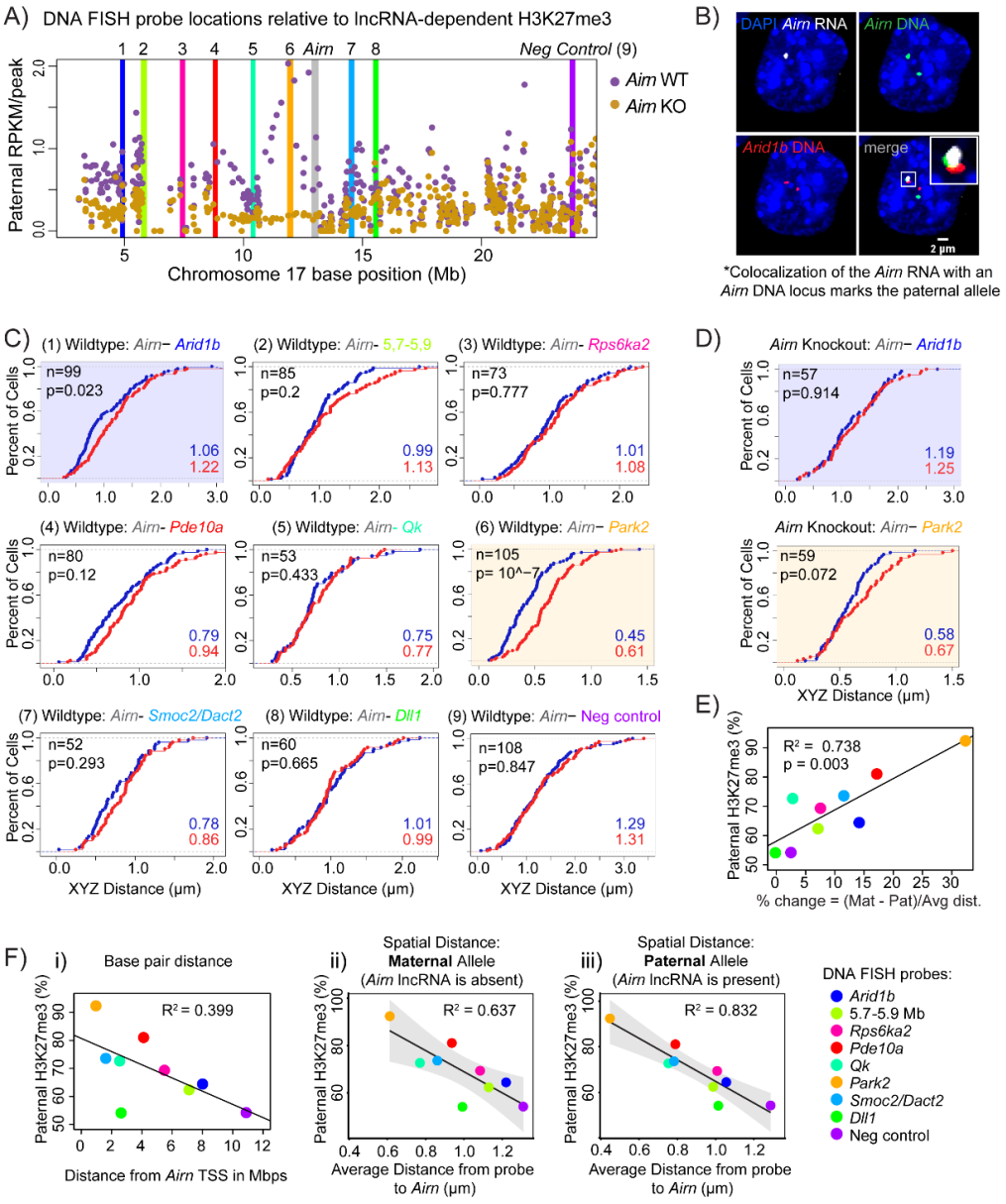


Figure 2. 2. Spread of H3K27me3 in the *Airn* domain is influenced by pre-existing genomic architecture and additional features on the paternal allele.

(A) DNA FISH probe location vs. paternal H3K27me3. (B) Representative FISH image. White box, signal overlap on paternal allele. Scale bar, 2µm. (C, D) Cumulative distribution plots for probes in (A). Spatial distance to *Airn* shown on paternal (blue) and maternal (red) alleles. Blue/red numbers, avg. distance on paternal/maternal alleles. n, # of cells. p-values from two-sample KS-tests. Shaded plots in (C; wild-type TSCs) correspond to those in (D; *Airn*-truncation TSCs). (E, F) Correlation between paternal H3K27me3 in regions probed for DNA FISH and (E) % avg. difference in distance between maternal and paternal alleles and (F): (i) distance in base pairs to *Airn* TSS, and avg. distance measured via FISH from probe to *Airn* on (ii) maternal and (iii) paternal alleles from (C). Grey in (ii) and (iii), 95% confidence intervals.

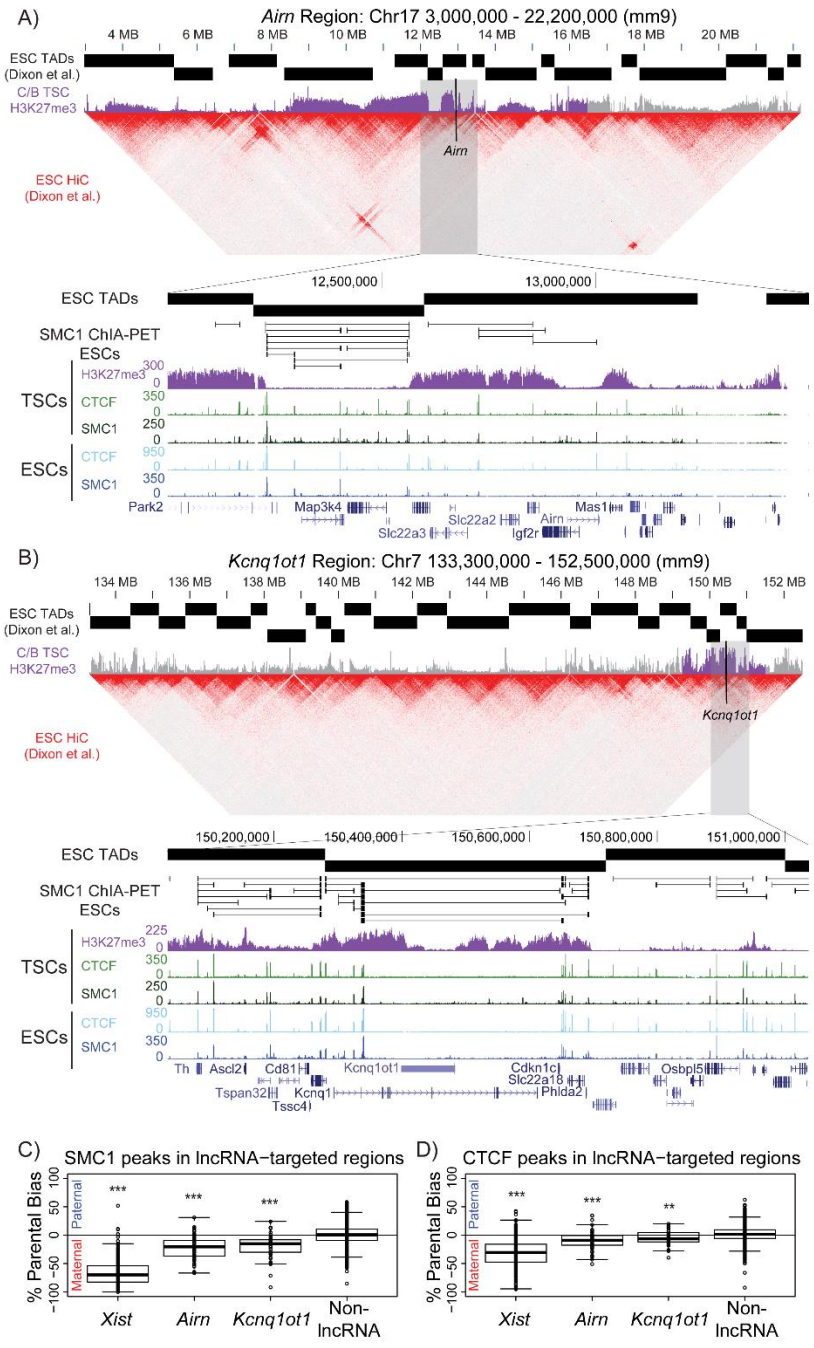


Figure 2. 3. Intensity of H3K27me3 in lncRNA target domains correlates with TADs, DNA loops, and SMC1 and CTCF binding.
(A, B) Hi-C data/TADs (Dixon et al., 2012), SMC1 loops (Dowen et al., 2014), SMC1/CTCF binding (Kagey et al., 2010; Stadler et al., 2011) in ESCs, SMC1/CTCF binding in TSCs around *Airn* (A) and *Kcnq1ot1* (B). Purple, non-allelic H3K27me3 signal (C/B TSCs). H3K27me3 shading turns gray at last detected peak of paternally-biased H3K27me3. (C, D) Avg. parent-of-origin bias from C/B and B/C TSCs of (C) SMC1 and (D) CTCF peaks in *Xist*, *Airn*, and *Kcnq1ot1* target regions. ***, p<0.001; **, p<0.01 relative to "Non-lncRNA" (all other autosomal peaks in genome); Tukey's HSD test. Scales as in Figure 1E, F.

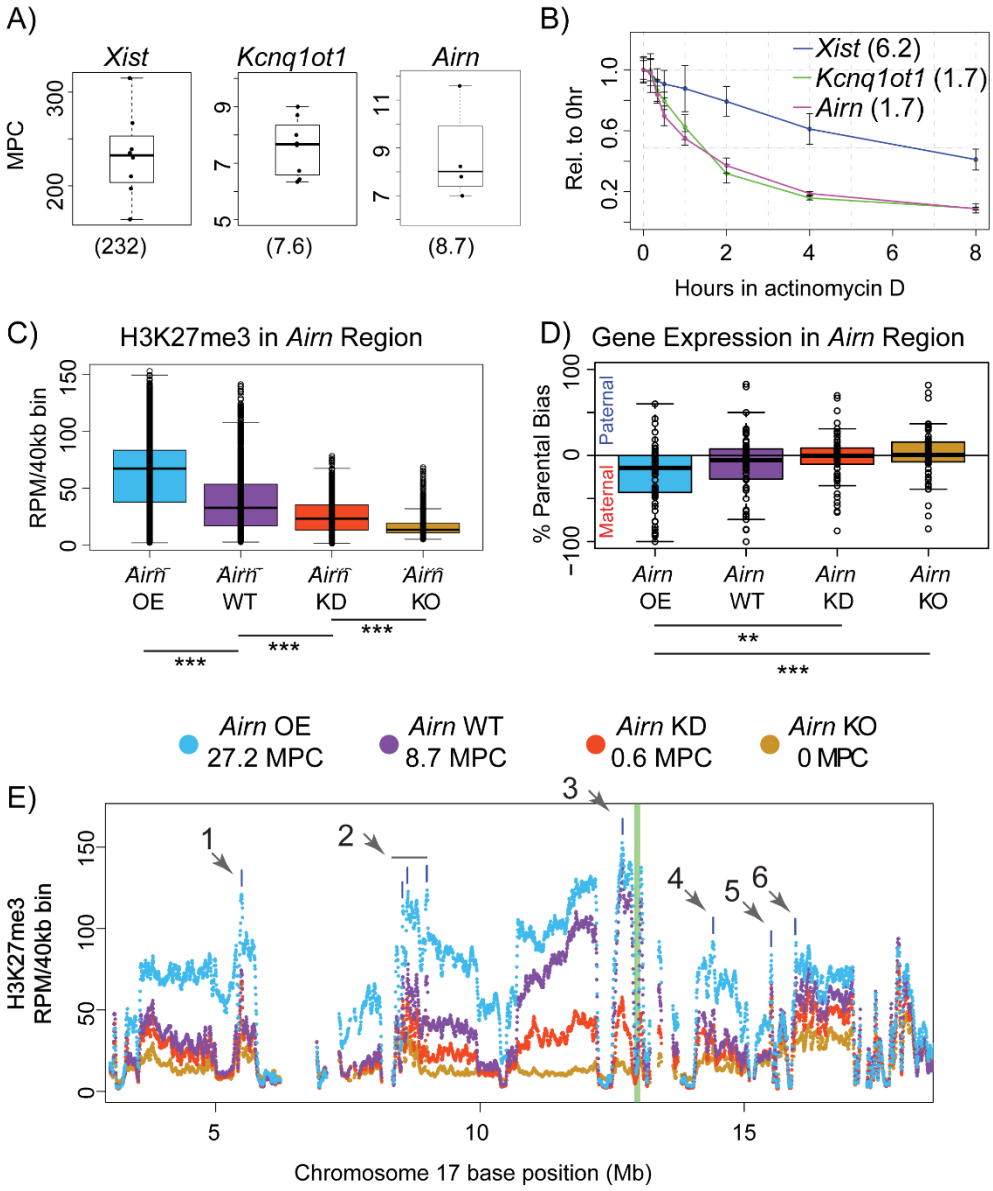


Figure 2. 4. LncRNA repressive potency correlates with abundance, stability, and underlying features of the genome.
(A) Molecules per cell (MPC) of *Xist*, *Airn*, *Kcnq1ot1* by RNA-Seq. (B) Stability of *Xist*, *Airn*, *Kcnq1ot1* in TSCs after 5µg/ml of Actinomycin D. Mean ±SD half-lives in parentheses. (C) Boxplots of H3K27me3 density in 40kb sliding bins across *Airn* domain, and (D) parent-of-origin expression for 61 considered genes (Figure 1) in *Airn*-overexpression (OE), -wild-type (WT), knockdown (KD), and truncation (KO) TSCs. ***, p<0.001; **, p<0.01, Tukey's HSD test. Y-axis as in Figure 1E, F. (E) H3K27me3 density in 40kb sliding bins across *Airn* domain in OE, WT, KD, and KO TSCs. *Airn* MPC is above the density plot. Blue ticks + grey arrows, CGIs in Figure 5A. Green bar, *Airn* locus. "WT", TSCs expressing dCas9-VP160 and a non-targeting sgRNA.

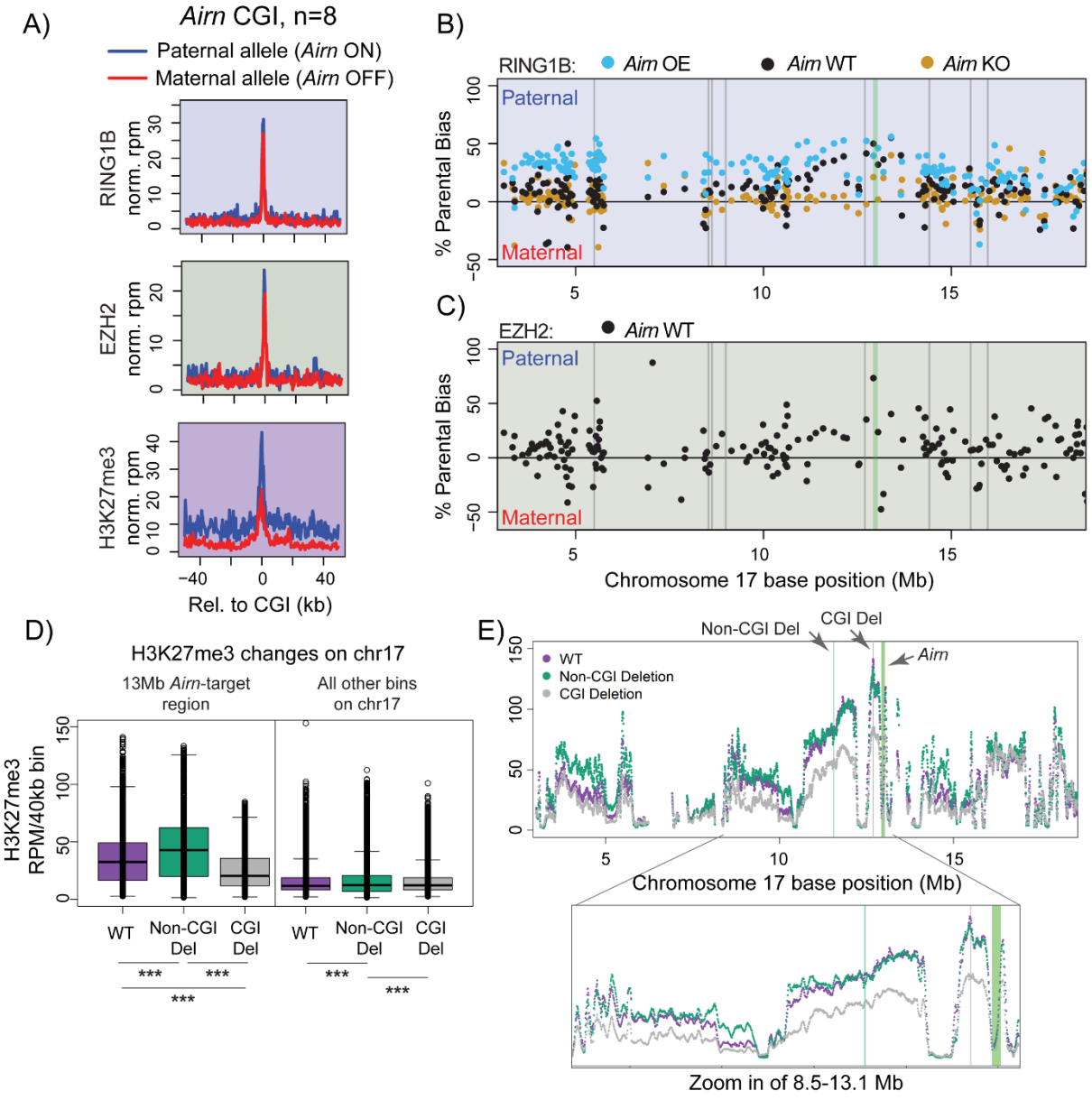


Figure 2. 5. CGIs bind PRCs autonomously and can nucleate spread of H3K27me3 by *Airn*. (A) Allelic metagene plots of RING1B, EZH2, H3K27me3 relative to 8 CGIs in H3K27me3 nucleation centers in *Airn* domain (i.e. blue ticks in Figure 4E). (B, C) Parent-of-origin bias in RING1B and EZH2 in peaks of H3K27me3 in *Airn* domain. Data from *Airn* wild-type (WT), overexpression (OE), and truncation (KO) TSCs shown. Green line, *Airn* locus. Grey lines, CGIs from in (A). Y-axis as in Figure 1E, F. Panels shaded in A-C for clarity. (D) Boxplot and (E) tiling plot of H3K27me3 density in 40kb bins sliding across *Airn* domain in WT, Non-CGI deletion, and CGI-deletion TSCs. (D) also shows non-*Airn* bins on chr17; note marginal increase in non-CGI relative to WT and CGI. ***, p<0.001, Tukey's HSD test. Vertical lines in (E), *Airn*, Non-CGI, and CGI deletion location. For Non-CGI and CGI, data shown is avg. of two clones.

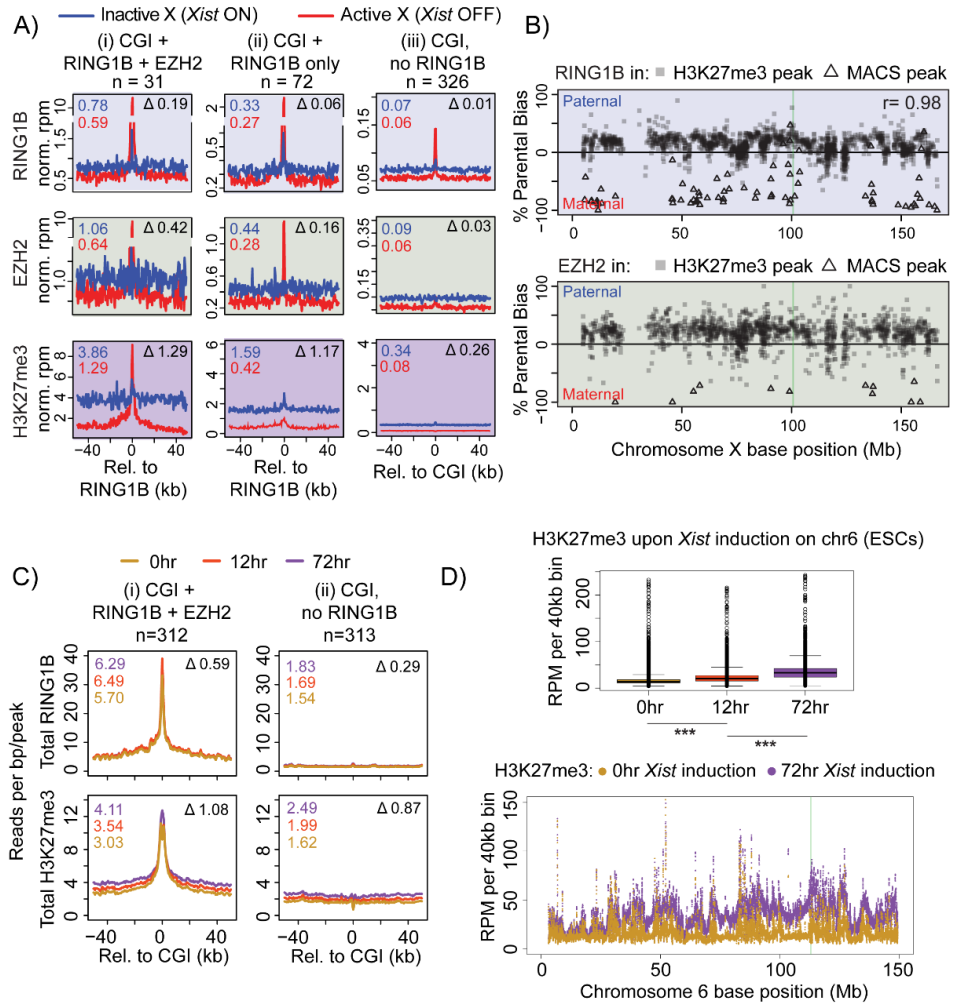


Figure 2. 6. Xist-induced H3K27me3 density is highest around CGIs that bind PRCs autonomously.

(A) Metagenes of allelic RING1B, EZH2, and H3K27me3 density at CGIs that coincide w/peaks of (i) RING1B and EZH2, (ii) RING1B only, or (iii) neither. Median signal on the active (red) and inactive X (blue) in the metagene window is shown in upper left, and difference in medians in upper right. Y-axes are broken in select plots to visualize trends on both X's. (B) Parent-of-origin bias of RING1B (top) and EZH2 (bottom) in peaks of H3K27me3 (squares) and peaks of RING1B and EZH2 (triangles). Green line, *Xist* locus. Y-axis as in Figure 1E, F. Pearson correlation between RING1B and EZH2 density in H3K27me3 peaks in upper right. Panels shaded in A+B for clarity. (C, D) H3K27me3 spreads from PRC-bound CGIs upon *Xist* induction in mouse ESCs. (C) Non-allelic RING1B and H3K27me3 density centered at chr6 CGIs bound by (i) RING1B+EZH2 or (ii) neither. Density is shown for three timepoints of *Xist* induction: 0hr, 12hr, and 72hr. Upper left, median at each timepoint. Upper right, difference in 72hr and 0hr medians. (D) Boxplot and tiling density of H3K27me3 across chr6 at 0hr (no *Xist* expression), 12hr, and 72hr *Xist* induction. 12hr tiling plot not shown for clarity. Green line, *Xist* insertion on chr6. ***, p<0.001, Tukey's HSD test.

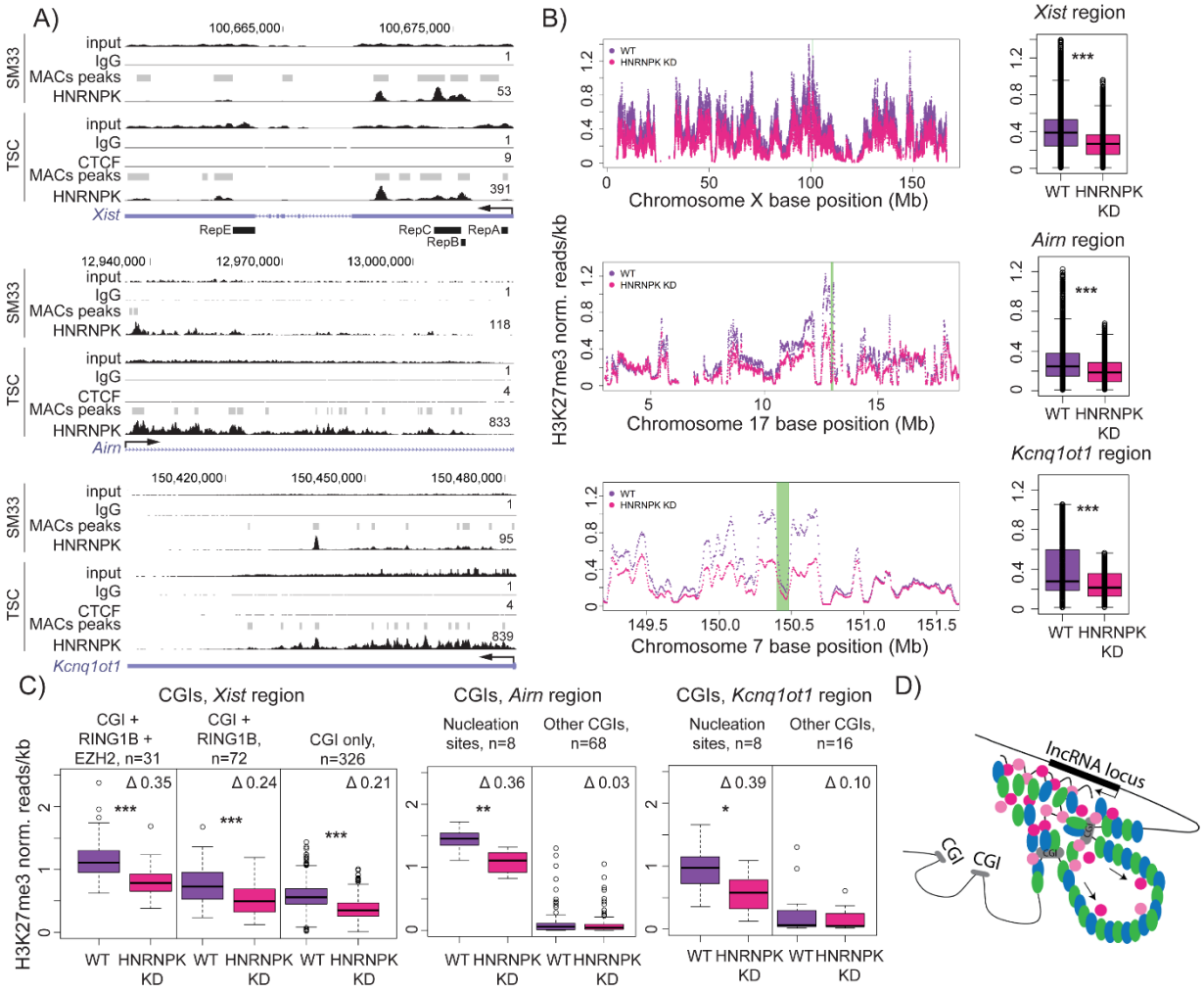


Figure 2. 7. Xist, *Airn*, and Kcnq1ot1 require HNRNPK to spread H3K27me3.
(A) Wiggle tracks of RNA-IP data for input, IgG, and HNRNPK in SM33 ESCs + TSCs and CTCF in TSCs across *Xist*, *Airn*, and *Kcnq1ot1*. Blocks above HNRNPK tracks, MACS peaks. Right-justified numbers, signal over IgG. *Xist* repeats are below *Xist* diagram. (B) H3K27me3 tiling density and boxplots in *Xist*, *Airn*, and *Kcnq1ot1* domains in WT and HNRNPK knockdown TSCs. Green bars, lncRNA loci. ***, p<0.001, two-tailed t-test. (C) Boxplots of H3K27me3 density +/- 2kb from CGI centers in *Xist*, *Airn*, and *Kcnq1ot1* domains. Difference in means between WT and knockdown, upper right corner. ***, p<0.001; **, p<0.01; *, p<0.05, two-tailed t-test. "Nucleation sites" in *Airn*/*Kcnq1ot1* plots, CGIs from Figures 5A/S6A. (D) Model: Super-stoichiometric interactions between proteins such as HNRNPK (pink circles) that bind lncRNAs (squiggles) and PRC1/2 (blue/green ovals) concentrate PRCs in lncRNA foci. These same interactions tether lncRNA foci to CGIs (grey ovals) pre-bound by PRCs, nucleating PRC spread in contacted regions.

2.5 Supplemental Table Legends

Supplemental Table 2.1. Autosomal peaks of H3K27me3 and their allelic biases. Columns A-C, "chr", "start", "end", give the coordinates for all autosomal H3K27me3 peaks in TSCs called using the hiddenDomain algorithm. Throughout the table, "cb" refers to a cell line derived from a CAST/EiJ (Cast) mother and C57BL/6J (B6) father. "bc" refers to the reciprocal cell lines derived from a B6 mother and Cast father. Columns D-G give raw allelic H3K27me3 read counts from wildtype C/B and B/C cell lines. "cb.wt.b6", B6 (paternal) allele in C/B cells. "cb.wt.cast", Cast (maternal) allele in C/B cells. "bc.wt.b6", B6 (maternal) allele in B/C cells. "bc.wt.cast", Cast (paternal) allele in B/C cells. These were the four columns read into the EdgeR script to determine significant biases in H3K27me3 peaks. Columns H and I, "cb.FDR" and "bc.FDR", give adjusted p-values derived from comparing read counts in columns D versus E and F versus G, respectively, which are output from EdgeR.

Columns J-S give raw allelic H3K27me3 read counts from the *Airn* and *Kcnq1ot1* truncation C/B TSCs. Columns "wt.b6" and "wt.cast" are the paternal B6 and maternal Cast read counts for an additional wildtype replicate that was prepared alongside the four truncation samples. Columns "ako10.b6" and "ako10.cast" are the paternal B6 and maternal Cast read counts for one *Airn* truncation ChIP-seq replicate. Columns "ako24.b6" and "ako24.cast" are the paternal B6 and maternal Cast read counts for the second *Airn* truncation ChIP-seq replicate. Columns "kko2.b6" and "kko2.cast" are the paternal B6 and maternal Cast read counts for one *Kcnq1ot1* truncation ChIP-seq replicate. Columns "kko3.b6" and "kko3.cast" are the paternal B6 and maternal Cast read counts for the second *Kcnq1ot1* truncation ChIP-seq replicate. Columns T and U, "k119.cb.wt.b6" and "k119.cb.wt.cast" give paternal B6 and maternal Cast H2AK119ub read counts from C/B TSCs within the hiddenDomains H3K27me3 called peaks. To make sorting peaks by bias easier, column V, "Bias", indicates the bias as "None", "Strain: B6", "Strain: CAST", "PO: Paternal", or "PO: Maternal", where PO stands for parent of origin bias.

Supplemental Table 2.2. Gene expression changes in *Airn* and *Kcnq1ot1* domains. Table shows gene expression measured via polyA RNA-sequencing for the 61 of 123 genes in the *Airn*-targeted region and the 27 of 43 genes in the *Kcnq1ot1*-targeted region that pass our expression threshold (average of > 10 allelic reads per dataset). Reads aligning to introns are included. Columns A-E, "gene", "chr", "strand", "start", "end", give the name coordinates of these genes. Columns F-U give the raw allelic RNA-seq read counts for eight C/B TSC datasets. "wt1.b6", B6 (paternal) allele in wildtype replicate number one. "wt1.cast", Cast (maternal) allele in wildtype replicate number one. "wt2.b6", B6 (paternal) allele in wildtype replicate number two. "wt2.cast", Cast (maternal) allele in wildtype replicate number two. Columns "ako2.b6" and "ako2.cast" are the paternal B6 and maternal Cast read counts for one *Airn* truncation RNA-seq replicate. Columns "ako10.b6" and "ako10.cast" are the paternal B6 and maternal Cast read counts for the second *Airn* truncation RNA-seq replicate. Columns "ako24.b6" and "ako24.cast" are the paternal B6 and maternal Cast read counts for the third *Airn* truncation RNA-seq replicate. Columns "kko2.b6" and "kko2.cast" are the paternal B6 and maternal Cast read counts for one *Kcnq1ot1* truncation RNA-seq replicate. Columns "kko3.b6" and "kko3.cast" are the paternal B6 and maternal Cast read counts for the second *Kcnq1ot1*

truncation RNA-seq replicate. Columns “kko4.b6” and “kko4.cast” are the paternal B6 and maternal Cast read counts for the third *Kcnq1ot1* truncation RNA-seq replicate. Columns V-AC, “wt1.pat”, “wt2.pat”, “ako2.pat”, “ako10.pat”, “ako24.pat”, “kko2.pat”, “kko3.pat”, and “kko4.pat”, give the proportion of paternal reads in each dataset. Columns AD-AK, “wt1.trans”, “wt2.trans”, “ako2.trans”, “ako10.trans”, “ako24.trans”, “kko2.trans”, “kko3.trans”, “kko4.trans”, give arcsine transformation of the paternal proportion for the respective dataset. Column AL, “ttest” gives the p-value from a two-tailed t-test comparing the three truncations versus the five-wildtype (two wildtypes plus the three truncation datasets for the other lncRNA) arcsine transformed values.

Supplemental Table 2.3. Properties of CGIs in the *Airn* domain. Columns A and B, “start” and “end”, give the coordinates of all CGIs within the *Airn*-targeted region on chromosome 17 (start < 16526000). Column C, “cpg.num”, gives the number of CpG’s in each island. Column D, “cgi.rank”, gives the order of each CGI from highest number of CpG’s within the island to lowest. Columns E-I, “k27.peak”, “ring1b.peak”, “ezh2.peak”, “smc1.peak”, and “ctcf.peak”, have a ‘Yes’ if the CGI overlaps a peak of H3K27me3, RING1B, EZH2, SMC1, or CTCF in wildtype TSCs, respectively. Columns J-N, “oe.k27.rpm”, “wt.k27.rpm”, “kd.k27.rpm”, “wt.ring1b.rpm”, and “wt.ezh2.rpm”, give the ChIP-seq reads per million □1kb from the center of the CGI in the following samples (all C/B TSCs): H3K27me3 in *Airn* overexpression TSCs, H3K27me3 in *Airn* wildtype TSCs, H3K27me3 in *Airn* knockdown TSCs, RING1B in *Airn* wildtype TSCs, and EZH2 in *Airn* wildtype TSCs. Columns O and P, “ring1b.b6” and “ring1b.cast” are the paternal B6 and maternal Cast ChIP-seq read counts for RING1B in wildtype TSCs. Columns Q and R, “ezh2.b6” and “ezh2.cast” are the paternal B6 and maternal Cast ChIP-seq read counts for EZH2 in wildtype TSCs. The final columns, “gene”, “wt.rpkm”, and “a10.rpkm”, give the gene name, *Airn* wildtype non-allelic gene expression, and *Airn* truncation non-allelic gene expression if that gene’s promoter overlaps that CGI. If the CGI overlaps two gene promoters, both genes are given.

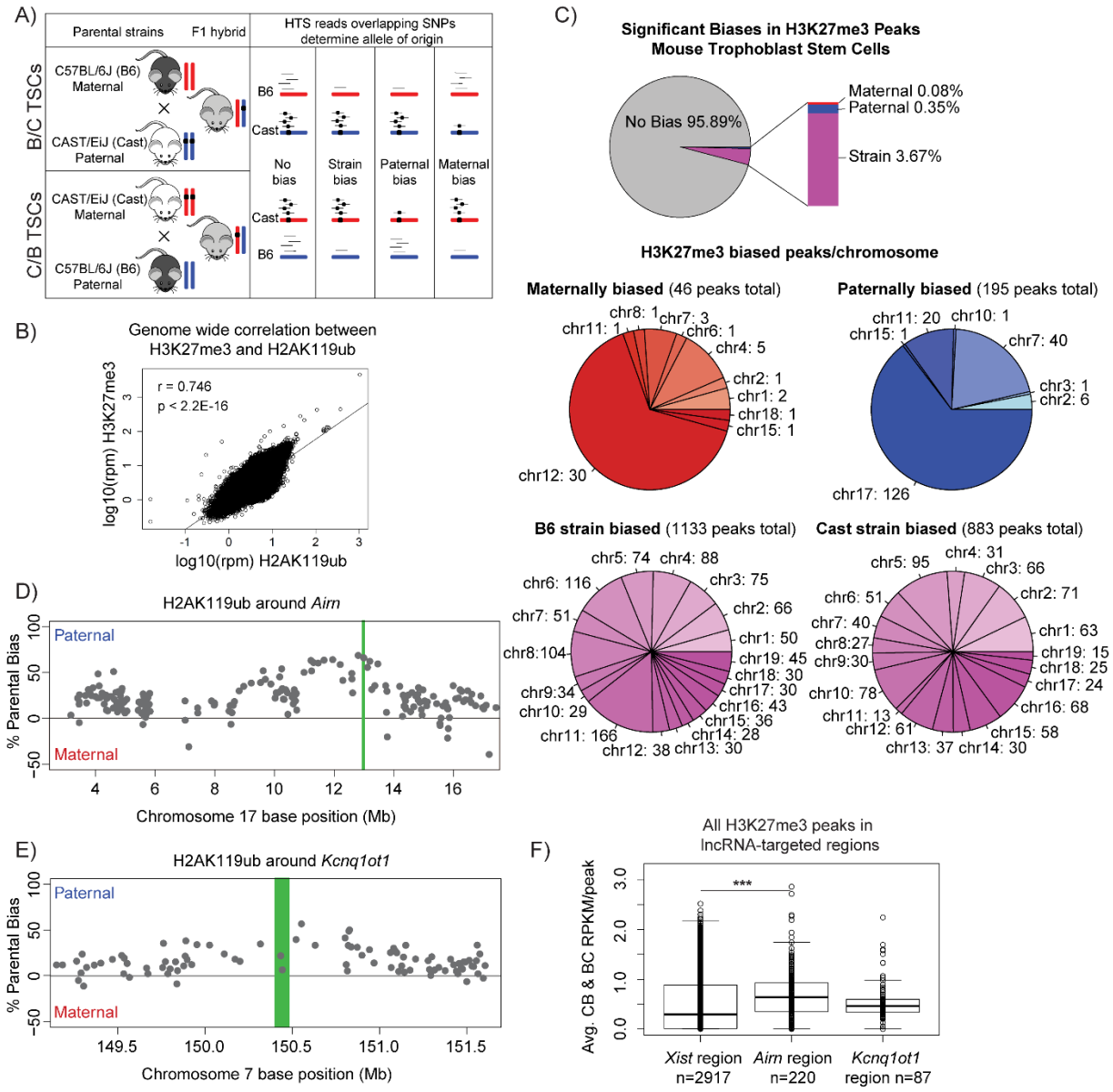
Supplemental Table 2.4. Properties of CGIs on the inactive X. Columns A and B, “start” and “end”, give the coordinates of all CGIs on the X chromosome. Column C, “cpg.num”, gives the number of CpG’s in each island. Columns D-H, “k27.peak”, “ring1b.peak”, “ezh2.peak”, “smc1.peak”, and “ctcf.peak”, have a ‘Yes’ if the CGI overlaps a peak of H3K27me3, RING1B, EZH2, SMC1, or CTCF in wildtype TSCs, respectively. Columns I-K, “wt.k27.rpm”, “wt.ring1b.rpm”, and “wt.ezh2.rpm”, give the ChIP-seq reads per million □1kb from the center of the CGI in the following samples (all C/B TSCs): H3K27me3 in wildtype TSCs, RING1B in wildtype TSCs, and EZH2 in wildtype TSCs. Columns L and M, “k27.b6” and “k27.cast”, are the paternal B6 and maternal Cast ChIP-seq read counts for H3K27me3 in wildtype TSCs. Columns N and O, “ring1b.b6” and “ring1b.cast”, are the paternal B6 and maternal Cast ChIP-seq read counts for RING1B in wildtype TSCs. Columns P and Q, “ezh2.b6” and “ezh2.cast”, are the paternal B6 and maternal Cast ChIP-seq read counts for EZH2 in wildtype TSCs. Column R, “inact”, has ‘Yes’ if the gene that overlaps the CGI was classified as an X-inactivated gene in TSCs in ((Calabrese et al., 2012); note that many of the CGI-containing genes with the highest levels of H3K27me3 were not classified as X-inactivated genes in 2012; at that time, we were

limited by our ability to call allele-specific expression owing to short read length (35nt) and limited read depth, and thus many lowly expressed genes on the X were invisible). The final columns, “gene”, “gene.rpkm”, “gene.b6”, and “gene.cast”, give the gene name, normalized wildtype non-allelic gene expression, raw B6 paternal RNA-seq read counts, and raw Cast maternal RNA-seq read counts if that gene’s promoter overlaps that CGI. If the CGI overlaps two gene promoters, both genes are given.

Supplemental Table 2.5. Sequencing counts per transcript for HNRNPK, CTCF, and IgG RNA-IP. Columns A-M are the output of featureCounts using UCSC mm9 annotated genes as features. “Geneid”, gene name. “Chr”, chromosome. “Start”, start of transcript. “End”, end of transcript. “Strand”, strand of transcript. “Length”, length of transcript. Columns G-I, “SM33 IgG”, “SM33 HNRNPK”, and “SM33 input”, give the unnormalized RNA-IP read counts per transcript in SM33 cells for IgG, HNRNPK, and input. Columns J-M, “TSC IgG”, “TSC HNRNPK”, “TSC CTCF”, and “TSC input” give the unnormalized RNA-IP read counts per transcript in C/B TSCs for IgG, HNRNPK, CTCF, and input. Upper quartile values from ERCC spike-in controls were used to normalized reads relative to the input for each cell type. Columns N-R, “sm.igg.norm”, “sm.hk.norm”, “ts.igg.norm”, “ts.hk.norm”, and “ts.ctcf.norm”, give these normalized values for each dataset. “sm”, SM33 cells. “ts”, TSCs. “hk”, HNRNPK.

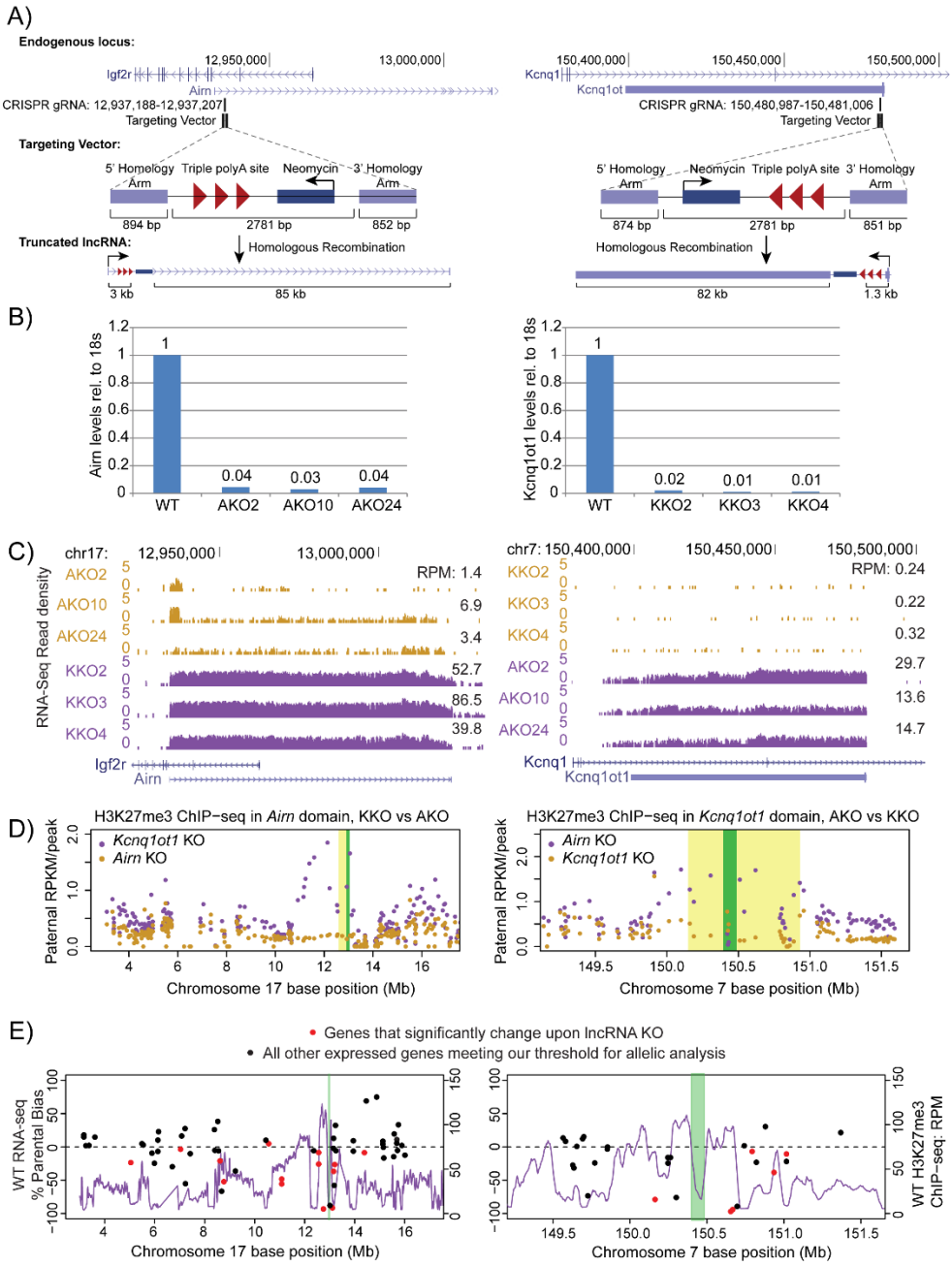
Supplemental Table 2.6. All genomic datasets used. Table is divided into 2 sections: “Genomic datasets generated in this study” and “Publicly available genomic datasets”. Under each section, “File ID” gives the name of the dataset, “Sequencing date” gives the date the samples were loaded onto the sequencer, “Data type” gives the type of experiment (RNA-seq, ChIP-seq, Hi-C, RNA-IP), “Cell type” gives the cell type and strain information when relevant, “Spike ins” says whether ERCC or *Drosophila* spike-ins were included, “Read length” gives information about 75bp versus 150 bp read length and single versus paired end sequencing, “Figures and tables” lists the figures and tables in the manuscript where each dataset was used. “Used to call peaks” has a “Yes” if the data was used to call MACS or hiddenDomain peaks, and “GEO” gives the GEO database reference for the data.

Supplemental Table 2.7. Oligonucleotides used. Table gives all oligonucleotide sequences used in the paper. “Oligo description” gives a descriptive name for the oligo, “Type” says whether the oligo was used for cloning, as for sgRNA annealing, qPCR, or genotyping, “Sequence” gives the oligo sequence, and “Location in the paper” includes either the title of the methods section or the specific figure where the oligo was used.



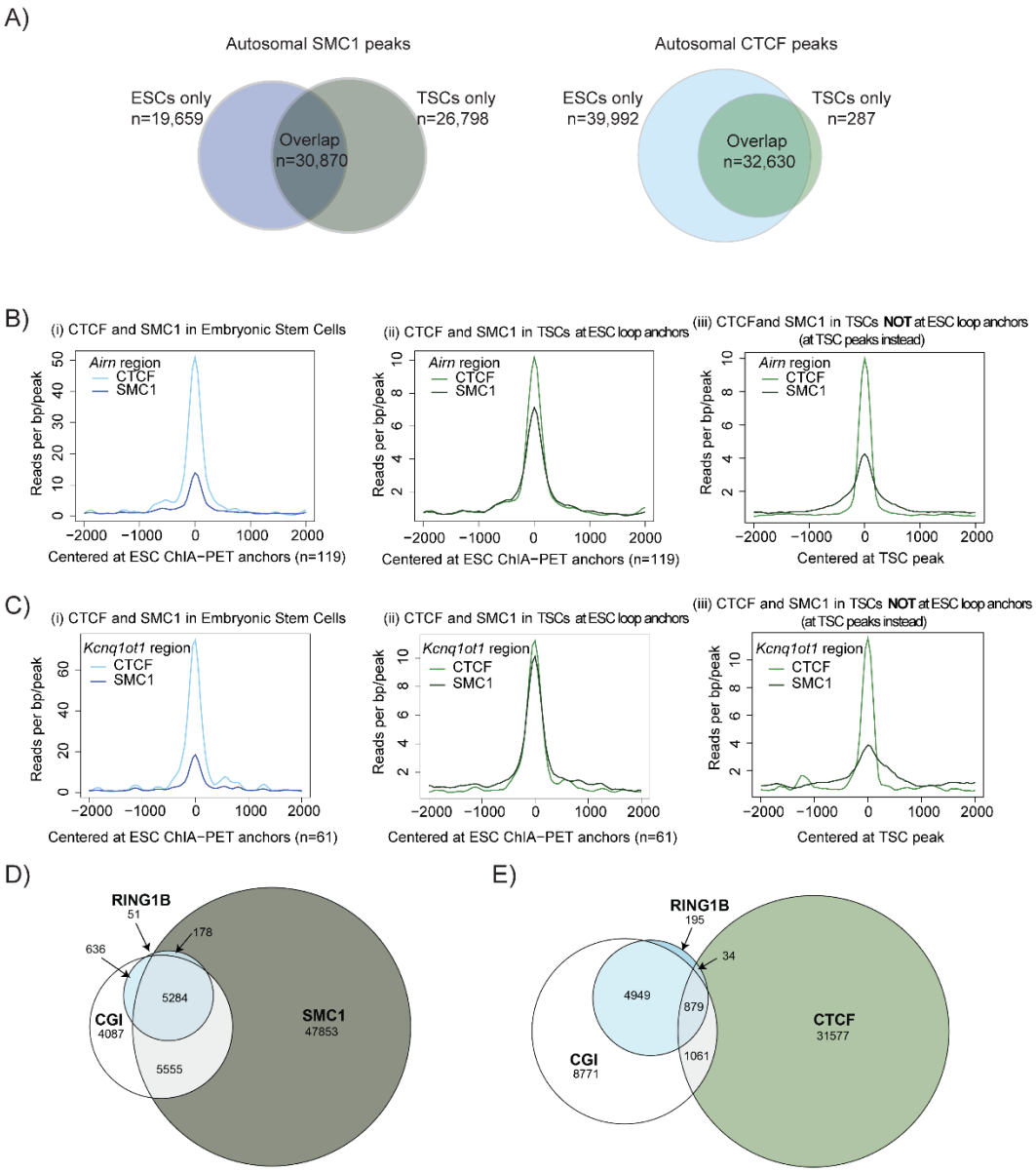
Supplemental Figure 2. 1. H3K27me3 biased peaks and correlations between H3K27me3 and H2AK119ub.

(A) Reciprocal F1-hybrid TSCs distinguish parent-of-origin bias from strain bias in high throughput sequencing experiments. (B) Correlation between H3K27me3 density and H2AK119ub density in 10kb bins genome wide in TSCs. r value, Pearson's correlation. (C) Top pie chart showing the percent of allelically biased autosomal H3K27me3 peaks in TSCs. The four lower pie charts show number of H3K27me3 peaks per autosome that have a maternal or paternal parent-of-origin bias or a B6 or Cast strain bias. (D, E) Parent-of-origin bias in H2AK119ub in C/B TSCs within H3K27me3 peaks in the *Airn* and *Kcnq1ot1* target domains. Green bars, lncRNA loci. Y-axis is the same as in Figure 1E, F. (F) Boxplots comparing paternal H3K27me3 density within peaks in *Xist*, *Airn*, and *Kcnq1ot1* targeted regions. ***, p<0.001; Tukey's HSD test.



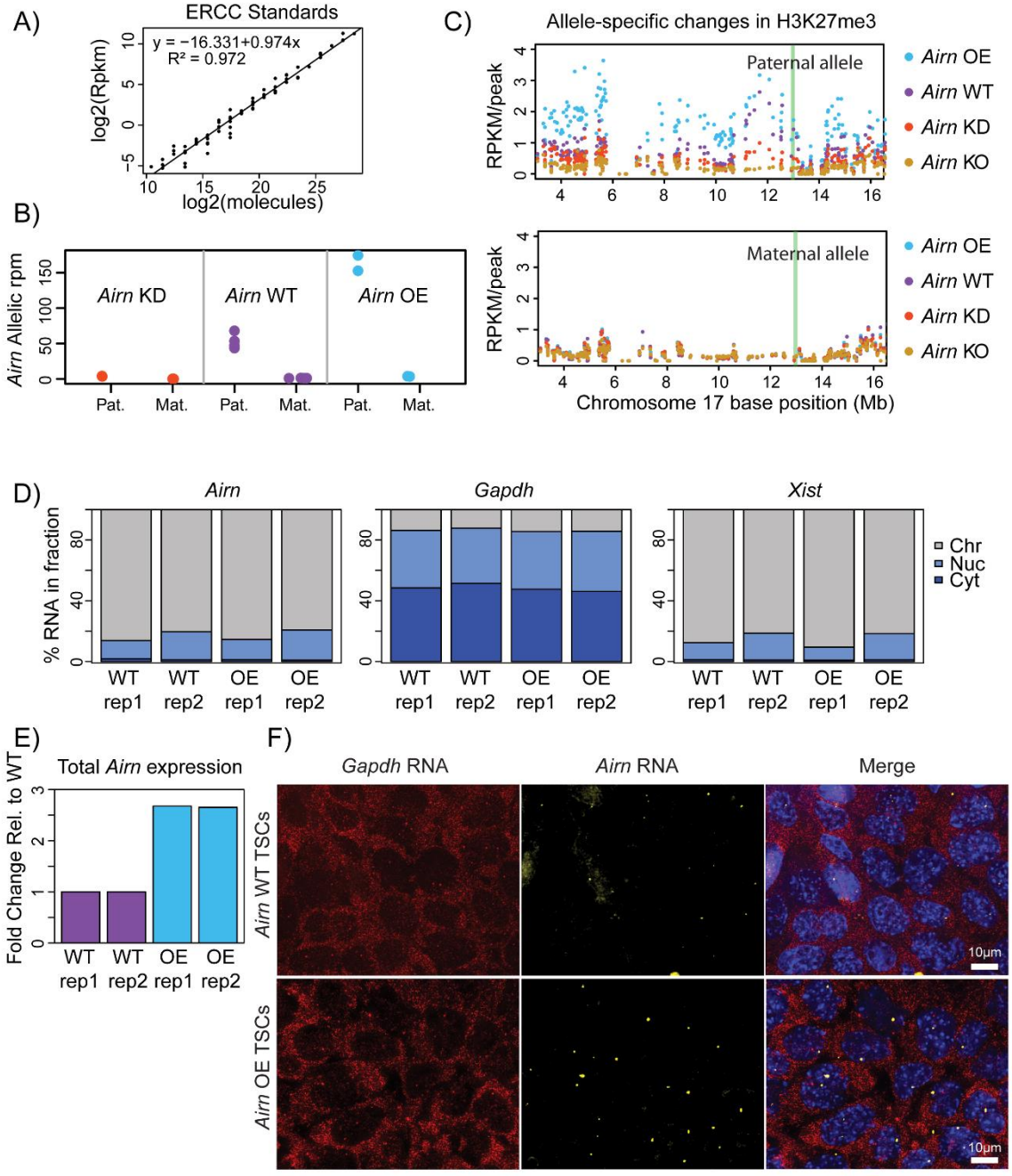
Supplemental Figure 2. 2. *Airn* and Kcnq1ot1 truncation and characterization.
(A) CRISPR-targeting strategy for *Airn* and *Kcnq1ot1*. (B) qPCR data showing >95% reduction in *Airn* or *Kcnq1ot1* expression in lncRNA truncation clones. (C) RNA-Seq data verifying successful lncRNA truncation. For each lncRNA, all three truncation clones were used for RNA-Seq, whereas only AKO2 + AKO10 and KKO2 + KKO3 were used for ChIP-Seq. (D) Allele-specific data demonstrating regional specificity of loss of H3K27me3 upon lncRNA truncation; *Airn* truncation (AKO) causes domain-wide loss of H3K27me3 in the *Airn* domain but not in the *Kcnq1ot1* domain, and *Kcnq1ot1* truncation (KKO) causes domain-wide loss of H3K27me3 in the *Kcnq1ot1* domain but not in the *Airn* domain. Each dot represents the paternal H3K27me3

ChIP-Seq reads pooled from two independently-derived, clonal AKO and KKO TSC lines (AKO2 and AKO10, and KKO2 and KKO3). H3K27me3 peak locations were defined by hiddenDomains using data from wild-type TSCs and are the same as those displayed in Figure 1. (E) Location and WT parental bias of the 61 and 27 expressed genes that met our threshold for allelic analysis in the *Airn* and *Kcnq1ot1* targeted regions, respectively, relative to WT H3K27me3 density. Red dots mark genes that significantly change upon lncRNA knockout and correspond to the individual genes plotted in Figures 1E and F. Black dots mark non-impacted genes. Green bars, lncRNA loci. Left y-axis is for gene expression bias and is the same as in Figure 1E, F and the right y-axis is for the ChIP-seq data.



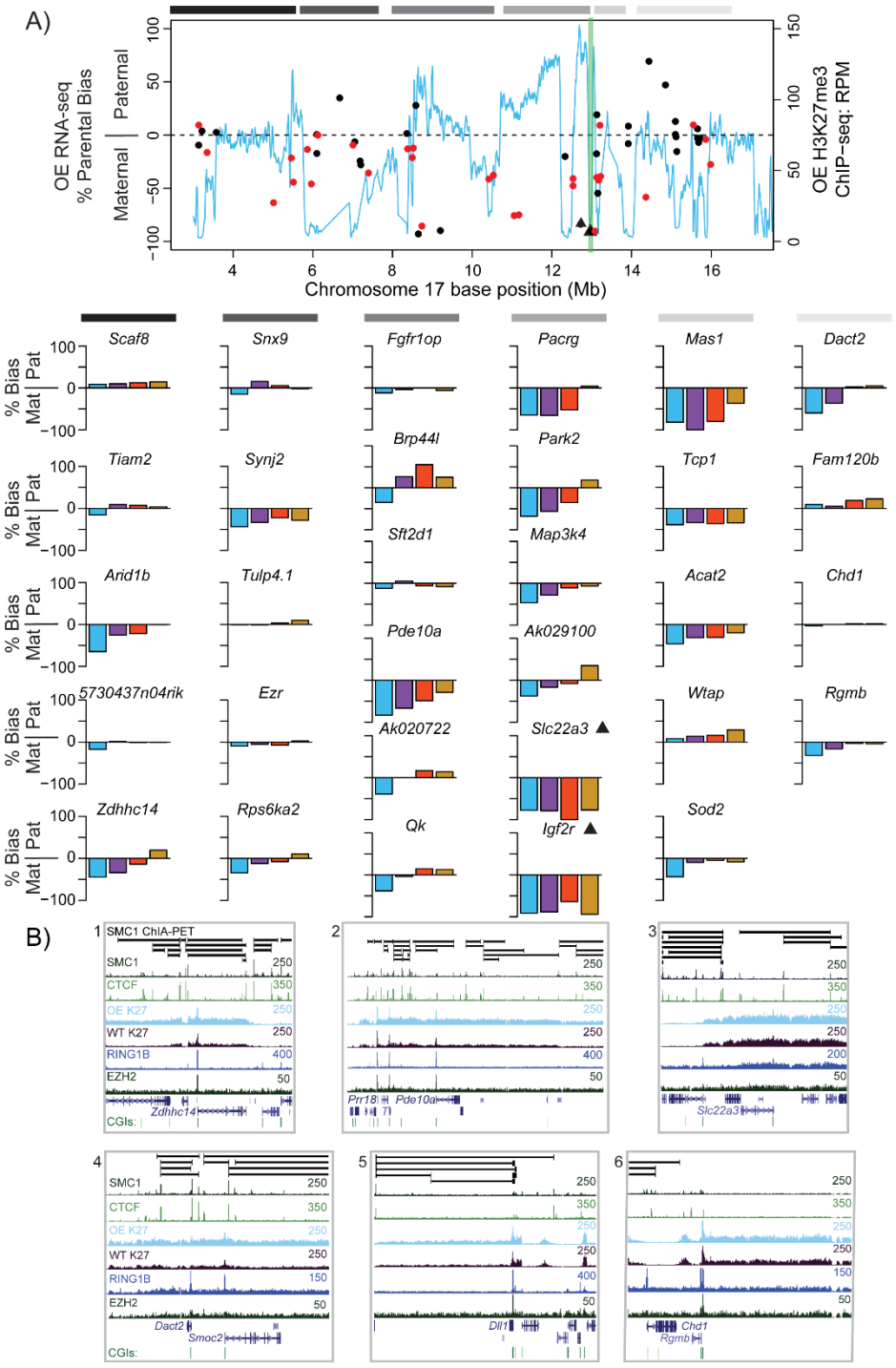
Supplemental Figure 2. 3. Comparison of CTCF and SMC1 data in ESCs and TSCs, and overlap with CGIs and RING1B.

(A) Venn diagrams showing overlap of CTCF and SMC1 ChIP-Seq peaks in ESCs (Kagey et al., 2010; Stadler et al., 2011) and in C/B and B/C TSCs. (B, C) CTCF and SMC1 ChIP-Seq signal in the (B) *Airn* and (C) *Kcnq1ot1* domains, from left to right: (i) using ESC data centered at ESC ChIA-PET anchors, (ii) using TSC data centered at ESC ChIA-PET anchors, and (iii) using TSC data centered at peaks of CTCF and SMC1 within the two domains that do not coincide with ESC ChIA-PET anchors. The similarity in signal intensity between (ii) and (iii) implies that a number of DNA loops that e*Xist* in the *Airn* and *Kcnq1ot1* domains in ESCs also e*Xist* in TSCs. (D, E) Venn diagrams showing overlap of SMC1 (D) or CTCF (E) with CGIs and RING1B peaks in TSCs. Peak numbers in each category are given on the plot.



Supplemental Figure 2. 4. Characterization of *Airn*-overexpressing and knockdown TSCs. (A) Representative relationship between ERCC spike-in control RNA-Seq read counts and copy number in TSCs. Data shown are from a single replicate of non-targeting gRNA control TSCs. ERCC spike-in controls, a series of commercially-available, synthetic polyadenylated RNAs whose individual abundance in solution spans five-orders of magnitude, were added to RNA from each sample just before initiating the protocol for RNA-Seq library preparation. Read counts were converted to molecules from the ERCC standard curve, then molecules-per-cell were calculated considering that the average TSC carries 30 picograms of RNA. (B) Allele-specific signal in knockdown (KD), wild-type (WT), and overexpressing (OE) TSCs shows *Airn*

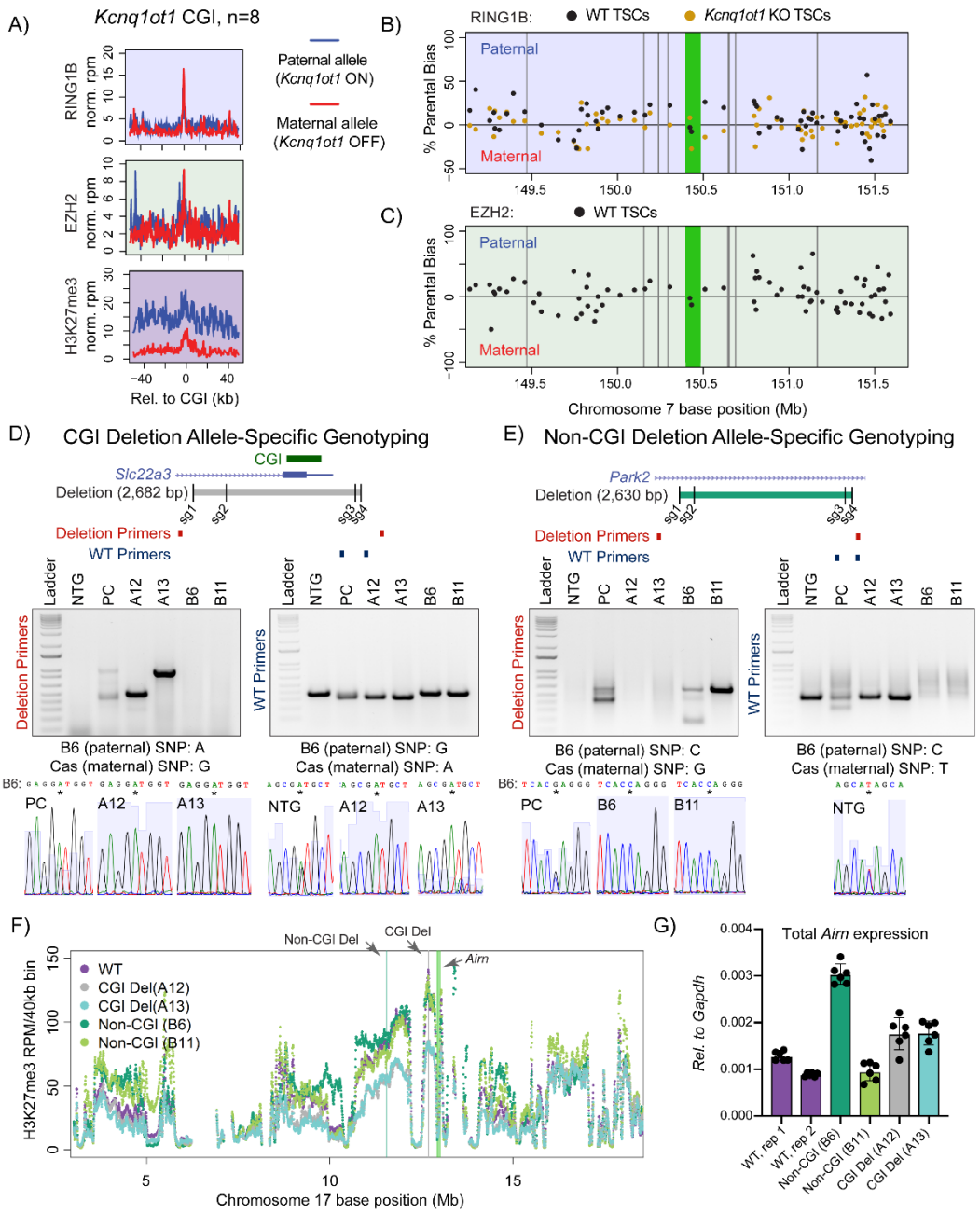
is specifically upregulated on the paternal allele. (C) Allele-specific H3K27me3 signal in *Airn* domain shows H3K27me3 is specifically increased on the paternal allele upon over-expression of *Airn*. Green vertical bar, *Airn* locus. (D) qPCR of *Airn*, Gapdh, and *Xist* RNA separated into cytoplasmic, free nuclear, and chromatin-bound fractions from two biological replicate preparations of RNA from *Airn*-WT and *Airn* OE TSCs (rep1 and rep2). Technical triplicates of qPCR were performed for each replicate and the average of those triplicates is plotted. (E) Change in total *Airn* RNA levels between WT and OE cells. Data are from two biological replicate preparations of RNA (rep1 and rep2) taken from samples in (D) before fractionation. Technical triplicates of qPCR were performed for each replicate and the average of those triplicates is plotted. (F) Representative single-molecule RNA FISH images for *Airn* (red) and Gapdh (yellow) RNAs in *Airn* WT and *Airn* OE. The increased dot size along with data from (D) indicate that the majority of *Airn* RNA remains chromatin localized upon *Airn* overexpression. Scale bars, 10μm.



Supplemental Figure 2. 5. Quantitation of gene expression and chromatin changes induced by *Airn* overexpression, repression, and knockout.

(A) Location and *Airn* OE parental bias of the 61 genes meeting our threshold for allelic analysis in the *Airn* target region, relative to OE H3K27me3 density. Red dots (n=29) mark genes that significantly change between *Airn* OE and *Airn* KO TSCs as assessed by a two-tailed t-test ($p < 0.05$). Individual boxplots for each of these genes is shown below the main plot, showing

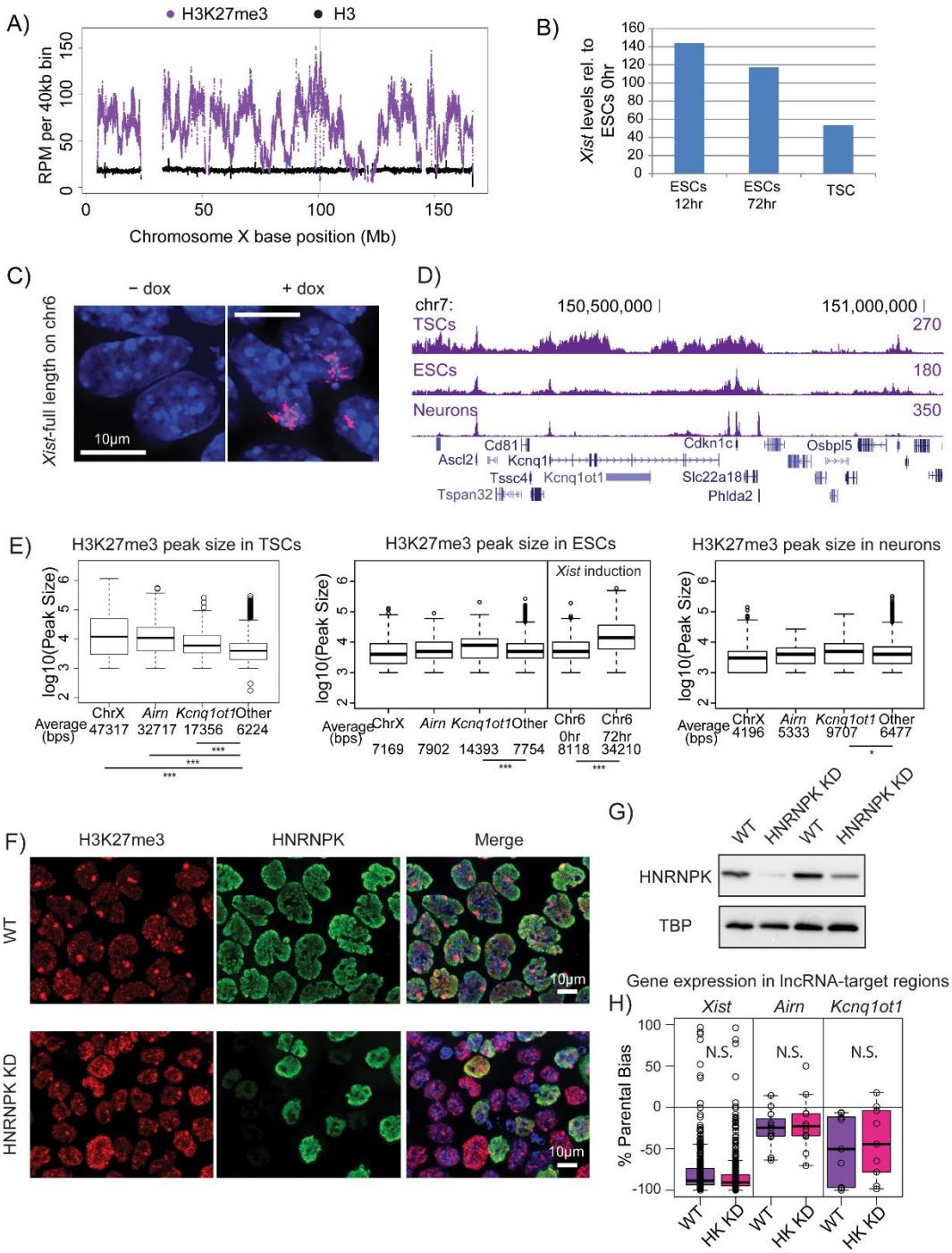
average parental bias in *Airn* OE, WT, KD, and KO TSCs. Gene boxplots are in order, columns first then rows, based on genomic location. Grey bars above the main plot correspond to grey bars above boxplot columns. Black triangles mark the *Slc22a3* and *Igf2r* genes, which are known targets of *Airn*, but did not significantly change in allelic expression upon *Airn* truncation (*Slc22a3* is barely expressed in TSCs, and the polyA tail from the G418 resistance expression cassette in the *Airn* truncation construct would silence the *Igf2r* gene on the paternal allele in *Airn* truncation TSCs regardless of its transcriptional status). Black dots mark non-impacted genes. Green bars, lncRNA loci. Left y-axis is for gene expression bias and is the same as in Figure 1E, F and the right y-axis is for the ChIP-seq data. (B) UCSC genome browser images depicting total SMC1, CTCF, OE H3K27me3, WT H3K27me3, RING1B, and EZH2 density, and ESC SMC1 ChIA-PET loop calls around the six proposed PRC nucleation regions in the *Airn* domain. We note that MACS did not call a peak of RING1B over the *Slc22a3* CGI. Nevertheless, visual inspection of total and allele-specific read density indicates enrichment of RING1B at levels above background over the *Slc22a3* CGI (see panel "3" and Table S3). In our interpretation, a peak was not called by MACS because of the above-background levels of RING1B in the broad regions flanking the *Slc22a3* CGI, which might have prevented MACS from detecting what otherwise appears to be a local enrichment.



Supplemental Figure 2. 6. RING1B and EZH2 bind CGIs prior to Kcnq1ot1 expression, and CGI deletion clone characterization.

(A) Metagene plots depicting RING1B, EZH2, and H3K27me3 read density relative to the center of all 8 CGIs that co-localize with RING1B peaks in the *Kcnq1ot1* domain. Island locations are shown as darkened lines in panels B and C. All 8 of these CGIs co-localize with peaks of SMC1 and two also overlap CTCF. (B, C) Parent-of-origin bias in RING1B and EZH2 in peaks of H3K27me3 in the *Kcnq1ot1* domain. RING1B data are from wild-type (WT) and *Kcnq1ot1* truncation (KO) TSCs, and EZH2 data are from wild-type TSCs. H3K27me3 peak locations are the same as in Figure 1. Green bar, *Kcnq1ot1* locus. Panels shaded in A-C for clarity. (D, E) Allele-specific genotyping of the *Slc22a3* CGI deletion clones (A12 and A13) and the non-CGI

deletion clones (B6 and B11). Upper diagrams show two the sets of genotyping primers relative to the location of the expected deletion. Each vertical black bar within deletion region marks the location of a sgRNA used to cut via CRISPR (sg1, sg2, sg3, sg4). Different combinations of sgRNA cuts give rise to PCR bands of different sizes. Inverse intensity images of ethidium bromide-stained agarose gels used for genotyping are shown below the diagrams. The sanger sequencing chromatograms that confirm allele-of-origin for the deletion clones are shown below agarose gels. * marks the locations of the informative SNPs in the PCR products. In both panels, "NTG" signifies DNA collected from non-targeting sgRNA control TSCs; these cells express doxycycline-inducible Cas9 but no functional sgRNA and therefore their genotype should be wild-type at the loci of interest. "PC" signifies DNA collected from polyclonal populations expressing either the *Slc22a3* or non-CGI sgRNA guides as well as Cas9; these cells serve as a form of positive control because deletion-product DNA arising from both B6 and Cast alleles should be present. In the left panel of (D), which shows the PCR to detect deletion of the *Slc22a3* CGI, deletion products of the expected size are detected in the PC control and in the two *Slc22a3* CGI deletion clones A12 and A13 but are not detected in the non-targeting sgRNA control ("NTG") nor in the two non-CGI deletion clones B6 and B11. Sanger sequencing of the deletion PCR products from the PC control confirms the ability to detect DNA from both alleles, and sequencing from the deletion clones A12 and A13 confirms deletion on the paternal allele. In the right panel of (D), PCR to detect wild-type DNA at the *Slc22a3* CGI detects signal in all lanes, consistent with the two deletion clones A12 and A13 being heterozygotes. Sanger sequencing of the wild-type PCR products from the NTG control confirms the ability to detect DNA from both alleles, and sequencing from the deletion clones A12 and A13 confirms that the wild-type DNA signal originates from the maternal allele, again consistent with A12 and A13 harboring paternal deletion of the *Slc22a3* CGI. In the left panel of (E), which shows the PCR to detect the non-CGI deletion, deletion products of the expected size are detected in the PC control and in the two non-CGI deletion clones B6 and B11 but not in the non-targeting sgRNA control ("NTG") nor in the two *Slc22a3* CGI deletion clones A12 and A13. Sanger sequencing of the deletion PCR products from the PC control confirms the ability to detect DNA from both alleles, and sequencing from the deletion clones B6 and B11 confirms the presence of deletion on the paternal allele. In the right panel of (E), PCR to detect wild-type DNA detects signal in all lanes save those from the deletion clones B6 and B11, suggesting that B6 and B11 are homozygous deletions that harbor a deletion of the expected size on the paternal allele and a deletion of the unexpected size on the maternal allele. (F) Tiling plot of H3K27me3 density in 40kb bins sliding across the *Airn* target region. H3K27me3 data are plotted separately for each CGI deletion and non-CGI deletion clone. Vertical bars mark the location of *Airn*, CGI deletion, and non-CGI deletion. (G) qPCR showing *Airn* expression in all four deletion clones and replicate RNA preparations from wild-type TSCs. Dots show individual qPCR technical replicates from separate RT and qPCR reactions.



Supplemental Figure 2. 7. Xist expression from chr6 in ESCs, H3K27me3 peak sizes in TSCs, ESCs, and neurons, and HNRNPK knockdown in TSCs.

(A) Tiling density of H3K27me3 and H3 on the X chromosome in TSCs. Axes are analogous to Figure 4E and 6D. (B) qPCR data showing levels of *Xist* 12hrs and 72hrs after doxycycline induction from chr6 in ESCs. *Xist* expression in TSCs is given for reference. Y-axis is relative to 0hr dox treatment (no *Xist* expression) in ESCs. (C) RNA FISH shows *Xist* cloud in ESCs upon addition of doxycycline. Scale bar, 10μm. (D) H3K27me3 read density in UCSC wiggle format in the *Kcnq1ot1* domain in TSCs, ESCs, and in cortical neurons. *Kcnq1ot1* is expressed in all 3

cell types whereas *Airn* is not highly expressed in ESCs and is not expressed in cortical neurons. Of the cell types examined, the distribution of H3K27me3 around *Kcnq1ot1* target genes is most even in TSCs (E) Size of H3K27me3 peaks defined by hiddenDomains in TSCs, ESCs, and in cortical neurons. In ESCs, 0hr and 72hr box plots show H3K27me3 peak sizes before and after *Xist* induction on chr6. ***, $p<0.001$; *, $p<0.05$; Tukey's HSD test. (F) Representative Immunofluorescence (IF) images for H3K27me3 and HNRNPK in WT TSCs and TSCs after 4 days of HNRNPK knockdown. Polyclonal cell population shows cells that have lost H3K27me3 enrichment on the X upon HNRNPK knockdown and some that maintain the enrichment. Scale bars, 10µm. (G) Western blot showing level of HNRNPK knockdown in two biological replicates. WT here refers to cells that were electroporated with non-targeting sgRNA control cassette and selected alongside of HNRNPK knockdown cells. (H) Boxplot showing parental bias of expressed genes in *Xist*, *Airn*, and *Kcnq1ot1* targeted regions in WT and HNRNPK knockdown TSCs. N.S., not significant; Tukey's HSD test.

CHAPTER 3: A piggyBac-based toolkit for inducible genome editing in mammalian cells[2]

3.1 Introduction

Within the last decade, the CRISPR (clustered regularly interspaced short palindromic repeat) bacterial immune system has provided researchers with multiple new methods to control gene expression in mammalian genomes. Co-expression of the Cas9 (CRISPR-associated protein 9) nuclease from *Streoptococcus pyogenes* along with an engineered single guide RNA (sgRNA) that targets a protein-coding exon is an effective way to introduce frameshift mutations in proteins of interest, owing to the fact that repair of the DNA break introduced by Cas9 often results in small deletions surrounding the cut site. Co-expression of Cas9 and multiple sgRNAs can also be used to excise larger regions from genes of interest, or to excise DNA regulatory elements (Aparicio-Prat et al., 2015; Canver et al., 2014; Gasperini et al., 2017; Ran et al., 2013; Zhu et al., 2016). Expression of a catalytically-dead Cas9 (dCas9) fused to a transcriptional activation or repression domain can be used to up- or down-regulate gene expression when sgRNAs are targeted to promoters or regulatory elements of interest (Hsu et al., 2014; Wright et al., 2016).

Owing to the broad utility of CRISPR, multiple methods have been developed to deliver the CRISPR-Cas9 machinery to mammalian cells. Transient transfection of Cas9- and sgRNA-expressing plasmids, or of Cas9 protein and *in vitro* synthesized sgRNAs, are useful when the

[2] This chapter previously appeared as an article in the journal, *RNA*. The original citation is as follows: Schertzer MD, Thulson E, Braceros KCA, Lee DM, Hinkle ER, Murphy RM, Kim SO, Vitucci EM, Calabrese JM (2019). A piggyBac-based toolkit for inducible genome editing in mammalian cells. *RNA*.

efficiency of transfection for the cell type of interest is high, and when the desired endpoint can be reached via transient expression of Cas9 and the sgRNA. Lentiviral delivery of Cas9/sgRNA vectors is also possible, and provides distinct advantages when transfection efficiency is low, or when the desired endpoint requires stable expression and or integration of Cas9/sgRNAs into the genome, such as for studies performed *in vivo* or for genome-wide phenotypic screens (Hartenian and Doench, 2015; Joung et al., 2017). However, delivery of the CRISPR machinery via lentivirus requires additional hands-on time, expertise, safety precautions, and cost relative to delivery via transient transfection.

The piggyBac transposon is a broadly used tool that allows DNA cargos up to 100 kilobases in length to be inserted into "AATT" sequences that are preferentially located in euchromatic regions of mammalian genomes (Cadiñanos and Bradley, 2007; Ding et al., 2005b; Li et al., 2011; Wang et al., 2008; Wilson et al., 2007). Owing to its high efficiency of transposition, piggyBac has been used in a wide range of applications, including in the stable expression of multi-subunit protein complexes, in the generation of transgenic mice and induced pluripotent stem cells, and in the large-scale production of recombinant proteins (Ding et al., 2005b; Kahlig et al., 2010; Kaji et al., 2009; Li et al., 2013; Yusa et al., 2009). Most recently, piggyBac has begun to be employed for CRISPR-based applications; piggyBac vectors have been used to study CRISPR off-target effects (Wu et al., 2014), to engineer mutations in human induced pluripotent stem cells (Wang et al., 2017), and to perform multiplexed activation of protein-coding and noncoding genes (Li et al., 2017b).

Herein, we describe the creation and validation of a piggyBac-based system for inducible editing of mammalian genomes by CRISPR-Cas9. In the system, which we call "CRISPR-Bac", two separate piggyBac cargo vectors, one that expresses an inducible Cas9 or dCas9 variant, and

another that expresses an sgRNA and the reverse-tetracycline transactivator [rtTA; (Gossen et al., 1995)], are transfected into cells along with a plasmid that expresses the piggyBac transposase. A short period of selection is used to obtain cells that stably express both the Cas9 and sgRNA cargo vectors. Our CRISPR-Bac vectors provide a simple way to rapidly insert the CRISPR-Cas9 machinery into mammalian genomes to knockdown proteins, delete kilobase-sized genomic regions, and activate or repress transcription of protein coding genes and long noncoding RNAs (lncRNAs), without the additional cost and labor involved in the packaging and delivery of lentiviral particles to cells.

3.2 Results

3.2.1 Cloning of CRISPR-Bac

We modeled CRISPR-Bac (Figure 1) after the pX330 plasmid system in which a humanized version of the Cas9 enzyme from Streoptococcus pyogenes is co-expressed with a chimeric sgRNA driven by a U6 promoter (Cong et al., 2013). We cloned the Cas9 from pX330 into a doxycycline-inducible expression cassette in a piggyBac cargo vector that also expresses a gene conferring resistance to Hygromycin B. We then converted the dual BbsI sites in pX330, which are used to clone the sgRNA targeting sequence into that vector, into dual BsmbI sites. Like BbsI, BsmbI is a Type IIS restriction enzyme, and it generates overhanging ends that are identical to those generated by BbsI. We cloned the BsmbI-modified sgRNA expression cassette into a piggyBac cargo from (Kirk et al., 2018) that expresses a bi-cistronic message which encodes the rtTA3 gene and a gene conferring resistance to G418 (originally cloned from Addgene plasmid #25735; (Shin et al., 2006)). The conversion of the pX330 BbsI sites, which

are not unique in the rtTA-expressing vector, to BsmbI sites, allows the exact sgRNA design and cloning protocol for pX330 (Cong et al., 2013) to be used to clone sgRNAs into CRISPR-Bac.

3.2.2 Knockdown of a protein-coding gene using CRISPR-Bac

We tested whether CRISPR-Bac could be used to knockdown a protein of interest in mouse embryonic stem cells (ESCs). We designed three sgRNAs targeting different exons of the Ezh2 gene (Figure S1, Table S1) and cloned them into our sgRNA-rtTA-expressing vector using the protocol outlined in (Cong et al., 2013). We then co-transfected our inducible Cas9-expressing piggyBac vector, a plasmid expressing the piggyBac transposase, and either each sgRNA-expressing vector separately, or a pool of all three sgRNAs into ESCs. As a control, we transfected an sgRNA-rtTA-expressing vector into which we did not clone a specific sgRNA-targeting sequence (our "no sgRNA" control). After selecting ESCs on Hygromycin B and G418 for 10 days, we removed the selection drugs and added 1µg/ml of doxycycline to the media for four days to induce the expression of Cas9. To assess the extent of EZH2 knockdown, we performed western blot and immunofluorescence (IF). Relative to the control ESCs, we observed greater than 60% reduction in EZH2 protein levels in the three lines expressing individual sgRNAs and more than 90% loss in the line expressing the pool of sgRNAs (Figure 2A). In repeat transfections of the sgRNA pool, we consistently observed greater than 90% loss of EZH2 protein levels (Figure 2B). IF to EZH2 confirmed our western blot analysis (Figure 2C). We compared the levels of EZH2 knockdown obtained via CRISPR-Bac to those obtained via transient transfection of the same sgRNA pool cloned into the widely-used pX330 vector (Cong et al., 2013). Four days after transfection of pX330, we harvested cells and performed western blot and IF. Relative to pX330 lacking a gene-targeting sgRNA ("No sgRNA" control), we

measured 25-35% reduction in EZH2 protein (Figure 2D, E). These results demonstrate that CRISPR-Bac can be used to inducibly knockdown a protein-coding gene of interest.

3.2.3 Targeted deletion of genetic elements using CRISPR-Bac

An important use of CRISPR-Cas9 is to create targeted deletions of regulatory elements (Aparicio-Prat et al., 2015; Canver et al., 2014; Gasperini et al., 2017; Hsu et al., 2014; Ran et al., 2013; Wright et al., 2016; Zhu et al., 2016). To test the utility of CRISPR-Bac in this application, we cloned into CRISPR-Bac pairs of sgRNAs that flank multiple different regulatory elements (RE1, RE2, RE3, RE4), to delete 2,331 bp, 2,480 bp, 1,222 bp, and 2,609 bp regions, respectively (Figure S1, Table S1). We created ESCs that stably express the different sgRNA pairs along with doxycycline-inducible Cas9. We induced expression of Cas9 for 4 days, collected genomic DNA, and performed quantitative PCR (qPCR) using amplicons within the deleted regions. By comparing qPCR results between the sgRNA-expressing ESCs and non-targeting sgRNA control ESCs, we approximated the extent that each targeted region was deleted in a polyclonal cell population. For 4 of 4 deletions, we observed more than 40% reduction in signal, indicating that close to half of the alleles in the cell population were deleted (Figure 3A; two-sided t-test).

To assess deletion efficiency in single cells, we isolated 36 individual colonies from cells transfected with sgRNAs to the RE3 element, and extracted their genomic DNA. To assess whether a deletion occurred on at least one allele, we performed PCR using primers that flanked the expected deletion. 21 of 36 clones (58%) showed a band within the expected size range for a deletion (310bp-426bps; Figure 3B). signifying that these clones were at least heterozygous for the deletion. To distinguish between clones that were heterozygous versus homozygous for the

deletion, we used a pair of primers that amplify inside the deletion. 12 out of the 21 clones (33% of the 36 clones) did not show a band, indicating that no wildtype allele was present and the cells were homozygous for the deletion (Figure 3C). Many homozygous clones showed weak wildtype bands, which we presume were due to genomic DNA from MEF feeder cells and not due to the presence of a wildtype allele in the clones. In support of this notion, we performed qPCR to detect the wildtype allele on 7 total clones: two clones classified as wildtype, two as heterozygous, and three as homozygous. Relative to heterozygous and wildtype clones, all three homozygous clones showed a 10 and 11 cycle difference, respectively, signifying that the clones that we genotyped as homozygous indeed lacked wild-type alleles (Figure 3D). These data demonstrate that CRISPR-Bac can be used to generate targeted genomic deletions with high efficiency.

3.2.4 Activation and repression of protein-coding gene transcription using CRISPR-Bac

In addition to creating targeted genomic deletions, the CRISPR-Cas9 system can be used to up- or down-regulate genes from their endogenous promoters, by targeting dCas9 fused to effector domains that recruit transcriptional co-activators or co-repressors. We cloned one such transcriptional activator fusion, dCas9-VP160 from (Cheng et al., 2013), and one such transcriptional repressor fusion, dCas9-KRAB from (Kearns et al., 2014), into the same piggyBac-based inducible expression vector that we used to express catalytically active Cas9 (Figure 1). We then tested our ability to upregulate *Ascl1*, a silent gene in ESCs, with dCas9-VP160, and we tested our ability to downregulate *Oct4*, an active gene in ESCs, with dCas9-KRAB (Figure 4A). Using CRISPR-Bac, we routinely observed 350-fold upregulation of *Ascl1* relative to non-targeting sgRNA control cells (Figure 4B). This level of activation was similar to

that obtained using transient transfection of dCas9-VP160 and SPgRNA vectors from ((Cheng et al., 2013; Perez-Pinera et al., 2013); Figure 4B). Using multiple sets of published and in-house-designed sgRNAs, the maximum level of *Oct4* down regulation we achieved was two- to three-fold (Figure 4C and data not shown). Relative to the 4 days used for protein knockdown and genomic deletions, we observed that for transcriptional modulation experiments, 2 days of doxycycline treatment was sufficient to detect effects maximal induced by dCas9-VP160 and –KRAB. These data show that CRISPR-Bac can be used to up- and down-regulate transcription of protein-coding genes of interest.

3.4.5 Activation and repression of lncRNA transcription using CRISPR-Bac

We next examined whether we could use CRISPR-Bac to activate and repress transcription of a lncRNA using dCas9-VP160 and dCas9-KRAB, respectively. We chose to target a lncRNA called *Airn* in two cell types: mouse ESCs, in which *Airn* is expressed at low levels, and mouse trophoblast stem cells (TSCs), in which *Airn* is more highly expressed and active ((Andergassen et al., 2017; Calabrese et al., 2015; Latos et al., 2009); Figure 5A). In ESCs, we were able to activate *Airn* ~15-fold above its levels in non-targeting sgRNA control cells (Figure 5B), but we were not able to repress *Airn*, likely due to its low endogenous expression (not shown; (Latos et al., 2009)). Compared to transient transfection of dCas9-VP160 and SPgRNA vectors from (Cheng et al., 2013; Perez-Pinera et al., 2013), we achieved a greater level of activation with CRISPR-Bac (Figure 5B). In TSCs, we were able to repress *Airn* to 10% of its normal expression and activate *Airn* 2.5-fold relative to non-targeting sgRNA control cells (Figure 5C). Therefore, CRISPR-Bac can be used to activate and repress transcription of lncRNAs.

Under normal physiological conditions, the *Airn* lncRNA is monoallelically expressed due to a process called genomic imprinting that leads to methylation of its promoter and gene silencing specifically on the maternally inherited allele (Lee and Bartolomei, 2013; Stöger et al., 1993). To assess whether activation of *Airn* via CRISPR-Bac led to mono- or bi-allelic activation of the lncRNA, we performed RNA Fluorescence In Situ Hybridization (FISH) in ESCs stably expressing dCas9-VP160 and either a non-targeting sgRNA or an *Airn*-targeting sgRNA. We performed a two-color RNA FISH experiment where one probe was complementary to the *Airn* lncRNA, and the other probe was complementary to the *Kcnq1ot1* lncRNA. *Kcnq1ot1,* like *Airn* is also imprinted and monoallelically expressed (Lee and Bartolomei, 2013). Unlike *Airn*, *Kcnq1ot1* it is robustly expressed in ESCs under normal conditions (Umlauf et al., 2004). *Kcnq1ot1* therefore served as a control to gauge the extent of *Airn* monoallelism upon activation by CRISPR-Bac. After taking z-stacks on a widefield microscope and deconvolving the resultant images, we used an automated pipeline to identify puncta whose RNA FISH signal surpassed a specified threshold. In two images taken of cells expressing the non-targeting sgRNA control, we counted zero puncta of *Airn* relative to 100 puncta of *Kcnq1ot1*, confirming prior data that show *Kcnq1ot1* is robustly expressed in ESCs while *Airn* is not (Latos et al., 2009; Umlauf et al., 2004). In contrast, in two images taken of cells expressing the *Airn*-targeting sgRNA, we counted 97 puncta of *Airn* relative to 130 puncta of *Kcnq1ot1* (Figure 5D). These data support the notion that CRISPR-Bac activates expression of *Airn* on the unmethylated paternal allele, and that the methylated maternal allele of *Airn* remains resistant to activation (Stöger et al., 1993).

3.4.6 sgRNA titration to achieve variable levels of lncRNA induction

The number of piggyBac cargos inserted into the genome can be controlled by altering the ratio of cargo vector to transposase plasmid (Cadiñanos and Bradley, 2007; Wang et al., 2008; Wilson et al., 2007). The CRISPR-Bac platform relies on simultaneous delivery of two cargo vectors: one vector expressing the sgRNA and rtTA/G418 resistance genes, and the other vector expressing the Cas9/dCas9 variant and hygromycin resistance genes (Figure 1). We sought to determine whether the extent of activation of a target gene of interest could be altered by altering the ratios of sgRNA, Cas9, and piggyBac transposase vectors in transfections. We tested a range of sgRNA-to-dCas9VP160-to-transposase ratios, using the *Airn* lncRNA as our target gene for activation (Table S3). We found modest but significant differences in the level of *Airn* activation when we transfected higher amounts of sgRNA and dCas9-VP160 plasmids relative to the piggyBac transposase plasmid (Figure 6A; see table of adjusted p-values from Tukey's HSD test), and these differences were accompanied by increased numbers of sgRNA and dCas9-VP160 cargo insertions per cell (Figure 6B). Thus, the extent of target gene activation using CRISPR-Bac can be partly controlled by changing the ratios of sgRNA/rtTA, Cas9, and transposase plasmids in transfections.

3.2.7 Simultaneous upregulation of two genes via CRISPR-Bac

By co-expression of multiple sgRNAs, CRISPR can be used to activate or repress multiple genes simultaneously (Cheng et al., 2013). To test if CRISPR-Bac is capable of multiplexed gene activation, we created ESCs expressing dCas9-VP160 and sgRNAs targeting the *Ascl1* and *Airn* promoters (same sgRNAs as in Figures 4B, 5B-C, and 6A). Relative to non-targeting sgRNA controls, qPCR demonstrated simultaneous 411-fold activation of *Ascl1* and

9.5-fold activation of *Airn* when sgRNAs for both targets were co-transfected (Figure 6C). This confirms that CRISPR-Bac can be used to target multiple genes in a single experiment.

3.2.8 CRISPR-Bac can be used in human cells

To determine if CRISPR-Bac system could be used in human cells, we tested Cas9-mediated knockdown and dCas9-VP160-mediated upregulation in SUM-159 cells, a commonly used cell line in breast cancer research (Grigoriadis et al., 2012). Similar to our experiments in ESCs, into SUM-159 cells we co-transfected the inducible Cas9-expressing piggyBac vector, a plasmid expressing the piggyBac transposase, and a pool of four sgRNAs targeting human EZH2 exons, then selected the cells with Hygromycin B and G418 for at least ten days, and induced Cas9 expression with 1μg/ml of doxycycline for 4 days. Via western blot, we detected greater than 70% reduction in EZH2 protein levels in both replicates, which was confirmed via IF (Figure 7A, B). In parallel, we co-transfected the dCas9-VP160, transposase, and a pool of two sgRNAs targeting the promoter of *IL1RN*, drug selected for at least ten days, and induced dCas9-VP160 expression for 2 days. *IL1RN* was activated ~54-fold relative to the no dCas9 control (Figure 7C). In these experiments, the Cas9/EZH2 sgRNA cells served as negative control in the dCasp-VP160/ *IL1RN* sgRNA experiment, and vice versa. These data show that the CRISPR-Bac system can be used in human cells.

3.3 Discussion

In Mus musculus-derived embryonic stem cells and trophoblast stem cells, we have shown that CRISPR-Bac can be used to knockdown proteins through frameshift/deletion, to delete kilobase-sized regulatory elements with high efficiency, and to up- and down-regulate the

transcription of protein-coding genes and an imprinted lncRNA. Levels of CRISPR-induced activation could partly be controlled through delivery of different ratios of CRISPR-Bac vectors. It seems likely that the use of different promoter elements within CRISPR-Bac (for example, a constitutive CMV promoter driving dCas9-VP160 instead of a TRE) might afford additional levels of control. It may also be possible to engineer CRISPR-Bac vectors that express multiple sgRNAs, as has been done elsewhere (Albers et al., 2015; Kabadi et al., 2014; Sakuma et al., 2015). Although in this work we only tested CRISPR-Bac in a limited number of cell types, it seems reasonable to presume that the CRISPR-Bac vectors or their modified derivatives would be functional in other mammalian cell types, given the broad activity of the piggyBac transposase (Cadiñanos and Bradley, 2007; Ding et al., 2005a; Kahlig et al., 2010; Kaji et al., 2009; Li et al., 2011, 2013; Wang et al., 2008; Wilson et al., 2007; Yusa et al., 2009). Indeed, CRISPR-Bac facilitated efficient knockdown and transcriptional upregulation in at least one human cell line, SUM159 (Grigoriadis et al., 2012). In our view, the main utility of CRISPR-Bac over other genome editing platforms is that CRISPR-Bac allows the generation of stable cell lines without the need to package CRISPR-Cas9 components into lentiviral delivery systems. It also preserves the sgRNA cloning strategy from the widely used pX330/335 systems, facilitating horizontal transfer of sgRNAs between the two platforms (Cong et al., 2013). Relative to prior studies that have used piggyBac to carry out CRISPR in mammalian cells (Li et al., 2017b; Wang et al., 2017; Wu et al., 2014), our study describes a single platform with interchangeable functionalities that has been optimized for protein knockdown, regulatory element deletion, and the up- and down-regulation of protein-coding and noncoding gene transcription.

3.4 Methods

3.4.1 Construction of CRISPR-Bac vectors

To create the doxycycline-inducible Cas9, dCas9-VP160, and dCas9-KRAB piggyBac vectors, a parent piggyBac vector was created in which a bGH-polyA signal and an EF1α promoter driving expression of a hygromycin resistance gene was ligated into the cumate-inducible piggyBac transposon vector from System Biosciences after its digestion with HpaI and SpeI, which cut just downstream of each chicken β-globin insulator sequence and removed all other internal components of the original vector. The TRE from pTRE-Tight (Clontech) was cloned upstream of the bGH-polyA site, and Cas9 from pX330 (Addgene plasmid # 42230; (Cong et al., 2013); gift from Feng Zhang), dCas9-VP160 from (Addgene plasmid # 48225; (Cheng et al., 2013); gift from Rudolf Jaenisch) and dCas9-KRAB from (Addgene plasmid # 50917; (Kearns et al., 2014); gift from Rene Maehr & Scot Wolfe) were each cloned behind the TRE by digestion with AgeI and SalI (NEB) followed by Gibson Assembly (NEB), to generate piggyBac cargo vectors capable of inducibly expressing Cas9, dCas9-VP160, and dCas9-KRAB, respectively, upon addition of doxycycline.

To create the rtTA-sgRNA expressing piggyBac vector, the dual BbsI sites in pX330 were converted to BsmbI sites using oligonucleotides, and the entire U6 expression cassette was cloned via Gibson assembly into the PacI site upstream of the rtTA3-IRES-Neo cassette in the rtTA-piggyBac-Cargo vector described in (Kirk et al., 2018). The rtTA3-IRES-Neo cassette was originally cloned from pSLIK-Neo and was a gift from Iain Fraser (Addgene plasmid # 25735). Oligonucleotides used for cloning are in Table S1.

We have submitted four plasmids to Addgene: (1) PB_rtTA_BsmBI, #126028, (2) PB_tre_Cas9, #126029, (3) PB_tre_dCas9_KRAB, #126030, and (4) PB_tre_dCas9_VP160, #126031.

3.4.2 sgRNA Design

Oligonucleotides used for sgRNA cloning are listed in Table S1 and their location relative to gene features are shown in Figure S1. Protein knockdown sgRNAs were designed using Desktop Genetics, and all other sgRNAs were designed using the CRISPOR program or taken from published sources ((Haeussler et al., 2016); Table S1).

3.4.3 Embryonic stem cell (ESC) culture

ESCs were grown on gelatin coated plates at 37°C in a humidified incubator at 5% CO2. Media was changed daily and consisted of DMEM high glucose plus sodium pyruvate, 0.1 mM non-essential AA, 100 u/mL penicillin-streptomycin, 2 mM L-glutamine, 0.1mM 2-mercaptoethanol, 15% ES-qualified FBS, and 1:500 LIF conditioned media produced from Lif—1Cα (COS) cells. ESCs were split at an approximate ratio of 1:6 every 48hr.

3.4.4 Trophoblast stem cell (TSC) culture

TSCs were cultured as in (Quinn et al., 2006). Briefly, TSCs were cultured at 37°C on pre-plated irradiated MEF feeder cells in TSC media [RPMI (Invitrogen), 20% Qualified FBS (Invitrogen), 100 u/mL penicillin-streptomycin, 1mM sodium pyruvate (Invitrogen), 100μM β-mercaptoethanol (Sigma), and 2mM L-glutamine] supplemented with Fgf4 (25ng/ml; Invitrogen) and Heparin (1μg/ml; Sigma) just before use. At passage, TSCs were trypsinized with 0.125%

Trypsin (Invitrogen) for 3 minutes at room temperature and gently dislodged from their plate with a sterile, cotton-plugged Pasteur pipette (Thermofisher). To deplete MEF feeder cells from TSCs prior to RNA isolation, TSCs were pre-plated for 40 minutes and cultured for three days in 70% MEF-conditioned TSC media supplemented with Fgf4 (25ng/ml; Invitrogen) and Heparin (1µg/ml; Sigma).

3.4.5 SUM-159 cell culture

SUM-159 cells were maintained in DMEM/F12 medium (Gibco, ThermoFisher Scientific) supplemented with 5% FBS, 5 µg/ml insulin, 1 µg/ml hydrocortisone, and antibiotic:antimycotic cocktail (Gemini Bio Products) as in (Zawistowski et al., 2017).

3.4.6 Stable transfections of CRISPR-Bac components

To generate stable CRISPR-Bac E14 embryonic stem cell lines, 5x10^5 cells were seeded in a single well of a 6-well plate, and the next day transfected with piggyBac cargo vectors and pUC19-piggyBac transposase from (Kirk et al., 2018), totaling 2.5 µg of plasmid DNA (see exact amounts in Table S3), using Lipofectamine 3000 (Invitrogen) according to manufacturer instructions. Cells were subsequently selected on Hygromycin [150µg/ml; Gibco] and G418 [200µg/ml; Gibco] for 7 to 12 days. Due to the efficiency of piggyBac cargo integration and the rapidity of Hygromycin selection, most observable death from drug selection occurred within ~3 days after addition of Hygromycin and G418 (i.e. cells with Hygromycin resistance were invariably resistant to G418).

To generate stable CRISPR-Bac trophoblast stem cell lines, 7.5x10^5 cells we co-electroporated using the Neon instrument (electroporation program: 1300V, 40ms, 1 pulse;

Invitrogen) with 5 µg of plasmid DNA at a 1:1:2 ratio of rtTA-sgRNA to dCas9 to transposase. Cells were selected on Hygromycin [150µg/ml; Gibco] and G418 [200µg/ml; Gibco] for 9 days.

To generate stable CRISPR-Bac SUM-159 cells, 5x10^5 cells were seeded per well of a 6-well plate and the next day transfected with 2.5 µg of plasmid DNA at a 1:1:2 ratio of rtTA-sgRNA to Cas9 to transposase using Lipofectamine 3000 (Invitrogen). Cells were subsequently selected on Hygromycin [250µg/ml; Gibco] and G418 [600µg/ml; Gibco] for at least 10 days.

3.4.7 Transient transfections

For transient transfections using pX330, 5x10^5 ESCs were seeded in a single well of a 6-well plate and transfected the next day using Lipofectamine 3000 with 2.5 µg of the pX330 empty vector ("no sgRNA") or a pool of pX330 vectors expressing the 3 sgRNAs to mouse EZH2. Cells were harvested 4 days after transfection. For transient transfections using VP160, 5x10^5 ESCs were seeded in a single well of a 6-well plate and transfected the next day using Lipofectamine 3000 with 1.25 µg of dCas9-VP160 (Cheng et al., 2013)and 1.25 µg of the SPgRNA empty vector ("no sgRNA"; (Perez-Pinera et al., 2013)) or SPgRNA containing sgRNAs targeting either *Airn* or *Ascl1*. Cells were harvested 2 days after transfection.

3.4.8 Protein isolation and western blotting

To isolate protein for western blotting, cells were washed with PBS, and then lysed with RIPA buffer (10 mM Tris-Cl (pH 7.5), 1 mM EDTA, 0.5 mM EGTA, 1% NP40, 0.1% sodium deoxycholate, 0.1% SDS, 140 mM NaCl) supplemented with 1 mM PMSF (Fisher scientific) and 1x protease inhibitor cocktail (Sigma) for 15 minutes at 4C, four days after induction with 1µg/ml doxycycline. Prior to western blotting, protein levels were quantified using the DC assay

from Biorad. For western blotting, primary and secondary antibody incubations were done for 1hr at room temperature. Antibodies used were EZH2 (Cell Signaling #5246, 1:1000 dilution), TBP (Abcam ab818, 1:2000 dilution), ERK2 (Santa Cruz; sc-1647; 1:500), donkey anti-mouse IgG-HRP secondary (Santa Cruz; sc-2314; 1:2500), and donkey anti-rabbit IgG-HRP secondary (Santa Cruz; sc-2313; 1:2500).

3.4.9 Genomic DNA isolation and qPCR

To isolate genomic DNA, 400ul of ESC lysis buffer (100 mM Tris-HCl, pH 8.1, 5mM EDTA, pH 8.0, 200mM NaCl, 0.2% SDS) supplemented with 80ul proteinase K (Denville) was used per 24-well well of ESCs, four days after induction with 1µg/ml doxycycline. Lysed ESCs were incubated at 55°C overnight, cells were boiled at 100°C for 1 hr to degrade RNA, and DNA was precipitated by addition of 2 volumes of 100% ethanol. DNA was pelleted and resuspended in 1x TE (10mM Tris-HCl 1mM EDTA pH 8.0) overnight at room temperature prior to qPCR. qPCR was performed using 100 ng of DNA per reaction and iTaq Universal SYBR Green Supermix (Biorad), with primers specified in Table S1. All related plots were generated using R version 3.4.1 (R Core Team, 2017).

3.4.10 qPCR for DNA copy number analysis

Genomic DNA was prepared as in Genomic DNA isolation and qPCR section above. qPCR signal (SsoFast, Biorad) from the genomic DNA was compared to signal from a molar standard amplified from increasing amounts of the corresponding dCas9-VP160 and rtTA plasmids. Primers used are listed in Table S1. All related plots were generated using R version 3.4.1 (R Core Team, 2017).

3.4.11 Generation of clonal ESCs with targeted genomic deletions and genotyping

After 4 days of dox induction, RE3 deletion E14 cells were cultured two days in the absence of dox to ensure that Cas9 was fully depleted. Then, 2,000 cells were plated on a 10cm plate with pre-plated irradiated MEF feeder cells. After 4 days, individual colonies were picked and plated on irMEFs. Clonal lines were passaged twice off of MEFs before genomic DNA was prepared as in Genomic DNA isolation and qPCR section above.

Genotyping PCR reactions were performed with gDNA using Apex Taq DNA Polymerase (Genesee Scientific). The first set of primers flanked the deletion and identified clonal lines with at least one allele deleted. The second set only amplified a wildtype allele, with both primers sitting inside the deletion. Many clones showed weak wildtype bands, likely due to MEF gDNA and not due to the presence of a wildtype allele in the ESC clone. Primers used are listed in Table S1.

3.4.12 RNA Isolation and qPCR

RNA was isolated using Trizol (Invitrogen). For RT-qPCR assays, 1-2µg of RNA was reverse transcribed using MultiScribe RT (Applied Biosystems), and qPCR was performed using iTaq Universal SYBR Green (Biorad) and primers specified in Table S1. All related plots were generated using R version 3.4.1 (R Core Team, 2017).

3.4.13 RNA FISH

Fosmid Wl1-2156F18 (*Airn*) and BAC RP23-101N20 (*Kcnq1ot1*) were ordered from the BACPAC resource center and fingerprinted with restriction digestion prior to use to verify inserted DNA. Fluorescent labeling was performed using BioPrime (Invitrogen). ESCs were

fixed on coverslips for 10 minutes in 4% paraformaldehyde/PBS, followed by a 10-minute permeabilization on ice in 0.5% TritonX-100 in PBS and 1:200 Ribonucleoside Vanadyl Complex (NEB). Coverslips were stored at -20C in 70% ethanol until use.

To initiate the RNA FISH protocol, coverslips were dehydrated by serial 3-minute incubations with 75%, 85%, 95%, and 100% ethanol, and air-dried for 5 minutes. RNA FISH probes were added and coverslips were placed cell-side down in a chamber humidified with 50% formamide/2xSSC overnight at 37°C. After overnight incubation, coverslips were washed 3x with 50% formamide/2xSSC at 42C and 3x with 1xSSC at 50C. Each wash was 5 minutes long. Coverslips were then rinsed 1x with PBS before a 2 minute incubation in DAPI stock diluted 1:1000 in water. Coverslips were rinsed twice more and affixed to glass slides using Vectashield (VectorLabs), then sealed with nail polish.

Four dimensional datasets were acquired by taking multi-channel Z-stacks on an Olympus BX61 widefield fluorescence microscope using a Plan-Aprochromat 60X/1.4 oil objective and a Hamamatsu ORCA R2 camera, controlled by Volocity 6.3 software. Excitation was provided by a mercury lamp and the following filters were used for the three fluorescent channels that were imaged: 377/25 ex, 447/30 em for DAPI (DAPI-5060B Semrock filter); 482/17 ex, 536/20 em for AlexaFluor488 (Semrock FITC-3540B filter); 562/20 ex, 642/20 em for Cy3 (Semrock TXRED-4040B filter). Pixel size was 0.108 μm, Z spacing was 0.2 μm, and images had 1344x1024 pixels. Between 46-49 Z-stacks were acquired for each image. Z-stacks were deconvolved using the iterative-constrained algorithm (Mediacy AutoQuantX3) with default algorithm settings. Sample settings for the deconvolution were: peak emissions for dyes (570 nm, 519 nm, 461 nm for Cy3, AlexaFluor 488 and DAPI respectively), widefield microscopy mode, NA = 1.4, RI of oil = 1.518, and RI of sample = 1.45. After deconvolution,

RNA FISH signals were located using the "Spots" function in Imaris software (version 8.3.1) and marked with equal sized spheres. To initially call spots on all images, spot detection values were set at 0.5µm for xy and 1.5µm for z, and background subtraction and auto quality settings were used. We manually optimized the quality/sensitivity setting to call *Kcnq1ot1* spots, and then used the same quality threshold to call *Airn* spots for the same image. Images are shown as maximum intensity projections made using ImageJ (Schindelin et al., 2012).

3.4.14 Immunofluorescence (IF)

Cells were fixed on coverslips the same as for RNA FISH (see above). To initiate the IF protocol, coverslips were washed twice in PBS and blocked for 30 minutes at room temperature in blocking solution (1x PBS with 0.2% Triton X-100, 1% goat serum, and 6 mg/mL IgG-free BSA). Then, coverslips were washed in 0.2% triton/1x PBS and incubated with EZH2 antibody (Cell Signaling #5246; 1:200 in block solution) for 1 hour at RT. Coverslips were washed 3x in 0.2% triton/1x PBS for 4 minutes each and incubated with secondary antibody (AlexaFluor 647 goat anti-rabbit, A-21245, 1:1000 in block solution for ESCs and AlexaFluor 488 goat anti-rabbit, A-11034, 1:1000 in block solution) for 30 minutes at RT. After incubation, coverslips were washed 3x in 0.2% triton/1x PBS for 4 minutes each and rinsed 1x with PBS before a 2 minute incubation in DAPI stock diluted to 5ng/ml in water. Coverslips were rinsed twice more and mounted to glass slides using Prolong Gold (Thermo Fisher Scientific P10144). Imaging and deconvolution was performed the same as described in the RNA FISH section with the below exceptions. The filters used for the two fluorescent channels that were imaged are 377/25 ex, 447/30 em for DAPI (DAPI-5060B Semrock filter), 482/17 ex, 536/20 em for AlexaFluor488, (Semrock FITC-3540B filter), and 628/20 ex, 692/20 em for AlexaFluor 647 (Semrock Cy5 4040A filter). Approximately 40 Z-stacks were acquired for each image. Sample settings for the

deconvolution included the following peak emissions for dyes: 670 nm, 519 nm, and 461 nm for AlexaFluor 647, AlexaFluor 488, and DAPI, respectively. Images are shown as maximum intensity projections made using ImageJ (Schindelin et al., 2012).

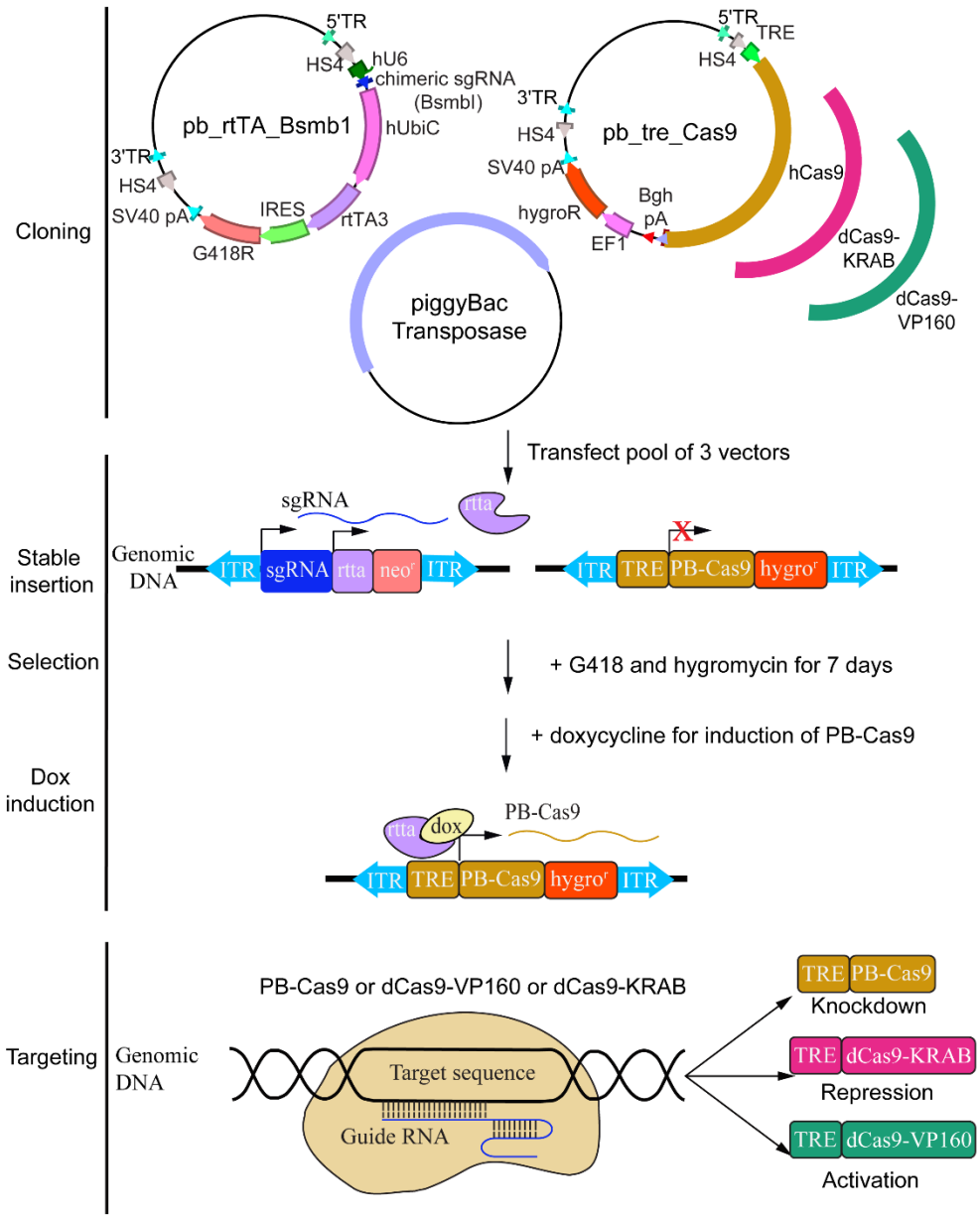


Figure 3. 1. Experimental pipeline used for CRISPR-Bac.
In a CRISPR-Bac experiment, a Cas9-expressing piggyBac cargo vector (or dCas9 variant) is co-transfected with a sgRNA- and rtTA-expressing piggyBac cargo vector and with a piggyBac transposase plasmid. Growth in Hygromycin B and G418 for 7 to 12 days selects for a population of cells that stably express an sgRNA of interest, and inducibly express a Cas9 or dCas9 variant. TR, piggyBac inverted terminal repeat. HS4, chicken β-globin insulator element. TRE, tetracycline responsive element (i.e. doxycycline inducible promoter). hCas9, dCas-VP160, dCas-KRAB from (Cheng et al., 2013; Cong et al., 2013; Kearns et al., 2014). EF1, EF1α promoter. HygroR, Hygromycin B resistance gene. SV40pA, SV40 polyadenylation signal. hU6-chimeric sgRNA from (Cong et al., 2013) with BsmbI sites replacing BbsI sites. hUbiC-rtTA3-IRES-G418 cassette was from pSLIK-Neo (Shin et al., 2006).

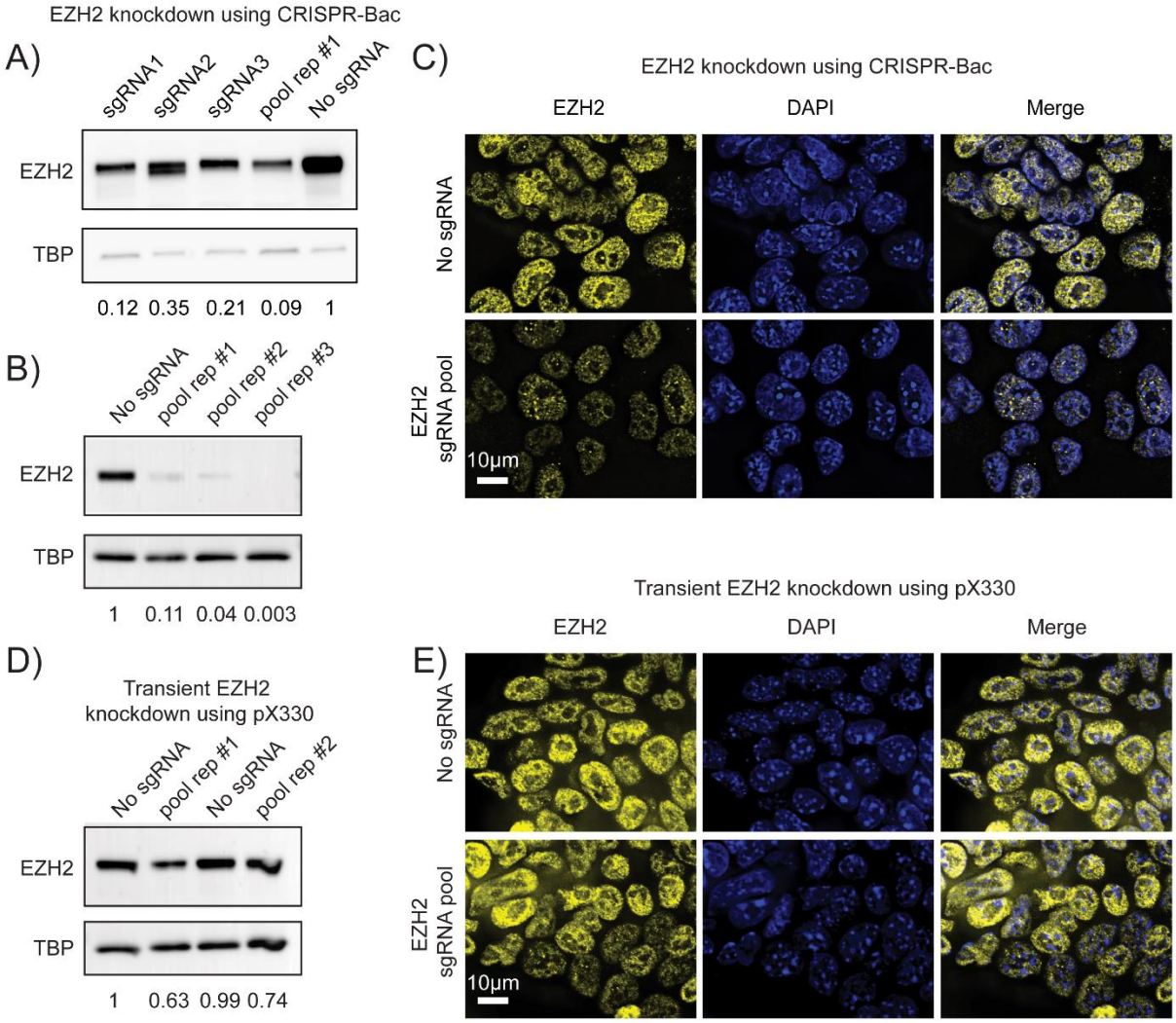


Figure 3. 2. Inducible protein-coding gene knockdown with CRISPR-Bac.
See Table S2 for details on replicates and experimental design for each figure panel. (A) Western blot of CRISPR-Bac assay in which three separate sgRNAs targeting Ezh2, a pool of all three sgRNAs, or a non-targeting sgRNA control ("No sgRNA") were co-transfected along with the inducible Cas9-expressing piggyBac cargo into E14 mouse ESCs. Western blots to EZH2 and TBP were performed on protein extracted from stably-selected cells, after four days of Cas9 induction with doxycycline. Values underneath blots represent knockdown of EZH2 relative to no sgRNA controls and normalized for loading with TBP protein levels. (B) Biological replicates of experiment in (A) for sgRNA pool. "Pool rep #1" is the same sample as in (A). (C) Representative immunofluorescence image showing EZH2 knockdown in non-targeting sgRNA control ("no sgRNA") or pooled sgRNA cells from replicate #2 in (B). Scale bar, 10μm. (D, E) Western blot and IF to EZH2 in transient transfection experiments with pX330.

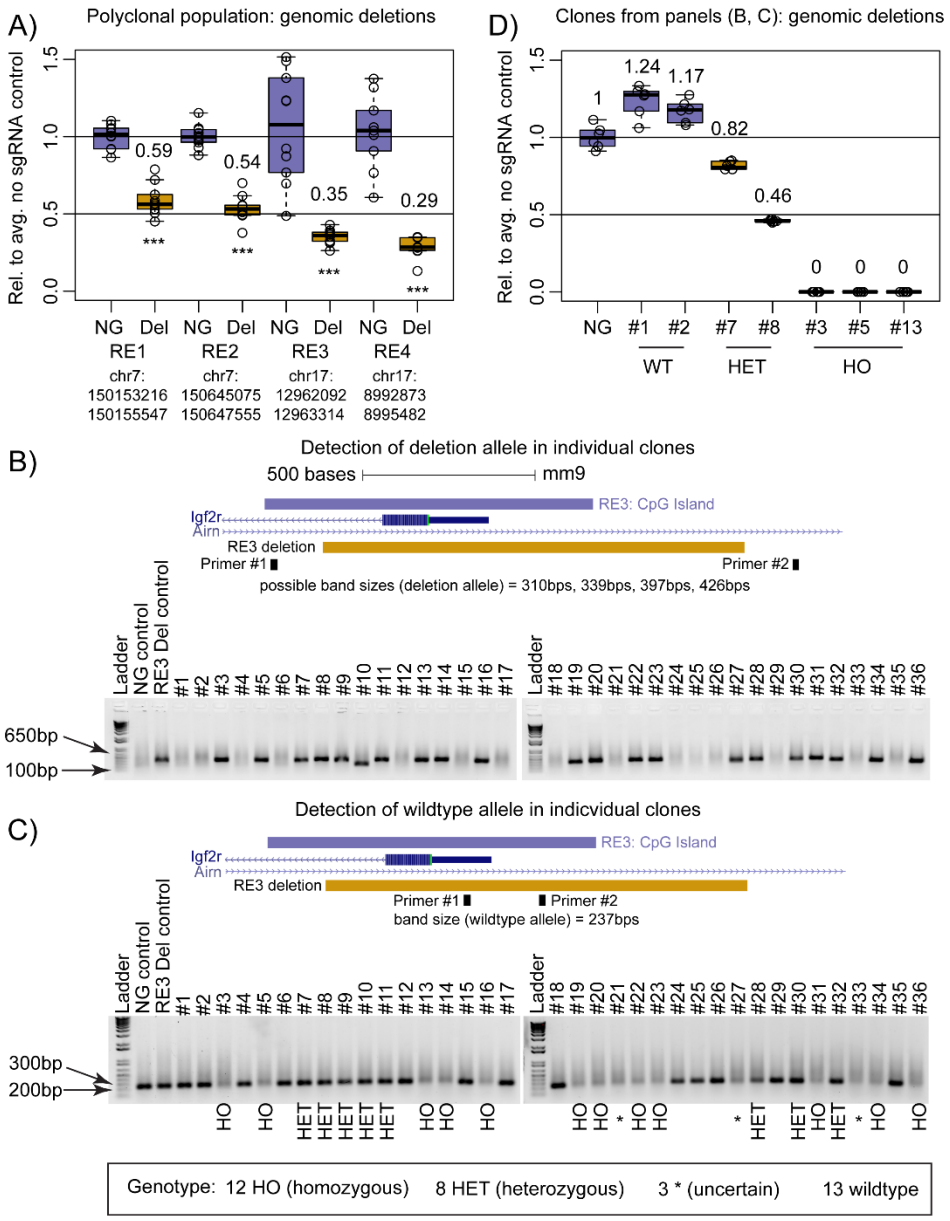


Figure 3. 3. Targeted deletion of DNA regulatory elements using CRISPR-Bac.
See Table S2 for details on replicates and experimental design for each figure panel. (A) qPCR results from polyclonal populations of ESCs expressing Cas9 and pairs of sgRNAs flanking four separate regulatory elements (RE). Primers for qPCR sit entirely inside of the expected deletion. Individual qPCR data points are shown in box-and-whisker format representing the mean and the interquartile range. Data from the non-targeting sgRNA control (NG) and sgRNA-expressing cells (Del) are plotted relative to the average of the signal in the NG control cells. ***, $p < 0.001$ from a two-sided t-test between NG and Del. (B, C) Agarose gels showing genotyping PCR products for the NG control and RE3 Del polyclonal populations from (A) and 36 clones isolated from RE3 deletion ESCs. The UCSC browser tracks above each gel show the location of RE3, the location of the expected deletion, and the location of primers used in the corresponding genotyping PCR. Scale bar, 500 bps. Gel in (B) identifies clones that have a deleted allele (four

possible band sizes based on the combination of sgRNAs that cut). Gel in (C) identifies clones that have a wildtype allele (primer pairs are the same as used for RE3 in (A)). Many clones showed weak wildtype bands, which we presume is due to MEF genomic DNA and not due to the presence of a wildtype allele. (D) qPCR results from RE3 clones in panels (B) and (C) classified as wildtype (#1 and #2), heterozygous (#7 and #8), and homozygous (#3, #5, and #13). Primers pairs are the same as in (A) and (C) to detect wildtype alleles. Data are plotted as in (A).

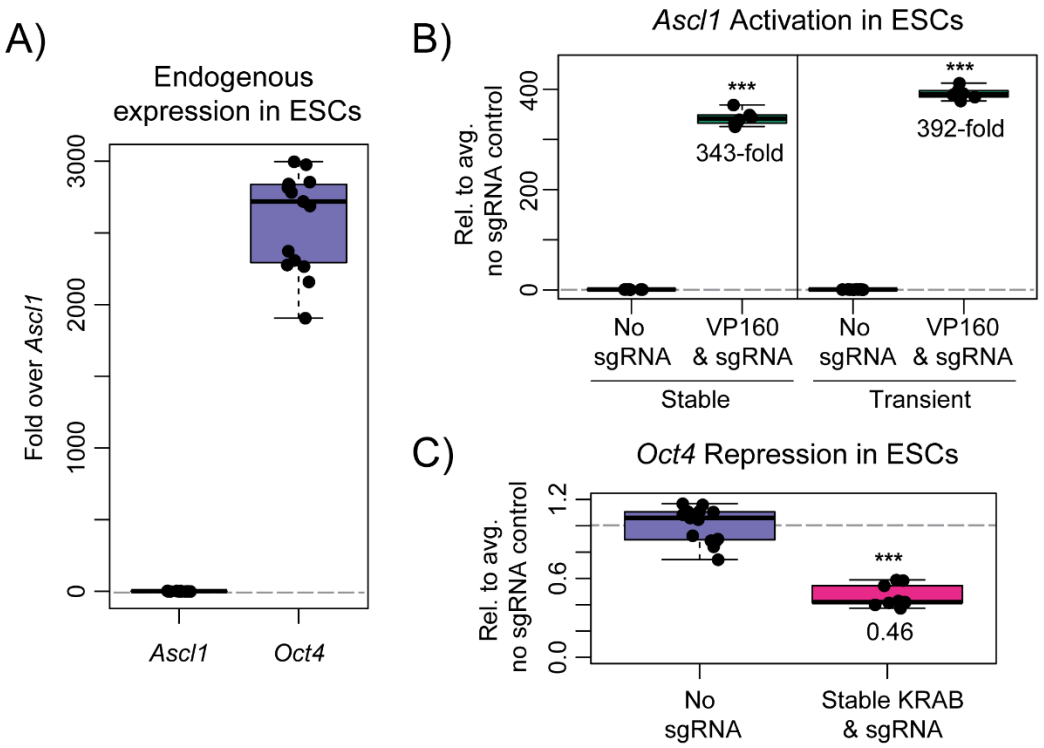


Figure 3. 4. Activation and repression of protein-coding gene transcription using CRISPR-Bac.

See Table S2 for details on replicates and experimental design for each figure panel. Individual qPCR data points are shown in box-and-whisker format representing the mean and the interquartile range. ***, $p < 0.001$ from a two-sided t-test between no sgRNA and sgRNA-expressing cells. (A) qPCR results showing endogenous expression of *Ascl1* and *Oct4* in ESCs relative to the average *Ascl1* signal. (B) qPCR showing transcriptional activation of *Ascl1* using CRISPR-Bac stable dCas9-VP160 and a pool of four *Ascl1*-targeting sgRNAs (Perez-Pinera et al., 2013) versus transient transfection of dCas9-VP160 and pooled SPgRNA-*Ascl1*. Data from the non-targeting sgRNA control (No sgRNA) and sgRNA-expressing cells are plotted relative to the average of the signal in the No sgRNA control cells. (C) qPCR showing transcriptional repression of *Oct4* using dCas9-KRAB. Data are plotted as in (B).

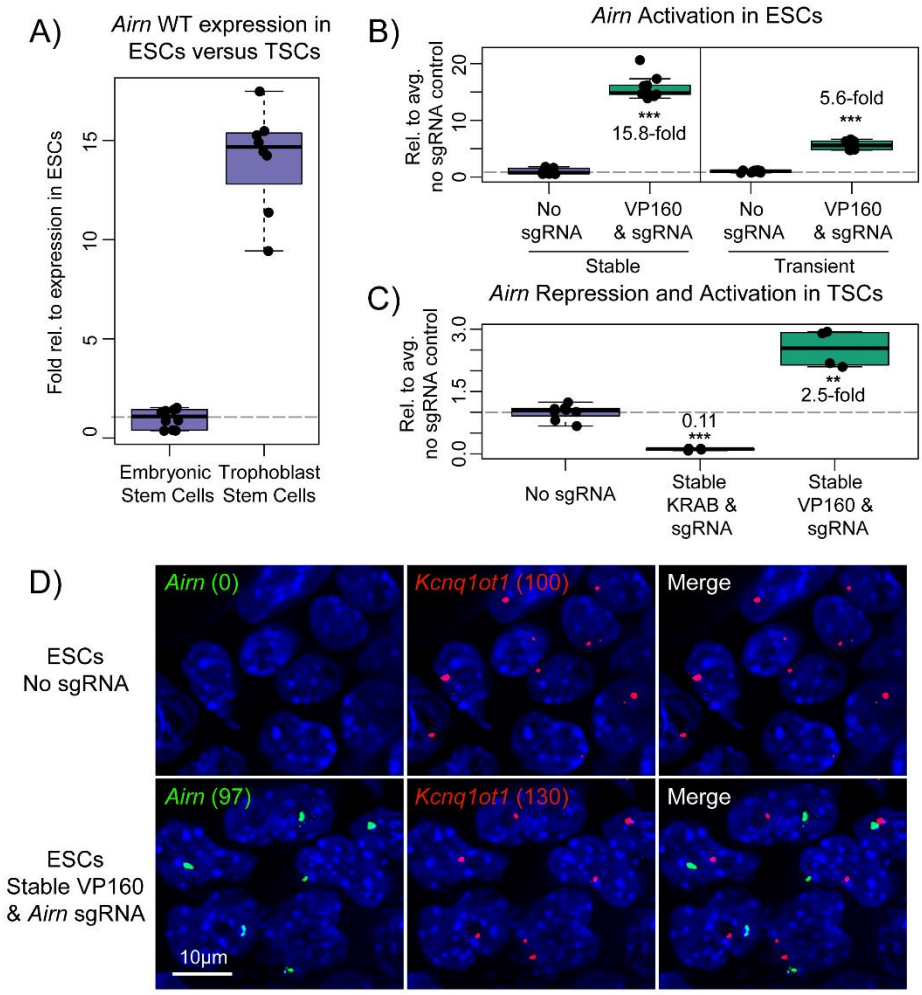
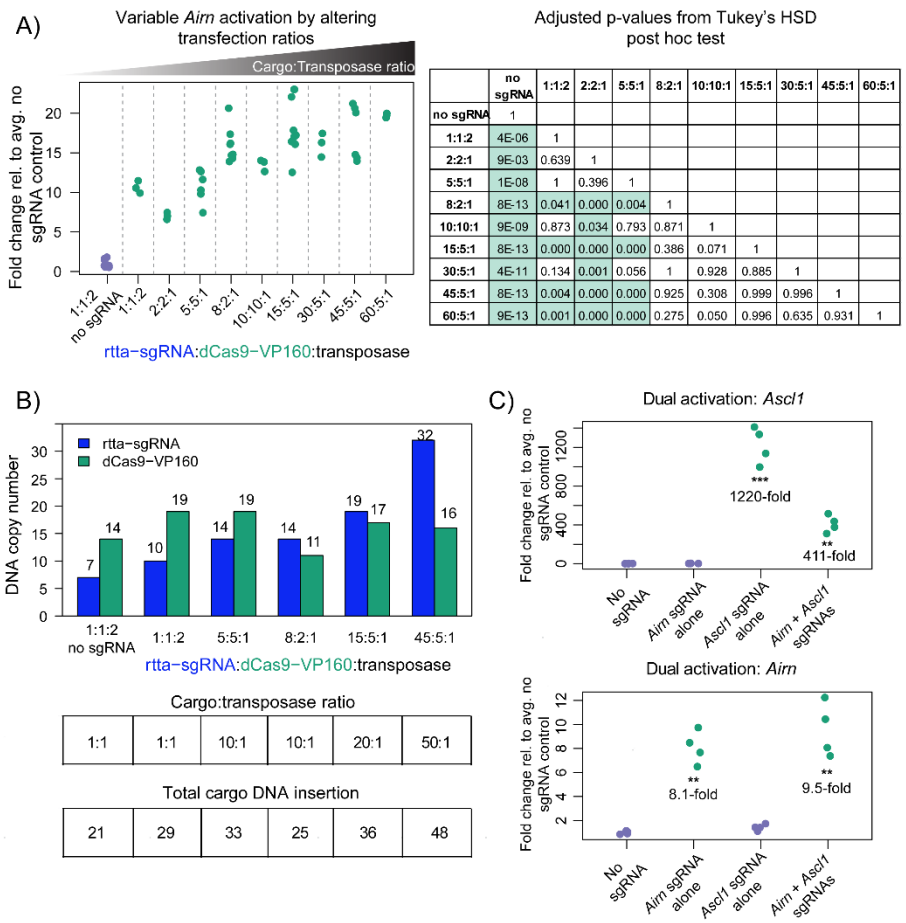


Figure 3. 5. Activation and repression of lncRNA transcription using CRISPR-Bac. See Table S2 for details on replicates and experimental design for each figure panel. All individual qPCR data points are shown in box-and-whisker format representing the mean and the interquartile range. ***, $p < 0.001$; **, $p < 0.01$ from a two-sided t-test between no sgRNA and sgRNA-expressing cells. (A) qPCR showing relative levels of *Airn* expression in wildtype ESCs and TSCs. (B) qPCR measuring *Airn* transcriptional activation in ESCs with stably transfected CRISPR-Bac constructs, dCas9-VP160 and *Airn*-targeting sgRNA, versus transient transfection of dCas9-VP160 and SPgRNA-*Airn*. Data from the non-targeting sgRNA control (No sgRNA) and sgRNA-expressing cells are plotted relative to the average of the signal in the No sgRNA control cells. (C) qPCR results for TSCs with stably transfected dCas9-KRAB or dCas9-VP160 and *Airn*-targeting sgRNA. Data are plotted as in (B). (D) Representative RNA FISH image showing *Airn* and *Kcnq1ot1* RNA in ESCs harvested alongside ESCs from (B). Numbers in parenthesis correspond to spots counted by Imaris software for *Airn* and *Kcnq1ot1* in each cell line. Scale bar, 10 μm.



	no sgRNA	1:1:2	2:2:1	5:5:1	8:2:1	10:10:1	15:5:1	30:5:1	45:5:1	60:5:1
no sgRNA	1									
1:1:2	4E-06	1								
2:2:1	9E-03	0.639	1							
5:5:1	1E-08	1	0.396	1						
8:2:1	8E-13	0.041	0.000	0.004	1					
10:10:1	9E-09	0.873	0.034	0.793	0.871	1				
15:5:1	8E-13	0.000	0.000	0.000	0.386	0.071	1			
30:5:1	4E-11	0.134	0.001	0.056	1	0.928	0.885	1		
45:5:1	8E-13	0.004	0.000	0.000	0.925	0.308	0.999	0.996	1	
60:5:1	9E-13	0.001	0.000	0.000	0.275	0.050	0.996	0.635	0.931	1

Figure 3. 6. Cargo to transposase ratio controls the extent of activation and multiplex gene activation by CRISPR-Bac.

See Table S2 for details on replicates and experimental design for each figure panel. Each replicate in the no sgRNA control and sgRNA-expressing cells is shown relative to the average of the signal in 'No sgRNA'. ***, $p < 0.001$; **, $p < 0.01$ from a two-sided t-test between no sgRNA and sgRNA-expressing cells. (A) qPCR measuring *Airn* activation. X-axis gives the transfection ratio of rtTA-sgRNA to dCas9-VP160 to transposase for each experiment. Ratios are plotted in ascending order based on the summed cargo (rtTA-sgRNA plus dCas9-VP160) to transposase ratio. "1:1:2 no sgRNA" and "8:2:1" data are the same as shown in Figure 5B. Corresponding table gives adjusted p-values from Tukey's HSD post hoc test for all comparisons, where p <= 0.05 are highlighted in green. (B) Bar plot showing DNA copy number per cell for the rtTA-sgRNA and dCas9-VP160 cargos under each transfection condition. Numbers over each bar give the average copy number calculated from 3 technical qPCR replicates. The tables below correspond to the bar plot showing the transfection ratio and total number of DNA cargos inserted (copy number of rtTA-sgRNA plus dCas9-VP160). (C) Simultaneous activation of *Ascl1* and *Airn* transcription upon co-transfection of a pool of four *Ascl1* sgRNAs from (Perez-Pinera et al., 2013) and one *Airn*. Data from the non-targeting sgRNA control (No sgRNA) and sgRNA-expressing cells are plotted relative to the average of the signal in the No sgRNA control cells. Individual qPCR data points are shown.

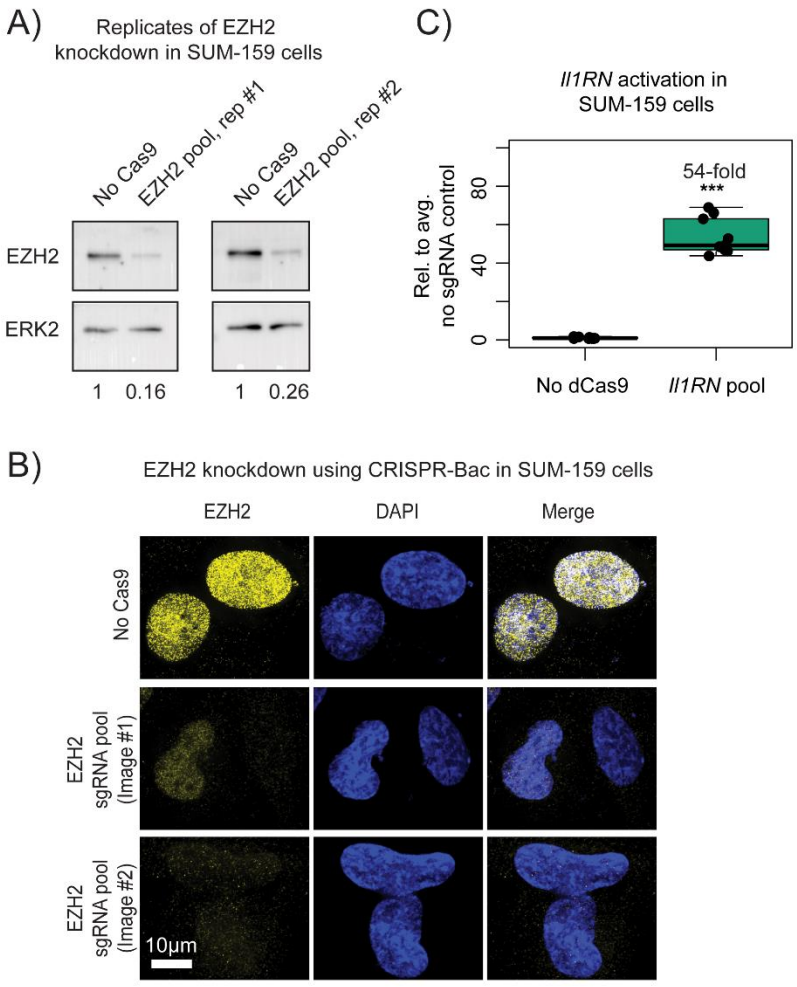


Figure 3. 7. The CRISPR-Bac system functions in human cell lines.
See Table S2 for details on replicates and experimental design for each figure panel. (A) Western blots to EZH2 and ERK2 loading control from two replicate CRISPR-Bac experiments in SUM-159 cells. "No Cas9" refers to measurements taken from dCas9-VP160/rtta-*IL1RN* sgRNA expressing SUM-159 cells that were cultured in parallel to those expressing Cas9 and the EZH2 sgRNA pool. Values underneath blots represent knockdown of EZH2 relative to No Cas9 control and normalized for loading with ERK2 protein levels. (B) Representative immunofluorescence images showing EZH2 knockdown from "No Cas9" and "EZH2 pool, rep #1" SUM-159 cells in (A). Image #1 shows a cell with partial knockdown next to a cell with full knockdown, and image #2 shows two cells with full knockdown. Scale bar, 10 µm. (C) qPCR results showing transcriptional activation of *IL1RN* in SUM-159 cells. In panel (C) "No dCas9" refers to measurements taken from Cas9/rtta-EZH2 sgRNA expressing SUM-159 cells that were cultured in parallel to those expressing dCas9 and the *IL1RN* sgRNAs. Data from the non-targeting sgRNA control (No sgRNA) and sgRNA-expressing cells are plotted relative to the average of the signal in the No sgRNA control cells. Individual qPCR data points are shown in box-and-whisker format. ***, $p < 0.001$ from a two-sided t-test between no sgRNA and sgRNA-expressing cells.

3.5 Supplementary Table Legend

Table S1. Oligonucleotides used. Oligos are divided into 3 groups: 'Oligonucleotides for CRISPR-Bac vector cloning', 'Oligonucleotides for sgRNA cloning', and 'Oligonucleotides for qPCR and PCR'. For all groups, the column 'Oligo' gives a unique name, 'Sequence' gives the full sequence, and 'Usage' says where the oligo was used in the paper. For the sgRNA and qPCR groups, the 'Source' column gives the PMID for the published paper where the sgRNA sequence was originally used, if not this study [(Kearns et al., 2014) for *Oct4* and (Perez-Pinera et al., 2013) for *Ascl1*]. For the sgRNA group, three additional columns give the mm9 coordinates for each sgRNA. For sgRNAs designed in this study, protein knockdown sgRNAs were designed using Desktop Genetics and all other sgRNAs were designed using CRISPOR (Haeussler et al., 2016).

Table S2. Experimental details per figure panel. For each figure panel in column 'Figure', information is given about the experiment and data obtained from the experiment. 'Type of data' refers to qPCR, western blot, etc. 'System' tells whether the CRISPR-Bas system generated in this study was used versus another system. For CRISPR-Bac experiments, 'Transfection ratio' for rtTA-sgRNA:Cas9/dCas9-fusion:transposase is given. 'Doxycycline induction' gives the number of days doxycycline was added to the media before cells were harvested. We found that 2 days was sufficient to detect changes in transcription, but 4 days was optimal for protein knockdown and genomic deletions. Information about the number and types of replicates are given in the 'Biological replicates' and 'Technical replicates' columns.

Table S3. DNA transfection ratios. For each ratio (rtTA-sgRNA:dCas9-VP160:transposase) in Figure 6, this table gives the amount of transfected DNA (in nanograms) for each individual plasmid.

Figure 2 sgRNAs: Protein knockdown

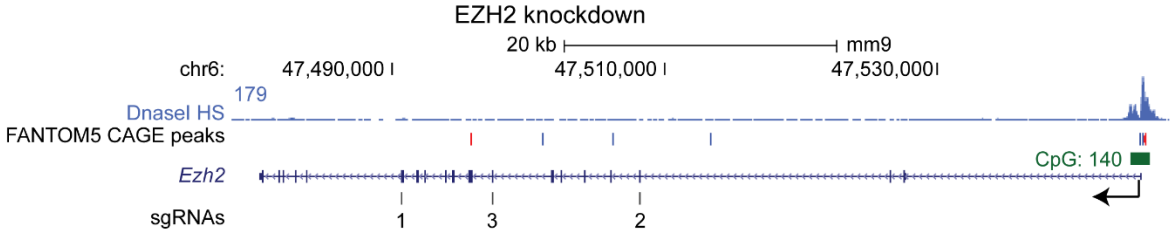


Figure 3 sgRNAs: Targeted deletion

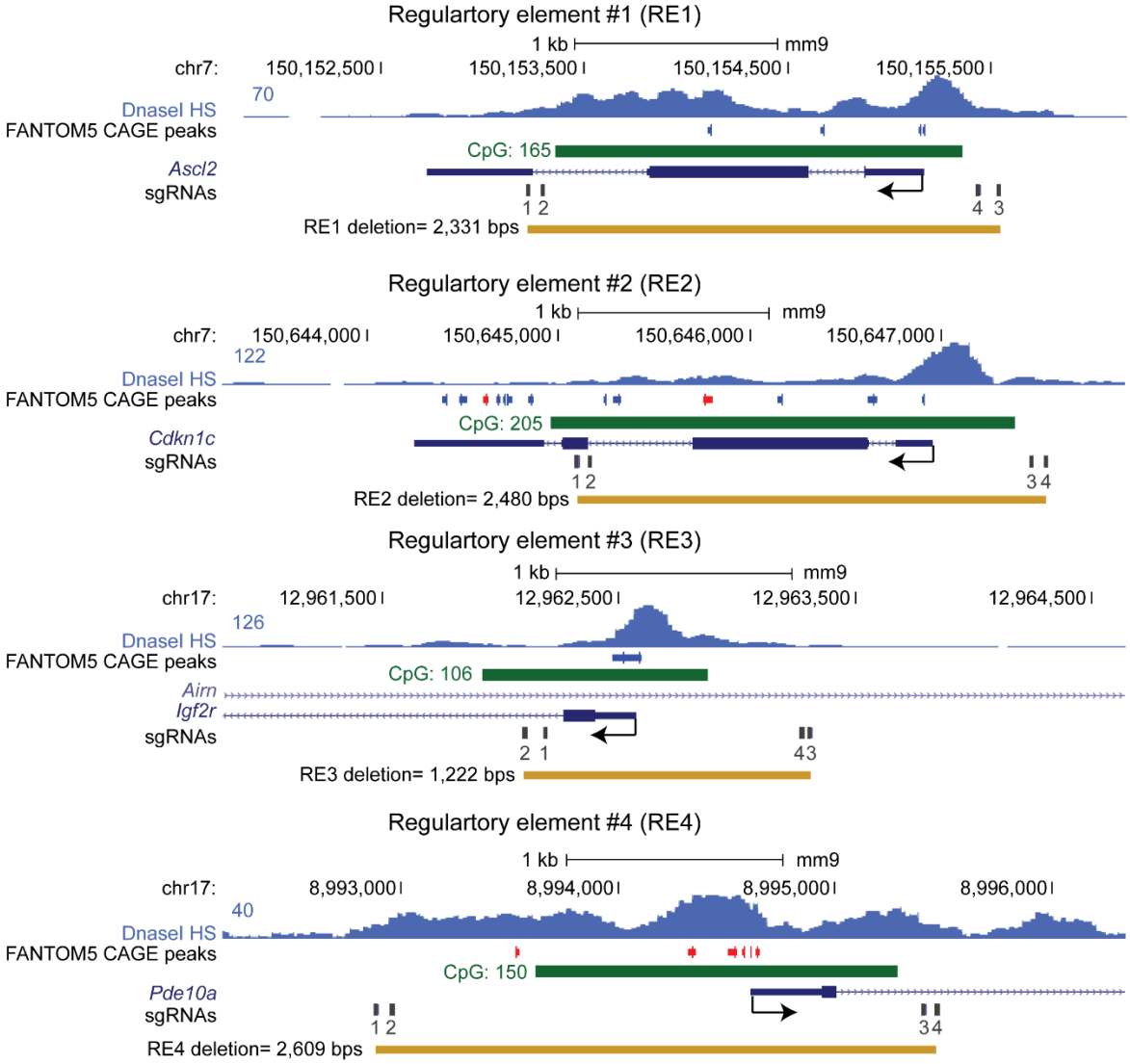


Figure 4 sgRNAs: Control of protein-coding gene transcription

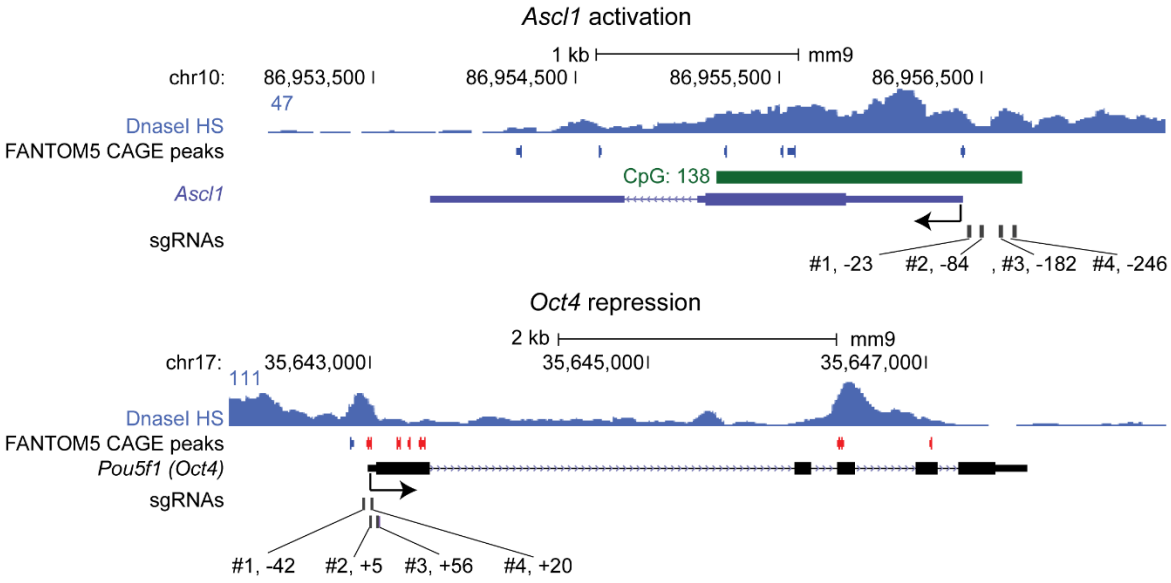


Figure 5 sgRNAs: Control of lncRNA transcription

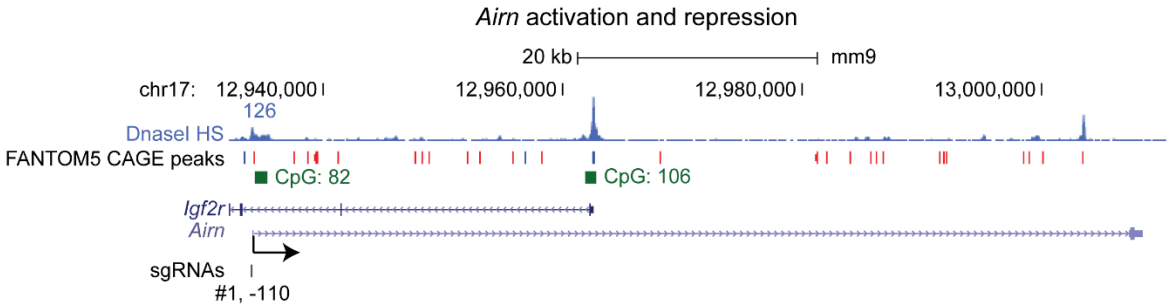
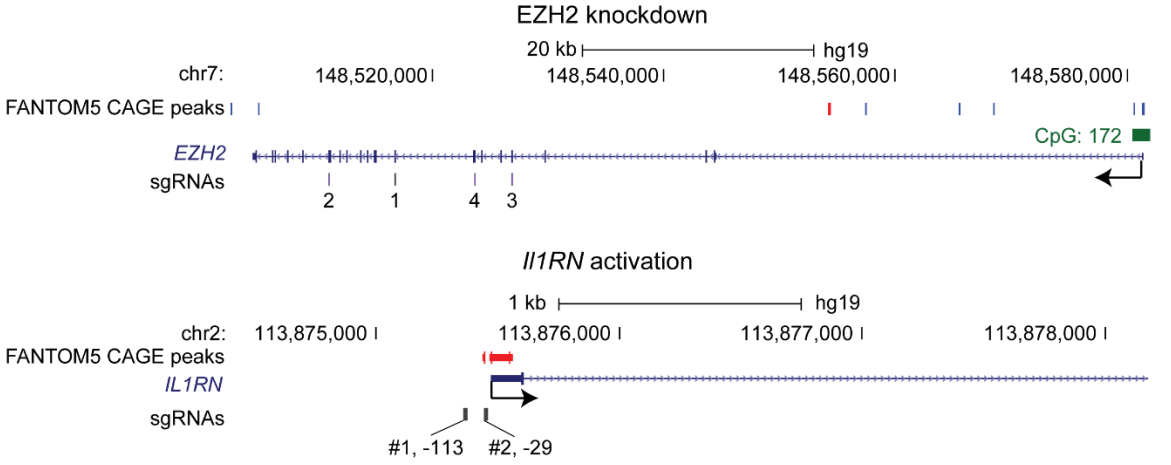


Figure 7 sgRNAs: CRISPR-Bac in SUM-159 cells



Supplemental Figure 3. 1. Diagrams of sgRNA location relative to genomic targets.
UCSC browser tracks are shown for each set of sgRNAs. For all mouse sgRNAs, DnaseI Hypersensitivity (HS) data in E14-ESCs are shown (John et al., 2011). CAGE peaks mark transcription start sites from the FANTOM5 resource (Lizio et al., 2015). CpG island locations and genes are shown underneath these tracks. Finally, positional information is given for each sgRNA. A negative value gives the number of basepairs upstream of the gene TSS and a positive value gives the number of bps downstream of the gene TSS. For regulatory element (RE) deletion sgRNAs, the size of the largest deletion possible is given in base pairs. For all human sgRNAs, the same information is given except for the DNase HS tracks. Scale bars are given at the top of each window. The NCBI37 genome build was used for both mouse and human data.

CHAPTER 4: Discussion and Future Directions

For decades, mechanistic insights on cis-repressive lncRNAs relied almost exclusively on the study of *Xist*. However, *Xist* is the most extreme example given that it is the only lncRNA capable of silencing an entire chromosome. Therefore, we took a different approach and studied *Airn* and *Kcnq1ot1*, two lncRNAs that function similarly to *Xist* but in smaller genomic regions. By comparing and contrasting features of these three lncRNAs and their targeted genomic environments, we gained several novel insights into their cis-repressive mechanism.

We showed that *Airn* and *Kcnq1ot1*, like *Xist*, spread H3K27me3 across megabases of DNA. Using ChIP-seq and a sliding window plotting approach, we were able to illuminate patterns of H3K27me3 at the highest resolution to date. In all three regions, H3K27me3 deposition is not uniform. For any given chromatin bound protein, ChIP-seq typically shows peaks with defined summits that decline in signal across several kilobases. In the lncRNA-targeted regions, H3K27me3 signal also has visible summits, but the signal declines from these points for megabases. This visual pattern for H3K27me3 became clearer upon *Airn* overexpression, the key experiment that suggested the presence of Polycomb nucleation sites.

The observed H3K27me3 patterns are at odds with the current model of lncRNA function *in cis*, again based primarily on studies of *Xist*. The current model posits that the lncRNA itself provides the specificity to bind and recruit Polycomb, and that DNA elements do not confer lncRNA target specificity (Cotton et al., 2014; Loda et al., 2017). Given this model, Polycomb patterns would not show “summits” or “nucleation sites”, but instead, uniformly spread within lncRNA-targeted regions. Our proposed model agrees with an important role for the lncRNA in

binding PRCs, albeit indirectly through RBPs, but importantly, works in concert with 3D structure and PRC-bound CGIs to confer target specificity on chromatin. Discussion of the various aspects of this model is detailed below.

A part of this model focuses on the contribution of 3D structure to lncRNA function. Similar to *Xist* (Engreitz et al., 2013), we showed that *Airn* uses pre-existing structure to target and spread Polycomb within specific TADs that are anchored by CTCF and cohesin proteins. Interestingly, we observed depletion of both CTCF and cohesin at TAD boundaries specifically on the lncRNA targeted allele. The X chromosome showed the strongest depletion, consistent with previous findings that the inactive X has poorly defined TADs (Deng et al., 2015; Nora et al., 2012; Rao et al., 2014; Splinter et al., 2011). *Airn* showed an intermediate level of CTCF/cohesin depletion between *Xist* and *Kcnq1ot1*, a trend which corresponds to lncRNA potency.

Pre-existing structure in the *Airn* and *Kcnq1ot1* regions has two main differences that may help explain their difference in repressive capacity. First, a 6.5 Mb loop within the *Airn* targeted domain is observed in four publicly available Hi-C datasets (Darrow et al., 2016; Deng et al., 2015; Dixon et al., 2012; Rao et al., 2014). High frequency contacts across a distance on this scale are rare; only 38 of the detected 9014 loops in the mouse embryonic kidney Patski cell line are larger than 6.5 Mb (Darrow et al., 2016). Conversely, the 2.3 Mb *Kcnq1ot1* targeted region does not contain long range loops but instead, contains a high density of CTCF and cohesin peaks relative to the *Airn* domain. From this, one could speculate that the presence of long-distance loops facilitate Polycomb spread over 13 Mb within the *Airn* domain, while the high frequency of smaller structures could restrict Polycomb spread to 2.3 Mb within the *Kcnq1ot1* domain. To better understand how pre-existing structure affects lncRNA function, I

propose future experiments to insert and express *Airn*, *Kcnq1ot1*, and *Xist* into the same genomic locus. Downstream experiments would include H3K27me3 ChIP-seq and RNA-seq to compare chromatin and gene expression changes in the presence of each lncRNA within a single environmental context.

In addition, the findings presented here and by others suggest an important role of PRC proteins in *Airn*'s ability to alter 3D structure. First, we show that *Airn* expression is required for compaction within its target domain. Second, previous PRC knockout experiments showed that compaction within the *Airn* and *Kcnq1ot1* regions is also dependent on PRCs (Mi Terranova et al., 2008). Third, deletion of a RING1B-bound CGI in the *Airn* region causes a 4.5 Mb loss of H3K27me3. This suggests that nucleation sites are co-dependent. There are several future experiments that would further elucidate this mechanism. Deletion of other proposed nucleation sites could highlight CGIs that are more or less important for silencing and compaction. DNA FISH experiments could be used to compare the frequency of contact between nucleation sites in the presence and absence of *Airn*.

Surprisingly, deletion of the CGI in the *Airn* domain also caused a significant loss of H3K27me3 in the *Kcnq1ot1* region. Upon further investigation of publicly available Hi-C datasets, we discovered high frequency interchromosomal contact between the *Airn* region on mouse chromosome 17 and the *Kcnq1ot1* region on chromosome 7 in lymphoblastoid cells (Rao et al., 2014). Both observations suggest that *Airn* and *Kcnq1ot1* domains may be co-dependent. To further test this idea, we could perform Hi-C in wildtype and *Airn* or *Kcnq1ot1* knockout cells. Additionally, we could use RNA-FISH to quantify the frequency of co-localization between the two lncRNAs. Finally, to test the importance of PRCs in this localization, we could measure *Airn* and *Kcnq1ot1* localization via FISH in RING1B knockout TSCs.

We showed that on the X chromosome, RING1B and EZH2 are depleted at CGIs on the paternal allele (Xi), but present on the maternal allele (Xa). Others have shown that RING1B and EZH2 are recruited specifically to unmethylated CGIs (Farcas et al., 2012; Wu et al., 2013). Additionally, CGIs on Xi are methylated while CGIs on Xa are unmethylated. This suggests that there could be a relationship between CGI methylation, RING1B/EZH2 depletion, and spread of Polycomb on the X and in *Airn* and *Kcnq1ot1* regions. We did not look at DNA methylation in this dissertation, but it is an important feature to measure moving forward.

Our findings support a potential relationship between lncRNAs and polycomb bodies. Polycomb bodies are nuclear foci of PRCs that are dependent on the SAM polymerization of PHC2 proteins of PRC1 (Isono et al., 2013). In support of lncRNA involvement, *Xist*, *Airn*, and *Kcnq1ot1* form clouds near their sites of transcription that co-localize with foci of PRC1 and PRC2 (Mak et al., 2002; Mi Terranova et al., 2008; Plath et al., 2003, 2004; Silva et al., 2003), where the foci surrounding the inactive X are the largest. HNRNPK mediates PRC1 interaction with Xist (Almeida et al., 2017; Pintacuda et al., 2017). Here, we showed that *Airn*, and *Kcnq1ot1* bind similar levels of HNRNPK as *Xist*, and all three ranked in the top 10 out of 23,366 UCSC genes in HNRNPK signal. Given these data, it is interesting to speculate that *Xist*, *Airn*, and *Kcnq1ot1* seed local concentration of PRCs, causing polymerization of nearby chromatin-bound PRCs and nucleoplasm PRCs to form polycomb bodies. At a lower frequency, these intrachromosomal polycomb bodies may contact one another.

In our experiments, we observed several differences in the spread of lncRNA-dependent PRCs between ESCs and TSCs, with TSCs showing more robust deposition and spreading of PRCs. Relative to ESCs, TSCs maintain higher levels of H3K27me3 and have broader H3K27me3 peaks. Furthermore, there are approximately 10-fold less RING1B/EZH2 co-bound

CGIs in TSCs, and RING1B and EZH2 are depleted at CGIs on the X in the presence of *Xist*. The cause of these differences is unclear. One idea is that one or more co-factors important for lncRNA-dependent PRC spread is present in TSCs but not in ESCs. Another interesting idea based on our model is that a higher density of PRC-bound CGIs in ESCs means that *Xist* contacts each individual CGI less frequently, suggesting less *Xist*-induced spreading. Thus, the model in ESCs may be more similar to a "hit and run" model that is proposed for *Airn* and *Kcnq1ot1* function in TSCs.

We showed that *Xist, Airn,* and *Kcnq1ot1* similarly depend on HNRNPK for the deposition of H3K27me3 within their targeted regions. This result provides further support that these lncRNAs function through a similar mechanism. I hypothesize that other *Xist* co-factors bind *Airn* and *Kcnq1ot1* and play significant roles in their cis-repressive function. To test this hypothesis, we could immunoprecipitate known proteins required for *Xist* function and determine their interactions with *Airn* and *Kcnq1ot1* using RNA-IP methods or RT-qPCR. Then, we could knockdown candidate proteins that interact with these lncRNAs to further decipher the role of specific protein in cis-repressive lncRNA mechanism.

Although there are many similarities between *Xist, Airn,* and *Kcnq1ot1,* there are also differences. In addition to spreading Polycomb across 165Mb and silencing ~1000 genes, *Xist* consistently functions as the most extreme: on the Xi, CTCF and cohesin are largely depleted, RING1B/EZH2 are depleted at CGIs, and H3K27me3 peaks are more broad (up to 1 Mb for a single peak). We found that both lncRNA expression and stability correlated with repressive potency; *Xist* is expressed 25-fold higher and is 5x more stable than *Airn* and *Kcnq1ot1*. We propose that these features influence the potency of *Xist*-induced Polycomb spreading on Xi in two main ways: (1) higher expression allows *Xist* to contact more sites and (2) stability may

increase contact time between *Xist* and nucleation sites. Conversely, we propose that *Airn* and *Kcnq1ot1* function through a "hit and run" mechanism due to their low abundance and stability.

We also speculate that length of the RNA is important for its function. Unspliced *Xist* is 23kb, while *Airn* and *Kcnq1ot1* are both greater than 80kb. Assuming a rate of transcription of 1-3kb/minute (Singh and Padgett, 2009; Wada et al., 2009), Xist is transcribed in <30 minutes, and able to diffuse away from its site of transcription to carry out its function. Conversely, *Airn* and *Kcnq1ot1* take between 1-1.5 hours to be transcribed, a number on par with their stability. These transcripts spend the bulk of their life tethered to their site of transcription by the transcription machinery, unable to diffuse and spread across a whole chromosome. I propose two future experiments to investigate the role of stability and length in lncRNA function. First, I would truncate *Airn* to approximately 40kb and quantify PRC spread relative to WT. Second, I would insert a stabilizing structure in the 3' end of the *Airn* transcript. Additionally, I would combine both aspects to measure *Airn*'s repressive potency following truncation and stabilization of the 40kb transcript. It is possible that altering these lncRNA features would allow *Airn* to function beyond its normal 13Mb boundaries, potentially elevating its repressive potency to match *Xist*.

Based on our measurements of *Airn* and *Kcnq1ot1* expression and function in TSCs, both lncRNAs are highly potent. At only 8 copies per cell, *Airn* silences 10 genes within a 13Mb region and *Kcnq1ot1* silences 7 genes within 2.3 Mb. Increasing *Airn*'s expression by 2.5-fold is sufficient to repress up to 20 additional genes within the 13 Mb region. LncRNAs are often expressed at lower levels relative to mRNAs (Ulitsky and Bartel, 2013). Our work suggests that other lowly expressed cis-acting lncRNAs may have significant effects on nearby genes that are currently unappreciated. We rationalize this potency because, unlike mRNAs that are exported, translated, and have protein products that function in multiple pathways, cis-acting lncRNAs

function locally near their site of transcription. Additionally, while lncRNAs are known to be mis-regulated across cancer types (Yan et al., 2015), many cancer-related lncRNAs may be improperly excluded from analysis pipelines due to low expression or minimal changes between normal and cancer cells. Our work suggests that current pipelines should be re-evaluated to analyze lncRNA relevance in cancer and disease.

There are tens of thousands of lncRNA transcripts annotated in the human genome, and very few have been studied beyond their annotation (Iyer et al., 2015). Here, we investigated the mechanism of three cis-repressive lncRNAs. *Xist*, *Airn*, and *Kcnq1ot1* are excellent models due to their mono-allelic expression, but there are likely many more unidentified cis-acting lncRNAs that are bi-allelically expressed, and thus, more difficult to classify. Many of the findings here can be applied to characterizing other cis-acting lncRNAs and can inform broader aspects of RNA biology.

REFERENCES

Albers, J., Danzer, C., Rechsteiner, M., Lehmann, H., Brandt, L.P., Hejhal, T., Catalano, A., Busenhart, P., Gonçalves, A.F., Brandt, S., et al. (2015). A versatile modular vector system for rapid combinatorial mammalian genetics. J. Clin. Invest. *125*, 1603–1619.

Almeida, M., Pintacuda, G., Masui, O., Koseki, Y., Gdula, M., Cerase, A., Brown, D., Mould, A., Innocent, C., Nakayama, M., et al. (2017). PCGF3/5-PRC1 initiates Polycomb recruitment in X chromosome inactivation. Science *356*, 1081–1084.

Andergassen, D., Dotter, C.P., Wenzel, D., Sigl, V., Bammer, P.C., Muckenhuber, M., Mayer, D., Kulinski, T.M., Theussl, H.-C., Penninger, J.M., et al. (2017). Mapping the mouse Allelome reveals tissue-specific regulation of allelic expression. Elife *6*, e25125.

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.

Barlow, D.P., and Bartolomei, M.S. (2014). Genomic imprinting in mammals. Cold Spring Harb. Perspect. Biol. *6*, a018382.

Barlow, D.P., Stöger, R., Herrmann, B.G., Saito, K., and Schweifer, N. (1991). The mouse insulin-like growth factor type-2 receptor is imprinted and closely linked to the Tme locus. Nature *349*, 84–87.

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

Blackledge, N.P., Farcas, A.M., Kondo, T., King, H.W., McGouran, J.F., Hanssen, L.L.P., Ito, S., Cooper, S., Kondo, K., Koseki, Y., et al. (2014). Variant PRC1 complex-dependent H2A ubiquitylation drives PRC2 recruitment and polycomb domain formation. Cell *157*, 1445–1459.

Brockdorff, N., Ashworth, A., Kay, G.F., McCabe, V.M., Norris, D.P., Cooper, P.J., Swift, S., and Rastan, S. (1992). The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. Cell *71*, 515–526.

Brown, C.J., Hendrich,', B.D., Rupert, J.L., Lafreniere,', R.G., Xing, Y., Lawrence, J., and Willard', H.F. (1992). The Human X/ST Gene: Analysis of a 17 kb Inactive X-Specific RNA That Contains Conserved Repeats and Is Highly Localized within the Nucleus.

Buchenau, P., Hodgson, J., Strutt, H., and Arndt-Jovin, D.J. (1998). The distribution of polycomb-group proteins during cell division and development in Drosophila embryos: impact on models for silencing. J. Cell Biol. *141*, 469–481.

Byron, M., Hall, L.L., and Lawrence, J.B. (2013). A multifaceted FISH approach to study endogenous RNAs and DNAs in native nuclear and cell structures. Curr. Protoc. Hum. Genet. *Chapter 4*, Unit 4.15.

Cadiñanos, J., and Bradley, A. (2007). Generation of an inducible and optimized piggyBac transposon system. Nucleic Acids Res. *35*, e87.

Calabrese, J.M., and Magnuson, T. Roles of Long Non-coding RNAs in X-Chromosome Inactivation.

Calabrese, J.M., Sun, W., Song, L., Mugford, J.W., Williams, L., Yee, D., Starmer, J., Mieczkowski, P., Crawford, G.E., and Magnuson, T. (2012). Site-specific silencing of regulatory elements as a mechanism of X inactivation. Cell *151*, 951–963.

Calabrese, J.M., Starmer, J., Schertzer, M.D., Yee, D., and Magnuson, T. (2015). A Survey of Imprinted Gene Expression in Mouse Trophoblast Stem Cells. G3: Genes|Genomes|Genetics *5*, 751–759.

Canver, M.C., Bauer, D.E., Dass, A., Yien, Y.Y., Chung, J., Masuda, T., Maeda, T., Paw, B.H., and Orkin, S.H. (2014). Characterization of genomic deletion efficiency mediated by clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9 nuclease system in mammalian cells. J. Biol. Chem. *289*, 21312–21324.

Cao, R., and Zhang, Y. (2004). SUZ12 Is Required for Both the Histone Methyltransferase Activity and the Silencing Function of the EED-EZH2 Complex. Mol. Cell *15*, 57–67.

Carrel, L., and Willard, H.F. (2005). X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature *434*, 400–404.

Cech, T.R., and Steitz, J.A. (2014). The Noncoding RNA Revolution—Trashing Old Rules to Forge New Ones. Cell *157*, 77–94.

Cerase, A., Smeets, D., Tang, Y.A., Gdula, M., Kraus, F., Spivakov, M., Moindrot, B., Leleu, M., Tattermusch, A., Demmerle, J., et al. (2014). Spatial separation of Xist RNA and polycomb proteins revealed by superresolution microscopy. Proc. Natl. Acad. Sci. U. S. A. *111*, 2235–2240.

Chadwick, B.P. (2007). Variation in Xi chromatin organization and correlation of the H3K27me3 chromatin territories to transcribed sequences by microarray analysis. Chromosoma *116*, 147–157.

Cheng, A.W., Wang, H., Yang, H., Shi, L., Katz, Y., Theunissen, T.W., Rangarajan, S., Shivalila, C.S., Dadon, D.B., and Jaenisch, R. (2013). Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system. Cell Res. *23*, 1163–1171.

Chittock, E.C., Latwiel, S., Miller, T.C.R., and Müller, C.W. (2017). Molecular architecture of polycomb repressive complexes. Biochem. Soc. Trans. *45*, 193–205.

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.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., et al. (2013). Multiplex genome engineering using CRISPR/Cas systems. Science *339*, 819–823.

Cooper, S., Dienstbier, M., Hassan, R., Schermelleh, L., Sharif, J., Blackledge, N.P., De Marco, V., Elderkin, S., Koseki, H., Klose, R., et al. (2014). Targeting Polycomb to Pericentric Heterochromatin in Embryonic Stem Cells Reveals a Role for H2AK119u1 in PRC2 Recruitment.

Cotton, A.M., Chen, C.-Y., Lam, L.L., Wasserman, W.W., Kobor, M.S., and Brown, C.J. (2014). Spread of X-chromosome inactivation into autosomal sequences: role for DNA elements, chromatin features and chromosomal domains. Hum. Mol. Genet. *23*, 1211–1223.

Cremer, T., and Cremer, M. (2010). Chromosome territories. Cold Spring Harb. Perspect. Biol. *2*, a003889.

Darrow, E.M., Huntley, M.H., Dudchenko, O., Stamenova, E.K., Durand, N.C., Sun, Z., Huang, S.-C., Sanborn, A.L., Machol, I., Shamim, M., et al. (2016). Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture. Proc. Natl. Acad. Sci. U. S. A. *113*, E4504-12.

Davidovich, C., Zheng, L., Goodrich, K.J., and Cech, T.R. (2013). Promiscuous RNA binding by Polycomb repressive complex 2. Nat. Struct. Mol. Biol. *20*, 1250–1257.

Davidovich, C., Wang, X., Cifuentes-Rojas, C., Goodrich, K.J., Gooding, A.R., Lee, J.T., and Cech, T.R. (2015). Toward a Consensus on the Binding Specificity and Promiscuity of PRC2 for RNA.

DeChiara, T.M., Robertson, E.J., and Efstratiadis, A. (1991). Parental imprinting of the mouse insulin-like growth factor II gene. Cell *64*, 849–859.

Deng, X., Ma, W., Ramani, V., Hill, A., Yang, F., Ay, F., Berletch, J.B., Blau, C.A., Shendure, J., Duan, Z., et al. (2015). Bipartite structure of the inactive mouse X chromosome. Genome Biol. *16*, 152.

Ding, S., Wu, X., Li, G., Han, M., Zhuang, Y., and Xu, T. (2005a). Efficient Transposition of the piggyBac (PB) Transposon in Mammalian Cells and Mice. Cell *122*, 473–483.

Ding, S., Wu, X., Li, G., Han, M., Zhuang, Y., and Xu, T. (2005b). Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. Cell *122*, 473–483.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Dixon, J.R., Gorkin, D.U., and Ren, B. (2016). Chromatin Domains: The Unit of Chromosome Organization. Mol. Cell *62*, 668–680.

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

Dowen, J.M., Fan, Z.P., Hnisz, D., Ren, G., Abraham, B.J., Zhang, L.N., Weintraub, A.S., Schuijers, J., Lee, T.I., Zhao, K., et al. (2014). Control of Cell Identity Genes Occurs in Insulated Neighborhoods in Mammalian Chromosomes.

Durand, N.C., Robinson, J.T., Shamim, M.S., Machol, I., Mesirov, J.P., Lander, E.S., and Aiden, E.L. (2016). Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. Cell Syst. *3*, 99–101.

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.

Eskeland, R., Leeb, M., Grimes, G.R., Kress, C.M., Boyle, S., Sproul, D., Gilbert, N., Fan, Y., Skoultchi, A.I., Wutz, A., et al. (2010). Ring1B Compacts Chromatin Structure and Represses Gene Expression Independent of Histone Ubiquitination. Mol. Cell *38*, 452–464.

Farcas, A.M., Blackledge, N.P., Sudbery, I., Long, H.K., McGouran, J.F., Rose, N.R., Lee, S., Sims, D., Cerase, A., Sheahan, T.W., et al. (2012). KDM2B links the Polycomb Repressive Complex 1 (PRC1) to recognition of CpG islands. Elife *1*, e00205.

Fatica, A., and Bozzoni, I. (2014). Long non-coding RNAs: new players in cell differentiation and development. Nat. Rev. Genet. *15*, 7–21.

Ferguson-Smith, A.C., Cattanach, B.M., Barton, S.C., Beechey, C. V., and Surani, M.A. (1991). Embryological and molecular investigations of parental imprinting on mouse chromosome 7. Nature *351*, 667–670.

Francis, N.J., Kingston, R.E., and Woodcock, C.L. (2004). Chromatin compaction by a polycomb group protein complex. Science *306*, 1574–1577.

Fursova, N.A., Blackledge, N.P., Nakayama, M., Ito, S., Koseki, Y., Farcas, A.M., King, H.W., Koseki, H., and Klose, R.J. (2019). Synergy between Variant PRC1 Complexes Defines Polycomb-Mediated Gene Repression. Mol. Cell.

Gao, Z., Zhang, J., Bonasio, R., Strino, F., Sawai, A., Parisi, F., Kluger, Y., and Reinberg, D. (2012). PCGF homologs, CBX proteins, and RYBP define functionally distinct PRC1 family complexes. Mol. Cell *45*, 344–356.

Gasperini, M., Findlay, G.M., McKenna, A., Milbank, J.H., Lee, C., Zhang, M.D., Cusanovich, D.A., and Shendure, J. (2017). CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for HPRT1 Expression via Thousands of Large, Programmed Genomic Deletions. Am. J. Hum. Genet. *101*, 192–205.

Geisler, S., and Coller, J. (2013). RNA in unexpected places: long non-coding RNA functions in

diverse cellular contexts. Nat. Rev. Mol. Cell Biol. *14*, 699–712.

Giorgetti, L., Lajoie, B.R., Carter, A.C., Attia, M., Zhan, Y., Xu, J., Chen, C.J., Kaplan, N., Chang, H.Y., Heard, E., et al. (2016). Structural organization of the inactive X chromosome in the mouse. Nature *535*, 575–579.

Gossen, M., Freundlieb, S., Bender, G., Müller, G., Hillen, W., and Bujard, H. (1995). Transcriptional activation by tetracyclines in mammalian cells. Science *268*, 1766–1769.

Grau, D.J., Chapman, B.A., Garlick, J.D., Borowsky, M., Francis, N.J., and Kingston, R.E. (2011). Compaction of chromatin by diverse Polycomb group proteins requires localized regions of high charge. Genes Dev. *25*, 2210–2221.

Grigoriadis, A., Mackay, A., Noel, E., Wu, P.J., Natrajan, R., Frankum, J., Reis-Filho, J.S., and Tutt, A. (2012). Molecular characterisation of cell line models for triple-negative breast cancers. BMC Genomics *13*, 619.

Haeussler, M., Schönig, K., Eckert, H., Eschstruth, A., Mianné, J., Renaud, J.-B., Schneider-Maunoury, S., Shkumatava, A., Teboul, L., Kent, J., et al. (2016). Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. Genome Biol. *17*, 148.

Hartenian, E., and Doench, J.G. (2015). Genetic screens and functional genomics using CRISPR/Cas9 technology. FEBS J. *282*, 1383–1393.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol. Cell *38*, 576–589.

Hsu, P.D., Lander, E.S., and Zhang, F. (2014). Development and Applications of CRISPR-Cas9 for Genome Engineering. Cell *157*, 1262–1278.

Hunkapiller, J., Shen, Y., Diaz, A., Cagney, G., McCleary, D., Ramalho-Santos, M., Krogan, N., Ren, B., Song, J.S., and Reiter, J.F. (2012). Polycomb-Like 3 Promotes Polycomb Repressive Complex 2 Binding to CpG Islands and Embryonic Stem Cell Self-Renewal. PLoS Genet. *8*, e1002576.

Isono, K., Endo, T.A., Ku, M., Yamada, D., Suzuki, R., Sharif, J., Ishikura, T., Toyoda, T., Bernstein, B.E., and Koseki, H. (2013). SAM Domain Polymerization Links Subnuclear Clustering of PRC1 to Gene Silencing.

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. Nat. Genet. *47*, 199–208.

John, S., Sabo, P.J., Thurman, R.E., Sung, M.-H., Biddie, S.C., Johnson, T.A., Hager, G.L., and Stamatoyannopoulos, J.A. (2011). Chromatin accessibility pre-determines glucocorticoid

receptor binding patterns. Nat. Genet. *43*, 264–268.

Joung, J., Konermann, S., Gootenberg, J.S., Abudayyeh, O.O., Platt, R.J., Brigham, M.D., Sanjana, N.E., and Zhang, F. (2017). Genome-scale CRISPR-Cas9 knockout and transcriptional activation screening. Nat. Protoc. *12*, 828–863.

Kabadi, A.M., Ousterout, D.G., Hilton, I.B., and Gersbach, C.A. (2014). Multiplex CRISPR/Cas9-based genome engineering from a single lentiviral vector. Nucleic Acids Res. *42*, e147.

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S., et al. (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430–435.

Kahlig, K.M., Saridey, S.K., Kaja, A., Daniels, M.A., George, A.L., and Wilson, M.H. (2010). Multiplexed transposon-mediated stable gene transfer in human cells. Proc. Natl. Acad. Sci. U. S. A. *107*, 1343–1348.

Kaji, K., Norrby, K., Paca, A., Mileikovsky, M., Mohseni, P., and Woltjen, K. (2009). Virus-free induction of pluripotency and subsequent excision of reprogramming factors. Nature *458*, 771–775.

Kalantry, S., Mills, K.C., Yee, D., Otte, A.P., Panning, B., and Magnuson, T. (2006). The Polycomb group protein Eed protects the inactive X-chromosome from differentiation-induced reactivation. Nat. Cell Biol. *8*, 195–202.

Kalantry, S., Purushothaman, S., Bowen, R.B., Starmer, J., and Magnuson, T. (2009). Evidence of Xist RNA-independent initiation of mouse imprinted X-chromosome inactivation. Nature *460*.

Kalb, R., Latwiel, S., Baymaz, H.I., Jansen, P.W.T.C., Müller, C.W., Vermeulen, M., and Müller, J. (2014). Histone H2A monoubiquitination promotes histone H3 methylation in Polycomb repression. Nat. Struct. Mol. Biol. *21*, 569–571.

Kanduri, C. (2011). Kcnq1ot1: A chromatin regulatory RNA. Semin. Cell Dev. Biol. *22*, 343–350.

Kearns, N.A., Genga, R.M.J., Enuameh, M.S., Garber, M., Wolfe, S.A., and Maehr, R. (2014). Cas9 effector-mediated regulation of transcription and differentiation in human pluripotent stem cells. Development *141*, 219–223.

Kelsey, A.D., Yang, C., Leung, D., Minks, J., Dixon-McDougall, T., Baldry, S.E.L., Bogutz, A.B., Lefebvre, L., and Brown, C.J. (2015). Impact of flanking chromosomal sequences on localization and silencing by the human non-coding RNA XIST. Genome Biol. *16*, 208.

Kirk, J.M., Kim, S.O., Inoue, K., Smola, M.J., Lee, D.M., Schertzer, M.D., Wooten, J.S., Baker, A.R., Sprague, D., Collins, D.W., et al. (2018). Functional classification of long non-coding RNAs by k-mer content. Nat. Genet. *50*, 1474–1482.

Kornienko, A.E., Dotter, C.P., Guenzl, P.M., Gisslinger, H., Gisslinger, B., Cleary, C., Kralovics, R., Pauler, F.M., and Barlow, D.P. (2016). Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. Genome Biol. *17*, 14.

Korostowski, L., Sedlak, N., and Engel, N. (2012). The Kcnq1ot1 Long Non-Coding RNA Affects Chromatin Conformation and Expression of Kcnq1, but Does Not Regulate Its Imprinting in the Developing Heart. PLoS Genet. *8*, e1002956.

Ku, M., Koche, R.P., Rheinbay, E., Mendenhall, E.M., Endoh, M., Mikkelsen, T.S., Presser, A., Nusbaum, C., Xie, X., Chi, A.S., et al. (2008). Genomewide Analysis of PRC1 and PRC2 Occupancy Identifies Two Classes of Bivalent Domains. PLoS Genet *4*.

Kundu, S., Ji, F., Sunwoo, H., Jain, G., Lee, J.T., Sadreyev, R.I., Dekker, J., and Kingston, R.E. (2017). Polycomb Repressive Complex 1 Generates Discrete Compacted Domains that Change during Differentiation. Mol. Cell *65*, 432-446.e5.

Kung, J.T., Kesner, B., An, J.Y., Ahn, J.Y., Cifuentes-Rojas, C., Colognori, D., Jeon, Y., Szanto, A., del Rosario, B.C., Pinter, S.F., et al. (2015). Locus-Specific Targeting to the X Chromosome Revealed by the RNA Interactome of CTCF. Mol. Cell *57*, 361–375.

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.

Latos, P.A., Stricker, S.H., Steenpass, L., Pauler, F.M., Huang, R., Senergin, B.H., Regha, K., Koerner, M. V, Warczok, K.E., Unger, C., et al. (2009). An in vitro ES cell imprinting model shows that imprinted expression of the Igf2r gene arises from an allele-specific expression bias. Development *136*, 437–448.

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, J.T., and Bartolomei, M.S. (2013). X-inactivation, imprinting, and long noncoding RNAs in health and disease. Cell *152*, 1308–1323.

Lee, J.T., and Jaenisch, R. (1997). Long-range cis effects of ectopic X-inactivation centres on a mouse autosome. Nature *386*, 275–279.

Lewis, A., Mitsuya, K., Umlauf, D., Smith, P., Dean, W., Walter, J., Higgins, M., Feil, R., and Reik, W. (2004). Imprinting on distal chromosome 7 in the placenta involves repressive histone methylation independent of DNA methylation. Nat. Genet. *36*, 1291–1295.

Lewis, A., Green, K., Dawson, C., Redrup, L., Huynh, K.D., Lee, J.T., Hemberger, M., and Reik, W. (2006). Epigenetic dynamics of the Kcnq1 imprinted domain in the earlyembryo. Development *133*, 4203–4210.

Li, G., Margueron, R., Ku, M., Chambon, P., Bernstein, B.E., and Reinberg, D. (2010). Jarid2 and PRC2, partners in regulating gene expression. Genes Dev. *24*, 368–380.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, 1000 Genome Project Data Processing (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078.

Li, H., Liefke, R., Jiang, J., Kurland, J.V., Tian, W., Deng, P., Zhang, W., He, Q., Patel, D.J., Bulyk, M.L., et al. (2017a). Polycomb-like proteins link the PRC2 complex to CpG islands. Nature *549*, 287–291.

Li, M.A., Turner, D.J., Ning, Z., Yusa, K., Liang, Q., Eckert, S., Rad, L., Fitzgerald, T.W., Craig, N.L., and Bradley, A. (2011). Mobilization of giant piggyBac transposons in the mouse genome. Nucleic Acids Res. *39*, e148.

Li, S., Zhang, A., Xue, H., Li, D., and Liu, Y. (2017b). One-Step piggyBac Transposon-Based CRISPR/Cas9 Activation of Multiple Genes. Mol. Ther. Nucleic Acids *8*, 64–76.

Li, Z., Michael, I.P., Zhou, D., Nagy, A., and Rini, J.M. (2013). Simple piggyBac transposon-based mammalian cell expression system for inducible protein production. Proc. Natl. Acad. Sci. U. S. A. *110*, 5004–5009.

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.

Lizio, M., Harshbarger, J., Shimoji, H., Severin, J., Kasukawa, T., Sahin, S., Abugessaisa, I., Fukuda, S., Hori, F., Ishikawa-Kato, S., et al. (2015). Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol. *16*, 22.

Loda, A., Brandsma, J.H., Vassilev, I., Servant, N., Loos, F., Amirnasr, A., Splinter, E., Barillot, E., Poot, R.A., Heard, E., et al. (2017). Genetic and epigenetic features direct differential efficiency of Xist-mediated silencing at X-chromosomal and autosomal locations. Nat. Commun. *8*, 690.

Luo, Y., Lin, L., Bolund, L., and Sørensen, C.B. (2014). Efficient construction of rAAV-based gene targeting vectors by Golden Gate cloning. Biotechniques *56*, 263–268.

Lyle, R., Watanabe, D., Vruchte, D. te, Lerchner, W., Smrzka, O.W., Wutz, A., Schageman, J., Hahner, L., Davies, C., and Barlow, D.P. (2000). The imprinted antisense RNA at the Igf2r locus overlaps but does not imprint Mas1. Nat. Genet. *25*, 19–21.

Lynch, M.D., Smith, A.J.H., De Gobbi, M., Flenley, M., Hughes, J.R., Vernimmen, D., Ayyub, H., Sharpe, J.A., Sloane-Stanley, J.A., Sutherland, L., et al. (2012). An interspecies analysis reveals a key role for unmethylated CpG dinucleotides in vertebrate Polycomb complex recruitment. EMBO J. *31*, 317–329.

Lyon, M.F. (1961). Gene Action in the X-chromosome of the Mouse (Mus musculus L.). Nature *190*, 372–373.

Mak, W., Baxter, J., Silva, J., Newall, A.E., Otte, A.P., and Brockdorff, N. (2002). Mitotically Stable Association of Polycomb Group Proteins Eed and Enx1 with the Inactive X Chromosome

in Trophoblast Stem Cells. Curr. Biol. *12*, 1016–1020.

Mancini-Dinardo, D., Steele, S.J.S., Levorse, J.M., Ingram, R.S., and Tilghman, S.M. (2006). Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes. Genes Dev. *20*, 1268–1282.

Margueron, R., Justin, N., Ohno, K., Sharpe, M.L., Son, J., Drury III, W.J., Voigt, P., Martin, S.R., Taylor, W.R., De Marco, V., et al. (2009). Role of the polycomb protein EED in the propagation of repressive histone marks. Nature *461*, 762–767.

Marks, H., Chow, J.C., Denissov, S., Françoijs, K.-J., Brockdorff, N., Heard, E., and Stunnenberg, H.G. (2009). High-resolution analysis of epigenetic changes associated with X inactivation. Genome Res. *19*, 1361–1373.

McGrath, J., and Solter, D. (1984). Completion of mouse embryogenesis requires both the maternal and paternal genomes. Cell *37*, 179–183.

Mendenhall, E.M., Koche, R.P., Truong, T., Zhou, V.W., Issac, B., Chi, A.S., Ku, M., and Bernstein, B.E. (2010). GC-Rich Sequence Elements Recruit PRC2 in Mammalian ES Cells. PLoS Genet. *6*, e1001244.

Meng, L., Person, R.E., Huang, W., Zhu, P.J., Costa-Mattioli, M., and Beaudet, A.L. (2013). Truncation of Ube3a-ATS unsilences paternal Ube3a and ameliorates behavioral defects in the Angelman syndrome mouse model. PLoS Genet. *9*, e1004039.

Mi Terranova, R., Yokobayashi, S., Stadler, M.B., Otte, A.P., van Lohuizen, M., Orkin, S.H., Peters, A.H.F.M., Terranova, R., Yokobayashi, S., Stadler, M.B., et al. (2008). Polycomb group proteins Ezh2 and Rnf2 direct genomic contraction and imprinted repression in early mouse embryos. Dev. Cell *15*, 668–679.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385.

Oksuz, O., Narendra, V., Lee, C.-H., Descostes, N., LeRoy, G., Raviram, R., Blumenberg, L., Karch, K., Rocha, P.P., Garcia, B.A., et al. (2018). Capturing the Onset of PRC2-Mediated Repressive Domain Formation. Mol. Cell *70*, 1149-1162.e5.

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. Mol. Cell *32*, 232–246.

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

Perez-Pinera, P., Kocak, D.D., Vockley, C.M., Adler, A.F., Kabadi, A.M., Polstein, L.R., Thakore, P.I., Glass, K.A., Ousterout, D.G., Leong, K.W., et al. (2013). RNA-guided gene

activation by CRISPR-Cas9–based transcription factors. Nat. Methods *10*, 973–976.

Pertea, M., Shumate, A., Pertea, G., Varabyou, A., Breitwieser, F.P., Chang, Y.-C., Madugundu, A.K., Pandey, A., and Salzberg, S.L. (2018). CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. Genome Biol. *19*, 208.

Pintacuda, G., Wei, G., Roustan, C., Kirmizitas, B.A., Solcan, N., Cerase, A., Castello, A., Mohammed, S., Moindrot, B., Nesterova, T.B., et al. (2017). hnRNPK Recruits PCGF3/5-PRC1 to the Xist RNA B-Repeat to Establish Polycomb-Mediated Chromosomal Silencing. Mol. Cell *68*, 955-969.e10.

Pinter, S.F., Sadreyev, R.I., Yildirim, E., Jeon, Y., Ohsumi, T.K., Borowsky, M., and Lee, J.T. (2012). Spreading of X chromosome inactivation via a hierarchy of defined Polycomb stations. Genome Res. *22*, 1864–1876.

Plath, K., Fang, J., Mlynarczyk-Evans, S.K., Cao, R., Worringer, K.A., Wang, H., de la Cruz, C.C., Otte, A.P., Panning, B., and Zhang, Y. (2003). Role of histone H3 lysine 27 methylation in X inactivation. Science *300*, 131–135.

Plath, K., Talbot, D., Hamer, K.M., Otte, A.P., Yang, T.P., Jaenisch, R., and Panning, B. (2004). Developmentally regulated alterations in Polycomb repressive complex 1 proteins on the inactive X chromosome. J. Cell Biol. *167*, 1025–1035.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Quinn, J., Kunath, T., and Rossant, J. (2006). Mouse trophoblast stem cells. Methods Mol. Med. *121*, 125–148.

R Core Team (2017). R: a language and environment for statistical computing.

Raab, J.R., Smith, K.N., Spear, C.C., Manner, C.J., Calabrese, J.M., and Magnuson, T. (2019). SWI/SNF remains localized to chromatin in the presence of SCHLAP1. Nat. Genet. *51*, 26–29.

Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. Nat. Protoc. *8*, 2281–2308.

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.

Redrup, L., Branco, M.R., Perdeaux, E.R., Krueger, C., Lewis, A., Santos, F., Nagano, T., Cobb, B.S., Fraser, P., and Reik, W. (2009). The long noncoding RNA Kcnq1ot1 organises a lineage-specific nuclear domain for epigenetic gene silencing. Development *136*, 525–530.

Regha, K., Sloane, M.A., Huang, R., Pauler, F.M., Warczok, K.E., Melikant, B., Radolf, M., Martens, J.H.A., Schotta, G., Jenuwein, T., et al. (2007). Active and Repressive Chromatin Are

Interspersed without Spreading in an Imprinted Gene Cluster in the Mammalian Genome. Mol. Cell *27*, 353–366.

Riising, E.M., Comet, I., Leblanc, B., Wu, X., Johansen, J.V., and Helin, K. (2014). Gene Silencing Triggers Polycomb Repressive Complex 2 Recruitment to CpG Islands Genome Wide. Mol. Cell *55*, 347–360.

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

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Sakuma, T., Nishikawa, A., Kume, S., Chayama, K., and Yamamoto, T. (2015). Multiplex genome engineering in human cells using all-in-one CRISPR/Cas9 vector system. Sci. Rep. *4*, 5400.

Schertzer, M.D., Thulson, E., Braceros, K.C.A., Lee, D.M., Hinkle, E.R., Murphy, R.M., Kim, S.O., Vitucci, E.C.M., and Calabrese, J.M. (2018). A piggyBac-based toolkit for inducible genome editing in mammalian cells. BioRxiv 448803.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). Fiji: an open-source platform for biological-image analysis. Nat. Methods *9*, 676–682.

Schoenfelder, S., Sugar, R., Dimond, A., Javierre, B.-M., Armstrong, H., Mifsud, B., Dimitrova, E., Matheson, L., Tavares-Cadete, F., Furlan-Magaril, M., et al. (2015). Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. Nat. Genet. *47*, 1179–1186.

Schwartz, Y.B., and Pirrotta, V. (2013). A new world of Polycombs: unexpected partnerships and emerging functions. Nat. Rev. Genet. *14*, 853–864.

Seidl, C.I.M., Stricker, S.H., and Barlow, D.P. (2006). The imprinted Air ncRNA is an atypical RNAPII transcript that evades splicing and escapes nuclear export. EMBO J. *25*, 3565–3575.

Shin, K.-J., Wall, E.A., Zavzavadjian, J.R., Santat, L.A., Liu, J., Hwang, J.-I., Rebres, R., Roach, T., Seaman, W., Simon, M.I., et al. (2006). A single lentiviral vector platform for microRNA-based conditional RNA interference and coordinated transgene expression. Proc. Natl. Acad. Sci. U. S. A. *103*, 13759–13764.

Silva, J., Mak, W., Zvetkova, I., Appanah, R., Nesterova, T.B., Webster, Z., Peters, A.H.F.., Jenuwein, T., Otte, A.P., and Brockdorff, N. (2003). Establishment of Histone H3 Methylation on the Inactive X Chromosome Requires Transient Recruitment of Eed-Enx1 Polycomb Group Complexes. Dev. Cell *4*, 481–495.

Simon, J.A., and Kingston, R.E. (2013). Occupying Chromatin: Polycomb Mechanisms for Getting to Genomic Targets, Stopping Transcriptional Traffic, and Staying Put. Mol. Cell *49*,

808–824.

Simon, M.D., Pinter, S.F., Fang, R., Sarma, K., Rutenberg-Schoenberg, M., Bowman, S.K., Kesner, B.A., Maier, V.K., Kingston, R.E., and Lee, J.T. (2013). High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. Nature *504*, 465–469.

Singh, J., and Padgett, R.A. (2009). Rates of in situ transcription and splicing in large human genes. Nat. Struct. Mol. Biol. *16*, 1128.

Sleutels, F., Zwart, R., and Barlow, D.P. (2002). The non-coding Air RNA is required for silencing autosomal imprinted genes. Nature *415*, 810–813.

Smeets, D., Markaki, Y., Schmid, V.J., Kraus, F., Tattermusch, A., Cerase, A., Sterr, M., Fiedler, S., Demmerle, J., Popken, J., et al. (2014). Three-dimensional super-resolution microscopy of the inactive X chromosome territory reveals a collapse of its active nuclear compartment harboring distinct Xist RNA foci. Epigenetics Chromatin *7*, 8.

Splinter, E., de Wit, E., Nora, E.P., Klous, P., van de Werken, H.J.G., Zhu, Y., Kaaij, L.J.T., van Ijcken, W., Gribnau, J., Heard, E., et al. (2011). The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. Genes Dev. *25*, 1371–1383.

Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D., Tiwari, V.K., et al. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature *480*, 490.

Starmer, J., and Magnuson, T. (2016). Detecting broad domains and narrow peaks in ChIP-seq data with hiddenDomains. BMC Bioinformatics *17*, 144.

Stöger, R., Kubička, P., Liu, C.-G., Kafri, T., Razin, A., Cedar, H., and Barlow, D.P. (1993). Maternal-specific methylation of the imprinted mouse Igf2r locus identifies the expressed locus as carrying the imprinting signal. Cell *73*, 61–71.

Sunwoo, H., Wu, J.Y., and Lee, J.T. (2015). The Xist RNA-PRC2 complex at 20-nm resolution reveals a low Xist stoichiometry and suggests a hit-and-run mechanism in mouse cells. Proc. Natl. Acad. Sci. U. S. A. *112*, E4216-25.

Surani, M.A., and Barton, S.C. (1983). Development of gynogenetic eggs in the mouse: implications for parthenogenetic embryos. Science *222*, 1034–1036.

Tada, T., Obata, Y., Tada, M., Goto, Y., Nakatsuji, N., Tan, S., Kono, T., and Takagi, N. (2000). Imprint switching for non-random X-chromosome inactivation during mouse oocyte growth. Development *127*.

Takagi, N., and Sasaki, M. (1975). Preferential inactivation of the paternally derived X chromosome in the extraembryonic membranes of the mouse. Nature *256*, 640–642.

Tang, Y.A., Huntley, D., Montana, G., Cerase, A., Nesterova, T.B., and Brockdorff, N. (2010). Efficiency of Xist-mediated silencing on autosomes is linked to chromosomal domain

organisation. Epigenetics Chromatin *3*, 10.

Tavares, L., Dimitrova, E., Oxley, D., Webster, J., Poot, R., Demmers, J., Bezstarosti, K., Taylor, S., Ura, H., Koide, H., et al. (2012). RYBP-PRC1 Complexes Mediate H2A Ubiquitylation at Polycomb Target Sites Independently of PRC2 and H3K27me3. Cell *148*, 664–678.

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

Umlauf, D., Goto, Y., Cao, R., Cerqueira, F., Wagschal, A., Zhang, Y., and Feil, R. (2004). Imprinting along the Kcnq1 domain on mouse chromosome 7 involves repressive histone methylation and recruitment of Polycomb group complexes. Nat. Genet. *36*, 1296–1300.

Wada, Y., Ohta, Y., Xu, M., Tsutsumi, S., Minami, T., Inoue, K., Komura, D., Kitakami, J., Oshida, N., Papantonis, A., et al. (2009). A wave of nascent transcription on activated human genes. Proc. Natl. Acad. Sci. U. S. A. *106*, 18357–18361.

Wang, G., Yang, L., Grishin, D., Rios, X., Ye, L.Y., Hu, Y., Li, K., Zhang, D., Church, G.M., and Pu, W.T. (2017). Efficient, footprint-free human iPSC genome editing by consolidation of Cas9/CRISPR and piggyBac technologies. Nat. Protoc. *12*, 88–103.

Wang, W., Lin, C., Lu, D., Ning, Z., Cox, T., Melvin, D., Wang, X., Bradley, A., and Liu, P. (2008). Chromosomal transposition of PiggyBac in mouse embryonic stem cells.

Wani, A.H., Boettiger, A.N., Schorderet, P., Ergun, A., Münger, C., Sadreyev, R.I., Zhuang, X., Kingston, R.E., and Francis, N.J. (2016). Chromatin topology is coupled to Polycomb group protein subnuclear organization. Nat. Commun. *7*, 10291.

Wilson, M.H., Coates, C.J., and George, A.L. (2007). PiggyBac Transposon-mediated Gene Transfer in Human Cells. Mol. Ther. *15*, 139–145.

Woo, C.J., Kharchenko, P. V., Daheron, L., Park, P.J., and Kingston, R.E. (2010). A Region of the Human HOXD Cluster that Confers Polycomb-Group Responsiveness. Cell *140*, 99–110.

Wright, A.V., Nuñez, J.K., and Doudna, J.A. (2016). Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. Cell *164*, 29–44.

Wu, X., Johansen, J.V., and Helin, K. (2013). Fbxl10/Kdm2b Recruits Polycomb Repressive Complex 1 to CpG Islands and Regulates H2A Ubiquitylation. Mol. Cell *49*, 1134–1146.

Wu, X., Scott, D.A., Kriz, A.J., Chiu, A.C., Hsu, P.D., Dadon, D.B., Cheng, A.W., Trevino, A.E., Konermann, S., Chen, S., et al. (2014). Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. Nat. Biotechnol. *32*, 670–676.

Wutz, a, and Jaenisch, R. (2000). A shift from reversible to irreversible X inactivation is triggered during ES cell differentiation. Mol. Cell *5*, 695–705.

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.

Yusa, K., Rad, R., Takeda, J., and Bradley, A. (2009). Generation of transgene-free induced pluripotent mouse stem cells by the piggyBac transposon. Nat. Methods *6*, 363–369.

Zawistowski, J.S., Bevill, S.M., Goulet, D.R., Stuhlmiller, T.J., Beltran, A.S., Olivares-Quintero, J.F., Singh, D., Sciaky, N., Parker, J.S., Rashid, N.U., et al. (2017). Enhancer Remodeling during Adaptive Bypass to MEK Inhibition Is Attenuated by Pharmacologic Targeting of the P-TEFb Complex. Cancer Discov. *7*, 302–321.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Zhao, J., Sun, B.K., Erwin, J.A., Song, J.-J., and Lee, J.T. (2008). Polycomb Proteins Targeted by a Short Repeat RNA to the Mouse X Chromosome. Science (80-.). *322*, 750–756.

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. (2010). Genome-wide identification of polycomb-associated RNAs by RIP-seq. Mol. Cell *40*, 939–953.

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. *34*, 1279–1286.

Żylicz, J.J., Bousard, A., Žumer, K., Dossin, F., Mohammad, E., da Rocha, S.T., Schwalb, B., Syx, L., Dingli, F., Loew, D., et al. (2019). The Implication of Early Chromatin Changes in X Chromosome Inactivation. Cell *176*, 182-197.e23.