Long non-coding RNAs interact with PRC1 to impact Polycomb group protein recruitment and expression of Polycomb regulated genes

Long non-coding RNAs interact with PRC1 to impact Polycomb group protein recruitment and expression of Polycomb regulated genes

A dissertation presented

by

Mridula Kumari Ray

to

The Division of Medical Sciences

In partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Genetics and Genomics

Harvard Medical School

Boston, Massachusetts

December 2013

Dissertation Advisor: Professor Robert E Kingston Mridula Kumari Ray

Long non-coding RNAs interact with PRC1 to impact Polycomb group protein recruitment and expression of Polycomb regulated genes

Abstract

Long non-coding RNAs (lncRNAs) are increasingly recognized as important regulators of genomic processes and cellular specification. Many lncRNAs regulate chromatin by functionally impacting the epigenetic state through direct interactions with chromatin-modifying proteins. We developed a protocol to enrich for chromatin-lncRNA interactions and used this technique to identify several candidate lncRNAs that interact with the Polycomb group (PcG) proteins. Our immunoprecipitation protocol uses a crosslinked chromatin fraction as the input and employs stringent washes and cross-validation techniques to dramatically decrease mRNA signal (as a metric of transient interactions or false positives), and increase the dynamic range of conventional RNA immunoprecipitation protocols. Applying this protocol to the PRC1 component *Bmi1,* we have identified 11 PcG-interacting lncRNA candidates whose expression impacts the transcription of many other chromatin factors and PcG targets. We focus on knockdown of one lncRNA candidate, CAT7, which increases expression of several homeobox-containing transcription factors as well as chromatin interacting proteins, including Trithorax group proteins, Jumanji-domain containing proteins, and PcG-like proteins in HeLa cells. Consistent with the observed increase in gene expression, knockdown of CAT7 decreases PcG binding (*Suz12, H3K27me3* and *Bmi1*) at the promoter of the homeodomain protein *Mnx1*, located at the boundary of an adjacent gene desert. During early motor neuron differentiation from embryonic stem cells, knockdown of CAT7 is accompanied by changes in expression of master regulators of neuronal specification: increased upregulation *Mnx1*, upregulation of *Isl1*, and downregulation of *Irx3,* as well as changes in

expression to several other PcG-regulated targets. Overall, this protocol is the first of its kind to efficiently identify *de novo* interactions between the PcG proteins and lncRNAs which impact PcG binding or PcG target gene expression.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I was very lucky in my PhD to have the support and overwhelming encouragement of my sister, mother and father. I know they will always be there for me, and that knowledge alone is my source of strength and adventure.

I also am very grateful to have spent countless hours with my lab-big-sisters Sara Miller and Jesse Cochrane, who, after 5 billion hours in lab together, have shaped me both scientifically and as a person. I want to thank them for their tireless help, humor, and support, and for making my time in lab, and my 20's, genuinely fun.

I feel very lucky to know Jesse Engreitz, who always manages to make me laugh and to challenge me to be my best at the same time.

I also want to thank Matt Simon, who helped immensely in my training, taking much of his own time to teach me in the beginning of my PhD.

My friends, Mohamed Haji, Kavita Radhakrishnan, Carissa Milliman, Alex Will, Adriana Tajonar, and Amanda Gorton, all played very important roles in my life as well, as did my cousins Sid and Shikha Sabharwal. They taught me to balance working hard with enjoying my free time, and my life has been so much more enjoyable with them beside me.

Finally, I would also like to thank my mentor, Bob Kingston, for always encouraging me to take the high road, and always asking me to open my eyes a little wider so I could see the big picture.

INTRODUCTION

Polycomb-Group Proteins and Long Non-Coding RNAs Contribute to Epigenetic Control of Transcription and Cellular Identity

Cellular specification is an intricate process that involves many layers of genomic regulation. Transcription, as the direct output of the genome, is the foundation for establishing cellular identity. The contribution of transcription to cellular identity is evidenced by the diverse spectrum of RNAs present in any given cell type [1-3], including transcripts originating from both protein-coding and non-coding regions of the genome. However, transcriptional diversity cannot arise from the DNA sequence alone, since, from fertilization onwards, nearly every cell in the body contains essentially the same DNA. Regulation of when, where, and how much of a gene is transcribed must then, in part, be "epigenetic": independent of DNA sequence, easily modified, and heritable through cell division. Here, I review protein-based modes of epigenetic inheritance and also explore the role of non-protein coding transcripts in directing these processes.

Epigenetics

While genetic information is a means to retain information from our parents, epigenetic control allows the body to alter its usage of that information in response to the environment. In this way, environmental cues (such as diet, development or disease) can leave a molecular impression to impact gene expression, even after the environment has changed.

Epigenetic control of transcription involves several layers of regulation. As one means of epigenetic control, the genome is organized into repeated units of DNA spooled around a core set of proteins called the histone proteins. The regular inclusion of histone proteins throughout the DNA provides a ubiquitous platform for an additional layer of genomic regulation. The placement of histones on the DNA, the higher order organization of the histones in the nucleus, and physical modification of the histones or the DNA itself, can epigenetically influence gene expression.

Histones are present in eukaryotes and Achaea, and are the most highly conserved of all proteins. The vertebrate somatic histones (*H2A, H2B, H3,* and *H4,* or their variants) form heterodimers that together compose an octameric core, around which a nearly equivalent mass of DNA (146 basepairs) is wrapped; this structure is called the nucleosome. Small stretches of linker DNA between nucleosomes are associated with the histone protein *H1* in some silenced regions of the DNA. The nucleosome is itself the fundamental repeated unit of a more highly ordered structure, chromatin, which also includes tethered proteins and RNAs. Finally, chromatin is further arranged in the 3D space of the nucleus, having a highly regulated, but still fluid conformation. In effect, organizing the genome into chromatin not only provides a platform for regulation, but protects the DNA from damage and condenses it to fit in the nucleus.

Chromatin organization also influences molecular processes, such as transcriptional initiation and elongation [4-6], which require direct interaction of protein complexes with the DNA. Specifically, binding of the transcription factor TBP to the DNA upstream of the transcription start site (TSS) is required for transcriptional initiation by RNA Polymerase II. Nucleosomal depletion at the TSS, which is a hallmark of active eukaryotic genes, permits direct access of TBP to the DNA so that the pre-initiation complex may assemble [7-9]. Conversely, the presence of a nucleosome occluding the TSS is often a feature of silenced genes [10,11]. For example, studies in mouse tissue report that a wide array of liver-specific genes, such as Cytochrome P450 and Murinoglobin1, display "on" (depleted TSS) or "off" (occluded TSS) modes of nucleosomal occupancy in matched liver and brain samples, respectively[12]. Similarly to the TSS, other regulatory regions that are bound by transcription factors may also differ in nucleosomal occupancy to influence tissue specific "on" versus "off" states [13]. Such loci may include sites adjacent to the TSS, called promoters, or distal sites, called enhancers. During hematopoiesis, the collinearly regulated globin genes display

coordinated changes in nucleosomal occupancy at both individual promoters and at a shared enhancer site[14,15].

The rate of proper transcriptional elongation can also be influenced by "remodeling" the nucleosomes: sliding a nucleosome along the DNA, sterically altering the DNA/histone interactions, or subjecting the nucleosome to histone replacement or ejection [6,8]. *In vitro* transcription of nucleosomal arrays reveals that the presence of a nucleosome greatly slows the rate of transcription, as compared to naked DNA [16]. During elongation of highly transcribed genes, such as *Hsp70* in heat shock response, entire histone octamers may be rapidly ejected from the gene body to facilitate immediate access of the large (1 MegaDalton) RNA polymerase to the DNA [17]. However, only one histone pair (H2A/H2B) is ejected during elongation of moderately transcribed genes. This topologically limits the rate of elongation because the polymerase must travel along the DNA that is partially constrained by the remaining histones[18-20]. Inclusion of specific histone variants, such as *H3.3* in the gene body or *H2A.Z* at the -1 nucleosome, is also correlated with active transcription [21].

Many remodeling events, such as those as above, require breaking hundreds of points of contact between the DNA and the histones, and overcoming biases for nucleosome positioning which may be driven by DNA sequence [22]. These processes may be carried out in an ATP-dependent manner, as an active means of transcriptional regulation [8,9,23]. Mutations in ATP-dependent chromatin remodelers display an expansive range of effects, such as global misregulation of splicing [24], or widespread developmental effects, as in CHARGE syndrome [25].

Physical epigenetic marks on the chromatin comprise another form of epigenetic regulation. Such marks consist of either methylation of the DNA itself at CG dinucleotides ("CpG") or covalent post-translational modifications, including methylation or acetylation (among others), of the N-terminal tails of histones[26]. Histone N-termini are structurally disordered and protrude from the

core of the nucleosome, allowing access to enzymes or transcription factors that can modify or recognize specific histone residues. These modifications alter the sterics of DNA/histone interactions, the stacking or compaction of nucleosomes, and the binding of non-histone proteins to the chromatin [6]. Consequently, histone marks are correlated with and, in some instances, are necessary for, altered levels of gene expression [27-29].

Histone marks, as well as the incorporation of histone variants, are believed to have combinatorial effects on both small regions of the genome, such as promoters, as well as large "domains" of chromatin. One broadly defined domain, heterochromatin, was first described as the regions of the nucleus that exhibited intense staining by the intercalating dye DAPI. Heterochromatin corresponds to transcriptionally silenced, gene-poor regions of DNA, in contrast to its gene-rich, transcriptionally active counterpart, euchromatin. On the molecular level, heterochromatin is characterized by highly compacted nucleosomes, and the presence of the H3K9me3 mark of silencing, the transcription factor *HP1* and the linker histone *H1* [30]. Euchromatin is generally less compacted, though not all parts of euchromatin are actively transcribed, and not all heterochromatin is strictly silent.

Chromatin may also be classified into domains in both a functional manner: displaying interdependent levels of gene expression, and/or a physical manner: co-localizing in 3D space[31,32]. Such domains may be comprised of segments of DNA that are not necessarily contiguous. While the precise mechanisms for establishing or maintaining chromatin structure are not well understood, these processes are largely modulated by DNA sequence, transcription factor binding, and chromatin remodeling.

The organization of chromatin into physical domains is a pervasive mechanism for transcriptional regulation. Domains of "active" or "silenced" chromatin can extend in physical space to impact expression of seemingly unrelated, but proximal genes. In flies, an inverted translocation of the X

chromosome causes the pigment gene *white* to be adjacent to heterochromatin. The result is a variegated change in eye color in males: though genetically identical, only a portion of the mutant cells show a null (colorless) phenotype, caused by variable spreading of heterochromatin to silence the *white* gene [33,34]. This quintessential example highlights the profound influence of nuclear architecture and epigenetic boundaries on gene expression.

Similarly, in wild-type cells, broad euchromatic domains are further subdivided and modularly regulated. For example, a group of developmental proteins called the HOX genes are organized co-linearly on the chromosome. Spatially and temporally coordinated expression (or silencing) of HOX genes is essential for proper body patterning during development. The active and silenced domains are often segregated from each other by insulator proteins, such as the architectural, sequence-specific DNA binding protein, CTCF. Upon perturbation of the binding site for CTCF between active HOX genes and a silenced enhancer, the organism can no longer silence the enhancer region, leading to aberrant expression of HOX genes and developmental defects [35]. This example shows the relevance of chromatin organization and boundaries to gene expression.

An important aspect of epigenetic inheritance is stability through cell division. For a cell to divide, the DNA must be replicated and then split evenly into each daughter cell. During DNA replication, chromatin is dismantled, and most proteins (including histones) and RNAs are stripped from the DNA to allow access to the replication machinery [36]. Upon completion of replication and subsequent cell division, each cell must re-establish its former nuclear architecture to maintain cellular identity. Rather than initiating this process *de novo* with respect to each cell type, the cell employs a mechanism to "remember" the former epigenetic state. While this process is poorly understood, it is known that a small portion of certain transcription factors remain bound to the DNA, demarcating the epigenetic features and boundaries which underlie a cell's identity[37,38].

The process of establishing chromatin architecture from such boundaries is also unknown, but is influenced by spreading of genomic features within the retained boundaries and preservation of DNA methylation on the parent strand.

Dividing cells (such as stem cells) have both the ability to self-renew and to develop into terminally differentiated cells. A stem cell not only expresses many different genes compared to a terminally differentiated cell, but also differs in its epigenetic signatures and chromatin architecture across much of the genome[3]. A stem cell may divide and then alter its epigenetic state in order to differentiate, or it may maintain its epigenetic state to remain a stem cell [39]. In this way, maintenance of epigenetic marks and chromatin domains is critical to both establishing and preserving cellular identity, and is a defining characteristic of epigenetic inheritance.

While the importance of epigenetic regulation is apparent from its impact on genomic regulation and a number of disease states, many aspects of epigenetic regulation remain unknown. A major question is how epigenetic regulators coordinate with each other to execute targeted changes to transcription, and how chromatin is regulated in response to various biological stimuli or through biological processes, such as differentiation, cell division, and cancer.

The Polycomb Group Proteins

The Polycomb group (PcG) proteins are a prominent group of transcription-modulating proteins important for epigenetic maintenance of gene silencing. *Polycomb* was first defined as a dominant genetic mutation in flies causing aberrant expression of the gene *Scr*, resulting in the formation of extra pairs of sex combs on the second and third legs of male flies[40,41]. Further genetic and biochemical investigation showed that *Polycomb,* in complexes with a handful of other proteins collectively termed the PcG proteins, is critical for proper gene silencing of the axial

development proteins, the HOX genes. Not all trancriptionally inactive genes are silenced by PcG proteins; rather, PcG target genes are heavily enriched for developmental regulators[42,43]. Since their initial discovery, PcG proteins and their homologs have been identified throughout metazoans, regulating thousands of genomic targets in every cell type and playing critical roles in cancer, cell cycle, and most notably in embryonic development.

In mammals, PcG proteins are critical for regulating cell plasticity by silencing certain developmental factors so that others may activate properly. Remarkably, cells are unable to progress from the embryonic stem (ES) cell state without the PcG proteins, and will die upon differentiation [44,45]. In addition, many multipotent stem cells show accelerated differentiation upon PcG depletion [46]. PcG proteins may also play a role in maintaining the state of terminally differentiated cells. Ablation of PcG proteins in adult mouse neurons (via a conditional knockout) leads to progressive neurodegeneration, memory loss, and impaired mobility, a phenotype similar to Huntington's Disease [47]. These changes have been previously associated with derepression of homeobox-domain containing genes and developmental transcription factors, the characteristic targets of PcG proteins.

Mechanistic studies have shown that PcG proteins maintain transcriptional silencing in an epigenetic fashion; they modify chromatin architecture by compaction of polynucleosomes and covalent modification of histones. There are at least two major functional core complexes of PcG proteins: PRC1 and PRC2, though many subcomplexes are still being identified and functionally defined [48]. Broadly, PRC2 has histone-methyltransferase activity, conferring tri-methylation of lysine 27 of *H3* (H3K27me3) as a mark of silenced euchromatin. PRC1 has a binding preference for H3K27me3 nucleosomes [49] and is recruited to many (but not all) of the same sites as PRC2, possibly through an independent recruitment mechanism [42,50,51]. H3K27me3 is not necessary for all PcG-mediated silencing. In fact, PRC1 binding is sufficient for compaction of polynucleosomes

in vitro [52,53] , and correlates with gene silencing *in vivo*. A specific subcomplex of PRC1, potentially exclusive from the compaction complex, also ubiquitinates histone H2A at lysine 119[54]. However, the significance of this mark to silencing in mammals is unknown. While the core proteins of PRC1 and PRC2 are conserved between flies and mammals, notable differences in the catalytic activity of each component [52] and in PcG recruitment exist between the two species.

In both mammals and flies, PcG proteins form "bodies" or large 3D structures composed non-contiguous regions of the chromatin whose silencing is interdependent[55]. While the formation of PcG bodies is not well understood, fly PcG bound loci are proposed to scan the nucleus in *trans* for similarly bound sites, creating PcG bodies. PcG repression in *cis* is partially explained in flies by (non-PcG) transcription factor binding at PcG-target promoter regions. Briefly, functionally defined stretches of DNA, called Polycomb Response Elements (PREs) dock the PcG proteins to cause silencing of the adjacent chromatin [56]. Fly PREs are non-uniformly marked by combinations of binding motifs of various sequence-specific transcription factors which together recruit PcG proteins. However, in mammals, none of the sequence-specific transcription factors are conserved with the exception of *Pho/YY1,* whose binding is not sufficient to define a PRE. Though the first mammalian PREs have recently been discovered [57,58], pinpointing additional mammalian PREs is further obfuscated by the broad regions (relative to regions in flies) of PcG binding and H3K27me3. This difference highlights the possibility of distinct mechanisms between the species in recruitment or spreading of PcG proteins.

Targeting of mammalian PcG complexes remains elusive, and different modes of targeting may exist in different biological contexts. In ES cells, PRC1 and PRC2 complexes bind almost exclusively to unmethylated CpG dinucelotides and primarily target developmental regulators[39,59]. As development progresses, PcG repression is selectively lost at specific developmental genes in order to facilitate differentiation of the cell down a defined lineage

pathway. As seen during differentiation of insulin-producing beta cells[60], canonical PcG signaling may govern the transcriptional state. An array of classical PcG-regulated, CpG-rich, developmental targets become derepressed from the progenitor state, losing H3K27me3 as the beta cells mature. These targets are essential master regulators of both beta cell and neuronal differentiation, and strongly influence the transcriptionally circuitry. In fact, though beta cells and neurons originate from different tissues (endoderm versus ectoderm) which diverge at the earliest stages of development, the transcriptome of beta cells is more similar to neuronal cells than to other endoderm-derived tissues [60]. This process is reflective of the essential role of the PcG proteins in development and cellular specification.

Concurrently, during beta cell differentiation, PRC2 mediated silencing (at least as evidenced by H3K27me3 signatures) also occurs *de novo* at genes which are not canonical PcG targets. Beta cell progenitors express several non-developmental proteins, such as *SLC16A1* (insulin hypersecretion/monocarboxylic acid transport), that impede differentiation or survival of beta cells. However, these genes are transcriptionally silent and are enriched for H3K27me3 in mature beta cells[60]. Unlike the canonical PcG targets, these *de novo* silenced genes are not enriched for CpG islands, and are not silenced by H3K27me3/PcG proteins in ES cells, other tissues, or at earlier developmental stages. Together, these data reveal the complex nature of mammalian PcG regulation.

The targeting of PcG proteins at regions transitioning from silent to active (or the reverse) is not well understood. A class of activating proteins called the Trithorax group proteins (TrX) functionally antagonize the PcG proteins, binding at many of the same sites, but conferring a mark of initiation (H3K4me3) at promoters. Genes which contain both H3K4me3 and H3K27me3 at their promoters (e.g. bivalent domains) are considered to be in a transitory chromatin state, where they are poised for activation or may already be transcribed. Importantly, there are

several H3K4me3 histone methyl-transferases, but to date, PRC2, and specifically, its core component EZH2 (or occasionally, the closely related EZH1), are the only known H3K27me3 methyltransferases *in vivo*.

Several studies have also implicated a class of tudor-domain containing proteins in the transition from an active to a PcG-repressed state [61]. These proteins bind specifically to marks of active chromatin, such as H3K36me3 or H3K4me3, and form complexes with histone demethylases. Removal of such marks may allow the PRC2 proteins to bind and confer the H3K27me3 mark. Additionally, binding partners of PRC2, such as the inactive histone demethylase Jarid2, may also contribute to its binding specificity [48]. In one model of PcG regulation, PRC2 samples the genome to identify nucleosome-dense [62], primed chromatin and/or a lack of mRNA transcription [63,64], before stably binding. PRC1 binding may follow PRC2 binding, though is generally considered a form of more stable silencing and might employ independent recruitment mechanisms.

This model is nevertheless unsatisfying: protein-based mechanisms are often correlative, and currently are too broad to explain the complex mechanisms precisely governing PcG activity. Historically, proteins, such as the PcG proteins and their binding partners, were considered the readers, writers and erasers of epigenetic marks. However, recent evidence potentiates another, rapidly generated, class of epigenetic effectors that may modulate activity of protein complexes: namely, RNA molecules which are never translated into protein *(ncRNAs)*. Specifically, these molecules have been shown to interact with PcG proteins in the nucleus, and are proposed to play roles in PcG recruitment, spreading, and organization of the chromatin.

Long non-coding RNAs

Non-coding RNAs (ncRNAs) are increasingly recognized as regulators of cellular specification and many biological processes. Their contribution to transcriptional diversity is astounding: while at least 80% of the genome is transcribed, less than 2% of the genome is translated into protein [1]. The significance of the non-coding transcriptome to life is perhaps best exemplified by staggering developmental defects or lethality induced by depletion of any one of a multitude of ncRNAs, several of which will be reviewed below. However, identification of functional ncRNAs and the mechanisms of how ncRNAs execute their functions in the cell are widely unknown.

Many ncRNAs function in protein complexes, acting in capacities often ascribed to protein components: sequestering proteins from other targets, allosterically regulating protein binding domains, targeting proteins to the DNA or mRNA, or acting as scaffolding for protein complexes[65]. The mechanisms underlying these protein/ncRNA interactions often necessitate strict sequence motifs, length, or structural features in the ncRNAs, such as are found in ribosomal RNAs, transfer RNAs and short RNAs (siRNAs/piRNAs/miRNAs/snoRNAs /snRNAs).

Most ncRNA species, however, do not fall into these well-established classes and their function, structural motifs, and protein binding partners, if any, are not well defined. Amidst this ambiguity, a very broad classification of long ncRNAs (lncRNAs) has emerged in recent years. These lncRNAs annotations are usually derived from deep transcriptome sequencing data (RNAseq) from a variety of tissues or cancers, with a focus on intergenic, intronic or long antisense transcripts with low coding potential [2,66,67]. Typically, lncRNAs are over 200 basepairs in length, spliced, expressed at lower copy number per cell compared to mRNAs (with some exceptions), and frequently display cell-type specific expression. Lists of putative lncRNAs have also been bioinformatically curated based on the genomic features of actively transcribed chromatin at protein coding RNAs (mRNAs) and a few known lncRNAs. These predictions preferentially

consider regions of the genome with chromatin marks of initiation and elongation (H3K4me3 and H3K36me3 respectively), RNA polymerase II binding, capping and polyadenylation. Additionally, some predictions rely heavily on conservation: based on the idea that functional lncRNAs will have sequence-based or syntenic conservation, and supported by the observation that many lncRNAs are transcribed from pseudogenes [66]. Conversely, other predictions contend that because lncRNAs are largely repetitive and do not code for amino acids, they are hypoconserved (compared to protein-coding genes) with the exception of small stretches of evolutionarily pressured interaction domains [67].

LncRNAs have been implicated in cancer, development, sex determination and various diseases in the body. Knockout of the canonical lncRNA *Xist*, is embryonic lethal in female mammals, as developing female cells lacking *Xist* cannot balance expression of RNA from both X-chromosomes [68]. Similarly, depletion of the lncRNAs *Fendrr* and *Braveheart* cause embryonic-lethal defects in murine heart formation [69,70], and depletion of *HOTTIP* [71]can induce limb malformations. In zebrafish development, the lncRNAs *Cyrano* and *Megamind* lead to widespread developmental defects: notochord and *Neurod*-related defects, and brain/eye malformations, respectively [67]. Surprisingly, morpholino depletion of *Cyrano* and *Megamind* can be partially rescued by ectopic expression of the syntenic human transcripts. The syntenic transcripts are not well conserved with respect to sequence, with the exception of small, putative protein binding domains. A plethora of other lncRNAs has also been correlated with cancer prognoses and congenital diseases such as Brachydactyly [72], and many lncRNAs show overexpression in specific tissues or diseases *in vivo* [73]. However, understanding the mechanisms of these lncRNAs remains technically challenging.

While lncRNAs have been implicated in countless biological processes, there is almost no understanding of which features make a lncRNA biologically relevant [74]. On the contrary, there are some cases where lncRNA depletion has no apparent effect on the cell [75]. Rather, the

opening of the chromatin through the process of transcription might be necessary for gene regulation of certain loci, rather than the transcript itself. Alternatively, a transcript may simply be a byproduct of open chromatin, where spurious promoters become accessible. Notably, spurious, mis-spliced, or abortive transcripts, as well as introns of highly transcribed genes, often have very short half lives in the cell [76,77]. This indicates that there might be specific characteristics to mark which transcripts are retained. Furthermore, it is not known if and how lncRNAs avoid translation, particularly because a number of lncRNAs were identified because they have similar features or genomic signatures as mRNAs (such as polyA tails). Features such as secondary structure, length of the polyA tail (if any) or association with a specific ribosomal protein [78] or snRNA [79] may control RNA stability or translational potential, though several of these hypotheses remain speculative or disconnected from true causation. Such inquiries will be difficult to ascertain until the lncRNAs are more conclusively annotated and classified.

Many lncRNAs, such as *Xist*, *HOTTIP*, *Braveheart* or *Fenderr*, have been shown to execute their widespread functions by modulating transcription through direct interactions with chromatin proteins. lncRNAs are involved in a broad scope of chromatin processes, affecting expression of both specific loci and entire chromosomes, and organizing the formation of nuclear structures or domains. *Xist* interacts directly with the X chromosome in female cells, silencing the entire chromosome from which it is transcribed. *Xist* initially localizes to gene-rich regions on the chromatin in a seemingly sequence-independent, proximity-driven manner [80,81]. Namely, it binds to gene-rich regions proximal in 3-D space to its transcription site, and spreads to coat and silence nearly the entire X-chromosome. Antisense-blocking of the *Rep*C region of *Xist* prevents *Xist* from nucleating, and therefore stops spreading of *Xist* and inactivation of the X chromosome. This phenotype is relieved as the blocking-oligo is diluted through cell division.

lncRNAs have also evolved to regulate the expression of individual loci by interaction with the chromatin [65]. Some genes, such as the imprinted genes, require that only one allele is transcribed. However, unlike the X chromosome, which is inactivated at random early in development, the expression of the imprinted genes is determined by parental origin: DNA methylation of imprinted enhancer elements persists through gametogenesis to control gene expression in the progeny. lncRNA transcripts have been isolated from many imprinted genes, including the *Kcnq1ot1* [82], *Gtl2* [83], *Airn* [84], and *h19* [32] lncRNAs, and are necessary for silencing of the Kcnq1, Dlk1, Igf2r, and Igf2 loci, respectively. An exception is *h19* which is thought to be transcribed from an enhancer region, but whose expression impacts silencing of several distal regions. Generally, imprinted lncRNAs colocalize with the chromatin, may or may not be involved in antisense regulation, and are necessary for silencing of large (often >100kb), contiguous segments of the DNA.

Another class of lncRNAs is a *cis* acting antisense-derived lncRNAs. Two such lncRNAs, *ANRIL* [85] and *Evf2* [86], act as a switch to determine which gene is expressed from a co-regulated locus. These lncRNAs may function by recruiting silencing (or activating) factors cotranscriptionally: *ANRIL* balances the expression of *InK4A* and *Arf* to regulate cell cycle and senescence, and *Evf2* regulates the homeotic *Dlx5/6* locus in neural development. In both instances, the lncRNAs contain sequence that is antisense to the mRNA in the respective loci, but also contain regions necessary to recruit chromatin proteins. In comparison to imprinted loci, antisense regulated genes silence much smaller genomic regions, indicating potential differences in recruitment and/or spreading mechanisms.

Establishment of nuclear domains or structures may also be dependent on lncRNAs. Telomeric silencing and some instances of heterochromatin formation have been shown to be dependent on lncRNAs[87,88], as has the formation of nuclear paraspeckles via the highly abundant lncRNA,

NEAT1[89]. Additionally, the HOX-locus encoded transcript *HOTTIP* influences HOX genes up to nearly 40 kb away, but physically near *HOTTIP* in 3D space[71]. Proximity to *HOTTIP* RNA has been shown as necessary and sufficient to organize long range interactions and specifically impact gene expression. Finally, knockdown, knockout and DNA-FISH studies implicate *Kcnq1ot1* in control of the expression and chromatin architecture of the ~1 Mb Kcnq1 locus[82,90-93]. Transgenic expression of *Kcnq1ot1* is sufficient to bidirectionally silence flanking genes *in vivo*[82]. *Kcnq1ot1* also organizes chromatin so that its targets are in proximity of silenced genes to perinucleolar regions, presumably to facilitate silencing[91]. Mechanistically, *Kcnq1ot1* has been proposed to act like *Xist*, to concurrently mediate gene expression and chromatin architecture[90,91,93].

Relationship between PcG proteins and lncRNAs

As has been seen in plants, several lncRNAs have been shown to directly contribute to PcG recruitment or silencing in mammals. Among these lncRNAs are *Xist* (PRC2) [94], *Braveheart* (PRC2)[70], the HOX encoded transcript *HOTAIR* (PRC2) [95], *Fendrr* (PRC2) [69], *Kcnq1ot1* (PRC2) [82], *Gtl2* (PRC2) [83] and *ANRIL* (PRC1 and PRC2) [85,96], though notably, these genes are not conserved in flies, or even necessarily between mouse and human. Depletion of these lncRNAs culminates a loss of silencing of the respective target loci (ref) and/or death.

Direct interactions of lncRNAs and the PcG proteins were primarily found through a protocol called RNA immunoprecipitation (RIP) or the closely related UV-crosslinked RIP (CLIP), where lncRNAs are pulled down via a protein interactor [97]. Several of the above studies have been supplemented by gel shift/EMSA (electric mobility shift assays), or by studies where PcG proteins bind to ectopic lncRNAs in nuclear lysate. However, such assays must be revisited, as recent

evidence suggests that the PRC2 protein, EZH2, binding strongly to RNAs without clear sequence specificity *in vivo* and *in vitro* [54,64]. Additional evidence to support these interactions comes from perturbation data, showing knockdown of a candidate impacts transcription of PcG target genes and PcG localization to target loci, such as for *Braveheart*, *HOTAIR,* or *Gtl2*. Similarly, both Alu-repeat deletion and competitive blocking interactions of the *ANRIL* transcript [49,98], where complementary oligos hybridized to *ANRIL* at putative PcG/RNA-binding sites, yield changes in gene expression and loss of PcG binding to the RNA and the regulated Ink4a/Arf DNA locus . Complementary mutations to PcG proteins (CBX7) showed a similar result. Finally, RNA and DNA FISH data reveal that in the absence of the PRC2 (e.g. Eed knockout), several genes in the Kcnq1 locus lose their silencing [91].

Perhaps the most extensive mechanistic studies on lncRNA/PcG interactions validate interactions between the PcG proteins and *Xist*. In a set of experiments, *Xist* was specifically pulled down via antisense oligos, alongside its associated chromatin [80,81]. These studies showed that PRC2 is recruited to the X chromosome in direct proportion to *Xist* binding, consistent with the hypothesis that *Xist* mediates PRC2 binding along the inactive X. EMSA and deletion analysis have also implicated the A-rich repeat *RepA* of *Xist*, as a region of PcG interaction [94]. *RepA* is essential for silencing and spatial organization of genes on the inactive X chromosome. Mutation or deletion of *RepA* leads to a loss of the *Xist*/EZH2 binding *in vitro*, and a loss of *Xist* binding and silencing at genic regions along the X-chromosome *in vivo*.

Finally, the PcG proteins have been shown to interact with components of the RNAi machinery and the RNA helicase, MOV10 [99]. Perturbation of MOV10 causes imbalances in INK4a/Arf expression, and is speculated to directly impact *ANRIL* functionality. Together, these data suggest a versatile relationship between many lncRNAs and the PcG proteins in gene silencing.

In addition to lncRNAs that bind to the PcG proteins, short (50-200nt) double-hairpin ncRNAs also are also transcribed from binding sites of PRC1 and PRC2 proteins, and bind the PRC2 protein Suz12 as assessed by RIP and EMSA [100]. These short ncRNAs are transcribed from sites of CpG rich regions at the 5' end of many PcG targets, and are often accompanied by paused RNA Polymerase II. The hairpins are thought to work upstream of PcG silencing, and might serve to fine-tune PcG proteins by recruiting them to target gene promoters, or to act as scaffolding for complex assembly.

Many intergenic regions which show changes in PcG binding during development are also transcribed [101]. The genomic boundaries of such lncRNA transcripts precisely coincide with conserved regions of Suz12 binding and/or H3K27me3 (at some developmental point), though the transcribed regions are often CpG-poor. Expression of these lncRNAs may either coincide with or oppose PcG binding and H3K27me3. Knockdown of several such lncRNA transcripts generated from PcG sites affects transcription of both flanking (*cis)* and distant (*trans*) PcG regulated genes (ref) in differentiating mouse neural precursor cells. Though a direct interaction of these lncRNA transcripts with the PcG proteins has not been thoroughly investigated, these data support the role of ncRNAs as major players in the transcriptional circuitry, particularly at PcG-regulated genes essential for embryonic development.

While many PcG interacting lncRNAs, such as *Xist* and *Kcnq1ot1*, are proposed to modify chromatin structure over long ranges in *cis*, the first candidate *trans* acting lncRNA has also been identified: the HOX encoded transcript, *HOTAIR* [95]. Expression of *HOTAIR* is important for PcG mediated silencing of a HOX gene on an entirely different chromosome. However, the low-expression of lncRNAs, such as *HOTAIR* or *HOTTIP*, raises mechanistic questions of how lncRNAs could locate their genomic targets in the nucleus.

The possibility of lncRNAs organized with both distant and nearby chromatin, such as is the case for the (highly expressed) *NEAT1*-dependent formation of paraspeckles [89], provides an attractive mechanistic hypothesis that could support the existence of *trans*-acting, low abundance lncRNAs. In the context of PcG proteins, lncRNAs might bind PcG proteins as they sample the genome, and tether, scaffold or recruit the PcG proteins to PcG bodies and *trans* loci. In this way, similar to *Xist* and *Kcnq1ot1* activity in *cis*, even low-expression, *trans* acting lncRNAs could micro-organize PcG activity and chromatin structure.

One lncRNA which controls expression in a colocalized region of the genome is the transcript from the *CISTR-ACT* locus [72]. This lncRNA is upregulated in patients exhibiting certain forms of Brachydactyly. While the *CISTR-ACT* transcript has not been shown to interact directly with the PcG proteins, overexpression of the transcript causes changes in PcG-regulated genes which are spatially co-localized (both genes in *trans* and in *cis)* with CISTR-ACT. This culminates in widespread changes in EZH2 binding and gene expression of developmental targets.

Several individual PcG proteins have shown *in vitro* binding to RNAs through gel shifts. The EZH2, EZH2 with EED, and Suz12 components of PRC2 have each shown binding to various RNAs. However, recent studies have demonstrated that EZH2 binds many RNAs *in vitro* and *in vivo* without strong sequence specificity [63,64]. Notably, structural conservation is very difficult to predict, and was not well accounted for in these studies.

While PRC2 binds promiscuously to RNAs around the genome, lncRNA function in the specific setting of PRC1-mediated stable silencing and/or compaction is relatively unexplored. Several PRC1 proteins, such as the chromodomain of CBX7 [49]and the Phe-Cys-Ser (FCS) domain of Polyhomeotic [102], have shown RNA binding without sequence specificity *in vitro*. Mutated FCS of Polyhomeotic in flies leads to lower levels of repression at an array of PcG targets *in vivo*. In the case of CBX7, the compaction subunit in mammalian PRC1, the chromodomain has shown both *in*

vitro binding to ssRNA, and to a lower extent dsRNA, as well as a minor affinity for dsDNA. Notably, the chromodomain also shows high affinity for H3K27me3 nucleosomes *in vitro*, and as suggested by immnoflourescence and ChIP studies, *in vivo*[54,85,103]. Mutation of CBX7 leads to decreased binding and silencing of *ANRIL in vitro* and *in vivo* respectively[49]. However, mutations to PcG proteins are often difficult to interpret as they effect wide-spread changes in the chromatin landscape. An underlying question remains as to how the PcG proteins bind specifically to RNA *in vivo*.

While there is an ever-growing body of literature suggesting interactions between lncRNAs and the PcG proteins, the precise nature of these interactions is relatively unknown. Namely, mechanistic studies and identification of lncRNA interactors are hindered by the high incidence of non-specific binding between PcG proteins and RNA, a lack of understanding of how the protein complexes specifically recognize partner lncRNAs, and the uncertainty of which PcG directly bind RNA *in vivo*. In my study, I developed a protocol that identifies novel, non-random lncRNA interactions with chromatin proteins, across a large range of transcript expression. This protocol does not require knowledge of which PcG protein(s) directly bind the lncRNA, or are necessary for binding specificity. By employing cross-validation and stringent washes, the protocol greatly reduces mRNA noise or transient interactions. I used my protocol to find lncRNAs that bind the PcG proteins in the context of stable silencing by the PRC1 complex, and found that a majority of the candidates assayed show widespread changes to the PcG-regulated transcriptional gene network upon siRNA knockdown. Finally, I also found that depletion of one candidate, *CAT7,* causes loss of PcG binding at the promoter of an upregulated gene *Mnx1*. Lastly, I showed depletion of *CAT7* also induces differential expression of several PcG-regulated master regulators of neural/pancreatic beta development during motor neuron differentiation from ES cells.

References

1 Mattick, J. Non-coding RNAs: the architects of eukaryotic complexity. *EMBO reports* **2**, 986-991, doi:10.1093/embo-reports/kve230 (2001).

2 Cabili, M. *et al*. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & development* **25**, 1915-1927, doi:10.1101/gad.17446611 (2011).

3 Consortium, E. P. *et al*. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

4 Knezetic, J. & Luse, D. The presence of nucleosomes on a DNA template prevents initiation by RNA polymerase II in vitro. *Cell* **45**, 95-104, doi:10.1016/0092-8674(86)90541-6 (1986).

5 Lorch, Y., LaPointe, J. & Kornberg, R. Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones. *Cell* **49**, 203-210, doi:10.1016/0092-8674(87)90561-7 (1987).

6 Workman, J. & Kingston, R. Alteration of nucleosome structure as a mechanism of transcriptional regulation. *Annual review of biochemistry* **67**, 545-579, doi:10.1146/annurev.biochem.67.1.545 (1998).

7 Felsenfeld, G. Chromatin as an essential part of the transcriptional mechanism. *Nature* **355**, 219-224, doi:10.1038/355219a0 (1992).

8 Kwon, H., Imbalzano, A., Khavari, P., Kingston, R. & Green, M. Nucleosome disruption and enhancement of activator binding by a human SW1/SNF complex. *Nature* **370**, 477-481, doi:10.1038/370477a0 (1994).

9 Imbalzano, A., Kwon, H., Green, M. & Kingston, R. Facilitated binding of TATA-binding protein to nucleosomal DNA. *Nature* **370**, 481-485, doi:10.1038/370481a0 (1994).

10 Yuan, G.-C. *et al*. Genome-scale identification of nucleosome positions in S. cerevisiae. *Science (New York, N.Y.)* **309**, 626-630, doi:10.1126/science.1112178 (2005).

11 Ercan, S. & Simpson, R. Global chromatin structure of 45,000 base pairs of chromosome III in a- and alpha-cell yeast and during mating-type switching. *Molecular and Cellular Biology* **24**, 10026-10035, doi:10.1128/mcb.24.22.10026-10035.2004 (2004).

12 Bargaje, R. *et al*. Proximity of H2A.Z containing nucleosome to the transcription start site influences gene expression levels in the mammalian liver and brain. *Nucleic Acids Research* **40**, 8965-8978, doi:10.1093/nar/gks665 (2012).

13 Hu, G. *et al*. Regulation of nucleosome landscape and transcription factor targeting at tissue-specific enhancers by BRG1. *Genome research* **21**, 1650-1658, doi:10.1101/gr.121145.111 (2011).

14 Kim, S.-I., Bresnick, E. & Bultman, S. BRG1 directly regulates nucleosome structure and chromatin looping of the alpha globin locus to activate transcription. *Nucleic Acids Research* **37**, 6019-6027, doi:10.1093/nar/gkp677 (2009).

15 Onishi, Y. & Kiyama, R. Enhancer activity of HS2 of the human beta-LCR is modulated by distance from the key nucleosome. *Nucleic Acids Research* **29**, 3448-3457, doi:10.1093/nar/29.16.3448 (2001).

16 Izban, M. & Luse, D. Factor-stimulated RNA polymerase II transcribes at physiological elongation rates on naked DNA but very poorly on chromatin templates. *The Journal of biological chemistry* **267**, 13647-13655 (1992).

17 Petesch, S. & Lis, J. Rapid, transcription-independent loss of nucleosomes over a large chromatin domain at Hsp70 loci. *Cell* **134**, 74-84, doi:10.1016/j.cell.2008.05.029 (2008).

18 Studitsky, V., Clark, D. & Felsenfeld, G. A histone octamer can step around a transcribing polymerase without leaving the template. *Cell* **76**, 371-382, doi:10.1016/0092-8674(94)90343-3 (1994).

19 Bednar, J., Studitsky, V., Grigoryev, S., Felsenfeld, G. & Woodcock, C. The nature of the nucleosomal barrier to transcription: direct observation of paused intermediates by electron cryomicroscopy. *Molecular cell* **4**, 377-386, doi:10.1016/s1097-2765(00)80339-1 (1999).

20 Felsenfeld, G., Clark, D. & Studitsky, V. Transcription through nucleosomes. *Biophysical chemistry* **86**, 231-237, doi:10.1016/s0301-4622(00)00134-4 (2000).

21 Jin, C. *et al*. H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. *Nature genetics* **41**, 941-945, doi:10.1038/ng.409 (2009).

22 Segal, E. *et al*. A genomic code for nucleosome positioning. *Nature* **442**, 772-778, doi:10.1038/nature04979 (2006).

23 Côté, J., Quinn, J., Workman, J. & Peterson, C. Stimulation of GAL4 derivative binding to nucleosomal DNA by the yeast SWI/SNF complex. *Science (New York, N.Y.)* **265**, 53-60, doi:10.1126/science.8016655 (1994).

24 Zraly, C. & Dingwall, A. The chromatin remodeling and mRNA splicing functions of the Brahma (SWI/SNF) complex are mediated by the SNR1/SNF5 regulatory subunit. *Nucleic Acids Research* **40**, 5975-5987, doi:10.1093/nar/gks288 (2012).

25 Vissers, L. *et al*. Mutations in a new member of the chromodomain gene family cause CHARGE syndrome. *Nature genetics* **36**, 955-957, doi:10.1038/ng1407 (2004).

26 Goldberg, A., Allis, C. & Bernstein, E. Epigenetics: a landscape takes shape. *Cell* **128**, 635-638, doi:10.1016/j.cell.2007.02.006 (2007).

27 Braunstein, M., Rose, A. B., Holmes, S. G., Allis, C. D. & Broach, J. R. Transcriptional silencing in yeast is associated with reduced nucleosome acetylation. *Genes & development* **7**, doi:10.1101/gad.7.4.592 (1993).

28 Jenuwein, T. & Allis, C. Translating the histone code. *Science (New York, N.Y.)* **293**, 1074-1080, doi:10.1126/science.1063127 (2001).

29 Pengelly, A., Copur, Ö., Jäckle, H., Herzig, A. & Müller, J. A histone mutant reproduces the phenotype caused by loss of histone-modifying factor Polycomb. *Science (New York, N.Y.)* **339**, 698-699, doi:10.1126/science.1231382 (2013).

30 Nielsen, A. *et al*. Heterochromatin formation in mammalian cells: interaction between histones and HP1 proteins. *Molecular cell* **7**, 729-739, doi:10.1016/s1097-2765(01)00218-0 (2001).

31 Schneider, R. & Grosschedl, R. Dynamics and interplay of nuclear architecture, genome organization, and gene expression. *Genes & development* **21**, 3027-3043, doi:10.1101/gad.1604607 (2007).

32 Holwerda, S. & de Laat, W. Chromatin loops, gene positioning, and gene expression. *Frontiers in genetics* **3**, 217, doi:10.3389/fgene.2012.00217 (2012).

33 Mueller, H. J. Types of visible variation induced by X-rays in Drosophila. *Journal of Genetics* **22**, 35 (1930).

34 Henikoff, S. Position-effect variegation after 60 years. *Trends in genetics : TIG* **6**, 422-426, doi:10.1016/0168-9525(90)90304-o (1990).

35 Holohan, E. *et al*. CTCF genomic binding sites in Drosophila and the organisation of the bithorax complex. *PLoS genetics* **3**, doi:10.1371/journal.pgen.0030112 (2007).

36 Budhavarapu, V., Chavez, M. & Tyler, J. How is epigenetic information maintained through DNA replication? *Epigenetics & chromatin* **6**, 32, doi:10.1186/1756-8935-6-32 (2013).

37 Alabert, C. & Groth, A. Chromatin replication and epigenome maintenance. *Nature reviews. Molecular cell biology* **13**, 153-167, doi:10.1038/nrm3288 (2012).

38 Francis, N., Follmer, N., Simon, M., Aghia, G. & Butler, J. Polycomb proteins remain bound to chromatin and DNA during DNA replication in vitro. *Cell* **137**, 110-122, doi:10.1016/j.cell.2009.02.017 (2009).

39 Gifford, C. *et al*. Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. *Cell* **153**, 1149-1163, doi:10.1016/j.cell.2013.04.037 (2013).

40 Lewis, E. A gene complex controlling segmentation in Drosophila. *Nature* **276**, 565-570, doi:10.1038/276565a0 (1978).

41 Lewis, P. H. Melanogaster-New Mutants: Report of Pamela H. Lewis. (1947).

42 Boyer, L. *et al*. Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* **441**, 349-353, doi:10.1038/nature04733 (2006).

43 Lee, T. *et al*. Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125**, 301-313, doi:10.1016/j.cell.2006.02.043 (2006).

44 Pasini, D., Bracken, A., Hansen, J., Capillo, M. & Helin, K. The polycomb group protein Suz12 is required for embryonic stem cell differentiation. *Molecular and Cellular Biology* **27**, 3769-3779, doi:10.1128/mcb.01432-06 (2007).

45 Chamberlain, S., Yee, D. & Magnuson, T. Polycomb repressive complex 2 is dispensable for maintenance of embryonic stem cell pluripotency. *Stem cells (Dayton, Ohio)* **26**, 1496-1505, doi:10.1634/stemcells.2008-0102 (2008).

46 Valk-Lingbeek, M., Bruggeman, S. & van Lohuizen, M. Stem cells and cancer; the polycomb connection. *Cell* **118**, 409-418, doi:10.1016/j.cell.2004.08.005 (2004).

47 Seong, I. *et al*. Huntingtin facilitates polycomb repressive complex 2. *Human molecular genetics* **19**, 573-583, doi:10.1093/hmg/ddp524 (2010).

48 Simon, J. A. & Kingston, R. E. Occupying Chromatin: Polycomb Mechanisms for Getting to Genomic Targets, Stopping Transcriptional Traffic, and Staying Put. *Molecular cell* **49**, doi:10.1016/j.molcel.2013.02.013 (2013).

49 Bernstein, E. *et al*. Mouse polycomb proteins bind differentially to methylated histone H3 and RNA and are enriched in facultative heterochromatin. *Molecular and Cellular Biology* **26**, 2560-2569, doi:10.1128/mcb.26.7.2560-2569.2006 (2006).

50 Oren, R. *et al*. Combinatorial Patterning of Chromatin Regulators Uncovered by Genome-wide Location Analysis in Human Cells. *Cell* **147**, doi:10.1016/j.cell.2011.09.057 (2011).

51 Tavares, L. *et al*. RYBP-PRC1 complexes mediate H2A ubiquitylation at polycomb target sites independently of PRC2 and H3K27me3. *Cell* **148**, 664-678, doi:10.1016/j.cell.2011.12.029 (2012).

52 Grau, D. *et al*. Compaction of chromatin by diverse Polycomb group proteins requires localized regions of high charge. *Genes & development* **25**, 2210-2221, doi:10.1101/gad.17288211 (2011).

53 Francis, N. J. Chromatin Compaction by a Polycomb Group Protein Complex. *Science* **306**, doi:10.1126/science.1100576 (2004).

54 Gao, Z. *et al*. PCGF homologs, CBX proteins, and RYBP define functionally distinct PRC1 family complexes. *Molecular cell* **45**, 344-356, doi:10.1016/j.molcel.2012.01.002 (2012).

55 Bantignies, F. & Cavalli, G. Polycomb group proteins: repression in 3D. *Trends in Genetics* **27**, doi:10.1016/j.tig.2011.06.008 (2011).

56 Ringrose, L. & Paro, R. Polycomb/Trithorax response elements and epigenetic memory of cell identity. *Development (Cambridge, England)* **134**, 223-232, doi:10.1242/dev.02723 (2007).

57 Vasanthi, D., Nagabhushan, A., Matharu, N. & Mishra, R. A functionally conserved Polycomb response element from mouse HoxD complex responds to heterochromatin factors. *Scientific reports* **3**, 3011, doi:10.1038/srep03011 (2013).

58 Woo, C., Kharchenko, P., Daheron, L., Park, P. & Kingston, R. A region of the human HOXD cluster that confers polycomb-group responsiveness. *Cell* **140**, 99-110, doi:10.1016/j.cell.2009.12.022 (2010).

59 Deaton, A. & Bird, A. CpG islands and the regulation of transcription. *Genes & development* **25**, 1010-1022, doi:10.1101/gad.2037511 (2011).

60 van Arensbergen, J. *et al*. Derepression of Polycomb targets during pancreatic organogenesis allows insulin-producing beta-cells to adopt a neural gene activity program. *Genome research* **20**, 722-732, doi:10.1101/gr.101709.109 (2010).

61 Ballaré, C. *et al*. Phf19 links methylated Lys36 of histone H3 to regulation of Polycomb activity. *Nature structural & molecular biology* **19**, 1257-1265, doi:10.1038/nsmb.2434 (2012).

62 Yuan, W. *et al*. Dense Chromatin Activates Polycomb Repressive Complex 2 to Regulate H3 Lysine 27 Methylation. *Science* **337**, doi:10.1126/science.1225237 (2012).

63 Davidovich, C., Zheng, L., Goodrich, K. & Cech, T. Promiscuous RNA binding by Polycomb repressive complex 2. *Nature structural & molecular biology* **20**, 1250-1257, doi:10.1038/nsmb.2679 (2013).

64 Kaneko, S., Son, J., Shen, S., Reinberg, D. & Bonasio, R. PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells. *Nature structural & molecular biology* **20**, 1258-1264, doi:10.1038/nsmb.2700 (2013).

65 Mercer, T. & Mattick, J. Structure and function of long noncoding RNAs in epigenetic regulation. *Nature structural & molecular biology* **20**, 300-307, doi:10.1038/nsmb.2480 (2013).

66 Guttman, M. *et al*. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223-227, doi:10.1038/nature07672 (2009).

67 Ulitsky, I., Shkumatava, A., Jan, C., Sive, H. & Bartel, D. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147**, 1537-1550, doi:10.1016/j.cell.2011.11.055 (2011).

68 Marahrens, Y., Panning, B., Dausman, J., Strauss, W. & Jaenisch, R. Xist-deficient mice are defective in dosage compensation but not spermatogenesis. *Genes & development* **11**, 156-166, doi:10.1101/gad.11.2.156 (1997).

69 Grote, P. *et al*. The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. *Developmental cell* **24**, 206-214, doi:10.1016/j.devcel.2012.12.012 (2013).

70 Klattenhoff, C. *et al*. Braveheart, a long noncoding RNA required for cardiovascular lineage commitment. *Cell* **152**, 570-583, doi:10.1016/j.cell.2013.01.003 (2013).

71 Wang, K. *et al*. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472**, 120-124, doi:10.1038/nature09819 (2011).

72 Philipp, G. M. *et al*. A misplaced lncRNA causes brachydactyly in humans. *Journal of Clinical Investigation* **122**, doi:10.1172/jci65508 (2012).

73 Wapinski, O. & Chang, H. Long noncoding RNAs and human disease. *Trends in cell biology* **21**, 354-361, doi:10.1016/j.tcb.2011.04.001 (2011).

74 Graur, D. *et al*. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome biology and evolution* **5**, 578-590, doi:10.1093/gbe/evt028 (2013).

75 Pease, B., Borges, A. & Bender, W. Non-Coding RNAs of the Ultrabithorax Domain of the Drosophila Bithorax Complex. *Genetics*, doi:10.1534/genetics.113.155036 (2013).

76 Clement, J., Qian, L., Kaplinsky, N. & Wilkinson, M. The stability and fate of a spliced intron from vertebrate cells. *RNA (New York, N.Y.)* **5**, 206-220, doi:10.1017/s1355838299981190 (1999).

77 Bashirullah, A., Cooperstock, R. & Lipshitz, H. Spatial and temporal control of RNA stability. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 7025-7028, doi:10.1073/pnas.111145698 (2001).

78 Kondrashov, N. *et al*. Ribosome-mediated specificity in Hox mRNA translation and vertebrate tissue patterning. *Cell* **145**, 383-397, doi:10.1016/j.cell.2011.03.028 (2011).

79 Almada, A., Wu, X., Kriz, A., Burge, C. & Sharp, P. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* **499**, 360-363, doi:10.1038/nature12349 (2013).

80 Engreitz, J. M. *et al*. The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science* **341** (2013).

81 Simon, M. *et al*. High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. *Nature*, doi:10.1038/nature12719 (2013).

82 Pandey, R. *et al*. Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Molecular cell* **32**, 232-246, doi:10.1016/j.molcel.2008.08.022 (2008).

83 Zhao, J. *et al*. Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Molecular cell* **40**, 939-953, doi:10.1016/j.molcel.2010.12.011 (2010).

84 Stricker, S. *et al*. Silencing and transcriptional properties of the imprinted Airn ncRNA are independent of the endogenous promoter. *The EMBO journal* **27**, 3116-3128, doi:10.1038/emboj.2008.239 (2008).

85 Kyoko, L. Y. *et al*. Molecular Interplay of the Noncoding RNA ANRIL and Methylated Histone H3 Lysine 27 by Polycomb CBX7 in Transcriptional Silencing of INK4a. *Molecular cell* **38**, doi:10.1016/j.molcel.2010.03.021 (2010).

86 Dinger, M. *et al*. Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome research* **18**, 1433-1445, doi:10.1101/gr.078378.108 (2008).

87 Bernstein, E. & Allis, C. RNA meets chromatin. *Genes & development* **19**, 1635-1655, doi:10.1101/gad.1324305 (2005).

88 Maison, C. *et al*. Higher-order structure in pericentric heterochromatin involves a distinct pattern of histone modification and an RNA component. *Nature genetics* **30**, 329-334, doi:10.1038/ng843 (2002).

89 Mao, Y., Zhang, B. & Spector, D. Biogenesis and function of nuclear bodies. *Trends in genetics : TIG* **27**, 295-306, doi:10.1016/j.tig.2011.05.006 (2011).

90 Higashimoto, K., Soejima, H., Saito, T., Okumura, K. & Mukai, T. Imprinting disruption of the CDKN1C/KCNQ1OT1 domain: the molecular mechanisms causing Beckwith-Wiedemann syndrome and cancer. *Cytogenetic and genome research* **113**, 306-312 (2006).

91 Redrup, L. *et al*. The long noncoding RNA Kcnq1ot1 organises a lineage-specific nuclear domain for epigenetic gene silencing. *Development* **136**, 525-530 (2009).

92 Mohammad, F., Mondal, T., Guseva, N., Pandey, G. & Kanduri, C. Kcnq1ot1 noncoding RNA mediates transcriptional gene silencing by interacting with Dnmt1. *Development (Cambridge, England)* **137**, 2493-2499, doi:10.1242/dev.048181 (2010).

93 Chandrasekhar, K. Kcnq1ot1: A chromatin regulatory RNA. *Seminars in Cell & Developmental Biology* **22**, doi:10.1016/j.semcdb.2011.02.020 (2011).

94 Zhao, J., Sun, B., Erwin, J., Song, J.-J. & Lee, J. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science (New York, N.Y.)* **322**, 750-756, doi:10.1126/science.1163045 (2008).

95 Rinn, J. *et al*. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**, 1311-1323, doi:10.1016/j.cell.2007.05.022 (2007).

96 Kotake, Y. *et al*. Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15(INK4B) tumor suppressor gene. *Oncogene* **30**, 1956-1962, doi:10.1038/onc.2010.568 (2011).

97 Brockdorff, N. Noncoding RNA and Polycomb recruitment. *RNA (New York, N.Y.)* **19**, 429-442, doi:10.1261/rna.037598.112 (2013).

98 Holdt, L. *et al*. Alu elements in ANRIL non-coding RNA at chromosome 9p21 modulate atherogenic cell functions through trans-regulation of gene networks. *PLoS genetics* **9**, doi:10.1371/journal.pgen.1003588 (2013).

99 El Messaoudi-Aubert, S. *et al*. Role for the MOV10 RNA helicase in polycomb-mediated repression of the INK4a tumor suppressor. *Nature structural & molecular biology* **17**, 862-868, doi:10.1038/nsmb.1824 (2010).

100 Kanhere, A. *et al*. Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2. *Molecular cell* **38**, 675-688, doi:10.1016/j.molcel.2010.03.019 (2010).

101 Hekimoglu-Balkan, B., Aszodi, A., Heinen, R., Jaritz, M. & Ringrose, L. Intergenic Polycomb target sites are dynamically marked by non-coding transcription during lineage commitment. *RNA biology* **9**, 314-325, doi:10.4161/rna.19102 (2012).

102 Wang, R. *et al*. Identification of nucleic acid binding residues in the FCS domain of the polycomb group protein polyhomeotic. *Biochemistry* **50**, 4998-5007, doi:10.1021/bi101487s (2011).

103 Morey, L. *et al*. Nonoverlapping functions of the Polycomb group Cbx family of proteins in embryonic stem cells. *Cell stem cell* **10**, 47-62, doi:10.1016/j.stem.2011.12.006 (2012).

CHAPTER 2

A Technology to Isolate Chromatin Associated Transcripts Reveals a Class of PRC1-interacting lncRNAs

CONTRIBUTIONS

All work presented was carried out by Mridula Kumari Ray, except the sequence alignment and peak analysis by Yanqun Wang. Both MKR and YW contributed to data analysis.

ABSTRACT

Long non-coding RNAs (lncRNAs) are increasingly recognized as important regulators of genomic processes and cellular specification. Many lncRNAs are hypothesized to regulate chromatin by functionally impacting the epigenetic state through interactions with chromatin-modifying proteins. Recently, numerous lncRNAs have been reported to play roles in the activity or recruitment of epigenetic factors such as the Polycomb group (PcG) proteins, to genomic sites. However, identification and functional validation of chromatin-interacting RNAs are technically challenging with respect to distinguishing true RNA interactors from artifacts. In order to identify new lncRNAs that interact with PcG-bound chromatin, we developed an immunoprecipitation protocol which dramatically decreases mRNA noise (as a metric of false positives), and increases the dynamic range of conventional RNA immunoprecipitation (RIP). Namely, we purified chromatin away from free nucleic acids and free proteins, performed an immunoprecipitation on the chromatin, and applied stringent washes geared at both RNA and protein specificity. We then applied this protocol to the PcG protein Bmi1 to generate a list of novel candidate lncRNAs interactors, including the functionally-elusive *RepE* region of *Xist*. Analyzing these candidates, we found that PRC1 putatively binds a class of nuclearly localized lncRNAs that show tissue-specificity in the body, and which may contain tandem repeats, possibly as structural elements.

Introduction

Interactions between lncRNAs and chromatin proteins, such as the PcG proteins, have been identified *in vivo* by a technique called RNA immunoprecipitation (RIP)[1]. In canonical RIP and its variations, cells may or may not be lightly crosslinked by formaldehyde or UV light, and RNAs are co-precipitated with a protein and sequenced. RIP conditions are generally optimized for protein/protein stability and specificity; however, these are precisely the conditions which promote non-specific RNA/RNA or RNA/protein interactions[2,3]. Such artifacts arise upon nuclear shearing, when distal RNAs are brought together and hybridize to one another via small stretches of complementarity[4]. In addition, the limited stringency of native washes, the inefficiency of UV crosslinking, and the low shearing resolution of most RIP protocols also contribute to a very low signal to noise ratio. As evidenced by the disparity of candidate lncRNAs found between various sources of PcG RIP data [5-8], there is a lack of consensus between RIPs from different groups, coupled with a high contamination of mRNA exons: a metric of false positives in RIP of many chromatin proteins.

RIP has been successfully used to verify lncRNA/protein interactions, which were suggested *a priori* by other sources of data. As exemplified by the lncRNAs interacting with PcG proteins, such as the essential cardiogenesis lncRNA *Braveheart* [9] or the HOX gene-regulator lncRNA *HOTAIR* [10], differential expression and knockdown of the transcripts were first observed to cause changes in expression of classical PcG target genes, and then sought out in PcG-RIP. Likewise, in the case of the highly abundant lncRNA *Xist* [11] or several lncRNAs associated with imprinted genes[8,12,13], PcG proteins were already known to be involved in silencing of the adjacent target regions. In these instances, RIP was used to verify, rather than to first indicate, the interactions with the PcG proteins.

The overwhelming false positive rate renders RIP ineffective as a means to identify protein/RNA interactions *de novo,* at least for lowly or moderately expressed RNAs. RIP often cannot readily discern true signal from noise of higher-expressed RNAs (such as mRNAs), essentially limiting the dynamic range of RIP to highly or overexpressed lncRNAs. Furthermore, since most lncRNAs are expressed at significantly lower levels than mRNAs[14,15], many lncRNAs are well outside this dynamic range. Indeed, RIP is not a robust technology for uncovering novel chromatin-interacting lncRNAs, and, a scarce number of RNAs from such studies have been biologically verified.

The PcG proteins also pose particular biochemical challenges for finding novel lncRNA interactions. Firstly, it is yet unclear which protein or set(s) of proteins directly bind to RNAs, or confer specificity for binding. Previously, the PcG protein EZH2 had been shown *in vitro* to directly bind to RNAs such as the *RepA* region of *Xist* [11] and has since been shown to spread along the inactive X in correlation with *Xist* spreading [16,17]. However, recent evidence shows that EZH2 does not have strong specificity for any RNA motif, and strongly binds many RNAs regardless of sequence, *in vivo* and *in vitro* [2,18]. Additionally, EZH2 is just one of several proteins in the methyltranferase complex of the PcG proteins, PRC2. It is not known whether the entire complex, or perhaps additional component(s) which interact with the complex, are important for proper lncRNA binding.

PcG mediated silencing is also executed by another PcG complex, PRC1, whose interactions with lncRNAs are yet uncertain. PRC1, or, more accurately, several PRC1-like complexes comprised of various combinations of subunits, are responsible for the compaction of chromatin and ubiquitylation of H2A. It is this compaction which is thought to block access of the transcriptional machinery to the DNA. PRC1 components are essential for the stabilization of silencing and are partially retained on the chromatin during mitosis to maintain epigenetic memory through cell division [19,20]. Though there is extensive overlap between PRC1 and PRC2 binding on the

chromatin[21,22], differences in activity of the complexes, recruitment of the complexes, and binding sites exist[23,24]. Therefore, it is mechanistically important to study lncRNAs in the presence of PRC1, the main engine of repression.

Previous data have suggested interactions between PRC1 and lncRNAs[25-27], though the specifics of such interactions remain unstudied. It is not well understood if PRC1 directly interacts with RNA *in vivo*, or how lncRNAs impact the activity or recruitment of PRC1 to the chromatin. Several PRC1 components, such as *Polyhomeotic* show *in vitro* binding to RNA [27], though notably, many of those same protein domains bind DNA as well. The PRC1 proteins Bmi1 and C*BX7* have been shown to interact with the lncRNA *ANRIL* [25,28] to modulate the Ink4a/Arf locus. However, additional candidates have been poorly studied in comparison to PRC2-interacting lncRNAs. It is not even known whether PRC1 is generally present at sites of lncRNA/PRC2 binding, or whether such interactions occur while PRC2 is bound to the chromatin. We therefore sought to develop a method that was better able to predict novel lncRNA interactions with chromatin, and specifically to investigate lncRNAs present at sites of PRC1 binding.

Results

Our goal is to uncover novel lncRNAs that interact with chromatin, with the hypothesis that interacting RNAs may modulate gene expression. While conventional RIP has been used to validate *a priori* hypotheses of lncRNA interactions, the variable results of such experiments between different groups, as well as strong mRNA contamination, indicate a high level of noise. Such noise interferes with identification of legitimate, stable lncRNA/chromatin interactions, and may arise from non-specific RNA interactions that occur during the purification, rather than in the cell.

To this end, we attempted to improve the signal to noise ratio of conventional RIP by purifying chromatin as the input to the immunoprecipitation (IP), and tailoring the washes around both RNA and protein specificity. Accordingly, we developed an IP protocol that uses a CsCl gradient to isolate chromatin, the substrate of the PcG proteins, away from free nucleic acids or free protein. We reasoned that by removing these sources of noise, we might also change the spectrum of RNAs that were pulled down to enrich for stable, *in vivo* interactions. We also chose to work in crosslinked cells, where complex components are covalently fixed together. In this manner, we could identify lncRNAs that act at the same genomic loci as PRC1, without necessarily knowing which component(s) of PRC1 or its binding partners (including PRC2) directly bound the lncRNAs. Moreover, the covalent fixing of the complexes allowed us to employ more stringent chromatin purification, IP, and wash conditions aimed at reducing RNA noise. In order to investigate the interactions of lncRNAs with chromatin proteins, such as the PcG proteins, we needed to expand the dynamic range of canonical RIP to include lowly expressed lncRNAs and exclude mRNAs.

Development of the Assay

We set up our system in HeLa cells stably expressing a FLAG-tagged PRC1 component, FLAG-Bmi1 (25% overexpression) to allow for cross-validation of results between the endogenous and tagged protein. The over-expressed FLAG-Bmi1 in this cell line was shown previously to interact with the core components of PRC1[29]. We further characterized the FLAG-Bmi1 protein by anti-FLAG immunofish, confirming the protein was indeed localized to punctate bodies on the chromatin (Figure 1), typical of endogenous Bmi1(Figure 2).

A brief description of our technique for discovering PRC1-associated lncRNAs on the chromatin is as follows: we isolated nuclei from crosslinked HeLa cells stably overexpressing FLAG-Bmi1, and

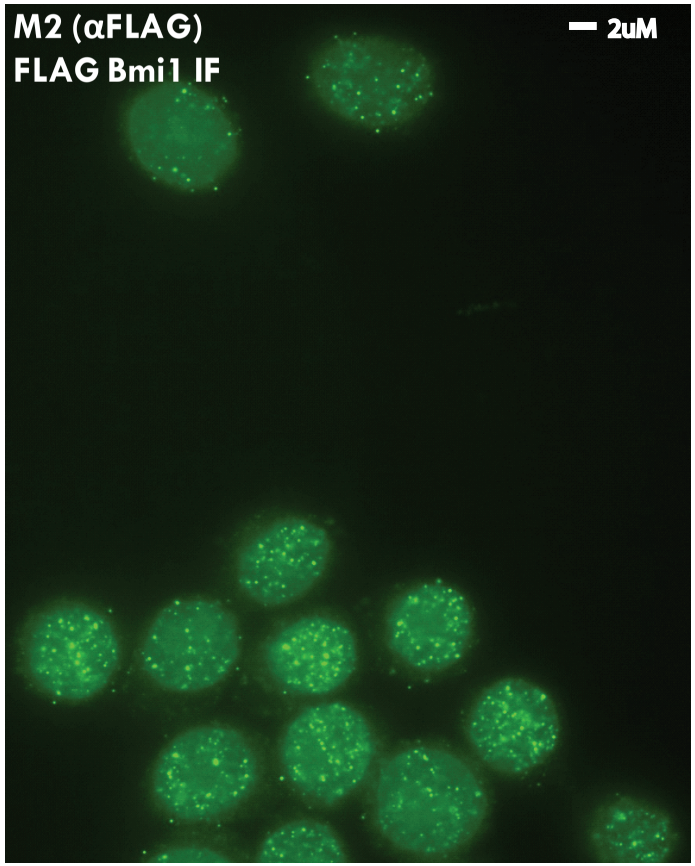

Figure 1. Immunoflourescence of FLAG-Bmi1
Immunoflourescence targeting FLAG was performed in Hela cells that stably express both endogenous Bmi1 and FLAG-Bmi1 (25% overexpression). Only FLAG-Bmi1 is recognized by the antibody. Punctate distribution is visible in all cells (100 cells counted, n=3).

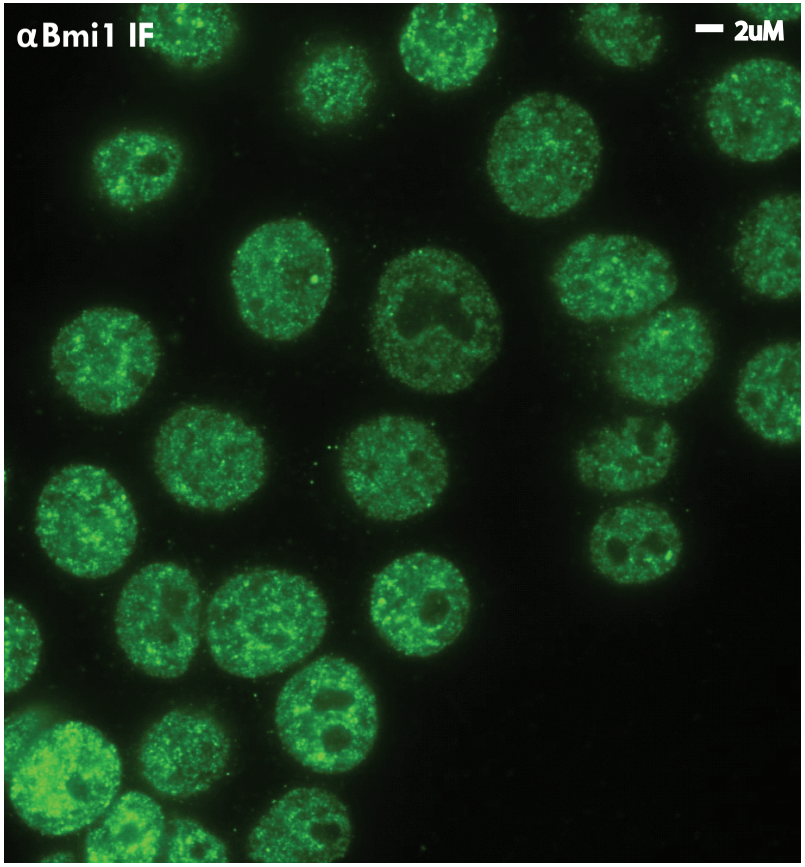

Figure 2. Immunoflourescence of Bmi1
Immunoflourescence targeting Bmi1 in HeLa cells that stably express both endogenous Bmi1 and FLAG-Bmi1 (25% overexpression). Both endogenous and FLAG-Bmi1 are recognized by the antisera. Punctate distribution and nucleolar exclusion is visible in all cells (100 cells counted, n=3).

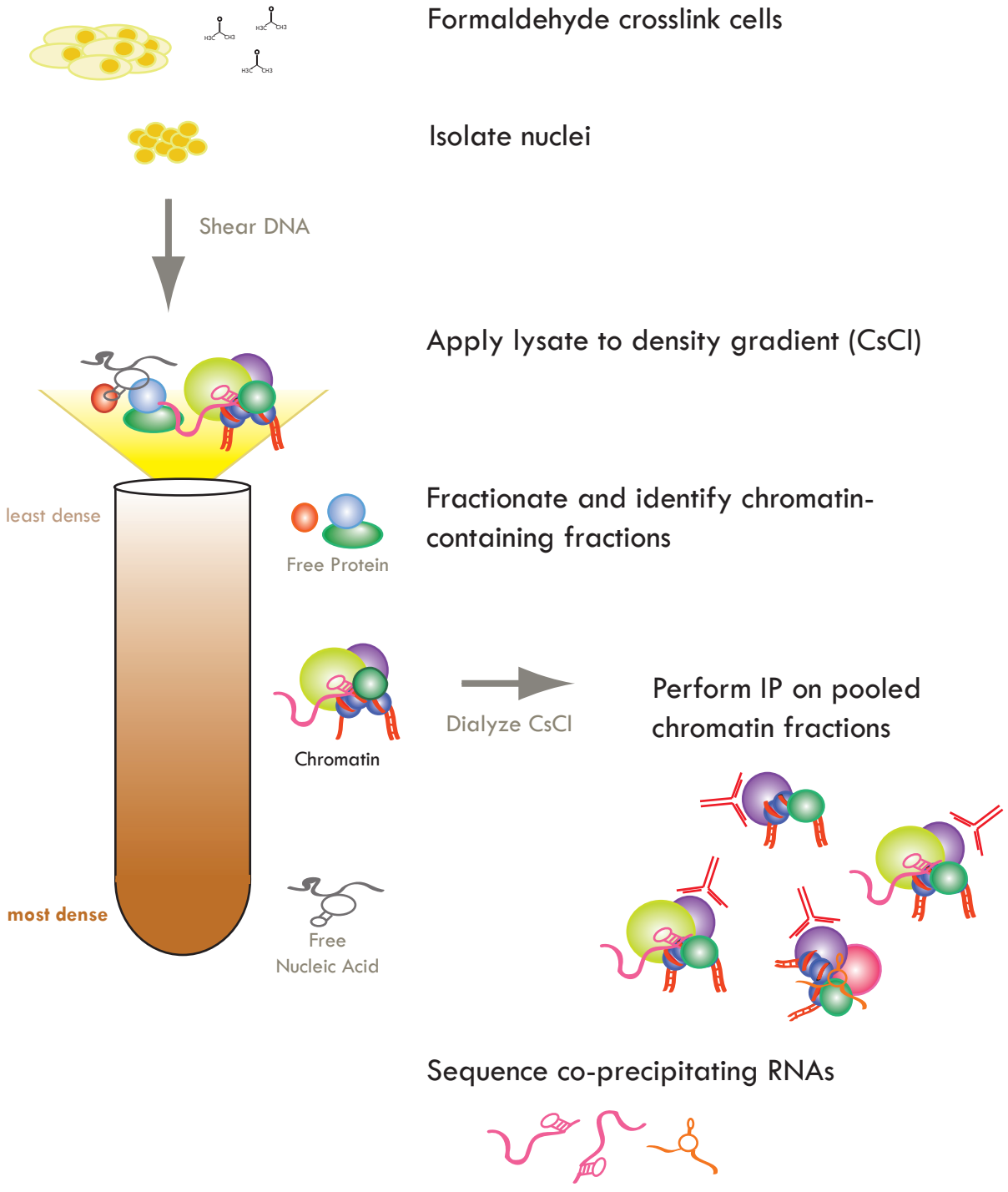

Figure 3. Schematic of protocol to isolate Chromatin Associated Transcripts (CATs)

sheared the nuclei. To isolate the chromatin, we applied the nuclear lysate to a CsCl density gradient. Chromatin-containing fractions from the gradient were identified by immunoblot and spectrophotometry, and pooled for further purification. CsCl was removed from the pooled fractions by dialysis to prepare the chromatin for IP against the PcG proteins (or various controls). Finally, the IP's were washed in both high and low salt for protein and RNA specificity, respectively, and the co-precipitating RNAs were isolated for sequencing (Figure 3). More detailed descriptions of these steps follow.

Migration of biomolecules through the CsCl Density Gradient

The isolation of chromatin by CsCl density gradient is a major purification step in the protocol. The density gradient, once a routine step in early mammalian ChIP, separates the bulk chromatin from sources of noise: free nucleic acids, free protein, lipids, and aggregates. As previously described in early ChIP studies [30], free protein is expected to run near the top of the gradient whereas free nucleic acids, which are much denser than proteins, are expected to collect at the bottom of the gradient. Chromatin, which is comprised of both nucleic acids (DNA and RNA) and proteins (histones, transcription factors, etc) is expected to migrate to the center of the gradient.

We confirmed previous results outlining the migration of various biomolecules through the density gradient. By Bradford assay, protein ran from the center of the gradient to nearly the top (Figure 4). Immunoblot analysis of specific targets revealed that the chromatin binding proteins Bmi1, Suz12, PHC1 and CTCF migrated slightly below the bulk protein, in the central fractions (fractions 4,5 and 6) with the histones (Figure 5). DNA, which is associated with protein and largely compacted into chromatin, migrated in the center of the gradient (Figure 6, Figure 7). The

tight co-migration of DNA and chromatin proteins suggested that the DNA was sufficiently crosslinked to remain associated with chromatin proteins through the purification.

In contrast, RNA collected at the bottom of the gradient, but was also spread throughout the bottom and middle of the gradient (Figure 8, Figure 9). RNA migration at the bottom of the gradient can be readily explained by the effect of high salts on RNA binding. Namely, the high concentrations of CsCl ablate weak or non-specific RNA/RNA interactions or non-specific interactions between basic patches of protein and RNA, and precipitate free RNA. RNA migration in the central fragments can be explained by crosslinking and protein association: Nucleic acids are inefficiently crosslinked to protein by formaldehyde, and are generally retained by being trapped in crosslinked protein "cages". The presence of RNAs in the bottom and central fractions is likely due to RNAs being caged by various protein interactors, and being sheared into non-uniform fragments based on RNA secondary structure and RNA/protein interactions (footprinting). Therefore, the crosslinked CsCl gradient serves not only as a means to reduce non-specific binding, but to then separate much of the contaminating RNA from the chromatin.

To ensure that lncRNAs indeed remained bound to the chromatin, we probed migration of the canonical chromatin bound lncRNAs *Xist* and *h19* RNAs as a proxy of lncRNA retention. RT-qPCR revealed that both *Xist* and *h19* lncRNAs migrated in fractions 4-6, with the bulk chromatin (Figure 9). This demonstrates that the *Xist* and *h19* present in the nuclear prep are protein-bound, and are sufficiently crosslinked to maintain the interaction through the purification.

We also examined mRNAs of various transcription levels to assay where mRNA noise might be generated from, and to show that retention of RNAs was specific to protein bound RNAs. We found that free nucleotides were precipitated to the bottom of the gradient, whereas highly transcribed RNAs (processed or unprocessed) migrated in a single, central fraction. P68, a processed mRNA of low/moderate transcription, was expected to accumulate in the lowest

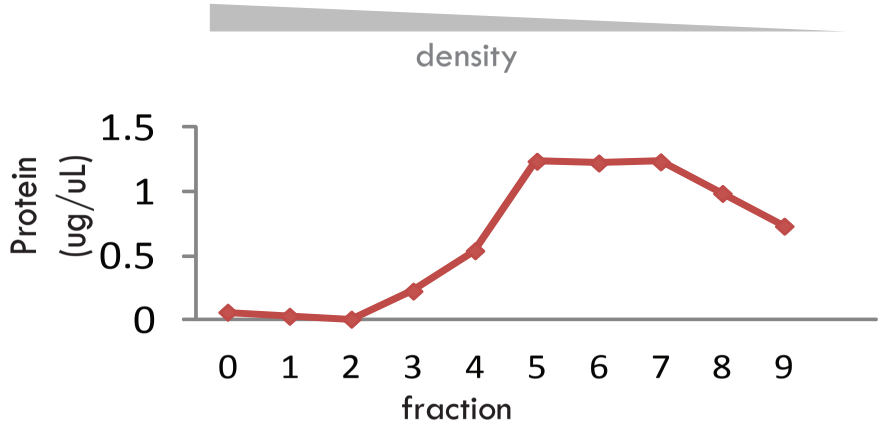

Figure 4. Bradford Assay Across Fractions
Protein migrates in the less dense, central to upper fractions.

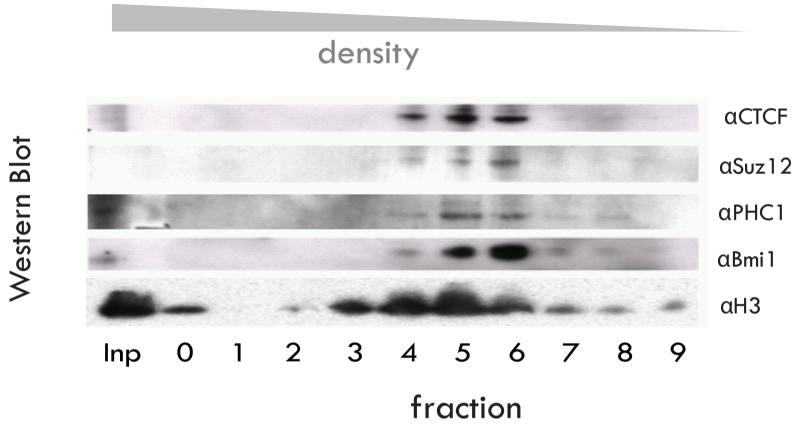

Figure 5. Immunoblot Across Fractions
Chromatin-bound transcription factors, including the PcG proteins, comigrate with histones in the central fractions.

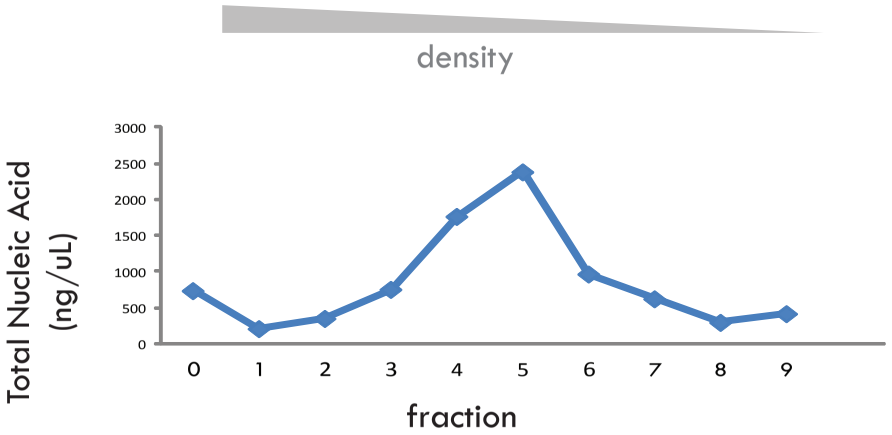

Figure 6. A260 Across Fractions
Nucleic acids collect either at the bottom of the gradient or in fractions

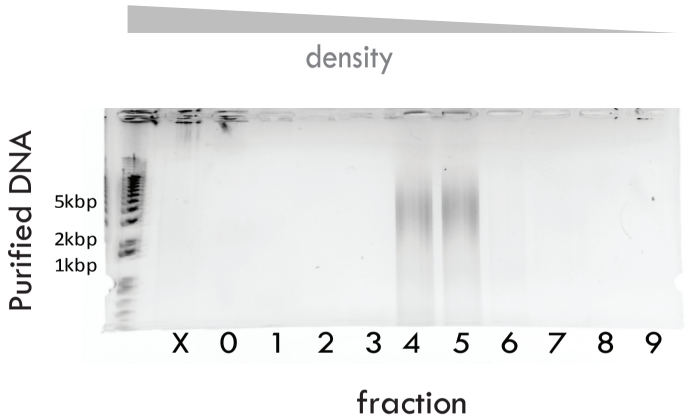

Figure 7. Purified DNA Across Fractions
DNA migrates in the central fractions. Though DNA is dense, its elevated migration is likely due to association with protein.

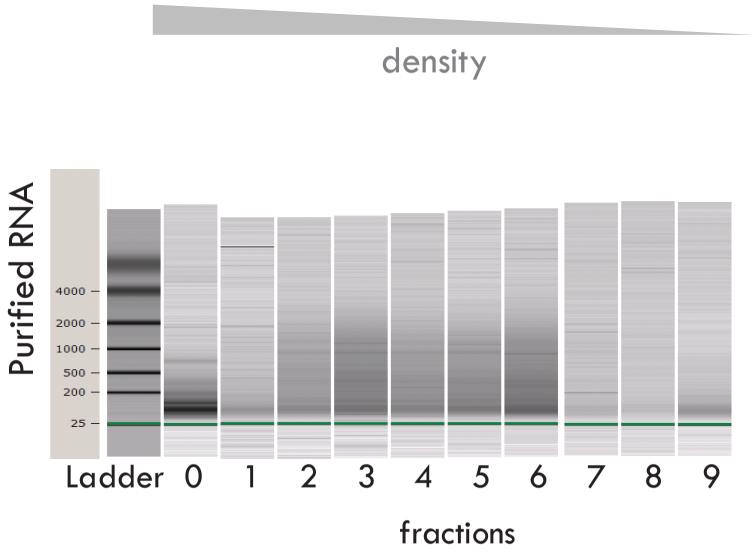

Figure 8. Total RNA Across Fractions
Free RNA is precipitated to the bottom of the gradient. RNAs are located in the lower and central fractions as well, likely a result of protein/RNA interaction.

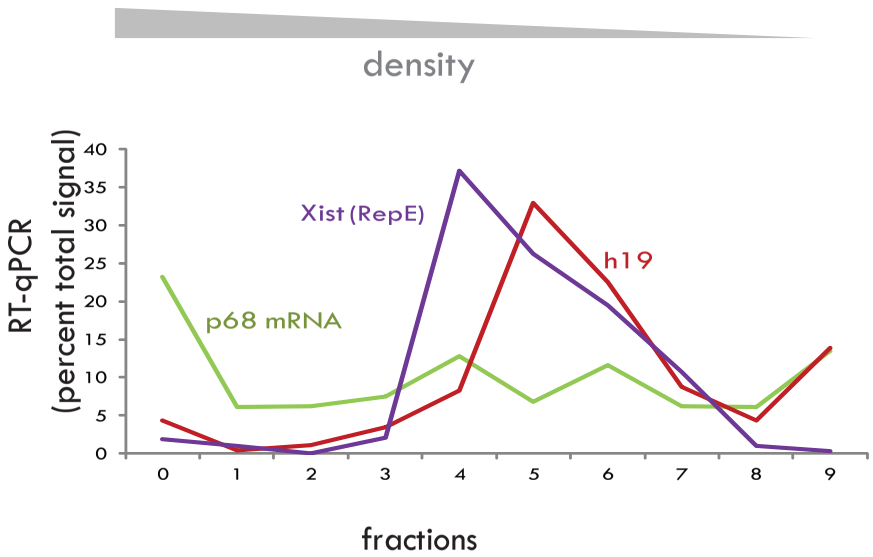

Figure 9. RT-qPCR of Select RNAs Across Fractions
Chromatin bound RNAs migrate in the central fractions, whereas free mRNAs migrate at the bottom fractions. Data normalized to total signal, so that area under each curve is 100 (%)

fractions if present due to cytoplasmic contamination (free of protein), or possibly in higher fractions if being transcribed or shuttled with proteins. P68 mostly accumulated in the bottom of the gradient or was present at levels indistinguishable from untranscribed controls (Oct4) (Figure 9, Figure 10), indicating that most of the P68 in the sample was free, and separated from the chromatin. Highly transcribed mRNAs such as *GAPDH* and unprocessed pre-mRNAs (GAPDH intron/exon junction normalized to a –RT control) were almost exclusively located in the most dense chromatin fraction (4), presumably still tethered to the chromatin and PolII (Figure 10). These data suggested that free RNA was indeed migrating to the bottom fraction, chromatin interacting lncRNAs were largely retained alongside chromatin proteins, and that mRNA in the chromatin fractions were likely tethered to the DNA, presumably at the respective genomic loci.

We optimized our shearing conditions using Bmi1 and *Xist* as positive controls to test the effects of shearing on PcG-bound lncRNAs. *Xist* has previously been shown to interact with PRC2 components, though direct binding to PRC1 has not been investigated. However, the high abundance of *Xist* in the cell made it an attractive candidate for RT-qPCR analysis. Interestingly, more intense shearing lead to migration of Bmi1 in the higher fractions, whereas total histone migration was not proportionally elevated (Figure 11). This is consistent with reported shearing-hypersensitivity of PcG binding sites, thought to be caused by nucleotide bias and a broad nucleosome-free region. Similarly, *Xist* but not *GAPDH* migration mirrored the elevated migration pattern of Bmi1 (Figure 12). Accordingly, shearing conditions (4.5 Kbp DNA fragments) were optimized to solubilize the DNA (Figure 17) and maintain RNA integrity, while keeping Bmi1 together with the bulk chromatin. Only fractions which contained both Bmi1 and the bulk chromatin (fractions 4,5, and 6) were pooled for dialysis and immunoprecipitation.

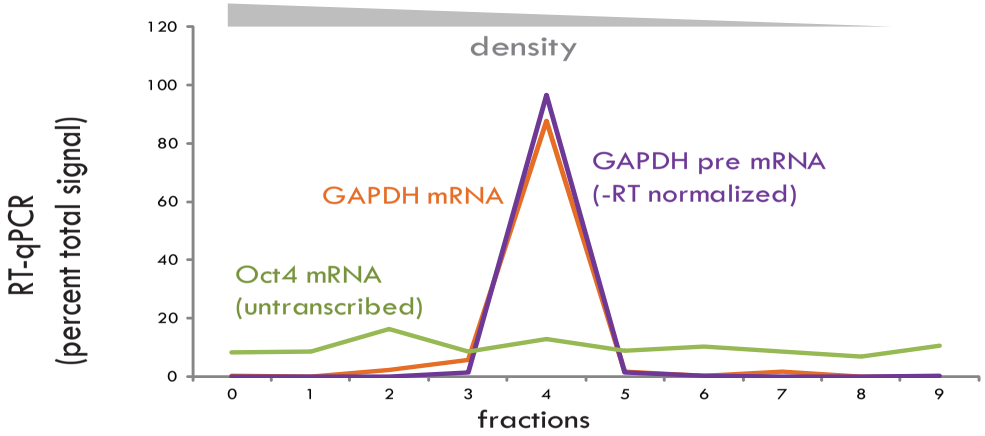

Figure 10. RT-qPCR of mRNAs Across Fractions
Both GAPDH (exon-spanning) and GAPDH pre-mRNA (exon-intron junction) co-migrate, almost entirely in fraction 4. Presumably, this is due to crosslinking at the transcription site. A non-transcribed control, Oct4, is shown. Data normalized to total signal, so that area under each curve is 100 (%).

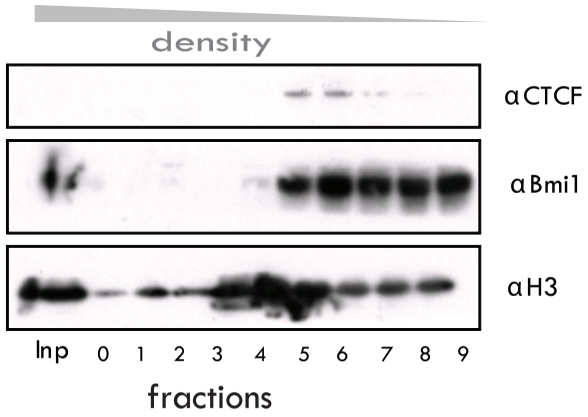

Figure 11. Immunoblot Across Fractions in Higher Shearing
Increased shearing conditions cause Bmi1, more so than CTCF or H3, to display an elevated migration pattern. This is consistent with shearing hypersensistivity at PcG binding sites

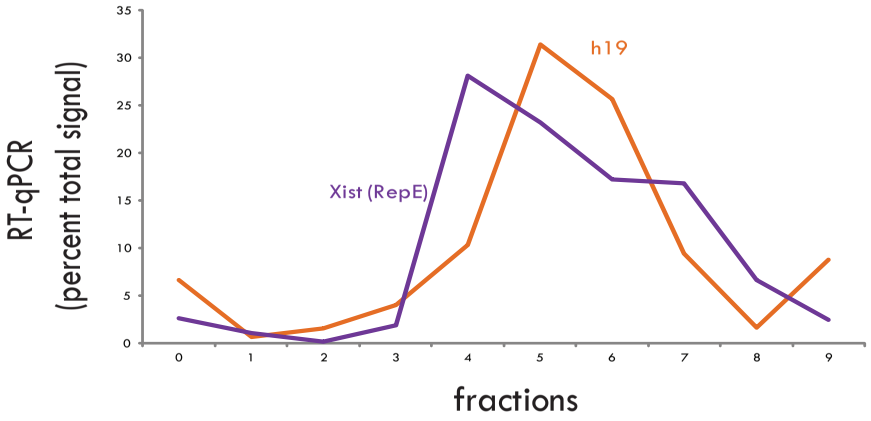

Figure 12. Chromatin-bound RNA Migration in Higher Shearing
Xist migrates in higher fractions upon increased shearing. h19 stays relatively unchanged. Data normalized to total signal, so that area under each curve is 100 (%).

IP specificity

We immunoprecipitated Bmi1 and FLAG-Bmi1 from the pooled chromatin fractions (fractions 4-6) expecting that true PcG-binding lncRNAs would be enriched over input, and cross-validated between the two samples. The FLAG-Bmi1 IP differed from the Bmi1 IP in a number of ways: it was performed with a mouse anti-FLAG antibody, was specific for only 20% of the Bmi1 in the cell (FLAG-Bmi1), was precipitated with covalently crosslinked agarose beads instead of ProteinA coated beads, was subjected to harsher IP conditions and washes (1M Urea), and was eluted from the beads via peptide elution (3X FLAG) instead of by SDS. We additionally performed an IP targeting the widely bound transcription factor CTCF (though in smaller scale), to show specificity for associations with Bmi1 versus general associations with chromatin. Finally we also performed a (smaller scale) IgG IP, as a universal negative control. We reserved portions of the input and IP eluates for immuno-blot (or silver stain) (Figure 13, Figure 14) and qPCR (Figure 15, Figure 16). These assays verified that the IP's targeting multiple chromatin proteins were specific at the protein and DNA levels. Of note, the mean length of the DNA from the Bmi1 IP was 3.5 kbp whereas the input DNA had a mean length of 4.5 kbp (Figure 17).

Identification of candidates: RNAs cross-validate in a non-random fashion

We sequenced RNA from the input, Bmi1 IP, and FLAG-Bmi1 IP and identified enriched peaks which cross-validated between the two samples. We first aligned uniquely mapping reads from each sample, using Bowtie. Using the program Model-based Analysis of ChIP-Seq[31] (MACS), we identified read pileups (peaks) in our sequencing data from individual IP samples. MACS called peaks based on read density, peak shape, amplitude, and width, to identify transcripts *de novo* from the data. These peaks represent exons of lncRNAs and potential protein binding domains.

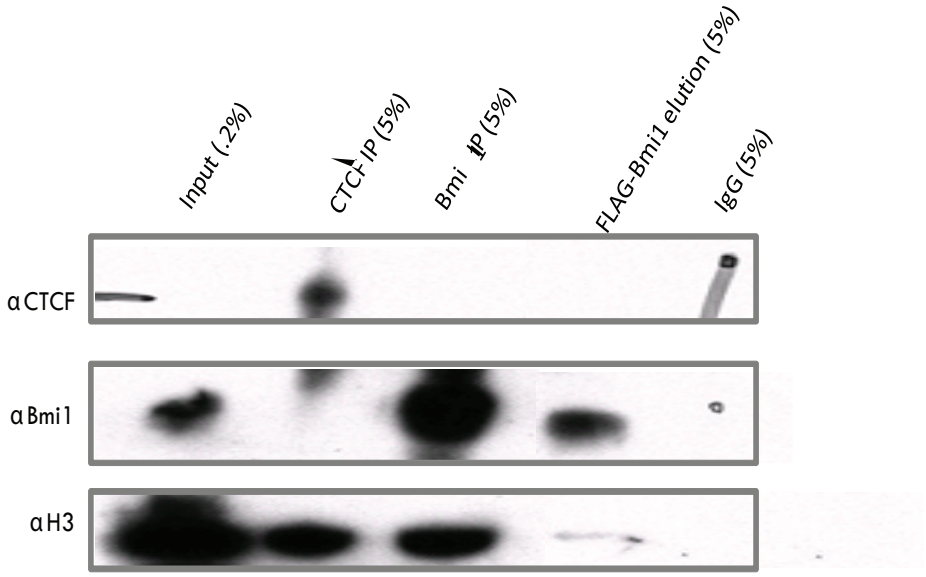

Figure 13. Immunoblot of IP Eluates
Eluates show protein specificity: CTCF and Bmi1 are selectively pulled down in the CTCF and Bmi1/FLAG-Bmi1 IP's. Similar results seen in >4 biological replicates. A non-related lane was electronically removed for figure clarity.

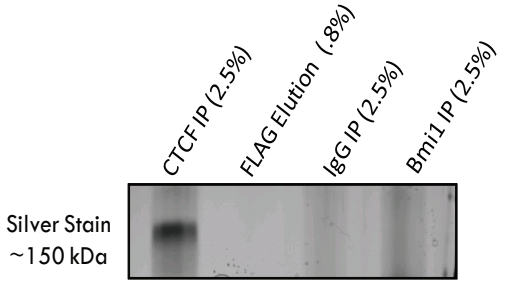

Figure 14. Silver Stain Analysis of IP Eluates
CTCF IP did not show a band in the Immunoblot input. We therefore confirmed that the IP shows a single band at the expected size by silver stain

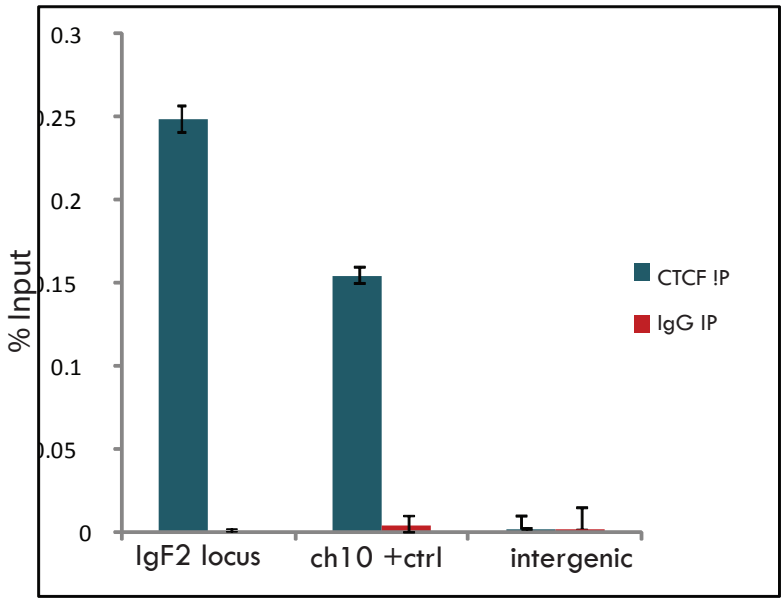

Figure 15. CTCF-IP shows specificity for CTCF targets
DNA was isolated from the a portion of the IP-eluate. qPCR was performed at two regions of published CTCF binding and an intergenic control. qPCR performed in triplicate. Similar results seen in >3 biological replicates.

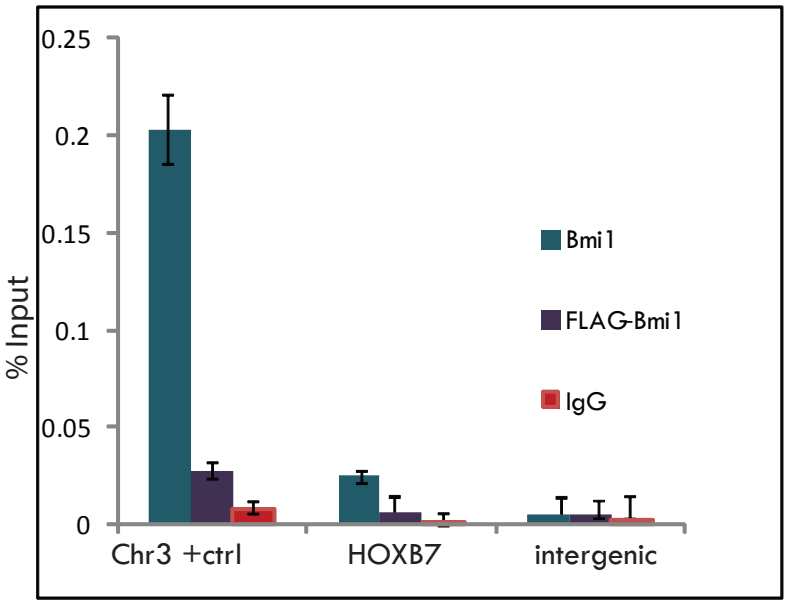

Figure 16. Bmi1 IP and FLAG-Bmi1 IP show specificity for Bmi1 targets.
DNA was isolated from the a portion of the IP-eluate. qPCR was performed at two regions of published Bmi1 binding and an intergenic control. Notably, FLAG-Bmi1 only accounts for 20% of the total Bmi1 in the cell. qPCR performed in triplicate. Similar results seen in >5 biological replicates.

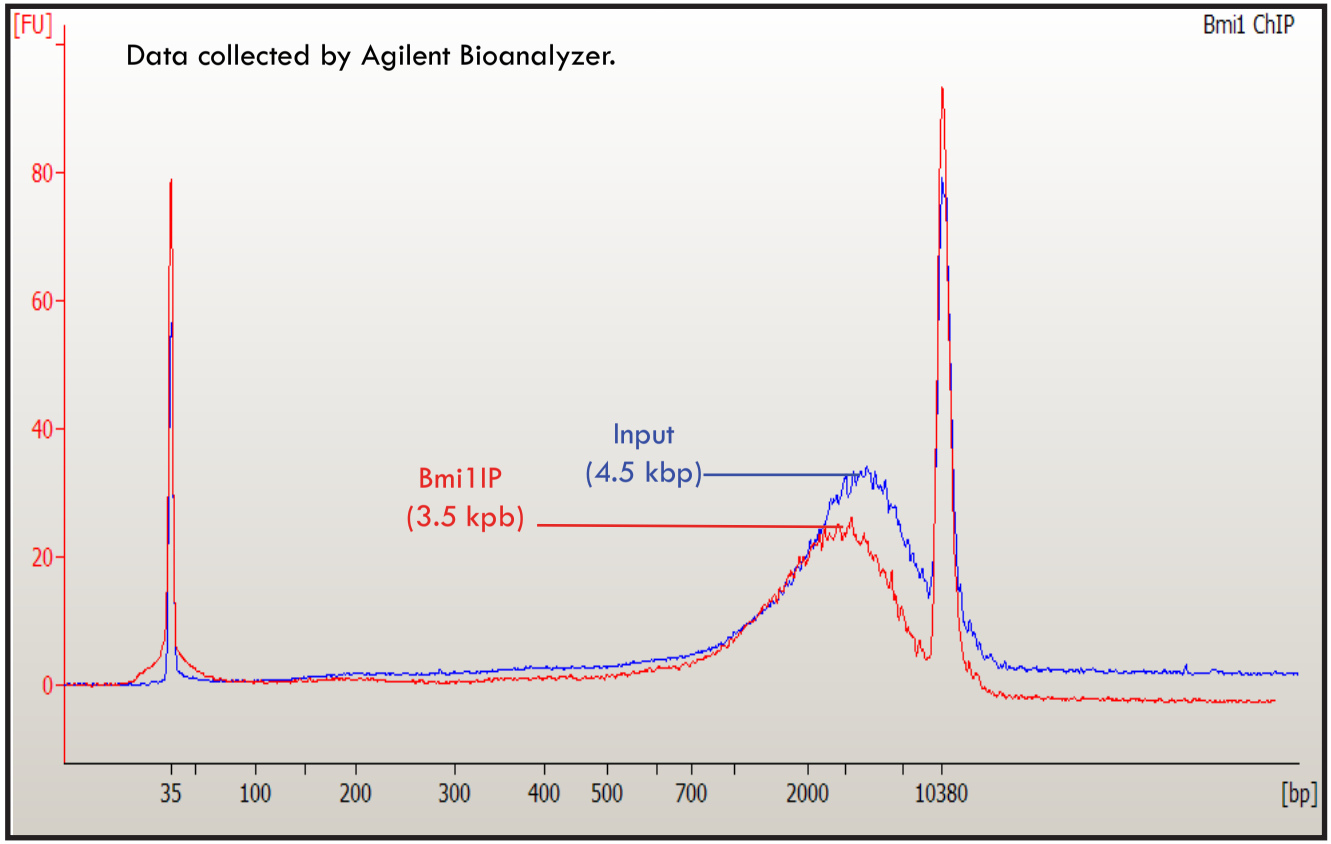

Figure 17. Length of DNA in Bmi1 IP and Input

A small portion of the eluate from each IP was reserved for analysis of the DNA. The Bmi1 IP had a mean shear length of 3.5 kbp, whereas the input was sheared to 4.5 kbp. FLAG-Bmi1 IP and IgG IP control DNA were not present in high enough amounts to be visible by this assay (data not shown).

The exons are unassembled and may not represent an entire lncRNA due to RNA shearing and possible footprinting effects. Therefore, to compare the samples, we simply considered overlapping peaks which were called in both Bmi1 and Bmi1 and FLAG-Bmi1 IPs, and had a pvalue from MACS less than 10^{-15} in both samples. For each overlapping peak, we then calculated the intensity ratios (reads in the IP over reads in the input) to normalize to transcript abundance in the input. We found that the intensity ratios across the entire set of overlapping peaks correlated with Rval=.61 (Figure 18). Significant enrichment over input was further determined at each peak using EdgeR[32], and overlapping, enriched peaks from Bmi1 and FLAG-Bmi1 IP's (pval <.01, fold change > 2) were selected for further study. These peaks represent cross-validated, enriched candidates for PcG interaction.

We also correlated the log-intensities at annotated regions, and found that the correlation occurred only at RNAs from specific regions of the genome. We first correlated all mRNA exons, as annotated by RefSeq[33]. Generally, processed mRNAs are expressed at higher levels than lncRNAs[14,15], and often are used as a metric of false-positives or transient interactions since no biology has implicated them in PcG function. We found that the log intensity ratios across mRNA exons were correlated with a low Rvalue of .30 (Figure 19). Of 8,731 mRNA exons expressed over background in both IP's, zero exons were significantly co-enriched between the replicates. Conversely, we found that intensity ratios across previously annotated lncRNAs (Ensembl)[33,34] had Rval = .80, and an Rval = .73 excluding all repetitive regions called by Repeatmasker[35] (Figure 20). Since lncRNAs are often poorly annotated, we also correlated only regions of lncRNAs with signal (MACS peaks within the lncRNAs), excluding repeats. These peaks had an Rval = .68 (Figure 21). Taken together, these data revealed that the correlations were non-random: strong at regions of predictive of true interactions, and weak at regions predictive of noise. Furthermore, the global depletion of mRNA exons is indicative of a major reduction in a known

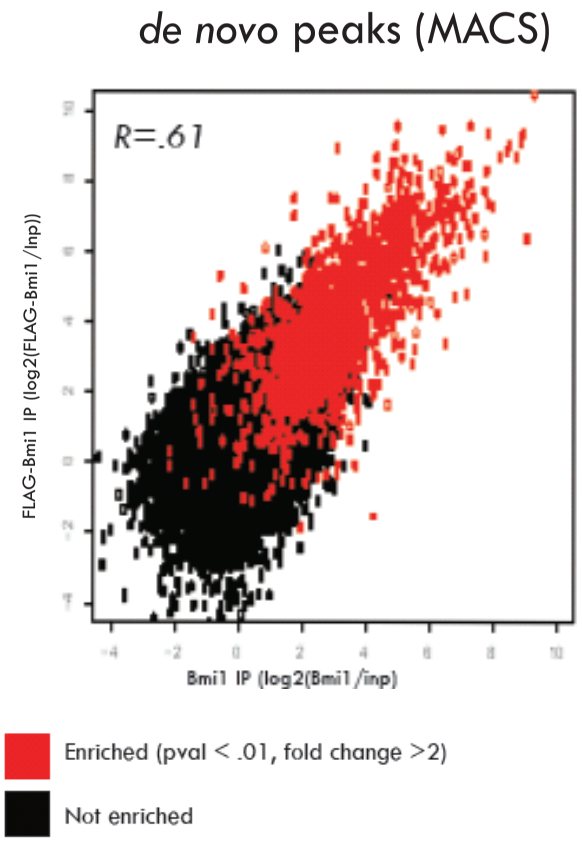

Figure 18. RNA from Bmi1 and FLAG Bmi1 IPs correlate at overlapping peaks

Count ratios are defined here as the total number of reads in an IP across a given region, divided by the total number of reads from the input sample across the same region. Bmi1 count ratios vs FLAG-Bmi1 count ratios were plotted in log2 scale across de *novo* peaks called by MACS. Only peaks with intensity above background are shown. Count ratios were analyzed by EdgeR. Peaks that are significantly enriched in both Bmi1 and FLAG-Bmi1 samples are shown (red).

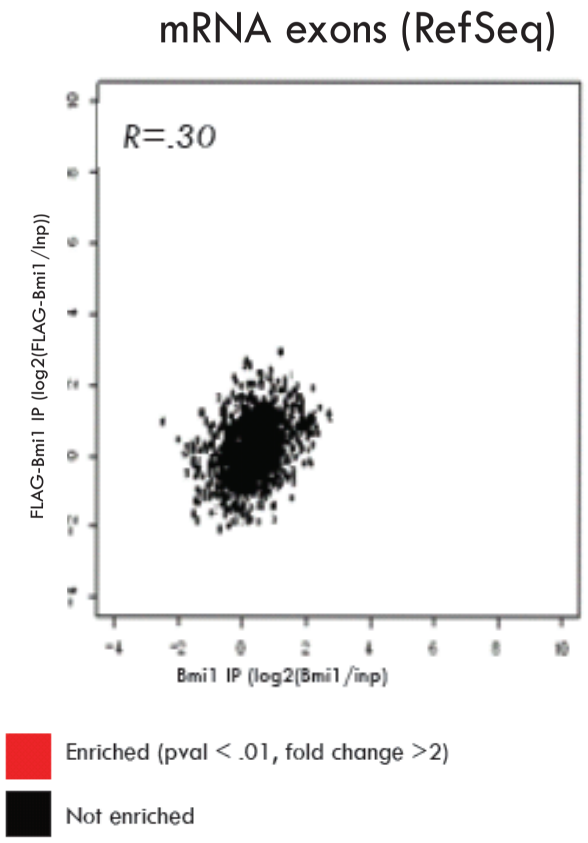

Figure 19. RNA from Bmi1 and FLAG-Bmi1 IPs correlate poorly at mRNA exons

Count ratios are defined here as the total number of reads in an IP across a given region, divided by the total number of reads from the input sample across the same region. Bmi1 count ratios vs FLAG-Bmi1 count ratios were plotted in log2 scale across mRNA exons as defined by RefSeq. Only exons with intensity above background are shown. Count ratios were analyzed by EdgeR: zero exons were significantly enriched in both Bmi1 and FLAG-Bmi1 samples.

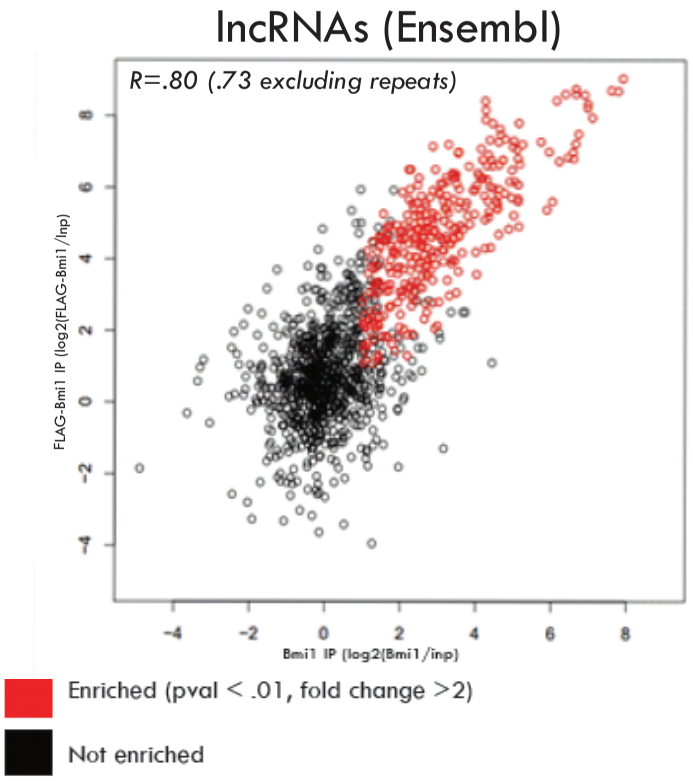

Figure 20. Bmi1 and FLAG-Bmi1 IPs correlate highly at lncRNAs

Count ratios are defined here as the total number of reads in an IP across a given region, divided by the total number of reads from the input sample across the same region. Bmi1 count ratios vs FLAG-Bmi1 count ratios were plotted in log2 scale across lncRNAs previously annotated by Ensembl. Only peaks with intensity above background are shown. Count ratios across lncRNAs were analyzed by EdgeR. lncRNAs that are significantly enriched in both Bmi1 and FLAG-Bmi1 samples with fold change over input >2 are shown (red).

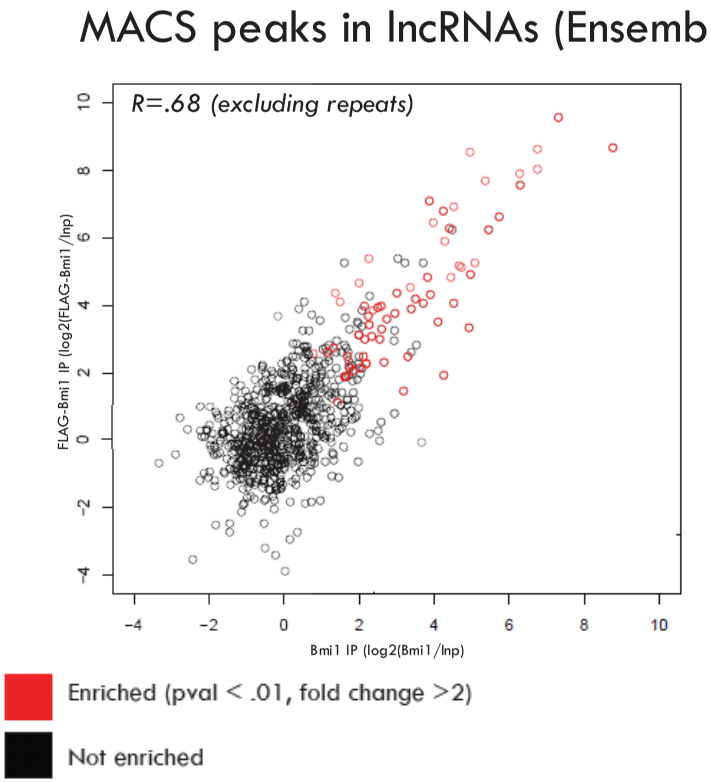

Figure 21. Bmi1 and FLAG-Bmi1 IPs correlate highly at MACS peaks mapping back to lncRNAs

Count ratios are defined here as the total number of reads in an IP across a given region, divided by the total number of reads from the input sample across the same region. Bmi1 count ratios vs FLAG-Bmi1 count ratios were plotted in log2 scale across de novo peaks, as called by MACS, and intesected with lncRNAs previously annotated by Ensembl. Only intersecting MACS peaks/lncRNAs with intensity above background are shown. Count ratios were analyzed by EdgeR. Peaks that are significantly enriched in both Bmi1 and FLAG-Bmi1 samples with fold change over input >2 are shown (red).

source of noise. Therefore, we had high confidence in our MACS-peak candidates, and the enriched, annotated lncRNAs also represented a second set of candidates for PcG interaction.

RT-qPCR validates individual candidates

We then validated individual RNA sequencing results by RT-qPCR, and incorporated additional controls. We performed RT-qPCR directly on the eluate from the Bmi1 IP, as well as from the CTCF IP and the IgG control from matched input (Figure 22). We examined *Xist*, which has been proposed previously to interact directly with *EZH2* of PRC2, and which, specifically the *RepE* domain, was also the top hit of our *de novo* peaks (Figure 23). RT-qPCR revealed that *Xist RepE* was highly enriched in the Bmi1 IP, but not in the IgG or CTCF IP. The functionality of *Xist* in this cell line was supported by RNA-FISH of *Xist*, showing that each cell had a single, distinct barr body (Figure 24). Conversely, *h19* (Figure 22, Figure 25), which has not been shown to interact directly with PcG proteins, was not enriched in the initial sequencing experiments or by RT-qPCR. *ANRIL*, which has been previously implicated to play a role in Bmi1 mediated regulation of the Ink4a/Arf locus, was mildly enriched in the Bmi1 IP sequencing data, but not the FLAG IP sequencing data. Because of this lack of cross-validation, and because the cell-cycling role of *ANRIL* in a cancer line is unclear, we could not interpret it as a control and did not pursue it by RT-qPCR. Overall, these data confirm that the sequencing results were specific in comparison to IgG and CTCF.

We next confirmed our results by performing a biological replicate of the FLAG-Bmi1, Bmi1 and IgG IP's. By comparison with the replicates (via IP/RT-qPCR) we identified reproducible candidates, removed IgG non-specific peaks, and demonstrated relatively low variability in the technique. We first generated cDNA similarly as we did in preparation for library construction

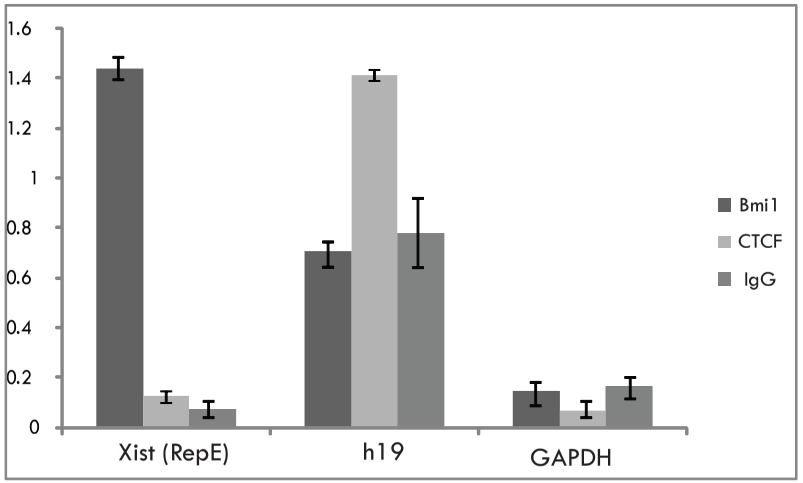

Figure 22. RT-qPCR of RNA from Bmi1, CTCF, and IgG IPs
Xist (RepE), but not h19 or GAPDH, is pulled down specifically by Bmi1 and not by CTCF. Performed in triplicate. SImilar results were seen in >3 biological replicates.

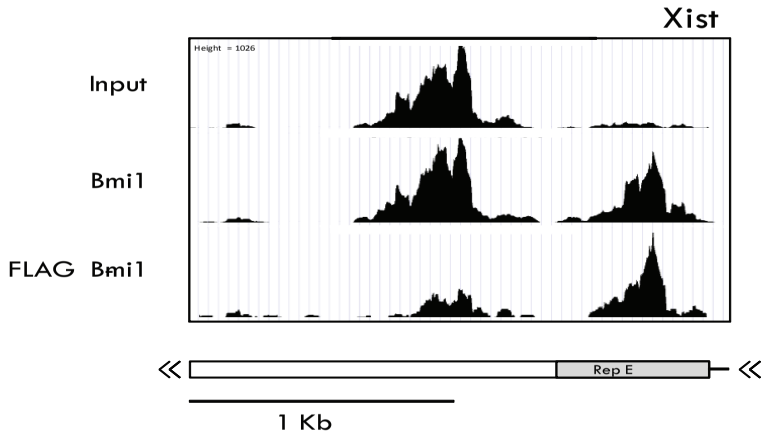

Figure 23. Sequencing reveals Xist RepE is selectively enriched in Bmi1 IP's
Bmi1 and FLAG-Bmi1 IP's are enriched for Xist precisely over RepE

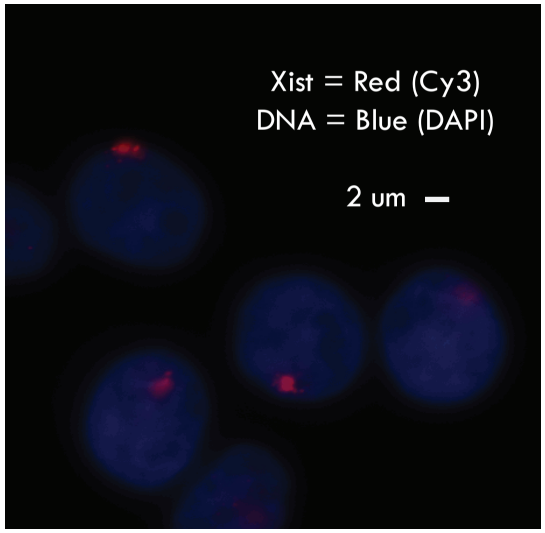

Figure 24. RNA-FISH of Xist in HeLa Bmi1F17 cells
Distinct Barr bodies appear in nuclei of 82% HeLa Bmi1F-17 cells. 100 cells counted. Similar results seen in n>5 experiments.

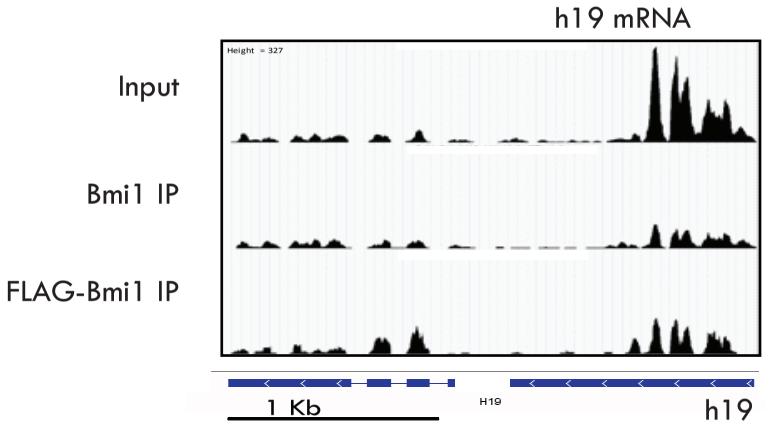

Figure 25. Sequencing data reveals h19 RNA is not enriched in the Bmi1 IP

above (NuGeN kit) and performed qPCR on a number of candidates from each IP/input in the replicate experiment. Once again, *Xist* was enriched in the pulldowns of Bmi1 or FLAG-Bmi1 as compared to the IgG; *GAPDH* and *h19* were not (Supplemental Table 1). We expanded our RT-qPCR screen to target sixty-three candidates in addition to *Xist, GAPDH,* and *h19*. Candidates were selected from co-enriched MACS peaks for validation based on: intergenic location, p-value as assigned by EdgeR, the absence of large repetitive regions (RepeatMasker), and peak width greater than 500 bp. We also gave higher weight to peaks aligning to previously annotated lncRNAs. Trying up to two amplicon pairs per peak, we found that 42 candidates were validated in the replicate pulldowns but not the IgG, 6 did not reproduce and were higher in the IgG, and 15 candidates consistently gave multiple melt curves, and were not interpretable. This yields a minimum 66.7% validation rate. These transcripts became our top candidates for PRC1 interaction. All results summarized in Supplemental Table 1.

We surveyed the 65 initial candidates for sequence motifs, though did not find any strong sequence bias. However, though we considered only uniquely mapping reads, many candidates contained tandem repeats[36]. More precisely, many candidates contained ~40-200bp sequences that were repeated with modest fidelity (~75%) along a single, contiguous region of the genome. These domains might account for the difficulty in obtaining single melt curves for many candidates by RT-qPCR. Furthermore, such domains may also be highly structured, and could represent a structural class of PcG-interacting lncRNAs.

Expression levels, nuclear localization, and tissue-specificity of the candidates

After showing non-random candidates had a high validation rate, we further characterized our highest-confidence candidates. We assayed the dynamic range of the experiment (with respect

to candidate expression level), localization of the candidates in the cell, and expression of the candidates in biological tissue.

The expression levels of our candidates varied widely. On average, the top 100 MACS peaks selected had an rpkm (reads per kilobase per million) of 0.48 in the input, yet about 7.71 in the Bmi1 IP (Figure 27, Figure 28, Figure 29). As a reference, *Xist Rep E*, which was our top candidate as ranked by p-value assigned by Edge-R, had an rpkm of 3.78 in the input and 34.43 in the IP (Figure 23). This showed us that our protocol was identifying candidates expressed at a wide range of abundances.

Based on prior evidence that lncRNAs modulate the PcG protein activity on chromatin, we verified that the candidate lncRNAs were indeed appreciably retained in the nucleus. We mined cell-fractionated RNA-seq data, publicly available from ENCODE. These data included matched whole-cell, nuclear and cytoplasmic RNAseq data, obtained from ten different human cell lines, including a HeLa line. For the 65 RT-qPCR verified targets, we verified 53 were primarily localized to the nucleus, 6 were both nuclear and cytoplasmic, and 7 had insufficient data to determine localization (e.g. 53/59 verified; 42/42 of which were replicated by RT-qPCR) (Supplementary Table 1).

Finally, we characterized expression of these lncRNAs, found in *HeLa* cells, in a biologically relevant setting. RT-qPCR was performed for selected candidates across RNA from 20 tissues in the body, placenta, and developing fetus, as well as in embryonic stem cells. We saw that these lncRNAs were indeed present in various tissues of the body, and were not simply artifacts exclusive to HeLa cells. Furthermore, each candidate was differentially expressed in a tissue-specific manner (Figure 26). These results show that we have generated a high-confidence, cross-validated list of non-random lncRNA candidates that may be working with the PcG proteins across various tissues of the body.

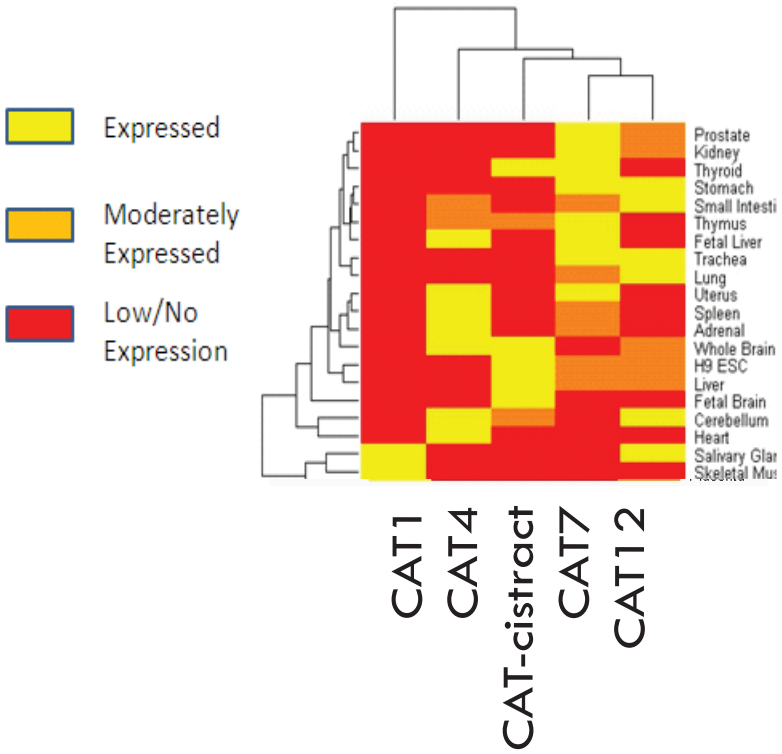

Figure 26. Relative expression of selected candidates across a variety of human tissue.

"Chromatin Associated Transcripts" (CATs) from our screen appear to be expressed in a tissue specific manner across the body. RT-qPCR was performed in triplicate and mean expression of select CATs is shown relative to GAPDH. Signal is further normalized to the mean value across each column. Hierarchical clustering using a Euclidean distance metric was used to generate the plot.

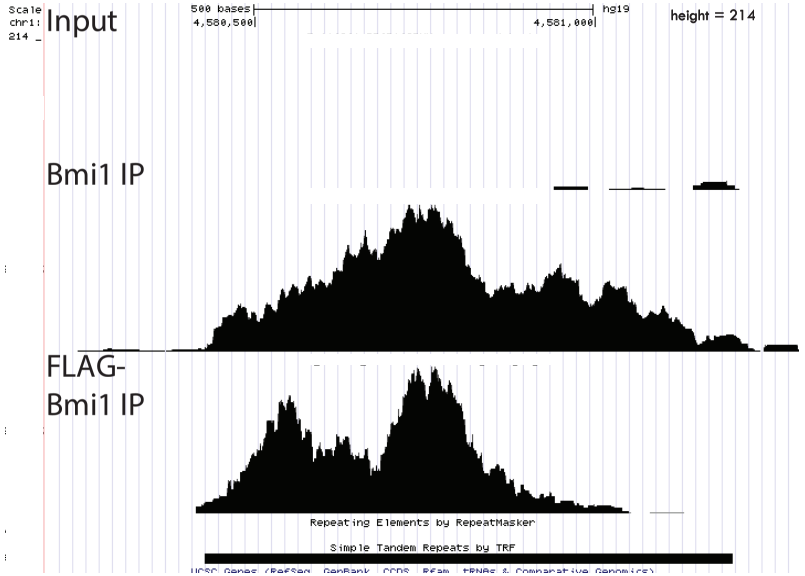

Figure 27. Representative Enriched Peak
An RNA peak from Bmi1 or FLAG Bmi1 IP
Chr1:4580194-4581300

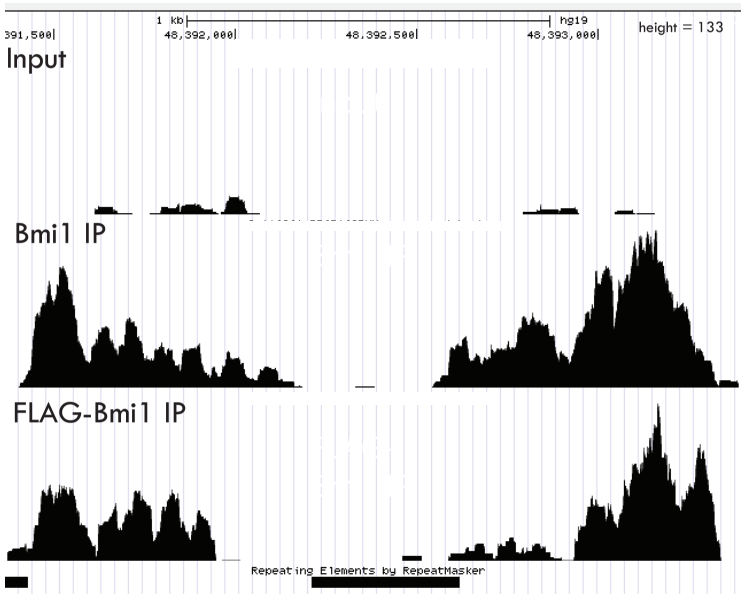

Figure 28. Representative Enriched Peak
An RNA peak from Bmi1 or FLAG Bmi1 IP
ChrX:489391290-48393443

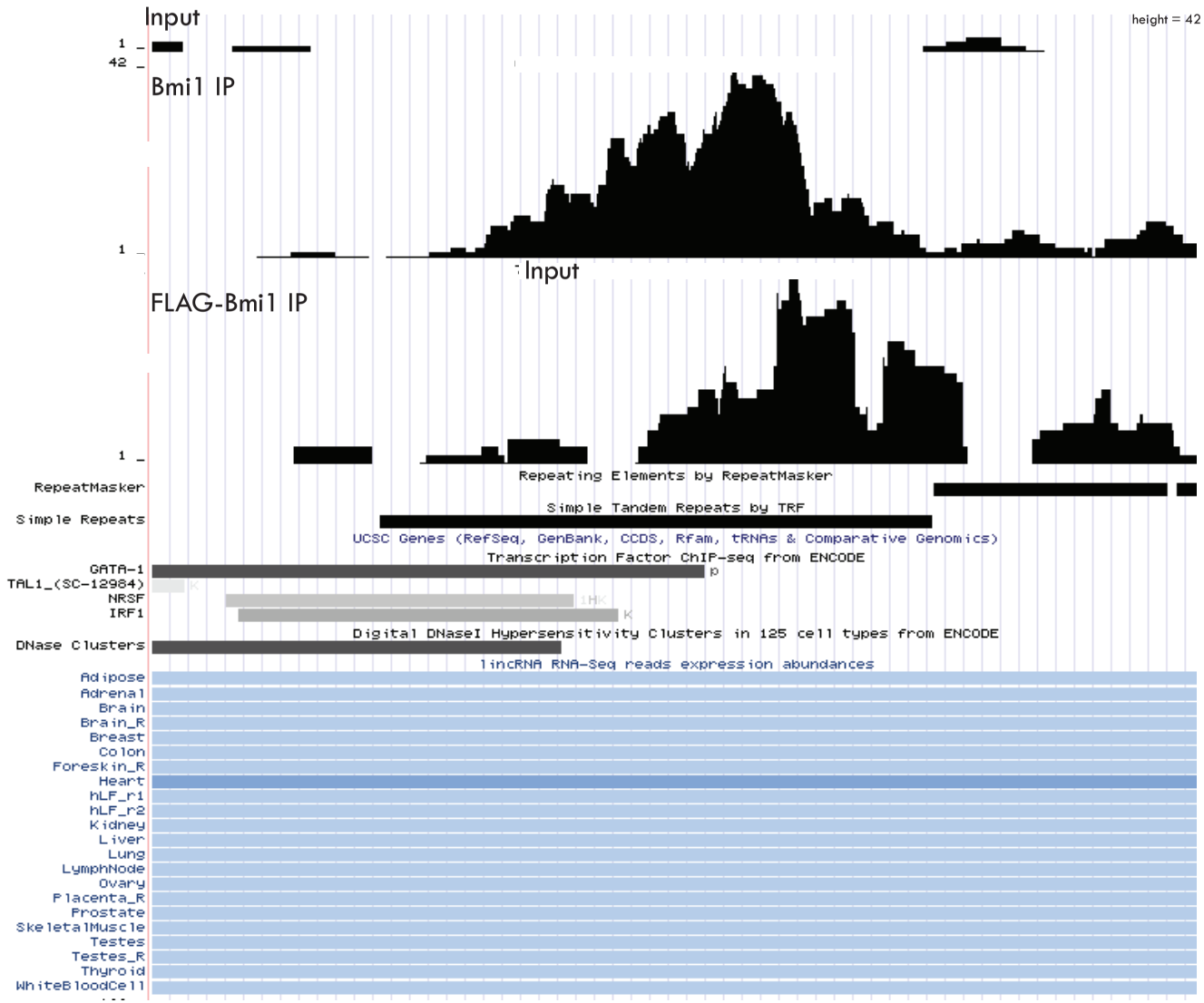

Figure 29. Representative Enriched Peak

An RNA peak from Bmi1 or FLAG Bmi1 IP: Chr16:58468257-58468920. Notably, this RNA region corresponds to the intron of an annotated, heart specific lncRNA. GATA1/NRSF binding sites coincide with the trancript.

Discussion

The molecular understanding of gene regulation has shifted to include lncRNAs as regulators of transcription. However, techniques to identify such lncRNAs remain limited in specificity and dynamic range. Our technology identifies a broad spectrum of chromatin-interacting lncRNAs, expressed at a variety of levels. By relying on a clean chromatin input, stringent washes geared at RNA specificity, and cross-validation techniques, we have nearly eliminated mRNA contamination from our top peaks: a metric of noise that plagues canonical RIP reactions. Applying our technology to the PRC1 protein, Bmi1, we generated a list of candidates that potentially interact with PRC1, and/or its binding partners.

Previous RIP or CLIP studies have targeted PRC2 components to study PcG/lncRNA binding. In particular, many sources have searched for RNAs by EZH2 binding. However, recent reports state that EZH2 promiscuously binds RNAs *in vitro* and *in vivo* [2,18]. It remains unknown which protein(s) confer specificity to PRC2 *in vivo*. Our IP protocol uses crosslinked complexes, allowing us to identify RNAs at sites of PRC1 binding, without assuming which PRC1 protein(s) – or which PRC1 binding partner, such as PRC2 – the RNA directly binds.

Our RNA candidates not only cross-validate between biological replicates, but also are enriched in a parallel immunoprecipition of a stable FLAG-Bmi1. Overall, the FLAG-Bmi1 and endogenous Bmi1 IP's correlate well, and, importantly, in a non-random fashion that is not driven by expression level. Namely, the correlation is strong at regions of putative positive control (lncRNAs) but much weaker at regions of noise (mRNA exons).

Among our candidates, we have identified a functionally novel region of *Xist* as the top hit. Though *Xist* has been shown to bind to EZH2 at its *RepA* locus *in vitro* [11], *RepA* deletion does not

result in PRC1-component delocalization from the inactive X chromosome (personal communication, Neil Brockdorff), and generally, PRC1 components are dispensable for X inactivation[37]. Furthermore, though localization of PRC1 components to the chromatin has been reported[28,37,38], the direct interaction of PRC1 with *Xist* has not been investigated. We found striking enrichment of *Xist* signal precisely over *RepE,* rather than at *RepA* . *RepE* is near the 3' end of *Xist*, and its function remains elusive. Our results suggest that *RepE* may play a role in recruitment of PRC1 proteins to the inactive X.

A large number of the top candidates contained tandem repetitive regions that were unique to a single, contained area of the genome – such as *RepE*. Sequence analysis of top candidates did not reveal any significant binding motif, though many candidates were modestly C-rich. Taken in the context of PRC2 lacking definite sequence specificity and yet binding to the highly structured domain *RepA* , it is possible that these tandem repetitive regions constitute a structural class of PcG target lncRNAs.

Conclusion:

Our method identifies nuclear, non-random RNAs that interact with chromatin proteins. As demonstrated with the PRC1 protein Bmi1, associated RNAs cross-validate at regions of expected signal, but not at noise. Our candidates are expressed in the body, largely localized to the nuclei, and often contain short tandem repeats. Moreover, they range from very high abundance in the cell, such as *Xist* RepE, to lower abundance transcripts, which are readily discernable from negligible mRNA noise. These RNAs constitute a class of lncRNAs, we refer to as Chromatin Associated Transcripts (CATs). Moreover, this protocol works for a variety of proteins/antisera, supporting its applicability to other systems, such as perhaps Oct4/Sox2/Nanog proteins in ES cells, to find CATs involved in maintaining the ES cell state. Our list of PRC1-interacting CATs are

strong contenders for true PcG interaction. Further validation is necessary in order to integrate them into the PcG gene-regulatory network.

Methods

Cell Culture and crosslinking

HeLa cells stably transduced with a copy of FLAG- tagged Bmi1 at approximately 25% overexpression[29] were cultured in DMEM supplemented with 10% FBS (Sigma), $NaHCO_3$ pH 7.5 and gentamycin. Cells were grown in suspension in a spinner flask (Matrical) to a density of approximately $3x10^8$ cells per liter. The cells were crosslinked at 1% formaldehyde for 20' in PBS at room temperature with light rocking.

Nuclei prep, Sonication and CsCl Gradient

All steps were performed in the presence of RNAsIn plus (Promega) RNAse inhibitor, DTT, .05 mM Spermine, .05 mM Spermidine, and Complete EDTA-free Protease inhibitor tablets (Roche).

Crosslinked HeLa cells were resuspended in LB1 (50 mM Tris 7.4, 140 mM KOAc, 1 mM EDTA, .5 mM EGTA, 10% glycerol, .5% NP-40, .25% Triton X-100, .1% digitonin) and spun to isolate nuclei and porate the nuclear membrane. The porated nuclei were then dounced in LB1 using a Wheaton homogenizer type A and applied over a glycerol pad (25% glycerol, 1 mM EDTA, .5 mM EGTA, 10 mM Tris pH7.4) to further separate cytoplasmic debris and expelled nucleoplasm. The nuclei were rinsed in LB2 (Tris pH 7.4, 10 mM EDTA, 5 mM EGTA and 200 mM KOAc), resuspended 1:1 in LB2 and sheared using the Covaris S2 machine (30 minutes at 20% Duty cycle, power 7 and 200 cycles/burst, in 30 second intervals; 700 uL aliquots). The sheared nuclei were then resuspended (dropwise) in LB2 with 3.37M CsCl and .1% Sodium Sarkosyl, and spun

for 48 hrs at 200,000 *g* at 8 degrees Celsius to form a density gradient. Fractions were collected by a peristaltic pump in aliquots of 1/10 the total volume.

Immunoprecipitation

Selected fractions were pooled and dialyzed (MWCO 3,500 kDa) for >5 hours in LB2 containing 5% glycerol at 4 degrees Celsius. Following dialysis, .05% NP-40 and .5% Triton X-100 was added as well as 100 mM urea (or 1 M urea for immunoprecipitations against FLAG). The material was then spun out and moved to a new tube to ensure removal of any aggregates. Less than 5% loss was checked by spectrophotometry at A260 and samples were normalized in the same buffer to 800 ng/uL. The material was then pre-cleared for 45 minutes with IgG-Agarose beads (Sigma). Dynal protein A beads (Invitrogen) were pre-bound to antibody with blocking by RNAse-free BSA (Ambion). The pre-bound beads, or covalently conjugated FLAG-agarose beads (Sigma) rinsed in LB2 were added to the input and incubated at 4° Celsius overnight with light rocking. Beads were washed at room-temperature four times in IP buffer and twice in IP buffer with reduced salt (25 mM KoAc). The immunoprecipitated material was eluted from Dynal beads in 1% SDS and 1mM EDTA or from M2 agarose beads with 3X FLAG peptide as described by Sigma. Crosslinks were reversed for 1.5 hrs at 65° Celsius in the presence (RNA isolation) or absence (Protein and DNA isolation) of 1U/uL proteinase K (Roche).

RNA isolation, cDNA generation and library construction

Either input RNA or RNA from the immunoprecipitation was stored in Trizol LS (Invitrogen) following crosslink reversal and proteinase K treatment. Chloroform was added and the sample was spun out according to the manufacturer's instructions. The aqueous phase was applied to

Zymo Clean-and concentrator 5 columns and DNAsed "in tube"as per the manufacturer's instructions for RNAs larger than 200 nucleotides. cDNA was generated using the NuGEN RNA Ovation FFPE kit (7150-08), or by SuperScript VILO (Invitrogen). For library generation, cDNA was blunted, A-tailed, and ligated to Illumina adaptors as described previously (Simon *et al.*, 2011).

Antibodies

Antibodies for immunoprecipitations targeted either FLAG (M2) (Sigma #M8823); Bmi1 (rabbit polyclonal antisera[39], or CTCF (Active Motif #39621). Immuno-blots and IF were carried out using primary antibodies as above, Suz12 (ab12703), pan-H3 (Abcam ab1791), or Clean-Blot HRP-conjugated secondary antibody.

RNA-FISH and Immunofluorescence (IF)

RNA FISH and IF was performed as previously published[40] using a CSK pre-extraction to permeabilize the cells. The construct targeting *Xist* was a gift from Dr. Judith Sharp, and was first used in the above work.

Immuno-blots and Silver Stain

Protein was either isolated from CsCl fractions by TCA precipitation and reverse-crosslinked in SDS as above, or loaded from the crosslink reversal stage from the immunoprecipitations. A final concentration of 0.5U/uL Benzonase (Novagen) was added to each sample with Lamelli Buffer and incubated at room temperature for 10 minutes followed by 10 minutes of boiling prior to SDS-PAGE. Antibodies were used at 1/4000 (Bmi1, Pan-H3) or 1/1000 (CTCF, Suz12). Silver

staining was performed with the SilverQuest kit (Invitrogen) according to the manufacturer's instructions.

DNA isolation and qPCR

Reverse-crosslinked eluate from immunoprecipitations was treated with DNAse-free RNAse (Roche) followed by Proteinase K (Roche) and subjected to phenol chloroform isolation. The aqueous phase was then ethanol precipitated with glycogen and NaCl. All qPCR was performed using Biorad iTaq with SYBR and ROX. A list of all primers used is compiled in Supplementary Table 1.

Peak calling and Significance using MACS and EdgeR

Uniquely mapped reads were aligned to the hg19 build of the genome. MACS[31] was used to call peaks over regions of read pileups. We used default parameters with the exception of the p-value, which was selected at 10E-15. Overlapping MACS peaks were defined as having as little as 1 bp overlap. Peaks were merged between Bmi1 and FLAG-Bmi1. Because MACS assumes a flat input (e.g. as for ChIP), we also assayed for significance relative to the input in two ways: by EdgeR[32] and by using MACS itself with an input control. All peaks reported were highly significant by p-value by either method. EdgeR also imposed intensity cuts as reflected in the p-value. Necessarily, we did not pay attention to the FDR because many peaks were expected in the input that were not in the IP data. We chose peaks with pval <.01 and fold change >2, as assessed by EdgeR.

References

1 Brockdorff, N. Noncoding RNA and Polycomb recruitment. *RNA (New York, N.Y.)* **19**, 429-442, doi:10.1261/rna.037598.112 (2013).

2 Kaneko, S., Son, J., Shen, S., Reinberg, D. & Bonasio, R. PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells. *Nature structural & molecular biology* **20**, 1258-1264, doi:10.1038/nsmb.2700 (2013).

3 Walstrom, K., Dozono, J. & von Hippel, P. Effects of reaction conditions on RNA secondary structure and on the helicase activity of Escherichia coli transcription termination factor Rho. *Journal of molecular biology* **279**, 713-726, doi:10.1006/jmbi.1998.1814 (1998).

4 Riley, K., Yario, T. & Steitz, J. Association of Argonaute proteins and microRNAs can occur after cell lysis. *RNA (New York, N.Y.)* **18**, 1581-1585, doi:10.1261/rna.034934.112 (2012).

5 Guil, S. *et al*. Intronic RNAs mediate EZH2 regulation of epigenetic targets. *Nature structural & molecular biology* **19**, 664-670, doi:10.1038/nsmb.2315 (2012).

6 Yang, L. *et al*. ncRNA- and Pc2 methylation-dependent gene relocation between nuclear structures mediates gene activation programs. *Cell* **147**, 773-788, doi:10.1016/j.cell.2011.08.054 (2011).

7 Khalil, A. *et al*. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 11667-11672, doi:10.1073/pnas.0904715106 (2009).

8 Zhao, J. *et al*. Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Molecular cell* **40**, 939-953, doi:10.1016/j.molcel.2010.12.011 (2010).

9 Klattenhoff, C. *et al*. Braveheart, a long noncoding RNA required for cardiovascular lineage commitment. *Cell* **152**, 570-583, doi:10.1016/j.cell.2013.01.003 (2013).

10 Rinn, J. *et al*. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**, 1311-1323, doi:10.1016/j.cell.2007.05.022 (2007).

11 Zhao, J., Sun, B., Erwin, J., Song, J.-J. & Lee, J. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science (New York, N.Y.)* **322**, 750-756, doi:10.1126/science.1163045 (2008).

12 Pandey, R. *et al*. Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Molecular cell* **32**, 232-246, doi:10.1016/j.molcel.2008.08.022 (2008).

13 Mercer, T. & Mattick, J. Structure and function of long noncoding RNAs in epigenetic regulation. *Nature structural & molecular biology* **20**, 300-307, doi:10.1038/nsmb.2480 (2013).

14 Mattick, J. Non-coding RNAs: the architects of eukaryotic complexity. *EMBO reports* **2**, 986-991, doi:10.1093/embo-reports/kve230 (2001).

15 Consortium, E. P. *et al*. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

16 Engreitz, J. M. *et al*. The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science* **341** (2013).

17 Simon, M. *et al*. High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. *Nature*, doi:10.1038/nature12719 (2013).

18 Davidovich, C., Zheng, L., Goodrich, K. & Cech, T. Promiscuous RNA binding by Polycomb repressive complex 2. *Nature structural & molecular biology* **20**, 1250-1257, doi:10.1038/nsmb.2679 (2013).

19 Francis, N., Follmer, N., Simon, M., Aghia, G. & Butler, J. Polycomb proteins remain bound to chromatin and DNA during DNA replication in vitro. *Cell* **137**, 110-122, doi:10.1016/j.cell.2009.02.017 (2009).

20 Francis, N. J. Chromatin Compaction by a Polycomb Group Protein Complex. *Science* **306**, doi:10.1126/science.1100576 (2004).

21 Boyer, L. *et al*. Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* **441**, 349-353, doi:10.1038/nature04733 (2006).

22 Lee, T. *et al*. Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125**, 301-313, doi:10.1016/j.cell.2006.02.043 (2006).

23 Gao, Z. *et al*. PCGF homologs, CBX proteins, and RYBP define functionally distinct PRC1 family complexes. *Molecular cell* **45**, 344-356, doi:10.1016/j.molcel.2012.01.002 (2012).

24 Simon, J. A. & Kingston, R. E. Occupying Chromatin: Polycomb Mechanisms for Getting to Genomic Targets, Stopping Transcriptional Traffic, and Staying Put. *Molecular cell* **49**, doi:10.1016/j.molcel.2013.02.013 (2013).

25 Maertens, G. *et al*. Several distinct polycomb complexes regulate and co-localize on the INK4a tumor suppressor locus. *PloS one* **4**, doi:10.1371/journal.pone.0006380 (2009).

26 El Messaoudi-Aubert, S. *et al*. Role for the MOV10 RNA helicase in polycomb-mediated repression of the INK4a tumor suppressor. *Nature structural & molecular biology* **17**, 862-868, doi:10.1038/nsmb.1824 (2010).

27 Wang, R. *et al*. Identification of nucleic acid binding residues in the FCS domain of the polycomb group protein polyhomeotic. *Biochemistry* **50**, 4998-5007, doi:10.1021/bi101487s (2011).

28 Bernstein, E. *et al*. Mouse polycomb proteins bind differentially to methylated histone H3 and RNA and are enriched in facultative heterochromatin. *Molecular and Cellular Biology* **26**, 2560-2569, doi:10.1128/mcb.26.7.2560-2569.2006 (2006).

29 Levine, S. *et al*. The core of the polycomb repressive complex is compositionally and functionally conserved in flies and humans. *Molecular and Cellular Biology* **22**, 6070-6078, doi:10.1128/mcb.22.17.6070-6078.2002 (2002).

30 Orlando, V., Strutt, H. & Paro, R. Analysis of chromatin structure by in vivo formaldehyde cross-linking. *Methods (San Diego, Calif.)* **11**, 205-214, doi:10.1006/meth.1996.0407 (1997).

31 Zhang, Y. *et al*. Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**, doi:10.1186/gb-2008-9-9-r137 (2008).

32 Robinson, M., McCarthy, D. & Smyth, G. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)* **26**, 139-140, doi:10.1093/bioinformatics/btp616 (2010).

33 Pruitt, K., Tatusova, T. & Maglott, D. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* **35**, *5*, doi:10.1093/nar/gkl842 (2007).

34 Flicek, P. *et al*. Ensembl 2012. *Nucleic Acids Research* **40**, 90, doi:10.1093/nar/gkr991 (2012).

35 Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in …*, doi:10.1002/0471250953.bi0410s25 (2009).

36 Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, *573*-580, doi:10.1093/nar/27.2.573 (1999).

37 Leeb, M. & Wutz, A. Ring1B is crucial for the regulation of developmental control genes and PRC1 proteins but not X inactivation in embryonic cells. *The Journal of cell biology* **178**, 219-229, doi:10.1083/jcb.200612127 (2007).

38 Ng, K., Pullirsch, D., Leeb, M. & Wutz, A. Xist and the order of silencing. *EMBO reports* **8**, 34-39, doi:10.1038/sj.embor.7400871 (2007).

39 Woo, C., Kharchenko, P., Daheron, L., Park, P. & Kingston, R. A region of the human HOXD cluster that confers polycomb-group responsiveness. *Cell* **140**, 99-110, doi:10.1016/j.cell.2009.12.022 (2010).

40 Xiao, C. *et al*. The XIST noncoding RNA functions independently of BRCA1 in X inactivation. *Cell* **128**, 977-989, doi:10.1016/j.cell.2007.01.034 (2007).

CHAPTER 3

lncRNAs Mediate Expression of PcG Gene-Regulatory Networks and Impact PcG Binding

CONTRIBUTIONS

All work presented was carried out by Mridula Kumari Ray, except sequencing alignment by Yanqun Wang and Ayla Ergun. Also, culture and electroporation of human ES cells, and differentiation to early motor neurons was performed by Ole Wiskow.

ABSTRACT

While a growing body of work has shown that the Polycomb group (PcG) proteins interact with long non-coding RNAs (lncRNAs), discerning true interactors from artifacts has proven extremely challenging. Here, we consider 17 novel candidates previously shown to interact with the PRC1 component Bmi1, and/or its binding partners *in vivo*, in HeLa cells. We find that siRNA depletion of 11 of the 17 individual candidates show widespread changes in the transcription of PcG-regulated genes. Furthermore, we show that depletion of one of our candidates, *CAT7*, leads to upregulation of the activating transcription factors the *Trithorax group proteins,* lysine-specific histone demethylases, and an array of homeodomain-containing genes. In addition, loss of *CAT7* causes derepression of the moderately close (400kbp away) gene *Mnx1*, as well as loss of PcG binding at the *Mnx1* promoter. Examining loss of *CAT7* during motor neuron differentiation from embryonic stem (ES) cells, we find that several essential regulators of neuronal development, such as the PcG/Shh regulated *Mnx1, Irx3* and *Isl1*, and many PcG-silenced genes in ES cells, show differential gene expression.

Introduction:

Long non-coding RNAs (lncRNAs) are emerging as genomic regulators which govern the transcriptional machinery. However, integration of lncRNAs into existing, protein-based models of gene-regulatory networks remains challenging. lncRNAs have been suggested to interact with protein complexes as tethers, allosteric switches, scaffolds, and protein evictors [1]. However, regulatory lncRNAs remain difficult to identify, and it is harder still to understand how they execute changes to gene expression.

Many lncRNAs have been shown to impact gene expression by modulating activity of transcription factors. Technologies such as RNA Immunoprecipitation (RIP) have been employed to identify transcription factor/RNA interactions. However, RIP is beleaguered with noise from non-specific RNA/RNA and RNA/protein binding [2,3]. These artifacts often overwhelm true signal, making RIP ineffective at generating novel hypotheses for biological validation.

In the previous chapter, we developed a technology to identify lncRNAs that interact with chromatin proteins. Our protocol drastically reduces mRNA contamination (a readout of noise) to expand the dynamic range of RNAs pulled down in a canonical RIP. We applied our technology to Bmi1, a Polycomb (PcG) protein that is part of a complex (PRC1) important for stable chromatin silencing and compaction. The output of our protocol was a list of lncRNAs candidates that interact with Bmi1 (PRC1), or its binding partners, potentially including another PcG complex, PRC2. These lncRNAs are of high interest because they represent candidate interactors of the PcG proteins at sites of stably silenced chromatin.

A growing body of work has implicated PRC2 in direct lncRNA binding [4-8], whereas PRC1 has remained largely unexamined. Though the precise role of lncRNAs in PcG mediated silencing has not been mechanistically established, much of the above evidence suggests that lncRNAs may

impact PRC2 component localization at specific sites on the chromatin. PRC1 binding is often preceded by PRC2 activity on the chromatin and in some instances, may require PRC2 and/or the product of PRC2 on the chromatin (H3K27me3) for proper localization. Notably, H3K27me3 is not necessary for all PcG-mediated silencing.

Recently, however, the veracity of many lncRNA/PRC2-interactions found by RIP has come into question. Several studies have shown that a component of PRC2, the methyltranferase EZH2, binds random RNAs *in vitro*, and also without sequence specificity *in vivo* [9,10]. Indeed, it is not known which components (if any) of PRC1/2 confer specificity in regards to RNA binding. These technical challenges demand that a higher standard of evidence is presented before a lncRNA is validated as part of the PcG gene network. Furthermore, there is a stark lack of biological validation from the resulting lncRNAs from PRC2-RIP screens. In order to assess the role of a lncRNA as it relates to PcG biology, not only should the lncRNA show interaction with the PcG proteins, but perturbation of the lncRNA should also exhibit an impact on the PcG-mediated gene-regulatory network or some aspect of PcG function [7,8]. Such biologically motivated criteria might include the ability of a lncRNA to affect PcG binding and/or expression of PcG target genes.

Our list of candidates from Chapter 2 represents a non-random set of putative PcG interactors that have been cross-validated, biochemically purified from a major source of noise (mRNAs), and found in nuclei of various tissues of the body. We therefore capitalized on our unique position to further investigate the role of lncRNAs in the PcG biology. Namely, we screened our candidates for effects on PcG-regulated targets and, after identifying interesting candidates, examined the role of a lncRNA in PcG recruitment at an affected locus. Finally, we probed the role of one of our candidates in a biological context, during motor neuron differentiation from embryonic stem cells.

Results:

Strategy to identify candidates for functional studies

To better understand the function of our lncRNAs in the cell, and particularly in relation to PcG biology, we designed a basic screen to identify which candidates functionally impact gene expression of PcG targets (Figure 30). Specifically, we reasoned that knocking down a lncRNA candidate might affect recruitment or assembly of PcG proteins at a repressed locus (or loci). As a result, the PcG proteins would no longer be able to repress expression of target gene(s) at the affected locus or loci, leading to an increase in gene expression at those site(s). Therefore, our screen consisted of knocking down RNAs and searching for changes in the transcriptome (RNAseq) relative to a scramble control.

We further considered that indirect effects of modifying PcG activity would also impact the output of the screen. Characteristically, PcG proteins silence master regulators of transcription. Moreover, several PcG-regulated genes might be directly affected by perturbing the same locus, such as genes organized in a PcG body. It follows that increased expression of such PcG targets could potentially enact a signaling cascade, affecting entire gene networks. Therefore, de-repression of PcG targets would not necessarily imply a direct interaction of a lncRNA with the chromatin at the target sites, or even that the change in expression was caused by loss of PcG binding. However, such a perturbation in the PcG gene network might be readily discernible from noise by broadly assaying changes in PcG target-gene expression.

Application and analysis of the screen

We selected seventeen of our top PRC1-interacting candidates in HeLa Bmi1F17 cells[11] (where they were first isolated), alongside Suz12 (of PRC2), Bmi1 (of PRC1), *Xist* lncRNA and scramble controls, to screen by this method. In choosing our lncRNA candidates, or Chromatin Associated Transcripts (CATs), we ranked candidates by p-value (EdgeR) but selectively included 9 candidates mapping back to previously annotated lncRNAs. CATs came from both unannotated and annotated lncRNA regions of the genome, but even the annotated CATs were primarily located in lncRNA introns, usually on the same strand. We also included one candidate which shared an exon with the transcript of CISTR-ACT, a chromatin interacting lncRNA upregulated in Brachydactyly [12]. The CISTR-ACT overlapping RNA was not among our top candidates, but was recently proposed to control changes in chromatin structure and expression, and to directly interact with chromatin near PcG binding sites. We also selected 9 of the top candidates which contained tandem repeats: short (20-200 nucleotides) repeated with ~75% fildelity across a single region of the genome, and called by Tandem Repeat Finder[13]. Such peaks are predicted to be highly structured, and were selected based on the previous finding that *RepA,* a tandem repeat found in *Xist*, binds to PRC2 components. Notably, *Xist* was also enriched in our screen and selected as a candidate for knockdown. However, our results suggested a strong enrichment precisely at the *RepE* locus of *Xist*, rather than at *RepA* (Figure 22-Figure 25).

We opted for an RNAi methodology that would increase the likelihood of knockdown in the nucleus. This decision was based on the hypothesis that our candidates acted directly on the chromatin, and also on prior cell fractionation data from ENCODE [14] suggesting many of our candidates are primarily localized to the nucleus. We chose to knockdown candidates using siRNAs, in order to avoid the need for shRNA processing. We also used nucleofection, which electroporates the cellular and nuclear membranes, to efficiently deliver the siRNAs into the

nucleus. After 48 hours (post nucleofection) in standard culture conditions, we harvested the total RNA from the cells to test knockdown efficiency by RT-qPCR. Following knockdown validation, we sequenced the (ribosomally depleted) total RNA to search for changes in transcriptional gene networks.

We developed a data pipeline to identify changes in gene expression following lncRNA knockdown (Figure 30). For each candidate, only uniquely mapping, non-duplicated reads, were aligned to the genome, yielding approximately 15M aligned reads per sample. The total aligned reads were normalized in the scramble sample for each comparison to a knockdown sample. We confined our comparisons to mRNAs annotated by Refseq[15]. We then analyzed the number of reads mapping back to each mRNA transcript, including various isoforms of the same gene. Transcripts were considered if they had signal above background in both the scramble and the knockdown transcriptomes. The list of genes (identified by gene name to avoid bias from multiple isoforms) represented the total set of expressed genes. We then broadly filtered the set of expressed genes, selecting genes with greater than 2-fold or less than 50% signal relative to the scramble control. This set was designated as the "changed-genes" for each knockdown.

We characterized each list of changed genes to find significant overlap with PcG-regulated genes. To analyze our data for enrichment of functional data sets, we submitted a list of changed genes to the data compendium, the Molecular Signatures database[16], for each knockdown experiment. We then identified which of 6,791 previously curated gene-sets showed significant overlap with the changed-genes for a knockdown. These gene sets included transcription factor motif binding sets, gene ontology sets, and curated pathways, such as protein reactome gene sets. Among these gene sets were PcG and H3K27me3 target gene sets from ChIP data, as well as sets of genes whose transcription changed after PcG protein depletion (RNAseq data), from an array of cell lines. To calculate significance for the intersection of a particular gene-set with a list

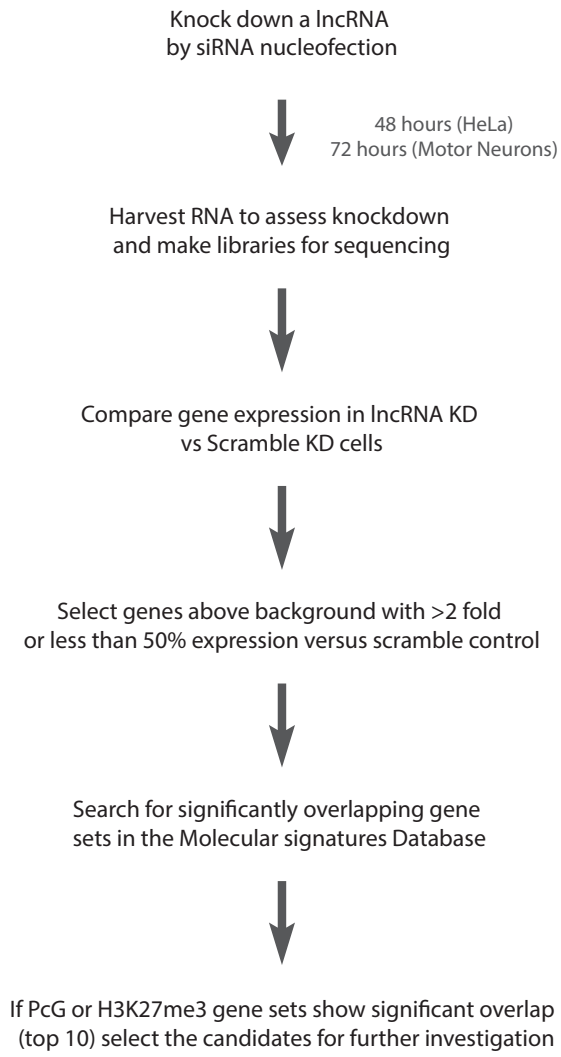

Figure 30. Strategy to identify lncRNAs that perturb the PcG gene-regulatory network

of changed genes, we only considered genes in any list (or gene background) that were expressed in the cell *and* were identified in the Molecular Signatures database. We then generated p-values and expected values for overlaps to ensure significance of our findings.

We first applied this strategy to identify expression changes from Bmi1 and Suz12 depletion, as positive controls (Figure 31). We independently knocked each mRNA down using single clones of previously published siRNA sequences [17,18]. However, despite showing 81.5% mRNA knockdown after 48 hours (versus a scramble), Bmi1 protein levels persisted near WT levels (Figure 32), presumably due to the long protein half-life in the cell. As a result, we did not expect to see many changes in gene expression. We found 157 changed-genes, largely biased towards membrane biology (pval = 4.07E-23) and probably a result of membrane damage during nucleofection. In effect, the Bmi1 knockdown served as a baseline control and validation of effective RNAi, rather than a comprehensive list of gene targets sensitive to loss of Bmi1.

Conversely, we observed significant mRNA and moderate protein reduction in Suz12 knockdowns: 69% of mRNA and more than 50% of protein was depleted (Figure 31, Figure 33). RNA-seq revealed 2-fold (or 50%) expression changes in 236 genes, several of which included other PRC2 components. This is consistent with previous reports of destabilization of PRC2 upon perturbation of any of its core components. Analyzing the set of changed genes, we saw a mild overlap with PRC2 target gene sets in the Molecular Signatures Database, which came from a wide range of cell types. The relatively moderate effect might be attributed to the residual protein in the cell (Figure 33). While this control also did not produce a comprehensive list of every gene impacted by Suz12, it revealed that many PcG targets were indeed susceptible to changes in expression due to loss of a PcG protein, and identifiable in the context of this assay.

We then verified that our siRNAs could enter the nucleus by knocking down *Xist* (RepE). *Xist* is a highly abundant, nuclear lncRNA which silences the X chromosome in female cells. It is notoriously

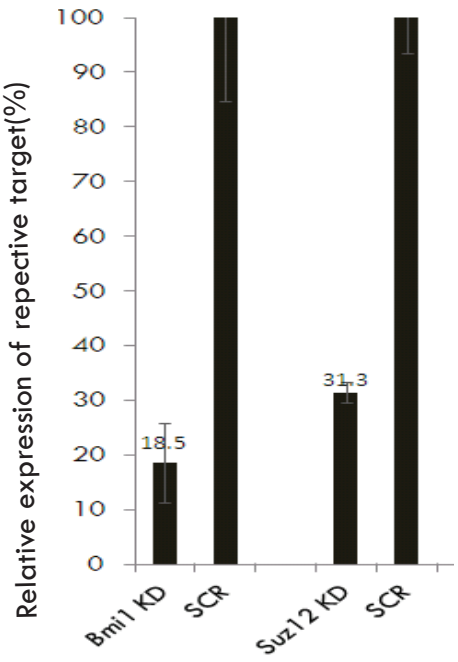

Figure 31. siRNA knockdown of Bmi1 and Suz12 mRNA after 48 hours.
Single siRNA clones were nucleofected into HeLa cells Target mRNAs are identified on the X-axis. After 48 hours, Suz12 or Bmi1 expression was examined by RT-qPCR (triplicate) in the respective knockdowns as well as a scramble control. Each lane was normalized to GAPDH signal, and then further normalized for each primer set such that the scramble was 100%. All primer sets spanned exons. Biological replicates showed similar results in n=3 (Bmi1) or n>5 (Suz12)

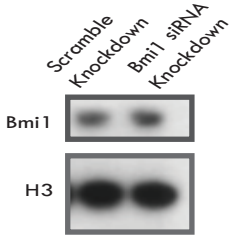

Figure 32. Immunoblot of Bmi1 knockdown after 48hrs
Bmi1 protein persisted even 48 hours after siRNA treatment.

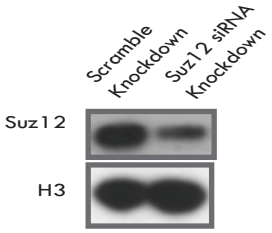

Figure 33. Immunoblot of Suz12 knockdown after 48hrs
Suz12 protein showed significant knockdown 48 hours after siRNA knockdown. However, there was still a significant protein remaining after knockdown.

difficult to deplete in the cell and yields minimal changes in expression, even with significant knockdown. However, we used *Xist* knockdown as a control to firstly confirm the siRNAs were entering the nucleus, and secondly to gauge background changes in expression. We saw a predictably modest 52% knockdown of *Xist*, and a relatively low number of changes in mRNA expression overall (141). There was no apparent bias towards changed expression of genes on the X chromosome (3/141). Once again, a significant portion of the changed genes were membrane proteins, likely a consequence of the nucleofection process, rather than a specific effect of the siRNA. This experiment demonstrated that our siRNAs were active in the nucleus, and gave a relative baseline of expression changes similar to the Bmi1 mRNA knockdown.

To investigate the role of our lncRNAs in the cell, we subjected seventeen of our novel candidates to this screen. We first verified that off-target effects were minimized in the cell for each knockdown. Explicitly, all knockdowns were performed using 2 unique siRNAs per candidate, versus a control scramble siRNA. The siRNAs were initially verified by BLAST[19] for sequence specificity (14 or fewer bp of homology to RefSeq-annotated transcripts). Whole transcriptome sequencing (RNAseq) of the knockdowns revealed that none of the most similar mRNA targets (by sequence) were significantly downregulated after 48 hours.

Applying the same analysis as for the control knockdowns above, we found that knockdown of most of the candidates affected PcG-related gene expression, without affecting expression of the core PcG proteins themselves. Of the 17 additional lncRNA candidates (besides *Xist*), 11 showed enrichment for gene sets relating to PRC2-regulated or H3K27me3 genes (Figure 34, Figure 35). In fact, excluding plasma membrane related gene sets, PcG-related sets were consistently among in the top ten most enriched gene sets in these knockdowns, and additional PcG-related gene sets were also enriched throughout the top 100 enriched gene sets. Importantly, the PcG related changed genes from each knockdown were not identical, expression of the core PcG proteins

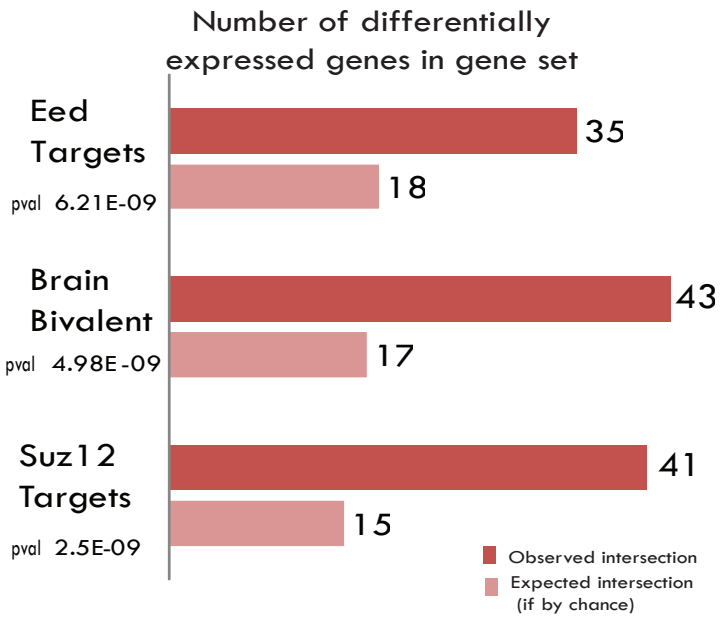

Figure 34. Genes with changed expression after CAT12 knockdown show significant overlap with PcG-related gene sets.

Total RNA from a lncRNA knockdown and scramble was sequenced. and aligned to genes. Genes expressed above background and showing >2fold (or less than 50%) expression relative to a scramble control were analyzed. Here, we show the gene set for CAT12 knockdown. Selected gene sets were identified by the Molecular signatures database. p-values and expected overlaps were calculated using a hypergemetric distribution, only considering genes that were expressed above background in the cell and found in the Molecular Signatures database. Isoforms were not included to avoid bias.

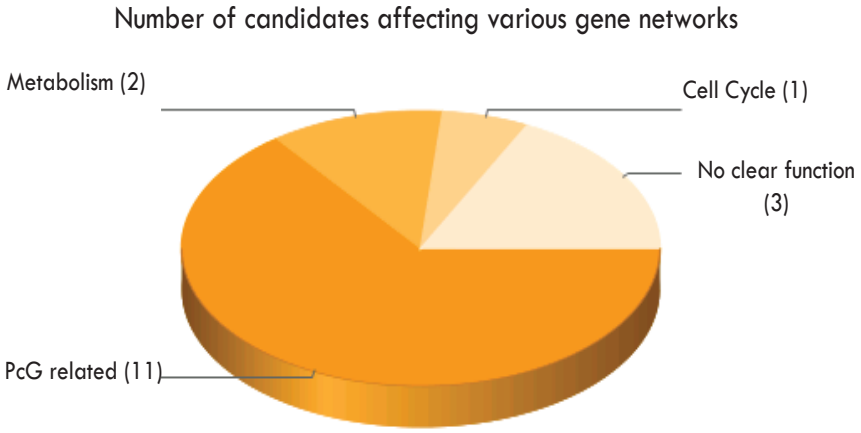

Figure 35. Knockdown of 11 of 17 candidates resulted in perturbations of the PcG gene network

Knockdown of individual candidates caused widespread changes to expression of PcG regulated genes or their targets. By this measure, the majority (11/17) candidates tested were found to be involved in PcG gene-regulatory networks. Knockdown of 3 individual candidates did not widely impact PcG gene networks, but may be involved in other biological functions. Knockdown of 3 other candidates had no clear effect.

themselves was unaffected, and no genes other than plasma membrane genes were changed in every assay. We ranked these 11 lncRNAs as high priority candidates for PcG interactions.

The top enriched gene sets in this analysis consistently suggested PRC2 involvement. However, we refrained from making any conclusions about these candidates regarding PRC1 versus PRC2 biology because the number and size of the gene sets included in the Molecular Signatures database were biased towards PRC2 (versus PRC1). This bias might be due to the (generally) poor quality of PRC1 antisera in regards to generating PRC1 component ChIP data sets, the availability of H3K27me3 data sets, and also by the redundancy and longer half-life of PRC1 proteins in knockdowns or knockout data sets. Instead, we chose to use PRC2 related gene sets as a proxy for general PcG activity.

Of the remaining six targets which did not show significant enrichment with PcG gene sets, one lncRNA was enriched for general metabolism genes, another for cell proliferation genes (perhaps S-phase), and a third for mitochondrial/electron transport chain related genes. The results are summarized in Supplemental Table 2. Lastly, the changed genes from final three targets did not show any overtly recognizable gene ontology. While a lncRNA may easily be regulating an aspect of PcG biology without exhibiting widespread changes to the PcG gene network, such pheonotypes are not readily recongnizable by this assay. Therefore, we considered these six candidates as less likely to be true PcG interactors. We conclude that 11 of the 17 lncRNAs that we tested indicate possible direct involvement with PRC1 regulation.

Impact of CAT7 KD on PcG binding

After identifying 11 candidates that disrupt the PcG gene network, we wanted to further investigate individual candidates for a role in PRC1 biology. We tested if upregulated PcG targets were coupled with a loss of PcG protein binding at the promoter sites. We were particularly intrigued by one lncRNA whose knockdown resulted in changes not only of PcG regulated gene targets, such as homeodomain proteins, but also caused upregulation of Trithorax group proteins (MLL1-4, SETD1B), several Jumanji-domain containing proteins, and an assortment of transcription factors and other chromatin proteins. This lncRNA, referred to as "Chromatin-Associated Transcript 7" (*CAT7*), is a 1.7 kbp, polyadenylated, capped RNA, composed of a single exon (Figure 36, Figure 37, Figure 38) (by RACE-PCR) and largely overlaps with a tandem repeat.

We next searched the genomic environment in the vicinity of *CAT7*, and found that *CAT7* was encoded on the same strand as an intron of a previously annotated EST/lncRNA of unknown function. This may be due to a poor annotation of the lncRNA, which is not expressed at annotated exons. Broadly, *CAT7* is located in an EZH2-rich gene desert between the developmental patterning gene *Sonic Hedgehog* (*Shh*) and the testis-specific gene, *RNF32* (Figure 39). Large deletion mutants of the syntenic region in mouse showed massive misregulation of Shh patterning in developing mouse embryos. These defects are thought to be due to loss of enhancer elements. Notably, *CAT7* is not conserved in mouse based on sequence or the presence of a tandem repeat in syntenic regions.

We also noticed that *CAT7* was transcribed directly 5' to another EST of the same strand, a liver-specific lncRNA transcript, and wanted to explore the role of *CAT7* in regulation of this transcript. Based on previous data from ENCODE, the putative promoter/enhancer region (directly 5' of the transcript, bounded by *CAT7*) showed DNAse sensitivity, a p300 like binding site, a Retinoid X

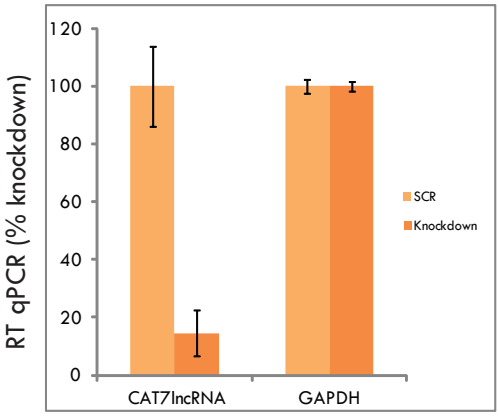

Figure 36. RT-qPCR of CAT7 lncRNA knockdown in HeLa (GAPDH normalized)
HeLa cells were nucleofected with siRNA targeting *CAT7* lncRNA or a scramble control. RNA was extracted and RT-qPCR for *CAT7* and GAPDH was performed.

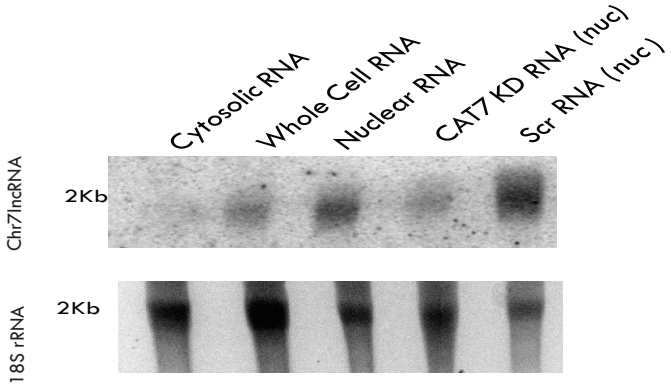

Figure 37. Northern Blot reveals *CAT7* is primarily nuclear and is efficiently knocked down.
Nuclei and cytoplasm were separated, and RNA was isolated from each. Northern of CAT7 in celluar compartments reveals that CAT7 is primarily nuclear. Similarly, Northern analysis shows that knockdown is observed in nuclear extracts upon siRNA treatment. All lanes have 20ug RNA/well; 18S EtBr loading control.

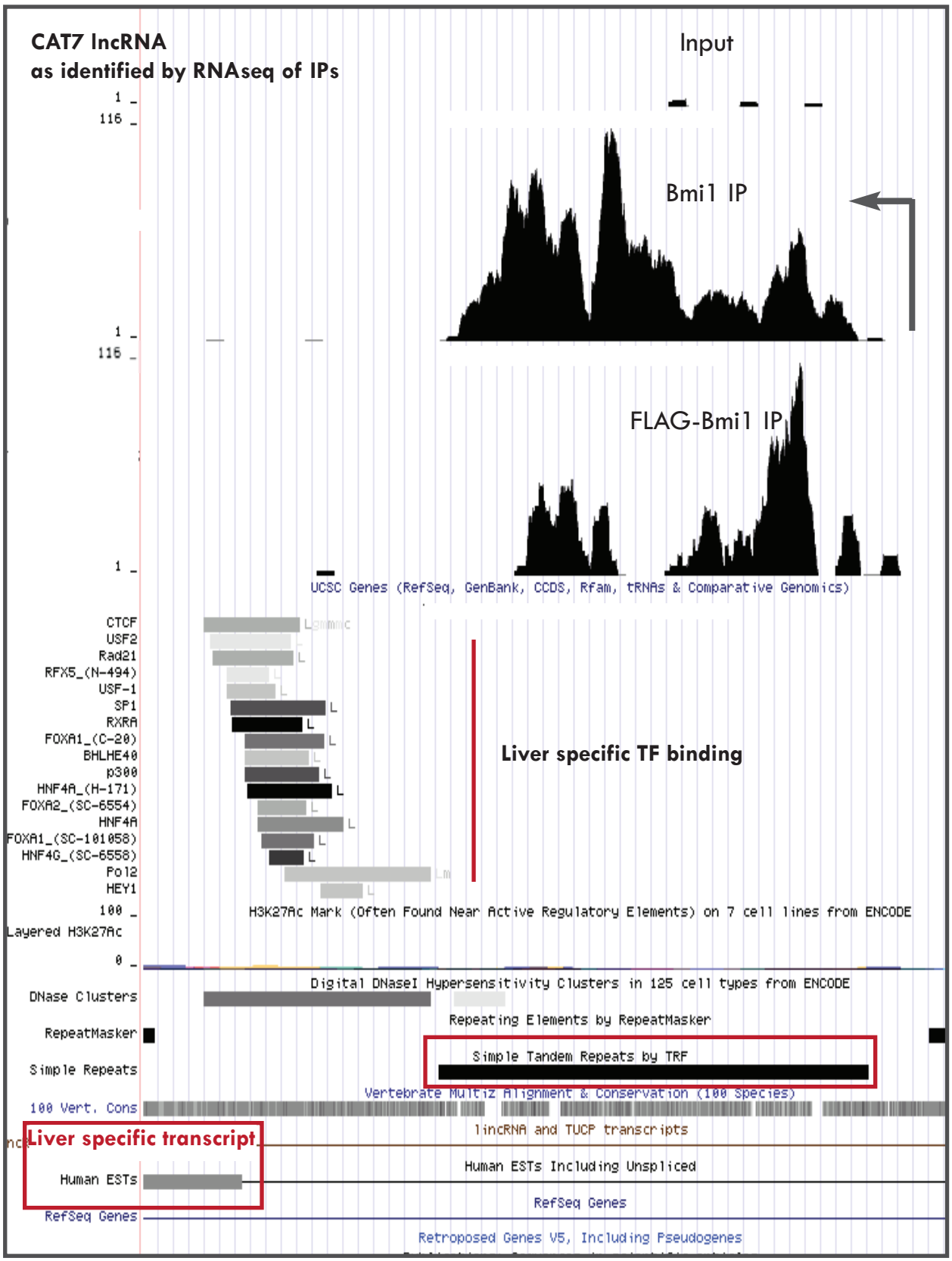

Figure 38. CAT7 Enrichment in IP vs Input by RNAseq, and the Surrounding Genomic Landscape
A nearby liver specific transcript and TF binding site, and a tandem repeat region are highlighted.

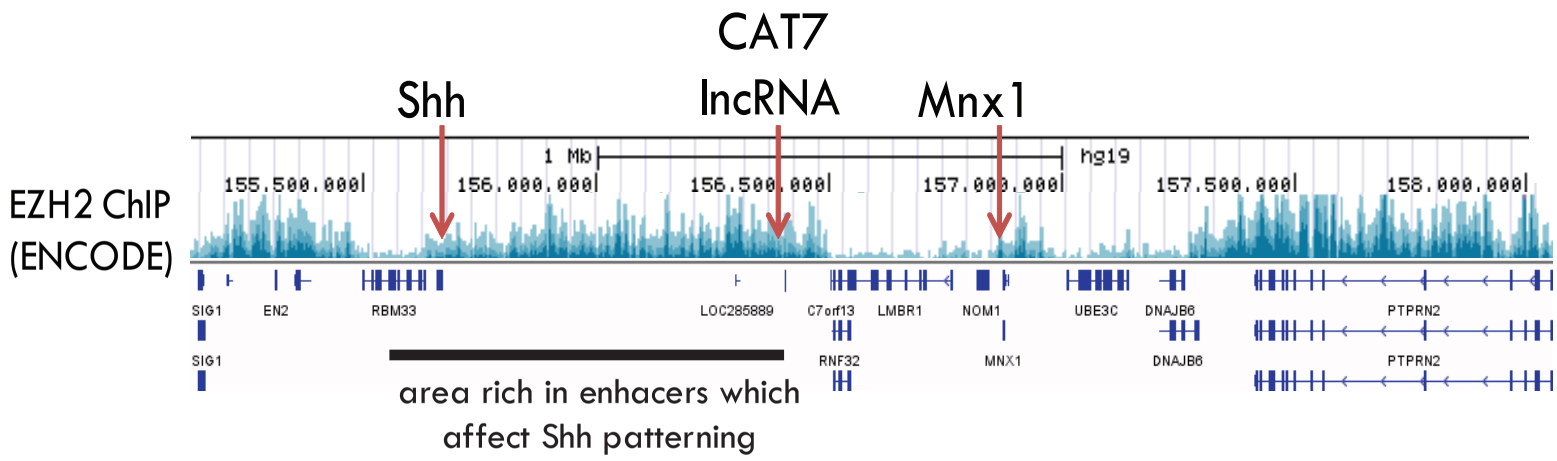

Figure 39. Broad view of CAT7 lncRNA in the genome

CAT7 lncRNA is located ~400kbp from Mnx1. It is transcribed from a gene desert between Shh and Rnf32. This region has been implicated as enhancer rich, and deletion analysis reveals (mouse) that it can affect Shh patterning. Shown above is a track of EZH2 localization in HeLa cells (ENCODE).

Receptor binding site, several HNF and Forkhead transcription factor binding sites, and a CTCF binding site (Figure 38). Presence of these proteins, as well as RNA PolII, were almost exclusively shown in liver with the exception of CTCF binding, which was present in a wide array of cell lines. We were surprised to see that genes with altered expression in the *CAT7* knockdown do not significantly overlap with gene sets pertaining to liver-related functions, at least in the context of HeLa cells. Additionally, the liver-specific lncRNA adjacent to *CAT7* is not transcribed in either the scramble or the knockdown samples. However, many differentially expressed genes are PcG regulated genes, such as *Mnx1 (up)* and *Irx3 (down),* which are also regulated by Shh signaling, or, in the case of *HoxA13 (up)*, control overlapping developmental processes [20].

We then tested whether overexpression of a PcG target gene in the *CAT7* knockdown was accompanied by a loss of PcG binding at its promoter. We were particularly intrigued by the overexpression of *Mnx1* (Figure 40, Figure 42) because it is located reasonably close to *CAT7* (~400kbp away on the same chromosome), on the border of a neighboring gene desert (Figure 39). In addition, *Mnx1* was also overexpressed (2.1 fold) in the Suz12 knockdown, but not as a consequence of any other lncRNA knockdown. To test if loss of *CAT7* caused a loss of PcG protein binding at *Mnx1*, we performed ChIP, targeting Bmi1, Suz12, and *H3K27me3* in *CAT7* knockdown versus scramble cells. We observed a 27% loss in Suz12 binding, a 34% loss in H3K27me3 signal, and a 22% loss in Bmi1 binding at the Mnx1 promoter in knockdown cells. Conversely, we saw almost no loss of ChIP signal at the control locus (HoxB6), and minimal signal in all IP's at an intergenic, negative control (Figure 41). These data show that knockdown of *CAT7* causes derepression of *Mnx1* and loss of PcG binding.

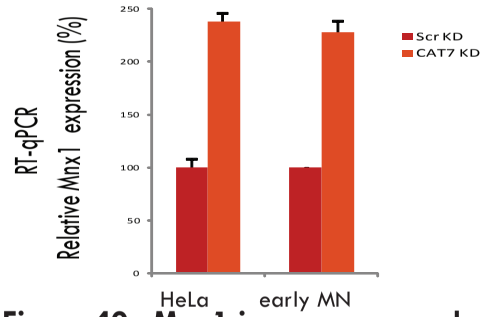

Figure 40. Mnx1 is overexpressed upon *CAT7* Knockdown

CAT7 lncRNA was knocked down in either HeLa cells or differentiating motor neurons (from ES cells). After 48 or 72 hours, respectively, Mnx1 expression was assayed by RT-qPCR. All samples are GAPDH normalized and performed in triplicate. Samples are further normalized so that the respective Scramble represents 100%. 3 (HeLa) or 2 (early motor neurons) biological replicates were performed.

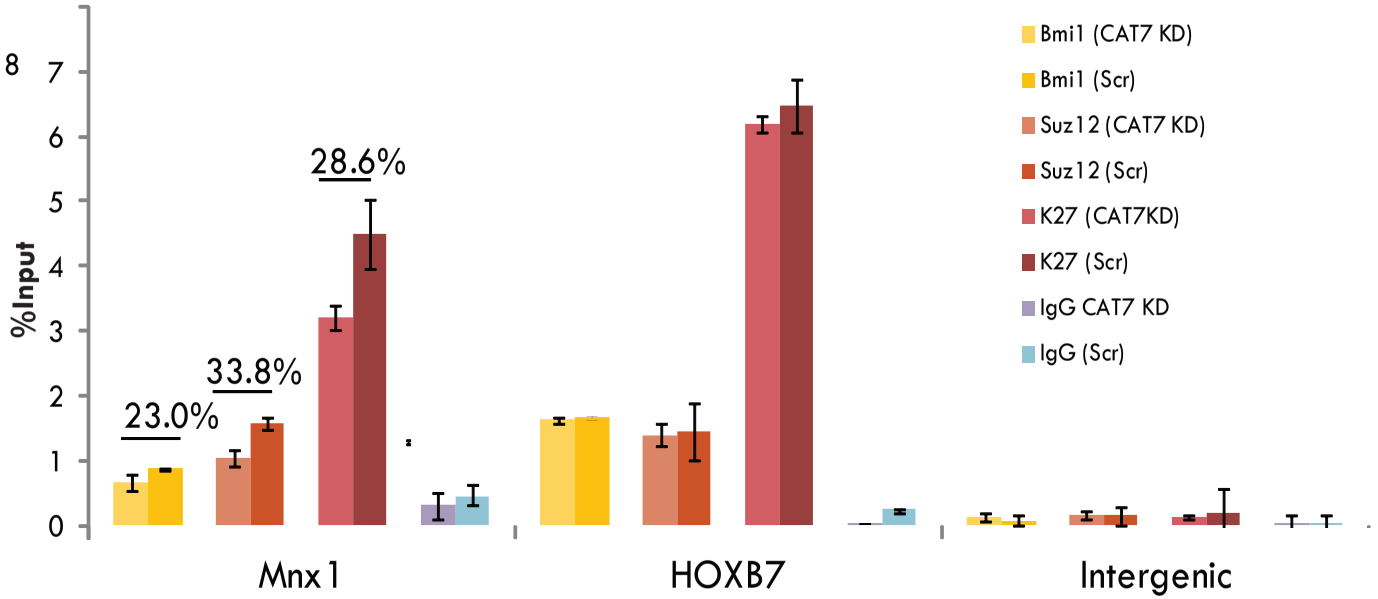

Figure 41. Polycomb Group Proteins are Selectively Lost from the Mnx1 Promoter in CAT7 Knockdown

CAT7 or a scramble control were individually knocked down in HeLa Bmi1F17 cells. ChIP-qPCR of Suz12, Bmi1, H3K27me3, and IgG negative control show that the PcG proteins are depleted at the Mnx1 promoter but not at the HOXB7 promoter. Significant depletions are quantified above the bars.

Perturbation of CAT7 during motor neuron differentiation

We wanted to further investigate the role of *CAT7* in a biological system where both *Mnx1* and PcG proteins are essential. *Mnx1* is expressed early in development as well as in adult tissue [21], and is causal of the developmental disorder *Curriano* syndrome [22]. Specifically, Mnx1 is essential for both motor neuron development and insulin-producing beta-cell formation in the pancreas. Interestingly, genomic analysis reveals that neuronal and beta-cell transcriptomes are closely related, despite their different origins in early embryogeneis (endoderm versus ectoderm) [23]. The similar expression patterns may largely be driven by an array of essential PcG-mediated regulators specific for both neuronal and beta cell differentiation, including *Mnx1, Isl1, Pax6* and *Neurod1*, which are silenced and H3K27me3 in most other cell types. We therefore were interested in neuronal formation based on the essential roles of Mnx1 and the PcG proteins.

While *Mnx1* is initially silenced in hES cells and bound by the PcG proteins [23-25], this repression is alleviated over the course of differentiation. We decided to use an *Mnx1* reporter human ES line to probe the role of *CAT7* through differentiation. The reporter cell line contains multiple insertions of eGFP under the control of a 9kb murine *Mnx1* promoter, and GFP expression has been shown previously to correspond to endogenous *Mnx1* expression [26]. In previous publications with this cell line, this protocol robustly generates GFP-positive early motor neurons after six days of differentiation. Notably, *CAT7* is expressed in human ES cells (Figure 26).

We designed a knockdown-assay in order to investigate the role of *CAT7* in motor neuron development from human embryonic stem cells (hES cells). Similar to our assay in HeLa cells, we knocked down *CAT7* in an ES line by electroporation with siRNAs. After allowing the cells to recover overnight in ES conditions, we directed the cells toward motor neuron differentiation by replacing the media with media containing neural growth factors and retinoic acid. Finally, we harvested cells at 72 hours and isolated the RNA for sequencing and analysis.

We performed the assay and saw that *CAT7* knockdown derepresses *Mnx1* expression in neuronal differentiation conditions. After verification of the *CAT7* knockdown (>71.3% in each replicate) by RT-qPCR, we tested for *Mnx1* expression relative to GAPDH in neurons (Figure 40). As seen in HeLa cells, *Mnx1* was upregulated 2.3 fold and 2.7 fold in biological replicates relative to scrambles. qPCR of eGFP revealed that GFP was expressed specifically in the knockdown cells, but at very low levels (data not shown).

Notably, *CAT7*-mediated overexpression of *Mnx1* was only observed in specific conditions. When *CAT7* was knocked down in ES cells and maintained in ES-cell media for 72 hours, *Mnx1* was not overexpressed relative to a scramble control, and was lowly expressed in both samples. Similarly, when *CAT7* was knocked down and the cells were placed in random differentiation conditions or in identical neural differentiation conditions without retinoic acid, *Mnx1* was neither highly expressed nor differentially regulated between the knockdown and scramble (data not shown). Together, these data indicate a role for the retinoic acid pathway in *CAT7* mediated derepression of Mnx1.

To more broadly test the effect of the knockdown on the PcG gene network, we sequenced the total RNA from the *CAT7* knockdown and scramble ES cells placed in neural+retinoic acid conditions. Applying the same analysis as above, we searched for genes with expression changed relative to a scramble control. We saw once again that PcG targets and developmental regulators were highly overrepresented in the list of changed genes. As expected from the RT-qPCRs, *Mnx1* overexpression (3.5 fold upregulation) was observed, as well as downregulation of the inversely related *Irx3* (51% decrease) (Figure 42, Figure 43, Figure 44). Notably, eGFP did not have much signal, and did not show changes in expression from the knockdown to the scramble at low depth (data not shown). We also saw that the neuronal master regulator *Isl1* and several pancreas or diabetes-type I genes were upregulated, including HLA proteins and

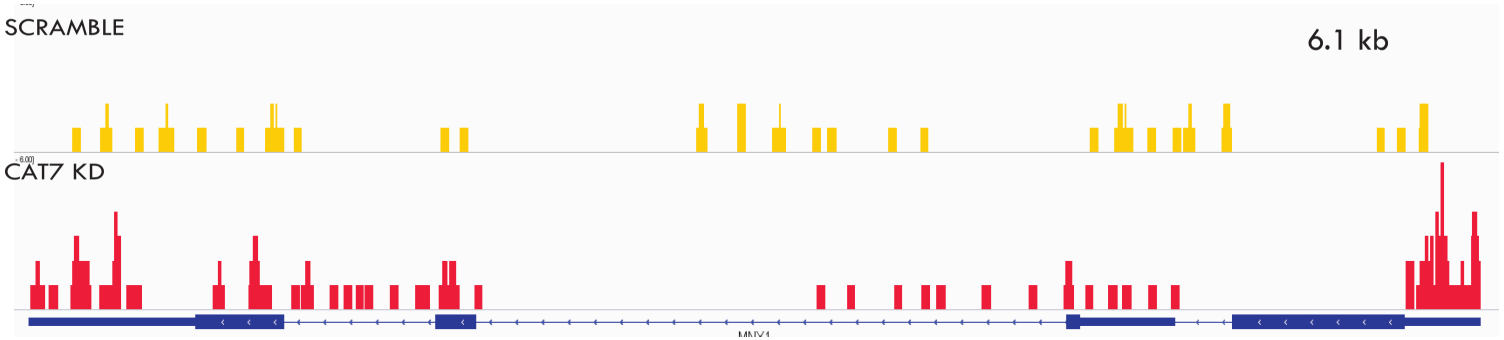

Figure 42. Mnx1 Expression in CAT7 Knockdown During Early Motor Neuron Differentiation from ES Cells
CAT7 lncRNA was knocked down in human embryonic stem cells and placed in media which promotes differentiation into motor neurons. After 72 hours, total RNA was harvested and compared to a scramble knockdown control. Several key regulators of motor neuron differentiation are affected. These regulators are controlled by both PcG proteins and Shh. Shown here is Mnx1, which is 3.5-fold upregulated.

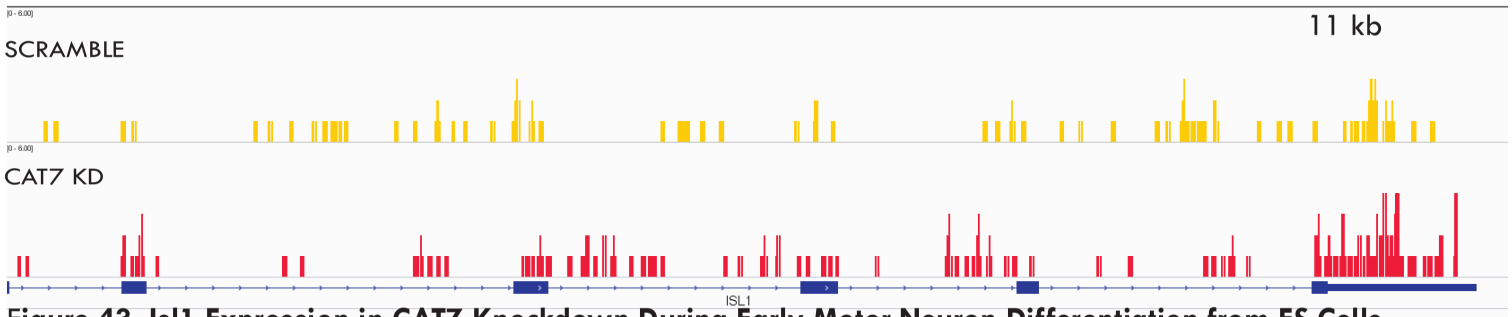

Figure 43. Isl1 Expression in CAT7 Knockdown During Early Motor Neuron Differentiation from ES Cells
Figure 44. Irx3 Expression in CAT7 Knockdown During Early Motor Neuron Differentiation from ES Cells
motor neurons. After 72 hours, total RNA was harvested and compared to a scramble knockdown control. Several key regulators of motor neuron differentiation are affected. These regulators are controlled by both PcG proteins and Shh. Shown here is Isl1, which is 2.3-fold upregulated.

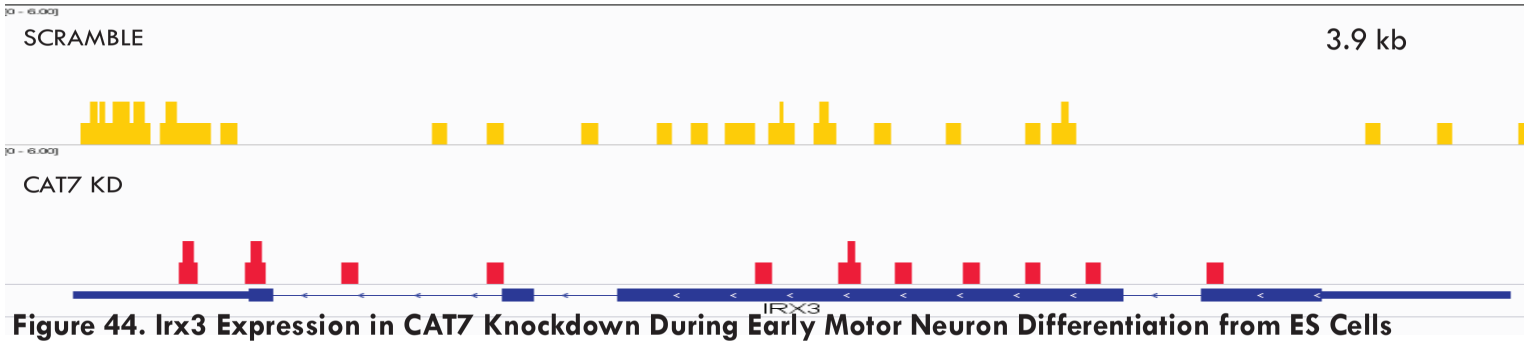

Figure 44. Irx3 Expression in CAT7 Knockdown During Early Motor Neuron Differentiation from ES Cells

CAT7 lncRNA was knocked down in human embryonic stem cells and placed in media which promotes differentiation into motor neurons. After 72 hours, total RNA was harvested and compared to a scramble knockdown control. Several key regulators of motor neuron differentiation are affected. These regulators are controlled by both PcG proteins and Shh. Shown here is Irx3, which is 69% downregulated.

GAB1. Many developmental targets listed in the Molecular Signatures database as H3K27me3 in ES cells, MLL-targets, or genes with bivalent promoters in an array of cell types were also differentially expressed. However, unlike the knockdown in HeLa cells, the Trithorax group proteins and Jumanji-domain containing proteins were not overexpressed. This indicates that *Mnx1* overexpression is not a result of elevated Trithorax group protein transcription. Furthermore, *CAT7* may be playing different roles in the context of cancer and development.

Discussion:

In our study, we investigated a list of potential PRC1-interacting lncRNAs for involvement in the PcG gene-regulatory network. We knocked down candidates in HeLa cells and searched for widespread changes in mRNA expression of PcG targets, relative to a scramble control. A striking 11 of the 17 candidates examined showed significant changes in PcG targets upon knockdown, 3 showed clear changes to other pathways such as metabolism and cell cycling, and 3 had no clear effect. In addition, the *RepE* region of *Xist*, a highly transcribed lncRNA shown previously to interact with PRC2, was also identified in this study, though the above assay was not an effective method to validate this interaction. Taken together, these data demonstrate that the wide dynamic range of the protocol used to identify these interactions (Chapter 2) allows capture of biologically relevant lncRNAs across a wide range of expression. Specifically, we have generated a technology which successfully identified multiple PRC1-interacting candidates that influence the PcG gene-regulatory network.

We also showed that knockdown of one of our candidates, *CAT7*, not only causes overexpression of the PcG/Shh-regulated gene *Mnx1*, but also leads to loss of a PRC1 component, a PRC2 component and H3K27me3 binding at the Mnx1 promoter. This is consistent with *Mnx1* derepression observed in our Suz12 knockdown. During early motor neuron differentiation from

hES cells, knockdown of *CAT7* causes significant changes in expression of several Shh/PcG regulated motor neuron master regulators. Namely, upregulation of *Mnx1* and *Isl1*, and downregulation of *Irx3,* were observed. These factors are initially repressed with H3K27me3 signatures in the ES-cell state and are activated during differentiation. Although *CAT7* is expressed in hES cells, knockdown of the transcript does not result in upregulation of Mnx1 in the absence of differentiation. Similarly, identical neural differentiation conditions which lack retinoic acid show low, non-differential *Mnx1* expression between candidates. This indicates that knockdown of *CAT7* is not sufficient to initiate derepression of *Mnx1*, and might require additional components from the retinoic acid pathway.

Furthermore, while *CAT7* depletion induced upregulation of activation factors such as the Trithorax group proteins and Jumanji domain containing proteins, in HeLa cells, none of these factors were upregulated as an effect of *CAT7* depletion in differentiating motor neurons. This could represent the different effects of *CAT7* in different cellular contexts.

The widespread transcriptional effects seen in this study evoke a mechanistic question of how lower abundance lncRNAs could elicit broad responses in PcG networks. We can further address this question by examination of one of our candidates that influences the PcG gene network. This candidate shares an exon with the previously studied transcript *DA125942*. During chondrogenesis, *DA125942* plays an important role in expression of master developmental regulators [12]. Namely, complementary to our findings, prior network analysis suggested that overexpression of *DA125942* enacts a negative feedback loop leading to decreased *PTHLH* (chondrogenic developmental regulator) expression (*cis*) and decreased *SOX9* (developmental gene) expression (*trans*). This leads to widespread changes in both EZH2 binding and gene expression of a variety of PcG targets. Though PTHLH and SOX9 are not expressed in our HeLa cells, the isoform of DA125945 found in HeLa cells may be influencing another developmental

regulator or perhaps may play tissue-specific developmental roles, based on the nuclear landscape.

DA125942 is transcribed from a regulatory locus, CISTR-ACT which is located in close physical (3D) proximity to PTHLH and Sox9 (in *trans*) during chondrogenesis. In certain forms of Brachydactyly, D125942 is upregulated, PTHLH is translocated far away from the CISTR-ACT/SOX9 locus and has reduced expression, and SOX9 expression is also reduced. Our study additionally suggests an interaction between an isoform of *DA125942* with the PcG proteins in HeLa cells, where the 3D contacts of CISTR-ACT has not been investigated. Collectively, these data could support a model where lncRNAs interact with the chromatin factors, such as the PcG proteins, to repress even *trans* targets that are in close physical proximity, and initiate a widespread signaling cascade. Such a model is consistent with previous reports of PcG proteins affecting co-localized, co-regulated regions of the chromatin (such as PcG bodies). It is also consistent with reports of various lncRNAs, such as *Xist* [5,27], *Kcnq1ot1* [28,29], and HOTTIP [30], which affect chromatin architecture and expression in close physical proximity to the RNA.

Broad application of this model can also explain how low-abundance nuclear lncRNAs reach distal targets to affect large genomic networks. Explicitly, low-expression lncRNAs could influence multiple targets by working at co-localized regions, such as in PcG bodies, to enact a signaling cascade. Expanding this model, we note that the PcG proteins are proposed to scan the genome for their targets, perhaps through an EZH2/RNA interaction. While surveying the chromatin, the PcG proteins might bind specific lncRNAs at their transcription sites, and then use the lncRNA to properly target the complex. Such a mechanism could serve to physically bring target lncRNAs to particular PcG regulated sites in both *cis* or *trans*, and in some instances might bypass the constraint that a low-expression lncRNA be transcribed in precisely the same 3D-space as its target gene.

Extending this model to *CAT7*, we see that *CAT7* and Mnx1 are located 400kb away from each other near the borders of adjacent PcG-rich gene desert regions. While *CAT7* mediated recruitment of the PcG proteins to the *Mnx1* promoter is perhaps the simplest mechanistic explanation for the interaction and derepression observed in this study, such we cannot verify that *CAT7* is localized to the *Mnx1* promoter. Technologies such as RNA/DNA-FISH, CHART, or RAP could be used to map the RNA to a distinct locus or loci on the chromatin, though these methods are not yet optimized for low abundance transcripts. Such studies could also be supplemented with DNA/DNA-FISH or 3C to gauge physical proximity of *CAT7* and *Mnx1*.

Conclusion

In our screen, we sought to validate the relevance of our candidates from Chapter 1 to the PcG gene-regulatory network. We found that the majority (11/17) of candidates tested did, in fact, yield widespread changes to PcG target-gene expression upon siRNA knockdown. Furthermore, perturbation of one candidate, *CAT7,* causes overexpression of the nearby master developmental protein *Mnx1*, and the Mnx1 promoter loses a significant portion of PcG binding. These results also validated in early motor neurons differentiating from ES cells, where siRNA depletion of *CAT7* caused misregulation of essential motor-neuron regulators. Taken together, our results show that the protocol we have developed in Chapter 1 identifies a class of lncRNAs which impact PcG recruitment, and gene-expression of PcG targets.

Methods:

HeLa Cell Culture

HeLa cells stably transduced with a copy of FLAG- tagged Bmi1 at approximately 25%

overexpression[11] were cultured in DMEM supplemented with 10% FBS (Sigma), $NaHCO_3$ pH 7.5 and gentamycin. Cells were grown in suspension in a spinner flask (Matrical) to a density of approximately $3x10^8$ cells per liter.

Human ES cell culture and differentiation to motor neurons

Human embryonic stem cells (hESCs) were maintained on plates coated with hESC-qualified Matrigel (BD Biosciences) in chemically defined mTeSR-1 medium (Stemcell technologies) and were passaged by manual picking or enzymatic digestion with TrypLE Express (Life Technologies) in the presence of 10 μM Y27632 (Sigma). For all experiments, passage number was less than 40. Media was changed daily.

hESCs were differentiated using a differentiation basal media containing 1:1 DMEM/F12 (Life Technologies) and NeuroBasal (Life Technologies), 2 mM glutamax-I (Life Technologies), 1x N2 supplement (Life Technologies), 1x B27 supplement (Life Technologies). For the non-directed differentiation, basal media was supplemented with 10% fetal calf serum (FCS) (Life Technologies). For neural differentiation, the media was supplemented with 10nM SB431542 (Sigma), 1 μM dorsomorphin (Millipore) and for certain experiments additionally with 0.1 μM retinoic acid (RA) (Sigma). Media was changed daily.

siRNA Knockdown

siRNAs were ordered from BioSciences. For HeLa cells, 10^6 cells were nucleofected in 100uL of Nucloefector R solution and 200nmoles of siRNA. Nucleofections were carried out using program I-013 on the Nucleofector II as per manufacturer's instructions. After 48 hours, cells were isolated for RNAseq or ChIP. For ES cells, the Neon torrent was used in place of the Nucleofector II.

$2x10^6$ cells were used in 1 pulse, 1200V, 20ms. All siRNA sequences can be found in Supplemental Table 2.

RNA purification, cDNA generation, and preparation for libraries for RNAseq

Whole cells, nuclei, or cytosolic were stored in Trizol (Invitrogen). Chloroform was added and the sample was spun out according to the manufacturer's instructions. The aqueous phase was applied to Zymo Clean-and concentrator 5 columns and DNAse was applied "in tube"as per the manufacturer's instructions for RNAs larger than 200 nucleotides. For RT-qPCR, cDNA was generated using SUPERscript VILO (Invitrogen). For sequencing, isolated RNA was ribo-depleted using the Ribo-Zero Magnetic Gold kit (Epicentre/Illumina) according to manufacturer instructions, and cDNA was generated with the TruSeq kit (Illumina). Libraries were constructed as previously described[5].

ChIP

Samples were dissociated from wells and pooled, and an aliquot of cells was reserved for RNA isolation (to test knockdown). ChIP was performed on the remaining cells, as previously published, but at a smaller scale[24,25]. Briefly, $5x10^6$ cells were crosslinked with 1% HCHO for 10 minutes in media, at room temperature. Bmi1 antisera[31], Suz12 (ab12703), H3K27me3 (ab6002), or rabbit IgG (Jackson Labs) was prebound to 20uL ProteinA beads in BSA (with 2ug Ab), and then added to sheared nuclear lysate from the crosslinked cells. Beads were washed in RIPA buffer four times and chromatin was eluted in Tris, SDS, and EDTA at 65C. After crosslink reversal, RNAse treatment, Proteinase K treatment and phenol-chloroform isolation, DNA was EtOH precipitated and resuspended in water for qPCR.

qPCR

qPCR was performed as per manufacturer's instructions using the Biorad iTaq 2X master mix). Primer sets can be found in Supplemental Table 1. For RT-qPCR, approximately 2ng cDNA/well was used.

Native Cell Fractionation

Nuclear isolation was performed in native cells, as previously published[32,33] and checked by hemocytomter (>97%) with Trypan blue. Cytosolic extract was also reserved, alongside whole cell extract. All extracts were immediately stored in Trizol.

Northern Blot and RACE Analysis

Nuclei were isolated from native cells to >97% purity by Trypan Blue staining on a hemocytometer. Northern blot analysis was performed using RNA probes as in previously published protocols[34], with the exception that a Hybond-N+ membrane was used instead of nitrocellulose. RACE was performed using the Ambion RLM-RACE kit (AM1700) as per manuafacturer's instructions.

The Northern probe sequence was as follows:

AACAAAGCCUGAGUCGAACACGAAAGGAAGAUGGUCGCUGAAGCGAAGGGGAGUCAUUU
GUGUCCGUUCCAUAAAUCAAGACUGUCGCCUUUCGAAAAGGGGAGGUGUCGCAGUCUGA
CAGCCUGAUCUGUUUCUAGGACGGCGUGUUUCCAGGAA

Gene Set Enrichment

RNAseq data was aligned to the hg19 build of the genome. Only uniquely mapped, non-duplicated reads were included. This yielded an average of 15M reads per lane. Samples were normalized by total reads, and total reads mapping back to an mRNA (only at exonic regions, as defined by RefSeq[15]) were compared in a scramble control versus an individual knockdown. Removing any mRNA that had fewer than 5 reads in either lane, we then searched for mRNAs that showed greater than 2-fold difference between a sample and the scramble control. This analysis included multiple isoforms. However, after identifying differentially expressed genes, we only considered one isoform per differentially expressed gene, to avoid bias. We entered differentially expressed genes into the Molecular Signatures Database[16] and compared our gene set to 6,791 others (default parameters, but excluding cancer gene sets and TF binding site motifs). We then identified if Suz12, PRC1, or H3K27me3 related gene-sets were significantly enriched using a broad null hypothesis (that all genes in the Molecular Signatures Database were expressed above background). We then calculated true p-values for the enrichment of these sets by considering the total number of genes actually expressed above background (as before) in the cell, which were also identified in the Molecular signatures database. We used a hypergeometric distribution to generate p-values and (rounded) expected values. A summary of the results is found in Supplemental Data Table 2. A schematic of this process is also found in Figure 30.

References

1 Mercer, T. & Mattick, J. Structure and function of long noncoding RNAs in epigenetic regulation. *Nature structural & molecular biology* **20**, 300-307, doi:10.1038/nsmb.2480 (2013).

2 Walstrom, K., Dozono, J. & von Hippel, P. Effects of reaction conditions on RNA secondary structure and on the helicase activity of Escherichia coli transcription termination factor Rho. *Journal of molecular biology* **279**, 713-726, doi:10.1006/jmbi.1998.1814 (1998).

3 Riley, K., Yario, T. & Steitz, J. Association of Argonaute proteins and microRNAs can occur after cell lysis. *RNA (New York, N.Y.)* **18**, 1581-1585, doi:10.1261/rna.034934.112 (2012).

4 Zhao, J., Sun, B., Erwin, J., Song, J.-J. & Lee, J. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science (New York, N.Y.)* **322**, 750-756, doi:10.1126/science.1163045 (2008).

5 Simon, M. *et al*. High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. *Nature*, doi:10.1038/nature12719 (2013).

6 Maertens, G. *et al*. Several distinct polycomb complexes regulate and co-localize on the INK4a tumor suppressor locus. *PloS one* **4**, doi:10.1371/journal.pone.0006380 (2009).

7 Klattenhoff, C. *et al*. Braveheart, a long noncoding RNA required for cardiovascular lineage commitment. *Cell* **152**, 570-583, doi:10.1016/j.cell.2013.01.003 (2013).

8 Rinn, J. *et al*. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**, 1311-1323, doi:10.1016/j.cell.2007.05.022 (2007).

9 Davidovich, C., Zheng, L., Goodrich, K. & Cech, T. Promiscuous RNA binding by Polycomb repressive complex 2. *Nature structural & molecular biology* **20**, 1250-1257, doi:10.1038/nsmb.2679 (2013).

10 Kaneko, S., Son, J., Shen, S., Reinberg, D. & Bonasio, R. PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells. *Nature structural & molecular biology* **20**, 1258-1264, doi:10.1038/nsmb.2700 (2013).

11 Levine, S. *et al*. The core of the polycomb repressive complex is compositionally and functionally conserved in flies and humans. *Molecular and Cellular Biology* **22**, 6070-6078, doi:10.1128/mcb.22.17.6070-6078.2002 (2002).

12 Philipp, G. M. *et al*. A misplaced lncRNA causes brachydactyly in humans. *Journal of Clinical Investigation* **122**, doi:10.1172/jci65508 (2012).

13 Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573-580, doi:10.1093/nar/27.2.573 (1999).

14 Consortium, E. P. *et al*. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

15 Pruitt, K., Tatusova, T. & Maglott, D. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* **35**, 5, doi:10.1093/nar/gkl842 (2007).

16 Liberzon, A. *et al*. Molecular signatures database (MSigDB) 3.0. *Bioinformatics (Oxford, England)* **27**, 1739-1740, doi:10.1093/bioinformatics/btr260 (2011).

17 Squazzo, S. *et al*. Suz12 binds to silenced regions of the genome in a cell-type-specific manner. *Genome research* **16**, 890-900, doi:10.1101/gr.5306606 (2006).

18 Jiang, Y. *et al*. Effect of siRNA-mediated silencing of Bmi-1 gene expression on HeLa cells. *Cancer science* **101**, 379-386, doi:10.1111/j.1349-7006.2009.01417.x (2010).

19 Tatusova, T. & Madden, T. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS microbiology letters* **174**, 247-250, doi:10.1111/j.1574-6968.1999.tb13575.x (1999).

20 Sakamoto, K. *et al*. Heterochronic shift in Hox-mediated activation of sonic hedgehog leads to morphological changes during fin development. *PloS one* **4**, doi:10.1371/journal.pone.0005121 (2009).

21 Hagan, D. *et al*. Mutation analysis and embryonic expression of the HLXB9 Currarino syndrome gene. *American journal of human genetics* **66**, 1504-1515, doi:10.1086/302899 (2000).

22 Köchling, J., Karbasiyan, M. & Reis, A. Spectrum of mutations and genotype-phenotype analysis in Currarino syndrome. *European journal of human genetics : EJHG* **9**, 599-605, doi:10.1038/sj.ejhg.5200683 (2001).

23 van Arensbergen, J. *et al*. Derepression of Polycomb targets during pancreatic organogenesis allows insulin-producing beta-cells to adopt a neural gene activity program. *Genome research* **20**, 722-732, doi:10.1101/gr.101709.109 (2010).

24 Boyer, L. *et al*. Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* **441**, 349-353, doi:10.1038/nature04733 (2006).

25 Lee, T. *et al*. Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125**, 301-313, doi:10.1016/j.cell.2006.02.043 (2006).

26 Hester, M. *et al*. Rapid and efficient generation of functional motor neurons from human pluripotent stem cells using gene delivered transcription factor codes. *Molecular therapy : the journal of the American Society of Gene Therapy* **19**, 1905-1912, doi:10.1038/mt.2011.135 (2011).

27 Engreitz, J. M. *et al*. The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science* **341** (2013).

28 Redrup, L. *et al*. The long noncoding RNA Kcnq1ot1 organises a lineage-specific nuclear domain for epigenetic gene silencing. *Development* **136**, 525-530 (2009).

29 Mohammad, F., Mondal, T., Guseva, N., Pandey, G. & Kanduri, C. Kcnq1ot1 noncoding RNA mediates transcriptional gene silencing by interacting with Dnmt1. *Development (Cambridge, England)* **137**, 2493-2499, doi:10.1242/dev.048181 (2010).

30 Wang, K. *et al*. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472**, 120-124, doi:10.1038/nature09819 (2011).

31 Woo, C., Kharchenko, P., Daheron, L., Park, P. & Kingston, R. A region of the human HOXD cluster that confers polycomb-group responsiveness. *Cell* **140**, 99-110, doi:10.1016/j.cell.2009.12.022 (2010).

32 Michael , Y. T. *et al*. Histone Variant H2A.Bbd Is Associated with Active Transcription and mRNA Processing in Human Cells. *Molecular cell* **47**, doi:10.1016/j.molcel.2012.06.011 (2012).

33 Tolstorukov, M., Kharchenko, P., Goldman, J., Kingston, R. & Park, P. Comparative analysis of H2A.Z nucleosome organization in the human and yeast genomes. *Genome research* **19**, 967-977, doi:10.1101/gr.084830.108 (2009).

34 Ardehali, M. *et al*. Spt6 enhances the elongation rate of RNA polymerase II in vivo. *The EMBO journal* **28**, 1067-1077, doi:10.1038/emboj.2009.56 (2009).

CONCLUSIONS
AND FUTURE DIRECTIONS

In our study, we developed a general protocol to identify lncRNAs that interact with chromatin proteins (CATs). By applying this protocol to Bmi1 of PRC1, we identified at least 11 novel candidates whose expression impacts PcG binding and/or PcG target gene expression. One such lncRNA, *CAT7,* also influences transcription of key neuronal factors during early motor-neuron differentiation from human embryonic stem cells.

Our final efforts in the lab will be to further characterize *CAT7* and a few other lncRNAs. We have generated libraries from the PcG ChIP-qPCR in *CAT7* knockdowns from Chapter 2. In this way, we can look genome-wide for the effects of knockdown on PcG recruitment. Similarly, we have knocked down 2 more CAT's and are awaiting ChIP sequencing. We will also compare our results with HiC data to look for co-localization of affected genes and the lncRNA transcription site.

Interestingly, *CAT7* is conserved by sequence in Zebrafish. Preliminary of results show that *CAT7*-targeted morpholino leads to a sinusoidal back defect in developing zebrafish. Experiments to try to rescue the phenotype with the human RNA are underway. In addition, we want to gauge the impact of Rep E deletion in mouse or human, on PRC1 recruitment to the inactive X by IF. This experiment will be done either by LNA-knockoff + IF, or by IF in deletion mutants.

Finally, others in the lab may try to optimize CHART or RAP to look at where on the chromatin the CATs are binding. As of yet, these technologies have only worked for very highly abundant lncRNAs, so the feasibility of these protocols on low abundance transcripts is yet unclear. The lab may also use this protocol to look at interactions with other chromatin proteins as well. It is with humility and great pleasure that I see my work continuing in others' hands and I look forward to seeing how it evolves.

Thank you for your time and attention.

SUPPLEMENTAL FIGURES

Supplementary Table 1: Candidates and RT-qPCR primers

Selected Candidate	Primer F	Primer R
chr1:1823225-1823877	GTGTCCTCAGAAGCCTCCAG	CGGAACTCAGCTGGAAAGAC
chr1:4580194-4581300	NA	NA
chr1:37897029-37897593	CTTGCAAGTGCTAGACTGGAG	GGTGAATTGGGTGTGCCTGTA
chr1:201501623-201502247	NA	NA
chr1:228258960-228260359	TGCTTTGTGTCTGGAGCATC	GATGGATGGGATTCTCAGGA
chr2:239861118-239862260	CAAGGAACCCATCCATCATC	GATGGATGGGATTCTCAGGA
chr2:7907588-7908632	GCTCTTCCACTTGTCCCTTG	TGAGAGCAGGTGAACCATGA
chr2:12302277-12302878	CCGAAATATGAAGGTCTCAGG	CTTTGCAGAGGTCACAGCAT
chr2:45414788-45416271	TATCGCACCTTCTTCCTGCT	GATTGAAAGGCACCCAAGAG
chr2:53109804-53110731	TATGTAATGCCTGGCTGCAC	TGATGGGAATAGTCAGGATGG
chr2:67898557-67900277	GCAGATTTGCACTTGGCTTT	TTCTTCCTGGCTTGCTGTTT
chr2:178016451-178017163	CATGGAAGATCTAGTCTTTCTATCTG	TCCAAATGAGAGTGGGTAGGTT
chr2:231555671-231555978	CACTCACACACACCCACACA	GTGGTTCCCTTGCATGATTT
chr3:13504501-13505115	GCAAACAGGGACATGGTTCT	TGCCTGACATCCTCTCTGTG
chr3:106848922-106849704	TCAGTGACCATCTCCTGCTG	GAGGTGGTGGCCATAAGAGA
chr3:126880354-126881469	ACCTCTGTCCCTCCTCCACT	TCTGGGAATGTTATTGCTCTGA
chr3:141416707-141417444	NA	NA
chr3:159814590-159815187	CCCTGGATTTACCACCTCCT	GTTAAGGCCAAATGGGTGAA
chr3:185680927-185681255	CCTATATTTCGCCACGCATC	GGAACAGAAGGGTACCGTGA
chr3:194492580-194493466	TGTGTTAGGGAAGGGCTGTC	ATCCCTGCTCCTGTCAGAAA
chr5:610361-611541	NA	NA
chr5:2705666-2706246	CTGGTGAGGTGTGAGCTGTG	CTGAGCCTGTACCACCTTCC
chr5:93903900-93906322	TAACAGGCCAACCTCGTTTC	CCGATTGTTAGGCGTTTGAT
chr7:141356194-141356525	ACAGGCTCTGCTCATGGAGT	CTGCAATCCCTTCTCCAGTC
chr7:156309548-156310990	CTGCATCAGGGAGGCTATGT	TCATGACAGCCTCCTTCACA
chr8:25148305-25148931	TCTGGAGCCTTGGAGAGAAA	ACACATTCCCAGCTGCATTT
chr8:38359987-38361309	CCCACACCTCTCCCTCCTAC	GCTGGACGACTTGACCTCTT
chr8:128269380-128271023	GGTGGGAGGGTAGGGTTTAC	CCTCCAACAACTGCTTCACA
chr8:134414697-134416102	CAGTGCTCAGGGATGTGAAA	ACAAGGGATGTGAGCAGCTT
chr8:144098779-144099360	NA	NA

Supplementary Table 1: Candidates and RT-qPCR primers (continued)

Selected Candidate	Primer F	Primer R
chr9:21995862-21996354	AGGAGCGGCAGACTTCTCTT	CTTGGGTTTGTACCGAGGTC
chr9:126108629-126110509	GTTCCAACTTTGGCATGTGA	GAACAACATGGCGCTCACT
chrX:45709542-45709894	CTCCTCCAAGTCCCACAGTC	AGGTGGTGGTGCTTGTGTCT
chrX:48392427-48393414	GTCACTCTCCACAGACCTGC	GACGGCAGGATGAGTGATGG
chrX:48909990-48910823	NA	NA
chrX:73047189-73047826	GAGGCAGAAAGGAAATGCAG	TTTCTGAGTCTGGCCTCCTT
chr10:3720830-3721905	TGTGCACCATTATCCACCTT	GGTGCACATTTGGGATGTAA
chr10:48322537-48323637	CTGTGATGGATGCGTGTTCT	TTCTCCTACAGGGTGCCAGT
chr10:119260104-119260831	NA	NA
chr10:125152436-125153312	TGGCAGAGGAACAGAGATCA	CTGAAAGCAAGTCTGCACAA
chr11:2018379-2019060	CGCTGCTGCCAGCTACACCTC	TGGCAGCTGGTTGGACGAGG
chr11:45516590-45517148	CCTCTTTCTCCTCCCTGTGG	CAACAAGTGGGTCTGTGTGG
chr11:47620608-47621318	NA	NA
chr11:63690379-63691148	AGAAACCGTTCCTCCAAGGT	TTGCATCACCCTCAGTGCCAG
chr11:74738077-74738590	AAGACAAACATGAGTAAGGAG	GTGGAGGCTGGTTTGAAGAG
chr12:2861268-2861837	CCCTACCGATGGTCTGGTTA	GGGATGACGAAGGAACGATA
chr12:53369800-53370300	TGATGGCCTCTCATTGTTGA	GGCACTTGAGTAGGGAGGTG
chr15:74058758-74059747	AAGCCTGGTGCTGATGGTAG	CCTGTATTGCATTCCTTCACC
chr15:98543931-98546173	GTGGCATGCAGTACGTGTTA	CACACATTGCCTACACCACA
chrX:48392427-48393414	TTGAGAGATGGGATGGTGGT	ACCTTGCAGAAGCCATCACT
chr16:56058009-56060021	CCTTTACAAACACTCCTTGGT	GGAGAGGACTGAGGAGTGTGG
chr16:80604005-80604515	NA	NA
chr17:38268301-38269007	CCTGGGTAGTGTGCTTCCTG	TGGACAGTGAATGGTCCAGA
chr17:72072155-72073142	TGGCGTTGGTGTTAGTCAAA	AGGAGAGGGTTTGGCATGAT
chr17:72094493-72095287	NA	NA
chr19:2361908-2363132	TGTGTCATGTCATGCGTGATT	GCCATCGACACACAACCATA
chr19:36476415-36477025	CATCTCCTGCTGCTGTTCAA	GAGGTGGTGGCCATAAGAGA
chr19:55959097-55960730	NA	NA
chr20:47130339-47131227	CCCTCTCTCCAGCTACACGA	GTTCAAATCCGAGGTCATGG
chr20:51039856-51041260	NA	NA

Supplementary Table 1: Candidates and RT-qPCR primers (continued)

Selected Candidate	Primer F	Primer R
chr20:59792756-59793915	GTGTAGGTGCAGGGACAGGT	CATCTACAGCTGCGTCTCCA
chr20:60523756-60524862	AGCCATTAACTACACTAACCG	TTACTGGACATAGTGAGTTGG
chr21:44599637-44600999	GCCCAGGTACTTGGTTTCTG	TGGCTGACACTGGACATTTG
chr21:44804640-44805877	NA	NA
GAPDH	GACAAGCTTCCCGTTCTCAG	GAGTCAACGGATTTGGTCGT
CISTR-ACT	GGAAGCTTTTGTGAGCTGGC	CTCTGACTGTGGAGAGGGGA

Supplementary Table 1: Candidates and RT-qPCR primers (continued)

Selected Candidate	Validates?	nuclear?
chr1:1823225-1823877	Y	y
chr1:4580194-4581300	mult melt	ND
chr1:37897029-37897593	Y	y
chr1:201501623-201502247	mult melt	y
chr1:228258960-228260359	**Y**	y
chr2:239861118-239862260	Y	y
chr2:7907588-7908632	N	y
chr2:12302277-12302878	Y	y
chr2:45414788-45416271	mult melt	y
chr2:53109804-53110731	mult melt	y
chr2:67898557-67900277	Y	y
chr2:178016451-178017163	N	y
chr2:231555671-231555978	Y	BOTH
chr3:13504501-13505115	**Y**	y
chr3:106848922-106849704	**Y**	y
chr3:126880354-126881469	Y	ND
chr3:141416707-141417444	mult melt	y
chr3:159814590-159815187	**Y**	y
chr3:185680927-185681255	**Y**	y
chr3:194492580-194493466	**Y**	y
chr5:610361-611541	mult melt	y
chr5:2705666-2706246	**Y - intronic**	y
chr5:93903900-93906322	**Y**	BOTH
chr7:141356194-141356525	Y	y
chr7:156309548-156310990	**Y**	y
chr8:25148305-25148931	Y	y
chr8:38359987-38361309	N	y
chr8:128269380-128271023	N	BOTH
chr8:134414697-134416102	**Y**	y
chr8:144098779-144099360	mult melt	y

Supplementary Table 1: Candidates and RT-qPCR primers (continued)

Selected Candidate	Validates?	nuclear?
chr9:21995862-21996354	ANRIL	y
chr9:126108629-126110509	Y	y
chrX:45709542-45709894	Y	y
chrX:48392427-48393414	**Y**	y
chrX:48909990-48910823	mult melt	y
chrX:73047189-73047826	**Y XIST**	y
chr10:3720830-3721905	N	ND
chr10:48322537-48323637	**Y**	y
chr10:119260104-119260831	mult melt	y
chr10:125152436-125153312	Y	ND
chr11:2018379-2019060	negative ctrl	BOTH
chr11:45516590-45517148	Y	ND
chr11:47620608-47621318	mult melt	y
chr11:63690379-63691148	Y	BOTH
chr11:74738077-74738590	Y	y
chr12:2861268-2861837	Y	y
chr12:53369800-53370300	N	BOTH
chr15:74058758-74059747	Y	y
chr15:98543931-98546173	Y	y
chr16:58468257-58468920	**Y**	y
chr16:56058009-56060021	**Y**	ND
chr16:80604005-80604515	mult melt	y
chr17:38268301-38269007	Y	y
chr17:72072155-72073142	Y	y
chr17:72094493-72095287	mult melt	y
chr19:2361908-2363132	Y	y
chr19:36476415-36477025	**Y**	y
chr19:55959097-55960730	mult melt	y
chr20:47130339-47131227	Y	y
chr20:51039856-51041260	mult melt	y

Supplementary Table 1: Candidates and RT-qPCR primers (continued)

Selected Candidate	Validates?	nuclear?
chr20:59792756-59793915	Y	ND
chr20:60523756-60524862	**Y**	BOTH
chr21:44599637-44600999	N	y
chr21:44804640-44805877	mult melt	y
GAPDH	N	CYTO
CISTR-ACT	Y	y

Supplementary Table 2

name	Chr location	Tandem rep	lncRNA
Xist RepE	chrX:73047189-73047826	RepE	lncRNA
CAT7	chr7:156309548-156310990	Y	lncRNA
CAT1	chr3:194492579-194493466	N	lncRNA
CAT4	chr16:58468257-58468920	Y	lncRNA
CAT12	chr5:2705665-2706246	Y	NO
CAT-CISTRACT	chr12:54,150,530-54,150,728	N	lncRNA
CAT16	chr16:56058009-56060021	Y	NO
CAT15	chr19:36476415-36477025	N	NO
CAT9	chr3:185680927-185681255	N	lncRNA
CAT11	chr1:228258959-228260359	Y	NO
CAT13	chr20:60523755-60524862	Y	NO
CAT14	chr3:13504500-13505115	N	NO
CAT10	chr10:48322537-48323637	Y	lncRNA
CAT5	chr5:93903900-93906322	N	INTRON
CAT3	chr3:159814590-159815187	N	NO
CAT8	chr8:134414697-134416102	Y	NO
CAT6	chrX:48392427-48393414	N	NO
CAT2	chr3:106848922-106849704	Y	lncRNA
Suz12		N	mRNA
Bmi1		N	mRNA
Scramble	NA	N	Neg Ctrl

Supplementary Table 2 (continued)

name	siRNA 1	siRNA 2	%KD vs Scr (GAPDH)
Xist RepE	UUCCCCUUCCCCAUUGUUUAU	UUCUCUCCCUCUAUCUAAGUA	48.0
CAT7	AACCUACAUGACAGCCUCCUU	GGAGGCUGUGUGGGAGGCUUU	85.6 HeLa; 67.3 hES
CAT1	UAGAUGUGCGUAUACCUUCUU	UGGAUGAACAAGAACAGCCUCUU	85.9
CAT4	CACUGCUGCAUUCAGUCCAUU	UGAUCAGCACCUAUCUCCCUU	83.2
CAT12	ACAGUAACCCUGCAUCCCAUU	AUGGUGGUGCAGGUUCACUGUUU	89.3
CAT-CISTRACT	UAGUUUCCAAGGGCCACAGUU	UUCACUCUUUGGUAGGUCCUGUU	91.3
CAT16	UUCACACACCACGUACUUACC	UUGGAGGAGUUAGUACAUGGU	90.7
CAT15	CUCCCCUGGAGAGUAAAUGGGUU	UAUCCAUUCUCUUAUUGCCCCUU	78.9
CAT9	UAUUAGCUGUCUACAAGGUUU	GCAUUUAGCAAGAUGUUGAUGUU	77.3
CAT11	UACACUAUCACCUACACUAUU	AUAGUAUAAAGGGAUAGUGUAUU	88.5
CAT13	UAGUGUACUUAAUGGGUUGUU	UCAACCAGCCCAUUAUCAACAUU	74.2
CAT14	ACUAGAGGCUCAUCCUAAGUU	UCUCAAAUGGAGGGUAUUCUGUU	71.3
CAT10	UUCCAUAAUUGUGCGUGAUGU	UUGGUAAGGGGUAUCCAGUGG	72.2
CAT5	UUGGUGUGGUGAUUGUUAGUU	UUUAAGCUCACGAUAACCGCG	97.5*
CAT3	GAAUUUGUUCGAAGUAGGUAC	CACCAUUUAGGUCCCUUAUUC	87.4*
CAT8	UUUGAGACAAGACGGAAGACU	UUCUCUUCACUUUCCCCUCCC	99.5*
CAT6	UGUAGAGUGAGUGAAGUGAUU	UAUGAUCAGGGGAUAAUAAACUU	86.3
CAT2	UCAUUCAUCCUCAGUGCCCUU	AGUACUAAAUGGAGGCCAGUGUU	84.9
Suz12	AAGCUGUUACCAAGCUCCUGU		70.3
Bmi1	AAGCAGAAAUGCAUCGAACAA		85.4
Scramble	UUCUCCGAACGUGUCACGUTT	ACGUGACACGUUCGGAGAATT	-

*high error = 10%-20%

Supplementary Table 2 (continued)

name	Genes in comparison	Select Changed Genes	Select Enriched Gene Sets	Overall
Xist RepE	141			N/A
CAT7	614 (hela), 728 (MN)	1, HOXA13, KDM6B, MLL1-4; MN: Mnx1, Isl1, Irx3		PcG/TrX
CAT1	185	SOX12, PAX9	ell H3K27me3, Brain Bivalent, MLL tar	PcG-related
CAT4	492	HOXC4, PITX2, HOXC5, HOXB8	27me3, Suz12 Targets, Organismal De	PcG-related
CAT12	409	PAX8, PHOX2A, PITX3	uz12 Targets, Eed Targets, PRC2 targe	PcG-related
CAT-CISTRACT	235	HOXC4, ZFPM1, HIC1	ES withK27me3, MCV6 with H3K27me3	PcG-related
CAT16	519	SOX9, SOX17, ZHX2	Brain Bivalent	PcG-related
CAT15	487	DLL4, RUNX1	H3K27me3 in ES, Suz12 Targets	PcG-related
CAT9	294	DLHX	uz12 Targets, Eed Targets, PRC2 targe	PcG-related
CAT11	324			PcG-related
CAT13	319			PcG-related
CAT14				PcG-related
CAT10	259			metabolism
CAT5	94			N
CAT3	441			cell cycle
CAT8	522			N
CAT6	295	ATP and NDU genes		mitochon/ETC
CAT2	214			N
Suz12	236	Mnx1, Suz12, PHC1		
Bmi1	167			
Scramble	0			