Supporting Information

Khalil et al. 10.1073/pnas.0904715106

SI Methods

Data Availability. The data concerning the large intergenic noncoding (linc)RNAs and the experiments here are freely available in Dataset S1 and public databases. This includes: coordinates of the K4-K36 domains in the 6 human cell types and associated codon substitution frequency (CSF) scores (table 1 in Dataset S1); coordinates of lincRNA exons defined by Nimblegen tiling microarrays and associated Pi LOD conservation scores (table 2 in Dataset S1); list of lincRNAs bound to polycomb repressive complex (PRC)2 in the 3 cell types examined (table 3 in Dataset S1); list of lincRNAs expressed in the 3 cell types examined (table 4 in Dataset S1); quantitative (q)PCR validation of PRC2-bound lincRNAs (table 5 Dataset S1); lincRNAs bound by CoREST (table 6 Dataset S1); gene ontology of TUG1 knock-downs (table 7 in Dataset S1); DNA probes used to visualize lincRNAs in situ (table 8 Dataset S1); DNA primers used to validate PRC2 enrichments and siRNA depletions (table 9 in Dataset S1); siRNA sequences used for each lincRNA depletion (table 10 in Dataset S1); and genes affected by siRNA-mediated depletion of lincRNAs are listed in table 11 of Dataset S1 . All microarray data including RNA hybridization to tiling arrays, RIP-chip experiments, and gene expression profiling of lincRNA knock downs is deposited at the Gene Expression Omnibus (GEO) under accession no. GSE16226.

Chromatin Map Data. Chromatin data in normal kidney were generated as previously described (1). Chromatin data for other cell types were downloaded from the GEO. Data for human embryonic stem cells were obtained from Ku et al. (GSE13084) (2), human lung fibroblasts (hLFs) from Guttman et al. (GSE13765) (3), human T cells from Barski et al. (4), and CD133 and CD36 hematopoietic stem cells from Cui et al. (GSE12646) (5).

Identifying K4-K36 Enriched Domains. K4-K36 domains were identified as previosuly described (3). Briefly, we used a sliding window approach across the genome and assessed significance of each window by computing the probability of observing the number of reads for any window of fixed size in the genome under a background model. We filtered the list of enriched domains to eliminate known protein-coding genes and miRNAs. Protein-coding genes were defined as all human, mouse, rat, and dog annotated genes as previously described (3), as well as additional genes identified by Clamp et al. (6).

CSF and Conservation Scores. CSF and conservation scores were calculated for K4-K36 domains and exonic structures as described (3, 7, 8).

Tilling Array Design, Hybridization, and Analysis. High-resolution DNA tiling arrays containing 2.1 million features were designed on the Nimblegen platform (HD2) to represent ≈1,100 lincRNA K4-K36 domains. We hybridized polyA amplified total RNA as previously described (3, 9). RNA was obtained from Ambion, and consisted of brain, breast, embryonic kidney, lung, ovary, skin, spleen, testis, and thymus. Also, we extracted total RNA from primary LFs, primary foot (F)Fs, HeLa cells, K562, and human H9 ES cells. After hybridization to our tiling array, we identified transcribed regions using a sliding window approach as previously described (3). An exon array was designed on the 385,000-feature Nimblegen array (WI) platform. The array tiled each of the 7,523 lincRNA exons detected on the 2.1 million feature array. We also tiled along exons from 1,000 expressed protein-coding genes. All array hybridization images were processed using Nimblescan software and normalized as described (3).

Cells Culture. Human female fetal lung fibroblasts, human male FFs, and HeLa cells were grown in DMEM supplemented with 10% FBS at 37 °C with 5% CO_2.

RNA Coimmunoprecipitation (RIP), Chip, and Analysis. RIP was carried out as previously described in Rinn et al. (9) with some modifications. Briefly, nuclear pellets were isolated, lysed and IPs were performed by incubating each antibody (below) overnight followed by stringent washing of protein A/G bead pellets with final resuspension in TRIzol (Invitrogen). RIP-Chip hybridizations were performed by isolating total coprecipitate RNA as described (9), amplified, labeled, and hybridized as described (3, 9). Antibodies for EZH2 (ab3748), SUZ12 (ab12073), CoREST (ab24166), H3K27me3 (ab6002), and rabbit IgG (ab37415) were purchased from Abcam. Antibody for H3K4me2 (07-030) was purchased from Millipore.

RIP-Chip hybridizations were analyzed as follows. We first normalized the data by dividing each probe value by the average normalized intensity across the array and log-transformed the ratios. We then identified significant regions that are enriched in the RIP experiments relative to IgG controls, using our previously described peak-calling algorithm (3). Briefly, we scanned the genome using sliding windows of consecutive probes of width w. We computed a score defined as the sum of the normalized probe intensities for each window. To determine the significance of this score, we permuted the normalized intensity values assigned to each probe and recalculated the statistic. We took the value for each permutation as the maximum score obtained for any random region. We performed 1,000 permutations, and assigned a P value to each region, corrected for multiple testing, based on its rank within this distribution. All regions with a familywise error rate (FWER) < 0.05 were retained.

To declare a region as enriched, we further required that the signal be at least 2-fold higher than the control (RIP/IgG). After identifying enriched regions, we aggregated them based on the K4-K36 domain in which they reside. Overlaps between replicates were computed based on the overlaps of the lincRNAs between conditions.

Nuclear Enrichment of mRNAs and lincRNAs. The nuclear pellet was isolated from HeLa cells (as described above), and the total nuclear RNA was extracted and hybridized to our custom tiling microarray. We filtered mRNAs and lincRNAs based on their expression in whole-cell extracted RNA to focus on mRNAs and lincRNAs normally expressed in HeLa cells. For each expressed mRNA and lincRNA, we computed a normalized absolute expression (as described above) level of nuclear abundance by computing the median of all probes tiling the lincRNA or mRNA transcript. We computed the distributions of nuclear abundance for mRNAs, expressed lincRNAs, and PRC2-bound lincRNAs. We computed the percentage of mRNAs with nuclear abundance levels greater than the median for the lincRNA distributions. To determine the significance of the distributions between PRC2-bound lincRNAs and non-PRC2-bound lincRNAs, we computed a nonparametric Kolmogorov–Smirnov (KS) test on the 2 observed distributions.

RNA FISH. We performed single molecule RNA FISH against the lincRNAs HOTAIR, Xist, MLKN1A, TUG1, SFPQ1, and FoxF1 using an established method (10). Briefly, we designed 48 fluorescently labeled oligonucleotides targeted to different regions of each lincRNA (table 8 in Dataset S1) and hybridized them to fixed FF and LF. We acquired the 3D image sections using a standard wide-field fluorescent microscope.

Western Blot Analysis. RIP was performed in HeLa cells with SUZ12, CoREST, or IgG as described above. However, instead of resuspending A/G beads with TRIzol, they were resuspended in 1× XT sample buffer (no. 161-0791; BioRad), incubated at 100 °C for 5 min, briefly spun down before loading equal amounts on a 4–20% Criterion precast Bis-Tris protein gel (no. 345-0123; BioRad) and running at 100 V for 2 h. Proteins were transferred to a nitrocellulose membrane for 2 h at 55 V. The membrane was incubated with primary antibody for either SUZ12 or CoREST for 1 h at RT with gentle shaking. Membrane was washed three times in PBS-T (PBS plus 0.1% Tween 20) before incubation with secondary antibody conjugated to HRP for 30 min. After addition of HRP substrate, the chemiluminescence signal was detected with X-ray film.

RNA Interference of PRC2-Associated lincRNAs. Pools of 4 siRNAs targeting each lincRNA were designed using the Dharmacon siRNA design algorithms. Each pool was transfected into a given cell type by electroporation using the Lonza Amaxa nucleofection technology according to Amaxa protocols using the appropriate Amaxa nucleofection kits. Specifically, following the instructions provided with the Human Dermal Fibroblast (NHDF) 96-well nucleofector (Lonza VHPD-1001), we used 2 × 10^5 cells per well and added siRNAs to a final concentration of 1 μM. Cells were nucleofected using program DT-130. Total RNA was isolated from each nucleofection reaction as described (9).

Differential Expression Analysis of lincRNA Depletion Experiments. Total RNA was isolated from each lincRNA depletion experiment as described (9) and hybridized to Affymetrix U133 Plus 2.0 gene expression arrays and processed as described (3). Significant differential gene expression was determined by comparing each expression of the gene after siRNA perturbation to a given lincRNA to its level in pair-matched scramble control and other knock down experiments. For each lincRNA, we considered the siRNA knockdowns of the target gene of interest as one class (sample class) and the controls plus the siRNA knockdowns against the other lincRNAs as another class (control class). We computed a t-statistic for each comparison and permuted class labels to control the false discovery rate (FDR) for each gene. For each lincRNA depletion experiment, we defined gene sets consistent of genes that were up-regulated (sample class versus control class) at an FDR $<$0.1. For each gene set, we identified enrichment of genes repressed by with EED1, EZH2, and SUZ12 as follows. Using published data (11), we ranked genes by their expression changes on knock down of each of these components compared with 3 scrambled RNAi controls. We then used GSEA (12) to compute a weighted KS like test for enrichment of the gene sets up-regulated by depletion of each lincRNA relative to these ranked profiles.

1. Mikkelsen TS, et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448:553–560.
2. Ku M, et al. (2008) Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet* 4:e1000242.
3. Guttman M, et al. (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458:223–227.
4. Barski A, et al. (2007) High-resolution profiling of histone methylations in the human genome. *Cell* 129:823–837.
5. Cui K, et al. (2009) Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell Stem Cell* 4:80–93.
6. Clamp M, et al. (2007) Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci USA* 104:19428–19433.
7. Garber M, et al. (2009) Identifying Novel Constrained Elements by Exploiting Biased Substitution Patterns. *Bioinformatics* 25:i54–i62.
8. Lin MF, Deoras AN, Rasmussen MD, Kellis M (2008) Performance and scalability of discriminative metrics for comparative gene identification in 12 Drosophila genomes. *PLoS Comput Biol* 4:e1000067.
9. Rinn JL, et al. (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* 129:1311–1323.
10. Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S (2008) Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Methods* 5:877–879.
11. Bracken AP, Dietrich N, Pasini D, Hansen KH, Helin K (2006) Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions. *Genes Dev* 20:1123–1136.
12. Subramanian A, et al. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 102:15545–15550.

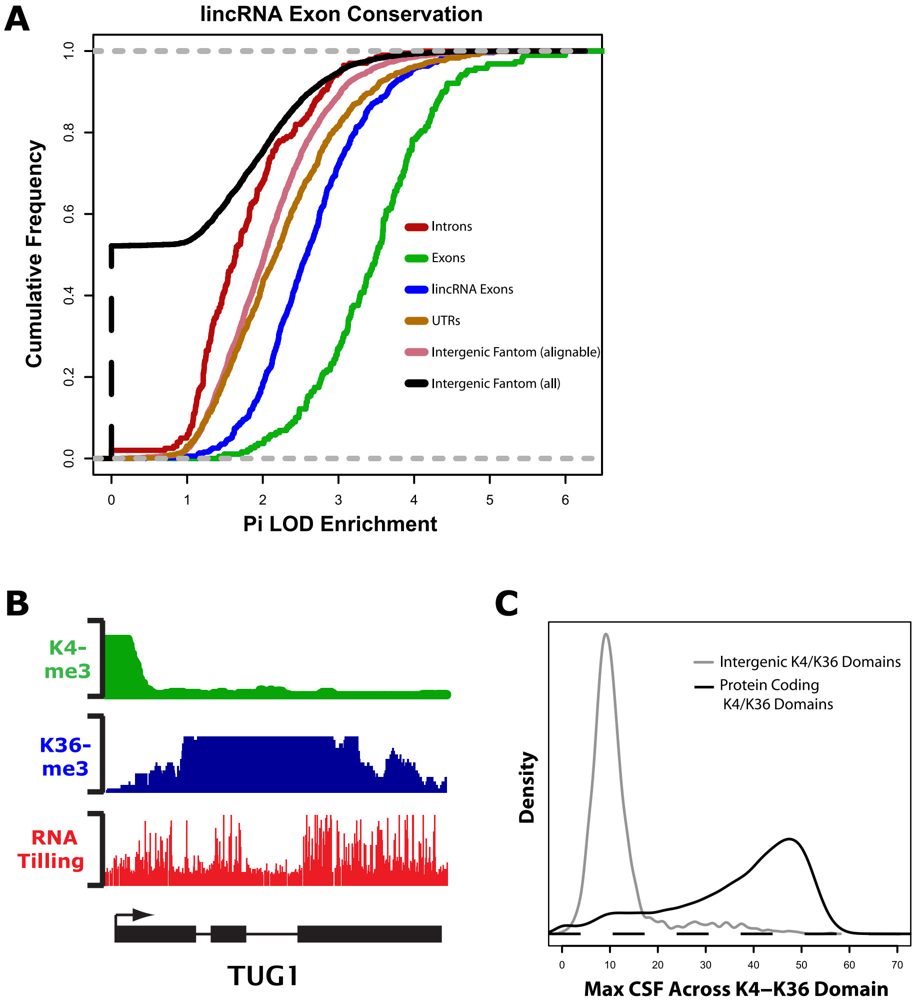

Fig. S1. Intergenic K4-K36 domains in the human genome produce multiexonic noncoding RNAs. (*A*) The lincRNA exon conservation compared with FANTOM and UTRs [figure adapted from Guttman et al. (1)]. Sequence conservation across 21 mammalian species is plotted cumulatively across each exon in the lincRNA transcript (blue), protein coding exons (green), and introns of protein coding genes (red), as well as alignable FANTOM exons (pink), all FANTOM exons (black), and UTRs (orange). The *x* axis is the enrichment of the log odds score of the Pi estimator (see *Methods*) normalized by random genomic regions; thus, larger LOD scores are more highly conserved. (*B*) Representative example of an intergenic K4-K36 domains for the lincRNA TUG1. For each histone modification (K4me3, green; K36me3, blue), the results of ChIP-sequence (seq) experiments are plotted as the number of DNA fragments obtained by ChIP-seq at each position divided by the average number across the genome. Intergenic K4-K36 domains were interrogated for presence of transcription by hybridizing RNA to DNA tiling arrays. The resulting RNA hybridization intensity (red) within each K4-K36 domain is plotted with respect to its genomic location. The start and stop of each exon, as determined by our RNA peak calling algorithm (see *Methods*), is indicated by a black bar. Arrowheads indicate the orientation of transcription. (*C*) CSF scores indicate that the vast majority of intergenic K4-K36 domains are noncoding. CSF scores were calculated across all intergenic K4-K36 domains (gray) and known protein-coding genes (black). The maximum CSF score for each domain is plotted along the *x* axis, and the percentage of K4-K36 domains with this score are plotted on the *y* axis. High and low CSF scores indicate protein-coding capacity and noncoding regions respectively.

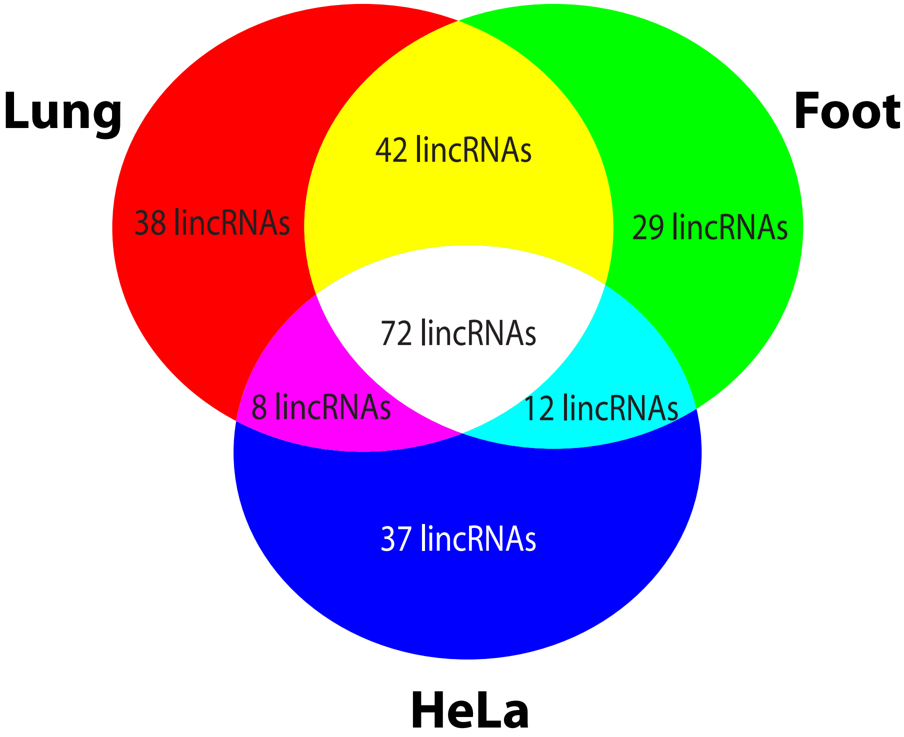

Fig. S2. Venn diagram demonstrating the number of lincRNAs bound to PRC2 in hLF (red), hFF (green) and HeLa (blue) cells.

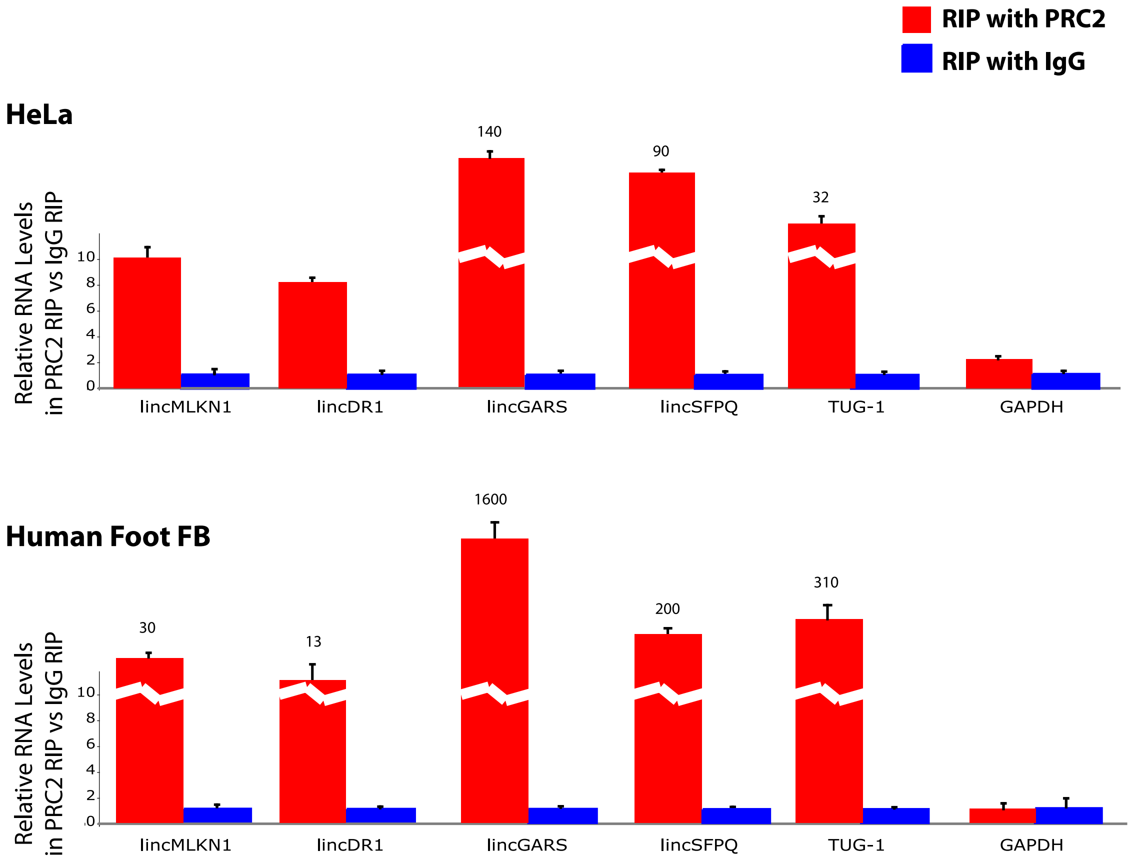

Fig. S3. Validation of PRC2-assocaited lincRNAs by qRT-PCR. Independent RIP experiments were performed in HeLa and hFF, and the coprecipitated RNA was subjected to qRT-PCR for several lincRNAs identified to be bound by DNA tiling array hybridizations. The fold enrichment of each lincRNA in the SUZ12 RIP (red bars) is relative to its matching IgG control RIP (blue bars). Each targeted lincRNA and GAPDH is represented on the *x* axis for HeLa RIP (*Top*) and hFF RIP (*Bottom*). Primers for lincRNAs used in qRT-PCR are provided in table 9 in Dataset S1.

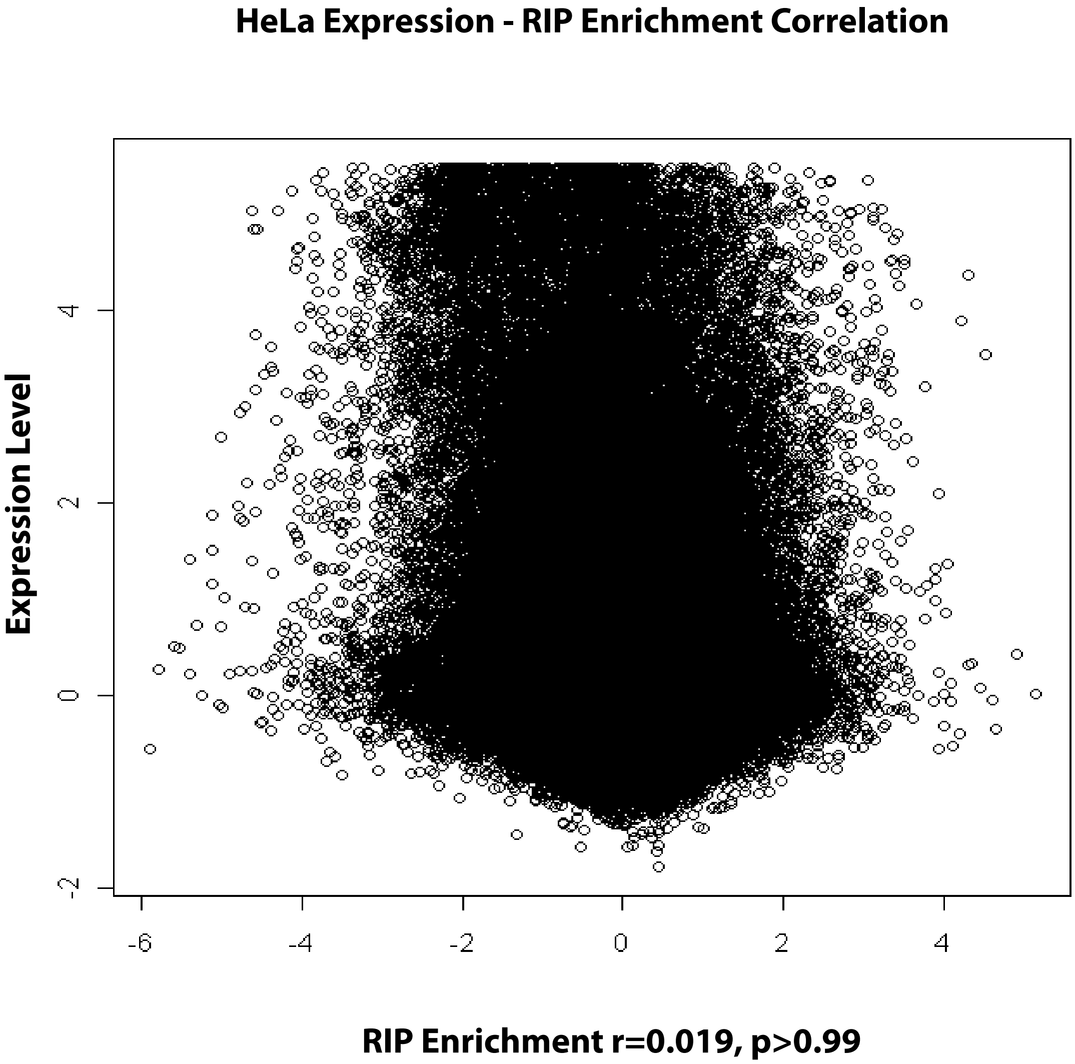

Fig. S4. RIP enrichment does not correlate with lincRNA transcript abundance. Each of the 385,000 probes on the array is represented as a circle. The *y* axis plots the relative expression level of each probe to the normalized array mean. The *x* axis represents the normalized fold enrichment of RIP relative to IgG. The correlation and nominal *P* value are indicated (−0.19 and $P > 0.99$, respectively).

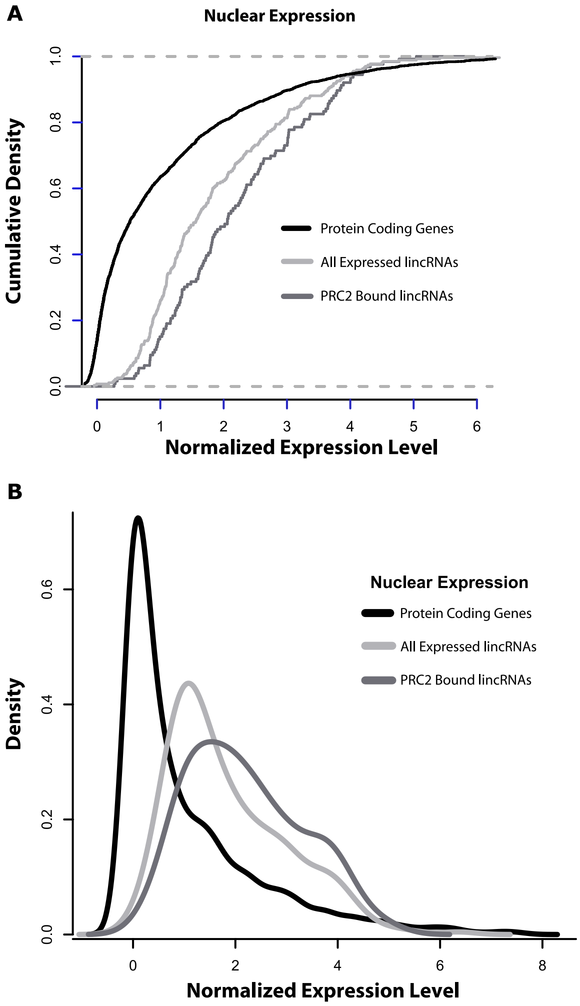

Fig. S5. Nuclear enrichment of protein coding genes and lincRNAs. (*A*) The cumulative density of normalized absolute RNA expression levels (see *Methods*) are shown for protein coding genes (black), all expressed lincRNAs (light gray), and expressed lincRNAs bound to PRC2 in HeLa (dark gray). (*B*) For each lincRNA and mRNA, we computed the expression in the nuclear fraction of HeLa cells. The distributions of normalized absolute RNA expression levels (see *Methods*) are shown for protein coding genes (black), all expressed lincRNAs (light gray), and expressed lincRNAs bound to PRC2 in HeLa (dark gray). Although protein-coding genes are less abundant in the nucleus than lincRNAs, ≈25% of the mRNAs are expressed at levels above the median for all lincRNAs and ≈19% for PRC2-associated lincRNAs.

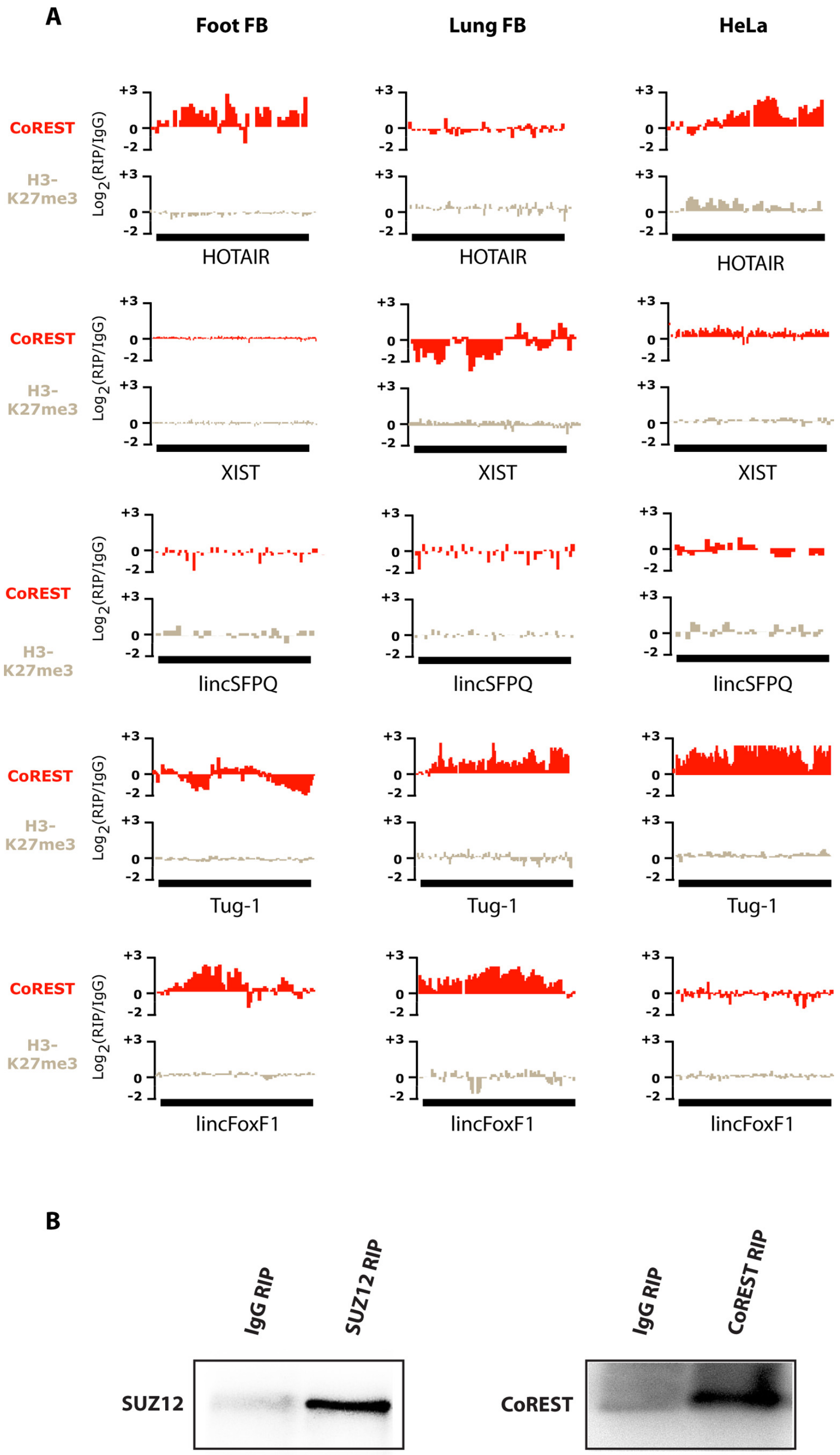

Fig. S6. Numerous lincRNAs are physically associated with CoREST. (*A*) Several examples of lincRNA exons (black box) that are enriched in RIP experiments relative to the IgG control in hFF (*Left*), hLF (*Center*), and HeLa (*Right*) cells. The lincRNAs were enriched in RIP experiments performed with antibody recognizing CoREST (red), but not with antibody recognizing the chromatin protein H3K27me3 (gray). Coprecipitated RNA for CoREST and for the respective control (IgG) was hybridized to the DNA tiling arrays. The hybridization values for each probe within a lincRNA exon are plotted as the log2 values for RIP hybirdization intensity divided by control (IgG) hybridization intensity. Note: TUG1 is coprecipitated with PRC2 in all 3 cell types (see Fig. 2), but only coprecipitated with CoREST in hLF and HeLa, but not in hFF. This is probably due to the fact that RIP reproducibility is 70–80%. (*B*) RIP was performed in HeLa cells with either SUZ12, CoREST, or IgG followed with Western blot analysis to determine specificity of antibodies. SUZ12 is detected in the SUZ12 RIP, but not in IgG RIP. Similarly, CoREST is detected in the CoREST RIP, but not in the IgG RIP.

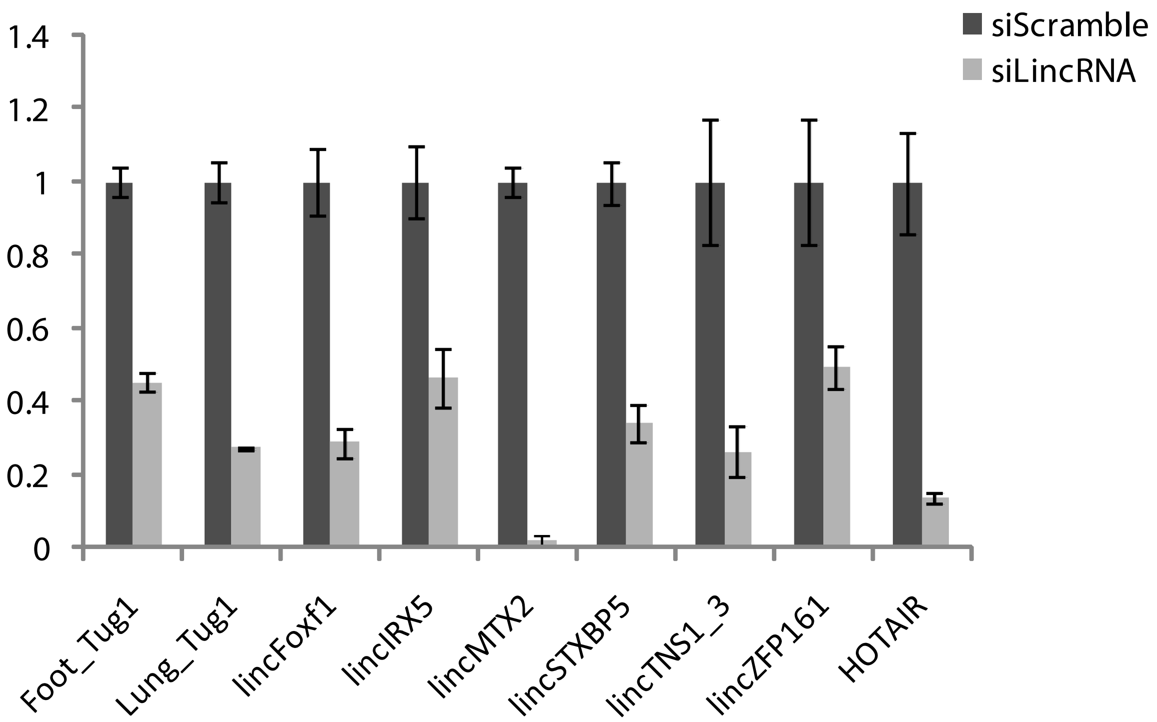

Fig. S7. Validation of siRNAs knockdowns of several lincRNAs by qRT-PCR. FF and LF were transfected with 20 nM of siRNAs for 1 of several lincRNAs of interest for 48 h before RNA isolation. RNA levels of each lincRNA were measured in mock transfected cells (control comprised of nonspecifc siRNA) in comparison with siRNA transfected cells. The *y* axis represents relative fold abundance of each lincRNA in the scramble control (black) and the siRNA pool targeting a given lincRNA (gray).

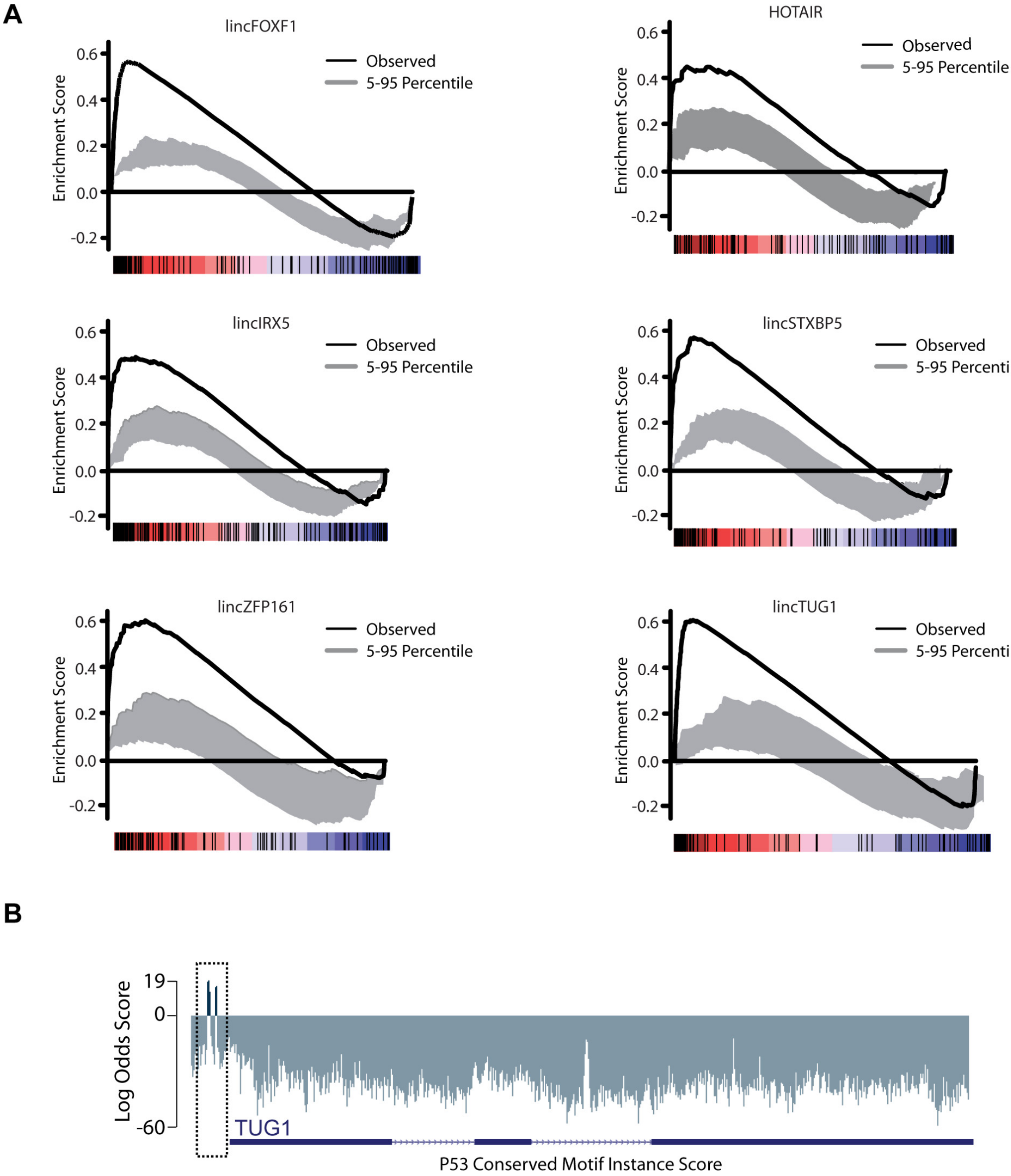

Fig. S8. Genes repressed by PRC2 associated lincRNA overlap with genes repressed by PRC2. (*A*) Gene set enrichment analysis (GSEA) comparing the protein-coding genes that are up-regulated upon depletion of a PRC2 bound lincRNA and those up-regulated upon depletion of various components of PRC2. The black line represents the observed enrichment score profile of protein-coding genes in the lincRNA gene set to the PRC2 gene set. To represent the significance of the black line, we permuted the enrichment score profiles for 100 random (size matched) gene sets. The dark gray region indicates the 5^{t}h to the 95th percentile confidence region; thus, results above the dark gray region are significant at $P < 0.05$. The enrichment profiles for all lincRNAs tested were significant at $P < 0.05$, whereas as the enrichment profile for an unrelated protein depletion (YY-1) was not significant (see Fig. 4). The rank of each gene in the lincRNA gene set is indicated by tick marks (below each enrichment score plot) on a schematic color bar indicating levels of differential expression, up-regulation in red and down regulation in blue. (*B*) The lincRNA TUG1 promoter exhibits highly conserved p53 binding motifs (boxed region), whereas the transcriptional unit does not exhibit enrichment. The log odds conservation score (Fig. 1 and *Methods*) is shown for the p53 binding motif at each position along the lincRNA TUG1 promoter.

Other Supporting Information Files

Dataset S1 (XLS)