Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression

Ahmad M. Khalil[a,b,1], Mitchell Guttman[a,c,1], Maite Huarte[a,b], Manuel Garber[a], Arjun Raj[d], Dianali Rivea Morales[a,b], Kelly Thomas[a,b], Aviva Presser[a], Bradley E. Bernstein[a,e], Alexander van Oudenaarden[d], Aviv Regev[a,c], Eric S. Lander[a,c,f,1,2], and John L. Rinn[a,b,1,2]

[a]The Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA 02142; [b]Department of Pathology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA 02215; Departments of [c]Biology and [d]Physics, Massachusetts Institute of Technology, Cambridge, MA 02139; [e]Molecular Pathology Unit and Center for Cancer Research, Massachusetts General Hospital, Charlestown, MA 02129; and [f]Department of Systems Biology, Harvard Medical School, Boston, MA 02114

Contributed by Eric S. Lander, May 3, 2009 (sent for review March 15, 2009)

We recently showed that the mammalian genome encodes >1,000 large intergenic noncoding (linc)RNAs that are clearly conserved across mammals and, thus, functional. Gene expression patterns have implicated these lincRNAs in diverse biological processes, including cell-cycle regulation, immune surveillance, and embryonic stem cell pluripotency. However, the mechanism by which these lincRNAs function is unknown. Here, we expand the catalog of human lincRNAs to ≈3,300 by analyzing chromatin-state maps of various human cell types. Inspired by the observation that the well-characterized lincRNA HOTAIR binds the polycomb repressive complex (PRC)2, we tested whether many lincRNAs are physically associated with PRC2. Remarkably, we observe that ≈20% of lincRNAs expressed in various cell types are bound by PRC2, and that additional lincRNAs are bound by other chromatin-modifying complexes. Also, we show that siRNA-mediated depletion of certain lincRNAs associated with PRC2 leads to changes in gene expression, and that the up-regulated genes are enriched for those normally silenced by PRC2. We propose a model in which some lincRNAs guide chromatin-modifying complexes to specific genomic loci to regulate gene expression.

histone modifications | epigenetic regulation | polycomb

Mammalian genomes produce a wide variety of noncoding RNA transcripts (1–3). In addition to classical RNAs (e.g., ribosomal RNAs, transfer RNAs, and others) and more recently discovered classes of small noncoding RNAs (e.g., microRNAs) (3, 4), there are many large noncoding RNAs of unknown function (3). Several, such large noncoding RNAs have been biologically characterized (including XIST, TSIX, HOTAIR, and AIR) (3), but shotgun cDNA sequencing and microarray hybridization have suggested that the vast majority of the mammalian genome can produce RNA transcripts under some circumstances (2). However, the biological significance of these transcripts has been highly controversial because most occur at extremely low levels and show little evolutionary conservation (5). A previous study reported that a subset of these ncRNAs shows evidence of evolutionary conservation (27), suggesting that at least some of these ncRNAs are functional.

Recently, we developed an approach for identifying large noncoding RNAs based on a distinctive chromatin signature that marks actively transcribed genes (1). The signature consists of a short region with histone H3 lysine 4 trimethylation (H3K4me3) (corresponding to the promoter) and a longer region with histone H3 lysine 36 trimethylation (H3K36me3, corresponding the transcribed region) (1, 6, 7). We refer to this chromatin signature as a K4-K36 domain. We generated chromatin-state maps across 4 mouse cell types, searched for K4-K36 domains, and then eliminated those corresponding to known protein-coding genes. We found 1,586 previously uncharacterized K4-K36 domains in the 4 mouse cell types, and showed that the vast majority encode large intergenic noncoding (linc)RNAs. These lincRNAs show similar expression levels as protein-coding genes, but lack any protein-coding capacity. Importantly, lincRNAs show significant evolutionary conservation relative to neutral sequences, providing strong evidence that they have been functional in the mammalian lineage (1). We note that nonconserved RNA sequences identified in other collections could be functional, but biological evidence such as loss-of-function experiments would be needed to establish their functionality (5) (Fig. S1*A*). Previous studies by us and others have demonstrated that groups of lincRNAs exhibit expression patterns across cell types and tissues that correlate with patterns seen for protein-coding genes involved in cellular processes such as cell-cycle regulation, innate immunity responses, and stem cell pluripotency (1, 14). Although these studies clearly demonstrate that there are many functional lincRNAs, key questions remain, including: How many lincRNAs are encoded in mammalian genomes? How do lincRNAs exert their functions? To begin to investigate the number of lincRNAs, we extended our approach of mapping K4-K36 domains to 6 human cell types. The results expand our catalog to 3,289 lincRNAs, which show clear evolutionary conservation within their transcripts. Extrapolation suggests that the total number may approach ≈4,500 lincRNAs.

To examine the biochemical mechanism by which lincRNAs function, we drew inspiration from one of the few well-studied lincRNAs: HOTAIR. We previously reported HOTAIR as a lincRNA transcribed from within the HOXC cluster, and showed that it acts to repress genes in the HOXD cluster, by binding to the polycomb repressive complex (PRC)2 and recruiting it to the locus (8). PRC2 is a methyltransferase that trimethylates H3K27 to repress transcription of specific genes (9, 10). Recently, several other large noncoding RNAs have been found to associate with chromatin-modifying complexes, including a large noncoding RNA encoded within the 5′ of XIST that can target PRC2 to the inactive X chromosome (11), the antisense transcript AIR that is associated with the chromatin-modifying complex G9a, an H3K9me2 meth-

Author contributions: A.M.K., M. Guttman, E.S.L., and J.L.R. designed research; A.M.K., M. Guttman, M.H., A. Raj, D.R.M., and K.T. performed research; A.M.K., M. Guttman, A.P., B.E.B., A.v.O., A. Regev, E.S.L., and J.L.R. contributed new reagents/analytic tools; A.M.K., M. Guttman, M. Garber, E.S.L., and J.L.R. analyzed data; and A.M.K., M. Guttman, A. Regev, E.S.L., and J.L.R. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The sequence reported in this paper has been deposited in the GEO database (accession no. GSE16226).

[1]A.M.K., M. Guttman, E.S.L., and J.L.R. contributed equally to this work.

[2]To whom correspondence may be addressed. E-mail: lander@broad.mit.edu or jrinn@broad.mit.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0904715106/DCSupplemental.

Downloaded at California Institute of Technology on August 6, 2021

GENETICS

yltransferase (12), and the Kcnq1ot1 transcript that binds both G9a and PRC2 (13). Some recent studies have demonstrated a few large noncoding RNAs bind chromatin proteins that add activating modifications (e.g., Trithorax) in mES cells (14). These few examples raised the possibility that many lincRNAs might be physically associated with chromatin-modifying complexes and might potentially target them to specific genomic regions. To test this hypothesis on a genomic scale, we performed RNA coimmunoprecipitation (RIP) with antibodies directed against several proteins involved in chromatin-modifying complexes (PRC2, CoREST, and SMCX). We find that as many as 38% of the lincRNAs expressed in the cell types studied are reproducibly associated with one of these complexes. Also, we show that RNA-interference-based depletion of various PRC2-associated lincRNAs results in activation of genes known to be repressed by PRC2. Together, our results indicate that thousands of functional lincRNAs are encoded in the human genome, and a significant proportion of lincRNAs are physically associated with chromatin-modifying complexes. We propose a shared mechanism for hundreds of lincRNAs that function in regulating the epigenetic landscape at distinctive target loci.

Results

Identification of Human lincRNAs. We recently identified lincRNAs in the mouse genome by analyzing chromatin-state maps of 4 mouse cell types: mES, lung fibroblasts (mLFs), neural precursor cells (NPCs), and mouse embryonic fibroblasts (MEFs) (1). Specifically, we used a computational algorithm to identify K4-K36 domains that do not overlap known protein-coding genes, and then eliminated the small proportion (≈5%) that showed any significant protein-coding capacity. We were left with 1,586 previously uncharacterized K4-K36 domains. We demonstrated that the vast majority of these regions encode lincRNAs, with 95% showing significant conservation to the human genome within their transcripts. To further extend the catalog of lincRNAs, we sought to analyze chromatin-state maps of 6 human cell types: hESC (10), 2 hematopoietic stem cells (CD133+ and CD36+) (15), T-cells (6), hLFs (1), and normal embryonic kidney (hEK). Using our previous computational approach, we identified K4-K36 domains that are well-separated from (*i*) the regions containing known protein-coding genes and all known classes of small noncoding RNAs in human, and (*ii*) the orthologous regions of known protein-coding genes in mouse, rat, and dog. We also eliminated the orthologous regions of our previously identified mouse lincRNAs. We previously showed that, for similar cell types in mouse and human, lincRNA loci show cross-species conservation not only at the level of nucleotide sequence, but also with respect to the presence of K4-K36 domains (1). We found a total of 1,833 previously uncharacterized intergenic K4-K36 domains in these 6 human cell types (Fig. 1*A*; Fig. S1*B* and table 1 in Dataset S1). We analyzed the coding potential of each such K4-K36 domain using the codon substitution frequency score (*SI Methods*), and found that <8% showed any evidence of protein-coding capacity (Fig. S1*C*) (16). After eliminating these cases, we were left with 1,703 loci encoding putative lincRNA genes.

To test whether these loci actually encode lincRNAs, we designed genomic tiling microarrays (at 10-base resolution) across 1,147 of the 1,703 loci (*SI Methods*) to determine their exonic structure. We hybridized poly(A^+)-amplified RNA from hES, brain, breast, hEK, hFF, hLF, K562, ovary, skin, spleen, testis, and thymus tissues. We analyzed the hybridization data using our previously reported peak-calling algorithm. This analysis revealed multiexonic RNA transcripts in 74% of the K4-K36 domains examined (Fig. 1*A*). There was an average of 4 exons per K4-K36 domain (total of 4,860 exons). We further focused on the 535 K4-K36 domains that were discovered in cell types in which RNA from the same cell type was hybridized. In these 3 cell types, RNA hybridization revealed multiexonic RNA transcripts in 85% of the tested loci; this detection rate is similar to that previously seen for K4-K36 domains

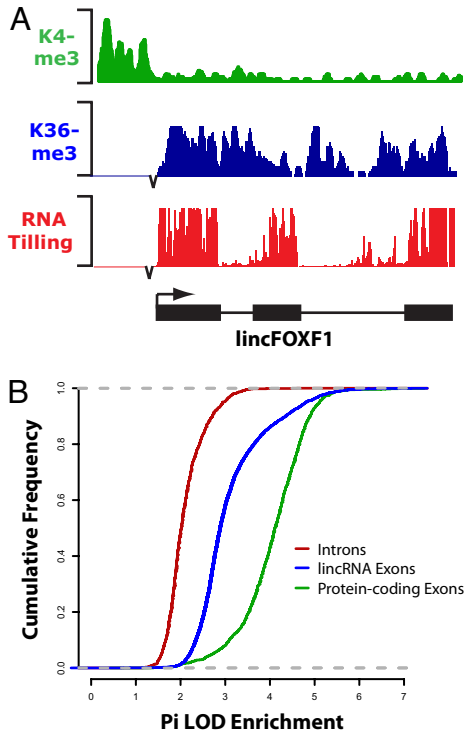

Fig. 1. Intergenic K4-K36 domains in the human genome produce multiexonic noncoding RNAs. (*A*) Representative example of an intergenic K4-K36 domain for lincFOXF1. For each histone modification (K4me3, green; K36me3, blue), the results of ChIP-sequence experiments are plotted as the number of DNA fragments obtained by ChIP-sequence 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 (*SI Methods*), is indicated by a black bar. Arrowheads indicate the orientation of transcription. (*B*) Sequence conservation scores (*SI Methods*) across 21 mammalian species indicates lincRNAs (blue) are much more conserved than neutrally evolving intronic sequences (red), although less so than protein-coding genes (green). For each lincRNA exon, protein-coding gene exon, and protein-coding gene intron, a conservation score was calculated and plotted along the *x* axis as a log-odd enrichment score (compared with random genomic regions of equivalent size). The cumulative number of exons with a given score or lower is represented on the *y* axis.

corresponding to known protein-coding genes and lincRNA loci in mouse (1). Given that such a high proportion of the human K4-K36 domains tested were validated as encoding lincRNAs, we conclude that the vast majority of the full set of 1,703 loci encode bona fide lincRNAs. We then studied the evolutionary conservation of the lincRNA loci. For each exon, we calculated the extent of sequence conservation across 21 mammalian species as previously described (*SI Methods*) (1). Human lincRNAs showed evolutionary conservation at levels similar to those seen for the lincRNAs in our previous study (Fig. 1*B*; table 2 in Dataset S1) (1).

Combining the 1,586 human orthologs of the lincRNA genes reported in our previous study with the 1,703 recently discovered human lincRNA genes identified in this study, our catalog of human lincRNA genes now includes 3,289 distinct loci. This catalog is certain to be incomplete, because it is based on chromatin-state maps of only 10 cell types (4 mouse and 6 human). Nonetheless, it is possible to make a rudimentary estimate of the total number of human lincRNAs based on the observation that 73% of all protein-coding genes are expressed in at least 1 of the 10 cell types analyzed here. If a similar proportion applies to lincRNAs, the total number of human lincRNAs would be estimated to be ≈4,500. If lincRNAs

Downloaded at California Institute of Technology on August 6, 2021

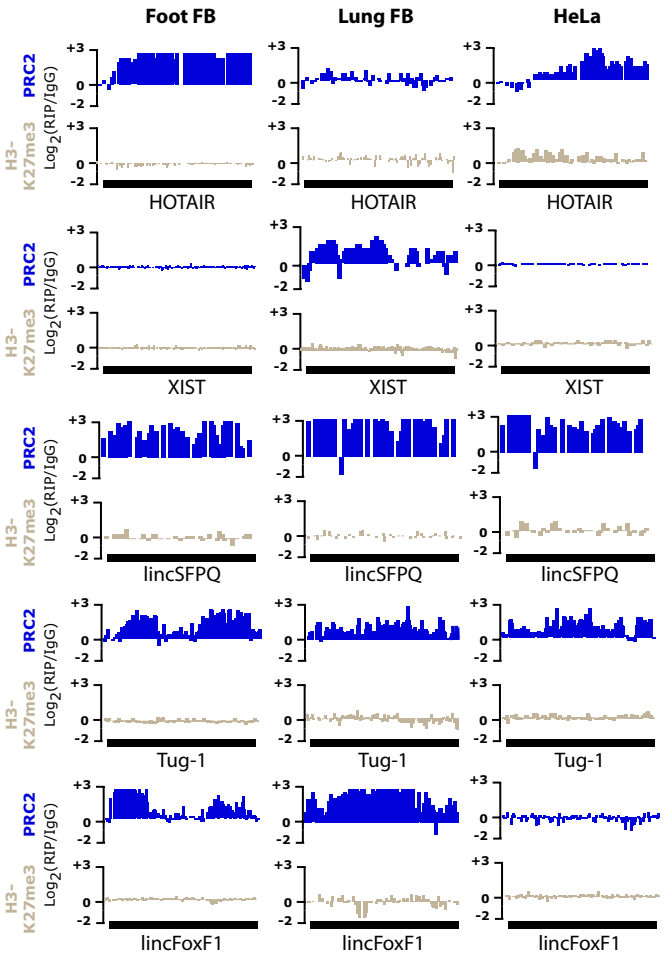

Fig. 2. Numerous lincRNAs are physically associated with PRC2. Several examples of lincRNA exons (black box) that are enriched in RIP experiments relative to the IgG control in hFF (*Left*), hLF (*Central*), and HeLa (*Right*) cells; lincRNAs were enriched in RIP experiments performed with antibodies recognizing the chromatin-modifying complexes: PRC2 (blue), but not with antibodies recognizing the chromatin protein H3K27me3 (gray). Coprecipitated RNA for each antibody 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.

have expression patterns that are more tissue- or condition-specific, the total number could be considerably higher. Obtaining a complete catalog will require generating chromatin-state maps across many more tissues.

Many lincRNAs Are Associated with PRC2. We next explored the mechanism by which lincRNAs function. As noted above, the lincRNA HOTAIR has been shown to physically associate with PRC2 (8). This physical association was shown by RIP-PCR assay: total (non-cross-linked) nuclear extract was incubated with an antibody against the SUZ12 protein, a component of PRC2; the extract was precipitated with Protein-A-coupled beads; and the coprecipitated RNA was then subjected to locus-specific RT-PCR to demonstrate the presence of HOTAIR. To test whether other lincRNAs are also associated with PRC2, we designed a "RIP-Chip" assay (*SI Methods*) to assay many lincRNAs simultaneously (Fig. 2). Briefly, we used antibodies against the proteins SUZ12 and EZH2, components of PRC2 (9, 10). The antibodies were incubated with non-cross-linked nuclear extracts from 3 human cell types: HeLa cells, h lung (L)F, and h foot (F)F; these cell types were chosen because they have previously been shown to have distinctive epigenetic landscapes and diverse gene expression patterns (8). We analyzed the coprecipitated RNAs by hybridization to a custom "exon-tiling" array (at 10-base resolution), containing exons from ≈900 human lincRNA loci and ≈1,000 human protein-coding genes; the protein-coding genes were previously known to be expressed in at least 1 of the 3 cell types. In parallel, we carried out a mock control with a nonimmune rabbit IgG polyclonal antibody to assess nonspecific interactions that may occur in RIP.

To identify lincRNAs and protein-coding genes that are coprecipitated with each of the PRC2 components, we analyzed the hybridization data with a peak-calling algorithm that finds regions in which the signal from the RIP assay is significantly enriched over the signal from the mock controls (*SI Methods*). Regions were defined based on a maximum familywise error rate (FWER) < 0.05 (*SI Methods* and table 3 in Dataset S1) (1). Given that RIP assays are known to show considerable variability (with typical reproducibility of ≈60%) (8), we performed several biological replicates for each cell type. We observed that ≈76% of the genes detected in one replicate are also detected in a second replicate (hLF, 70%; hFF, 75%; HeLa, 83%; see table 3 in Dataset S1). As a positive control, we checked whether HOTAIR and XIST were detectably coprecipitated in our RIP-Chip data. Consistent with previous reports, HOTAIR coprecipitated with PRC2 in both HeLa and hFFs, but not in hLFs. Similarly, XIST, which is expressed only in female cells, was detectably coprecipitated in the hLF cells (which came from a female source), but not the hFF cells (which came from a male source) (Fig. 2). These results were consistent across all replicates.

In addition to the RIP assay, we also assayed expression patterns of lincRNAs and protein-coding genes on the custom exon-tiling array. We extracted total RNA from the same 3 human cell types (HeLa, hLF, and hFF), prepared poly(A^+)-amplified cDNA, and hybridized the product to the exon-tiling array. Of the lincRNA genes on the array, we found that 47% were detectably expressed in at least 1 of the 3 cell types (HeLa, 25%; hLF, 37%; and hFF, 33%; see table 4 in Dataset S1). Consistent with the design of the tiling array, essentially all of the protein-coding genes were detectably expressed in the relevant cell type. Analysis of the RIP-Chip results, in conjunction with the expression analysis, suggests that a significant proportion of all lincRNAs expressed in 1 of these 3 cell types are physically associated with PRC2. Specifically, we find that ≈30% of expressed lincRNAs are detected in at least 1 of the replicates. As a conservative estimate, we only considered lincRNAs detected in at least 2 replicates. Using this criterion, we observe that 24% of lincRNAs (114 of 469) expressed in 1 of the 3 cell types is detected as physically associated with PRC2 (Fig. 2; Fig. S2).

As an independent validation of the association with PRC2, we selected 5 lincRNAs that were detected in our RIP-Chip data as associated with PRC2 in both HeLa and hFF, and performed RIP-quantitative (q)PCR assays for these transcripts, using qRT-PCR. In all 10 tests (5 lincRNAs in 2 cell types), the results were confirmed (Fig. S3 and table 5 in Dataset S1). Notably, the RIP-qPCR assays showed a higher degree of enrichment than the RIP-Chip assays, consistent with the fact that arrays have a narrower dynamic range. As a validation that the associations of lincRNAs with PRC2 are specific, we tested whether the enrichment in the RIP-Chip experiment was simply a reflection of transcript abundance, which would suggest nonspecific interactions. We found no significant correlation between transcripts levels of the lincRNAs and their level of PRC2-enrichment ($r = -0.109$, $P > 0.99$; Fig. S4).

As a second approach to assess the specificity of PRC2 binding to lincRNAs, we examined the proportion of mRNAs bound to PRC2. In sharp contrast to the lincRNAs, very few of the mRNAs assayed in the RIP-Chip experiments showed physical association with PRC2. Of the 1,000 mRNAs represented on the array, only 16 (<2%) were detected in 2 replicates (Fig. 3*A*). We suspect that

Downloaded at California Institute of Technology on August 6, 2021

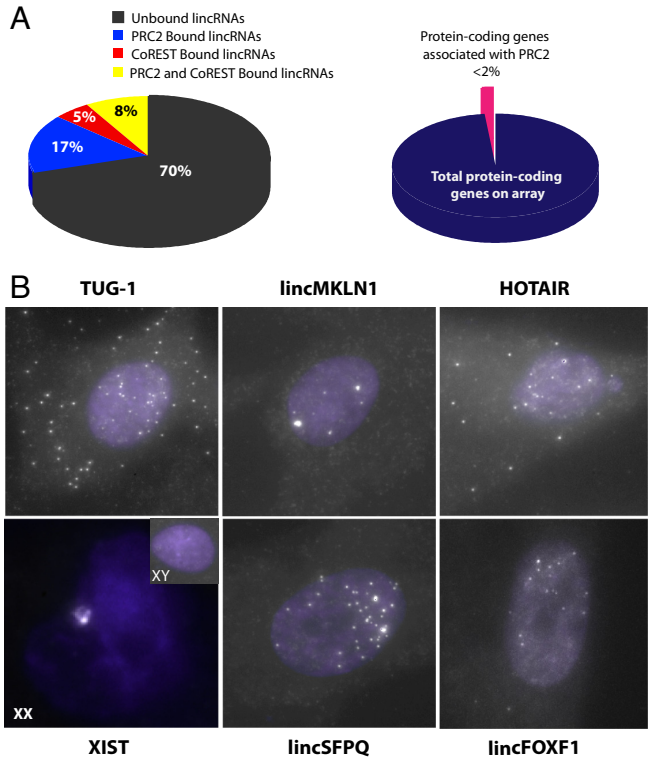

Fig. 3. Diversity and nuclear localization of chromatin associated lincRNAs. (*A*) Many lincRNAs, but not protein-coding mRNAs are physically associated with chromatin-modifying complexes. (*Left*) Pie chart representing the percentage of unique lincRNAs expressed in all 3 tested cell types (hFF, hLF, and HeLa) that are bound only to PRC2 (blue), only to CoREST (red), bound by both PRC2 and CoREST (yellow), and those not bound by either complex (black). (*Right*) Pie chart indicating the percentage of protein-coding genes (pink) reproducibly bound to PRC2 and or CoREST in all 3 cell types relative to the total number of expressed protein-coding genes (dark blue). (*B*) Subcellular localization analysis of lincRNAs by RNA FISH demonstrates localization of lincRNAs to the nucleus. Each panel represents the in situ hybridization of ≈40 fluorescently labeled DNA oligos with complementarity to the interrogated lincRNA. RNA FISH experiments were performed in male hFF for each represented lincRNA (XIST, HOTAIR, TUG-1, lincMKLN-1, lincFOXF1, and lincSFPQ), and also in female hLF for XIST (XX). White "speckles" indicate the subcellular localization of each lincRNA. The nuclear compartment is demarked by DAPI staining (purple).

many of these 16 cases are artifacts, because only a small proportion (<1% of expressed mRNAs) are detected in 3 replicates. Thus, the proportion of transcripts associated with PRC2 is much higher for lincRNAs than for mRNAs. To demonstrate that this result is not simply due to a low concentration of mRNAs in the nucleus, we compared the concentration of lincRNAs and mRNAs in the nucleus (*SI Methods*). Although lincRNAs tend to have greater abundance in the nucleus than mRNAs, we find that the distributions of nuclear abundance of lincRNAs and mRNAs show substantial overlap, with at least 25% of mRNAs being expressed at levels comparable with the 50th percentile level for lincRNAs (Fig. S5). We also reasoned that lincRNAs associated with PRC2 should have significant representation in the nucleus. Therefore, we examined the abundance of lincRNAs in the nucleus, and we found that PRC2-bound lincRNAs show a significantly higher abundance in the nucleus than non-PRC2-bound lincRNAs (*SI Methods* and Fig. S5). We also performed RNA-FISH on HOTAIR, XIST, and 4 additional lincRNAs detected as associated with PRC2. In all cases, the lincRNAs showed either exclusively nuclear or nuclear and cytoplasmic localization (Fig. 3*B*). Last, we explored whether a lincRNA that is expressed in 2 cell types (A and B), and associated with PRC2 in one cell type (A), is also associated with PRC2 in the second cell type (B). Considering all pairs of cell types, we found that was the case for ≈85% of lincRNAs (table 3 in Dataset S1). Collectively, these results provide strong evidence that a substantial portion (20–30%) of lincRNAs is specifically bound by PRC2 (Figs. 2 and 3*A*).

Association of lincRNAs with Other Chromatin-Modifying Complexes. Having found that many lincRNAs are associated with PRC2, we then explored whether they might be associated with other repressive chromatin-modifying complexes. We examined CoREST, a repressor of neuronal genes (17). We performed RIP-Chip using an antibody against CoREST in the same 3 cell types (HeLa, hLF, and hFF). Applying the analysis above, we found that 63 of the 469 lincRNAs expressed in HeLa cells were reproducibly detected as bound to CoREST (Fig. 3*A*; Fig. S6 and table 6 in Dataset S1). As with PRC2, <2% of mRNAs coprecipitated with CoREST. We note that ≈60% of the lincRNAs associated with CoREST are not associated with PRC2 in HeLa cells, indicating that each complex has specific lincRNAs associated with it. The observation that 40% of the lincRNAs associated with CoREST are also associated with PRC2 suggests that the 2 complexes share some regulatory targets. Considering PRC2 and CoREST together, we find that ≈38% of lincRNAs expressed in at least 1 of the cell types examined are reproducibly bound to at least 1 of the 2 complexes (180 of 469 expressed). This proportion is likely to be an underestimate, because we only count lincRNAs that were detected in at least 2 replicates; the proportion could be as high as 52%. These results raise the possibility that lincRNAs may be associated with additional chromatin-modifying proteins. Intriguingly, preliminary results involving RIP-Chip with SMCX, a histone H3K4me3 demethylase, suggest that this enzyme also binds a significant number of lincRNAs. We also tested whether chromatin proteins themselves (rather than chromatin-modifying proteins) are associated with lincRNAs. Specifically, we performed RIP-Chip with antibodies against the modified histones H3K27me3 and H3K4me2. We found no significant enrichment of lincRNAs (Fig. 2). These findings are consistent with other studies that identified XIST to coprecipitate with PRC2, but not H3K27me3, despite their immediate nuclear proximity (11).

Functional Evidence That lincRNAs Act Through the PRC2 Pathway. Having found that a substantial fraction of lincRNAs is physically associated with PRC2, we sought evidence that they have a functional role in polycomb-mediated repression. Previously, we have shown that depletion of HOTAIR causes up-regulation of genes normally repressed by PRC2 (8). To test whether other lincRNAs have a similar effect, we studied HOTAIR and 5 additional lincRNAs found to be associated with PRC2. For each of these 6 lincRNAs, we designed pools consisting of 4 siRNAs targeting each lincRNA (*SI Methods*). We also used standard control siRNA pools that do not correspond to any human sequence. We transfected the siRNA pools into hLF (3 pools), hFF (3 pools), or both (1 pool), with each experiment performed in duplicate. We measured the level of lincRNA knock-down by qRT-PCR, and compared the results to the control siRNA pool; we only used experiments in which we achieved >2-fold depletion (Fig. S7).

We hybridized the total RNA from these experiments to standard gene-expression arrays to measure the resulting changes in gene expressions. Specifically, for each of the 6 lincRNAs, we determined the gene sets (S_1, S_2, … , S_6) that were up-regulated relative to the control siRNA pools [at a false discovery rate (FDR) < 0.1]. These sets contained between 103 and 352 genes (Fig. 4*A*; Fig. S8*A*). The sets of genes affected by each lincRNA did not show significant overlap, suggesting that each lincRNA has distinct target sets. We searched for, but found no common motifs enriched among the up-regulated genes for each lincRNA. However, given the small number of target genes and the inability to

Downloaded at California Institute of Technology on August 6, 2021

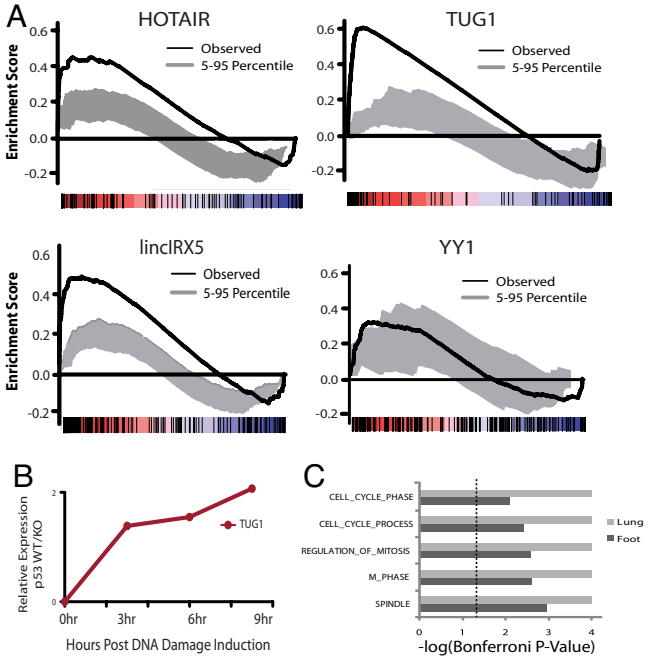

Fig. 4. Genes repressed by PRC2 associated lincRNA overlap with genes repressed by PRC2. (*A*) GSEA comparing the protein-coding genes that are up-regulated on depletion of a PRC2 bound lincRNA and those up-regulated on depletion of various components of PRC2. The black line represents the observed enrichment 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 5th 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. 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*) lincRNA TUG1 is transcriptionally regulated by p53 in response to DNA damage. The *y* axis indicates the log2 ratio of lincRNA TUG1 expression in p53 wild-type cells divided by the expression value in p53 knock-out cells. The *x* axis indicates time after induction of DNA damage. (*C*) Gene ontology (GO) enrichment analysis identified numerous cell-cycle regulation pathways that were specifically derepressed on knock down of lincRNA TUG1. The enrichment FDR is plotted as $-\log(\text{FDR})$ on the *x* axis. Results are shown from knockdown experiments in hLFs (gray) and in hFFs (black). Dashed line denotes FDR < 0.05.

distinguish between direct and indirect targets, this result may simply reflect the low statistical power in analyzing a relatively small set of genes. Also, no lincRNA knock-down significantly affected the expression level of nearby genes (a window of at least 10 genes in either direction), suggesting that these lincRNAs are not likely to function via a *cis*-acting mechanism. Rather, our observation suggests that influence on gene regulation by PRC2 associated lincRNAs is likely exerted by a *trans* mechanism, similarly to what we have previously shown for HOTAIR (8).

We then sought to determine whether the up-regulated gene sets were highly enriched in genes normally repressed by PRC2 in human fibroblasts. Toward this end, we analyzed published data (9) in which the investigators measured gene expression changes in human fibroblasts in response to depletion of 3 key components of PRC2 (EZH2, SUZ12, and EED-1) with shRNAs. For each component, we ranked all genes based on their change in expression level; the ranked lists are similar for each of the 3 components. We then used gene set enrichment analysis (GSEA) (18) to test whether the gene sets up-regulated in response to depletion of the 6 lincRNAs (S_1, S_2, … , S_6) were enriched among the genes up-regulated in response to depletion of the PRC2 components. The resulting enrichments were highly significant (FDR < 0.01) for each of the 6 lincRNAs and each of the 3 PRC2 components (18 analyses in all, Fig. 4*A*; Fig. S8*A*). As a negative control we examined the genes affected by the shRNA-mediated depletion of YY1 (19), a transcription factor associated with chromatin. In contrast to depletion of the lincRNAs, we found no significant enrichment of PRC2 target genes. These results show that depletion of lincRNAs associated with PRC2 causes changes in gene expression, and these genes are strongly enriched for genes normally repressed by PRC2. This finding provides functional evidence that many lincRNAs likely function through their interaction with PRC2.

An Example: TUG1 Represses p53-Dependant Cell-Cycle Regulation. Last, we decided to focus on a specific PRC2-associated lincRNA, TUG1. TUG1 was originally identified as a transcript up-regulated by taurine, and siRNA-based depletion of TUG1 in the developing mouse eye was found to block retinal development (20); the mechanism by which TUG1 depletion produces this phenotype is unknown. In our study, we found that TUG1 is ubiquitously expressed in human and mouse cell types and tissues, and is bound to PRC2 in all 3 of the cell types examined. Previously, we studied regulation of lincRNAs in response to DNA damage, and found that TUG1 was among the 39 lincRNAs specifically induced in p53-wild type, but not p53-mutant cells (Fig. 4*B*) (1). Also, TUG1 promoter contains many highly conserved binding sites for p53 (Fig. S8*B*). We selected TUG1 as 1 of the 6 lincRNAs above that we depleted with siRNA pools. Depletion of TUG1 led to significant up-regulation of 120 genes, which were strongly enriched for those involved in cell-cycle regulation (regulation of mitosis, spindle formation, and cell-cycle phasing) (Fig. 4*C*; table 7 in Dataset S1). Thus, TUG1 is induced by p53, binds to PRC2, and has a role in repressing specific genes involved in cell-cycle regulation. Interestingly, p53 is well known to cause both activation and repression of many genes. Although p53 has been shown to be a direct activator of many genes, the mechanism of p53-induced repression remains unknown. Our results suggest the intriguing hypothesis that TUG1, and perhaps other lincRNAs, may function as downstream repressors in transcriptional pathways.

Discussion

It is becoming clear that the mammalian genome encodes thousands of lincRNAs that are highly conserved and, thus, biologically functional (1, 27). The results of our previous article (1) and this study together identify 3,289 lincRNAs, and suggest that the total may be in the range of ≈4,500. Expression patterns suggest that these lincRNAs are involved in diverse biological processes, including cell-cycle regulation, innate immunity, and ES pluripotency, but the mechanisms by which they have their roles were completely unknown. Inspired by studies of the lincRNAs HOTAIR (8) and XIST (11), we investigated the idea that many lincRNAs are involved in the establishment of chromatin states. In this study, we report that a substantial proportion (24%) of lincRNAs expressed in a cell type are physically associated with the repressive chromatin-modifying complex PRC2, and the proportion is even larger (38%) when additional chromatin-modifying proteins (CoREST and SCMX) are included. Thus, it seems likely that a significant fraction of lincRNAs will be associated with chromatin-modifying proteins. Beyond the physical association, our functional analysis demonstrates that siRNA-mediated depletion of these lincRNAs results in preferential derepression of PRC2 regulated genes at distant loci, consistent with a *trans*-acting mechanism. Collectively, these results suggest that many lincRNAs collaborate with chromatin-modifying proteins to repress gene expression at specific loci.

There is a growing body of literature from yeast to mammals suggesting the noncoding RNAs have an important role in chromatin-state formation (4, 21). In *Schizosaccharomyces pombe*, a

Downloaded at California Institute of Technology on August 6, 2021

process known as RNA induced transcriptional silencing (RITS) has been shown to have an important role in heterochromatin formation over centromeric repeats (22). Similarly, short RNAs have been shown to have an important role in the establishment of heterochromatic silencing in plants. In *Caenorhabditis elegans*, genetic screens have identified polycomb homologs to be required for proper gene silencing in an RNA-dependent manner (4). In mammals, only a few specific RNAs (such as HOTAIR and XIST) have been implicated in directing chromatin modification. However, there is evidence that RNA has a key role in shaping mammalian epigenetic landscapes. For example, depletion of single-stranded (ss)RNA in mouse fibroblasts inhibits global heterochromatin formation (23). Similarly, ssRNA, but not ssDNA, is required for the maintenance of the histone modifications H3K27me3 and H3K9me3 (24).

Our results suggest an intriguing hypothesis that lincRNAs bind to chromatin-modifying complexes to guide them to specific locations in the genome. Whereas chromatin-modifying proteins are often ubiquitously expressed, they establish epigenetic states that differ markedly among cell types and conditions. Under our model, differentially expressed lincRNAs could bind to these complexes and help establish cell type specific epigenetic states. In particular, PRC2 is involved in establishing repressive chromatin states involving H3K27me3. Together, PRC2 and a lincRNA might have the role of a transcriptional repressor by directing silencing to specific loci. Such a mechanism could function within a larger regulatory program. Specifically, a newly induced transcription factor might establish a particular cellular state by (*i*) directly activating some downstream genes, and (*ii*) activating lincRNAs that (with PRC2) repress genes involved in a previous or competing cellular state. Our observations concerning the lincRNA TUG1 suggest that it may function in such a program. On DNA damage, TUG1 is induced in a p53-dependent manner, likely through direct binding of p53, in view of many p53-binding sites in its promoter. It then binds PRC2 (based on our RIP-Chip data) and is involved in repressing important cell-cycle related genes (based on siRNA-based depletion of TUG1). Thus, we speculate that TUG1 may serve as a downstream transcriptional repressor in the p53 pathway to repress cell-cycle progression in response to DNA damage. Similarly, we have recently shown that HOTAIR serves as a transcriptional repressor of HOXD genes. We now know that HOXA13, the key distal regulator, directly transcribes HOTAIR to establish positional identity by repressing the appropriate HOX clusters. Thus, HOTAIR serves as a downstream repressor in the HOXA13 transcriptional network. This model raises many mechanistic questions, including (*i*) whether most lincRNAs associated with chromatin-modifying complexes directly guide the complexes to specific loci, and (*ii*) if so, how the guidance is accomplished (e.g., by direct base pairing at specific sequence motifs).

Our experiments have focused on chromatin-modifying complexes that add repressive chromatin marks. It is possible that many additional lincRNAs are associated with chromatin-modifying complexes that confer activating modifications, as has been recently reported in a few cases (14). These questions can be addressed by performing RIP experiments with a wide range of antibodies across a wide range of cell types, to create a catalog of lincRNA-protein interactions. Last, although we have found that a substantial proportion of lincRNAs are associated with repressive chromatin-modifying complexes, we do not mean to suggest that all lincRNAs necessarily function in this manner. There may be classes of lincRNAs that function in entirely different ways. For example, the lincRNAs NEAT1 and NEAT2 have been recently shown to be important in the formation of paraspeckles (25), and the lincRNA NRON has a role in repressing nuclear import (26). It is possible that additional lincRNAs have roles in these and numerous other cellular pathways. The full range of biological diversity of lincRNAs and their mechanisms clearly remains to be explored.

Materials and Methods

Identification of K4-K36 domains was performed as previously described (1). RIP was performed as previously described (8) with some modifications. Hybridization to tilling arrays was performed as previously described (1). For detailed methods, see *SI Methods*. The data concerning the lincRNAs and the experiments here are available in Dataset S1 and public databases. All microarray data including RNA hybridization to tiling arrays, RIP-Chip experiments, and gene expression profiling of lincRNA knockdowns is deposited at the Gene Expression Omnibus (GEO) under accession no. GSE16226.

ACKNOWLEDGMENTS. We thank Yang Shi and Shigeki Iwase (Harvard Medical School) for antibodies to SMCX and their input on the manuscript; J. P. Mesirov and Pablo Tamayo [The Broad Institute of Harvard and Massachusetts Institute of Technology (MIT)] for discussions and statistical insights; and Miguel Rivera for access to hEK ChIP-Seq data. A.M.K. is supported by National Institutes of Health Training Grant HL007893. M. Guttman is a Vertex scholar. J.L.R is a Damon Runyon-Rachleff Innovation and Smith Family Foundation Scholar. J.L.R. and A. Regev are Investigators of the Richard Merkin Foundation for Stem Cell Research at the Broad Institute. A.R. is supported by the Pioneer award and by the Burroughs Wellcome Fund. This work was supported by the National Human Genome Research Institute, and the Broad Institute of MIT and Harvard.

1. Guttman M, et al. (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458:223–227.
2. Birney E, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447:799–816.
3. Mattick JS (2009) The genetic signatures of noncoding RNAs. *PLoS Genet* 5:e1000459.
4. Bernstein E, Allis CD (2005) RNA meets chromatin. *Genes Dev* 19:1635–1655.
5. Pang KC, Frith MC, Mattick JS (2006) Rapid evolution of noncoding RNAs: Lack of conservation does not mean lack of function. *Trends Genet* 22:1–5.
6. Barski A, et al. (2007) High-resolution profiling of histone methylations in the human genome. *Cell* 129:823–837.
7. Mikkelsen TS, et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448:553–560.
8. Rinn JL, et al. (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* 129:1311–1323.
9. 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.
10. Ku M, et al. (2008) Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet* 4:e1000242.
11. Zhao J, Sun BK, Erwin JA, Song JJ, Lee JT (2008) Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* 322:750–756.
12. Nagano T, et al. (2008) The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science* 322:1717–1720.
13. Pandey RR, et al. (2008) Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Mol Cell* 32:232–246.
14. Dinger ME, et al. (2008) Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res* 18:1433–1445.
15. 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.
16. 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.
17. Andres ME, et al. (1999) CoREST: A functional corepressor required for regulation of neural-specific gene expression. *Proc Natl Acad Sci USA* 96:9873–9878.
18. 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.
19. Affar el B, et al. (2006) Essential dosage-dependent functions of the transcription factor yin yang 1 in late embryonic development and cell cycle progression. *Mol Cell Biol* 26:3565–3581.
20. Young TL, Matsuda T, Cepko CL (2005) The noncoding RNA taurine upregulated gene 1 is required for differentiation of the murine retina. *Curr Biol* 15:501–512.
21. Mattick JS (2001) Non-coding RNAs: The architects of eukaryotic complexity. *EMBO Rep* 2:986–991.
22. Moazed D (2009) Small RNAs in transcriptional gene silencing and genome defence. *Nature* 457:413–420.
23. Maison C, et al. (2002) Higher-order structure in pericentric heterochromatin involves a distinct pattern of histone modification and an RNA component. *Nat Genet* 30:329–334.
24. Bernstein E, et al. (2006) Mouse polycomb proteins bind differentially to methylated histone H3 and RNA and are enriched in facultative heterochromatin. *Mol Cell Biol* 26:2560–2569.
25. Sunwoo H, et al. (2009) MEN varepsilon/beta nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. *Genome Res* 19:347–359.
26. Willingham AT, et al. (2005) A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science* 309:1570–1573.
27. Ponjavic J, Ponting CP, Lunter G (2007) Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res* 17:556–565.

Downloaded at California Institute of Technology on August 6, 2021