Functional classification of long non-coding RNAs by k-mer content

Jessime M. Kirk1,2, Susan O. Kim1,8, Kaoru Inoue1,8, Matthew J. Smola3,9, David M. Lee14, Megan D. Schertzer1,4, Joshua S. Wooten1,4, Allison R. Baker10, Daniel Sprague1,5, David W. Collins6, Christopher R. Horning6, Shuo Wang6, Qidi Chen6, Kevin M. Weeks10, Peter J. Mucha7 and J. Mauro Calabrese8,1*

The functions of most long non-coding RNAs (lncRNAs) are unknown. In contrast to proteins, lncRNAs with similar functions often lack linear sequence homology; thus, the identification of function in one lncRNA rarely informs the identification of function in others. We developed a sequence comparison method to deconstruct linear sequence relationships in lncRNAs and evaluate similarity based on the abundance of short motifs called k-mers. We found that lncRNAs of related function often had similar k-mer profiles despite lacking linear homology, and that k-mer profiles correlated with protein binding to lncRNAs and with their subcellular localization. Using a novel assay to quantify Xist-like regulatory potential, we directly demonstrated that evolutionarily unrelated lncRNAs can encode similar function through different spatial arrangements of related sequence motifs. K-mer-based classification is a powerful approach to detect recurrent relationships between sequence and function in lncRNAs.

This problem extends to the thousands of lncRNAs that lack characterized functions.

Results
K-mer-based quantitation as a means to compare IncRNA sequence content. We hypothesized that lncRNAs with shared functions should harbor sequence similarities that confer shared functions, even if conventional alignment algorithms do not detect the similarity. Our rationale follows. First, most lncRNAs probably have no catalytic activity, suggesting that the proteins they bind in cells define their function. Second, proteins often bind RNA through short motifs, or k-mers, that are between three and eight bases in length, where ‘k’ specifies the length of the motif. Third, the mere presence of a set of protein binding-motifs may be more important than their relative positioning within an lncRNA, meaning that functionally related lncRNAs could harbor related motif contents and still lack linear sequence similarity.

To test our hypothesis, we developed a method of sequence comparison, called SEEKR (sequence evaluation from k-mer representation). In SEEKR, all k-mers of a specified length k (that is, k = 4, 5, or 6, etc.) are counted in one-nucleotide increments across each lncRNA in a user-defined group, such as the GENCODE annotation set43. K-mer counts for each lncRNA are then normalized by lncRNA length and standardized across the group to derive a matrix of k-mer profiles, which consist of z-scores for each k-mer in each lncRNA. The relative similarity of k-mer profiles between any pair of lncRNAs can then be determined via Pearson’s correlation (see Fig. 1a,b and Methods).
SEEKR offers advantages relative to existing alignment algorithms. Foremost, SEEKR does not consider positional information in similarity calculations, allowing it to quantify non-linear sequence relationships. For reasons described above, this functionality might suit IncRNAs better than traditional alignment algorithms developed to detect linear sequence homology between evolutionarily related entities\(^{11,21,22,29,38}\). Second, whereas traditional alignment algorithms can only quantify similarity, SEEKR can quantify similarities and differences using Pearson’s correlation. Third, SEEKR can quantify relationships in groups of IncRNAs despite differences in overall length, whereas length differences can confound traditional alignment algorithms. For example, conventional alignment of a 20-kb and 4-kb RNA is barely informative (80% of the 20-kb RNA would not align), but their k-mer contents can be compared via SEEKR. Lastly, SEEKR is algorithmically efficient; all pair-wise comparisons between human GENCODE IncRNAs can be computed in under 1 min.

Initially, we assessed whether SEEKR could detect previously identified sequence similarities in IncRNAs. We compared k-mer profiles via SEEKR for all pair-wise combinations in a set of 161 IncRNAs recently described to be conserved between human and mouse\(^2\). We also aligned the IncRNAs to each other using two existing alignment algorithms, the hidden Markov model based nhmmer\(^1\), and Stretcher, an implementation of the global alignment algorithm Needleman–Wunsch\(^2\). In this test, SEEKR detected known IncRNA homologs nearly as well as, or better than, both algorithms (Fig. 1c). We defined signal-to-background in this assay as the ratio between the median similarity of homologous and non-homologous IncRNAs. By this metric, nhmmer detected homologs the most clearly, as expected (signal-to-background ratio of 0.695: 0.000), followed by SEEKR (signal-to-background of 0.015: 0.003 at k-mer length k = 6) and Stretcher (signal-to-background of 0.025: 0.307; Fig. 1d). We conclude that k-mer-based classification can detect sequence similarity between evolutionarily related IncRNAs.

We next examined whether SEEKR could detect novel forms of similarity between IncRNAs with no known sequence homology. We created k-mer profiles for all IncRNAs in the human and mouse GENCODE databases\(^3\), as well as for select IncRNAs that were not included in GENCODE. Next, we compared k-mer profiles between all IncRNAs in each organism using Pearson’s correlation and hierarchically clustered the resulting matrices to examine the patterns that emerged. Consistent with our hypothesis, clustering IncRNAs by SEEKR grouped many by known function in human and mouse (Fig. 2). Several known cis-repressive IncRNAs, including XIST, TSIX, KCNQ1OT1, UBE3A-ATS, ANRIL/CDKN2B-AS1, and Airn, clustered together due to high abundance of AU-rich k-mers, whereas several cis-activating IncRNAs, including PCAT6, HOXT10, LINCO00570, DBE-T, and HOTAIR1, clustered separately due to high abundance of GC-rich k-mers (Fig. 2a,d). These patterns were robust over differing k-mer lengths (Supplementary Fig. 2). To determine whether this level of clustering was significant, we curated lists of human and mouse cis-regulators and compared average pair-wise k-mer similarities between IncRNAs in each list to pair-wise similarities of 10,000 size-matched lists of randomly selected IncRNAs from the respective organism. Human and mouse cis-regulators, and human cis-activators (but not mouse cis-activators), were significantly more similar to each other than expected by random chance (Supplementary Table 2). Concordantly, SEEKR detected significant similarity between the cis-repressive Kcnq1ot1 and Xist IncRNAs where none was found by conventional alignment algorithms (Supplementary Fig. 1). We conclude that IncRNAs of related function can have related k-mer profiles even if they lack linear sequence similarity.

Unexpected relationships also emerged in the hierarchical clusters of Fig. 2. Most notably, the IncRNAs NEAT1 and MALAT1 showed greater than average similarity to XIST in both human and mouse. Among all human lncRNA pair-wise comparisons, their Pearson’s r values fell in the 99.99th and 99.60th percentile, respectively. Likewise, in mouse the similarities were in the 97.15th and 95.32nd percentiles. The meaning of the similarity between the three IncRNAs is unclear, but we note that all three IncRNAs seed the formation of subnuclear compartments and engage with actively transcribed regions of the genome\(^3\). We speculate that their k-mer similarity is related to these shared actions.

LncRNAs can be partitioned into communities of related k-mer content. We next used a network-based approach to partition IncRNAs into communities of related k-mer profiles, reasoning that such communities would provide a framework to understand the
predictive value of lncRNA k-mer content. We created networks of relationships between all human and mouse lncRNAs in which weighted edges connected lncRNAs in an organism if the Pearson’s correlation between their standardized k-mer profiles met a threshold for similarity (see Methods). We then used the Louvain method to assign lncRNAs within the largest connected component of the network representations to communities of related k-mer profiles. Approximately half of all GENCODE lncRNAs grouped into five major communities in both human and mouse. LncRNAs not assigned to the five most populated communities were assigned to a ‘null’ community. Our network-based approach and hierarchical clustering grouped lncRNAs in similar ways ($P < 1 \times 10^{-324}$, chi-squared; Supplementary Tables 3 and 4), signaling community robustness. LncRNA community assignments and associated summary statistics are provided in Supplementary Tables 5–12 and Supplementary Fig. 3. Differences in human and mouse community structures may be due in part to differences in completeness of lncRNA annotation. In the versions of GENCODE used for this work, there were about twice as many lncRNAs annotated in human (v22, n = 15,953) as there were annotated in mouse (vM5, n = 8,245; ref. 12).

**K-mer content correlates with localization and protein binding.** We next examined whether lncRNAs with related k-mer profiles shared biological properties. For this analysis, we focused on human lncRNAs, where data from the ENCODE project allowed us to examine lncRNA subcellular localization and protein associations, transcriptome wide. To determine whether k-mer content provides information about lncRNA localization, we examined ENCODE subcellular fractionation RNA-seq (RNA-seq) experiments.
performed in HepG2 and K562 cells\(^5\). For each lncRNA expressed in each cell type, we computed its nuclear ratio and determined whether the distributions of nuclear ratios differed between communities. The majority of communities showed slight but significant differences in their distribution of nuclear ratios, with the largest differences found between communities 1 and 3 (Fig. 3a and Supplementary Tables 13–16). Concordantly, lncRNAs that associate with poly-somes in K562 cells\(^5\) were also non-uniformly distributed between communities (\(P = 3.5 \times 10^{-5}\), chi-squared), and were the most over-and under-represented in the most cytoplasmic and nuclear lncRNA communities, respectively (communities 3 and 1 being the most cytoplasmic and nuclear, respectively; Supplementary Table 17). Lastly, we used ENCODE data to identify the most cytoplasmic and nuclear lncRNAs in HepG2 and K562 cells and determine which k-mers were asymmetrically distributed between lncRNAs in the two compartments. We found that 360 and 27 k-mers were significantly enriched in cytoplasmic and nuclear lncRNAs, respectively (\(P < 0.05\); Kolmogorov–Smirnov test; Supplementary Table 18). Consistent with our RNA-seq and polysome analyses, 58% and 93% of the cytoplasmic- and nuclear-biased k-mers were the most enriched in the most cytoplasmic and nuclear lncRNA communities, respectively (communities 3 and 1; see Supplementary

---

**Fig. 3 | LncRNA localization and protein-binding correlate with k-mer content.** a. Violin plots of IncRNA localization by k-mer community in K562 (orange) and HepG2 (green) cells, as determined from RNA-seq of polyA-selected and ribosome-depleted RNA. N, the null community. Lines show the lower, median, and upper quartile of values (see Supplementary Figs. 13–16 for samples sizes). b. From left to right: \(-\log_{10}(P)\) significance of increase in likelihood, percentage increase in precision, and percentage increase in recall obtained when IncRNA community information is included in a logistic regression to predict protein association. Black line in the left-hand plot corresponds to a \(-\log_{10}(P)\) value of 0.05 (\(n = 3,747\) IncRNAs for HepG2, \(n = 3,278\) IncRNAs for K562). c. Eleven of the 17 proteins with experimentally determined PWMs from ref. \(^23\) show significantly increased abundance of motif-matching k-mers (\(n = 4,096\)) in IncRNA communities that are enriched for binding to the protein in question (*\(P < 0.01\), permutation test). d. The most enriched k-mers in 300-nucleotide windows surrounding motif matches in CLIP peaks do not always match the motif. PWMs from ref. \(^23\) are shown above average z-scores for the top five most enriched k-mers in true-positive relative to false-positive binding regions for the protein in question. PWMs and top k-mers are shown for all 17 proteins in Supplementary Fig. 5.
Table 18, last column). We conclude that k-mer content provides information about the subcellular localization of an lncRNA.

To determine whether k-mer content provides information about protein binding in lncRNAs, we examined ENCODE data for 156 enhanced cross-linking immunoprecipitation (eCLIP) experiments performed for 109 proteins in HepG2 and K562 cells\(^{37}\). We created binary vectors for each experiment that recorded whether the lncRNAs bound or did not bind a given protein, then built separate logistic regression models for each protein to determine whether k-mer community assignments could improve prediction of lncRNA–protein associations over a null model that only included lncRNA length and expression as covariates. LncRNA community assignments significantly increased the log-likelihood of detecting lncRNA–protein associations for the majority of proteins examined (\(P\) adjusted <0.05; 146 of 156, \(\sim 94\%\); Fig. 3b and Supplementary Table 19). Increases in precision and recall in community-informed models were generally modest but significant (Fig. 3b and Supplementary Table 20). In total, \(\sim 17\%\) (25 of 146) of our models had an increase in precision and/or recall of 5% or more. Notably, in all cases in which recall increased, precision also increased, indicating that k-mer community information increased the ability to predict true lncRNA–protein associations and simultaneously increased the fidelity of those predictions. When we used individual 6-mers instead of lncRNA communities as predictive features, results were no better than the null model that used only lncRNA length and expression as predictive features. Models with more features than samples are prone to learning noise in their training set, and often lose predictive power due to overfitting\(^{38}\). Using individual 6-mers brought the number of features being evaluated to 4,099, more than the number of lncRNAs expressed in HepG2 and K562 cells (3,745). We conclude that k-mer content provides information about the protein-binding potential of an lncRNA, but that no single k-mer provides an overwhelming portion of that information, and that k-mer communities provide a way to collapse high-dimensional k-mer matrices down to representative variables for predictive purposes.

Protein binding to RNA is difficult to assess from motif content alone due to the degeneracy of most motifs and the challenge of predicting the effects of RNA structure\(^{34,35,39–41}\). Supporting this notion, we found that the abundance of motif-matching k-mers was consistently, but not always, higher in the communities enriched for binding of specific proteins than in the cognate communities not enriched for binding, indicating that factors in addition to motif abundance control protein–lncRNA associations (Fig. 3c). We therefore sought to determine whether k-mer content could distinguish between motif matches in lncRNAs that coincide with protein binding events and those that do not. We searched the lncRNAs expressed in HepG2 and K562 cells for matches to binding motifs of the 17 proteins in Fig. 3c, whose position weight matrices (PWMs) were determined from biochemical assays in ref.\(^{23}\). We annotated motif matches that fell inside and outside of eCLIP peaks as true and false-positive matches, respectively. As expected, the majority of motif matches fell outside of eCLIP peaks (that is, they were false-positive matches; Supplementary Table 21). We then used SEEKR to compare regional k-mer content in 300-nucleotide windows surrounding true and false-positive motif matches. Remarkably, for 13 of 17 proteins examined, k-mer profiles of true-positive binding regions were more similar to each other than k-mer profiles of randomly selected, size-matched sets of false-positive regions (\(P\) value <0.005; Supplementary Fig. 4). These data support the notion that binding modules for the same protein in different RNAs often have sequence similarity that extends beyond the protein-binding motif, and that this similarity can be quantified, in part, by local k-mer content.

Moreover, SEEKR provides a simple way to visualize the density of specific k-mers within eCLIP-enriched regions. We compared the most overrepresented k-mers in true-positive binding regions to protein-binding motifs measured in vitro\(^{37}\), and found that their relationships differed substantially from protein to protein (Fig. 3d and Supplementary Fig. 5). For certain proteins, such as HNRNPC, KHDRBS1, and QKI, the most enriched k-mers in true-positive regions matched the PWM for the protein that was determined in vitro\(^{37}\). We interpret this observation to mean that, for these proteins, motif density plays a dominant role in determining RNA binding in vivo, because our k-mer data show that motif-matching k-mers are more abundant in true-positive regions than they are in false-positive regions. For other proteins, such as FXR1, IGFBP1, and TIA1, the most enriched k-mers in true-positive regions did not match the PWM determined in vitro\(^{37}\). For these proteins, sequence beyond the binding motif may play a dominant role in dictating association with RNA, possibly due to effects from RNA structure. When PWMs were extracted from eCLIP peaks, similar relationships between k-mers and in vitro-defined motifs were observed (Supplementary Fig. 5). These results show how SEEKR can be used to augment traditional motif-based analyses and provide insights into mechanisms of RNA–protein interaction. SEEKR provides a way to quantify sequence similarities between any number of protein-binding regions, which, in turn, can provide predictive power and identify shared characteristics that are not apparent from PWM-based motif analyses.

**Similarities in lncRNA communities between organisms.** Given (1) that k-mer content provides some indication of protein-binding potential in an lncRNA, (2) that sequence specificities of many RNA binding proteins are conserved\(^{37}\), and (3) that protein binding probably dictates lncRNA function, we hypothesized that k-mer contents between communities of functionally related lncRNAs could be conserved even if the lncRNAs themselves lack known evolutionary relationships. In support of this idea, we identified extensive similarity between certain human and mouse lncRNA communities via SEEKR (see Methods and Supplementary Fig. 6). Most notably, lncRNAs in human community 1 (the ‘XIST’ community) had k-mer profiles that were, as a group, nearly indistinguishable from lncRNAs in mouse community 1 (the ‘Xist’ community) and were also similar to lncRNAs in mouse community 4 (\(P<0.0001\) for both comparisons). Human community 2 and community 3 (the ‘HOTTIP’ community) were both similar to mouse community 2 (the ‘Hottip’ community; \(P<0.0001\)). No other major similarities between mouse and human were apparent. Extending this analysis across greater evolutionary distance, we found HOTTIP-like lncRNA communities in ten of ten vertebrates examined as well as in the sea urchin *Strongylocentrotus purpuratus*, and XIST-like lncRNA communities in seven of ten vertebrates examined (Supplementary Figs. 7–9; ref.\(^{10}\)). These analyses demonstrate that, at the level of k-mers, subsets of human lncRNAs are more similar to lncRNAs in other genomes than they are similar to lncRNAs in their own genome, supporting the idea that groups of lncRNAs have similar function in different organisms despite lacking obvious linear sequence similarity.

**SEEKR can predict Xist-like regulatory potential in lncRNAs.** We next directly tested whether k-mer profiles could be used to predict lncRNA regulatory potential. We focused on the ability of certain lncRNAs to repress transcription in cis. cis-Repression was one of the earliest characterized functions of lncRNAs, and is essential for normal human health and development. In the most striking example, the *XIST* lncRNA silences nearly all genes across an entire chromosome during X-chromosome inactivation\(^{‘}\). cis-Repression is also one of most straightforward lncRNA functions to study because, by definition, cis-acting lncRNAs act near their site of transcription.
We developed a reductionist assay to study IncRNA cis-repressive activity in a normalized genomic context, called TETRIS (transposable element to test RNAs effect on transcription in cis). TETRIS enables the sequence of an IncRNA and an adjacent reporter gene to be manipulated in a plasmid, but then rapidly inserted into chromosomes via the piggyBac transposase, so that effects of the IncRNA on the reporter can be studied in genomic chromatin (see Fig. 4a and Methods). Under our assay conditions, piggyBac catalyzes 4–7 insertions of each cargo per stably selected cell, and cell density estimates suggest that between 100,000 and 500,000 cells receive insertions and survive selection (Fig. 4b and data not shown). Thus, each TETRIS assay probably surveys 400,000–3,500,000 insertion events. Insertion-site-dependent variations in IncRNA-induced effects are averaged out in the population, bypassing the need to isolate clones of modified cells, and providing the means to quantify IncRNA regulatory potential without influence from genomic position.

We validated TETRIS by comparing effects that expression of different IncRNAs had on luciferase activity. A cell line created from a vector that lacked an IncRNA insert (TETRIS-Empty) showed an approximately twofold increase in luciferase activity upon addition of doxycycline, representing our baseline for the assay (Fig. 4c). We attribute this mild activation to the close proximity of the dox-inducible and luciferase promoters, and to the fact that both promoters are contained within the same insulated domain. By contrast, expression of the first 2 kb of Xist repressed luciferase fivefold relative to no dox, demonstrating that Xist-2kb cell lines, respectively (mean ± s.d. of 2.03 ± 0.50 and 0.23 ± 0.08), showing that TETRIS assays result in reproducible effects on luciferase activity. For its repressive effect, Xist requires ‘Repeat A’, a 425-nucleotide-long element contained within its first 2 kb. In the context of TETRIS, deletion of Repeat A resulted in a significant, but not
Fig. 5 | Mapping of elements required for repression by Xist-2kb in TETRIS. a, b. Minimum free energy (MFE) (a) and arc-based structural models (b) of the first 2 kb of Xist from ref. 41; green and blue bars mark starts and stops of indicated regions; locations of Xist repeats and predicted stable structures; low S/S, regions of low SHAPE reactivity and Shannon entropy from ref. 41) are also shown. c. Deleted regions. d. Effects on luciferase after dox addition. *Bonferroni-corrected P <0.001 relative to Wild-type/Xist-2kb via Student’s t test. Tukey boxplots show the lower, median, and upper quartile of values, and ±1.5× interquartile range (see Supplementary Table 22 for sample sizes and exact P values). Min., Minimal; br., broad.
complete, de-repression of luciferase, whereas expression of Repeat A alone resulted in repression relative to control, but at reduced levels compared to Xist-2kb (‘ΔrepA’ and ‘repA only’; Fig. 4c). Similarly, expression of the first 5.5 kb of Xist caused a fivefold repression of luciferase, whereas deletion of the first 2 kb from the 5.5 kb construct caused complete loss of repressive activity (‘Xist-5.5kb’ and ‘Xist-2-5.5’; Fig. 4c). Expression of either the final 3.3 kb of Xist or the Hottip lncRNA had no repressive effect (Fig. 4c). These experiments demonstrate (1) that TETRIS is a suitable assay to measure repression by cis-acting lncRNAs in a normalized genomic context, and (2) in the assay, sequence elements in addition to Repeat A cooperate to encode repressive function in the 5’ end of Xist.

We next used TETRIS and SEEKR to test our hypothesis that k-mer content can predict lncRNA regulatory potential. We reasoned that we could design entirely synthetic lncRNAs that lacked linear sequence similarity to any known lncRNA but nonetheless had robust Xist-like repressive activity. We generated six synthetic lncRNA sequences in silico with varying levels of k-mer similarity to the first 2 kb of Xist, and cloned them into TETRIS to measure their effects on luciferase activity. As measured by SEEKR, the lncRNAs had Pearson's similarities to Xist that ranged from average (a Pearson's r of ~0) to 3 s.d. above the mean similarity for all mouse lncRNAs (a Pearson's r of 0.19, more similar to Xist-2kb than all other lncRNAs in the mouse genome; see Fig. 4d). Using nhmmer or Stretcher to align the synthetic lncRNAs to the first 2 kb of Xist produced either no alignments (nhmmer) or alignments that differed by only 3% across all six synthetic lncRNAs (Stretcher; see Fig. 4c, grid below graph). Via BLAST, the lncRNAs had no significant similarity to the mouse genome or to each other (not shown). The lack of informative alignments was expected because the synthetic lncRNAs have no evolutionary relationship with Xist, any region in the genome, or each other. Nevertheless, as envisioned, the synthetic fragments that SEEKR classified to be most similar to Xist had the highest repressive activity (Fig. 4c). These data directly demonstrate that evolutionarily unrelated lncRNAs can encode similar function through different spatial arrangements of related sequence motifs. Thus, k-mer content can be used to predict lncRNA regulatory potential.

We next examined whether SEEKR could predict Xist-like repressive activity in endogenous lncRNAs. We cloned into TETRIS 33 lncRNAs or lncRNA fragments that had a range of k-mer similarities to the first 2 kb of Xist. Included in our final set of fragments were several conserved lncRNAs and/or shorter fragments contained within them (Airn, Hottip, Kcnqlot1, Malat1, Noat1, and Pvt1), as well as many lncRNAs with uncharacterized functions (Supplementary Table 22). Again, the more Xist-like an lncRNA fragment was at the level of k-mers, the more likely it was to repress in TETRIS; the Pearson's r value between Xist-like Xist at a k-mer length of 6 and luciferase activity upon dox addition was −0.41 (P = 0.02). Including the 6 synthetic lncRNAs in the correlation brought the Pearson's r value to −0.52 (P = 0.0007; Fig. 4f). Nhmer and Stretcher had no ability to predict repressive activity, demonstrating that these algorithms cannot detect sequence signatures correlated with repressive activity in this setting (P = 0.32 and 0.91, respectively; Fig. 4g,h). LncRNA fragment length also had no ability to predict repressive activity (r = 0.03, P = 0.84).

Lastly, we examined whether k-mer profiles associated with sequence elements required for repression by Xist-2kb might increase our ability to predict repressive activity in other lncRNAs. To determine the elements in Xist-2kb required for repression, we made a series of 26 deletions (Fig. 5). Surprisingly, 15 of the deletions, including ones that removed predicted stable structures, pseudoknots, and ~40% of Repeat A (‘ASS1’, ‘ASS2’, ‘APK2’, ‘ASS3’, and ‘ASS4’; see bottom panel in Fig. 5; ref. 41), had no significant effect on repression. However, removal of all 8 GC-rich portions of Repeat A, but not its U-rich linkers, caused an approximately threefold reduction in repression (‘ΔGC repeat in rA’ versus ‘ΔU spacer in rA’), as did removal of 3 predicted stable structures and their intervening sequences in the 742 nucleotides immediately downstream of Repeat A (‘ASS23/4 broad’; ref. 41). Co-deletion of Repeat A and the stable structures had an additive effect, causing a near complete loss of repression (the ‘ΔrAASS234 b’ mutant), whereas expression of Repeat A or the stable structures alone had half the repressive potency of Xist-2kb (‘Only rA’ and ‘Only SS234’). Expression of both regions together had the same repressive potency as Xist-2kb (‘Minimal’). Thus, in TETRIS, the major elements required for repression are contained between nucleotides 308 and 1,476 of Xist. Based on prior structural models4,48, we infer that the elements are comprised of protein binding sites, spacer sequences, and stable structures.

Having mapped the elements responsible for repression in Xist-2kb, we attempted to extract subsets of 6-mers from them that increased our ability to predict Xist-like repression. We also examined whether k-mer variance across lncRNA communities or k-mer nucleotide composition could be used to extract subsets of outperforming 6-mers, and whether different k-mer lengths had better predictive power than k = 6. No rationally designed subset of 6-mers could predict repression better than the full 6-mer profile of Xist-2kb, nor could any other k-mer length (Supplementary Fig. 10). These results support the ideas that different lncRNAs can encode similar function through related, but not necessarily identical, sequence solutions, and that the full complement of 6-mers may be a broadly effective search tool to identify such similarities (not too relaxed, not too stringent).

Discussion

Collectively, our data support the notion that many lncRNAs function through recruitment of proteins that harbor degenerate RNA-binding motifs, and that spatial relationships between protein-binding motifs in these lncRNAs are often of secondary importance to the concentration and effectiveness of the motifs themselves. By this logic, an lncRNA may merely need to present the appropriate motifs embedded within the appropriate structural contexts to achieve a specific function. Thus, different lncRNAs probably encode similar function through vastly different sequence solutions, and non-linear sequence comparisons can be used to discover similarities between them. By extension, because the RNA-binding motifs of many proteins are conserved42-48, it is plausible that groups of lncRNAs rely on similar motifs to encode related function in different organisms even though they lack direct evolutionary relationships. This concept is supported by our observation that lncRNA communities with related k-mer contents exist in human, mouse, and other organisms. We propose that non-linear sequence homology—in which the relative abundance of a set of protein-binding motifs is conserved, but the sequential relationships between them are not—is prevalent in lncRNAs. To quantify non-linear homology, we introduce SEEKR, a method to compare sequence content between any group of lncRNAs, regardless of the size of the group, the evolutionary relationships between the lncRNAs being analyzed, or the differences in their lengths. Each lncRNA (and each functional domain within each lncRNA) has its own k-mer signature, which can encode information about protein binding and RNA structure. SEEKR provides a simple way to tie this information to a biological property.

URLs. SEEKR, https://github.com/CalabreseLab/seekr; nhmmer, http://hmmer.org/download.html; Biopython, http://biopython.org; CHAMP, https://github.com/weiri827/champ; Gephi, https://gephi.org/; ENCODE RNA-seq, http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCshlLongRnaSeq/; MEME, http://meme-suite.org/doc/fimo.html.
Online content
Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41588-018-0207-8.

Received: 24 July 2017; Accepted: 24 July 2018; Published online: 17 September 2018

References
1. Iyer, M. K. et al. The landscape of long noncoding RNAs in the human transcriptome. Nat. Genet. 47, 199–208 (2015).
2. Geisler, S. & Coller, J. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. Nat. Rev. Mol. Cell Biol. 14, 699–712 (2013).
3. Holoch, D. & Moazed, D. RNA-mediated epigenetic regulation of gene expression. Nat. Rev. Genet. 16, 71–84 (2015).
4. Liu, X., Hao, L., Li, D., Zhu, L. & Hu, S. Long non-coding RNAs and their biological roles in plants. Genetics Proteomics Bioinformatics 13, 137–147 (2015).
5. Rinn, J. L. & Chang, H. Y. Genome regulation by long noncoding RNAs. Annu. Rev. Biochem. 81, 145–166 (2012).
6. Gutschner, T. & Diederichs, S. The hallmarks of cancer: a long non-coding RNA point of view. RNA Biol. 9, 703–719 (2012).
7. Lee, J. T. & Bartel, D. M. X-inactivation, imprinting, and long noncoding RNAs in health and disease. Cell 152, 1308–1323 (2013).
8. Wu, X. & Sharp, P. A. Divergent transcription: a driving force for new gene origin? Cell 155, 990–996 (2013).
9. Cech, T. R. & Steitz, J. A. The noncoding RNA revolution-trashing old rules to forge new ones. Cell 157, 77–94 (2014).
10. Henrion, H. et al. Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. Cell Rep. 11, 1110–1122 (2015).
11. Cabili, M. N. et al. Integrative annotation of human large intergenic non-coding RNAs reveals global properties and specific subclasses. Genes Dev. 29, 1915–1927 (2015).
12. Derrien, T. et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 22, 1775–1789 (2012).
13. Bateman, A. et al. Uniprot: a hub for protein information. Nucleic Acids Res. 43, D204–D212 (2015).
14. Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. Nat. Struct. Biol. 10, 980 (2003).
15. Ulitsky, I. & Bartel, D. P. IncRNAs: genomics, evolution, and mechanisms. Cell 154, 26–46 (2013).
16. Kutter, C. et al. Rapid turnover of long noncoding RNAs and the evolution of gene expression. PLoS Genet. 8, e1002841 (2012).
17. Necsulea, A. et al. The evolution of IncRNA repertoires and expression patterns in tetrapods. Nature 505, 635–640 (2014).
18. Eddy, S. R. Computational analysis of conserved RNA secondary structure in transcriptomes and genomes. Annu. Rev. Biophys. 43, 433–456 (2014).
19. Quinn, J. J. et al. Rapid evolutionary turnover underlies conserved IncRNA-genome interactions. Genes Dev. 30, 191–217 (2016).
20. Eddy, S. R. Homology searches for structural RNAs: from proof of principle to practical use. RNA 21, 605–607 (2015).
21. Wheeler, T. J. & Eddy, S. R. nhmmer: DNA homology search with profile HMMs. Bioinformatics 29, 2487–2489 (2013).
22. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16, 276–277 (2000).
23. Ray, D. et al. A compendium of RNA-binding motifs for decoding gene expression. Nature 499, 172–177 (2013).
24. Stell, R., Skrivskova, L. & Allan, F. H. RNA sequence- and shape-dependent recognition by proteins in the ribonucleoprotein particle. EMBO Rep. 6, 33–38 (2005).
25. Edgar, R. C. & Batzoglou, S. Multiple sequence alignment. Curr. Opin. Struct. Biol. 16, 368–373 (2006).
26. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. J. Mol. Biol. 215, 403–410 (1990).
27. Pervouchine, D. D. et al. Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. Nat. Commun. 6, 5903 (2015).
28. Chadwick, B. P. Variation in Xi chromatin organization and correlation of the H3K27me3 chromatin territories to transcribed sequences by microarray analysis. Chromosoma 116, 147–157 (2007).
29. Engebrecht, J. M. et al. RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent Pre-mRNAs and chromatin sites. Cell 159, 188–199 (2014).
30. Mak, W. et al. Mitotically stable association of polycistrain group proteins eed and enxl with the inactive X chromosome in trophoblast stem cells. Curr. Biol. 12, 1016–1020 (2002).
31. Wutz, A., Rasmussen, T. P. & Jaenisch, R. Chromosomal silencing and rapid turnover of the piggyBac (PB) transposon in mammalian cells and mice. Cell 122, 473–483 (2005).
32. Hezroni, H. et al. Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. Cell Rep. 11, 1110–1122 (2015).
33. Eddy, S. R. Homology searches for structural RNAs: from proof of principle to practical use. RNA 21, 605–607 (2015).
34. Wheeler, T. J. & Eddy, S. R. nhmmer: DNA homology search with profile HMMs. Bioinformatics 29, 2487–2489 (2013).
35. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16, 276–277 (2000).
36. Ray, D. et al. A compendium of RNA-binding motifs for decoding gene expression. Nature 499, 172–177 (2013).
37. Stell, R., Skrivskova, L. & Allan, F. H. RNA sequence- and shape-dependent recognition by proteins in the ribonucleoprotein particle. EMBO Rep. 6, 33–38 (2005).
38. Edgar, R. C. & Batzoglou, S. Multiple sequence alignment. Curr. Opin. Struct. Biol. 16, 368–373 (2006).
39. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. J. Mol. Biol. 215, 403–410 (1990).
40. Pervouchine, D. D. et al. Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. Nat. Commun. 6, 5903 (2015).
41. Chadwick, B. P. Variation in Xi chromatin organization and correlation of the H3K27me3 chromatin territories to transcribed sequences by microarray analysis. Chromosoma 116, 147–157 (2007).
42. Engebrecht, J. M. et al. RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent Pre-mRNAs and chromatin sites. Cell 159, 188–199 (2014).
Methods

Kcnqlot1 versus Xist comparison. Kcnqlot1 was aligned to Xist using nimmer and Stretcher with default parameters. To assess significance of the alignments, we generated 1,000 pseudo-Kcnqlot1s that were the same length as real Kcnqlot1 but composed of nucleotides randomly selected from a distribution of the mononucleotide content of Kcnqlot1 (0.202C:0.235A:0.250G:0.313T). We then aligned the pseudo-lncRNAs to Xist with nimmer and Stretcher and as compared their k-mer contents relative to all other mouse lncRNAs at k-mer length k = 6 via SEEKR.

SEEKR. In SEEKR, a matrix of k-mer counts for a user-defined set of lncRNAs is created by counting all occurrences of each k-mer in each lncRNA in one-nucleotide increments, and then dividing those counts by the length of the corresponding lncRNA. Z-scores are then derived for each k-mer in each lncRNA by subtracting the mean-length-normalized abundance of each k-mer in the group of lncRNAs being analyzed from the length-normalized abundance of the k-mer in the lncRNA of interest. Typically, divisions that differ by more than 3 standard deviations in abundance of that k-mer in the group of lncRNAs being analyzed. We refer to the array of z-scores for each k-mer in a given lncRNA as its k-mer profile. Similarity between any two lncRNAs can be calculated by comparing their k-mer profiles with Pearson’s correlation.

Our rationale for length normalization in SEEKR follows. Without length normalization, k-mer profiles become difficult to interpret for lncRNAs of different lengths. For example, an RNA that is ten times longer than another RNA will have ten times the number of k-mers. Without normalization, these lncRNAs would be considered dissimilar by SEEKR, regardless of the similarity in their relative concentrations of k-mers. By length normalizing, SEEKR creates a list of relative k-mer content in a given lncRNA that is robust to differences in length. The idea that length normalization is important is supported by studies of known cis-repressive lncRNAs. At 18 kb, the Xist lncRNA is the most potent cis-repressive lncRNA known. At least three other known cis-repressive lncRNAs are longer than Xist, but less potent: Airn, Kcnqlot1, and Ube3a-ATS are 90 kb 85 kb, and 1.1 Mb, respectively. These, the longest lncRNA, Ube3a-ATS, is the least potent, arguing that length alone does not account for lncRNA potency. In certain biological contexts, lncRNA length may be relevant, or it may have varying influence on lncRNA function. However, what these contexts might be and to what extent length does or does not affect lncRNA function in them are not known and difficult to predict. We also note that Pearson’s correlation inherently normalizes for length. Thus, comparisons of k-mer content that use Pearson’s correlation will eliminate length as a variable.

GENCODE lncRNA annotations. All GENCODE annotations used in this work were from human build v22 and mouse build vMs. For each lncRNA, only the major splice annotation was considered (the -001 isoform). In total, there were 15,953 human and 8,245 mouse transcripts. The heat maps in Fig. 2 were generated with GENCODE annotations plus the additional lncRNA sequences downloaded from the UCSC genome browser25: SAMSMSN, XACT, UBE3A-ATS, MORRBD, and NESPAS (human); and unspliced Airn, Arml, Bvht, Haunt, Morrbd, unspliced Tiss, Ube3a-ATS, Xistar, and Uppherhand (mouse).

Conservation analysis. Ninety-three pairs of human and mouse GENCODE lncRNAs were recently identified as putative homologs due to their high conservation at the DNA level34. These 93 lncRNAs, plus an additional 68 lncRNA pairs that had equivalent names in mouse and human GENCODE annotations, formed the final set of 161 homologs that were used for the conservation analysis of Fig. 1c. For the Fig. 1c experiment, ‘signal’ values were computed as the mean of the 161 homolog-to-homolog measurements in each of the three algorithms; likewise, background values were computed as the mean of the remaining 12,880 non-homologous comparisons. Homologous pairs were defined as being ‘detected’ if the signal value/average similarity (as determined via SEEKR, nhmmer, or Stretcher) was greater than 0.13. We also compared communities generated with 5-mers and 7-mers to those generated with 6-mers. We created contingency tables that compared the distribution of lncRNAs in each of the five major 6-mer communities plus the null to the distribution of lncRNAs in each of the five major 5-mer and 7-mer communities plus their respective nulls. P values comparing communities between the k-mer lengths were all <1 × 10−110 (chi-squared), indicating that community definitions are largely stable when 5-mers, 6-mers, or 7-mers are used (Supplementary Tables 9 and 10). This stability, the quality of our TETRIS predictions when using 6-mers (Supplementary Fig. 10), and the computational inefficiency of performing operations on matrices of k-mers with length k = 7 or greater provided additional support for our decision to use 6-mers for the bulk of our analyses.

We also compared communities within the five major 6-mer communities between the two sets of communities were significantly similar via a chi-squared test. In both human and mouse, the P value was <1 × 10−100 (Supplementary Tables 3 and 4).

LncRNA localization analysis. Localization data were downloaded from ENCODE (see URLs) as fastq files and aligned to GRCh38 with STAR using default parameters27. FeatureCounts was used to tabulate the number of reads aligning to our set of lncRNAs28. We then filtered out all lncRNAs with <0.1 reads per kilobase of transcript per million aligned reads (RPKM) from each community, and calculated the number of reads in the nuclear fraction over the total number of reads from both the nuclear and cytosolic fractions for each lncRNA.

To determine whether specific k-mers were enriched in cytosolic or nuclear lncRNAs, we selected cytosolic- and nuclear-enriched subgroups of lncRNAs that were expressed in HepG2 or K562 cells. Because the subcellular distribution values for HepG2 or K562 expressed lncRNAs were not normally distributed (Fig. 3a), we needed to employ different thresholds to define cytosolic and nuclear so that the two groups would include similar numbers of lncRNAs. ‘Cytosolic’ lncRNAs were defined as any lncRNA that was more than 50% cytosolic, which resulted in 5,756 transcripts. Then, the average difference in k-mer abundance between lncRNAs in the two compartments, we calculated the mean value of the z-scores for each k-mer in each group, and then used the difference between the means as the metric to calculate the nuclear-enrichment score (Supplementary Table 18). To test for significant differences between the distributions of z-scores between lncRNAs in the two compartments, we used a KS-test and calculated a Bonferroni correction. This analysis yielded 387 k-mers whose distributions differed significantly between cytosolic and nuclear lncRNAs (P value <0.05; Supplementary Table 18).
Using only the lncRNAs from community 3, we repeated the process of applying the Louvain algorithm to define communities and measure cellular localization in order to rule out the possibility that potential subcommunities were responsible for the cytosolic nature of community 3. The Louvain algorithm found four main subcommunities and all smaller subcommunities were grouped into a fifth community. The results of analysis of variance tests indicated that there were no significant differences between any of the communities for either the polyA-selected or ribosome-depleted RNA-seq data. We performed this analysis again for community 1, but no subcommunities were found to be significantly different (Supplementary Fig. 11). This uniformity of cellular localization among possible subcommunities provides biological support for our original community definitions.

LncRNA polysome association. A recent study found 229 lncRNAs in GENCODE v22 that were polysome-associated in K562 cells. A chi-squared test showed that these 229 lncRNAs were non-randomly distributed between the communities (P value = 3.5 × 10⁻⁶; Supplementary Table 17). The expected values for the chi-squared test were calculated by filtering all communities for lncRNAs expressed in K562 cells, dividing the number of incRNAs in each community by the total number of expressed lncRNAs (3,277), and multiplying by the number of polynomials for each incRNA (229).

LncRNA–protein association data. eCLIP data were downloaded from ENCODE [35,36]. For each of the 156 eCLIP experiments 'bed narrowPeak' data (representing sites of protein binding that passed an ENCODE-defined threshold for enrichment over background; refs [35,36]) were pooled from available biological duplicates. Genomic coordinates were overlapped with lncRNA exon coordinates annotated by GENCODE. Any lncRNA that overlapped with one or more eCLIP peak was considered as having a true binding interaction with the given protein. LncRNA expression data were collected from ENCODE RNA-seq experiments in the same cell type as that of the eCLIP experiment (HepG2 or K562).

For each protein, a vector was built for each lncRNA that encoded whether the protein-lncRNA pair did or did not interact. Next, two feature matrices (null and full) were constructed. The null matrix included the log normalized values for length and expression of each of the lncRNAs. The full matrix included log normalized length and expression, as well as an additional five columns that corresponded to each of the five incRNA communities. Each lncRNA was assigned a value of 1 in the column representing its community.

Models of protein associations. To address whether lncRNA communities contained information about lncRNA–protein associations, we used a machine learning model [37]. We tested whether providing the model with the community data allowed it to predict interactions better than a corresponding null model that was not given the community data but still included lncRNA length and expression allowed it to predict interactions better than a corresponding null model that was not given the community data but still included lncRNA length and expression.

Calculating the abundance of motif-matching k-mers in lncRNA communities. The data for the bar graph in Fig. 3c were generated by the following approach. Of the 109 proteins on which eCLIP was performed in ref. 23, 79 showed significant association with at least one k-mer community over the null (Supplementary Table 19). Of these 79 proteins, binding motifs for 17 were determined via an in vitro binding assay in ref. 23. The PWMs for each of these 17 proteins contained relative weights for each motif matching 6-mer, representing the likelihood that the k-mer in question would bind the protein in question. We multipled the weight of each motif matching 6-mer by its average standardized abundance in each of the six communities, including the null, to obtain k-mer abundances that were scaled by the likelihood that the k-mer in question matched the binding motif in question. For each of the 17 proteins, sums of the weighted abundance for all motif-matching k-mers were created for the communities in which protein binding was enriched and not enriched over the null, respectively, then divided by the number of communities in each group to obtain the average weighted abundance in each of the six communities. To address whether lncRNA communities

Measuring k-mer similarity surrounding motif matches in lncRNAs. The lncRNAs expressed in HepG2 and K562 cells were examined for motif matches to the 17 proteins for which eCLIP data was reported in ref. 23 and whose PWMs were determined via a high-throughput in vitro assay in ref. 23 by using FIMO at a threshold of P < 0.01 (from the MEME suite, see URLs; ref. 39; Supplementary Table 21). Each motif match was then labeled as a true positive if it overlapped the eCLIP peak, or a false positive if it did not. For each protein, the sequences surrounding the center of each true- and false-positive motif match (up to 150 bp on either side of the center, or up to the end of the gene, whichever came first) were collected and their k-mer contents were analyzed with SEEKR. Significance of the similarity between true-positive regions was measured by permutation test against randomly selected sets of false-positive regions controlling for both the size of the set and the number of overlapping regions in the set (Supplementary Fig. 4).

Identifying motifs from eCLIP peaks. To find motifs in eCLIP peaks for the 17 proteins listed in Fig. 3c, we extracted the subset of sequences from eCLIP peaks whose CLIPper-defined P value was <0.001 (peaks with the highest read densities relative to control; ref. 39). We searched these sequences for motifs using DREME at default parameters as a part of the MEME-CLIP package [40].

Human-to-mouse and human-to-other community similarity calculations. To evaluate the similarity between human and mouse lncRNA communities, we calculated the distribution of similarities between all pair-wise combinations of lncRNAs within each human k-mer community (‘human-to-self’), and compared this distribution to: (1) a distribution of pair-wise comparisons made between all other human lncRNAs excepting lncRNAs from the community in question (‘human-to-other-human’), (2) distributions of all pair-wise comparisons made between all lncRNAs in each of the five mouse lncRNA communities (‘human-to-mouse’), and (3) distributions of all pair-wise comparisons made between all human and mouse lncRNAs that did not fall into one of the five major communities (‘human-to-null’). We then performed a permutation test to determine whether a given human community was similar enough to a mouse community to overcome its intrinsic similarity to other lncRNAs in the human genome. The expectation was that, for related communities, the human-to-mouse distribution would be more similar to the human-to-self distribution than it would be to the human-to-other-human and human-to-null distributions. Bonferroni-adjusted P values were calculated by permutation tests where we iteratively subsampled 0.1–1% of each distribution, re-measured the mean pair-wise similarities, counted the number of trials in which the human-to-mouse mean PWM sample was closer to the human-to-other-human mean PWM than to the human-to-self mean, and, finally, divided by the total number of trials performed (36,000). This bootstrapping procedure provided a statistical framework to determine whether the similarities uncovered between human and mouse communities were greater than would have been expected from random chance. For example, the distribution of similarities between a randomly selected subset of IncRNAs from human community 1 and size-matched subsets of IncRNAs from mouse community 1 was always more similar to the distribution of similarities between all pair-wise comparisons of the human community 1 subset than it was to the distribution of similarities between the human community 1 subset and size-matched subsets of IncRNAs from human community 1 human IncRNAs (see upper left panel in Supplementary Fig. 6, ‘H-1 versus M-1’ plot—the H-1 versus H-1 distribution in red is nearly indistinguishable from the H-1 versus M-1 distribution in purple).

NATURE GENETICS | www.nature.com/naturegenetics
To generate the plots in Supplementary Figs. 8 and 9, identical analyses were performed that compared human lncRNA communities to lncRNA communities from rabbit, dog, opossum, chicken, lizard, coelacanth, zebrafish, stickleback, Nile tilapia, elephant shark, and sea urchin. In these latter cases, the human XIST and HOTTIP lncRNAs were doped into the lncRNA annotation set from the organism in question to find the homologous communities that were the most XIST- and HOTTIP-like (Supplementary Fig. 7).

Generation of plasmids for TETRIS assays. The pTETRIS-cargo vector was created from components of a cumate-inducible piggyBAC transposon vector (System Biosciences), pGL4.10-Luciferase (Promega), and pTRE-Tight (Clontech). Briefly, a 567-bp fragment containing a minimal mouse PGK promoter was cloned into a SacI site in pGL4.10-Luciferase to generate pGL4-PGK-Luc-pa. The reverse complement of PGK-Luc-pa was cloned into a vector containing the bovine growth hormone polyA site. The entire bGHp-[reversePGK-Luc-pa] was cloned into NotI and Sall sites of the piggyBAC vector (System Biosciences). The cumate-inducible promoter in the piggyBAC vector was then replaced with the Tetracycline Responsive Element (TRE) from pTRE-Tight (Clontech) via Gibson assembly to generate pTETRIS-cargo in Fig. 4a, in which the lncRNA, the luciferase gene, and a gene encoding puromycin resistance are all flanked by chicken HS4 insulator elements, and inverted terminal repeats recognized by the piggyBAC transposase. The rTtA-cargo vector from Fig. 4a was generated by cloning the hUbic-rTtA3-IRES-Neo cassette from pSLIK-Neo (Addgene Plasmid no. 25735) into Sall and Sall sites in a piggyBAC transposon vector (System Biosciences). The piggyBAC transposase from System Biosciences was cloned into Smal and HindIII sites in pUC19 (NEB) to allow propagation of the transposase on ampicillin plates.

Generation of TETRIS-lncRNA Cargo vectors. lncRNA fragments were PCR-amplified from genomic DNA or bacterial artificial chromosomes using Phusion DNA Polymerase (NEB), or commercially synthesized (Genewiz, IDT), and cloned via Gibson assembly into the Swal site of pTETRIS-Cargo. Insert size was verified by restriction digestion, and the 5’ and 3’ end of each insert was verified by Sanger sequencing. To generate mutant Xist-2kb constructs, the 2-kb fragment of Xist was subcloned into pGEM-T-Easy, and the regions in question were deleted using site-directed mutagenesis, or by synthesis of a mutated fragment and re-cloning back into compatible sites in pGEM-Xist-2kb (Genewiz). Deletions were verified by Sanger sequencing and then assembled into the Swal site of pTETRIS-Cargo. The sequences of all inserted fragments, including Xist-2kb mutations, are listed in Supplementary Table 22.

Estimation of TETRIS copy number per cell. Genomic DNA was prepared from biological triplicate derivations of TETRIS-GFP and TETRIS-Xist-2kb cell lines. qPCR signal (SsoFast, Biorad) from the genomic DNA was compared to signal from a molar standard amplified from increasing amounts of the corresponding TETRIS plasmid (Supplementary Table 23).

TETRIS assays. To generate stable TETRIS-lncRNA cell lines, 8 × 10^4 embryonic stem cells (ATCC CRL-1821) were seeded in a single well of a 6-well plate, and the next day transfected with 0.5 µg TETRIS cargo, 0.5 µg rTtA-cargo, and 1 µg of pUC19-piggyBAC transposase. Cells were subsequently selected on puromycin (2 µg/ml) and G418 (200 µg/ml) for 6–12 d. Due to the efficiency of piggyBAC cargo integration and the rapidity of puromycin selection, all observable death from drug selection occurred within ~3 d after addition of puromycin and G418 (that is, cells with puromycin resistance were invariably resistant to G418). For luciferase assays, 1 × 10^4 cells per well of a 24-well plate were seeded in triplicate from each biological replicate preparation of a stable TETRIS-lncRNA cell line. At 24 h post seeding, medium was changed to include doxycycline at a final concentration of 1 µg/ml. After 2 d growth in dox-containing media, cells were lysed with 100 µl lysis buffer (Promega), and luciferase activity was measured using Bright-Glo Luciferase Assay reagents (Promega) on a PHERaStar FS plate reader (BMG Labtech). Luciferase activity was normalized to protein concentration in the lysates via Bradford assay (Biorad). Each lncRNA fragment was assayed at least in triplicate from at least two independent biological replicate preparations of stable TETRIS-lncRNA cell lines.

Synthetic lncRNA design. Synthetic lncRNAs were designed by generating 10,000,000, 1,650-nucleotide-long lncRNAs in silico that compared human lncRNA communities to lncRNA communities selected and synthesized as geneBlocks (Integrated DNA Technologies) and Gibson assembled into the Swal site in TETRIS. Similarities in k-mer content to the 2-kb fragment of Xist are relative to all other mouse GENCODE lncRNAs.

Visualization of Xist structural models. Minimum free energy and probability-arc structural models of Xist-2kb were generated using SHAPE-MaP data from ref. 48, the visualization package VARNA10, and a modified version of the IGV browser11. Predicted pseudoknots and regions of low SHAPE reactivity and low Shannon entropy in Xist-2kb are from ref. 57.

TETRIS predictions for k-mer sizes and subsets. We measured SEEKR’s ability to capture the relationship between an lncRNA’s Xist-likelihood and its repressive ability in the TETRIS assay using k-mers from size one to eight. In each case, the correlation is measured using the means of all biological and technical replicates of each real and synthetic lncRNA, and by normalizing k-mer counts of Xist-2kb and the lncRNA in question in context with all mouse GENCODE lncRNAs. This process was repeated for select subsets of k-mers that had the potential to increase our ability to predict repressive activity in TETRIS. Individual subsets were created by counting and normalizing k-mers as normal with SEEKR then removing columns of the resulting count matrix that were not included in a given subset. Additionally, we randomly generated 100,000 k-mer subsets each containing between 2 and 4,095 k-mers, and measured each of the subsets’ Pearson’s r values relative to our TETRIS data (Supplementary Fig. 10).

Statistical analyses. All statistics were performed in Python or R. Details of statistical analyses are described in the corresponding sections. All multiple comparison tests were adjusted using a Bonferroni correction. P values are reported as exact values except in cases where the P value was calculated using a permutation test, and no random samples were found to be more extreme than the observed value. In these cases, P values are reported as (P ≤1/n), where n is the number of permutations performed.

Data availability. The datasets generated during and/or analyzed during the current study are available within the article and its supplementary information files.

References
47. Tyner, C. et al. The UCSC Genome Browser database: 2017 update. Nucleic Acids Res. 45, D626–D634 (2017).
48. The R Core Team. R: a Language and Environment for Statistical Computing (The R Foundation for Statistical Computing, 2017).
49. Saldanha, A. J. TreeView—Extensible visualization of microarray data. Bioinformatics 20, 3246–3248 (2004).
50. Weir, W. H., Emmons, S., Gibson, R., Taylor, D. & Mucha, P. J. Post-processing partitions to identify domains of modularity optimization. Algorithms 10, 93 (2017).
51. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013).
52. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–929 (2014).
53. Pedregosa, F. et al. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011).
54. Bailey, T. L. et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 37, W202–W208 (2009).
55. Machanick, P. & Bailey, T. L. MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics 27, 1696–1697 (2011).
56. Darty, K., Denise, A. & Ponty, Y. VARNA: interactive drawing and editing of the RNA secondary structure. Bioinformatics 25, 1974–1975 (2009).
57. Busan, S. & Weeks, K. M. Visualization of RNA structure models within the Integrative Genomics Viewer. RNA 23, 1012–1018 (2017).