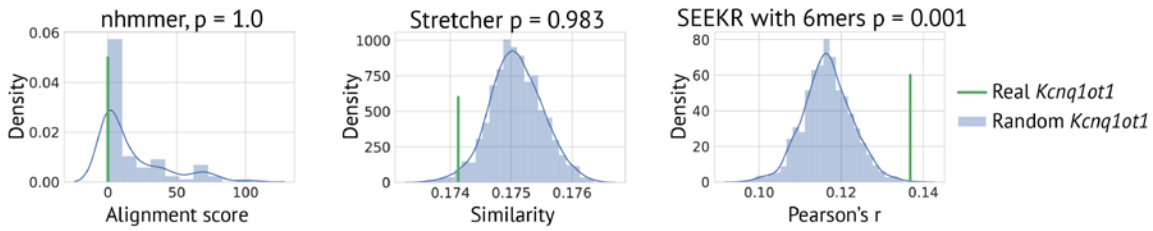
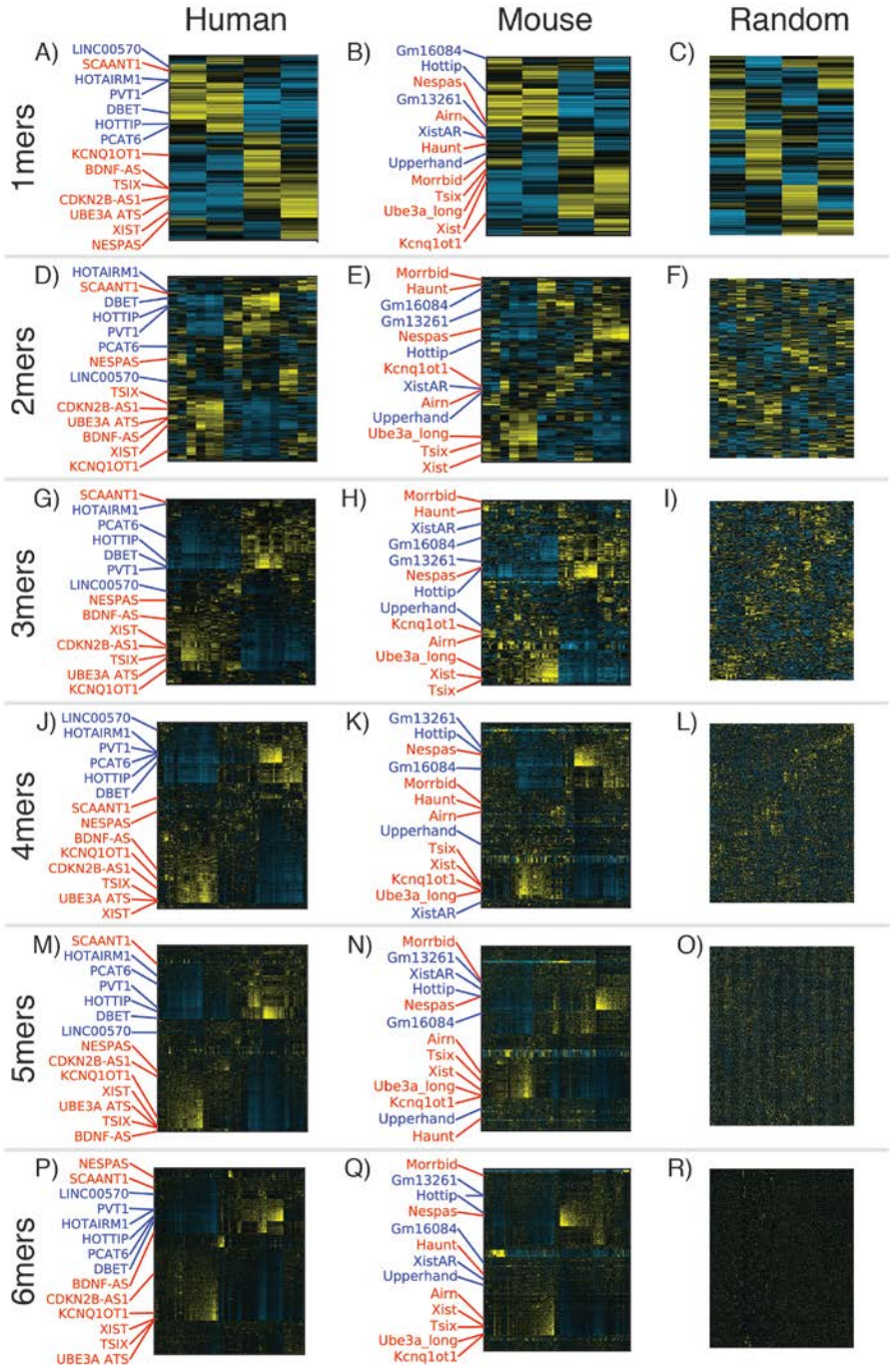
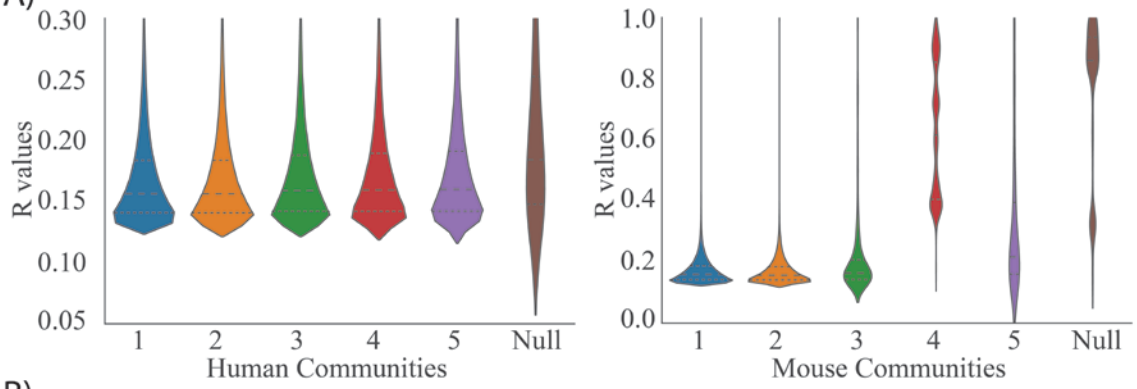


Supplementary Fig. 1. Comparison of *Xist* to *Kcnq1ot1* via nhmmer, Stretcher, and SEEKR, relative to 1,000 randomly generated lncRNAs of length/mononucleotide content identical/similar to *Kcnq1ot1*. Only SEEKR is able to detect a significant level of similarity between *Xist* and *Kcnq1ot1*.



Supplementary Fig. 2. Hierarchical clusters of human and mouse GENCODE lncRNAs, and sequences randomly generated using the nucleotide composition of the human set, at varying kmer lengths. Axes and label colors are the same as in Fig. 2. Locations of cis-repressing and cis-activating lncRNAs are marked in red and blue, respectively.

A)



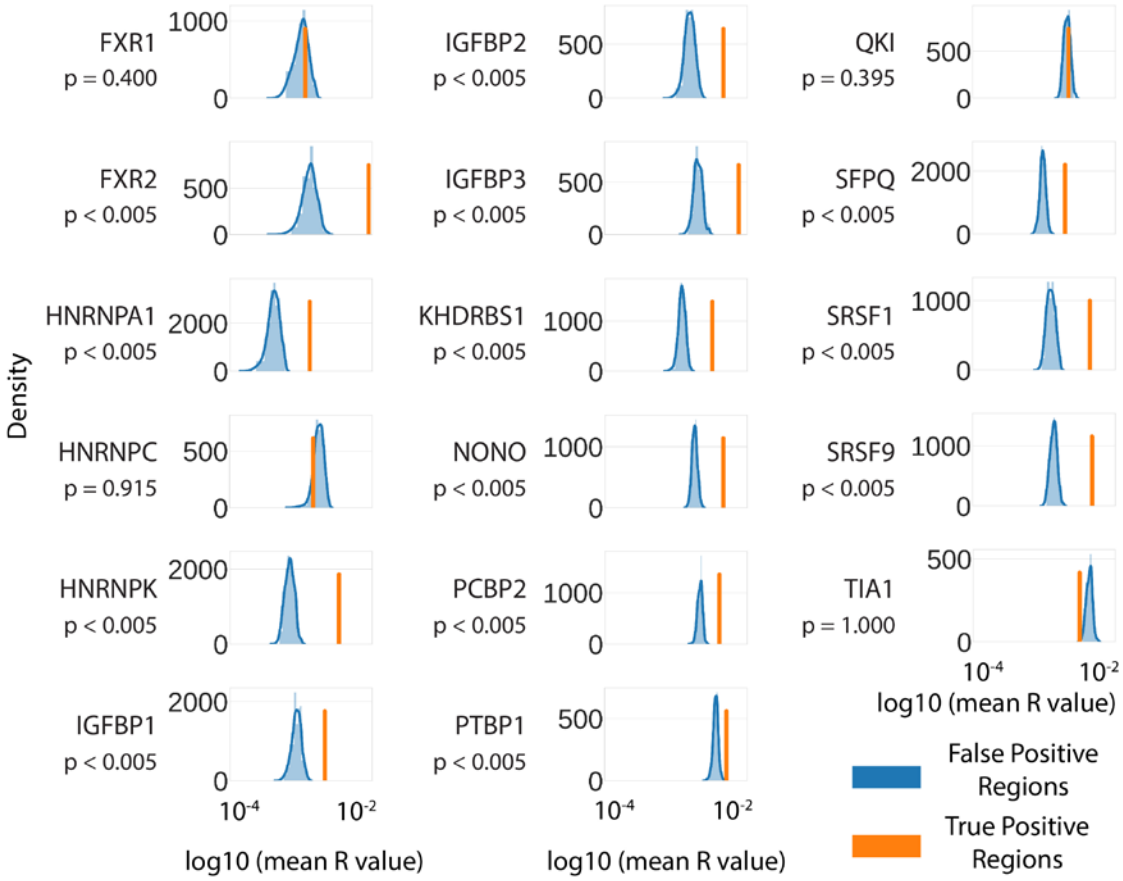
B)

Com.	Count	Mean	Std	Min	25%	50%	75%	Max
1	294126	0.169	0.041	0.130	0.141	0.156	0.184	1
2	108281	0.170	0.044	0.130	0.141	0.156	0.184	1
3	120144	0.172	0.045	0.130	0.142	0.159	0.188	1
4	36378	0.173	0.046	0.130	0.142	0.159	0.189	1
5	23268	0.175	0.053	0.130	0.142	0.160	0.191	1
Null	3841	0.251	0.190	0.130	0.148	0.184	0.252	1

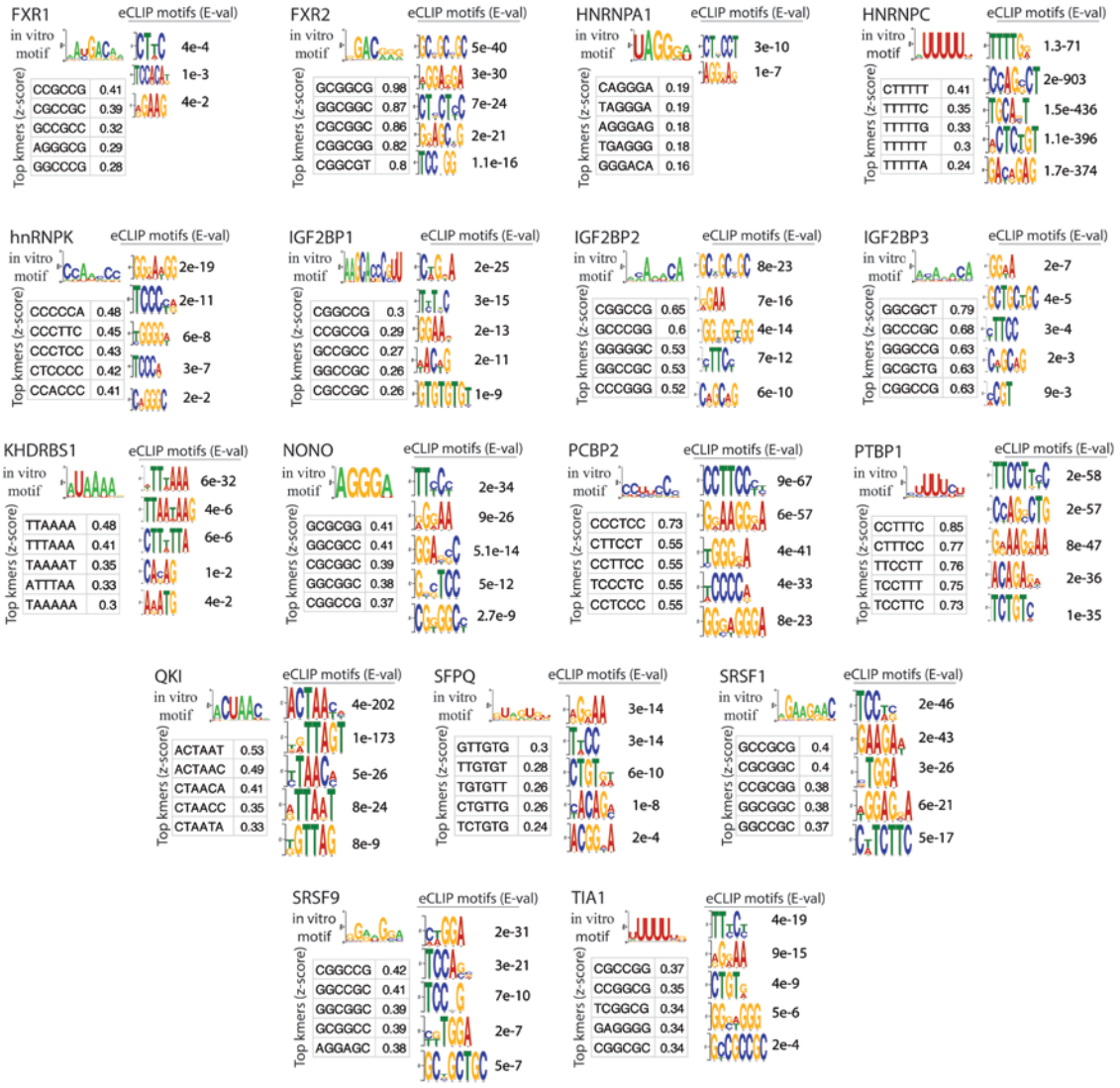
C)

Com.	Count	Mean	Std	Min	25%	50%	75%	Max
1	179939	0.171	0.041	0.130	0.142	0.160	0.188	1
2	41759	0.171	0.051	0.130	0.141	0.157	0.185	1
3	1251	0.206	0.128	0.130	0.143	0.164	0.207	1
4	52939	0.626	0.217	0.154	0.405	0.615	0.854	1
5	240	0.323	0.238	0.131	0.161	0.217	0.395	1
Null	9844	0.760	0.254	0.130	0.637	0.859	0.935	1

Supplementary Fig. 3. Relationships between lncRNAs in human and mouse communities. **(A)** Violin plots of the distribution of Pearson's r values for the similarities between lncRNAs in each community. Lines show the lower, median, and upper quartile of values (see "Count" column of tables for the sample size). **(B)** Summary statistics of Pearson's r values between human lncRNAs in each community. "Comm.", community assignment. "Count", number of edges (i.e. comparisons between pairs of lncRNAs) in community. "Mean", average Pearson's r value of edges. "Std", standard deviations. "Min", smallest Pearson's r value. "25%", Pearson's r value of the 25th percentile. "50%", Pearson's r value of the 50th percentile. "75%", Pearson's r value of the 75th percentile. "Max", largest Pearson's r value. **(C)** Summary statistics of Pearson's r values between mouse lncRNAs in each community. Column labels are the same as in (B).

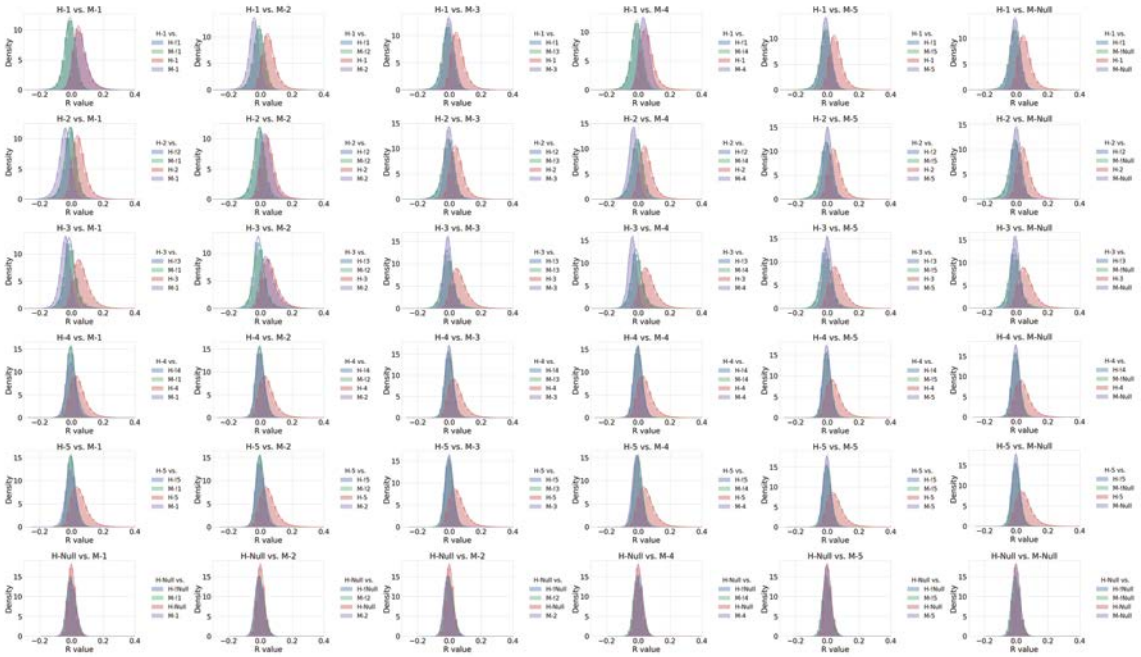


Supplementary Fig. 4. The distributions of average pairwise similarities of kmer profiles from random and size matched sets of false positive binding regions compared to the average pairwise similarity of the experimentally confirmed true positive regions for each protein (n=2000 regions, p-value determined by unadjusted permutation test). Because the average pairwise similarities of true positive regions are consistently an order of magnitude or more above those for the false positive regions, the x-axes are plotted on a log scale. For 13 of 17 proteins, the true positive regions are more similar to each other than any randomly generated set of false positive regions.

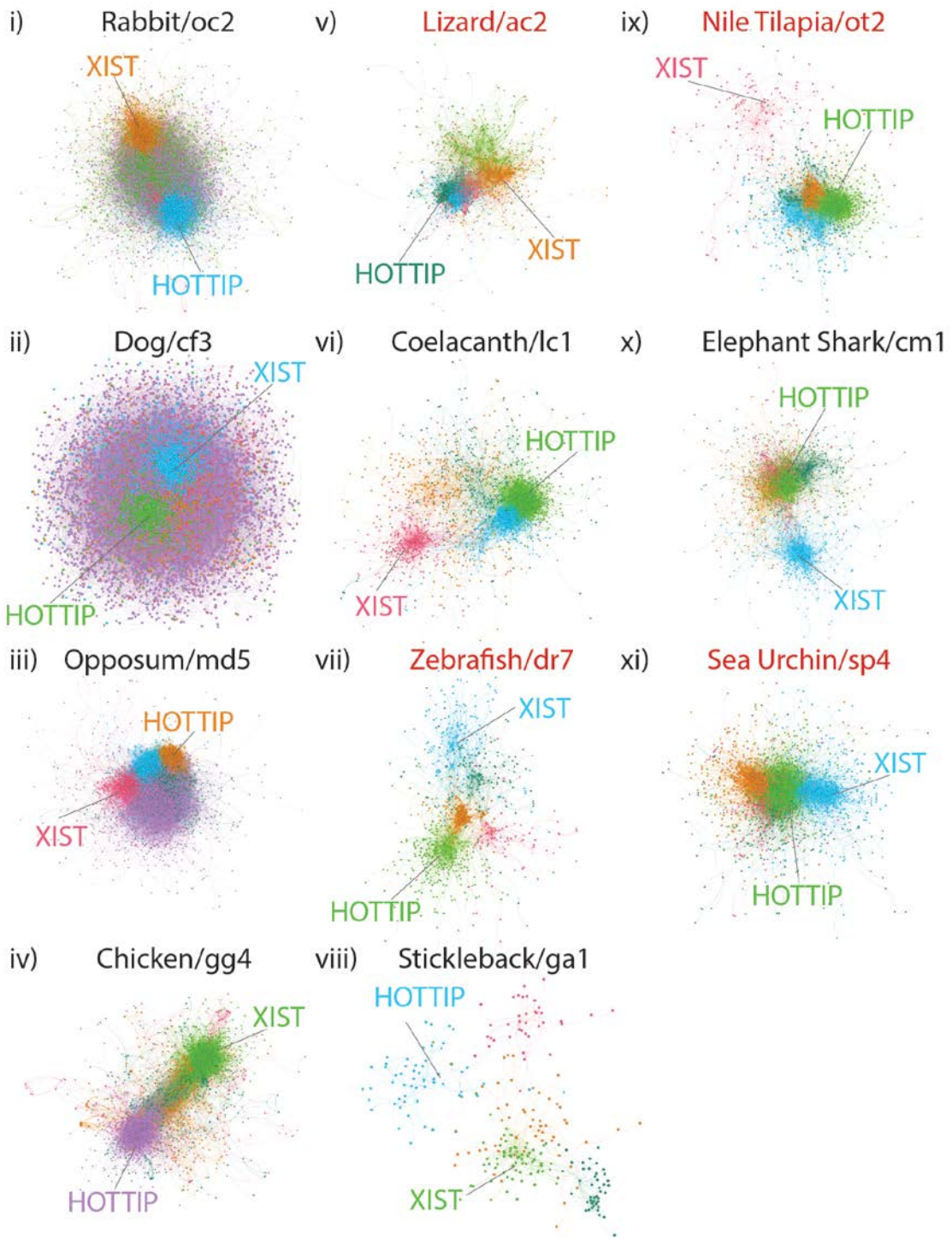


Supplementary Fig. 5. Biochemically measured PWMs (from (*1*); "*in vitro motifs*") for the 17 proteins with CLIP data in HepG2 and K562 cells (from (*2*)) are shown above the five kmers that were the most enriched in true positive regions (motif matches falling inside of CLIP peaks) relative to false positive regions (motif matches falling outside of CLIP peaks) for each protein in question. The z-scores associated with kmer enrichment in the true positive regions are also shown. Adjacent to that information are the top 5 motifs identified from eCLIP peaks using DREME (*3*). Only 3 and 2 motifs were identified

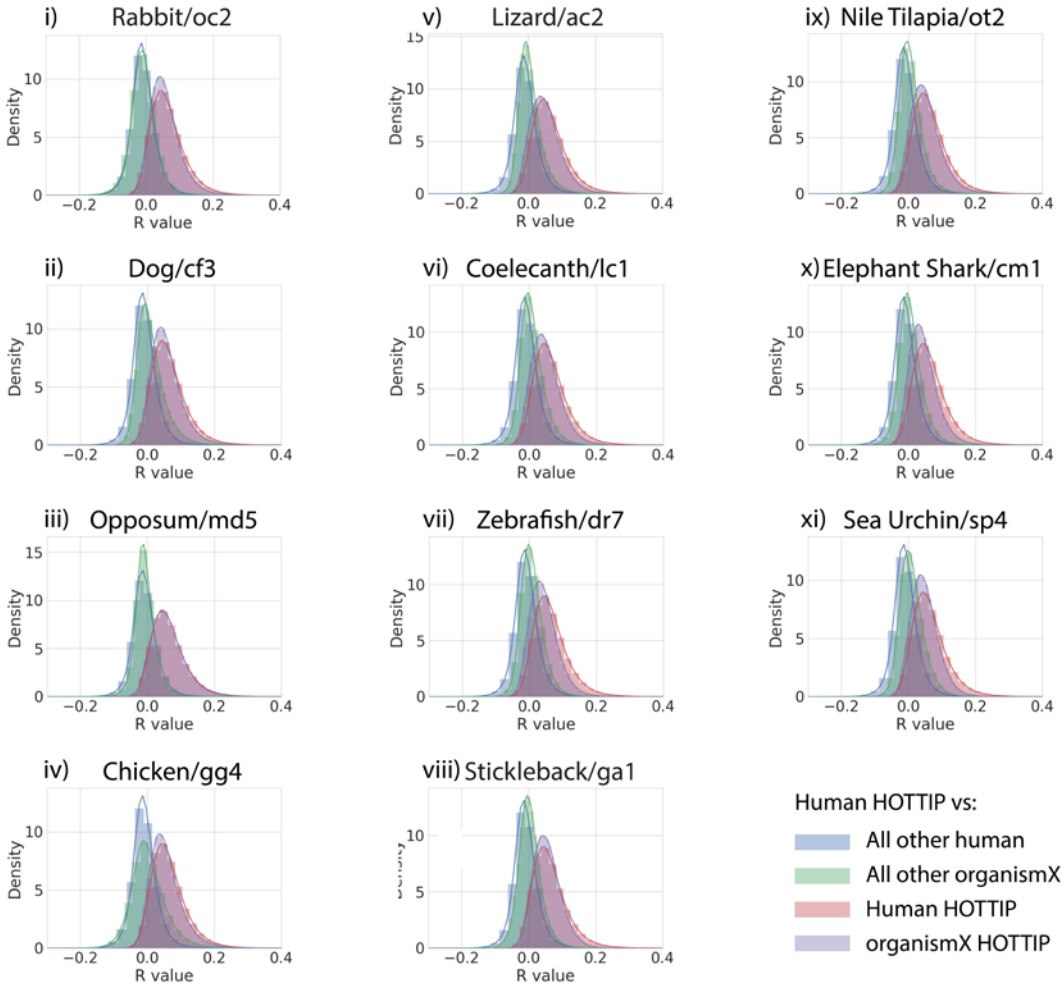
by DREME from FXR1 and HNRNPA1 eCLIP data, respectively. For HNRNPC, the first motif listed ranked outside of the top 5 (ranked 11th by E-value), but it is shown because it is the eCLIP-derived motif that best matched the *in vitro*-derived motif. For all other proteins, the eCLIP-derived motif that in our evaluation best matched the *in vitro*-derived motif fell in the top 5. By our evaluation, *in vitro*-derived motifs, top eCLIP-derived motifs, and top enriched 6mers from SEEKR showed some level of concordance for 11 of 17 proteins (FXR2, HNRNPA1, HNRNPC, HNRNPK, KHDRBS1, NONO, PCBP2, PTBP1, QKI, SFPQ, and SRSF9), and the *in vitro*-derived motifs and top eCLIP-derived motif showed concordance for an additional protein (TIA1). For the remaining five proteins, the *in vitro*-derived motifs, top eCLIP-derived motifs, and top enriched 6mers from SEEKR showed substantial differences.



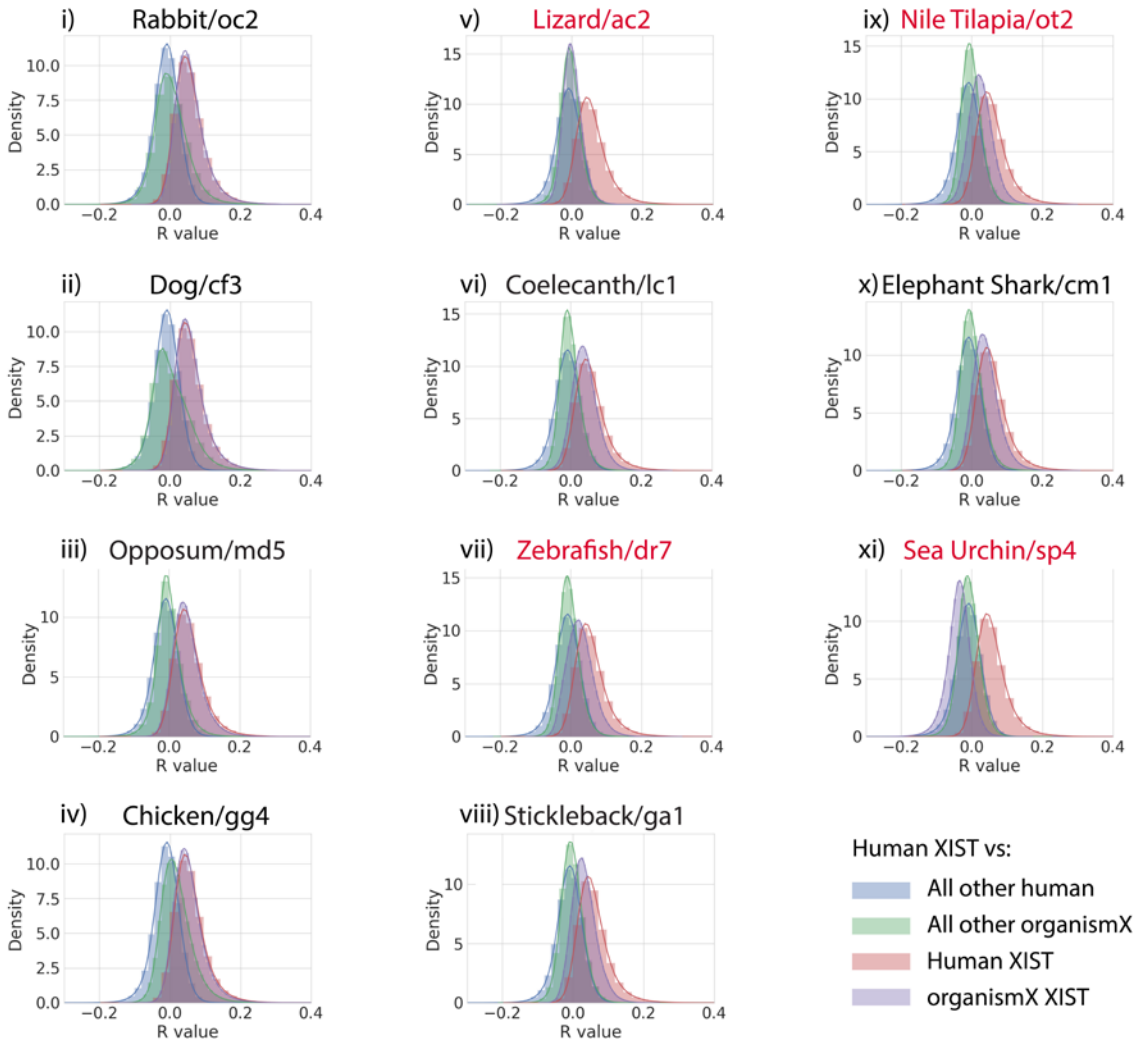
Supplementary Fig. 6. Similarity between human and mouse lncRNA communities. “H-#” refers to human lncRNAs and their corresponding community number (1, 2, etc.). “H-!#” refers all human lncRNAs excepting the community number shown. “M-#” and “M-!#”, same as for human but with mouse lncRNAs. Significant similarity was observed between communities H-1 and M-1, H-1 and M-4, H-2 and M-2, and H-3 and M-2 (note the clear overlap of red and purple histograms).



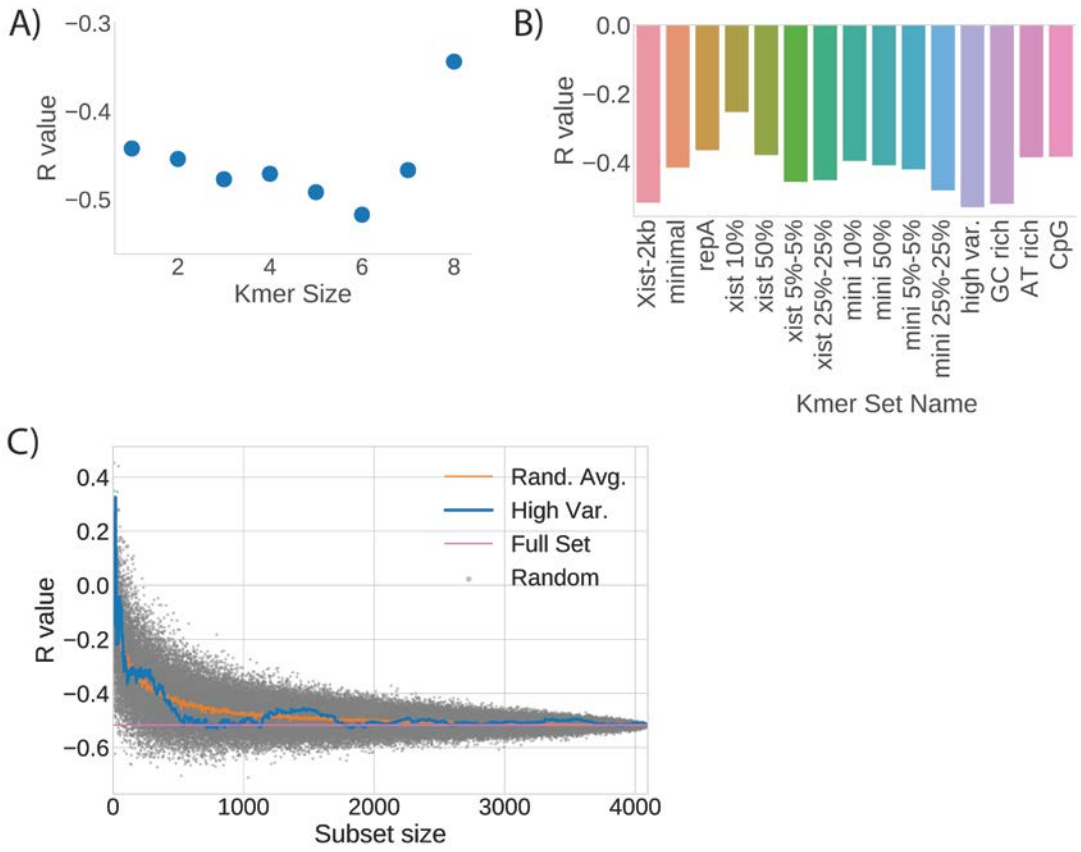
Supplementary Fig. 7. Louvain defined communities in other organisms. Human XIST and HOTTIP have been added to each set of lncRNAs.



Supplementary Fig. 8. Similarity between the human *HOTTIP* community (community #3 in Fig. 2A) and cognate lncRNA communities in other organisms. A *HOTTIP*-like community was found all organisms examined (red and purple histograms show significant overlap).

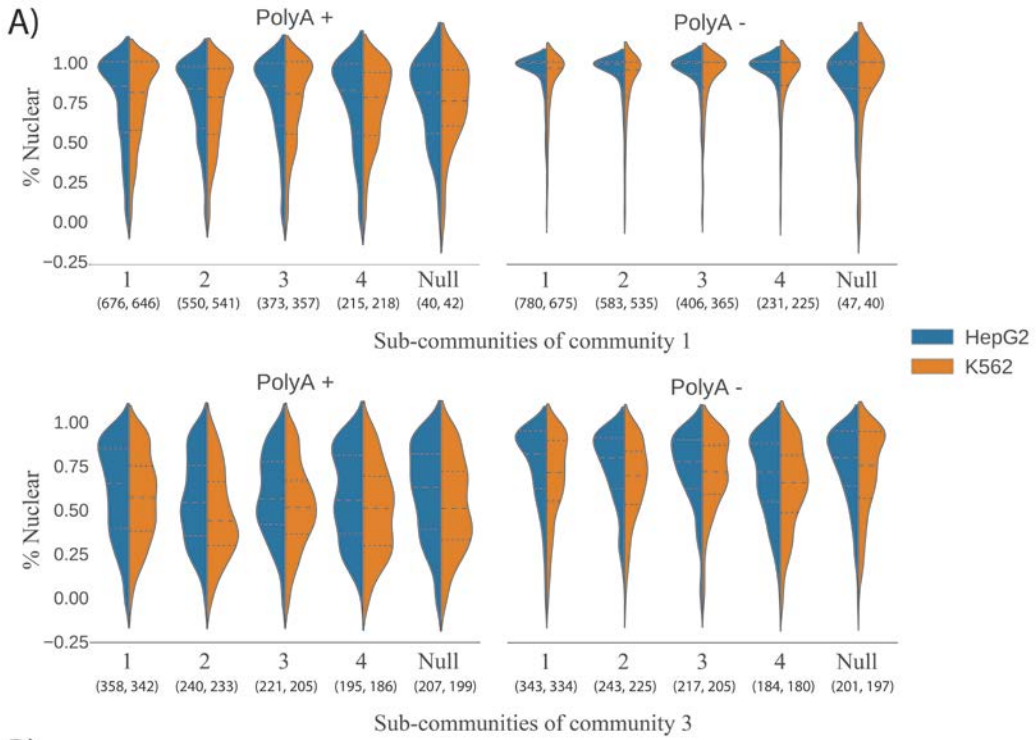


Supplementary Fig. 9. Similarity between the human *XIST* community (community #1 in Fig. 2A) and cognate lncRNA communities in other organisms. An *XIST*-like community was found in seven of the ten vertebrate species examined (names in black; red and purple histograms show significant overlap).



Supplementary Fig. 10. Additional correlations with TETRIS data show that the full set of 4096 6mers is the kmer set that is most likely to provide the greatest predictive value in SEEKR. **(A)** The correlation between *Xist*-likeness and repressive ability in the TETRIS assay is reported (y-axis) for kmer sizes 1 through 8 when running SEEKR (x-axis). 6mers provide the best correlation (-0.52). **(B)** Subsets of 6mers were selected in an attempt to improve the correlation between *Xist*-likeness and repressive ability. "*Xist*-2kb" contains the full set of 4096 6mers, which represents the Pearson's r value (-0.52) on which other 6mer sets could improve. "minimal" also uses the full set of 6mers, but measures each lncRNA inserted into TETRIS for its similarity to the minimal repressive fragment found in Fig. 5. Similarly, "repA" uses the full set of 6mers, but measures lncRNAs for their similarity to the repeat A region of Xist (Fig. 5). All 6mer sets to the right of the "repA" bar are subsets of 6mers that used the Xist-2kb transcript to calculate

correlations between lncRNAs and TETRIS data. "*xist* 10%" is the set of 410 6mers are the most overabundant in *Xist*-2kb relative to all other mouse lncRNAs. Likewise, "*xist* 50%" is the set of 2048 6mers that have the largest z-score in *Xist*-2kb. "*xist* 5%-5%" contains the 210 6mers with the largest z-scores, plus the 210 6mers with the lowest z-scores. These low abundance z-scores were added to ensure that not all 6mers in the subset were correlated with each other. "xist 25%-25%" contains the 1024 6mers with the largest z-scores, plus another 1024 6mers with the lowest z-scores. "mini 10%", "mini 50%", "mini 5%-5%", and "mini 25%-25%", are the same 6mer subsets as their "xist" counterparts, except that the 6mers are the most over- and under-represented in the minimal fragment of *Xist*, instead of the *Xist*-2kb fragment. "high var." 6mers are the 800 6mers with the highest standard deviations across the six communities defined in Fig. 3. "GC rich" and "AT rich" are 6mer subsets that contain at least four "GC" nucleotides or "AT" nucleotides, respectively. Each set contains 1408 kmers. "CpG" contains the 1185 6mers that contain a "CG" dinucleotide in their sequence. No rationally designed subsets of 6mers were significantly more predictive of lncRNA repressive activity than the baseline "Xist-2kb" fragment. **(C)** 100,000 subsets of randomly generated kmers, across the full range of subset sizes, are plotted relative to their Pearson's r values for our TETRIS data (grey circles). The average Pearson's r value at each subset size was calculated (orange line). The kmers with the largest standard deviation across lncRNA communities are also plotted for each kmer subset size (blue line). At no kmer subset size were either the average random sets or the most highly variable kmers significantly more predictive of TETRIS data than the full set of 6mers (pink line).



B)

Community	Method	Count	ANOVA p-value
1	PolyA +	3658	0.811
1	PolyA -	3887	0.07
3	PolyA +	2386	0.229
3	PolyA -	2329	0.61

Supplementary Fig. 11. Need to change HepG to HepG2 Lack of significant differences in subcellular localization in sub-communities of human community 1 and 3 lncRNAs. To determine if sub-communities of lncRNAs harbor significantly different biological properties within the five major lncRNA communities in human, lncRNAs from community #1 and #3 were extracted from the set of human GENCODE lncRNAs, and the Louvain algorithm run at default resolution parameter was used to identify the five most likely sub-communities within each. Because communities #1 and #3 were the most nuclear and cytoplasmic communities respectively, we examined if their respective sub-communities harbored significant differences in subcellular localization. **(A)** Violin plots of the distributions of cellular localization ratios for lncRNAs in communities 1 and 3. Lines show the lower, median, and upper quartile of values**.** The sample size of each

distribution is indicated below the distribution, indicating the number of lncRNAs in the distribution for HepG2 and K562, respectively. **(B)** Results of ANOVA tests examining if the nuclear distributions amongst sub-communities was different. “Community”, the original community from which sub communities were created. “Method”, the RNA-seq method used to generate the ENCODE dataset. “Count”, number of lncRNAs in each data set and the sample size used to calculate the p-value. “ANOVA p-value”, results of the ANOVA test, where p-values < .05 indicate a significant difference between distributions. Unlike that observed for the major lncRNA communities in Fig. 3A, no significant differences in subcellular localization between sub-communities were detected.

Supplementary Table 1. List of curated *cis*-regulatory lncRNAs in human and mouse. “Name”, common name of lncRNA. “Function”, indicates if lncRNA is a *cis*-activator or repressor. “H/M”, the species in which the lncRNA has been shown to regulate transcription. “References”, the PMID number for the relevant article(s).

Supplementary Table 2. Relationship between lncRNAs with known transcriptional regulatory function as measured by SEEKR. “Species”, GENCODE set of lncRNAs. “Function”, the literature reported regulatory role of the lncRNAs. “Count”, the number of lncRNAs curated from the literature with a given function for a given species (full lists in Supplemental Table 1). “Mean”, the average Pearson’s correlation of all pairwise comparisons of lncRNAs in the set. “p-value”, the results of a permutation test of 10,000 random, sized matched sets of lncRNAs. SEEKR predicts that the lncRNAs in each of these classes are significantly more similar to each other than would be expected, with the exception of the mouse *cis*-activators.

Species	Function	Count	Mean	p-value
human	cis-repression	9	0.079	<0.0001
human	cis-activation	6	0.060	0.0014
mouse	cis-repression	8	0.072	0.0011
mouse	cis-activation	5	0.011	0.1592

Supplementary Table 3. Contingency table of Louvain communities and hierarchical clusters definitions in human. Each cell represents the number of lncRNAs that are found in both the corresponding row and column labels when groups of lncRNAs are defined using either the Louvain or hierarchical method. The large values along the diagonal indicate that the group definitions are stable with respect to the particular algorithm used for detection ($p < 1E-324$; Chi-squared).

		Human Clusters					
		1	**2**	**3**	**4**	**5**	**Null**
Human Communities	**1**	2784	5	0	56	22	153
	2	8	1278	361	23	32	310
	3	8	94	1202	7	12	197
	4	84	37	30	796	17	133
	5	83	14	11	28	536	105
	Null	2164	295	243	121	133	4571

Supplementary Table 4. Contingency table of Louvain communities and hierarchical clusters definitions in mouse. Each cell represents the number of lncRNAs that are found in both the corresponding row and column labels when groups of lncRNAs are defined using either the Louvain or hierarchical method. The large values along the diagonal indicate that the group definitions are stable with respect to the particular algorithm used for detection ($p < 1E-324$; Chi-squared).

		Mouse Clusters					
		1	**2**	**3**	**4**	**5**	**Null**
Mouse Communities	**1**	1555	6	41	0	4	203
	2	5	751	17	0	7	499
	3	88	4	156	0	3	209
	4	0	0	1	326	0	0
	5	0	1	2	0	42	31
	Null	204	191	630	0	167	3102

Supplementary Table 5. Summary statistics of human lncRNA communities. “Comm.”, community assignment; number of lncRNAs in each community is in parentheses. “N”, lncRNAs not assigned to a community at the specified threshold of similarity. “Length”, average length and (standard deviation). “GC”, average GC content and (standard deviation). “CpG”, proportion of lncRNAs that overlap CpG islands. “Proteins”, proportion of lncRNAs that overlap protein-coding genes. “Exons”, average number of exons in the lncRNA and (standard deviation).

Comm.	Length	GC	CpG	Proteins	Exons
1 (3023)	**1715** (3207)	**0.37** (0.04)	**0.08**	**0.41**	**2.28** (1.88)
2 (2021)	**1629** (2245)	**0.56** (0.04)	**0.27**	**0.52**	**2.64** (2.67)
3 (1529)	**1068** (879)	**0.58** (0.06)	**0.87**	**0.61**	**2.35** (1.76)
4 (1109)	**1469** (8177)	**0.47** (0.05)	**0.18**	**0.48**	**2.70** (1.65)
5 (789)	**1316** (1670)	**0.48** (0.05)	**0.20**	**0.55**	**2.13** (1.19)
N (7545)	**755** (710)	**0.46** (0.04)	**0.15**	**0.41**	**2.79** (2.50)

Supplementary Table 6. Summary statistics of mouse lncRNA communities. “Comm.”, community assignment; number of lncRNAs in each community is in parentheses. “N”, lncRNAs not assigned to a community at the specified threshold of similarity. “Length”, average length and (standard deviation). “GC”, average GC content and (standard deviation). “CpG”, proportion of lncRNAs that overlap CpG islands. “Proteins”, proportion of lncRNAs that overlap protein-coding genes. “Exons”, average number of exons in the lncRNA and (standard deviation).

Comm.	Length	GC	CpG	Proteins	Exons
1 (1824)	**2430** (3160)	**0.39** (0.03)	**0.07**	**0.57**	**1.88** (1.64)
2 (1288)	**1610** (1282)	**0.55** (0.05)	**0.62**	**0.67**	**2.72** (3.34)
3 (463)	**1475** (1080)	**0.46** (0.04)	**0.16**	**0.51**	**2.51** (1.50)
4 (327)	**1192** (422)	**0.41** (0.01)	**0.00**	**0.01**	**3.91** (0.33)
5 (76)	**1276** (1070)	**0.49** (0.04)	**0.03**	**0.18**	**2.43** (1.76)
N (4297)	**1048** (833)	**0.47** (0.04)	**0.14**	**0.42**	**2.75** (1.68)

Supplementary Table 7. Human lncRNA community assignments and descriptions. “LncRNA”, the common identifier for each lncRNA. “Community”, the community to which SEEKR assigned the lncRNA (Fig. 2). “Top 6mers”, each cell contains a list of the top ten most overabundant 6mers, and their corresponding z-scores. “Polysomes”, the status of the lncRNA’s polysomal association. lncRNAs are assigned 1 if the lncRNA is polysome associate, 0 if they are, and left blank if the lncRNA is not expressed in K562 cells. “Proteins”, the set of proteins that are found to be overrepresented in the community to which the lncRNA is assigned. “Localization”, an indicator of the cellular localization of the transcript in K562 and HepG2 cells (Fig. 3). The lncRNA is assigned ‘C’ is it is more than 50% cytosolic on average, ‘N’ if it is more than 95% nuclear on average, ‘N & C’ if it is expressed but meets neither of the previous criteria, and left blank if it is expressed in neither K562 or HepG2 cells.

Supplementary Table 8. Mouse lncRNA community assignments and descriptions. “LncRNA”, the common identifier for each lncRNA. “Community”, the community to which SEEKR assigned the lncRNA (Fig. 2). “Top 6mers”, each cell contains a list of the top ten most overabundant 6mers, and their corresponding z-scores.

Supplementary Table 9. Contingency table comparing 5mer based human communities to 6mer based communities. Each cell represents the number of lncRNAs that are found in both the corresponding row and column labels when community detection is run using either 5mers or 6mers as a similarity measure. The large values along the diagonal indicate that the community definitions are similar to one another ($p < 1E\text{-}324$; Chi-squared).

		5mer Human Communities					
		1	**2**	**3**	**4**	**5**	**Null**
6mers Human Communities	**1**	2835	1	0	2	3	179
	2	1	1785	28	8	8	182
	3	1	52	1343	1	0	123
	4	107	81	23	491	3	392
	5	57	67	24	371	8	250
	Null	226	133	34	13	84	7037

Supplementary Table 10. Contingency table comparing 7mer based human communities to 6mer based communities. Each cell represents the number of lncRNAs that are found in both the corresponding row and column labels when community detection is run using either 7mers or 6mers as a similarity measure. The large values along the diagonal indicate that the community definitions are similar to one another ($p < 1E-324$; Chi-squared).

		7mer Human Communities					
		1	**2**	**3**	**4**	**5**	**Null**
6mers Human Communities	**1**	1629	50	5	70	68	1198
	2	3	53	988	53	47	868
	3	0	11	1026	32	27	424
	4	5	0	11	888	38	155
	5	1	2	0	5	671	98
	Null	84	255	76	80	74	6958

Supplementary Tables 11 and 12. Human and mouse community kmer profiles, comprised of the average z-score per kmer in each human and mouse lncRNA community. Columns "1", "2", "3", "4", "5", and "Null" are the mean z-score values for each kmer across all lncRNAs in the given community. "std" is the standard deviation of the kmer across the six communities.

Supplementary Table 13. Results of HSD tests between lncRNA localization of communities using polyA-selection in HepG2 cells. “Comm1” and "Comm2", community assignment for first and second set of lncRNAs compared, respectively. “n1” and “n2”, number of lncRNAs in Comm1 and Comm2, respectively. "meandiff", the mean difference in localization values between the two communities. "lower", the lower bound of the 95% confidence interval (CI) of the "meandiff" value. "upper", the upper bound of the 95% CI. "p < 0.05", result of HSD test indicating whether or not the means of "Comm1" and "Comm2" are significantly different. The test is significant if '0' is not contained within the CI.

Comm1	Comm2	n1	n2	meandiff	lower	upper	p < 0.05
1	2	1719	1397	0.0128	-0.0153	0.041	
1	3	1719	1180	-0.11	-0.1395	-0.0804	Yes
1	4	1719	775	-0.0214	-0.0552	0.0123	
1	5	1719	561	-0.0252	-0.0632	0.0127	
1	null	1719	685	-0.0476	-0.0829	-0.0124	Yes
2	3	1397	1180	-0.1228	-0.1537	-0.0919	Yes
2	4	1397	775	-0.0343	-0.0693	0.0007	
2	5	1397	561	-0.0381	-0.0771	0.0009	
2	null	1397	685	-0.0605	-0.0969	-0.0241	Yes
3	4	1180	775	0.0885	0.0524	0.1246	Yes
3	5	1180	561	0.0847	0.0447	0.1247	Yes
3	null	1180	685	0.0623	0.0248	0.0998	Yes
4	5	775	561	-0.0038	-0.0471	0.0395	
4	null	775	685	-0.0262	-0.0671	0.0147	
5	null	561	685	-0.0224	-0.0668	0.0221	

Supplementary Table 14. Results of HSD tests between lncRNA localization of communities using ribosome-depletion in HepG2 cells. “Comm1” and "Comm2", community assignment for first and second set of lncRNAs compared, respectively. “n1” and “n2”, number of lncRNAs in Comm1 and Comm2, respectively. "meandiff", the mean difference in localization values between the two communities. "lower", the lower bound of the 95% confidence interval (CI) of the "meandiff" value. "upper", the upper bound of the 95% CI. "p < 0.05", result of HSD test indicating whether or not the means of "Comm1" and "Comm2" are significantly different. The test is significant if '0' is not contained within the CI.

Comm1	Comm2	n1	n2	meandiff	lower	upper	p < 0.05
1	2	1864	1285	-0.0904	-0.1111	-0.0698	Yes
1	3	1864	1152	-0.1464	-0.1678	-0.125	Yes
1	4	1864	786	-0.0553	-0.0796	-0.031	Yes
1	5	1864	565	-0.0449	-0.0722	-0.0175	Yes
1	null	1864	740	-0.0156	-0.0404	0.0091	
2	3	1285	1152	-0.0559	-0.0791	-0.0328	Yes
2	4	1285	786	0.0351	0.0093	0.061	Yes
2	5	1285	565	0.0456	0.0168	0.0744	Yes
2	null	1285	740	0.0748	0.0485	0.1011	Yes
3	4	1152	786	0.0911	0.0647	0.1175	Yes
3	5	1152	565	0.1015	0.0722	0.1308	Yes
3	null	1152	740	0.1307	0.1039	0.1576	Yes
4	5	786	565	0.0104	-0.021	0.0419	
4	null	786	740	0.0397	0.0105	0.0689	Yes
5	null	565	740	0.0292	-0.0026	0.0611	

Supplementary Table 15. Results of HSD tests between lncRNA localization of communities using polyA-selection in K562cells. “Comm1” and "Comm2", community assignment for first and second set of lncRNAs compared, respectively. “n1” and “n2”, number of lncRNAs in Comm1 and Comm2, respectively. "meandiff", the mean difference in localization values between the two communities. "lower", the lower bound of the 95% confidence interval (CI) of the "meandiff" value. "upper", the upper bound of the 95% CI. "p < 0.05", result of HSD test indicating whether or not the means of "Comm1" and "Comm2" are significantly different. The test is significant if '0' is not contained within the CI.

Comm1	Comm2	n1	n2	meandiff	lower	upper	p < 0.05
1	2	1651	1289	-0.0344	-0.0625	-0.0062	Yes
1	3	1651	1125	-0.1571	-0.1864	-0.1278	Yes
1	4	1651	758	-0.051	-0.0842	-0.0178	Yes
1	5	1651	537	-0.0466	-0.0842	-0.0089	Yes
1	null	1651	659	-0.0423	-0.0772	-0.0074	Yes
2	3	1289	1125	-0.1227	-0.1536	-0.0918	Yes
2	4	1289	758	-0.0166	-0.0513	0.018	
2	5	1289	537	-0.0122	-0.0511	0.0267	
2	null	1289	659	-0.0079	-0.0442	0.0284	
3	4	1125	758	0.1061	0.0705	0.1417	Yes
3	5	1125	537	0.1105	0.0708	0.1502	Yes
3	null	1125	659	0.1148	0.0777	0.152	Yes
4	5	758	537	0.0044	-0.0383	0.0472	
4	null	758	659	0.0087	-0.0316	0.0491	
5	null	537	659	0.0043	-0.0397	0.0483	

Supplementary Table 16. Results of HSD tests between lncRNA localization of communities using ribosome-depletion in K562 cells. “Comm1” and "Comm2", community assignment for first and second set of lncRNAs compared, respectively. “n1” and “n2”, number of lncRNAs in Comm1 and Comm2, respectively. "meandiff", the mean difference in localization values between the two communities. "lower", the lower bound of the 95% confidence interval (CI) of the "meandiff" value. "upper", the upper bound of the 95% CI. "p < 0.05", result of HSD test indicating whether or not the means of "Comm1" and "Comm2" are significantly different. The test is significant if '0' is not contained within the CI.

Comm1	Comm2	n1	n2	meandiff	lower	upper	p < 0.05
1	2	1636	1170	-0.1335	-0.1607	-0.1063	Yes
1	3	1636	1086	-0.1345	-0.1623	-0.1067	Yes
1	4	1636	703	-0.0929	-0.1249	-0.0608	Yes
1	5	1636	510	-0.0962	-0.1323	-0.0602	Yes
1	null	1636	621	-0.0297	-0.0632	0.0038	
2	3	1170	1086	-0.001	-0.031	0.0289	
2	4	1170	703	0.0406	0.0067	0.0745	Yes
2	5	1170	510	0.0373	-0.0004	0.075	
2	null	1170	621	0.1038	0.0685	0.1391	Yes
3	4	1086	703	0.0416	0.0072	0.076	Yes
3	5	1086	510	0.0383	0.0001	0.0764	Yes
3	null	1086	621	0.1048	0.069	0.1406	Yes
4	5	703	510	-0.0033	-0.0447	0.038	
4	null	703	621	0.0632	0.024	0.1023	Yes
5	null	510	621	0.0665	0.024	0.109	Yes

Supplementary Table 17. Distributions of polysome associated lncRNAs between communities. “Community”, the name of the community. “Observed”, the number of literature reported lncRNAs associated with polysomes, in a given community. “Expected”, the number of lncRNAs that would be associated with polysomes if the lncRNAs were randomly distributed between the communities. “Ratio” Observed divided by Expected. Polysomal lncRNAs are not uniformly distributed across communities (p = 3.5e-5, Chi-squared); they are most enriched in community 3 and most depleted in community 1, providing additional support for the hypothesis that kmer content provides information about lncRNA cellular localization.

Community	Observed	Expected	Ratio
1	24	51	0.47
2	32	36	0.89
3	52	39	1.33
4	25	21	1.19
5	11	17	0.65
Null	85	65	1.31

Supplementary Table 18. Kmer abundance in nuclear and cytosolic lncRNAs. “cyto”, the mean z-score of a given kmer across the 2801 cytosolic lncRNAs. “nuc”, the mean z-score of a given kmer across the 4576 nuclear lncRNAs. “nuc – cyto”, the difference between the “nuc” and “cyto” columns. “p-value”, results of a KS-test comparing the distributions of “cyto” and “nuc” kmers. “adjusted p-value”, a Bonferroni correction of the p-value (p-value * 4096). “community”, the community in which the kmer is most overrepresented.

Supplementary Table 19. Protein log likelihood results comparing the predictive power of null versus full logistic regression models. “Protein" HUGO identifier for the protein. "Cell Type", the cell type in which the eCLIP experiment was performed. "Adjusted p-value", adjusted p-values, indicating if the full model provided significantly more information than the null model, were calculated using a likelihood ratio test followed by a Bonferroni correction (n=3747 lncRNAs for HepG2, n=3278 lncRNAs for K562). "Sig", whether or not the test came back as significant at a p-value threshold of $p<0.05$. "Communities", a list of communities in which the odds ratio of a lncRNA/protein interaction was significantly increased relative to the null community ($p<0.05$); cells are left empty if the protein is not significantly enriched in any particular community (even if the full model as a whole was found significant).

Supplementary Table 20. Protein logistic regression (LR) precision and recall results. “Protein", HUGO identifier for the protein. "Cell Type", the cell type in which the eCLIP experiment was performed. "Communities", a list of communities in which the odds ratio of a lncRNA/protein interaction was significantly increased relative to the null (at a 95% confidence level). "Precision_Null", Precision score for the LR null model. "Recall_Null", Recall score for the null model. "Precision_Full", Precision score for the full LR model. "Recall_Full", Recall score for the full model. "Precision_Diff", difference in precision score between the full and null models. "Recall_Diff", difference in recall scores between the full and null models. "PR_Diff_Avg", Mean of the "Precision_Diff" and "Recall_Diff" columns.

Supplementary Table 21. Protein binding motif counts across lncRNAs expressed in HepG2 or K562 cells. “Protein”, the name of the RNA binding protein. “True Positives”, motifs identified by FIMO that were experimentally validated by eCLIP data. “Total”, all motifs identified by FIMO. "TP%", the True Positive Rate is the number of True Positive regions divided by the Total number of regions. "(0.01)" indicates that FIMO was run at a threshold of 0.01. "(0.0001)" indicates that FIMO was run at a threshold of 0.0001. "% Diff." is the percent difference between "TP% (0.01)" and TP% (0.0001)". The True Positive percentage is low for both the 0.01 and 0.0001 threshold, and there is little to no difference between the percentages at each threshold. The 0.01 threshold was chosen for our analysis performed as part of Fig. 3D since the number of True Positive samples were multiple orders of magnitude more numerous at that threshold.

Proteins	True Positive (0.01)	Total (0.01)	TP% (0.01)	True Positive (0.0001)	Total (0.0001)	TP% (0.0001)	% Diff.
FXR1	399	23669	1.7	7	323	2.2	-0.5
FXR2	278	20539	1.4	12	484	2.5	-1.1
HNRNPA1	5486	64872	8.5	52	424	12.3	-3.8
HNRNPC	3474	38076	9.1	211	2611	8.1	1.0
hnRNPK	1204	22075	5.5	92	1355	6.8	-1.3
IGF2BP1	1311	35404	3.7	25	504	5.0	-1.3
IGF2BP2	340	22963	1.5	9	906	1.0	0.5
IGF2BP3	525	23434	2.2	20	906	2.2	0.0
KHDRBS1	1125	19565	5.8	54	895	6.0	-0.3
NONO	1046	26786	3.9	27	1065	2.5	1.4
PCBP2	1017	20455	5.0	188	3008	6.3	-1.3
PTBP1	2339	66007	3.5	102	2262	4.5	-1.0
QKI	1780	44428	4.0	63	498	12.7	-8.6
SFPQ	1119	40233	2.8	2	233	0.9	1.9
SRSF1	19781	238759	8.3	1072	12118	8.8	-0.6
SRSF9	3005	76606	3.9	131	2934	4.5	-0.5
TIA1	4008	77512	5.2	211	3877	5.4	-0.3

Supplementary Table 22. TETRIS-lncRNA fragment information. “ID”, lncRNA or lncRNA fragment inserted. “cloned fragment”, sequence cloned into the SwaI site of the TETRIS cargo. “spliced”, a column indicating whether the lncRNA inserted is spliced; where “n” indicates that the lncRNA is not known to be spliced, “y” indicates that the lncRNA’s spliced product was cloned into TETRIS, and a sequence indicates the predicted sequence of the spliced product from the genomic DNA that was cloned into TETRIS. “mm10_chr”, the chromosome from which the fragment originated relative to UCSC genome build mm10; linc00651 is a human lncRNA on hg19 chr20. “mm10_start”, “mm10_end”, the start and end coordinates of the cloned lncRNA relative to UCSC genome build mm10. “strand”, the genomic strand of the cloned lncRNA. “rel. luc.”, the average luciferase activity upon induced expression of the lncRNA relative to no dox. “SEEKR”, similarity to the *Xist-2kb* fragment at kmer length k=6. “nhmmer”, nhmmer alignment score relative to a perfect match to the *Xist-2kb* fragment (maximum of 1). “Stretcher”, proportion of nucleotides aligning to the *Xist-2kb* fragment using Stretcher. “Assays”, the total number of TETRIS assay technical replicates. “Biological Replicates”, the number of independent biological replicate derivations of TETRIS cell lines. “Length”, number of nucleotides for each transcript. “p-value”, p-value of a Student’s t-test between TETRIS values for *Xist-2kb* and TETRIS values for a given mutant, as shown in Fig. 5D. “Adjusted p-value”, a Bonferroni corrected p-value. “Sig”, True if the adjusted p-value for the t-test is less than 0.05, and False otherwise.

Supplemental Table 23. Oligonucleotide primers for the TETRIS assay. “Name”, labels for the primers. “Forward Primer”, sequence of the first of the primer pair. “Reverse Primer”,

Supplemental Files

seekr_py.zip

seekr_py.zip contains a python implementation of primary code used to create kmer count matrices. The .zip file contains five additional files:

README.md

Documentation on how to run the program.

requirements.txt

Describes other python modules needed to run SEEKR.

src/kmer_counts.py

The core script to generate a kmer count matrix. Can be used from the command-line or as a python module.

src/my_tqdm.py

Code for a progress bar.

src/fasta_reader.py

Code for reading fasta formatted files into Python.

supp_table_1_regulatory_lncRNAs.xlsx

supp_table_7_human_lncRNAs.xlsx

supp_table_8_mouse_lncRNAs.xlsx

supp_table_11_community_kmer_stats_v22.xlsx

supp_table_12_community_kmer_stats_M5.xlsx

supp_table_18_local_count_means.xlsx

supp_table_19_protein_ll.xlsx

supp_table_20_protein_LR.xlsx

supp_table_22_tetris-frag-table2.xlsx

supp_table_23_tetris_copy_num_primers.xlsx

1. D. Ray *et al.*, A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172-177 (2013).
2. E. L. Van Nostrand *et al.*, Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods* **13**, 508-514 (2016).
3. P. Machanick, T. L. Bailey, MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696-1697 (2011).