ARTICLE

Improved functions for nonlinear sequence comparison using SEEKR

SHUANG LI,[1,2,3] QUINN E. EBERHARD,[1,2,3,4] LUKE NI,[5] and J. MAURO CALABRESE[1,2,3]

[1]Department of Pharmacology, University of North Carolina, Chapel Hill, North Carolina 27599, USA
[2]RNA Discovery Center, University of North Carolina, Chapel Hill, North Carolina 27599, USA
[3]Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, North Carolina 27599, USA
[4]Curriculum in Bioinformatics and Computational Biology, University of North Carolina, Chapel Hill, North Carolina 27599, USA
[5]Chapel Hill High School, Chapel Hill, North Carolina 27516, USA

ABSTRACT

SEquence Evaluation through *k*-mer Representation (SEEKR) is a method of sequence comparison that uses sequence substrings called *k*-mers to quantify the nonlinear similarity between nucleic acid species. We describe the development of new functions within SEEKR that enable end-users to estimate *P*-values that ascribe statistical significance to SEEKR-derived similarities, as well as visualize different aspects of *k*-mer similarity. We apply the new functions to identify chromatin-enriched lncRNAs that contain *XIST*-like sequence features, and we demonstrate the utility of applying SEEKR on lncRNA fragments to identify potential RNA-protein interaction domains. We also highlight ways in which SEEKR can be applied to augment studies of lncRNA conservation, and we outline the best practice of visualizing RNA-seq read density to evaluate support for lncRNA annotations before their in-depth study in cell types of interest.

Keywords: XIST; eCLIP; *k*-mer; lncRNA; sequence comparison

INTRODUCTION

Mammalian genomes produce thousands of long noncoding RNAs (lncRNAs) (Mattick et al. 2023). While the majority remain functionally uncharacterized, lncRNAs have been shown to carry out a diversity of molecular functions, which can be specified by their sequence and structural properties, or even merely through the act of their transcription. However, relative to most protein-coding RNAs, lncRNAs are poorly conserved, evolve rapidly, and rarely harbor long stretches of linear sequence similarity. As a result, identifying the sequence elements that give rise to molecular functions in lncRNAs remains a challenge (Mattick et al. 2023).

In 2018, we reported the development of a simple algorithm to perform nonlinear sequence comparison, called SEquence Evaluation through *k*-mer Representation (SEEKR). SEEKR estimates the similarity of pairs of lncRNAs by standardizing their length-normalized *k*-mer frequencies relative to those found in a larger, user-specified set of background sequences. Using SEEKR, we showed that *k*-mer content can be used to identify similarities in protein-binding and localization between lncRNAs, and to identify lncRNA sequences that may share molecular functions. The main outputs of SEEKR are lists of standardized *k*-mer contents within the sequences being compared (i.e., their "*k*-mer profiles"), along with matrices of Pearson's r-values that quantify similarity in *k*-mer profiles between sequences. These outputs can help to identify features shared within lncRNAs of interest, even if those features are not detectable by linear alignment (Kirk et al. 2018, 2021; Sprague et al. 2019).

However, in its current form, SEEKR lacks a framework to assess the significance of *k*-mer similarity relative to a background expectation. In contrast, the most broadly used linear alignment algorithms present end-users with *P*-values that describe the significance of similarity between pairs of sequences (e.g., BLAST) (Altschul et al. 1990). These *P*-values enable rapid parsing of results and prioritization of similarities that may be biologically significant.

Herein, we describe the implementation of a *P*-value function along with other updates that improve the

Corresponding author: jmcalabr@med.unc.edu
Handling editor: Ling-Ling Chen
Article is online at http://www.rnajournal.org/cgi/doi/10.1261/rna.080188.124.

interpretability of *k*-mer similarities defined by SEEKR. As a framework to illustrate the updates, we compare the lncRNA *XIST* to the set of human lncRNAs annotated by GENCODE (Frankish et al. 2023). SEEKR updates are accessible through GitHub, the Python Package Index, and the Docker Hub.

RESULTS

New SEEKR functions including those to estimate significance of *k*-mer similarity

We developed a series of functions within the SEEKR package that enable end-users to estimate the significance of SEEKR-derived Pearson's r-values. We also created new functions to visualize different aspects of *k*-mer similarity, and deprecated other functions that were either redundant with new ones or difficult to install using current versions of Python (Fig. 1A). Updates to SEEKR are available for download through the Python Package Index, GitHub, and the Docker Hub (https://pypi.org/project/seekr/; https://github.com/CalabreseLab/seekr; https://hub.docker.com/r/calabreselab/seekr). A separate GitHub page contains the code and console commands used to generate the figures in this manuscript (https://github.com/CalabreseLab/seekr2.0_update_manuscript).

As a framework to illustrate the new SEEKR functions, we describe their use in the context of identifying similarities between the lncRNA *XIST* and other human lncRNAs. We first outline the process of estimating the significance of SEEKR-derived Pearson's r-values before highlighting other functions in SEEKR. To estimate the significance of Pearson's r-values, SEEKR relies on a user-defined set of background sequences to determine what level of similarity would be expected between any two lncRNAs selected by chance. This set of background sequences should be large enough that the total number of pairwise comparisons between sequences in the set will generate a well-populated distribution. For studies in human or mouse, lncRNA sequences curated by GENCODE are a suitable resource (Frankish et al. 2023), and SEEKR provides a function that enables download from current and past versions of the GENCODE database. For the set of background sequences used below, we used SEEKR to download and deduplicate the list of "Ensembl_canonical" GENCODE lncRNAs that are greater than 500 nt in length (henceforth referred to as GENCODE canonical lncRNAs).

After selecting a set of background sequences, the next step in estimating significance using SEEKR is to identify what probability distribution best fits the list of SEEKR-derived Pearson's r-values generated from all possible

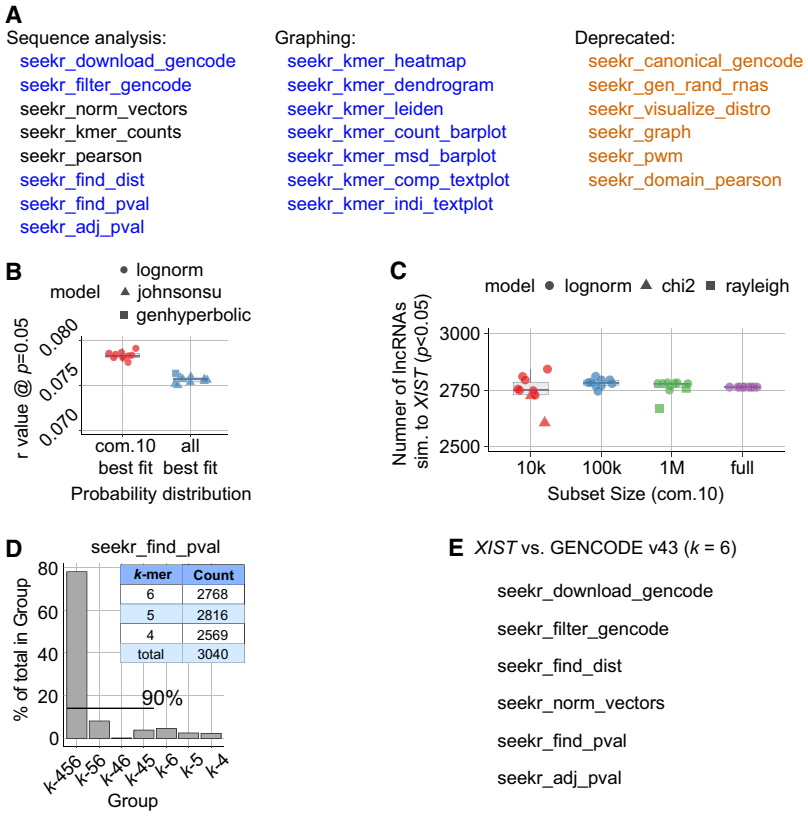

FIGURE 1. New SEEKR functions including those to estimate significance of *k*-mer similarity. (*A*) Console/command-line functions in SEEKR. Blue text, new or updated functions. Black text, retained functions. Orange text, deprecated functions. "seekr_filter_gencode" replaces "seekr_canonical_gencode." "seekr_kmer_leiden" replaces "seekr_graph." The "fasta-shuffle-letters" utility in the MEME Suite (Bailey et al. 2015) can be used in place of "seekr_gen_rand_rnas." (*B*) Documentation of the Pearson's r-value threshold at a *P*-value of $P = 0.05$ for each probability distribution that best fits the list of r-values derived from 100,000 randomly selected pairwise comparisons from the set of GENCODE v43 Ensembl_canonical lncRNAs at $k = 6$. "Com.10 best fit": pairwise comparisons fit to the common10 distributions in scipy.stats. "All best fit": pairwise comparisons fit to all 124 probability distributions in scipy.stats. The Kolmogorov–Smirnov test was used to assess goodness-of-fit to each distribution. Fitting was repeated 10× for each of the common10 and all distributions. johnsonSU, Johnson's SU distribution. genhyperbolic, generalized hyperbolic distribution. (*C*) Number of lncRNAs detected as significant at the $P < 0.05$ threshold when fitting the set of background sequences to the common10 and at the specified subsetting sizes. (*D*) Number and identity of lncRNAs detected as significantly similar to *XIST* using $k = 4$, 5, or 6 when fit to a constant distribution (log-normal). (*E*) The SEEKR console commands used to download GENCODE lncRNAs and calculate adjusted *P*-values for a sequence comparison of interest. See also https://github.com/CalabreseLab/seekr2.0_update_manuscript.

pairwise comparisons of background sequences. Parameters describing the shape of this probability distribution are then input into the cumulative distribution function to calculate the *P*-value describing the likelihood that any pair of lncRNAs is more similar to each other than would be expected by chance. The function "seekr_find_dist" is used to find the probability distribution that best fits the set of background sequences, and the function "seekr_find_pval" is used to calculate the *P*-values from this probability distribution. Lastly, if desired, the function "seekr_adj_pval" can be used to correct *P*-values for multiple-testing.

For distribution fitting, "seekr_find_dist" defaults to fit the "common10" distributions defined in Python's Fitter library, and "seekr_find_pval" defaults to selecting the best-fitting distribution from the distributions examined by the user in "seekr_find_dist." If desired, instead of fitting to the common10, end-users may fit a custom subset or all 124 of the distributions available in scipy.stats (Virtanen et al. 2020); they may also elect to select a distribution other than the best-fitting one for *P*-value calculations. For background sequences whose comparisons populate a normal-like distribution, such as the GENCODE canonical lncRNAs, we find that fitting to the common10 distributions is efficient and practical. However, if the accuracy of fitting at the threshold of significance is of the highest priority, using all available models in "seekr_find_dist" may be preferable to the common10. Using the set of GENCODE canonical lncRNAs at $k = 6$, we observed the best-fitting probability distribution from the common10 was most frequently log-normal (Fig. 1B). The best-fitting distribution from the full set of 124 in scipy.stats was most frequently Johnson's SU distribution, but still yielded Pearson's r-values at the $P = 0.05$ threshold that were on average only ∼3% different from the *P*-value threshold calculated by the log-normal (Fig. 1B). For all *P*-value calculations in the work below, we fit our background lncRNAs to the log-normal distribution.

In the case of large background sets, such as the set of GENCODE canonical lncRNAs, which contains 15,500 sequences, fitting a background distribution to all possible pairwise comparisons is computationally intensive and unnecessary for obtaining relatively accurate parameters that describe a distribution's shape (all possible pairwise comparisons from the GENCODE canonical set will yield $15{,}500^2$, or 240,250,000 data points in a distribution). In these instances, end-users may wish to fit distributions on a subset of pairwise comparisons that are randomly selected from the complete list. Tests on the set of GENCODE canonical lncRNAs indicate that subsetting on 100,000 values provides stable estimates of significance at low computational expense (Fig. 1C). We therefore performed subsetting on 100,000 values for all *P*-value calculations in our work below.

Lastly, in prior work with SEEKR, we observed that biological signals, including similarities in gene regulatory function and protein-binding, are often detected to similar extents at *k*-mer lengths of $k = 4$, 5, or 6 (Kirk et al. 2018; Sprague et al. 2019). We therefore sought to determine the extent to which significance estimates change as a function of *k*-mer length. We compared the lncRNA *XIST* to all GENCODE canonical lncRNAs at values of *k* ranging from $k = 4$–6. At a threshold of $P < 0.05$, these comparisons yielded a sum of 3040 distinct lncRNAs across the three values of *k*; 90% of the 3040 lncRNAs were detected as significantly similar to *XIST* using at least two different values of *k*, and 78% were detected as significant at all three values of *k* (Fig. 1D). Thus, when analyzing the significance of SEEKR-derived Pearson's r-values, different *k*-mer lengths detect similar sets of lncRNAs.

Putting these recommendations to use, selecting a *k*-mer length $k = 6$ and using the set of GENCODE canonical lncRNAs as both the background and the set of lncRNAs to search, we identified 2768 lncRNAs whose overall 6-mer contents are more similar to *XIST* than would be expected by chance; after correcting for multiple-testing using the Benjamini–Hochberg method, this list was reduced in size to 1265 lncRNAs (Fig. 1E; Table 1; Supplemental Table S1; https://github.com/CalabreseLab/seekr2.0_update_manuscript).

SEEKR functions to visualize features of *k*-mer similarity

To demonstrate graphing functions within SEEKR, we next investigated *LINC00632*, *DLX6-AS1*, and *PCDH10-DT*, which are highly similar to *XIST* (Table 1) and whose Ensembl_canonical transcript isoforms are strongly supported by short-read RNA-seq data (de Goede et al. 2021). *XIST* is thought to encode its repressive functions through the cumulative action of separate domains that interact with different subsets of proteins (Trotman et al. 2021). While full repression by *XIST* requires its entire sequence, some of its regions most critical for repression are comprised of tandemly repeated sequences, referred to as Repeats A, B, D, E, and F (Dixon-McDougall and Brown 2021, 2022). Therefore, to determine whether our select *XIST*-like lncRNAs harbored regional similarity to *XIST*, we separated the lncRNAs into ∼500 nt fragments and compared the fragments in each lncRNA to each fragment within *XIST*. Because data suggest Repeats A, B, D, E, and F function as discrete domains, we evaluated each *XIST* Repeat as a single intact fragment and separated the intervening *XIST* intervals into ∼500 fragments (Fig. 2A; Supplemental Files S1 and S2 contain fragments of *XIST* and all other GENCODE canonical lncRNAs used in this study, respectively). From these analyses, we found that *LINC00632*, *DLX6-AS1*, and *PCDH10-DT* each harbor significant similarity to Repeat E but lack similarity to the other Repeats in *XIST* (Fig. 2B–D). Moreover, these three lncRNAs also harbor significant similarity to fragments

TABLE 1. Top 10 most *XIST*-similar lncRNAs, as assessed by overall *k*-mer content

Transcript ID	Gene name	Transcript length	r-value	Adj. *P*-value
ENST00000626826.1	*HELLPAR*	205,012	0.476	0
ENST00000570269.2	*ENSG00000259976*	15,214	0.359	5.54×10^{-08}
ENST00000663028.1	*ENSG00000286473*	6197	0.344	2.27×10^{-07}
ENST00000623833.1	*ENSG00000279717*	2679	0.330	7.94×10^{-07}
ENST00000648200.2	*LINC00632*	21,234	0.327	8.96×10^{-07}
ENST00000562952.1	*ENSG00000261654*	5985	0.325	9.22×10^{-07}
ENST00000617901.1	*ENSG00000277151*	5305	0.324	9.22×10^{-07}
ENST00000430027.3	*DLX6-AS1*	15,364	0.321	1.08×10^{-06}
ENST00000566193.1	*ENSG00000260197*	2666	0.320	1.10×10^{-06}
ENST00000667505.1	*PCDH10-DT*	8044	0.314	1.74×10^{-06}

We evaluated the set of nonidentical "Ensembl_canonical" GENCODE v43 lncRNA transcripts that were ≥500 nt in length. Transcript length is shown in nucleotides. SEEKR-derived *P*-values were adjusted using the Benjamini–Hochberg method.

distributed across the final exon of *XIST* (Figs. 2B–D). A closer examination revealed that the similarities could be attributed to a uniform enrichment of *k*-mers rich in A and T nucleotides, whereas *k*-mers rich in G and C nucleotides were among the most variably enriched (Figs. 2E–G). We also observed that many of the 500 nt fragments within *XIST*'s final exon (i.e., Int.6 in Fig. 2A) were significantly more similar to each other than would be expected by chance (Fig 2H,I; Supplemental Fig. S1). We selected a *k*-mer length of $k = 4$ for these analyses, so as to reduce the number of *k*-mers that have zero count values in the analyzed fragments, following prior guidance described in Kirk et al. (2018) and Sprague et al. (2019). However, similar trends were detected at *k*-mer lengths $k = 5$ and 6 (Supplemental Fig. S2).

Domain-based search identifies chromatin-associated lncRNAs that harbor *XIST*-like fragments

Given that the tandem repeats within *XIST* are some of the regions most essential for its ability to induce and maintain gene silencing, we next used SEEKR to perform a parallel search for *XIST*-like lncRNAs. In this search, we separated all 15,550 GENCODE canonical lncRNAs into ~500 nt fragments (including *XIST*, as a positive control), and then used SEEKR to identify lncRNAs that contain fragments with significant *k*-mer similarity to *XIST* Repeats A, B, D, E, and F (at an unadjusted *P*-value of <0.05). We then summed the number of *XIST*-similar fragments in each lncRNA and used these sums to rank lncRNAs by their overall *XIST*-likeness. This analysis identified several intriguing lncRNAs (Table 2; Supplemental Table S2). As might have been expected, *XIST* ranked high (third). However, the known repressive lncRNA *KCNQ1OT1* ranked eighth (in a tie with five other transcripts) (Schertzer et al. 2019; Quinodoz et al. 2021). Moreover, 18 of the top 22 lncRNAs were expressed at detectable levels in K562 cells, and of those 18 lncRNAs, 17 were chromatin-enriched, including the architectural lncRNA *NEAT1* (Table 2; Dunham et al. 2012; Obuse and Hirose 2023). Thus, with this simple fragment-based search, we identified several chromatin-enriched lncRNAs that harbor domains that resemble those required for repression by *XIST*, including *KCNQ1OT1* and *NEAT1*. Additional *XIST*-like transcripts of interest include *ENST00000605862.6*, *ENST00000506640.3*, and *ENST00000622550.2*, functionally uncharacterized lncRNAs which are both spliced and expressed in multiple cell types (Dunham et al. 2012; de Goede et al. 2021). We also identified *RENO1*, a conserved lncRNA whose depletion in mouse embryonic stem cells causes changes in gene expression and a failure to differentiate properly into neurons (Hezroni et al. 2020).

The tandem repeats in *XIST* appear in a specific order and are thought to serve as recruitment centers for specific RNA-binding proteins (RBPs) (Trotman et al. 2021). With the possible exception of *KCNQ1OT1*, *XIST*-like fragments within the lncRNAs of Table 2 appeared in an order that was different from that found in *XIST* (Fig. 3). However, we hypothesized that at least some of the fragments might still associate with the same RBPs as their cognate domains in *XIST*. Therefore, in three lncRNAs of interest, we examined the eCLIP read density profiles of six RBPs that have been subject to eCLIP in K562 cells and exhibit characteristic enrichment in the tandem repeats in *XIST* (Van Nostrand et al. 2020). These RBPs included RBM15 (enriched over Repeat A); HNRNPM (enriched over Repeat F region); HNRNPK (enriched over Repeats B and D); and MATR3, PTBP1, and TIA1 (enriched over Repeat E; Fig. 3A). In all three lncRNAs, we observed *XIST*-like fragments that colocalized with the eCLIP signal from the expected RBPs. In *KCNQ1OT1*, Repeat A-, F-, and B-like fragments colocalized with RBM15, HNRNPM, and HNRNPK, respectively (Fig. 3B). In *NEAT1*, Repeat F-, B/D-, and E-like fragments colocalized

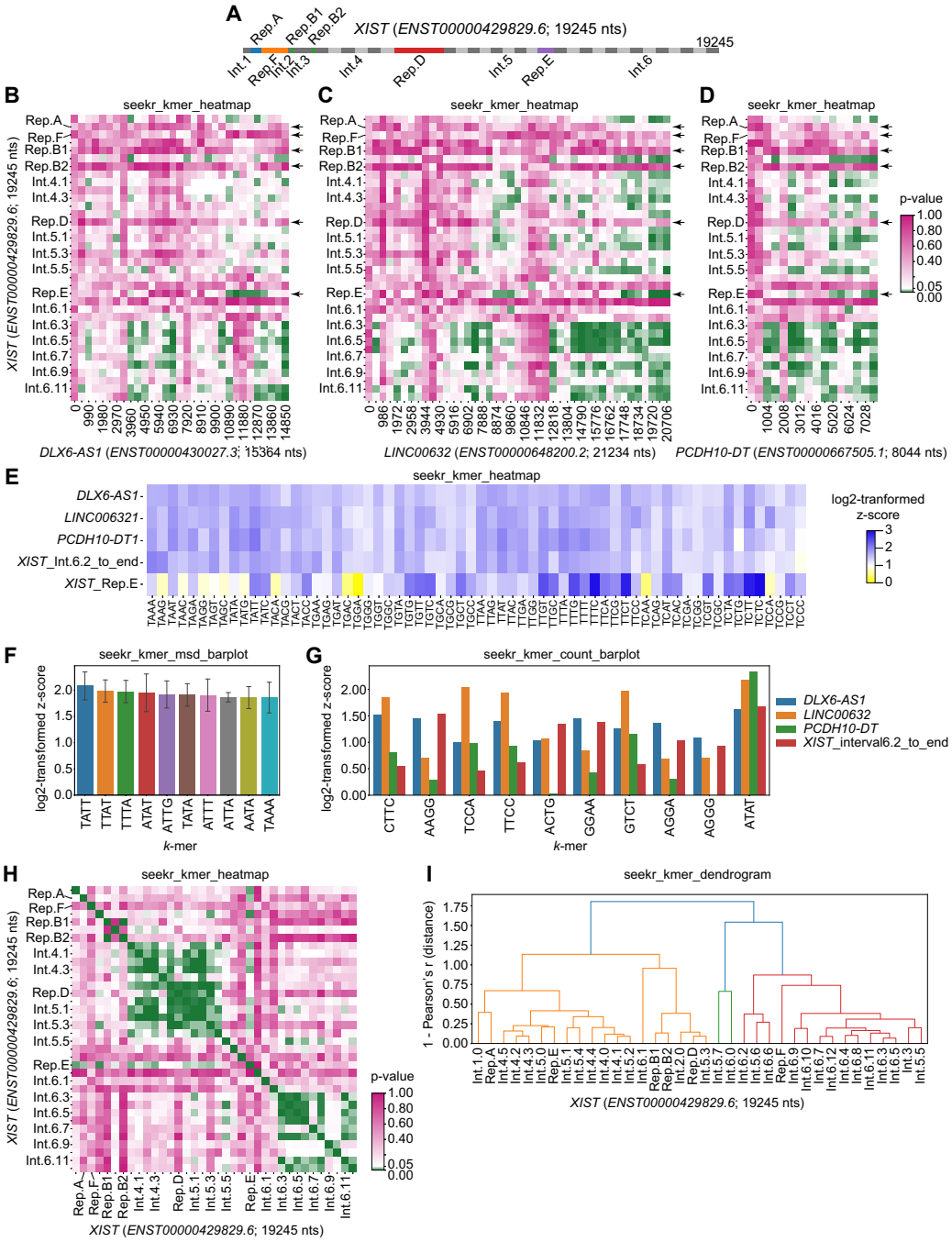

FIGURE 2. SEEKR functions to visualize features of *k*-mer similarity. (*A*) Intervals/chunks of *XIST* used to identify regional similarities. Repeats A–F in *XIST* are colored in non-gray shades and were each evaluated as single intervals, regardless of length. "Repeat F" is annotated here as a region that spans nt 743–1807 in *XIST* (*ENST00000429829.6*). (*B–D*) Regional similarities between *XIST* and *LINC00632*, *DLX6-AS1*, and *PCDH10-DT*. The latter three lncRNAs were separated into ~500 nt consecutive intervals and compared to the *XIST* intervals in (*A*). Heatmaps show *P*-values for each comparison. Comparisons were made using $k = 4$. (*E*) Heatmap displaying log2-transformed *z*-scores at $k = 4$, for the 3′ portion of *XIST* (Interval 6.2 to its transcript end), Repeat E, and *LINC00632*, *DLX6-AS1*, and *PCDH10-DT* relative to the set of nonidentical "Ensembl_canonical" GENCODE v43 lncRNA transcripts ≥500 nt in length. Log2-transformed *z*-scores were derived using seekr_kmer_counts. Only those *k*-mers beginning with "T" nucleotides are displayed for clarity. (*F*) Top 10 4-mers that are the most enriched in the 3′ portion of *XIST* and the full-length lncRNAs from (*E*) relative to the set of "Ensembl_canonical" also from (*E*). (*G*) Top 10 4-mers that exhibit the most variable enrichment in the 3′ portion of *XIST* and the full-length lncRNAs from (*E*) (i.e., highest standard deviation across the RNAs). (*H*) Regional similarities within *XIST*. Heatmap shows *P*-values for each pairwise comparison of *XIST* intervals shown in (*A*). Comparisons were made using $k = 4$. (*I*) Dendrogram showing clusters of *XIST* intervals from (*A*) that harbor related *k*-mer contents. Clusters (different colors) were defined as the nodes whose correlation distances are <0.7*[max_distance_in_dendrogram]. Comparisons were made using $k = 4$. The SEEKR console/command line functions that were used to generate graphs are displayed *above* each figure panel.

TABLE 2. Top 22 *XIST*-like lncRNAs from GENCODE v43

Transcript ID	Gene name	Transcript length	Tot_RNA	chrom_RNA	cyt_RNA	%_chrom	Rep.A	Rep.F	Rep.B1	Rep.B2	Rep.D	Rep.E	SUM
ENST00000626826.1	*HELLPAR*	205,012	0.117	1.955	0.008	99.5	16	3	0	0	2	137	158
ENST00000624628.1	*ENSG00000279072*	24,137	0.002	0.004	0.001	70.3	0	0	0	35	0	0	35
ENST00000429829.6	*XIST*	19,245	4.83	403.8	0.776	99.8	3	3	1	1	12	14	34
ENST00000501122.2	*NEAT1*	22,743	3.82	884.2	3.867	99.6	2	5	0	0	1	20	28
ENST00000623075.1	*ENSG00000280383*	23,112	0.008	0.051	0.098	33.8	3	0	3	3	0	16	25
ENST00000568752.1	*ENSG00000261200*	3971	0.000	0.245	0.005	97.5	5	0	0	3	5	5	18
ENST00000587762.2	*MIR23AHG*	19,049	0.750	12.83	1.334	90.6	2	1	5	6	0	2	16
ENST00000624209.1	*ENSG00000279036*	3578	0.000	0.000	0.000	NA	3	0	5	5	0	1	14
ENST00000609439.2	*ENSG00000291208*	10,247	0.072	3.426	0.290	92.2	1	0	1	12	0	0	14
ENST00000597346.1	*KCNQ1OT1*	91,667	0.793	41.87	0.084	99.8	1	2	2	6	1	2	14
ENST00000648200.2	*LINC00632*	21,234	0.000	0.000	0.000	NA	1	0	0	0	1	12	14
ENST00000458178.2	*PPM1F-AS1*	37,852	0.000	0.000	0.000	NA	1	0	2	8	1	2	14
ENST00000605862.6	*ENSG00000242588*	17,286	2.556	69.70	1.624	97.7	1	0	2	7	0	2	12
ENST00000604411.1	*TSIX*	37,027	0.003	0.267	0.000	99.6	1	4	1	0	0	6	12
ENST00000624919.1	*ENSG00000280434*	14,262	0.000	1.185	0.006	99.4	1	2	0	4	0	4	11
ENST00000671643.1	*ENSG00000286288*	4575	0.000	0.000	0.000	NA	2	0	4	1	2	2	11
ENST00000610177.1	*LINC02604*	3441	0.085	22.85	1.329	94.5	2	0	0	6	0	3	11
ENST00000635449.1	*PRNCR1*	12,727	0.000	0.080	0.001	97.5	0	1	0	0	0	10	11
ENST00000657488.1	*ENSG00000286379*	4878	0.032	4.110	0.095	97.7	0	0	6	2	0	2	10
ENST00000622550.2	*ENSG00000290535*	7213	0.084	4.137	0.145	96.6	1	0	2	7	0	0	10
ENST00000506640.3	*ENSG00000291215*	6469	0.248	25.48	0.825	96.9	2	0	1	7	0	0	10
ENST00000665286.1	*RENO1*	6784	0.342	5.097	0.383	93.0	4	0	4	1	1	0	10

"tot_RNA," total/ribosome-depletion RNA-seq data; "chrom_RNA," chromatin fraction RNA-seq data; "cyt_RNA," cytosolic fraction RNA-seq data; all RNA-seq data are displayed in TPM units and were collected from K562 cells as part of the ENCODE project (Dunham et al. 2012). "%_chrom," 100*[chrom_RNA_TPM]/[chrom_RNA_TPM + cyt_RNA_TPM + 0.0001]. "Rep.[A–E]," count of significant hits to each *XIST* repeat when analyzing the 500 nt chunks of each listed lncRNA. "SUM" sum of counts in "Rep.[A–E]" columns.

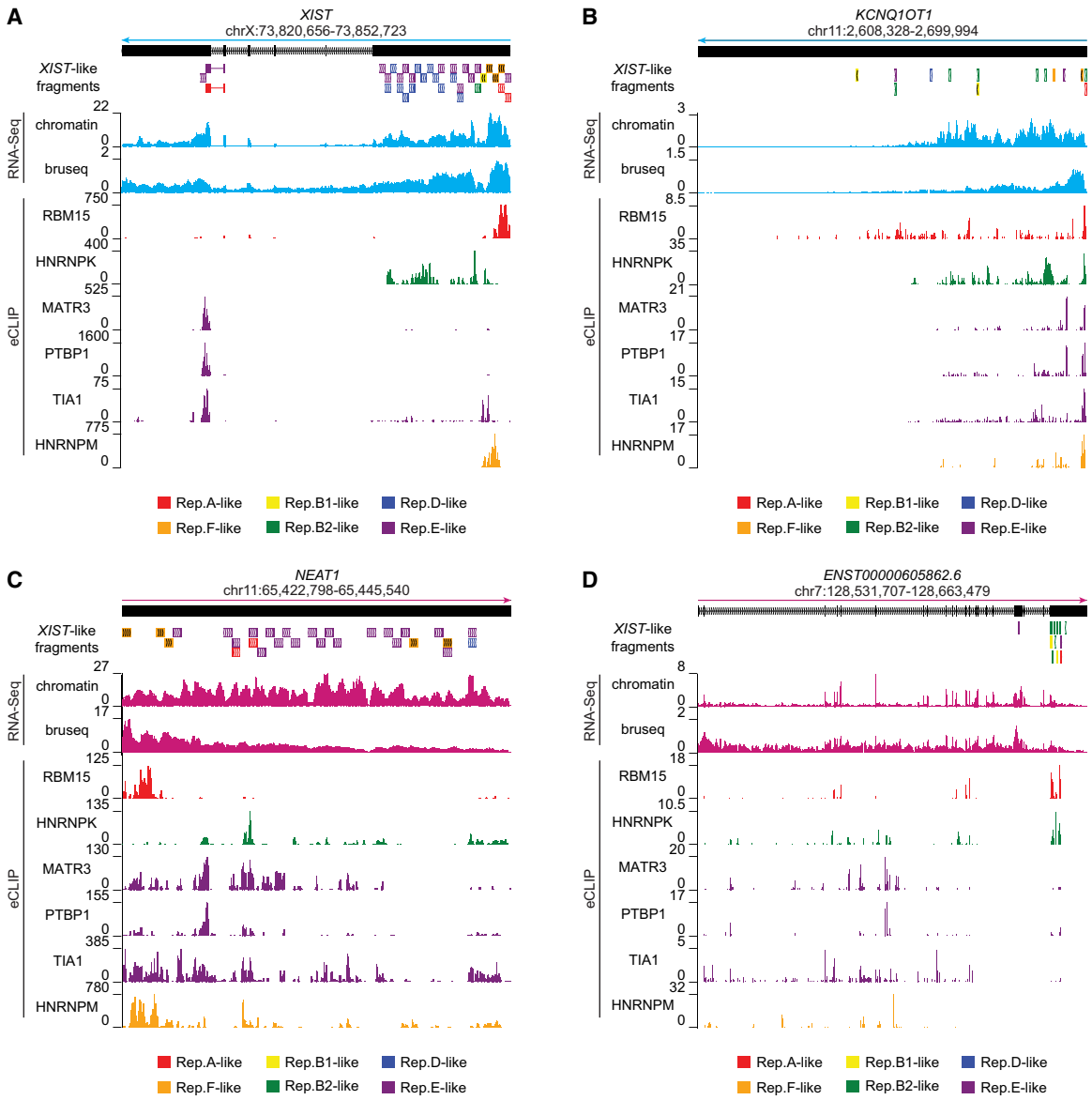

FIGURE 3. Domain-based search identifies chromatin-associated lncRNAs that harbor *XIST*-like fragments. (*A–D*) Screen images from the UCSC Genome Browser displaying gene annotations, location and identity of *XIST*-like fragments, read density from chromatin-associated RNA-seq and from Bru-seq, and background-corrected eCLIP signal for a subset of RBPs enriched over the tandem repeats of *XIST*.

with HNRNPM, HNRNPK, and MATR3/PTBP1/TIA1, respectively (Fig. 3C). In *ENST00000605862.6*, A- and B-like fragments colocalized with RBM15 and HNRNPK, respectively (Fig. 3D). Thus, while the arrangement and number of *XIST*-like fragments in each of those lncRNAs differs from *XIST*, several of their *XIST*-like fragments exhibited enriched association for the expected RBPs.

XIST-like fragments in K562-expressed lncRNAs associate with *XIST*-binding proteins

We next examined the extent to which SEEKR-defined *XIST*-like fragments colocalized with *XIST*-binding proteins in all lncRNAs expressed in K562 cells ($n = 3068$; expression defined as >0.0625 TPM from total RNA-seq data [Dunham et al. 2012]). We used the Wilcoxon signed-rank test to compare eCLIP read counts in each *XIST*-like fragment to eCLIP read counts in a paired set of fragments that were randomly shuffled among the 3068 K562-expressed lncRNAs (excluding the regions defined by SEEKR to be *XIST*-like). Strikingly, we found that the cognate binding proteins of each *XIST* Repeat were significantly enriched over *XIST*-like fragments when compared to randomly shuffled controls (Fig. 4). Moreover, for all Repeats except Repeat A, the expected cognate protein (s) was the one that was the most significantly enriched

A

XIST Rep.	# of Frags.	RBM15	HNRNPK	MATR3	PTBP1	HNRNPM
Rep.A	270	2e-5	4e-8	2e-6	5e-8	2e-8
Rep.B1	364	0.007	2e-12	1.000	0.648	1.000
Rep.B2	532	5e-12	2e-25	1.000	5e-4	0.115
Rep.D	80	1.000	5e-5	0.003	0.012	1.000
Rep.E	952	1.000	4e-6	2e-50	1e-9	2e-4
Rep.F	114	0.235	1.000	0.017	0.305	2e-7

B

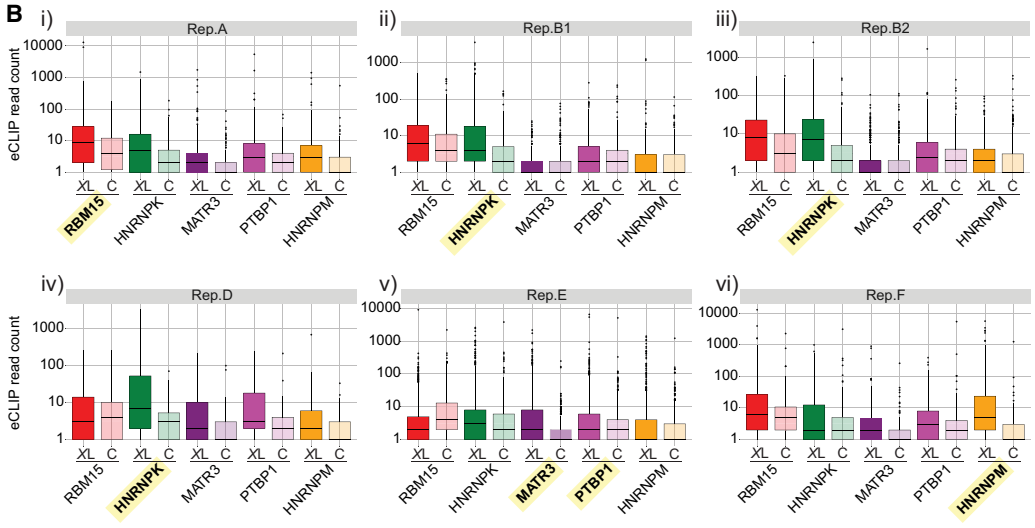

FIGURE 4. *XIST*-like fragments in K562-expressed lncRNAs associate with *XIST*-binding proteins. (*A*) Summary table displaying the number of K562 lncRNA fragments detected as significantly similar ($P < 0.05$) to each *XIST* Repeat (no. of Frags.), along with Bonferroni-corrected adjusted *P*-values from Wilcoxon signed-rank tests describing the comparison between eCLIP read density in the set of *XIST*-like fragments versus shuffled control fragments, for each of five *XIST*-binding proteins. (*B*) Box plots displaying eCLIP read density for select *XIST*-like proteins under XL, the set of detected *XIST*-like fragments, and C, the set of shuffled controls, for each of five *XIST*-binding proteins. Yellow highlights, the protein(s) that associate most robustly with each *XIST* Repeat (Van Nostrand et al. 2020).

over the set of corresponding *XIST*-like fragments. Specifically, HNRNPK was the most significantly enriched protein over Repeat B- and D-like fragments, MATR3 and PTBP1 were the most significantly enriched over Repeat E-like fragments, and HNRNPM was the most significantly enriched over Repeat F-like fragments (Fig. 4). For Repeat A-like fragments, we observed that all five proteins were significantly enriched over shuffled controls (Fig. 4); these findings can be rationalized because Repeat A is arguably the most complex of all *XIST* repeats, and it contains consensus binding motifs for each of the proteins we analyzed. Thus, fragment-based analyses with SEEKR can be used to identify domains in other lncRNAs that associate with proteins that exhibit enriched binding to a domain of interest.

Not all expressed, *XIST*-like lncRNA annotations are supported by RNA-seq data in K562 cells

In carrying out these analyses, we also noted that transcript annotations for certain highly ranked, expressed lncRNAs in Table 2 were not strongly supported by short-read RNA-seq data from K562 or HepG2 cells (Dunham et al. 2012). The most highly ranked *XIST*-similar lncRNA, *HELLPAR*, is annotated as an unspliced, 200 kb lncRNA that begins at the 3′ end of the protein-coding gene *PARPBP* and terminates near the 5′ end of the protein-coding gene *IGF1* (van Dijk et al. 2012; Frankish et al. 2021). Using standard approaches to quantify transcript abundance from RNA-seq data, *HELLPAR* registers as being expressed and chromatin-associated in K562 cells (Table 2; Supplemental Table S2). However, RNA-seq read density suggests that transcription across the *HELLPAR* locus is not due to the expression of a single lncRNA. Rather, it appears to result from a combination of imprecise termination of the upstream *PARPBP* gene, leading to read-through transcription in the 5′ half of the *HELLPAR* locus, and the transcription of a separate lncRNA, *LINC02456*, whose promoter is found in the center of the locus and whose transcribed product runs antisense through the *IGF1* protein-coding gene (Fig. 5A; Dunham et al. 2012; de Goede et al. 2021; Zhu et al. 2023).

Likewise, again visualizing short-read RNA-seq data from Dunham et al. (2012), we found examples of other *XIST*-like lncRNAs that register as being expressed and chromatin-associated in K562 cells but whose transcript annotations were not strongly supported. These included

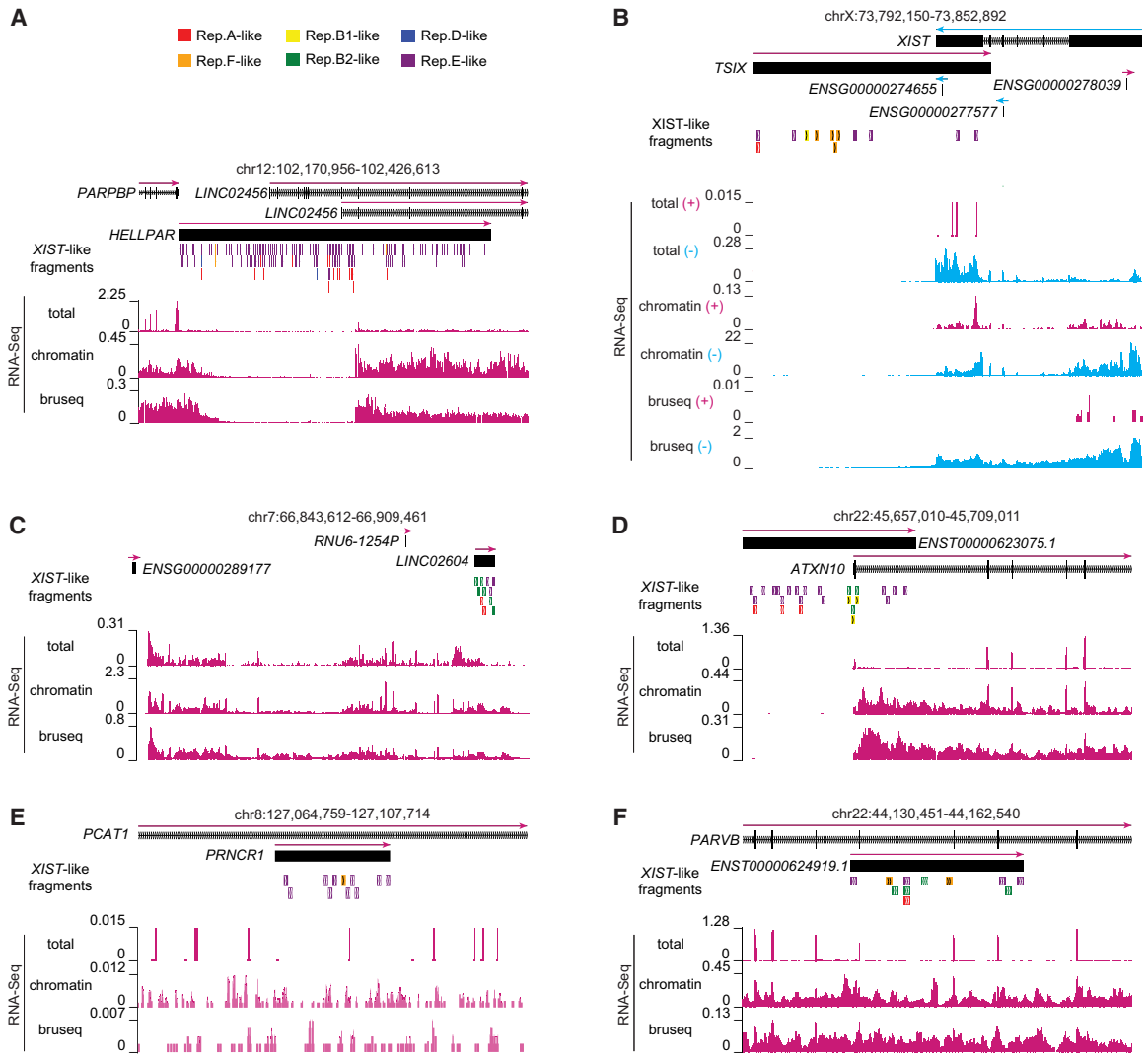

FIGURE 5. Not all expressed, *XIST*-like lncRNA annotations are supported by RNA-seq data in K562 cells. (*A–F*) Screen images from the UCSC Genome Browser displaying gene annotations, location and identity of *XIST*-like fragments, and read density from ribosome-depletion total RNA-seq (total), chromatin-associated RNA-seq (chromatin), and Bru-seq experiments (bruseq) performed in K562 cells (Dunham et al. 2012; Luo et al. 2020).

TSIX, whose only RNA-seq read density coincided with the gene structure of *XIST* on the opposite strand, and may have arisen due to imperfect strand-specificity of the dUTP RNA-seq protocol (Fig. 5B; Levin et al. 2010); *LINC02604*, which is annotated as a monoexonic lncRNA but appears to be part of a much larger transcribed and spliced region (Fig. 5C); and monoexonic lncRNAs that overlap in the sense direction relative to longer protein-coding or lncRNA genes: *ENST00000623075.1*, *PRNCR1*, and *ENST00000624919.1* (Fig. 5D–F). In each of these examples, RNA-seq abundance estimators register the lncRNAs as being expressed, underscoring the importance of examining RNA-seq read density relative to lncRNA transcript annotations before investigating SEEKR-derived sequence similarities.

Using the SEEKR *P*-value function to augment evolutionary studies of lncRNAs

We previously demonstrated that SEEKR can be applied to detect *k*-mer similarity in lncRNAs between different species (Kirk et al. 2018; Sprague et al. 2019). We reasoned that the newly developed *P*-value function in SEEKR could help to prioritize results from between-species lncRNA comparisons, just as above we demonstrated that it can help prioritize results from within-species lncRNA comparisons. Therefore, we investigated the SEEKR similarity scores of pairs of human and mouse lncRNAs that a recent study identified as potentially homologous owing to their conserved genomic locations and shared patterns of protein-binding motifs (Huang et al. 2024). Specifically, we

examined two sets of lncRNAs: a set of 5564 human lncRNAs that have predicted homologs in mouse (the "HM" set); and a set of 570 human lncRNAs that have predicted homologs in mouse and zebrafish (the "HMZ" set) (Huang et al. 2024). Comparing human and mouse lncRNA pairs against the distribution of pairwise comparisons between all human lncRNAs at *k*-mer length $k = 6$, we observed that the distributions of SEEKR similarity scores in both the HM and HMZ sets exhibited significant rightward shifts, as would be expected for conserved sequences ($P < 2\times10^{-16}$ for both comparisons; Wilcoxon rank sum test [Fig. 6; Kirk et al. 2018]). About 20% of lncRNA pairs from the HM and HMZ sets harbored SEEKR-derived 6-mer contents that were more similar to each other than would be expected by chance, using the set of all human lncRNAs as background ($P < 0.05$; 1222 of 7355 pairs in the HM set and 165 of 570 lncRNA pairs in the HMZ set; Fig. 6; orange bars). At the level of 6-mer content, most of the significantly similar pairs were less similar to each other than lncRNAs known to exhibit strong conservation, such as *H19*, *NEAT1*, *XIST*, *MALAT1*, and *KCNQ1OT1* (Fig. 6; orange bars vs. green lines). Examining further the set of 165 HMZ lncRNA pairs that harbored significant *k*-mer similarity to each other, we identified two pairs of human/mouse lncRNAs that each harbored a high number of fragments similar to *XIST*'s tandem repeat regions: *HRAT92* (human) and *6330403L08Rik* (mouse); and *SNHG14* (human) and *9330162G02Rik* (mouse). We defined a "high" number of fragments in this analysis as any lncRNA that had greater than seven *XIST*-like fragments, which corresponds to a fragment count that ranks in the 99.5% or greater when considering all GENCODE canonical human lncRNAs (Supplemental Tables S2 and S3). These analyses demonstrate ways in which SEEKR can be used to prioritize similarity between lncRNAs across species to augment other approaches to study evolutionary relationships.

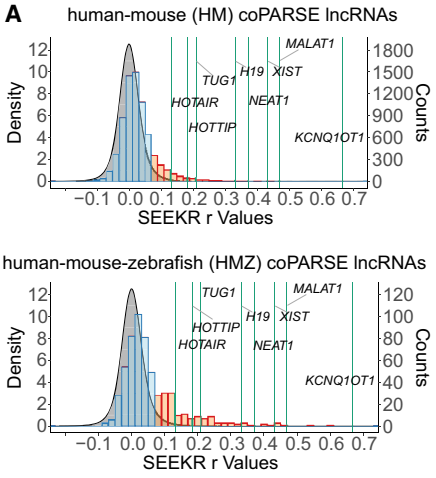

FIGURE 6. Using the SEEKR *P*-value function to augment evolutionary studies of lncRNAs. (*A*,*B*) Distributions of SEEKR-derived r-values calculated using $k = 6$ for all human–mouse lncRNA pairs defined as coPARSE in human and mouse (*A*; HM set) and human, mouse, and zebrafish (*B*; HMZ) set (Huang et al. 2024). HM and HMZ r-values are binned in discrete bars, with corresponding counts shown on the *right-hand y*-axes. Orange bars, binned counts of lncRNA pairs whose *P*-values of similarity are <0.05; blue bars, all other bins. Green lines mark SEEKR-derived r-values between exemplar pairs of conserved human and mouse lncRNAs. SEEKR-derived r-values for all pairwise comparisons of human lncRNAs are shown as continuous gray distributions whose densities are displayed on the *left-hand y*-axes.

DISCUSSION

We describe a series of updates to the SEEKR package that enable enhanced interpretation of SEEKR-derived similarity metrics. The updates were designed for application to lncRNAs but can be applied to the study of any nucleotide sequence. SEEKR is written in Python and can be installed via the Python Package Index, GitHub, or the Docker Hub (https://pypi.org/project/seekr/; https://github.com/CalabreseLab/seekr; https://hub.docker.com/r/calabreselab/seekr). The major functions of SEEKR can be implemented via the UNIX console, facilitating their use by biologists with little or no experience with Python. The SEEKR, Python, and UNIX commands used in the analyses above can be found in the GitHub page associated with this manuscript (https://github.com/CalabreseLab/seekr2.0_update_manuscript).

To illustrate the use of the new SEEKR functions, we applied them to the discovery and study of *XIST*-like lncRNAs in the human transcriptome. With a minimal set of commands and python code, we identified several lncRNAs that harbor *XIST*-like sequence features. Searches for whole-transcript similarity to *XIST* followed by fragment-based analyses highlighted three lncRNAs—*LINC00632*, *DLX6-AS1*, and *PCDH10-DT*—that contained regions similar to *XIST* Repeat E and its terminal exons, which likely serve architectural roles in *XIST* but may also enable recruitment of certain histone-modifying enzymes (Yamada et al. 2015; Sunwoo et al. 2017; Yue et al. 2017; Pandya-Jones et al. 2020; Dixon-McDougall and Brown 2022). We also observed an overall enrichment for A- and T-rich *k*-mers contained within the final long exon of *XIST*, although the biological significance of this enrichment is unclear.

A fragment-based search for regions similar to the tandem repeats in *XIST* identified a different set of *XIST*-like lncRNAs. Here, the known repressive lncRNA *KCNQ1OT1* ranked eighth among 15,550 lncRNAs queried, and nearly every lncRNA in the top 22 was enriched in the chromatin fraction of K562 cells. A significant number of *XIST*-like fragments in these and other lncRNAs colocalized with expected sets of *XIST*-associated RBPs.

Our results underscore the utility of fragment-based *k*-mer searches, particularly when the query and target

lncRNAs are substantially longer than the presumed functional modules within them (Sprague et al. 2019). For example, the tandem repeats in *XIST* are each locally enriched in different *k*-mers that presumably underlie their ability to interact with different protein cofactors. However, local *k*-mer enrichments are diluted when the ~19 kb long *XIST* transcript is analyzed as a whole, as are the local enrichments of potentially analogous domains in other lncRNAs. Instead, by using *XIST*'s tandem repeats as query features in fragment-based searches, we were able to detect regional similarities between *XIST* and other lncRNAs that were not apparent in whole-transcript similarity searches. The likelihood that these regional similarities have biological relevance is underscored by the fact that on the whole, *XIST*-like domains in other lncRNAs exhibited enriched associations with known *XIST*-binding proteins. These same searches highlight the utility of *P*-value assessments as a way to threshold SEEKR-derived r-values and focus on the most significant similarities.

Additionally, several prior studies have used different approaches to infer evolutionary relationships between pairs of lncRNAs, which remains a significant challenge in the field (Necsulea et al. 2014; Hezroni et al. 2015; Ross 2021; Huang et al. 2024). SEEKR can be applied in conjunction with any of these approaches to help contextualize as well as prioritize predicted evolutionary relationships for further study.

Lastly, as a corollary to demonstrating SEEKR's new functions, we highlight the best practice of using RNA-seq data to evaluate lncRNA annotations for support after identifying sequence similarities using SEEKR. From our analyses, we identified *XIST*-like domains in several lncRNA transcript annotations that were expressed at detectable levels in K562 cells but were not optimal representations of the lncRNAs produced from those genomic regions. In the GitHub page associated with this manuscript, we describe how to use custom scripts and standard genomic tools to convert RNA-seq alignments into wiggle tracks for display in the UCSC Genome Browser (i.e., STAR, Samtools, and BEDtools (Li et al. 2009; Quinlan and Hall 2010; Dobin et al. 2013; Raney et al. 2014; https://github.com/CalabreseLab/seekr2.0_update_manuscript). We have found that visual inspection of RNA-seq alignments in this way is a simple yet powerful approach to evaluate the support for lncRNA transcript annotations before their experimental characterization. Our analyses also highlight a need for continued efforts to improve lncRNA annotations, even within the human genome, which is the best annotated among vertebrates.

MATERIALS AND METHODS

SEEKR analyses

The Python code and console commands used for *k*-mer analyses in each figure and table can be found in https://github.com/CalabreseLab/seekr2.0_update_manuscript. SEEKR updates can be installed via the Python Package Index, via GitHub, or via the Docker Hub (instructions found on https://github.com/CalabreseLab/seekr). Supplemental Files S1 and S2 contain fragments of *XIST* and all other GENCODE canonical lncRNAs, respectively. The community graph in Supplemental Figure S1 was made using Gephi (Bastian et al. 2009).

Analysis of ENCODE RNA- and Bru-seq data

RNA-seq and Bru-seq data sets were downloaded from the ENCODE portal (https://www.encodeproject.org/) (Dunham et al. 2012; Luo et al. 2020). From K562 cells, the data sets used were: total RNA-seq (ENCSR885DVH), fractionated RNA-seq (chromatin: ENCSR000CPY; poly(A) cytosolic: ENCSR000COK), Bru-seq (ENCSR729WFH), and Bru-Chase (2 h: ENCSR633UIR; 6 h: ENCSR762OPQ). From HepG2 cells, the data sets used were: total RNA-seq (ENCSR181ZGR), Bru-seq (ENCSR974AQD), Bru-Chase (2 h: ENCSR295FEH; 6 h: ENCSR135DZR). Using the STAR aligner, a genomic index was generated using the GRCh38 primary assembly genome FASTA and the GENCODE v43 basic annotation GTF using the --sjdbGTFfile FLAG. Individual replicates from each experiment were then aligned (Dobin et al. 2013). Alignments were filtered for quality using Samtools view with the -q 30 FLAG, and the FASTQ sequences for mate pairs R1 and R2 were extracted by samtools fasta, providing the alignments as paired end inputs and the -s FLAG to filter out any singletons (Li et al. 2009). Using kallisto, FASTQ files were aligned to a set of sequences extracted from the GENCODE GRCh38 v43 basic annotation GTF combined with additional transcript sequences that contained all exons and introns from the earliest start to the latest end of each nonmonoexonic v43 gene (Bray et al. 2016; Frankish et al. 2023). Expression values for each GENCODE canonical lncRNA are reported in Supplemental Table S2.

Visualization of RNA-seq data in the UCSC Genome Browser

Using the quality-filtered STAR alignments from the K562 chromatin fraction RNA-seq (ENCSR000CPY) and K562 Bru-seq (ENCSR729WFH) data sets, replicates were merged with Samtools using the samtools merge command (Quinlan and Hall 2010). To extract negative-stranded data from the merged replicate files (both of which were reverse-stranded, paired-ended RNA-seq experiments), the samtools view -h -f 99, and samtools view -h -f 147 commands were used followed by samtools merge. Conversely, to extract positive-stranded data from the merged replicate files, the samtools view -h -f 83 and samtools view -h -f 163 were used followed by samtools merge. These filtered and merged BAM files were converted to BED12 files using BEDtools (Quinlan and Hall 2010). The number of aligned reads were counted from the original filtered and merged alignments using Samtools (Li et al. 2009). A Python script (make_wiggle_tracks_1_11_24.py) was then used to generate wiggles from BED12 files. Wiggles were converted into bigWigs using the ucsctools/320 wigToBigWig command (Kent et al. 2010). See the GitHub page for line-by-line code (https://github.com/CalabreseLab/seekr2.0_update_manuscript).

Visualization of (s)eCLIP data in the UCSC Genome Browser

BAM and BED files from (s)eCLIP data experiments (E) and their matched mock input control experiments (C) were downloaded from the ENCODE portal (https://www.encodeproject.org/) for six RBPs: RBM15 (E: ENCSR196INN, C: ENCSR454EER); MATR3 (E: ENCSR440SUX, C: ENCSR183FVK); PTBP1 (E: ENCSR981WKN, C: ENCSR445FZX); TIA1 (E: ENCSR057DWB, C: ENCSR356GCJ); HNRNPK (E: ENCSR953ZOA, C: ENCSR143CTS); and HNRNPM (E: ENCSR412NOW, C: ENCSR212ILN) (Dunham et al. 2012; Van Nostrand et al. 2016; Luo et al. 2020).

Replicates from CLIP experiments were downsampled before merging for visualization. The number of aligned reads from each replicate (GRCh38-aligned BAM files) was determined using samtools view -c (Li et al. 2009). Replicates were downsampled first by determining the minimum read count between replicates, and then the downsampling_fraction was calculated as follows: [1 – (replicate 1 read counts – replicate 2 read counts)/(minimum read count between the replicates)]. The downsampling_fraction was used in samtools view -b -s < downsampling_fraction > to reduce the size of the larger data set. Downsampled replicates were merged, filtered for MAPQ > 30, and split by strand using Samtools (Li et al. 2009). For eCLIP experiments, samtools view -b -q 30 -f 144 was used to retrieve the second mate of the negative-stranded data, while samtools view -b -q 30 -f 160 was used to retrieve the second mate of the positive-stranded data. For the seCLIP experiments, samtools view -b -q 30 -f 16 was used to retrieve the negative-stranded data, while samtools view -b -q 30 -F 16 was used to retrieve the positive-stranded data. The ENCODE peak BED files for replicates from each experiment were downloaded and sorted with BEDtools (Quinlan and Hall 2010). For each downsampled and merged eCLIP BAM file, the reads under the peaks from each replicate were extracted using bedtools intersect -split -ubam, and the resultant BAM files were merged with Samtools (Quinlan and Hall 2010). The reads under the same set of peaks were likewise extracted and merged from the corresponding mock input controls.

Wiggle tracks were then created as described in the RNA-seq section above. Specifically, peak-extracted BAM files were converted to BED12 files using BEDtools (Quinlan and Hall 2010). The number of aligned reads from the original filtered, downsampled, and merged alignments (prepeak filtering) were counted with Samtools (Li et al. 2009). A custom Python script (make_wiggle_tracks_1_11_24.py) was then used to generate wiggles from BED12 files. Signal for each eCLIP wiggle file was then normalized relative to its corresponding control by subtracting the signal in each bin of the control wiggle from the signal in the same bin of the experiment wiggle. Negative values were excluded. This process was completed with the control_normalize_wiggles_2_20_24.py script. Wiggles were then converted into bigWigs using the ucsctools/320 wigToBigWig command (Kent et al. 2010). See the GitHub page for line-by-line code (https://github.com/CalabreseLab/seekr2.0_update_manuscript).

Evaluation of (s)eCLIP density over *XIST*-like lncRNA fragments

A list of lncRNAs expressed in K562 cells (>0.0625 TPM) was compiled from Supplemental Table S2, using total RNA-seq data from https://www.encodeproject.org/ (Dunham et al. 2012; Luo et al. 2020) (ENCSR885DVH). Expressed lncRNAs were separated into ~500 nt fragments using Python code outlined in https://github.com/CalabreseLab/seekr2.0_update_manuscript. Each lncRNA fragment was searched for similarity to each *XIST* repeat region using SEEKR, and those fragments whose SEEKR-derived r-values passed a *P*-value threshold of <0.05 (unadjusted) were retained as significant. LncRNA fragment coordinates were converted into genome coordinates (hg38) by adding (or subtracting, for negative-stranded genes) the fragment start position within its host transcript to the genomic coordinate of the corresponding exon of the host transcript. Fragments overlapping splice junctions were split into one fragment per exon before converting into genomic coordinates. A corresponding set of shuffled controls for each *XIST* repeat was created by starting with the same list of lncRNAs expressed in K562 cells, excluding those lncRNAs that harbored a fragment with significant similarity to the *XIST* repeat in question, and randomly selecting one fragment from this set of lncRNA sequences that was equal in length to its corresponding *XIST*-similar fragment. (s)eCLIP reads under the *XIST*-similar and control fragments were counted using BEDtools multicov (Quinlan and Hall 2010). (s)eCLIP data sets used were RBM15 (ENCSR196INN); MATR3 (ENCSR440SUX); PTBP1 (ENCSR981WKN); TIA1 (ENCSR057DWB); HNRNPK (ENCSR953ZOA); and HNRNPM (ENCSR412NOW) (Dunham et al. 2012; Van Nostrand et al. 2016; Luo et al. 2020).

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (DBI-2228805), the National Institutes of Health (R35GM153293), and the Yang Biomedical Scholars Award to J.M.C.

Received July 11, 2024; accepted August 4, 2024.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403–410. doi:10.1016/S0022-2836(05)80360-2

Bailey TL, Johnson J, Grant CE, Noble WS. 2015. The MEME suite. *Nucleic Acids Res* **43:** W39–W49. doi:10.1093/nar/gkv416

Bastian M, Heymann S, Jacomy M. 2009. Gephi: an open source software for exploring and manipulating networks. Proceedings of the International AAAI Conference on Web and Social Media, 3(1), 361–362. https://doi.org/10.1609/icwsm.v3i1.13937

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34:** 525–527. doi:10.1038/nbt.3519

de Goede OM, Nachun DC, Ferraro NM, Gloudemans MJ, Rao AS, Smail C, Eulalio TY, Aguet F, Ng B, Xu J, et al. 2021. Population-scale tissue transcriptomics maps long non-coding RNAs to complex disease. *Cell* **184:** 2633–2648.e19. doi:10.1016/j.cell.2021.03.050

Dixon-McDougall T, Brown CJ. 2021. Independent domains for recruitment of PRC1 and PRC2 by human XIST. *PLoS Genet* **17:** e1009123. doi:10.1371/journal.pgen.1009123

Dixon-McDougall T, Brown CJ. 2022. Multiple distinct domains of human XIST are required to coordinate gene silencing and subsequent heterochromatin formation. *Epigenetics Chromatin* **15:** 6. doi:10.1186/s13072-022-00438-7

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29:** 15–21. doi:10.1093/bioinformatics/bts635

Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, et al. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74. doi:10.1038/nature11247

Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland JE, Mudge JM, Sisu C, Wright JC, Armstrong J, Barnes I, et al. 2021. Gencode 2021. *Nucleic Acids Res* **49:** D916–D923. doi:10.1093/nar/gkaa1087

Frankish A, Carbonell-Sala S, Diekhans M, Jungreis I, Loveland JE, Mudge JM, Sisu C, Wright JC, Arnan C, Barnes I, et al. 2023. GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res* **51:** D942–D949. doi:10.1093/nar/gkac1071

Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, Ulitsky I. 2015. Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep* **11:** 1110–1122. doi:10.1016/j.celrep.2015.04.023

Hezroni H, Ben-Tov Perry R, Gil N, Degani N, Ulitsky I. 2020. Regulation of neuronal commitment in mouse embryonic stem cells by the Reno1/Bahcc1 locus. *EMBO Rep* **21:** e51264. doi:10.15252/embr.202051264

Huang W, Xiong T, Zhao Y, Heng J, Han G, Wang P, Zhao Z, Shi M, Li J, Wang J, et al. 2024. Computational prediction and experimental validation identify functionally conserved lncRNAs from zebrafish to human. *Nat Genet* **56:** 124–135. doi:10.1038/s41588-023-01620-7

Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. 2010. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26:** 2204–2207. doi:10.1093/bioinformatics/btq351

Kirk JM, Kim SO, Inoue K, Smola MJ, Lee DM, Schertzer MD, Wooten JS, Baker AR, Sprague D, Collins DW, et al. 2018. Functional classification of long non-coding RNAs by k-mer content. *Nat Genet* **50:** 1474–1482. doi:10.1038/s41588-018-0207-8

Kirk JM, Sprague D, Calabrese JM. 2021. Classification of long noncoding RNAs by *k*-mer content. *Methods Mol Biol* **2254:** 41–60. doi:10.1007/978-1-0716-1158-6_4

Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A. 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* **7:** 709–715. doi:10.1038/nmeth.1491

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, Genome Project Data Processing 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25:** 2078–2079. doi:10.1093/bioinformatics/btp352

Luo Y, Hitz BC, Gabdank I, Hilton JA, Kagda MS, Lam B, Myers Z, Sud P, Jou J, Lin K, et al. 2020. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res* **48:** D882–D889. doi:10.1093/nar/gkz1062

Mattick JS, Amaral PP, Carninci P, Carpenter S, Chang HY, Chen LL, Chen R, Dean C, Dinger ME, Fitzgerald KA, et al. 2023. Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nat Rev Mol Cell Biol* **24:** 430–447. doi:10.1038/s41580-022-00566-8

Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grutzner F, Kaessmann H. 2014. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505:** 635–640. doi:10.1038/nature12943

Obuse C, Hirose T. 2023. Functional domains of nuclear long noncoding RNAs: insights into gene regulation and intracellular architecture. *Curr Opin Cell Biol* **85:** 102250. doi:10.1016/j.ceb.2023.102250

Pandya-Jones A, Markaki Y, Serizay J, Chitiashvili T, Mancia Leon WR, Damianov A, Chronis C, Papp B, Chen CK, McKee R, et al. 2020. A protein assembly mediates XIST localization and gene silencing. *Nature* **587:** 145–151. doi:10.1038/s41586-020-2703-0

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842. doi:10.1093/bioinformatics/btq033

Quinodoz SA, Jachowicz JW, Bhat P, Ollikainen N, Banerjee AK, Goronzy IN, Blanco MR, Chovanec P, Chow A, Markaki Y, et al. 2021. RNA promotes the formation of spatial compartments in the nucleus. *Cell* **184:** 5775–5790.e5730. doi:10.1016/j.cell.2021.10.014

Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, Nguyen N, Paten B, Zweig AS, Karolchik D, et al. 2014. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics* **30:** 1003–1005. doi:10.1093/bioinformatics/btt637

Ross CJ, Rom A, Spinrad A, Gelbard-Solodkin D, Degani N, Ulitsky I. 2021. Uncovering deeply conserved motif combinations in rapidly evolving noncoding sequences. *Genome Biol* **22:** 29. doi:10.1186/s13059-020-02247-1

Schertzer MD, Braceros KCA, Starmer J, Cherney RE, Lee DM, Salazar G, Justice M, Bischoff SR, Cowley DO, Ariel P, et al. 2019. lncRNA-induced spread of polycomb controlled by genome architecture, RNA abundance, and CpG Island DNA. *Mol Cell* **75:** 523–537.e510. doi:10.1016/j.molcel.2019.05.028

Sprague D, Waters SA, Kirk JM, Wang JR, Samollow PB, Waters PD, Calabrese JM. 2019. Nonlinear sequence similarity between the XIST and RSX long noncoding RNAs suggests shared functions of tandem repeat domains. *RNA* **25:** 1004–1019. doi:10.1261/rna.069815.118

Sunwoo H, Colognori D, Froberg JE, Jeon Y, Lee JT. 2017. Repeat E anchors XIST RNA to the inactive X chromosomal compartment through CDKN1A-interacting protein (CIZ1). *Proc Natl Acad Sci* **114:** 10654–10659. doi:10.1073/pnas.1711206114

Trotman JB, Braceros KCA, Cherney RE, Murvin MM, Calabrese JM. 2021. The control of polycomb repressive complexes by long noncoding RNAs. *Wiley Interdiscip Rev RNA* **e1657**. doi:10.1002/wrna.1657

van Dijk M, Thulluru HK, Mulders J, Michel OJ, Poutsma A, Windhorst S, Kleiverda G, Sie D, Lachmeijer AM, Oudejans CB. 2012. HELLP babies link a novel lincRNA to the trophoblast cell cycle. *J Clin Invest* **122:** 4003–4011. doi:10.1172/JCI65171

Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, Blue SM, Nguyen TB, Surka C, Elkins K, et al. 2016. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods* **13:** 508–514. doi:10.1038/nmeth.3810

Van Nostrand EL, Freese P, Pratt GA, Wang X, Wei X, Xiao R, Blue SM, Chen JY, Cody NAL, Dominguez D, et al. 2020. A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583:** 711–719. doi:10.1038/s41586-020-2077-3

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* **17:** 261–272. doi:10.1038/s41592-019-0686-2

Yamada N, Hasegawa Y, Yue M, Hamada T, Nakagawa S, Ogawa Y. 2015. XIST Exon 7 contributes to the stable localization of XIST RNA on the inactive X-chromosome. *PLoS Genet* **11:** e1005430. doi:10.1371/journal.pgen.1005430

Yue M, Ogawa A, Yamada N, Charles Richard JL, Barski A, Ogawa Y. 2017. XIST RNA repeat E is essential for ASH2L recruitment to the inactive X and regulates histone modifications and escape gene expression. *PLoS Genet* **13:** e1006890. doi:10.1371/journal.pgen.1006890

Zhu X, Du M, Gu H, Wu R, Gao M, Xu H, Tang J, Li M, Liu X, Zhong X. 2023. Integrated analysis of lncRNA and mRNA expression profiles in patients with unexplained recurrent spontaneous abortion. *Am J Reprod Immunol* **89:** e13691. doi:10.1111/aji.13691

MEET THE FIRST AUTHOR

Shuang Li

***Meet the First Author* is an editorial feature within *RNA*, in which the first author(s) of research-based papers in each issue have the opportunity to introduce themselves and their work to readers of *RNA* and the RNA research community. Shuang Li is the first author of this paper, "Improved functions for nonlinear sequence comparison using SEEKR." Shuang is a research associate in the lab of J. Mauro Calabrese, in the Department of Pharmacology at the UNC Chapel Hill School of Medicine. Shuang's research focuses on the development and application of statistics and visualization tools of the SEEKR package, which quantifies *k*-mers-based nonlinear similarity.**

What are the major results described in your paper and how do they impact this branch of the field?

The statistical and visual updates to SEEKR enable easier and clearer ways to quantify and locate nonlinear sequence similarities that bear biological relevance in long noncoding RNAs, which could also be applied across species to query evolutionary conservations. With three ways to implement it, the new SEEKR package accommodates users with little to no coding experience as well as professional bioinformaticians.

What led you to study RNA or this aspect of RNA science?

Long noncoding RNAs (lncRNAs) carry out various essential molecular functions through their sequence features, structural properties, or even merely the act of their transcription. Because of the rapidly evolving nature of lncRNAs, it is challenging to connect sequence features with biological functions. As a scientist fascinated with complicated data analysis, I was drawn to the challenge of putting in place this one piece of the puzzle.

During the course of these experiments, were there any surprising results or particular difficulties that altered your thinking and subsequent focus?

When we went through highly ranked candidate lncRNAs that bear *XIST*-like domains, we came across multiple transcripts that passed our expression threshold in K562 cells, but their annotations were not supported by short-read RNA-seq data, which was obvious when illustrated by wiggle tracks. As the human genome is the best annotated among vertebrates, the amount of unsupported lncRNAs we encountered during the analysis surprised us. Quinn Eberhard, who is among the authors of this paper, is taking the leadership to tackle this problem. Please stay tuned for her updates.

What are your subsequent near- or long-term career plans?

For the next phase, I want to explore and investigate how deep learning can be applied to our biological questions. We have several well-built models, such as DNABERT-2 and AlphaFold3, which integrate enormous sequence data with physical, chemical, and structural properties. These models could then serve as a basis for developing new models that are easier to train and more focused to a particular question. My goal would be to accomplish a small machine learning project to improve the SEEKR package.

Improved functions for nonlinear sequence comparison using SEEKR

Shuang Li, Quinn E. Eberhard, Luke Ni, et al.

RNA 2024 30: 1408-1421 originally published online August 26, 2024
Access the most recent version at doi:10.1261/rna.080188.124