**Predicting noncoding RNA function in cancer**

Nearly 60,000 regions in the human genome produce RNAs that have little to no potential to encode for protein, and one seventh of these lncRNAs have altered expression in cancer. Early works suggest, as a class, lncRNAs play significant roles in many aspects of the disease, largely through the regulation of transcription. However, the overwhelming majority of lncRNAs remain unstudied, and their biological functions and potential roles in cancer remain unknown. We developed a computational method to predict lncRNA function – to our knowledge the first of its kind – that suggests one quarter of all lncRNAs may regulate transcription in their local genomic environment via its activation or its repression, respectively. As proof of principle, we have validated our method by showing it can be used to classify lncRNAs with known function, as well as to design synthetic lncRNAs with pre-programmed function. Remarkably, many of the lncRNAs that we predict have local regulatory function are encoded in the genome in close proximity to some of the most recurrently mutated, overexpressed, or silenced genes in cancer. This discovery has led to our hypothesis that lncRNAs play a major role in maintaining the homeostatic levels of many oncogenes and tumor suppressors across tumor types. If our early predictions are correct, they could transform the study of lncRNAs in cancer by suggesting direct paths to understand functions and mechanisms for thousands of unstudied lncRNAs whose expression strongly correlates with different subtypes of the disease. The focus of my laboratory and this Kimmel proposal is to validate our predictions, to delineate mechanisms by which lncRNA-mediated transcriptional control occurs, and to determine how the process might be targeted to selectively kill cancer cells. The pathological regulation of transcription plays a central role in the development and maintenance of most cancers; thus, our work has potential to aid in improved strategies for treatment, detection, and diagnosis of cancers that span spectrums of the disease.

The human genome is pervasively transcribed, and many of the RNAs produced have little or no potential to encode for protein (1, 2). These so-called long noncoding RNAs (lncRNAs) remain enigmatic entities in cancer. A handful have been shown to affect diverse biological processes relevant to the development and maintenance of cancer, including cell division, apoptosis, and DNA damage signaling, yet the molecular mechanisms through which they function in these processes are not well understood (3-5). Upwards of 8,000 lncRNAs are differentially expressed across the cancer spectrum (2), and many are found in close proximity to some of cancer’s most frequently mutated, overexpressed, or silenced protein-coding genes, including *AKT*, *EGFR*, *KRAS*, *MYB*, *MYC*, *PTEN*, and *RB1*. However, the function and molecular mechanisms of nearly all of these lncRNAs are unknown, and, equally important, are unpredictable by existing methods.

Nevertheless, studies in organisms that span from yeast to man highlight a recurrent theme in lncRNA biology: lncRNA transcription often induces the local activation or repression of protein-coding gene expression. In the most striking example, expression of a conserved lncRNA called XIST leads to transcriptional silencing of nearly all genes along the entire inactive X chromosome (6). More recent works have shown similar transcription-regulating lncRNAs act in different ways to promote cancer, including the lncRNAs *PVT1*, required for sustained expression of MYC from focal amplifications (7), *SChLAP1*, which antagonizes BRG1-dependent transcription in metastatic prostate cancer (8), and *HOTAIR*, thought to alter transcription through its influence on Polycomb activity in metastatic breast, prostate, and lung cancers (9-11). It remains an open question as to which other, if any, of the thousands of lncRNAs expressed in humans have similar impacts on tumorigenesis. Considering the known ability of RNA to control local transcription across kingdoms of life (6, 12-16), and the genomic proximity of many lncRNAs to oncogenes and tumor suppressors, ***we hypothesize that a significant subset of lncRNAs function to regulate transcription in regions of the human genome that are critical for the development of cancer.*** Such lncRNAs may serve as molecular rheostats that control gene expression states essential for development or maintenance of cancer. As most lncRNAs are expressed with high levels of tissue specificity (1, 2, 17), ***devising systematic approaches to understand the biology of lncRNAs and their mechanisms of action could suggest new avenues to alter transcription in a way that is selectively toxic to tumors*.** LncRNA presence can also be measured in blood and urine (18-20), and so a deeper understanding of lncRNA function should also aid in the molecular classification of cancers and may facilitate development of biomarkers that can be measured by non-invasive methods. With these goals in mind, our research aims to address two basic questions. First, which genes central to the development and maintenance of cancer are subject to lncRNA-mediated control in select cell lines from the Cancer Cell Line Encyclopedia (21)? Second, what are the mechanisms by which this regulation is achieved? Our studies lay the groundwork for future, more detailed investigations into how lncRNA-mediated regulatory events might be leveraged to selectively kill cancer cells.

**BACKGROUND**

*PARK – A computational tool to predict lncRNA function in cancer*: The cellular functions of lncRNAs are currently impossible to predict through computational analysis, preventing a systematic understanding of their roles in cancer.We therefore set out to develop a computational tool capable of predicting lncRNA function based on the following assumptions. First, lncRNA structure may be more dynamic than protein structure, and functional elements within lncRNAs may not be constrained by linear relationships. Second, with few exceptions, lncRNAs have no defined catalytic activity, suggesting that associated RNA binding proteins (RBPs) play major roles in defining function. Third, RBPs often bind motifs, or k-mers, between 3 to 8 bases in length, and these specificities are deeply conserved (22, 23).

We developed a method, *P*redictive *A*nalysis from *R*epresented *K*mers (PARK), capable of classifying lncRNAs by known function. In the current version of PARK, counts of all kmers (possible RBP binding sites) are collected for each lncRNA in a group. Counts are then standardized and compared to all other lncRNAs to calculate the relative similarity of all pairs of lncRNAs (**Fig 1**). Preliminary analysis found PARK could detect known homologues between mouse and human lncRNAs as well as or better than broadly used Needleman-Wunsch alignments (**Fig 1**). Remarkably, PARK correctly grouped lncRNAs by known function in an analysis of all human and mouse lncRNAs (1): those lncRNAs known to locally repress transcription, such as *XIST* and *KCNQ1OT1*, clustered towards the top of the dendogram, whereas those known to locally activate transcription, such as *PVT1* or *LINC00651*, clustered towards the bottom (**Fig 2**). We are collaborating with Professors Peter Mucha and Kevin Weeks, experts in network and RNA structural analyses, respectively, to further develop PARK. We envision the method’s predictive power will increase when we incorporate more sophisticated network models as well as structural characteristics of RNA (24). We hypothesize that PARK will allow, for the first time, accurate prediction of function for chromatin-associated lncRNAs expressed in cancer cells and we will directly test this hypothesis in **Aim 1**.

*A new assay to study the local effects of lncRNAs on transcription in cancer*: To test predictions made by PARK and to create a flexible platform to study mechanisms of RNA-mediated transcriptional control in cancer, we developed a novel assay that we call *TETRIS* (*T*ransposable *e*lement to *t*est *R*NA’s effect on transcription in c*is*; **Fig 3A,B**). The assay leverages the PiggyBAC transposase (25-30) to insert an inducible lncRNA next to a luciferase reporter into chromosomal sites randomly distributed throughout the genome of any cell type. As expected, when expressed in *TETRIS*, a 2kb fragment of *XIST* repressed luciferase activity in mouse and human cells, and *LINC00651*, a transcriptional activator of the SNAIL oncogene (31), enhanced it (**Fig 3C**). Thus, *TETRIS* can measure local activation and repression by lncRNAs in mouse and human cells.

*Design of a synthetic lncRNA with pre-specified function*: Using the *TETRIS* system, we validated our ability to predict lncRNA function through PARK. We generated 10 million random lncRNAs *in silico* with a length and nucleotide content similar to the 2kb fragment of XIST used in **Fig 3**, and used PARK to find the most and least similar synthetic lncRNAs, respectively. BLAST analysis indicated the selected lncRNAs were truly synthetic as they had no homology greater than 20bp to any region in the mouse or human genomes. Three lncRNAs were synthesized, cloned into *TETRIS*, and examined for repressive activity. As predicted, the most similar synthetic lncRNA repressed luciferase, whereas the least similar lncRNA did not (**Fig 4**). To our knowledge, this is the first synthetic lncRNA ever to be created with a pre-specified biological output. Our success supports our view that we can predict functions of uncharacterized lncRNAs whose PARK profiles pass a threshold of similarity to lncRNAs of known function. We anticipate that the successful development of PARK in **Aim 1** will highlight direct paths to understand the functions and mechanisms of lncRNAs that are expressed across the cancer spectrum, greatly reducing the cost and time needed to characterize their functions in the disease and to discover the means to manipulate their biological activities for therapeutic gain.

**AIMS**

***Aim 1:*** *To discover lncRNAs capable of local activation and repression of transcription in cancer cells.*

Approximately 60,000 lncRNAs without annotated function are found throughout the human genome and ~8,000 are differentially expressed in cancers (2). It remains central challenge to determine which of these, if any, impact the biology of cancer in a meaningful way. Our PARK method allows, for the first time, the functions of lncRNAs to be predicted computationally. Based on PARK profiles that are similar to lncRNAs of known function, we predict upwards of one quarter of all human lncRNAs have potential to regulate the transcription of genes in their local genomic environments; half of these are predicted to activate and half are predicted to repress transcription, respectively. Remarkably, several are encoded within 5kb of some of the most frequently mutated, over-, or under-expressed genes in human cancer, including *AKT*, *BCL2*, *CD44*, *CDKN2A*, *EGFR*, *FOXP1*, *GATA6*, *KIT*, *KRAS*, *MYB*, *MYC*, *NKX2-1*, *NOTCH2*, *PTEN*, *RB1*, and *SOX2*. Thus, if even a fraction of our predictions are correct, they have the potential to unearth regulatory mechanisms whose perturbation may be toxic to a wide range of cancers.

To test this hypothesis, we will profile cytoplasmic and chromatin-associated RNA via RNA-Seq in ~20 cell lines that have been genetically and pharmacologically interrogated as part of the CCLE (21). Selected lines originate from defined subtypes of breast, prostate, lung, and colon cancers, the four tissues from which cancer most commonly originates in the United States. Prior works, including those from TCGA, have profiled RNA expression in cancer, but have not examined chromatin-associated lncRNAs (2). We expect lncRNAs whose natural function is to locally alter transcription will be enriched in chromatin relative to cytoplasmic fractions; hence our need to re-profile RNA expression in these cells. The cell lines selected for profiling are those that most closely resembled tumors from their primary tissue of origin as assessed by CCLE analysis of chromosomal copy number and gene expression patterns (21). This filter ensures our profiling approximates tumor biology in the most reasonable way that can be expected from analysis of immortalized cancer cell lines. HT1080 fibrosarcoma cells will also be profiled because of their utility in testing predictions via TETRIS below. This experiment will generate a list of chromatin-associated lncRNAs expressed in cancer cells that we will then analyze for predicted regulatory ability.

To find which chromatin-associated lncRNAs are capable of locally regulating transcription, we will clone ~30 of the most relevant, as assessed by PARK similarity and their genomic proximity to a cancer-essential gene, into *TETRIS* and perform luciferase assays as in **Fig 3**. Our success in **Figs. 2** and **3**, and our decision to limit **Aim** **1**’s study to chromatin-enriched lncRNAs gives us confidence that we will identify novel lncRNAs that regulate transcription of genes essential for the maintenance of cancer. However, we do not expect all of our predictions to be correct, and efforts are underway to improve PARK with new network and structural models. Therefore, testing of predictions will be iterative; a subset of predicted regulatory lncRNAs will be tested in *TETRIS*, computational models will be adjusted based on results obtained, and new predictions will be made, tested, and so on. To verify that the lncRNAs that alter transcription in the context of *TETRIS* do so endogenously, we will utilize CRISPRa/i (32, 33) to repress or activate their expression, respectively, from their endogenous promoters in cell lines where they are and are not naturally expressed. We will then find the effect that lncRNA perturbation has on nearby mRNA and protein production via qPCR and Western. Last, we will examine how alteration of lncRNA expression changes the ability of these modified cell lines to grow in soft agar assays and colonize nude mice through tail vein injections. Aim 1 thus serves a dual purpose. First, it refines our method to predict lncRNA function that has promise to revolutionize the study of lncRNAs in cancer biology. Second, leveraging PARK, subcellular profiling, the *TETRIS* assay, and endogenous manipulation via CRISPR, we expect to discover a set of lncRNAs whose function is to regulate transcription of critical oncogenes and tumor suppressors. This work represents the first step towards defining the prevalence and physiological impact of RNA-mediated transcriptional control in cancer.

***Aim 2:*** *Systematic discovery and study of RBPs required for RNA-mediated transcriptional control in cancer*

The mechanisms by which lncRNAs control local transcriptional states are poorly defined, making it difficult to predict how their regulatory activities might be leveraged to kill cancer cells. Recent studies have identified ~20 RBPs required for transcriptional regulation induced by different lncRNAs (34-41). However, the molecular role each RBP plays in this regulation has not been defined. Moreover, lncRNA mechanism has never been examined systematically; it is not clear if different lncRNAs require shared or separate subsets of RBPs for transcriptional control. Last, altered expression of many of the RBPs identified, such as HNRNPAB, HNRNPA2B1, HNRNPK, HNRNPU, RBM15, MED12, MED13, WTAP, SF1, SPEN, PTBP1, and TOP1, contribute to cancer in ways that are incompletely defined (42-47). We hypothesize that a major, undescribed function of these RBPs in cancer is to mediate transcriptional regulation by lncRNAs.

Aim 2 will determine which of the 12 cancer-associated RBPs above are required for specific lncRNAs to regulate transcription, and determine how knockdown of these RBPs affects lncRNA-mediated transcriptional control, genome-wide. To discover RBPs required for lncRNA-induced regulation of transcription, we will study a subset of activating and repressing lncRNAs from **Aim 1**, selected based on how frequently altered their natural target genes are in cancer (for example, predicted regulatory lncRNAs near KRAS, EGFR, RB1, and PTEN would be priorities), as well as *XIST*, a known repressor, and *LINC00651* and *PVT1*, known activators of SNAIL and MYC, respectively [**Fig 3**;(7, 31, 48)]. We will use CRISPR to deplete RBPs in TETRIS lines expressing each lncRNA and in TETRIS-GFP control cells. About half of the RBPs to be profiled have been shown to be required for *XIST* and/or *LINC00651* function (37, 39-41); these lncRNAs thus serve as standards against which to compare RBP dependencies of other lncRNAs. To rule out non-specific effects of RBP knockdown on basal transcription from TETRIS, levels of GFP and luciferase mRNA will be measured after knockdown of each RBP in TETRIS-GFP cells. To examine if depletion of the same RBPs disrupts endogenous lncRNA-mediated transcriptional control, we will deplete RBP “hits” using CRISPR in CCLE cell lines of diverse anatomical origin in which prior studies have shown lncRNAs play important roles (MCF7, LNCaP, A549, HCT-116) and perform RNA-Seq. We hypothesize that we will be able to predict where in the genome transcription will be affected and in what direction (either reduced or increased), based on the locations of chromatin-associated lncRNAs that PARK predicted to regulate transcription. Future studies will examine how transcriptional and epigenetic changes induced by RBP depletion alter sensitivity to existing or novel small molecule anticancer agents (21). We hypothesize that the RBPs that are essential for lncRNAs to function represent handles through which lncRNA regulatory capacity can be modulated in cancer.

**LncRNAs** are emerging as major regulators of genome function in some of the most malignant forms of cancer, yet their regulatory prevalence and mechanisms of action remain enigmatic. We describe novel methods to predict and study lncRNA function that we expect will revolutionize the way lncRNAs are studied in cancer. Our long-term goal is to devise methods to manipulate lncRNAs and their effector RBPs in ways that selectively poison transcriptional events essential for cancer cells to survive.

1. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, Lagarde J, Veeravalli L, Ruan X, Ruan Y, Lassmann T, Carninci P, Brown JB, Lipovich L, Gonzalez JM, Thomas M, Davis CA, Shiekhattar R, Gingeras TR, Hubbard TJ, Notredame C, Harrow J, Guigo R. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 2012;22(9):1775-89. Epub 2012/09/08. doi: 22/9/1775 [pii]

10.1101/gr.132159.111. PubMed PMID: 22955988; PMCID: 3431493.

2. Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, Barrette TR, Prensner JR, Evans JR, Zhao S, Poliakov A, Cao X, Dhanasekaran SM, Wu YM, Robinson DR, Beer DG, Feng FY, Iyer HK, Chinnaiyan AM. The landscape of long noncoding RNAs in the human transcriptome. Nat Genet. 2015. doi: 10.1038/ng.3192. PubMed PMID: 25599403.

3. Cheetham SW, Gruhl F, Mattick JS, Dinger ME. Long noncoding RNAs and the genetics of cancer. Br J Cancer. 2013;108(12):2419-25. doi: 10.1038/bjc.2013.233. PubMed PMID: 23660942; PMCID: 3694235.

4. Gutschner T, Diederichs S. The hallmarks of cancer: a long non-coding RNA point of view. RNA Biol. 2012;9(6):703-19. doi: 10.4161/rna.20481. PubMed PMID: 22664915; PMCID: 3495743.

5. Prensner JR, Chinnaiyan AM. The emergence of lncRNAs in cancer biology. Cancer discovery. 2011;1(5):391-407. doi: 10.1158/2159-8290.CD-11-0209. PubMed PMID: 22096659; PMCID: 3215093.

6. Lee JT, Bartolomei MS. X-inactivation, imprinting, and long noncoding RNAs in health and disease. Cell. 2013;152(6):1308-23. Epub 2013/03/19. doi: S0092-8674(13)00205-5 [pii]

10.1016/j.cell.2013.02.016. PubMed PMID: 23498939.

7. Tseng YY, Moriarity BS, Gong WM, Akiyama R, Tiwari A, Kawakami H, Ronning P, Reuland B, Guenther K, Beadnell TC, Essig J, Otto GM, O'Sullivan MG, Largaespada DA, Schwertfeger KL, Marahrens Y, Kawakami Y, Bagchi A. PVT1 dependence in cancer with MYC copy-number increase. Nature. 2014;512(7512):82-+. doi: 10.1038/nature13311. PubMed PMID: WOS:000339908000036.

8. Prensner JR, Iyer MK, Sahu A, Asangani IA, Cao Q, Patel L, Vergara IA, Davicioni E, Erho N, Ghadessi M, Jenkins RB, Triche TJ, Malik R, Bedenis R, McGregor N, Ma T, Chen W, Han SM, Jing XJ, Cao XH, Wang XJ, Chandler B, Yan W, Siddiqui J, Kunju LP, Dhanasekaran SM, Pienta KJ, Feng FY, Chinnaiyan AM. The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. Nature Genetics. 2013;45(11):1392-+. doi: 10.1038/ng.2771. PubMed PMID: WOS:000326384100023.

9. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai MC, Hung T, Argani P, Rinn JL, Wang Y, Brzoska P, Kong B, Li R, West RB, van de Vijver MJ, Sukumar S, Chang HY. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature. 2010;464(7291):1071-6. Epub 2010/04/16. doi: nature08975 [pii]

10.1038/nature08975. PubMed PMID: 20393566; PMCID: 3049919.

10. Kim K, Jutooru I, Chadalapaka G, Johnson G, Frank J, Burghardt R, Kim S, Safe S. HOTAIR is a negative prognostic factor and exhibits pro-oncogenic activity in pancreatic cancer. Oncogene. 2013;32(13):1616-25. Epub 2012/05/23. doi: onc2012193 [pii]

10.1038/onc.2012.193. PubMed PMID: 22614017; PMCID: 3484248.

11. Liu XH, Liu ZL, Sun M, Liu J, Wang ZX, De W. The long non-coding RNA HOTAIR indicates a poor prognosis and promotes metastasis in non-small cell lung cancer. BMC Cancer. 2013;13:464. doi: 10.1186/1471-2407-13-464. PubMed PMID: 24103700; PMCID: 3851855.

12. Geisler S, Coller J. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. Nat Rev Mol Cell Biol. 2013;14(11):699-712. doi: 10.1038/nrm3679. PubMed PMID: 24105322.

13. Morris KV, Mattick JS. The rise of regulatory RNA. Nat Rev Genet. 2014;15(6):423-37. doi: 10.1038/nrg3722. PubMed PMID: 24776770; PMCID: 4314111.

14. Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. Annu Rev Biochem. 2012;81:145-66. Epub 2012/06/06. doi: 10.1146/annurev-biochem-051410-092902. PubMed PMID: 22663078.

15. Grant J, Mahadevaiah SK, Khil P, Sangrithi MN, Royo H, Duckworth J, McCarrey JR, VandeBerg JL, Renfree MB, Taylor W, Elgar G, Camerini-Otero RD, Gilchrist MJ, Turner JM. Rsx is a metatherian RNA with Xist-like properties in X-chromosome inactivation. Nature. 2012;487(7406):254-8. Epub 2012/06/23. doi: nature11171 [pii]

10.1038/nature11171. PubMed PMID: 22722828.

16. Melamed E, Arnold AP. Regional differences in dosage compensation on the chicken Z chromosome. Genome Biol. 2007;8(9):R202. doi: 10.1186/gb-2007-8-9-r202. PubMed PMID: 17900367; PMCID: 2375040.

17. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 2011;25(18):1915-27. Epub 2011/09/06. doi: gad.17446611 [pii]

10.1101/gad.17446611. PubMed PMID: 21890647; PMCID: 3185964.

18. Li J, Wang Y, Yu JJ, Dong RF, Qiu HF. A high level of circulating HOTAIR is associated with progression and poor prognosis of cervical cancer. Tumor Biol. 2015;36(3):1661-5. doi: 10.1007/s13277-014-2765-4. PubMed PMID: WOS:000351884000039.

19. Vausort M, Wagner DR, Devaux Y. Long noncoding RNAs in patients with acute myocardial infarction. Circ Res. 2014;115(7):668-77. doi: 10.1161/CIRCRESAHA.115.303836. PubMed PMID: 25035150.

20. Wang R, Chinnaiyan AM, Dunn RL, Wojno KJ, Wei JT. Rational Approach to Implementation of Prostate Cancer Antigen 3 Into Clinical Care. Cancer. 2009;115(17):3879-86. doi: 10.1002/cncr.24447. PubMed PMID: WOS:000269230700013.

21. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehar J, Kryukov GV, Sonkin D, Reddy A, Liu M, Murray L, Berger MF, Monahan JE, Morais P, Meltzer J, Korejwa A, Jane-Valbuena J, Mapa FA, Thibault J, Bric-Furlong E, Raman P, Shipway A, Engels IH, Cheng J, Yu GK, Yu J, Aspesi P, Jr., de Silva M, Jagtap K, Jones MD, Wang L, Hatton C, Palescandolo E, Gupta S, Mahan S, Sougnez C, Onofrio RC, Liefeld T, MacConaill L, Winckler W, Reich M, Li N, Mesirov JP, Gabriel SB, Getz G, Ardlie K, Chan V, Myer VE, Weber BL, Porter J, Warmuth M, Finan P, Harris JL, Meyerson M, Golub TR, Morrissey MP, Sellers WR, Schlegel R, Garraway LA. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483(7391):603-7. Epub 2012/03/31. doi: nature11003 [pii]

10.1038/nature11003. PubMed PMID: 22460905; PMCID: 3320027.

22. Stefl R, Skrisovska L, Allain FH. RNA sequence- and shape-dependent recognition by proteins in the ribonucleoprotein particle. EMBO Rep. 2005;6(1):33-8. doi: 10.1038/sj.embor.7400325. PubMed PMID: 15643449; PMCID: 1299235.

23. Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, Na H, Irimia M, Matzat LH, Dale RK, Smith SA, Yarosh CA, Kelly SM, Nabet B, Mecenas D, Li W, Laishram RS, Qiao M, Lipshitz HD, Piano F, Corbett AH, Carstens RP, Frey BJ, Anderson RA, Lynch KW, Penalva LO, Lei EP, Fraser AG, Blencowe BJ, Morris QD, Hughes TR. A compendium of RNA-binding motifs for decoding gene regulation. Nature. 2013;499(7457):172-7. doi: 10.1038/nature12311. PubMed PMID: 23846655; PMCID: 3929597.

24. Mathews DH. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. RNA. 2004;10(8):1178-90. doi: 10.1261/rna.7650904. PubMed PMID: 15272118; PMCID: 1370608.

25. Ding S, Wu X, Li G, Han M, Zhuang Y, Xu T. Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. Cell. 2005;122(3):473-83. doi: 10.1016/j.cell.2005.07.013. PubMed PMID: 16096065.

26. Chaudhury A, Kongchan N, Gengler JP, Mohanty V, Christiansen AE, Fachini JM, Martin JF, Neilson JR. A piggyBac-based reporter system for scalable in vitro and in vivo analysis of 3' untranslated region-mediated gene regulation. Nucleic Acids Res. 2014;42(10):e86. doi: 10.1093/nar/gku258. PubMed PMID: 24753411; PMCID: 4041432.

27. Woltjen K, Michael IP, Mohseni P, Desai R, Mileikovsky M, Hamalainen R, Cowling R, Wang W, Liu P, Gertsenstein M, Kaji K, Sung HK, Nagy A. piggyBac transposition reprograms fibroblasts to induced pluripotent stem cells. Nature. 2009;458(7239):766-70. doi: 10.1038/nature07863. PubMed PMID: 19252478; PMCID: 3758996.

28. Woltjen K, Kim SI, Nagy A. The piggyBac Transposon as a Platform Technology for Somatic Cell Reprogramming Studies in Mouse. Methods Mol Biol. 2015. doi: 10.1007/7651\_2015\_274. PubMed PMID: 26126450.

29. Ivics Z, Li MA, Mates L, Boeke JD, Nagy A, Bradley A, Izsvak Z. Transposon-mediated genome manipulation in vertebrates. Nat Methods. 2009;6(6):415-22. doi: 10.1038/nmeth.1332. PubMed PMID: 19478801; PMCID: 2867038.

30. Nakazawa Y, Huye LE, Dotti G, Foster AE, Vera JF, Manuri PR, June CH, Rooney CM, Wilson MH. Optimization of the PiggyBac transposon system for the sustained genetic modification of human T lymphocytes. J Immunother. 2009;32(8):826-36. doi: 10.1097/CJI.0b013e3181ad762b. PubMed PMID: 19752751; PMCID: 2796278.

31. Orom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, Lai F, Zytnicki M, Notredame C, Huang Q, Guigo R, Shiekhattar R. Long noncoding RNAs with enhancer-like function in human cells. Cell. 2010;143(1):46-58. doi: 10.1016/j.cell.2010.09.001. PubMed PMID: 20887892; PMCID: 4108080.

32. Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, Whitehead EH, Guimaraes C, Panning B, Ploegh HL, Bassik MC, Qi LS, Kampmann M, Weissman JS. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. Cell. 2014. doi: 10.1016/j.cell.2014.09.029. PubMed PMID: 25307932.

33. Hilton IB, D'Ippolito AM, Vockley CM, Thakore PI, Crawford GE, Reddy TE, Gersbach CA. Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. Nat Biotechnol. 2015;33(5):510-7. doi: 10.1038/nbt.3199. PubMed PMID: 25849900; PMCID: 4430400.

34. Li Z, Chao TC, Chang KY, Lin N, Patil VS, Shimizu C, Head SR, Burns JC, Rana TM. The long noncoding RNA THRIL regulates TNFalpha expression through its interaction with hnRNPL. Proc Natl Acad Sci U S A. 2014;111(3):1002-7. doi: 10.1073/pnas.1313768111. PubMed PMID: 24371310; PMCID: 3903238.

35. Lin N, Chang KY, Li Z, Gates K, Rana ZA, Dang J, Zhang D, Han T, Yang CS, Cunningham TJ, Head SR, Duester G, Dong PD, Rana TM. An evolutionarily conserved long noncoding RNA TUNA controls pluripotency and neural lineage commitment. Mol Cell. 2014;53(6):1005-19. doi: 10.1016/j.molcel.2014.01.021. PubMed PMID: 24530304; PMCID: 4010157.

36. Carpenter S, Aiello D, Atianand MK, Ricci EP, Gandhi P, Hall LL, Byron M, Monks B, Henry-Bezy M, Lawrence JB, O'Neill LA, Moore MJ, Caffrey DR, Fitzgerald KA. A long noncoding RNA mediates both activation and repression of immune response genes. Science. 2013;341(6147):789-92. doi: 10.1126/science.1240925. PubMed PMID: 23907535.

37. Chu C, Zhang QC, da Rocha ST, Flynn RA, Bharadwaj M, Calabrese JM, Magnuson T, Heard E, Chang HY. Systematic discovery of xist RNA binding proteins. Cell. 2015;161(2):404-16. doi: 10.1016/j.cell.2015.03.025. PubMed PMID: 25843628; PMCID: 4425988.

38. Hasegawa Y, Brockdorff N, Kawano S, Tsutui K, Nakagawa S. The matrix protein hnRNP U is required for chromosomal localization of Xist RNA. Dev Cell. 2010;19(3):469-76. Epub 2010/09/14. doi: S1534-5807(10)00381-3 [pii]

10.1016/j.devcel.2010.08.006. PubMed PMID: 20833368.

39. McHugh CA, Chen CK, Chow A, Surka CF, Tran C, McDonel P, Pandya-Jones A, Blanco M, Burghard C, Moradian A, Sweredoski MJ, Shishkin AA, Su J, Lander ES, Hess S, Plath K, Guttman M. The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. Nature. 2015;521(7551):232-6. doi: 10.1038/nature14443. PubMed PMID: 25915022.

40. Moindrot B, Cerase A, Coker H, Masui O, Grijzenhout A, Pintacuda G, Schermelleh L, Nesterova TB, Brockdorff N. A Pooled shRNA Screen Identifies Rbm15, Spen, and Wtap as Factors Required for Xist RNA-Mediated Silencing. Cell Rep. 2015;12(4):562-72. doi: 10.1016/j.celrep.2015.06.053. PubMed PMID: 26190105; PMCID: 4534822.

41. Lai F, Orom UA, Cesaroni M, Beringer M, Taatjes DJ, Blobel GA, Shiekhattar R. Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. Nature. 2013;494(7438):497-501. doi: 10.1038/nature11884. PubMed PMID: 23417068; PMCID: 4109059.

42. Mercher T, Coniat MB, Monni R, Mauchauffe M, Nguyen Khac F, Gressin L, Mugneret F, Leblanc T, Dastugue N, Berger R, Bernard OA. Involvement of a human gene related to the Drosophila spen gene in the recurrent t(1;22) translocation of acute megakaryocytic leukemia. Proc Natl Acad Sci U S A. 2001;98(10):5776-9. doi: 10.1073/pnas.101001498. PubMed PMID: 11344311; PMCID: 33289.

43. Arcaini L, Rossi D, Lucioni M, Nicola M, Bruscaggin A, Fiaccadori V, Riboni R, Ramponi A, Ferretti VV, Cresta S, Casaluci GM, Bonfichi M, Gotti M, Merli M, Maffi A, Arra M, Varettoni M, Rattotti S, Morello L, Guerrera ML, Sciarra R, Gaidano G, Cazzola M, Paulli M. The NOTCH pathway is recurrently mutated in diffuse large B-cell lymphoma associated with hepatitis C virus infection. Haematologica. 2015;100(2):246-52. doi: 10.3324/haematol.2014.116855. PubMed PMID: WOS:000351278900023.

44. Gao R, Yu Y, Inoue A, Widodo N, Kaul SC, Wadhwa R. Heterogeneous nuclear ribonucleoprotein K (hnRNP-K) promotes tumor metastasis by induction of genes involved in extracellular matrix, cell movement, and angiogenesis. J Biol Chem. 2013;288(21):15046-56. doi: 10.1074/jbc.M113.466136. PubMed PMID: 23564449; PMCID: 3663525.

45. McCutcheon IE, Hentschel SJ, Fuller GN, Jin W, Cote GJ. Expression of the splicing regulator polypyrimidine tract-binding protein in normal and neoplastic brain. Neuro Oncol. 2004;6(1):9-14. doi: 10.1215/S1152851703000279. PubMed PMID: 14769134; PMCID: 1871971.

46. Ahuja HG, Felix CA, Aplan PD. The t(11;20)(p15;q11) chromosomal translocation associated with therapy-related myelodysplastic syndrome results in an NUP98-TOP1 fusion. Blood. 1999;94(9):3258-61. PubMed PMID: 10556215.

47. Shitashige M, Naishiro Y, Idogawa M, Honda K, Ono M, Hirohashi S, Yamada T. Involvement of splicing factor-1 in beta-catenin/T-cell factor-4-mediated gene transactivation and pre-mRNA splicing. Gastroenterology. 2007;132(3):1039-54. doi: 10.1053/j.gastro.2007.01.007. PubMed PMID: 17383426.

48. Wutz A, Rasmussen TP, Jaenisch R. Chromosomal silencing and localization are mediated by different domains of Xist RNA. Nat Genet. 2002;30(2):167-74. Epub 2002/01/10. doi: 10.1038/ng820

ng820 [pii]. PubMed PMID: 11780141.

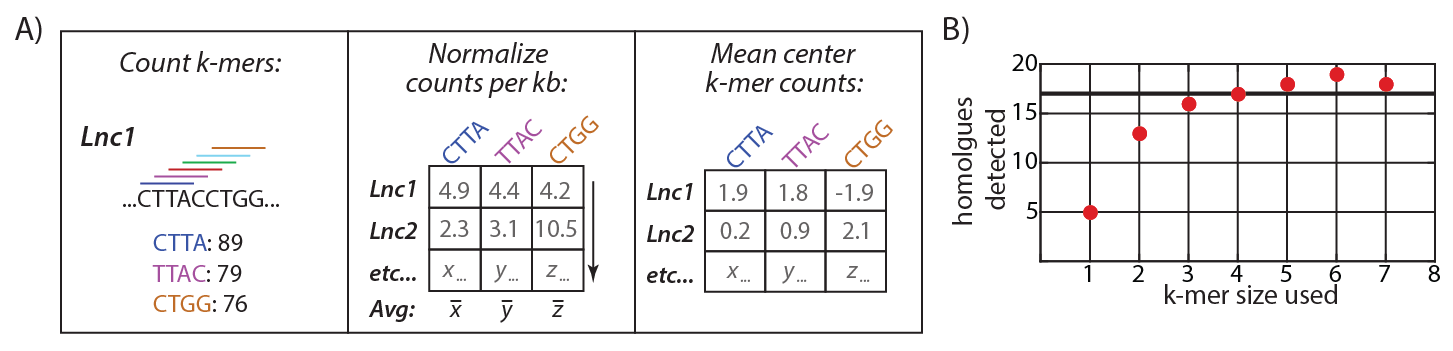


Figure 1. (A) Overview of PARK. (B) PARK using k-mers of length four and above detects homologues as or more efficiently than Needleman-Wunsch alignments (black line) in a training set of 36 lncRNAs that are conserved between human and mouse gcREF.

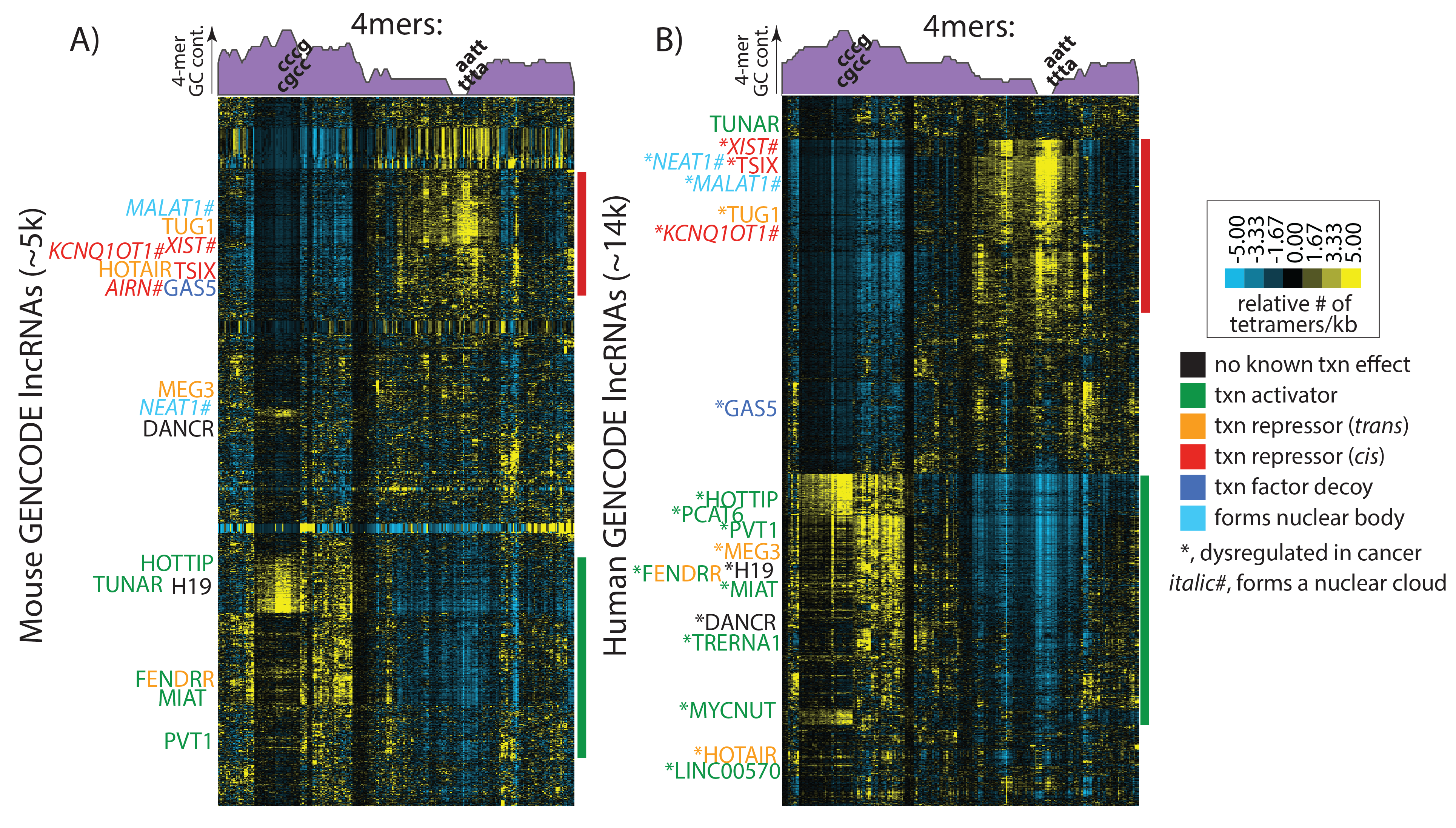


Figure 2. Mouse (A) and human (B) lncRNAs hierarchically clustered via PARK. LncRNAs with known functions are shown relative to dendogram position. Red and green bars to the right of the dendograms mark predicted repressive and activating *cis*-acting lncRNAs. Purple density plots show GC content summed in five 4-mer windows across dendograms. Near identical patterns and lncRNA clusters are observed with k-mer length 5 and 6 (not shown).

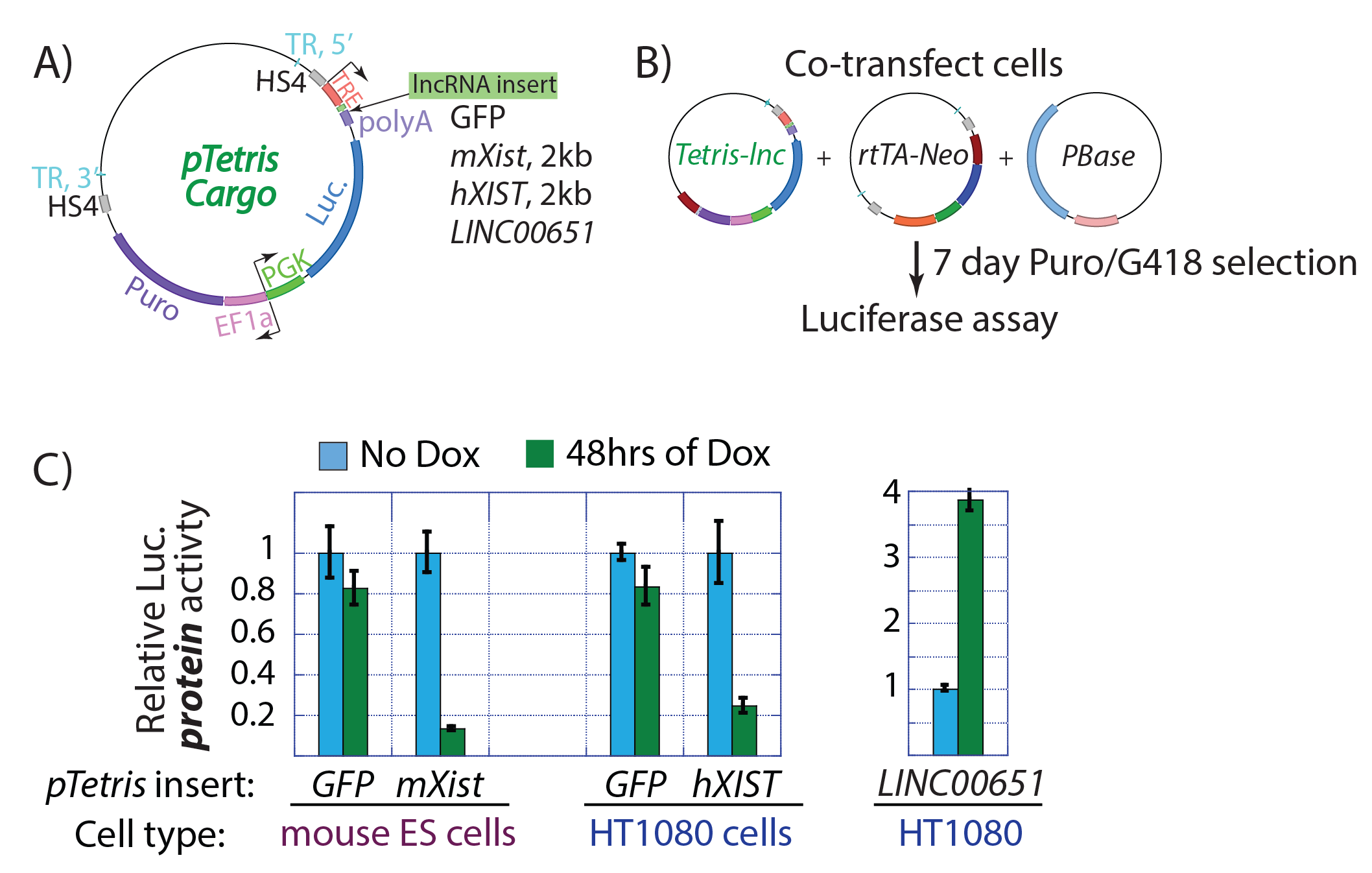
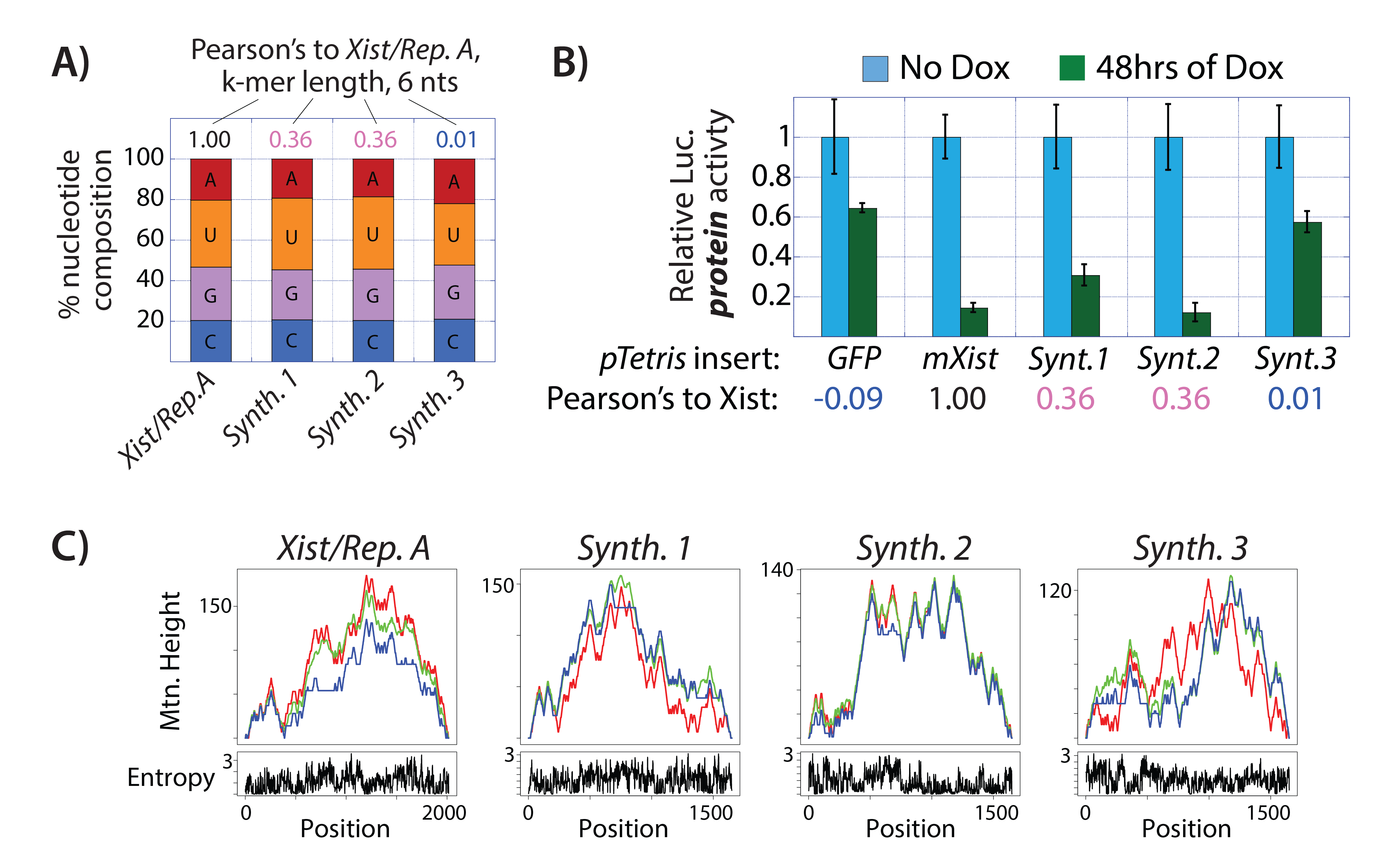
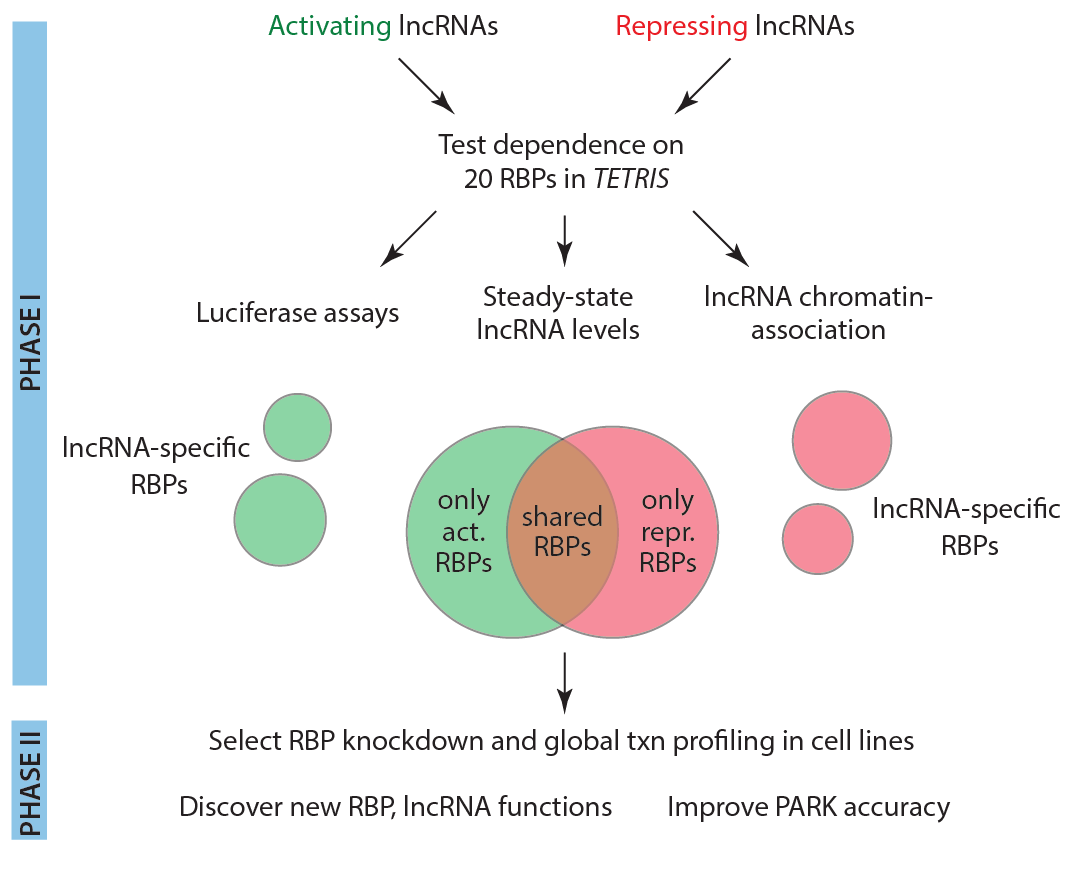


Figure 3. (A) *pTetris* cargo vector. LncRNAs are cloned behind a tetracycline-inducible promoter (the TRE) and effect on luciferase expression is measured after addition of the tetracycline analogue, doxycycline. TR, PBase recognition elements. HS4, chicken beta-globin insulator. (B) To perform a *pTetris* assay, three requisite plasmids are co-transfected and luciferase activity can be measured 7 days after selection in Puro/G418. (C) GFP, the first 2kb of mouse and human Xist [which encompass the “Repeat A” region; repaREF], and *LINC00651* [a known activating lncRNA; oromREF] were cloned into *pTetris*. *Xist* induction repressed luciferase protein activity 5 to 10-fold after 48 hrs of expression in mouse ES cells and human HT1080 fibrosarcoma cells. Induction of *LINC00651* (a known activating lncRNA) increases luciferase protein activity 4-fold after 48 hrs. GFP induction had a mild repressive effect, possibly due to read-through transcription.



**Figure 4. (A)** Three 1,650 nucleotide long synthetic lncRNAs (Synth.1, 2, and 3) were synthesized to have a nucleotide content similar to that found in the first 2kb of mouse Xist, but whose k-mer contents varied. As quantified by Pearson’s correlation, Synth.1 and 2 had k-mer contents highly similar to Xist, scoring in the top 10 out of 10 million randomly generated synthetic lncRNAs, whereas Synth. 3 was in the bottom 10. **(B)** The synthetic lncRNA “Synth.2” repressed luciferase activity to the same extent as the first 2kb of mouse Xist, and represents, to our knowledge, the first synthetic lncRNA ever created with a pre-specified biological output. Interestingly, “Synth.1” did not repress as efficiently despite having an equal level of k-mer similarity to Xist as defined by PARK. As predicted, Synth. 3 did not repress luciferase when compared to the GFP control. **(C)** ViennaFOLD mountain and positional entropy plots viennaREF highlight predicted structural similarities and differences between the four lncRNAs. Red, blue, and green lines in mountain plots represent the minimum free energy structure, the thermodynamic ensemble of predicted structures, and the centroid structure, respectively. We are collaborating with Professors Peter Mucha and Kevin Weeks at UNC to incorporate metadata such as these into PARK to augment its predictive power.



**Figure 5. Systematic discovery and study of RBPs required for lncRNA-mediated transcriptional control in cancer**

**(Aim 2 overview).** Five activating and five repressive lncRNAs discovered in Aim 1 will be selected for study in Aim 2 based on their regulatory potency and the intuited likelihood that they control gene expression important for cancer cell survival. Using *TETRIS*-lncRNA HT1080 cell lines created in Aim 1, we will examine to what extent each of these lncRNAs (plus *PVT1*, *LINC00651* and *XIST*) depend on 12 cancer-associated RBPs to exert transcriptional control. HT1080 cells are optimal for TETRIS assays because they are easily transfected and express the luciferase reporter at levels that allow robust detection of both transcriptional activation and repression induced by lncRNAs. From these experiments, we anticipate finding that certain RBPs will be required for all lncRNAs tested, certain RBPs will required for activating or repressive lncRNAs only, and others may be required for individual lncRNAs or not at all. Phase II will examine the roles of select RBPs in regulating endogenous lncRNA function in select CCLE cell lines profiled in Aim 1. These experiments may uncover novel regulatory roles for the RBPs and chromatin-associated lncRNAs analyzed, and will also be used to improve predictions made by PARK. We anticipate that Phase II experiments will be completed outside of the 2-year funding period of the Kimmel award.