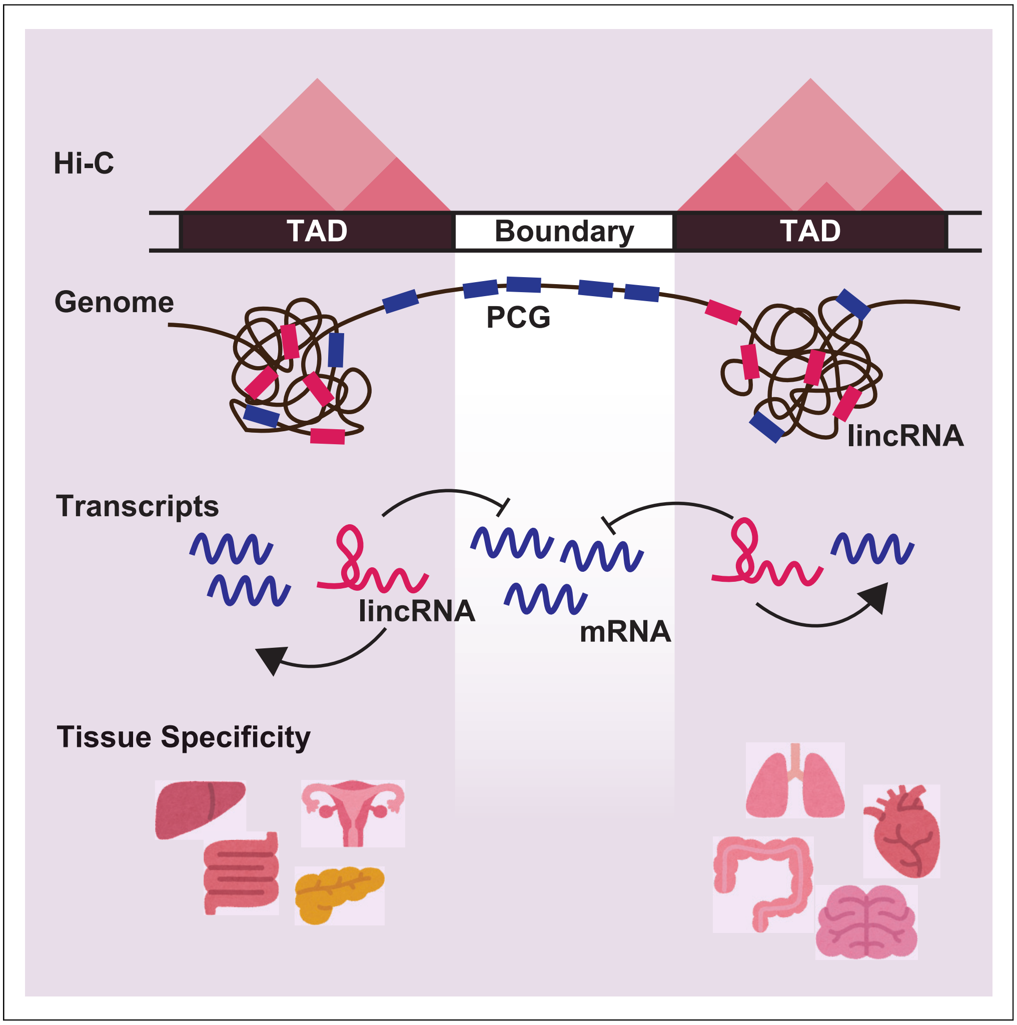
iScience

CellPress OPEN ACCESS

Article

Topologically associating domain underlies tissue specific expression of long intergenic non-coding RNAs



Yu Hamba, Takashi Kamatani, Fuyuki Miya, Keith A. Boroevich, Tatsuhiko Tsunoda

tsunoda@bs.s.u-tokyo.ac.jp

Highlights

lincRNA loci are found more enriched within TADs than TAD boundaries

lincRNAs are found more inside TADs than PCGs

Expression patterns of lincRNAs in TADs have higher tissue-specificity

lincRNAs analysis with TAD coordinates captures pathological transcriptional states

Hamba et al., iScience *26*, 106640
May 19, 2023 © 2023 The Author(s).
https://doi.org/10.1016/j.isci.2023.106640

iScience

CellPress
OPEN ACCESS

Article

Topologically associating domain underlies tissue specific expression of long intergenic non-coding RNAs

Yu Hamba,[1,2] Takashi Kamatani,[1,3,4,5] Fuyuki Miya,[6] Keith A. Boroevich,[2] and Tatsuhiko Tsunoda[1,2,7,8,*]

SUMMARY

Accumulating evidence indicates that long intergenic non-coding RNAs (lincRNAs) show more tissue-specific expression patterns than protein-coding genes (PCGs). However, although lincRNAs are subject to canonical transcriptional regulation like PCGs, the molecular basis for the specificity of their expression patterns remains unclear. Here, using expression data and coordinates of topologically associating domains (TADs) in human tissues, we show that lincRNA loci are significantly enriched in the more internal region of TADs compared to PCGs and that lincRNAs within TADs have higher tissue specificity than those outside TADs. Based on these, we propose an analytical framework to interpret transcriptional status using lincRNA as an indicator. We applied it to hypertrophic cardiomyopathy data and found disease-specific transcriptional regulation: ectopic expression of keratin at the TAD level and derepression of myocyte differentiation-related genes by E2F1 with down-regulation of LINC00881. Our results provide understanding of the function and regulation of lincRNAs according to genomic structure.

INTRODUCTION

A large proportion of the human genome is transcribed into RNA, yielding many non-coding RNAs (ncRNAs) in addition to mRNAs.[1] Among ncRNAs, long non-coding RNAs (lncRNAs) are broadly defined as transcripts longer than 200 nucleotides that lack coding capacity.[2,3] Accumulating evidence indicates that without being translated into proteins, lncRNAs are functional in many cellular processes, such as gene imprinting,[4] development[5] and immune response.[6] Among those various mechanisms reported, many lncRNAs are associated with chromatin-modifying complexes and guide conformational changes in nuclear domains or regulation of transcriptional enhancer activity.[7] Moreover, lncRNA dysfunction has been associated with a variety of human diseases, including cancer, cardiovascular disease, and neurodegenerative disorders.[8]

lncRNAs are classified based on their genomic origin: Antisense lncRNAs are transcribed from the opposite strand of protein-coding genes (PCGs), bidirectional eRNAs are transcribed from the enhancer region, and lincRNAs have transcription units independent of promoters and enhancers of PCGs.[9] lincRNAs possess many characteristics of mRNAs: lincRNA transcription is regulated by key transcription factors and typical histone modifications, and lincRNA transcripts are processed by the canonical spliceosomal machinery.[10–13] Compared to protein-coding genes, lincRNA expression is highly specific to cell types and tissues.[10,13,14] Compared to mRNA, which requires translation into protein to exert gene function, lincRNA transcripts exert their function directly on their own. This means that lincRNA transcriptional regulation serves as a functional control, emphasizing the importance of elucidating the molecular basis of specific expression. Although the mechanisms that cause tissue-specific gene expression patterns have been extensively investigated in terms of promoter sequences[15] and epigenomic modifications,[16] the mechanisms that can comprehensively explain the nature of lincRNAs and this difference in expression specificity between mRNAs and lincRNAs are still unclear.

One of the highest order regulatory mechanisms that determine transcription patterns is genomic structure that allows for genome-wide regulation; chromatin compaction, as seen in heterochromatin regions, limits access to gene promoters. In the nucleus, chromosomes form higher-order structures that govern gene

[1]Laboratory for Medical Science Mathematics, Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Tokyo 113-0033, Japan

[2]Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama 230-0045, Japan

[3]Department of AI Technology Development, M&D Data Science Center, Tokyo Medical and Dental University, Tokyo 101-0062, Japan

[4]Division of Precision Cancer Medicine, Tokyo Medical and Dental University Hospital, Tokyo 113-8519, Japan

[5]Division of Pulmonary Medicine, Department of Medicine, Keio University School of Medicine, Tokyo 160-8582, Japan

[6]Center for Medical Genetics, Keio University School of Medicine, Tokyo 160-8582, Japan

[7]Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Chiba 277-8562, Japan

[8]Lead contact

*Correspondence: tsunoda@bs.s.u-tokyo.ac.jp

https://doi.org/10.1016/j.isci.2023.106640

regulation.[17] Recent advances in chromatin conformation capture technology have revealed the three-dimensional structure of the genome. Adjacent chromosomal regions were shown to fold into topological associating domains (TADs) ranging in size from a few hundred kb to 1–2 Mb.[18] Chromatin structures that comprehensively regulate the tissue-specific expression patterns of lincRNAs must be tissue and genome-wide regulatory systems; TADs fulfill this requirement by confining enhancer-promoter interactions within each TAD, thus enabling gene expression regulation. Literature shows that lincRNAs expression are positively correlated with those of nearby genes,[19] suggesting that regulation by enhancers insulated within TADs determine when and in which tissues lincRNAs are transcribed with neighboring genes.

In this study, first, we combined TAD coordinates from 13 primary human tissues with expression data from the GTEx project[20] to investigate into the distribution of lincRNA loci in the human genome in relation to TADs and clarify how they are associated with tissue-specific expression patterns. Next, by defining and utilizing two groups of PCGs: Co-expressed *cis* PCGs spatially close to the lincRNA locus and *trans* PCGs regulated by post-transcriptional lincRNA molecules, we proposed an analytical framework to infer biological processes relevant to traits such as tissue phenotypes and diseases. Because TAD structure is maintained throughout the cell lineage, our method should be applicable to a broad range of human gene expression data. Last, as an example for its medical applications, we analyzed hypertrophic cardiomyopathy (HCM) data, and identified lincRNA signatures specific to HCM. Overall, our findings demonstrate that interpreting the genome at the 3D level can provide insight into lincRNA functionality.

RESULTS

Characteristics of lincRNA expression in primary human tissues

First, we investigated tissue specificity of lincRNA expression by using expression data from 54 human tissues of GTEx. t-distributed stochastic neighbor embedding (t-SNE) analysis of the expression data showed a clear clustering of samples by tissue of origin (Figure S1A), and the expression pattern of 7345 lincRNAs appeared to be at least as tissue-specific as that of 18358 PCGs (Figure S1B). To quantitatively compare tissue specificity of expression between lincRNAs and PCGs, we calculated tissue specificity scores (0 = ubiquitous; 1 = specific to a single tissue) based on the Shannon entropy specificity of each gene (Figure 1A). Overall, lincRNA has higher tissue specificity than PCG, consistent with previous findings.[10,11,21] The distribution of the tissue specificity scores showed bimodal distributions with high and low specificity groups for both lincRNA and PCG. The percentage of the completely tissue-specific group of lincRNAs was larger than that of PCGs (8.81% versus 1.06%), and the tissue specificity score of lincRNAs was significantly higher than that of PCGs ($p < 10^{-4}$; permutation test; Figure 1A right). This suggests that there is likely a genome-wide regulatory mechanism that results in a tissue-specific expression pattern of lincRNAs that is differentiated from the pattern of expression of PCGs.

To characterize lincRNA expression in the human tissues, we conducted an agglomerative clustering analysis on the lincRNA expression profiles from the GTEx data. This resulted in the generation of three major clusters: class 1, 2, and 3 (Figure 1B). The distribution of the tissue specificity scores was lowest for class 1 and highest for class 3 (Figure 1C). Next, we examined the lincRNA expression intensities and found that class 1 was the highest and class 3 was the lowest (Figure 1D). We also calculated correlation coefficients between expression levels of lincRNA and those of neighboring genes for each cluster, and its result showed that lincRNA expression levels were positively correlated with those of neighboring genes in most cases (Figure 1E).

lincRNA loci show functional localization to TADs

Using TAD coordinates detected by Hi-C on 13 primary human tissues,[22] we tested the hypothesis that tissue-specific expression and function of lincRNAs is regulated in a manner dependent on the higher-order structure of the genome. When PCGs and lincRNAs were compared, lincRNAs were found to have a significantly higher percentage of loci within TADs across multiple tissues (Figures 2A and S2A). More interestingly, the localization of lincRNAs in TADs was more significant than those of other ncRNA subtypes such as antisense, miRNA, and rRNA. This result suggests that beyond just being a set of untranslated transcripts, ncRNAs, particularly lincRNAs, can be clearly differentiated in the context of the chromatin environment, such as TADs. To rule out the possibility that the differences of the gene locus enrichment for TADs were because of differences in gene length, we conducted a permutation test in which gene loci were randomly replaced in the genome (Figure S2B). The results showed that although the PCG loci were less enriched in TADs relative to the random distributions in almost all tissues ($p < 10^{-4}$; permutation test;

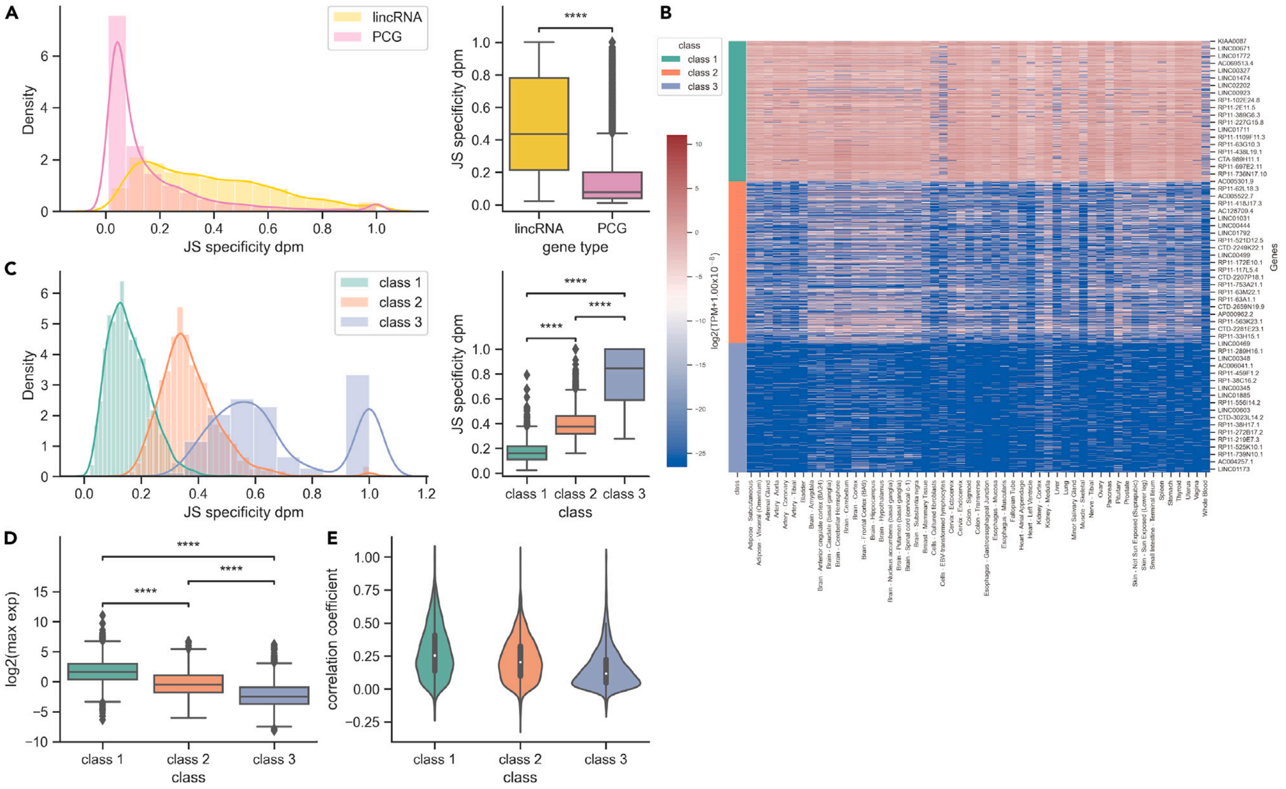


Figure 1. The expression of lincRNAs is specific in human tissues

(A) Distribution of tissue specificity scores calculated for 7345 lincRNA (yellow) and 18358 PCGs (pink) across GTEx 54 tissues ($p< 10^{-323}$).

(B) Heatmap generated by unsupervised cluster analysis of all lincRNA expression profile across all GTEx tissues. Only partial gene names are shown.

(C) Distribution of tissue specificity scores calculated for lincRNA across the three clusters.

(D) Boxplots of max expression levels of lincRNAs across all GTEx tissues for the three lincRNA clusters (class 1 versus class 2: $p = 1.89 \times 10^{-88}$, class 2 versus class 3: $p = 9.11 \times 10^{-33}$, class 1 versus class 3: $p = 1.46 \times 10^{-249}$).

(E) Distribution of the Pearson correlation coefficients between adjacent PCG-lincRNA pairs across all GTEx samples for the three lincRNA clusters. A, D ****$p< 10^{-4}$, Mann-Whitney-Wilcoxon test with Bonferroni correction, two-sided.

Figure S3A), lincRNA loci were significantly more enriched in TADs than compared to the random distributions in all tissues ($p< 10^{-4}$; permutation test; Figure S3B). Incidentally, antisense ncRNAs showed a wide variation and no clear trends in their localization across tissues (Figure S3C).

To examine the relationship between gene loci enrichment in TADs and tissue specificity of their expressions, we defined coordinates of TADs that could be analyzed across tissues. First, we extracted TAD regions common across all 13 tissues (Figure 2B). Although the common TAD regions had a lower overall genome coverage (total coverage: 60.1%) compared to the TAD regions in each tissue, they had a median length of 0.64M bp per region and median gene count of 10, which did not deviate significantly from the general definition for a TAD (Figures S4A and S4B).[18] To confirm whether these regions function as effective TAD insulators, we examined the differences in correlations between gene expression levels according to the gene positions in relation to the TAD regions by using the GTEx expression data. More specifically, we calculated the moving average of the expression correlations according to the intergenic distance for three conditions: gene pairs belonging to the same TAD, those belonging to TAD and TAD boundary, and those belonging to different TADs. Gene pairs belonging to the same TAD maintained higher correlations than other pairs across long distances ~1M bp (Figure 2C), indicating that the gene expression regulation by TADs was observed on a genome scale, supporting previous studies.[23–25] Next, we defined the expression level of each TAD as the average of the expression levels of the included genes, and examined the correlations of the expression levels among the TADs of each chromosome using all of the GTEx expression data. They showed almost entirely positive correlations except for a small portion that had differences in TAD transcription levels among tissues (Figures S5A and S5B). These results confirmed that the common

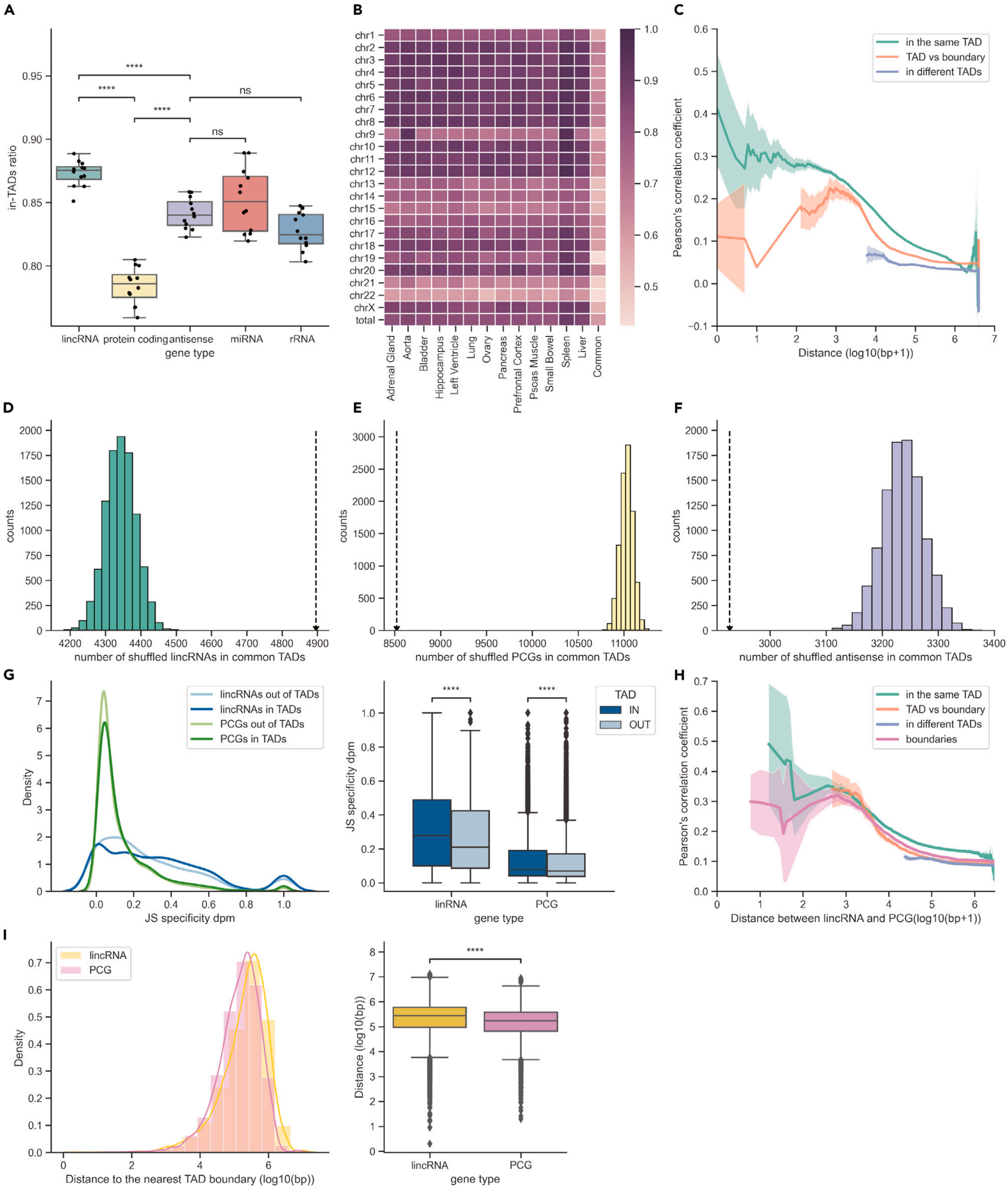


Figure 2. Functional enrichment of lincRNAs in the higher-order structure of genome

(A) Ratio of genes resided in TAD of 12 tissues (adrenal grand, aota, bladder, hippocampus, left ventricle, lung, liver, ovary, pancreas, prefrontal cortex, psoas muscle and small bowel) across lincRNA, PCG, antisense, miRNA and rRNA. Spleen data were excluded here because they showed outliers. lincRNA versus protein_coding: $p = 1.40 \times 10^{-13}$, lincRNA versus antisense: $p = 2.68 \times 10^{-6}$, protein_coding versus antisense: $p = 2.62 \times 10^{-9}$, t-test independent samples with Bonferroni correction.

Figure 2. *Continued*

(B) Heatmap of TAD coverage of all autosomes, X chromosomes and whole genome for 13 tissues and common regions.

(C) Moving average of expression correlation coefficients according to intergenic distances for gene pairs belonging to the same TAD, TAD and boundary, and different TADs. Error bands represent 95% confidence intervals.

(D–F) The number of genes that resided in TADs compared to the distribution of controls (histogram) of 10^4 sets of randomly replaced lincRNA loci for (D) PCG loci (E) antisense loci and (F) in common TAD region. The numbers for the true gene loci are indicated by black arrows with dashed lines.

(G) Distribution of tissue specificity scores calculated across GTEx tissues for lincRNAs in TADs, lincRNAs outside TADs, PCGs in TADs and PCGs outside TADs.

(H) Moving average of expression correlation coefficients according to intergenic distances for lincRNA-PCG pairs belonging to the same TAD, TAD and boundary different TADs and boundaries. Error bands signify 95% confidence intervals.

(I) Distributions of distance between gene and the nearest TAD boundary for lincRNAs in TADs and PCGs in TADs. G, I ****$p< 10^{-4}$, Mann-Whitney-Wilcoxon test with Bonferroni correction, two-sided.

TAD regions derived from the TAD coordinates across the 13 tissues maintain the basic characteristics of TAD, which previous studies have suggested.[24,26,27] Similar to the TAD regions identified from the primary tissues, the lincRNA loci were found to be strongly enriched within the common TAD regions against the random distribution by permutation, whereas PCG and antisense loci were less enriched in TAD regions ($p< 10^{-4}$; permutation test; Figures 2D–2F and S3A–S3C).

To confirm these results, we performed a permutation test using circular randomization that preserves the size and number of gene regions as well as their structure in genomic relationship to each other ($p< 10^{-4}$; permutation test; Figure S6A). We also removed possible confounders by excluding housekeeping (HK) genes,[28] setting a threshold for expression levels (Excluding genes with a median TPM less than 0.5 in all tissue), and restricting lincRNAs to those with experimentally proven functional roles.[29] Because most of the HK genes are PCGs and are known to be located at the TAD boundary,[24,25] exclusion of HK genes could have affected patterns of PCG localization and expression specificity. However, the absence of HK genes did not significantly affect the trends in genomic location and specificity patterns for each gene type ($p< 10^{-4}$; permutation test; Figure S6B). Filtering out low expression genes eliminated groups with particularly high tissue specificity scores (Figure S6E), which includes a subset of class 3 (Figures 1B–1E), genes with low expression levels and exclusive expression patterns. When restricted to lincRNAs with experimental functional annotations, permutation pvalue was higher than in the other conditions ($p = 5.0 \times 10^{-4}$; permutation test; Figure S6F) and there was no statistically significant difference in the tissue specificity scores (Figure S6G). It should be noted that the number of experimentally proven functional lincRNA genes is limited to a few hundred, and there may be an artificial bias that makes lincRNAs with large condition-specific expression variations, such as human disease, more likely to be selected for functional evaluation. In either condition, lincRNA loci were consistently localized to TADs and PCG loci were localized to TAD boundaries, with higher tissue specificity for gene groups in TADs.

We also examined the tissue specificity of gene expressions inside and outside the TAD regions. Because the broad genomic environment in testis, where dramatic chromatin architectural reorganization throughout spermatogenesis and temporary disappearance of TADs are observed,[30] does not ensure maintenance of TAD structure and regulation, we excluded the expression data of testis from the following analyses. Without testis expression data, lincRNA expression maintained higher tissue specificity than those of PCGs (Figure S7A). When lincRNAs and PCGs were divided into groups of those which were located inside and outside TADs, expressions of the groups included in TADs showed higher tissue specificities for both types of genes (Figure 2G). The relationship between the TAD localization and tissue specificity of lincRNA expressions was also examined for the three previously described classes (class 1, 2, and 3) of lincRNAs identified from their expression patterns. In class 1 and 2, which had relatively low overall tissue specificity, the expression of lincRNAs inside TADs showed significantly higher tissue specificity compared to those outside TADs (Figure S7B; class 1: $p = 5.078 \times 10^{-9}$, class 2: $p = 3.16 \times 10^{-3}$). The difference in the expression specificities between groups within and outside TADs was the largest in class 2. On the other hand, in class 3, which has the highest overall tissue specificity, there was no significant difference in expression specificities between lincRNAs inside and outside TADs. This implies that the direct contribution of localization in TADs to the tissue specificity is limited and that additional regulatory mechanisms are required to explain the increased tissue specificity. Comparison of the proportion of genes present in TADs between classes showed that class 2 had the highest ratio, class 3 had a ratio as high as class 2, and class 1 had the lowest ratio (Figure S7C; class 1: ratio = 0.594, class 2: ratio = 0.684, class 3: ratio = 0.676), suggesting that localization of loci to TADs possibly serves as a basis for downstream regulation of expression.

We examined the correlation of gene expression between lincRNAs and PCGs with intergenic distance and position in the genome structure, and found that lincRNA-PCG pairs belonging to the same TAD maintained a slightly higher correlation than other positional gene pairs at distances greater than 10K bp. Of interest, contrary to the expectation that gene pairs located at the TAD boundaries would inevitably have higher correlation coefficients because of their ubiquitous expression patterns, in-TAD gene pairs located a few hundred bp closer together tended to have higher correlation coefficients than those at the TAD boundaries (Figure 2H). The distribution of the distance from the loci of lincRNAs and PCGs inside TADs to the nearest TAD boundary, respectively, showed that lincRNAs were located farther from the boundary and more internal to TADs than PCGs ($p = 3.90 \times 10^{-134}$; Figure 2I).

Analytical framework for trait-specific transcriptional status of TAD units indicated by lincRNAs

Because the expression pattern of lincRNA is more cell- and tissue-specific than that of PCG, lincRNAs could be useful and sensitive biomarkers; their expression profiles can be used for identifying expression states specific to particular biological phenomena. Therefore, we proposed the following analysis scheme for biomedical applications (Figure S8). First, lincRNAs that vary specifically for a biological phenomenon of interest are identified with a differential expression (DE) analysis. Next, the *cis*- and *trans*-target genes of lincRNAs are analyzed. Based on the results above, genes that have loci in the same TAD and show similar expression patterns to lincRNAs are identified as *cis*-target genes. Because lincRNAs and *cis*-target genes are specifically regulated by regulatory sequences insulated within TADs, gene ontology (GO) term enrichment analysis are used to identify trait-specific functions common across the *cis*-target genes. For *trans*-target genes, under consideration of their functionality, genome-wide genes that have expression patterns negatively correlated with the lincRNAs are identified. lincRNAs often regulate *trans*-target genes with co-factors. Therefore, if there is a common upstream regulator of the *trans*-target genes, it is likely to be a co-factor of the lincRNA. Further information on the subcellular localization of the lincRNAs would enable identification of the upstream factors that interact directly with lincRNAs. Because lincRNAs are polyA-modified, albeit inefficiently,[31] they are detectable by standard RNA-Seq as well as PCGs. TADs are also maintained in most tissues and cells.[32] Therefore, this analysis framework is applicable to a wide range of samples profiled with conventional RNA sequencing methods.

lincRNA indicated transcriptional analysis highlighted HCM-specific transcriptional regulations

As one of biomedical applications of the above analysis scheme, we applied it with a dataset of expression profiles from HCM[33] to identify pathologically relevant molecular profiles common to patients with heterogeneous genetic backgrounds. We identified 6,095 DEGs in human HCM by using RNA-Seq data of myocardial tissues from healthy human volunteers and myocardial tissues from HCM patients with multiple genetic backgrounds (Table S1, FDR<0.05, Figure S9A). The results of GO analysis for all DEGs showed that GO enrichment related to translational processes was detected at a top level in HCM, with decreased expression of translation-related genes consistent with a recent report,[34] and GO terms such as mitochondria and energy production, which are important for myocardial function, also showed significant enrichment, indicating that the overall DEGs captured the characteristics of HCM (Figure S9B).

Next, we identified lincRNAs with differing expression levels in HCM. We found that 132 lincRNAs were significantly up-regulated and 140 lincRNAs were significantly down-regulated in the myocardium tissues from patients compared to those from healthy volunteers (Figure 3A). We found that 113 of these lincRNAs resided within 11 TADs with altered expression levels, indicating HCM-specific *cis*-association by lincRNAs (Figure 3B). We performed GO enrichment analysis of these 113 lincRNAs and DEGs in the same TADs according to three patterns: DEGs in all TADs with altered expression levels, DEGs in TADs with up-regulated expression levels, and DEGs in TADs with down-regulated expression levels. Significant GO enrichment was observed only for DEGs in TADs with down-regulated expression levels, which indicated a relevance of a GO category in biological pathways related to keratinization (Figures 3C, S10A, and S10B). Among the genes used in the GO analysis, keratin associated proteins were found to be downregulated in HCM (Figure 3D). On the other hand, we could observe that some genes of the keratin family were markedly up-regulated (Figures 3E and S10C). It has been reported that the K8/K18 pair is ectopically expressed downstream of TNF-α in a desmin-deficient mouse model of heart failure and exerts a cardioprotective effect in the myocardium after tissue injury, and that K8/K18 is also ectopically expressed in human failing myocardium,

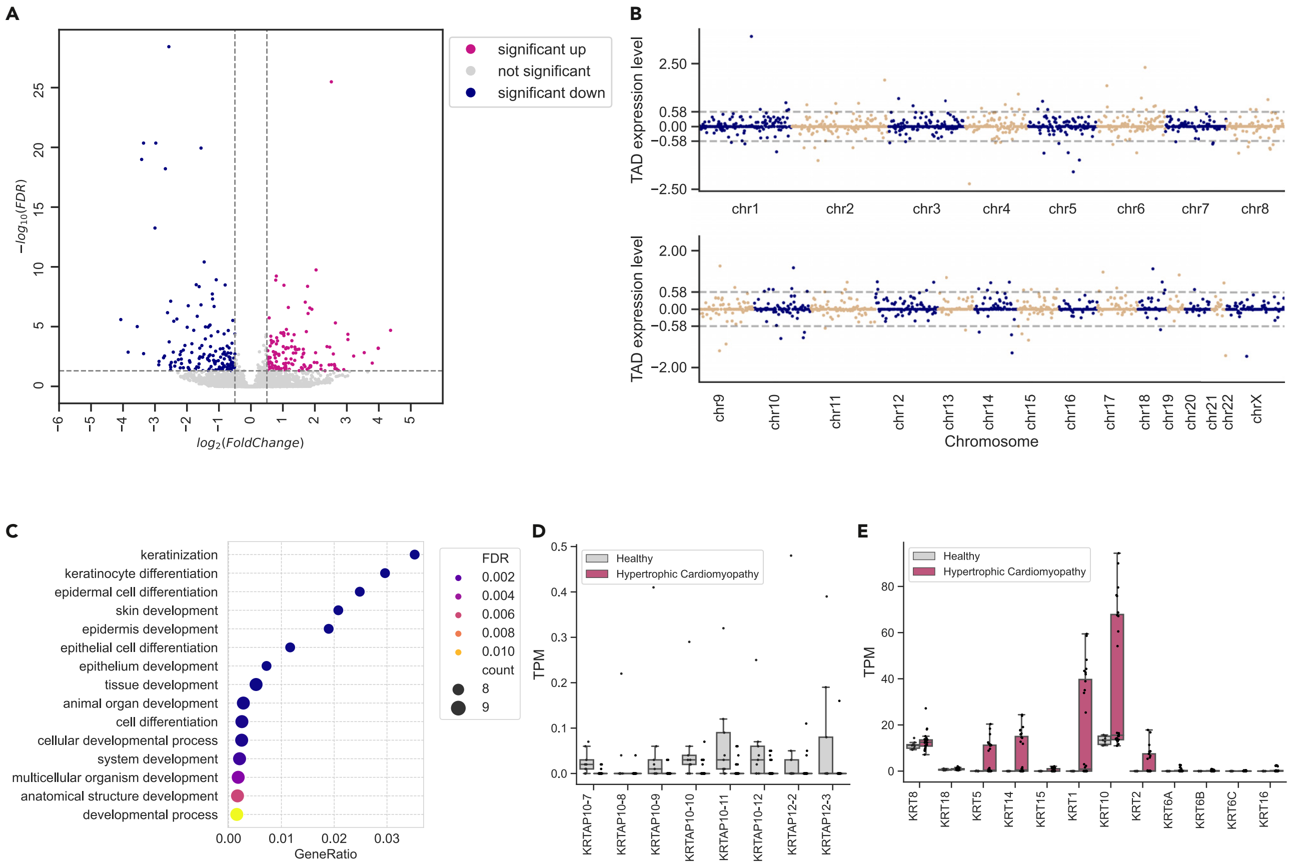


Figure 3. Analysis of TAD-related cis function of lincRNA

(A) Volcano plot showing differentially expressed lincRNAs between in HCM and healthy control tissues (log2 fold-change (FC) > 0.58 or log2 FC <-0.58, FDR <0.01).

(B) Dot plots showing the distribution of TAD expression levels and their chromosomal locations, that were calculated as the median of log2 FCs of all genes in the same TAD.

(C) Enriched biological processes based on DEGs in TADs with significantly lower expression levels (n = 9; FDR <0.05, Fisher's Exact test).

(D and E) Gene expression levels (TPMs) for (D) keratin associated proteins and (E) keratin family genes in healthy and HCM tissues.

where TNF-α expression is up-regulated.[35] Because the symptoms of HCM are also caused by mechanical stress because of abnormal physiological power production in sarcomeres,[36] it is possible that the ectopic expression of intermediate filaments acts to protect the myocardium. Ectopic expression of intermediate system filaments may have a protective effect on the myocardium. However, in HCM, we did not observe any changes in the expression of the K8/K18 pair as observed in the heart failure mouse model, and the activity of the upstream factor TNF was also suppressed compared to controls, whereas the expression of the K5/K14 and K1/K10 pairs was significantly upregulated (Figures 3D,3E, and S10C). This suggests that several different keratin pairs contribute to myocardial protection depending on the type of myocardial stress of different intensity and duration in humans.

Finally, we analyzed genome-wide *trans*-regulations of the post-transcriptional lincRNAs in HCM. To identify potential pairs of lincRNAs functional in HCM and their *trans*-target genes, we calculated the correlation of gene expression between pairs of 272 lincRNAs and 1930 PCGs, both with significantly altered expression levels (Figures 3A and S9A). Among the pairs with significantly correlated expression levels (absolute value of correlation coefficient >0.8), 303 pairs were positively correlated whereas 22 pairs were negatively correlated (Figure 4A). Furthermore, more than half of the negatively correlated pairs (15 pairs) were between PCGs and LINC00881. LINC00881 is known to be expressed specifically in cardiomyocytes after differentiation,[37] whereas it was significantly reduced in HCM (Figure 4B). Gene ontology analysis of the 15 PCGs that were negatively correlated with LINC00881 showed an enrichment of ontology terms

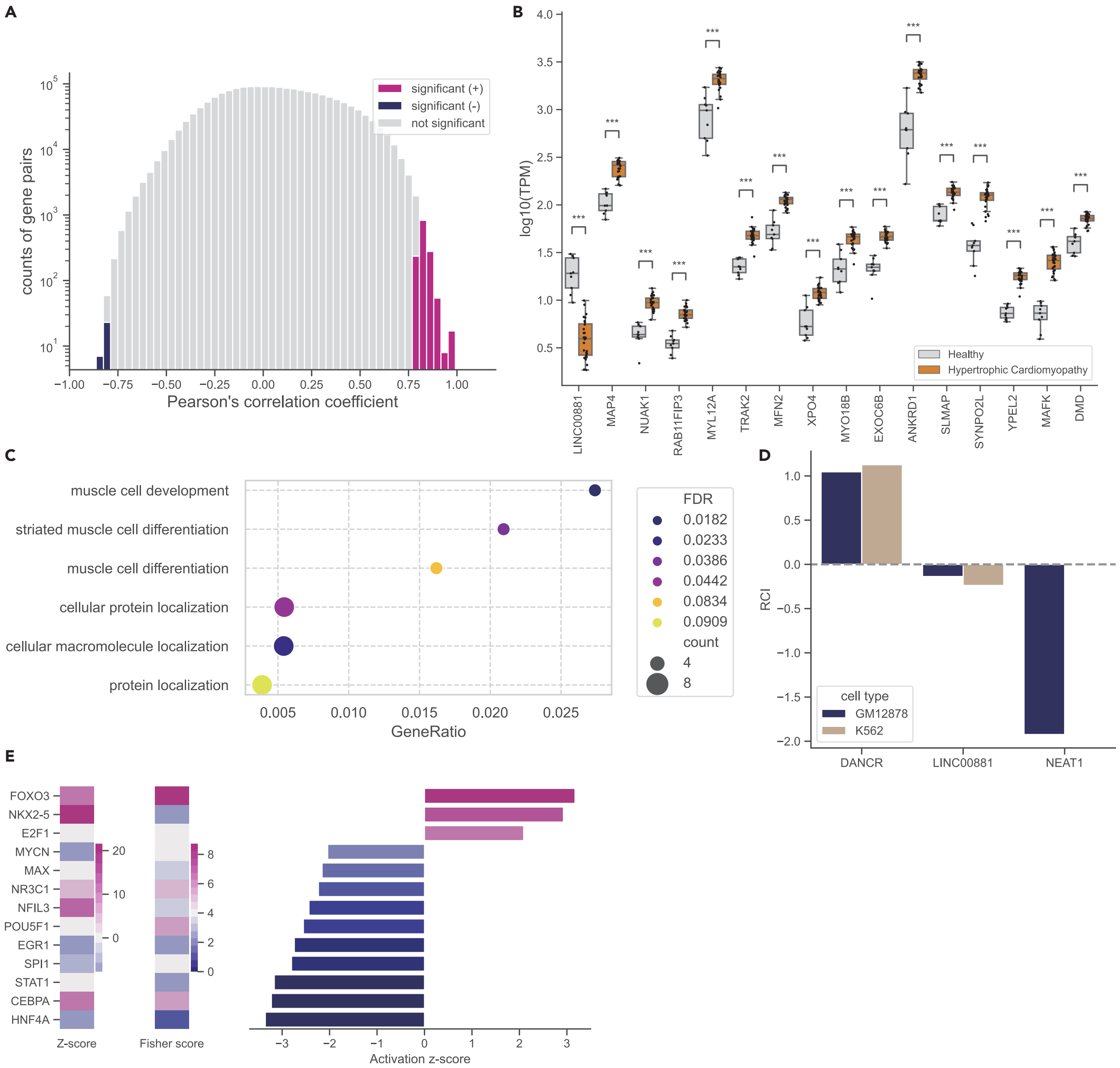


Figure 4. Analysis of *trans* function of lincRNA

(A) Distributions of the Pearson correlation coefficients between differentially expressed PCGs and lincRNAs.

(B) Gene expression levels of LINC00881 and the PCGs negatively correlated with LINC00881 in HCM and healthy tissues (***p< 0.001, Mann-Whitney-Wilcoxon test two-sided with Bonferroni correction).

(C) Enriched biological processes among PCGs negatively correlated with LINC00881 (n = 15; FDR <0.05, Fisher's Exact test).

(D) Bar plot representing CN-RCI values for DANCR (cytoplasmic RNA control), LINC00881 and NEAT1 (nuclear RNA control) across GM12878 and K562.

(E) Transcriptomic footprints inferred with oPOSSUM (left and center) and Ingenuity Pathway Analysis (IPA: right) for 13 transcription factors.

related to myocyte development (Figures 4B and 4C). According to lncATLAS, a database of subcellular localization of lncRNAs, LINC00881 is relatively localized in the nucleus, suggesting that it functions in the nucleus (Figure 4D). Because the 15 negatively correlated PCGs are located on different chromosomes and not near to the locus of LINC00881 (chromosome 3), it is highly likely that LINC00881 regulates these genes through nuclear factors. Therefore, we applied oPOSSUM (ver.3.0) to the promoter region sequence data to search for nuclear factors that commonly regulate the 15 PCGs (Table S2). In addition, to narrow down the number of plausible factors with different approaches, we used Ingenuity Pathway Analysis

(IPA) to search for nuclear factors that function upstream of the set of PCGs that vary in HCM, and 13 regulatory factors were commonly identified in these two analyses (Figure 4E). Of the three transcription factors with positive IPA Activation Zscore, E2F1 was shown to bind well to the promoter region of the target gene (Figures S11A–S11E). These results support the importance of E2F1 as a regulator of HCM traits and suggest that down-regulation of LINC00881 may be involved in the regulation of myocardial-specific pathways in HCM.

DISCUSSION

In this study, we demonstrated that lincRNA loci are significantly enriched within TAD regions and their localization is associated with tissue specificity at a genome-wide level. More specifically, PCG loci were enriched in TAD boundaries, whereas lincRNAs showed locus enrichment for TADs. lincRNAs enriched in TADs had higher tissue specificity than other lincRNAs or PCG, suggesting that TADs could be the basis for tissue-specific expression of lincRNAs. The increased tissue specificity associated with TAD localization complements the mechanisms of tissue-specific expression by promoters[15] and epigenomes[16] that have been investigated. Importantly, we validated the *cis* functions previously reported for individual lincRNA loci, such as ANRIL[38] and ANRASSF1,[39,40] as whole gene type properties and demonstrated its significance. As shown in Figure S7C, lincRNAs expressed in multiple tissues have increased tissue specificity according to TAD localization, and these lincRNAs' regulation or function can be associated with a sub-TAD state in which the structure within the TAD changes from tissue to tissue, whereas one-to-one recognition of transcription factors and regulatory sequences may be preferable for the highly specific lincRNA regulation. Often, the unknown function of lincRNAs has been inferred from the function of their neighboring genes because of the expression correlation between lincRNAs and their neighboring genes.[41] We show at the whole genome level that expression correlation is higher for genes in the same TAD than for gene pairs at other locations regardless of intergenic distance, as has been suggested so far.[26,27,42] Especially *cis*-association between lincRNAs and neighboring PCGs in TADs, which are highly tissue-specific genes, could suggest their functional relationship. Furthermore, among genes in TADs, lincRNAs may function as core TAD factors because lincRNAs are located farther from the boundary than PCGs. Recent study shows that lncRNA *Eleanor* activates ESR1 gene and neighboring genes in the same *Eleanor*-TAD in breast cancer cells.[43] Such lincRNA *cis*-activation at TAD level may occur in a broad genomic region. PCGs were significantly enriched at TAD boundaries with low expression specificity; McArthur and Capra reported that house-keeping genes are abundant at the highly conserved TAD boundary across multiple cell types,[44] which is consistent with our results.

In addition to lincRNAs and PCGs, we also examined enrichment in TADs for various gene types registered in GENCODE and found that the rate of enrichment varies among gene types. As this study shows, there is potential to gain new understanding of function and its regulation in many gene types from the perspective of genome structure. Future reevaluation of various genomic elements from the perspective of higher order structure can be expected to yield new insights into genomic function and evolution.

We further surmised that the identification of two groups of *cis*- and *trans*-regulated genes of lincRNAs would allow us to infer the mechanisms that produce differences in traits between samples. To demonstrate this, we analyzed data from human HCM to identify a group of genes associated with HCM-specific expression variation of lincRNAs and their upstream genes, which provided insight into the pathways relevant to HCM pathogenesis. HCM is often caused by mutations in sarcomere-related genes. However, there are many cases in which genetic abnormalities are not identified, and there is a need to understand the molecular mechanisms of the pathogenesis of HCM and to develop therapeutic methods. Here, based on our new approach and data from patients with multiple genetic backgrounds, we found elevated keratin gene expression and an HCM-specific transcriptional pathway by E2F1 with down-regulation of LINC00881. Ectopic expression of keratin in myocardium has been reported in studies of desmin-deficient mouse models of heart failure,[35] although little is known about its relevance in human cardiac disease. Although only expression of the keratin K8/K18 pair was detected in the mouse model, the human HCM samples showed little expression of K8/K18 and elevated expression of the K5/K14 and K1/K10 pairs instead. In mice, K8/K18 is thought to function as an alternative cytoskeleton to compensate for desmin deficiency, whereas what the differences in keratin pairing mean in human stressed myocardium has not yet been investigated. Future work is needed to clarify its precise background mechanisms.

In conclusion, this study demonstrated that, at a genome-wide level, the lincRNA loci were significantly localized to TADs compared to PCGs and that this localization was associated with the tissue-specificity of gene expression. Based on this, we proposed a new analysis framework for diseases/traits that are relevant to lincRNA expression, applied it to HCM data, and identified transcriptional regulations specific to HCM pathogenicity. This study will contribute to a fundamental understanding of the molecular basis underlying tissue-/disease-specific expression of lincRNAs.

Limitations of the study

It should be noted that there is some limitations of this study: it uses data from existing tissue-derived TADs as information on 3D structures, and it is not possible to refer to sub-TADs or local loop structures that are finer than information on a given TAD region. Although more TAD coordinates from primary human tissues are needed, there are often different algorithm-based tools used in Hi-C data analysis to identify TADs resulting in the introduction of considerable technical noise. This makes it challenging to ensure robustness, especially with multiple external TAD datasets. It is difficult to clearly identify molecular mechanisms by which lincRNA expression specificity is increased in TADs or that cause the expression correlation between lincRNA and coding gene pairs in TADs. Future accumulation of time-series expression data and experimental corroboration is awaited to elucidate those mechanisms. Furthermore, the relatively small sample size (n = 37) for the analysis of HCM-specific transcriptional states could potentially introduce bias, requiring further external validation and experimental corroboration.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - GTEx data processing
  - lincRNA clustering
  - Co-expression with neighboring genes
  - Tissue specificity
  - Topologically associating domains
  - RNA-seq quality control and data analyses of human HCM
  - GO enrichment analysis
  - Analysis of disease-specific expression status with lincRNA expression as an indicator
  - Identification of lincRNA candidates suppressing the target genes
  - Subcellular localization of *LINC00881*
  - Transcription factor analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.106640.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number JP19J22115.

AUTHOR CONTRIBUTIONS

Y.H. perceived the study, analyzed and interpreted the data, and wrote the manuscript. T.K. and F.M. reviewed the study and the manuscript and advised. K.B. contributed to the writing and revision of the manuscript. T.T. perceived and supervised the study, interpreted the data, and wrote the manuscript. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 14, 2022
Revised: December 15, 2022
Accepted: April 5, 2023
Published: April 26, 2023

REFERENCES

1. Kapranov, P., Cawley, S.E., Drenkow, J., Bekiranov, S., Strausberg, R.L., Fodor, S.P.A., and Gingeras, T.R. (2002). Large-scale transcriptional activity in chromosomes 21 and 22. Science *296*, 916–919. https://doi.org/10.1126/science.1068597.

2. Kapranov, P., Cheng, J., Dike, S., Nix, D.A., Duttagupta, R., Willingham, A.T., Stadler, P.F., Hertel, J., Hackermüller, J., Hofacker, I.L., et al. (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science *316*, 1484–1488. https://doi.org/10.1126/science.1138341.

3. Mattick, J.S., and Rinn, J.L. (2015). Discovery and annotation of long noncoding RNAs. Nat. Struct. Mol. Biol. *22*, 5–7. https://doi.org/10.1038/nsmb.2942.

4. Nagano, T., Mitchell, J.A., Sanz, L.A., Pauler, F.M., Ferguson-Smith, A.C., Feil, R., and Fraser, P. (2008). The air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. Science *322*, 1717–1720. https://doi.org/10.1126/science.1163802.

5. Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., and Chang, H.Y. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell *129*, 1311–1323. https://doi.org/10.1016/j.cell.2007.05.022.

6. Peng, X., Gralinski, L., Armour, C.D., Ferris, M.T., Thomas, M.J., Proll, S., Bradel-Tretheway, B.G., Korth, M.J., Castle, J.C., Biery, M.C., et al. (2010). Unique signatures of long noncoding RNA expression in response to virus infection and altered innate immune signaling. mBio *1*, e00206–e00210. https://doi.org/10.1128/mBio.00206-10.

7. Lai, F., Orom, U.A., Cesaroni, M., Beringer, M., Taatjes, D.J., Blobel, G.A., and Shiekhattar, R. (2013). Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. Nature *494*, 497–501. https://doi.org/10.1038/nature11884.

8. Mercer, T.R., and Mattick, J.S. (2013). Structure and function of long noncoding RNAs in epigenetic regulation. Nat. Struct. Mol. Biol. *20*, 300–307. https://doi.org/10.1038/nsmb.2480.

9. Ulitsky, I., and Bartel, D.P. (2013). lincRNAs: genomics, evolution, and mechanisms. Cell *154*, 26–46. https://doi.org/10.1016/j.cell.2013.06.020.

10. Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. *25*, 1915–1927. https://doi.org/10.1101/gad.17446611.

11. Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. *22*, 1775–1789. https://doi.org/10.1101/gr.132159.111.

12. Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223–227. https://doi.org/10.1038/nature07672.

13. Ravasi, T., Suzuki, H., Pang, K.C., Katayama, S., Furuno, M., Okunishi, R., Fukuda, S., Ru, K., Frith, M.C., Gongora, M.M., et al. (2006). Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. Genome Res. *16*, 11–19. https://doi.org/10.1101/gr.4200206.

14. Mercer, T.R., Dinger, M.E., Sunkin, S.M., Mehler, M.F., and Mattick, J.S. (2008). Specific expression of long noncoding RNAs in the mouse brain. Proc. Natl. Acad. Sci. USA *105*, 716–721. https://doi.org/10.1073/pnas.0706729105.

15. Mattioli, K., Volders, P.J., Gerhardinger, C., Lee, J.C., Maass, P.G., Melé, M., and Rinn, J.L. (2019). High-throughput functional analysis of lncRNA core promoters elucidates rules governing tissue specificity. Genome Res. *29*, 344–355. https://doi.org/10.1101/gr.242222.118.

16. Marques, A.C., Hughes, J., Graham, B., Kowalczyk, M.S., Higgs, D.R., and Ponting, C.P. (2013). Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. Genome Biol. *14*, R131. https://doi.org/10.1186/gb-2013-14-11-r131.

17. Bonev, B., and Cavalli, G. (2016). Organization and function of the 3D genome. Nat. Rev. Genet. *17*, 661–678. https://doi.org/10.1038/nrg.2016.112.

18. Gibcus, J.H., and Dekker, J. (2013). The hierarchy of the 3D genome. Mol. Cell *49*, 773–782. https://doi.org/10.1016/j.molcel.2013.02.011.

19. Kopp, F., and Mendell, J.T. (2018). Functional classification and experimental dissection of long noncoding RNAs. Cell *172*, 393–407. https://doi.org/10.1016/j.cell.2018.01.011.

20. GTEx Consortium, Laboratory Data Analysis &Coordinating Center LDACC—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx eGTEx groups, NIH Common Fund, Jo, B., Mohammadi, P., Park, Y.S., Parsana, P., Biospecimen Collection Source Site—NDRI, et al.. (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213. https://doi.org/10.1038/nature24277.

21. Molyneaux, B.J., Goff, L.A., Brettler, A.C., Chen, H.-H., Hrvatin, S., Rinn, J.L., Arlotta, P., and Arlotta, P. (2015). DeCoN: genome-wide analysis of in vivo transcriptional dynamics during pyramidal neuron fate selection in neocortex. Neuron *85*, 275–288. https://doi.org/10.1016/j.neuron.2014.12.024.

22. Schmitt, A.D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C.L., Li, Y., Lin, S., Lin, Y., Barr, C.L., and Ren, B. (2016). A compendium of chromatin contact maps reveals spatially active regions in the human genome. Cell Rep. *17*, 2042–2059. https://doi.org/10.1016/j.celrep.2016.10.061.

23. Krefting, J., Andrade-Navarro, M.A., and Ibn-Salem, J. (2018). Evolutionary stability of topologically associating domains is associated with conserved gene regulation. BMC Biol. *16*, 87. https://doi.org/10.1186/s12915-018-0556-x.

24. Dixon, J.R., Gorkin, D.U., and Ren, B. (2016). Chromatin domains: the unit of chromosome organization. Mol. Cell *62*, 668–680. https://doi.org/10.1016/j.molcel.2016.05.018.

25. Li, L., Lyu, X., Hou, C., Takenaka, N., Nguyen, H.Q., Ong, C.T., Cubeñas-Potts, C., Hu, M., Lei, E.P., Bosco, G., et al. (2015). Widespread rearrangement of 3D chromatin organization underlies polycomb-mediated stress-induced silencing. Mol. Cell *58*, 216–231. https://doi.org/10.1016/j.molcel.2015.02.023.

26. Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., Van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385. https://doi.org/10.1038/nature11049.

27. Flavahan, W.A., Drier, Y., Liau, B.B., Gillespie, S.M., Venteicher, A.S., Stemmer-Rachamimov, A.O., Suvà, M.L., and Bernstein, B.E. (2016). Insulator dysfunction and oncogene activation in IDH mutant gliomas. Nature *529*, 110–114. https://doi.org/10.1038/nature16490.

28. Eisenberg, E., and Levanon, E.Y. (2013). Human housekeeping genes, revisited. Trends Genet. *29*, 569–574. https://doi.org/10.1016/j.tig.2013.05.010.

29. Zhao, H., Shi, J., Zhang, Y., Xie, A., Yu, L., Zhang, C., Lei, J., Xu, H., Leng, Z., Li, T., et al. (2020). LncTarD: a manually-curated database of experimentally-supported functional lncRNA-target regulations in human diseases. Nucleic Acids Res. *48*, D118–D126. https://doi.org/10.1093/nar/gkz985.

30. Luo, Z., Wang, X., Jiang, H., Wang, R., Chen, J., Chen, Y., Xu, Q., Cao, J., Gong, X., Wu, J., et al. (2020). Reorganized 3D genome structures support transcriptional regulation in mouse spermatogenesis. iScience *23*, 101034. https://doi.org/10.1016/j.isci.2020.101034.

31. Schlackow, M., Nojima, T., Gomes, T., Dhir, A., Carmo-Fonseca, M., and Proudfoot, N.J. (2017). Distinctive patterns of transcription and RNA processing for human lincRNAs. Mol. Cell *65*, 25–38. https://doi.org/10.1016/j.molcel.2016.11.029.

32. Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., and Aiden, E.L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680. https://doi.org/10.1016/j.cell.2014.11.021.

33. Liu, X., Ma, Y., Yin, K., Li, W., Chen, W., Zhang, Y., Zhu, C., Li, T., Han, B., Liu, X., et al. (2019). Long non-coding and coding RNA profiling using strand-specific RNA-seq in human hypertrophic cardiomyopathy. Sci. Data *6*, 90. https://doi.org/10.1038/s41597-019-0094-6.

34. Li, N., Wu, H., Geng, R., and Tang, Q. (2018). Identification of core gene biomarkers in patients with diabetic cardiomyopathy. Dis. Markers *2018*, 6025061. https://doi.org/10.1155/2018/6025061.

35. Papathanasiou, S., Rickelt, S., Soriano, M.E., Schips, T.G., Maier, H.J., Davos, C.H., Varela, A., Kaklamanis, L., Mann, D.L., Capetanaki, Y., et al. (2015). Tumor necrosis factor-α confers cardioprotection through ectopic expression of keratins K8 and K18. Nat. Med. *21*, 1076–1084. https://doi.org/10.1038/nm.3925.

36. Green, E.M., Wakimoto, H., Anderson, R.L., Evanchik, M.J., Gorham, J.M., Harrison, B.C., Henze, M., Kawas, R., Oslob, J.D., Rodriguez, H.M., et al. (2016). A small-molecule inhibitor of sarcomere contractility suppresses hypertrophic cardiomyopathy in mice. Science *351*, 617–621. https://doi.org/10.1126/science.aad3456.

37. Li, Y., Lin, B., and Yang, L. (2015). Comparative transcriptomic analysis of multiple cardiovascular fates from embryonic stem cells predicts novel regulators in human cardiogenesis. Sci. Rep. *5*, 9758. https://doi.org/10.1038/srep09758.

38. Yap, K.L., Li, S., Muñoz-Cabello, A.M., Raguz, S., Zeng, L., Mujtaba, S., Gil, J., Walsh, M.J., and Zhou, M.M. (2010). Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. Mol. Cell *38*, 662–674. https://doi.org/10.1016/j.molcel.2010.03.021.

39. Beckedorff, F.C., Ayupe, A.C., Crocci-Souza, R., Amaral, M.S., Nakaya, H.I., Soltys, D.T., Menck, C.F.M., Reis, E.M., and Verjovski-Almeida, S. (2013). The intronic long noncoding RNA ANRASSF1 recruits PRC2 to the RASSF1A promoter, reducing the expression of RASSF1A and increasing cell proliferation. PLoS Genet. *9*, e1003705. https://doi.org/10.1371/journal.pgen.1003705.

40. Alecki, C., Chiwara, V., Sanz, L.A., Grau, D., Arias Pérez, O., Boulier, E.L., Armache, K.J., Chédin, F., and Francis, N.J. (2020). RNA-DNA strand exchange by the Drosophila Polycomb complex PRC2. Nat. Commun. *11*, 1781. https://doi.org/10.1038/s41467-020-15609-x.

41. Zhang, Z., Zhang, J., Fan, C., Tang, Y., and Deng, L. (2019). KATZLGO: large-scale prediction of LncRNA functions by using the KATZ measure based on multiple networks. IEEE ACM Trans. Comput. Biol. Bioinf *16*, 407–416. https://doi.org/10.1109/TCBB.2017.2704587.

42. Sarnataro, S., Riba, A., and Molina, N. (2021). Regulation of transcription reactivation dynamics exiting mitosis. PLoS Comput. Biol. *17*, e1009354. https://doi.org/10.1371/journal.pcbi.1009354.

43. Abdalla, M.O.A., Yamamoto, T., Maehara, K., Nogami, J., Ohkawa, Y., Miura, H., Poonperm, R., Hiratani, I., Nakayama, H., Nakao, M., and Saitoh, N. (2019). The Eleanor ncRNAs activate the topological domain of the ESR1 locus to balance against apoptosis. Nat. Commun. *10*, 3778. https://doi.org/10.1038/s41467-019-11378-4.

44. McArthur, E., and Capra, J.A. (2021). Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched for heritability. Am. J. Hum. Genet. *108*, 269–283. https://doi.org/10.1016/j.ajhg.2021.01.001.

45. Mas-Ponte, D., Carlevaro-Fita, J., Palumbo, E., Hermoso Pulido, T., Guigo, R., and Johnson, R. (2017). LncATLAS database for subcellular localization of long noncoding RNAs. RNA *23*, 1080–1087. https://doi.org/10.1261/rna.060814.117.

46. Zou, Z., Ohta, T., Miura, F., and Oki, S. (2022). ChIP-Atlas 2021 update: a data-mining suite for exploring epigenomic landscapes by fully integrating ChIP-seq, ATAC-seq and Bisulfite-seq data. Nucleic Acids Res. *50*, W175–W182. https://doi.org/10.1093/nar/gkac199.

47. Camargo, A.P., Vasconcelos, A.A., Fiamenghi, M.B., Pereira, G.A.G., and Carazzolle, M.F. (2020). tspex: a tissue-specificity calculator for gene expression data. Preprint at Research Square. https://doi.org/10.21203/rs.3.rs-51998/v1.

48. Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics *32*, 3047–3048. https://doi.org/10.1093/bioinformatics/btw354.

49. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

50. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

51. Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinf. *12*, 323. https://doi.org/10.1186/1471-2105-12-323.

52. Soneson, C., Love, M.I., and Robinson, M.D. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Res. *4*, 1521. https://doi.org/10.12688/f1000research.7563.2.

53. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.

54. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

55. Kwon, A.T., Arenillas, D.J., Worsley Hunt, R., and Wasserman, W.W. (2012). oPOSSUM-3: advanced analysis of regulatory motif over-representation across genes or ChIP-Seq datasets. G3 *2*, 987–1002. https://doi.org/10.1534/g3.112.003202.

56. Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A. (2003). PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. *13*, 2129–2141. https://doi.org/10.1101/gr.772403.

57. Gel, B., Díez-Villanueva, A., Serra, E., Buschbeck, M., Peinado, M.A., and Malinverni, R. (2016). regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. Bioinformatics *32*, 289–291. https://doi.org/10.1093/bioinformatics/btv562.

58. Oki, S., Ohta, T., Shioi, G., Hatanaka, H., Ogasawara, O., Okuda, Y., Kawaji, H., Nakaki, R., Sese, J., and Meno, C. (2018). ChIP-Atlas: a data-mining suite powered by full integration

of public ChIP-seq data. EMBO Rep. *19*, e46255. https://doi.org/10.15252/embr.201846255.

59. McKinney, W. (2010). Data structures for statistical computing in python *445*, 51–56. https://doi.org/10.25080/Majora-92bf1922-00a.

60. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

61. Virtanen, P., Gommers, R., Oliphant, T.E., 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. https://doi.org/10.1038/s41592-019-0686-2.

62. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al. (2020). Array programming with NumPy. Nature *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2.

63. Waskom M.L. (2021). seaborn: statistical data visualization. J. Open Source Softw., 3021. https://doi.org/10.21105/joss.03021.

64. Hunter J.D. (2007). Matplotlib: A 2D graphics environment. Computing in science & engineering 9, 90–95. https://doi.org/10.1109/MCSE.2007.55.

65. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat Biotechnol. *29*, 24–26. https://doi.org/10.1038/nbt.1754.

66. Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

67. Weber M. statannot. 2019. Available online at: https://github.com/webermarcolivier/statannot

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Human reference genome GENCODE Release 25	GENCODE project	https://www.gencodegenes.org
GTEx v8 release gene expression data	GTEx v8	https://gtexportal.org/home/datasets
TAD coordinates on hg38	3D Genome Browser	http://3dgenome.fsm.northwestern.edu/publications.html
A list of housekeeping genes	E. Eisenberg and E.Y. Levanon[28]	http://www.tau.ac.il/~elieis/HKG/
Experimentally supported functional lncRNA data	LncTarD 2.0[29]	https://lnctard.bio-database.com
lncRNA subcellular localization data	LncATLAS[45]	lncatlas.crg.eu
ChIP sequencing data	ChIP-Atlas[45,46]	https://chip-atlas.org
Raw RNA sequencing data of HCM and control myocardial tissues	Liu et al.[33]	GEO: GSE130036
Software and algorithms		
Python version 3.8.0	Python Software Foundation	https://www.python.org
pandas version 0.25.3	W. McKinney et al.[59]	https://doi.org/10.5281/zenodo.3509134
scikit-learn version 0.22	F. Pedregosa et al.[60]	https://scikit-learn.org
scipy version 1.5.2	P. Virtanen et al.[61]	https://scipy.org
numpy version 1.17.4	C. R. Harris et al.[62]	https://numpy.org
statannot version 0.2.3	M. Weber[67]	https://github.com/webermarcolivier/statannot
tspex version 0.6.2	Camargo et al.[47]	https://doi.org/10.21203/rs.3.rs-51998/v1
seaborn version 0.11.0	M. L. Waskom[63]	https://doi.org/10.21105/joss.03021
matplotlib version 3.1.2	J. D. Hunter[64]	https://doi.org/10.1109/MCSE.2007.55
R version 3.6.2	R Foundation for Statistical Computing	https://www.r-project.org
FastQC version 0.11.8	S. Andrews[66]	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
MultiQC version 1.9	Ewels et al.[48]	https://doi.org/10.1093/bioinformatics/btw354
fastp version 0.20.1	Chen et al.[49]	https://doi.org/10.1093/bioinformatics/bty560
STAR version 2.7.3a	Dobin et al.[50]	https://doi.org/10.1093/bioinformatics/bts635
RSEM version 1.3.1	Li and Dewey[51]	https://doi.org/10.1186/1471-2105-12-323
Tximport	Soneson et al.[52]	https://doi.org/10.12688/f1000research.7563.1
DESeq2	Love et al.[53]	https://doi.org/10.1186/s13059-014-0550-8
Bedtools version 2.29.2	Quinlan and Hall[54]	https://doi.org/10.1093/bioinformatics/btq033
IGV version 2.16.0	J. Robinson et al.[65]	https://doi.org/10.1038/nbt.1754
oPOSSUM version 3.0	Kwon et al.[55]	https://doi.org/10.1534/g3.112.003202
Ingenuity Pathway Analysis	QIAGEN Inc.	https://digitalinsights.qiagen.com/IPA
PANTHER	Thomas et al.[56]	http://www.pantherdb.org

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and data should be directed to and fulfilled by the Lead Contact, Tatsuhiko Tsunoda (tsunoda@bs.s.u-tokyo.ac.jp).

Materials availability

This study did not generate new unique materials.

Data and code availability

All data analyzed during this study were retrieved from public repositories. GTEx v8 release gene expression data are available through the GTEx Portal [https://gtexportal.org/home/datasets]. TAD coordinates on hg38 estimated from human Hi-C data are available through 3D Genome Browser [http://3dgenome.fsm.northwestern.edu/publications.html]. A list of housekeeping genes is available at [http://www.tau.ac.il/∼elieis/HKG/]. Experimentally supported functional lncRNA data are available through LncTarD 2.0 Browser [https://lnctard.bio-database.com]. ChIP sequencing data are available through ChIP-Atlas [https://chip-atlas.org]. Raw RNA sequencing data of HCM and control myocardial tissues are available through NCBI Sequence Read Archive (SRA) under accession GSE130036.[33] Code used for all processing and analysis is available upon request. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

GTEx data processing

GTEx expression data (gene transcripts per kilobase million (TPM)) used in this study was downloaded from the GTEx portal [https://gtexportal.org/home/datasets; accessed 12.20.2019, file ''GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_tpm.gct.gz''], which contained 17,382 postmortem samples. The editing level was called on the GTEx v8p release (study accession phs000424.v8.p2).

Based on gene type classification of GENCODE Release 25 annotation file, expression values of 7345 lincRNA and 18358 protein-coding genes (PCGs) were extracted from GTEx data and clustering was performed for each gene types. For sample clustering, we performed a t-SNE and a Euclidean distance-based clustering with scikit-learn (v.0.22) using the following parameters: sklearn.manifold.TSNE(n_components = 2, init = 'random', random_state = 101, method = 'barnes_hut', n_iter = 1000, verbose = 2, perplexity = 40).fit_transform(sample_expression_data).

lincRNA clustering

Clustering based on lincRNA expression patterns across tissued was performed using scikit-learn (v.0.22) agglomerative clustering method with default parameters. Median gene-level TPM by tissue was log2-transformed and clustered into three classes to avoid imbalances in the number of genes per cluster. Gene expressions based on the three clusters determined were plotted on a heatmap.

Co-expression with neighboring genes

All gene loci were annotated with GENCODE Release 25 annotation file. The basic gene annotation file was converted to BED file format using a custom bash script. Neighboring genes of lincRNAs were identified by bedtools (v2.29.2)[54] using the distance between gene bodies. Pearson's correlation coefficients for expression correlations were calculated between adjacent genes with the GTEx expression data.

Tissue specificity

To assess the specificity of gene expression across all tissues, an entropy-based specificity score was calculated using the Jensen-Shannon specificity variance metric (JSS DPM) in the tspex package.[47] Using GTEx expression data, the score for each gene was quantified as a continuous value between 0 (ubiquitous expression) and 1 (tissue-specific expression).

Topologically associating domains

We downloaded and used TAD coordinates on hg38 estimated from human Hi-C data from 3D Genome Browser [http://3dgenome.fsm.northwestern.edu/publications.html; accessed 02.09.2020, file ''hg38'']. For 13 primary tissues (adrenal gland, aorta, bladder, small bowel, prefrontal cortex, hippocampus, liver, lung, ovary, pancreas, psoas muscle, spleen and left ventricle).

For TADs from each of the primary tissues, we identified all genes that reside within TADs for lincRNA, protein coding, antisense, processed pseudogene, transcribed processed pseudogene, processed transcript, sense intronic, To be Experimentally Confirmed (TEC), unprocessed pseudogene, transcribed unprocessed pseudogene, unitary pseudogene, transcribed unitary pseudogene, polymorphic pseudogene, miRNA, snRNA, snoRNA, rRNA and miscRNA. Only those genes of which the entire gene body was contained within TAD regions were considered. We used a permutation scheme to obtain randomized distribution of gene positions for evaluating whether or not loci of each gene type are enriched within TADs: the gene loci were re-shuffled (10,000 permutations genome-wide) to count the number of genes in TADs and to estimate permutation p-values. Each permutation consists of the same TAD coordinates and gene sets, in which TAD regions were kept whereas the gene positions were randomly shuffled using the shuffle utility of bedtools (v.2.29.2).[54] This was repeated for each of the 13 tissues.

To obtain a list of common TAD regions, TAD coordinates which overlapped across all the 13 tissues were determined using the intersect utility of bedtools (v.2.29.2).[54] The resulting list of 2,326 common regions were used as common TADs for the statistical framework described below.

Genes contained in the common TADs were identified using the intersect utility of bedtools (v.2.29.2).[54] To confirm the insulating effect of TADs on gene expression, we examined the moving average of the expression correlations according to the intergenic distance for gene pairs with different positions in relation to the TAD and the correlation of TAD expression levels for each chromosome. Pearson's correlation coefficients were calculated for gene expression values in GTEx, and the intergenic distances were calculated with GENCODE Release 25 annotation file. Moving averages were used to examine trends in expression correlations and smooth out short-term variation by calculating the average of the data over an interval of log10(bp+1) intergenic distance. Confidence intervals were calculated from standard error of correlation coefficients within the sliding window. To quantify the expression levels of each TAD region, GTEx TPM data were log2-transformed and standardized, and median expression levels were calculated among genes included in each TAD. Pearson's correlation coefficients were calculated for expression levels between TADs. For all lincRNAs and PCGs, in addition to the permutation tests and tissue specificity score calculations described above, a circular randomized permutation test was performed (10,000 permutations chromosome-wide) that preserves the size and number of gene regions plus their structure in genomic relationship to each other, using the regionR package (v.1.30.0)[57] and R (v.3.6.2). Permutation tests and tissue specificity scores described above were performed for lincRNAs and PCGs under the following conditions: exclusion of HK genes,[28] restriction to lincRNAs with experimentally proven functional roles,[29] setting a threshold for expression levels (median TPM more than 0.5 at least in one tissue).

In addition to the position patterns of lincRNAs and PCGs within and outside TADs, we compared the distribution of the shortest distances between genes and TAD boundaries in order to examine their distribution within TADs. The distance from each lincRNA and PCG within a TAD to the nearest TAD boundary was calculated using the distanceToNearest utility of regionR packages (v.1.30.0)[57] and R(v.3.6.2).

RNA-seq quality control and data analyses of human HCM

Publicly available RNA-Seq datasets with accession code GSE130036,[33] which contained 37 fastq files (myocardial tissues from 28 HCM patients and 9 healthy donors) were downloaded from the NCBI GEO database.

Quality control of the reads was performed with FastQC (v.0.11.8) and MultiQC (v.1.9).[48] The reads were trimmed with fastp (v.0.20.1)[49] and aligned to the human hg38 reference (GENCODE Release 25 basic gene annotation) using STAR (v.2.7.3a)[50] with default parameters. Quantification of gene expression was performed using RSEM (v.1.3.1)[51] with default parameters. The raw count matrix containing all the samples was created using tximport package (v.1.14.2)[52] and R (v.3.6.2). Normalization and differential expression analysis were carried out using DESeq2 package (v1.26.0)[53] and R (v.3.6.2). Difference of gene expression was considered significant if their log2 fold change (HCM/control) was >0.58 or <-0.58 and its false discovery rate (FDR) was <0.05 upon correction for multiple testing with the Benjamini–Hochberg method.

GO enrichment analysis

For determining DEGs between myocardial tissues of HCM patients and healthy donors, after confirming that 41567 genes with observed background expression (TPM >0 in at least one sample) covered 20595 reference genes registered in PANTHER, GO term enrichment analysis was performed using the

PANTHER Overrepresentation Test (release 02.24.2021) and the GO ontology database (release 02.01.2021).[56]

Analysis of disease-specific expression status with lincRNA expression as an indicator

We calculated median log2 fold change of genes within each TAD that had at least one differentially expressed lincRNA according to the common TAD coordinates, setting the log2 fold change of genes with FDR ≥0.05 to 0. TADs with median log2 fold change >0.58 or <-0.58 for DEGs within them were defined as having significant differences in expression levels. We performed GO analysis on the genes included in TADs with significant differences in expression levels according to three gene sets: all genes, up-regulated genes, and down-regulated genes.

Identification of lincRNA candidates suppressing the target genes

Correlations of expression levels between differentially expressed lincRNAs and PCGs from all 37 samples were calculated. For PCG and lincRNA expression levels, we used normalized TPM values. These normalized values were obtained with RSEM (v.1.3.1).[51] Only gene pairs with Pearson's correlation coefficient >0.8 were considered for further study.

Subcellular localization of *LINC00881*

To characterize the molecular function of *LINC00881*, we ascertained its subcellular localization using LncATLAS database (lncatlas.crg.eu).[45] LncATLAS provides relative concentration index (RCI) to define and quantify RNA localization, which is defined as the log2-transformed ratio of FPKM (fragments per kilobase per million mapped) in the cytoplasm vs. nucleus.

Transcription factor analysis

To identify transcription factors predicted to be involved in regulatory function of *LINC00881*, we searched over-representation of predicted transcription factor binding sites (TFBS) in gene promoters and proximal 5′ regulatory regions (±5,000 bp from TSS) by Single Site Analysis (SSA) of oPOSSUM ver.3.0.[55] The CORE collection of JASPAR database was used as the source of DNA binding profiles. A set of 15 coding genes negatively correlated with *LINC00881* were input to SSA to obtain Z-scores and Fisher scores for each predicted transcription factor. To functionally filter the sequence-based predicted transcription factors, we selected nuclear factors involved in transcriptional regulation from those predicted to be upstream regulators (categories: "transcriptional regulator" and "ligand-dependent nuclear receptor") by the use of Ingenuity Pathway Analysis (IPA, Qiagen). Z-Score (ZS), used to calculate the predicted activation state from IPA, was used to infer likely activation states of upstream regulators based on comparison with a model that assigns random regulation directions. Only those hits found in both analyses were considered for the overlap.

To support the ability of the identified co-factor candidates to actually regulate their target genes, human ChIP-seq data registered in ChIP-Atlas[46,58] for 6 genes (OXO3, NKX2-5, E2F1, MYCN, MAX and NR3C1) were specified and the corresponding results were streamed to IGV (v.2.16.0) to obtain snapshots.

QUANTIFICATION AND STATISTICAL ANALYSIS

Student's t-test was used to determine the percentage of each gene type in the TAD. Mann-Whitney-Wilcoxon test two-sided with Bonferroni correction was used to compare gene specificity between the two groups. ns: $P > 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 10^{-4}$, Student's t-test and Mann-Whitney-Wilcoxon test. $P < 0.05$ and FDR <0.05 were considered to be statistically significant. Unless otherwise stated, all statistical analyses were performed in Python 3.8.0 statistical environment, using the packages pandas (v.0.25.3), scikit-learn (v.0.22), scipy (v.1.5.2), numpy (v.1.17.4), statannot (v.0.2.3) and tspex (v.0.6.2) and all plots were generated in python using the packages seaborn (v.0.11.0) and matplotlib (v.3.1.2).