Functional annotation of human long noncoding RNAs using chromatin conformation data

Saumya Agrawal[1], Tanvir Alam[2], Masaru Koido[1,3], Ivan V. Kulakovskiy[4,5], Jessica Severin[1], Imad Abugessaisa[1], Andrey Buyan[5,6], Josee Dostie[7], Masayoshi Itoh[1,8], Naoto Kondo[9], Yunjing Li[10], Mickaël Mendez[11], Jordan A. Ramilowski[1,12], Ken Yagi[1], Kayoko Yasuzawa[1], Chi Wai Yip[1], Yasushi Okazaki[1], Michael M. Hoffman[11,13,14,15], Lisa Strug[10], Chung Chau Hon[1], Chikashi Terao[1], Takeya Kasukawa[1], Vsevolod J. Makeev[4,16], Jay W. Shin[1], Piero Carninci[1], Michiel JL de Hoon[1]

[1]RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [2]College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar. [3]Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [4]Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia. [5]Institute of Protein Research, Russian Academy of Sciences, Pushchino, Russia. [6]Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia. [7]Department of Biochemistry, Rosalind and Morris Goodman Cancer Research Center, McGill University, Montréal, Québec, Canada. [8]RIKEN Preventive Medicine and Diagnosis Innovation Program, Wako, Japan. [9]RIKEN Center for Life Science Technologies, Yokohama, Japan. [10]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada. [11]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. [12]Advanced Medical Research Center, Yokohama City University, Yokohama, Japan. [13]Princess Margaret Cancer Centre, Toronto, Ontario, Canada. [14]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [15]Vector Institute, Toronto, Ontario, Canada. [16]Moscow Institute of Physics and Technology, Dolgoprudny, Russia.

Abstract

Transcription of the human genome yields mostly long non-coding RNAs (lncRNAs). Systematic functional annotation of lncRNAs is challenging due to their low expression level, cell type-specific occurrence, poor sequence conservation between orthologs, and lack of information about RNA domains. Currently, 95% of human lncRNAs have no functional characterization. Using chromatin conformation and Cap Analysis of Gene Expression (CAGE) data in 18 human cell types, we systematically located genomic regions in spatial proximity to lncRNA genes and identified functional clusters of interacting protein-coding genes, lncRNAs and enhancers. Using these clusters we provide a cell type-specific functional annotation for 7,651 out of 14,198 (53.88%) lncRNAs. LncRNAs tend to have specialized roles in the cell type in which it is first expressed, and to incorporate more general functions as its expression is acquired by multiple cell types during evolution. By analyzing RNA-binding protein and RNA-chromatin interaction data in the context of the spatial genomic interaction map, we explored mechanisms by which these lncRNAs can act.

Introduction

Human cells express tens of thousands of long non-coding RNAs (lncRNAs) [1]. LncRNAs are defined as RNA transcripts of at least 200 nt with no or limited protein-coding potential. Though this class of RNAs has been known for almost 50 years [2–4] , ~95% of lncRNAs still lack functional annotation [5]. While some lncRNAs have defined important roles in transcriptional regulation [6,7], chromatin maintenance [8,9], translation [10], and other biological processes, annotating the function of lncRNAs systematically is challenging due to their low expression [11], rapid degradation compared to mRNAs [12,13], high cell type-specificity [11], and lack of conservation across organisms [14]. In contrast to protein-coding genes, the absence of families of lncRNAs with related sequences greatly hinders systematic inference of lncRNA function.

LncRNAs can regulate the expression of genes that are up to several megabases away in linear genomic distance, as distal regions can be brought into spatial proximity by chromatin folding [15–17]. One can examine these three-dimensional relationships in the nucleus using Hi-C, a chromosome conformation capture technique [18,19] that involves cross-linking spatially proximal DNA segments, followed by ligation and sequencing. Hi-C can identify interacting genomic regions at 5 kbp - 10 kbp resolution depending on the protocol and sequencing depth, and has been used to study spatially proximal genomic regions in several cell types [19–26].

Here, we present a high-resolution genomic interaction map to systematically annotate the biological role of lncRNAs using one newly produced and 17 publicly available deep sequenced Hi-C data from 18 human cell types and tissues. We used Cap Analysis Gene Expression (CAGE; [27]) data to identify expressed genes and their precise transcription start sites (TSSs), as well as FANTOM5 bi-directionally transcribed enhancers [28]. At CAGE-identified promoters, we used histone mark enrichment to distinguish between canonical (H3K4me3-enriched) promoters and enhancer-like (H3K4me-enriched) promoters. Using the map of spatially close genomic regions and the CAGE data, we identified potential interacting partners of 14,198 lncRNAs across the 18 cell types and tissues. These partners included mRNA genes, FANTOM5 enhancers, and other lncRNAs (intergenic and non-intergenic lncRNAs). Based on the genes associated in 3D space, we have annotated the lncRNA properties and biological role of 7,651/14,198 lncRNAs (53.88%). We provide a visualization platform that allows users to browse and compare the functional annotations for each lncRNA in individual cell types (https://fantom.gsc.riken.jp/zenbu/reports/#FANTOM6_HiC).

Results

Data processing for 18 cell types used in the study

We uniformly processed Hi-C data for 18 cell types: newly generated data for induced pluripotent stem cells (iPSCs), and previously published Hi-C data for 17 cell types (one

embryonic cell line, five primary cell types, nine cell lines, one tissue, and one model cell) **(Figure 1; Table 1–2)**. Comparison of A/B compartments (open and closed chromatin, respectively) [29] derived from the Hi-C data demonstrated the wide distribution of chromatin architectures across these cell types. The iPSC was close to embryonic stem cells and differences between replicates negligible compared to differences between cell types **(Figure 2a)**. This demonstrated the reproducibility of the Hi-C data. We merged the Hi-C data across biological replicates for each cell type and from the merged replicates, we identified genomic interactions at 10 kbp resolution. The genomic region interactions with read count ≥ 5 and q-value ≤ 0.05 were defined as significant genomic interactions and from here on are referred to as genomic interactions. The number of genomic interactions per cell type varies from 2,540,361 to 46,901,774, with Hi-C sequencing depth and includes interactions in both A and B compartments **(Table 3)**. We recalculated A/B compartments at 1 Mbp resolution from the merged data for each cell type.We determined the expression of promoters and FANTOM5 enhancers in a given cell type using CAGE, and selected the strongest expressed promoter of every gene in each cell type. For the majority of genes (65.86% mRNA, 84.93% intergenic lncRNAs and 81.55% non-intergenic lncRNAs), the same promoter had the strongest expression in all 18 cell types **(Supplementary Figure 1)**. The strongest promoter and expressed FANTOM5 enhancers in each cell type were then mapped to the corresponding genomic interactions **(Figures 1, 2b; Table 4)**. We used the FANTOM CAT catalog of transcripts [1] to classify genes into mRNAs, intergenic lncRNAs, antisense lncRNAs, divergent lncRNAs, and sense intronic lncRNAs, with the last three classes collectively referred to here as non-intergenic lncRNAs. Promoters were classified into the promoter types: H3K4me3 enriched (canonical promoters), H3M4me1 enriched (enhancer like), or neither, based on the overlapping ChromHMM states [30]. Most 10 kbp genomic windows contained no more than one promoter **(Supplementary Figure 2)** from either the intergenic lncRNA or mRNA gene classes **(Figure 2c; Supplementary Figure 3)**, indicating that the 10 kbp resolution was sufficient for our purposes. The promoters of non-intergenic lncRNA (mainly consisting of divergent and antisense lncRNAs) were typically found in the same window as their associated mRNA promoter **(Figure 2c)**.

We defined genomic interactions for which both genomic windows contain at least one promoter or FANTOM5 enhancer as annotated interactions. The number of annotated interactions varies from 50,170 to 619,848 **(Table 3)** among the cell types, with 70.54 - 99.52% of expressed promoters and enhancers having at least one annotated interaction in the cell type in which they are expressed **(Table 6)**.

Number of interacting genes depends on promoter type and gene class

Next, we analyzed the connectivity for different gene classes and promoter types (defined based on histone mark enrichment) in the selected cell types. In iPSC, we found that genomic regions containing H3K4me1 (enhancer-like) promoters have more annotated interactions

compared to regions containing H3K4me3 (canonical) promoters **(Figure 2d, e)**. For both promoter types, mRNA promoters interacted with significantly more promoters than intergenic lncRNAs **(Figure 2d, f)**. This may be explained by the fact that mRNA promoters tend to be expressed in many cell types, requiring interactions with multiple cis-regulatory elements that provide cell type-dependent control, while lncRNAs tend to be expressed in few cell types and therefore represent cell type-specific regulatory elements. The same patterns were found in other cell types **(Figure 2e, f)** demonstrating that the dependence of the number of interacting genes on gene class and promoter type is a general feature of the 3D regulatory interaction network. Further, mRNA promoters, which typically are enriched in the H3K4me3 histone mark, form longer range interactions with intergenic lncRNA promoters with the H3K4me1 histone mark compared to those with the H3K4me3 histone mark **(Supplementary Figure 4a)**. A correlation analysis for each gene between the promoter expression and the number of annotated interactions across the 18 cell types demonstrated that for all three gene classes the number of interacting genes increases as expression level increases (**Figure 2g; Table 7)**. Correlation values were significantly higher for intergenic lncRNA promoters compared to mRNA coding promoters (**Figure 2h)**, suggesting that expression of intergenic lncRNAs and number of genomic interactions are interdependent. However, some promoters have a negative or no significant correlation between expression and the number of annotated interactions (**Table 7**).

LncRNA Hi-C gene clusters and their properties

For every expressed lncRNA in each cell type, we created a Hi-C genomic region cluster with interacting regions that extends to a specific degree of separation from the reference lncRNA, as outlined in **Figure 3a**. Next, expressed promoters from different gene classes and FANTOM5 enhancers expressed in the corresponding cell type were mapped to the Hi-C genomic region clusters to obtain Hi-C gene clusters for each lncRNA in each cell type. The generated Hi-C gene clusters were used to explore the biological role of lncRNAs in each cell type.

First, Gene Ontology (GO) enrichment analysis was performed for each lncRNA Hi-C gene cluster **(Table 8)**. While non-intergenic lncRNAs by definition have an mRNA gene in their immediate neighborhood that could be used to functionally annotate the lncRNA, functional annotation of intergenic lncRNAs must rely on longer-range gene interactions as provided by Hi-C data **(Figure 2c)**. We therefore assessed the performance of Hi-C data analysis primarily by its ability to provide functional annotation for intergenic lncRNAs. Depending on cell type, 21.96 - 39.40% of intergenic lncRNAs **(Figure 3b lower right panel)** and 25.60 - 44.73% of non-intergenic lncRNAs **(Supplementary Figure 5)** have at least one significantly enriched (FDR adjusted p-value ≤ 0.1) biological processes GO term.

We compared this functional annotation based on chromatin conformation to a functional

annotation based on the linear genomic neighborhood of each lncRNA (a strategy commonly used in the absence of Hi-C data). We performed GO enrichment analysis on linear genomic clusters consisting of all expressed mRNA genes within a +/- 1MB genomic distance from the reference lncRNA. For both strategies, we selected the statistically most significant GO term as the representative term for each cluster **(Figure 3b top and lower left panel)**. Across cell types, at a significance threshold of FDR adjusted p-value ≤ 0.1, on average 84 more intergenic lncRNAs and 252 non-intergenic lncRNAs can be annotated with a significantly enriched GO term in the Hi-C gene cluster analysis compared to the linear gene cluster analysis **(Figure 3b lower panel, Supplementary Figure 6, Supplementary Figure 7)**. Depending on cell type, 5.01 - 16.15% of intergenic lncRNAs could be associated with at least one GO term only by Hi-C, 6.38 - 13.02% could be annotated both by Hi-C and by linear gene cluster analysis, and only 3.72 - 10.96% could be annotated only by using linear clusters **(Figure 3b lower right panel)**. All together, depending on cell type we annotated upto 3.97 - 17.50% more lncRNAs using Hi-C gene clusters compared to linear gene clusters **(Figure 3b).** Overall, this demonstrates that lncRNA functional annotation provided by Hi-C data analysis can only be partially obtained by analyzing genes located in the genomic neighborhood of a lncRNA.

We further compared the Hi-C results with expression correlation analysis for the reference intergenic lncRNA and mRNA promoters in their cluster. Depending on cell type, 27.98 - 48.60% of intergenic lncRNA Hi-C gene clusters have a statistically significant preference for mRNA genes in the clusters with either positive or negative expression correlation with the reference lnRNA (**Figure 3c)**. However, as the majority of clusters have no specific expression correlation preference, analyzing expression correlation alone cannot reveal the clusters of associated genes found by chromatin conformation data analysis **(Figure 3c)**. To explore the effectiveness of Hi-C gene cluster GO analysis, we analyzed the expression correlation between GO-annotated genes (the subset of mRNA genes in Hi-C clusters with a GO annotation) and reference lncRNAs. We found only up to 14.28 - 30.48% of the reference lncRNAs that are associated with at least one GO term have a statistically significant non-zero average correlation with the GO genes **(Figure 3d)**. This demonstrates that the Hi-C cluster GO analysis can identify functional associations that cannot be obtained from expression correlation analysis **(Figure 3d)**.

Next, we performed Transcription factor binding sites (TFBSs) enrichment analysis for each intergenic lncRNA Hi-C gene cluster to identify regulatory modulators associated with promoters and enhancers. Depending on cell type, 328 to 2,253 Hi-C gene clusters were significantly enriched (FDR adjusted p-value ≤ 0.1) for at least one motif **(Supplementary Figure 8)**, including both cell type-specific and ubiquitous motifs **(Supplementary Figure 9)**, which may reflect cell type-specific or general functions, respectively, of the lncRNA.

Next, we analyzed the association between the promoter type of an intergenic lncRNA and the GO terms associated with its Hi-C gene cluster. Most GO terms were equally associated with intergenic lncRNAs with a H3K4me1 or a H3K4me3 promoter type **(Figure 4a left panel, Table 9)** suggesting that both are involved in regulating specific biological processes. Further, the cell type-specificity of a Hi-C gene cluster GO term showed that the majority of cell types have similar association patterns for both promoter types, demonstrating that there is no distinctive difference between them in regard to the associated biological functions across all the cell types **(Figure 4a right panel).**

We then categorized lncRNAs based on their evolutionary age [31] into human-specific lncRNAs and lncRNAs also expressed in other lineages (older lncRNAs) and found that the number of mRNA genes interacting with intergenic lncRNAs increases with the evolutionary age of the lncRNA **(Figure 4b; p = 2e-5)**, suggesting that younger lncRNAs have more specific interactions compared to older lncRNAs. Comparing the Hi-C cluster GO terms, the human-specific lncRNAs tend to be associated with GO terms that have more cell type related functions while older lncRNAs are more associated with GO terms related to general cell maintenance **(Figure 4c left panel, Table 10)**. For the majority of cell types, human-specific lncRNAs are associated with GO terms that have a higher cell type specificity compared to older lncRNAs **(Figure 4c right panel; Figure 4d).** This suggests lncRNAs were involved in cell type-specific functions when first acquired, and expanded their regulatory repertoire during evolution to encompass more general cellular functions.

We also compared GO and motif enrichment analysis results between cell type-enriched and broadly expressed intergenic lncRNAs. Out of 4,911 expressed intergenic lncRNAs, differential expression analysis identified 2,621 as cell type-enriched and 368 as broadly expressed lncRNAs; with the remaining 1,922 lncRNAs lowly expressed in all selected cell types **(Figure 5a)**. GO analysis comparison among the cell types showed that Hi-C clusters of 1,168 out of 2,621 cell type-enriched **(Figure 5c)** and 268 out of 368 broadly expressed intergenic lncRNAs **(Figure 5d)** have at least one GO term enriched in the Hi-C gene cluster in at least one out of 16 cell types. Next, for each cell type pair, we calculated the correlation between the number of clusters enriched for binding sites for specific TFs **(Figure 5b)**. We found that genes in Hi-C clusters of cell type-enriched and of broadly expressed lncRNAs tend to have binding sites for different TFs, suggesting that they have different transcription regulatory profiles. Further, for the majority of cell types, cell type-enriched lncRNAs are associated with GO terms that have a higher cell type specificity compared to broadly expressed lncRNAs **(Figure 5e, f).**

Regulator lncRNAs based on chromatin association

Nostrand *et al.* [32] have shown that knockdown of an RNA binding protein (RBP) results in differential expression and alternative splicing for genes with DNA binding sites for the RBP,

suggesting that it mediates pre- and post-transcriptional regulation at the chromatin level. As lncRNAs have been known to regulate mRNA splicing and post-translational modifications by interacting with the RBPs [33], lncRNAs may play a role in recruiting RBPs to their target sites. Based on eCLIP data for K562 and HepG2 cells [32] we found 6,239 and 6,216 lncRNAs, respectively, with at least one RBP binding to it. On the chromatin, RBP ChIP-Seq data shows enrichment of binding sites at promoters and enhancers in 1,192 lncRNA gene clusters (241 intergenic lncRNAs + 951 non-intergenic lncRNAs) in K562 and 1,133 lncRNA gene clusters (243 intergenic lncRNAs + 870 non-intergenic lncRNAs) in HepG2 **(Figure 6a)**. Combined these two results, we could identify 147 lncRNAs in K562 (36 intergenic lncRNAs + 111 non-intergenic lncRNAs) and 126 lncRNAs in HepG2 (27 intergenic lncRNAs + 99 non-intergenic lncRNAs) with RNA binding sites for specific RBPs that also show enrichment for DNA binding sites in their corresponding Hi-C gene clusters, suggesting that these lncRNAs may have a role in RBP recruitment in-cis at the promoters and enhancers to assist in transcription regulation.

Complementary to this analysis, we explored if lncRNAs can recruit RBPs on other RNA transcripts produced in each cluster. We performed enrichment analysis based on eCLIP signal on transcripts of Hi-C gene clusters and identified 1,168 lncRNAs gene clusters (270 intergenic lncRNAs + 898 non-intergenic lncRNAs) in K562 and 1,153 lncRNAs gene clusters (252 intergenic lncRNAs + 901 non-intergenic lncRNAs) in HepG2 that show enrichment for at least one RBP binding to transcripts of cluster genes. Further, 439 lncRNAs in K562 (121 intergenic lncRNAs + 318 non-intergenic lncRNAs) and 381 lncRNAs in HepG2 (101 intergenic lncRNAs + 280 non-intergenic lncRNAs) have RNA binding sites for specific RBPs that show enrichment for RNA binding sites at their corresponding Hi-C gene clusters. This suggests that lncRNA may bind specific RBPs to mediate their recruitment to both DNA and RNA. Clusters enriched for ChIP-Seq peaks and for genes with eCLIP signal overlapped significantly the RBPs FUS, NONO, ILF3, PCBP1, PCBP2 and HNRNPK **(Figure 6a)**, in agreement with previous results for RBPs PCBP1, PCBP2 and HNRNPK [32].

We then used RNA-DNA interaction data from Red-C, a proximity-based assay to capture and sequence RNA-DNA interactions, to identify RNA-chromatin interactions in K562 [32], yielding 58 reference lncRNAs with enrichment for chromatin interactions within their Hi-C gene clusters **(Figure 6b)**. Comparison with Hi-C GO analysis identified 25 (8 intergenic lncRNAs, 9 divergent lncRNAs, and 8 sense intronic lncRNAs) out of these 58 lncRNAs with enrichment for at least one GO term in their Hi-C cluster. Comparison of the Red-C and RBP ChIP-Seq enrichment analysis results for these 58 lncRNAs found that nine lncRNAs (LINC00470, LINC00963, RP1-167A14.2, AC018890.6, RP11-304L19.5, RP3-329A5.8, U91328.19, STAG3L5P-PVRIG2P-PILRB, CATG00000032680.1) out 58 lncRNAs have RNA binding sites for one or more RBPs for which their Hi-C gene clusters show DNA binding site enrichment at the promoters and FANTOM5 enhancers, suggesting these lncRNAs may acting as scaffolds for RBP recruitment at promoters

and enhancers.

As an example, the unannotated broadly expressed lncRNA U91328.19 (ENSG00000272462) is located in a Hi-C gene cluster containing multiple histone genes **(Figure 7)**. The GO analysis shows that this lncRNA may have a role in chromatin organization; its expression is positively correlated with the GO-annotated genes in its cluster. RBP HNRNPL, which plays a crucial role in repressing cryptic exons and is known to bind to the lncRNA transcript [34] , shows enrichment of DNA binding in the respective Hi-C cluster in K562. As Red-C data analysis revealed enrichment of this lncRNA binding at promoters and enhancers of its gene cluster, we conclude the lncRNA binds to HNRNPL and may guide its recruitment to chromatin.

Interactive querying and visualization of lncRNA functional annotations across cell types

We created an interactive system, accessible at https://fantom.gsc.riken.jp/zenbu/reports/#FANTOM6_HiC, to query and visualize the lncRNA functional annotations derived from our Hi-C analysis and to compare them across the 18 cell types. For each lncRNA, its 1) Hi-C gene cluster, 2) associated GO terms, TFBSs, RBPs, and TFs, and 3) genomic interactions are shown. The interactive system also allows comparing lncRNA expression, GO terms, GO genes-reference lncRNA expression correlation, genome wide association study (GWAS) traits, and TFBS motif enrichment across the cell types. As an example, the intergenic lncRNA CATG00000020001 is expressed specifically in the human liver cancer cell line HepG2 **(Supplementary Figure 10a)**. Genes in the lncRNA Hi-C cluster in HepG2 are enriched for GO terms related to blood physiology **(Supplementary Figure 10b; Supplementary Figure 10c bottom right heatmap)**. Expression of the lncRNA and genes in this GO category are positively correlated **(Supplementary Figure 10c; top heatmap)**, with all GO-annotated genes in the cluster expressed in HepG2 **(Supplementary Figure 10c; middle heatmap)**. Most of these genes are members of the Serpin family of serine protease inhibitors, involved in blood coagulation and angiogenesis [35]. Some GO-annotated genes are shared among the terms while others are specific to a term **(Supplementary Figure 10c; bottom left heatmap)**. Blood associated GWAS traits are enriched in the Hi-C cluster A/B compartments **(Supplementary Figure 10d)**. We conclude that both Hi-C and expression data indicate a role of this lncRNA in blood coagulation.

Discussion

In this study, based on Hi-C supported Gene Ontology (GO) enrichment analysis, we have provided functional annotations for 7,651 out of 14,198 lncRNAs (53.88%) (2,173/4,911 intergenic lncRNAs and 5,478/9,287 non-intergenic lncRNAs) expressed across 18 cell types. This is important especially for intergenic lncRNAs as their genomic environment lacks genes that could otherwise be used to suggest a functional role. Functional annotation of lncRNAs by

combining Hi-C interactions and GO analysis provides a robust methodology to understand the role of lncRNAs in cells. This analysis is independent of the co-expression of the genes, an alternative method to predict gene function, which is strongly affected by tissue composition, technical variation, and normalization issues that negatively affect the comparability of measured gene expression levels, provides only a single annotation of lncRNAs for all cell types, and therefore cannot identify cell type-specific functions of lncRNAs [36–41]. In contrast, Hi-C-based GO analysis avoids many of these pitfalls and provides a functional annotation for each lncRNA in each cell type separately.

Benabdallah *et al.* [42] have shown that during mouse embryonic cells to neural progenitor cells differentiation the promoters are in close proximity but not in direct contact with the enhancers regulating its activity. Our Hi-C clustering strategy includes not only the lncRNA directly interacting with the lncRNA promoter but also promoters that are spatially proximal to the lncRNA but not directly interacting, and thereby takes the mobility of the lncRNA in the nucleus into account.

We found several promoters pairs for H3K4me1 intergenic lncRNAs - H3K4me3 mRNAs in each cell type **(Table 11).** Yan *et al.* [43] have shown that histone methyltransferases MLL3 and MLL4 stimulate H3M4me1 methylation at enhancer regions, which in turn promotes promoter-enhancer interaction by recruiting the Cohesin complex. Close proximity of promoter-enhancer regions enables sharing transcription machinery between them. Studies have shown that noncoding RNAs facilitate the expression of promoters by terminating RNA PolII pausing [44] and recruiting transcription factors [45]; H3K4me1 intergenic lncRNA-H3K4me3 mRNA promoters pairs may rely on similar mechanisms.

Our analysis demonstrated that human-specific lncRNAs have more cell type-specific functions, while older lncRNA have more general functions. Won et al. [46] have shown that human-specific bi-directional enhancers in the human brain interact with genes associated with human cognitive function. Our results extend this pattern from bi-directional enhancers to include long non-coding RNAs. To our knowledge, the relation between the lineage of a lncRNA and its spatially colocalized genes has not been explored. However, our analysis is limited to the lncRNAs expressed in the restricted number of tissues in which the evolutionary lineage of the lncRNA has been analyzed [31]. Future improved lineage annotations can reveal finer details about the interdependence of lncRNA function and its evolutionary origin.

Studies have shown that lncRNAs can modulate gene expression by interacting with chromatin and changing the genome architecture, scaffolding the regulatory proteins, and facilitating the recruitment of the regulatory proteins like RBPs [33]. RBPs are known to bind to chromatin-associated RNAs and recently emerged as not only transcription regulators but also as playing a role in genome architecture maintenance. Our results suggest that many of the lncRNAs play a

role in enabling RBPs to perform their function by recruiting them at chromatin and RNAs.

Our analysis also found that certain lncRNAs are enriched for RNA-chromatin interactions at the genes of their Hi-C cluster in the K562 cell line. With lncRNAs known to have a role in chromatin structure organization [47–49] , the enrichment of RBPs in some of these clusters may indicate that the post-transcriptional machinery may be shared among the genes associated with a biological process. The presence of RNA-chromatin interactions in these clusters may facilitate cluster formation and bring genes of similar function together. While several techniques have been developed in recent years that probe RNA-chromatin interactions [50–54], lncRNAs may be underrepresented in these data due to their low expression level compared to mRNAs. In contrast, as Hi-C depends on DNA, it is equally powerful in providing annotations for lowly expressed lncRNAs as for highly expressed mRNAs. Hi-C has thus allowed us to develop the first broad cell type-specific functional annotation of lncRNAs.

Material and Methods

Hi-C data generation and processing

iPSC Hi-C data was generated as described in Ramilowski *et al.* [55] while data for the remaining cell types was downloaded from previously published studies (**Table 1**). Data for each replicate was processed using HICUP ver. 0.5.10 [56] which involved read truncation, mapping, filtering experimental artifacts, and de-duplication. The alignment files for all the replicates for each cell type were merged to perform downstream analysis. GRCh38 primary human genome assembly was used for the analysis.

Gene and enhancer models and primary annotations

FANTOM CAT gene models [1] and GRCh38 FANTOM 5 bidirectional enhancers [57] (https://zenodo.org/record/556775) were used as the primary genome annotation. The FANTOM CAT gene classes used in this study are mRNA (protein-coding), intergenic lncRNAs, antisense lncRNAs, divergent lncRNAs, and sense intronic lncRNAs.

Expression level of promoters and bi-directional enhancers

Expression levels of promoters and enhancers were determined using CAGE data. iPSC CAGE data was generated for this study using the nAnT-iCAGE protocol [27] while for remaining cell types publicly available matched CAGE libraries were used **(Table 5)**. Promoter CAGE tag counts were estimated by intersecting CTSS files for individual libraries with the promoter bed file using bedtools (v2.26.0) and were normalized to calculate promoter expression in tags per million (tpm). The expression for a promoter in a cell type was determined by calculating the mean expression across all CAGE libraries for that cell type. For each gene, the promoter with the highest expression level, requiring a minimum expression of 0.5 tpm, and minimum 3 tag

counts in at least one CAGE library was used for the downstream analysis. Next, CAGE expression of enhancers was calculated by summing the CAGE tag counts across the libraries for each cell type, ignoring the tag directionality. All enhancers with an aggregate tag count of at least 5 were used for the downstream analysis.

In the absence of matched CAGE libraries, FANTOM5 data for pancreatic tissues was repurposed for Islets and BetaH1 cells. First, expressed transcripts in Islets and BetaH1 cells were determined using the RNA-seq data from published studies [21,58] (**Table 5**) using Kallisto ver. 0.45.0 [59]. Next, the strongest promoter for each gene was determined based on expressed transcripts in each cell type and pancreatic tissue CAGE data. Expressed enhancers in pancreatic tissues were assigned to both Islets and BetaH1 cell types.

Promoter types

The precalculated chromatin state models for the selected cell types were downloaded from the previous studies listed in **Table 5**. The strongest promoters were intersected with corresponding chromatin state models and were assigned a promoter type: H3K4me3 enriched, H3K4me1 enriched, or Neither (undetermined) depending on the overlapping state **(Table 5)**. As chromatin state data were not available for THP-1, THP1-PMA, and RPE-1, and the FANTOM CAT promoter type was used to annotate promoters in these cell types **(Table 5)**.

A/B compartments, TADs and loops

Hi-C alignment .bam files for individual replicates and merged data were converted into .hic format using an in-house awk script and the Pre command from the Juicer package [60]. A/B compartments were identified at 1 Mbp resolution using the function eigenvector from Juicer package with options: -p VC <hic file> <chromosome name> BP 1000000. The positive value represents compartment A (transcriptionally active compartment) and the negative value represents compartment B (compartment with lower transcriptional activity compared to compartment-A). The compartments were redefined by reassigning the signs (+/-) to eigenvalues in cases where average expression values (determined using CAGE peaks) of compartment B (negative value) was higher than compartment A (positive value). Further, TADs and chromatin loops were calculated as a resource for the research community using functions from the Juicer package.

Determining significant genomic interactions

The intra-chromosomal genomic interactions were identified using the Bioconductor package GOTHiC [61] at 10 kbp resolution. The alignment .bam files were converted into .gothic files using format conversion script hicup2gothic from the HiCUP package. The interactions were calculated using .gothic together with corresponding restriction enzyme files **(Table 1)** for each cell type. All the interactions supported by at least 5 read pairs and q-value ≤ 0.05 were defined as significant genomic interactions. The interactions were annotated by overlaying the expressed promoters and enhancers in the selected cell types to identify the promoters interacting with each other. In cases where promoters for more than one gene overlapped the same 10 kbp region, the interactions were counted multiple times, with one interaction for each gene. Interactions with annotations on both sides were used for the downstream pairwise analysis.

Hi-C clusters and linear clusters

Hi-C genomic clusters for each lncRNA were defined using significant genomic interactions. The genomic window overlapping with the reference lncRNA promoter was selected as the reference genomic region. To define the Hi-C genomic clusters, all the genomic regions connected to reference genomic region by Hi-C interaction upto 2 degrees separation [reference lncRNA genomic region → genomic region that have Hi-C interaction connecting to the reference lncRNA genomic region (Genomic region with 1° degree of separation from reference genomic region) → genomic regions that are connected to Genomic region with 1° degree of separation by Hi-C interactions (Genomic region with 2° degree of separation from reference window)] within the A or B compartment (extended by +/- 100 kbp) in which the reference lncRNA regions is situated were used to define the HiC genomic region clusters **(Figure 3a)**. In the case of DMFB, Ery, and HUVEC where the Hi-C sequencing depth was relatively low, we also included genomic windows connected within 3 degrees of separation. The expressed promoters and enhancers were mapped to the Hi-C genomic region clusters to identify the Hi-C gene clusters. Linear gene clusters for each expressed lncRNA were defined by aggregating all mRNA genes within 1 Mbp upstream or downstream from the reference lncRNA locus.

Gene ontology (GO) enrichment analysis

The GO term database from NCBI was downloaded on Nov 28th 2019 (file: gene2go from ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/). All mRNA genes with an entrezID that are expressed in a given cell type and belong to at least one HiC cluster were used for the analysis. The GO term enrichment analysis for each Hi-C cluster was performed using one-sided Fisher's exact test

(details are given in **Table 12**). The background consisted of all genes in all clusters other than the genes in the reference lncRNA cluster. The analysis was performed for the GO terms that have at least 1 mRNA gene in their geneset in common with the cluster. The p-value was corrected for multiple testing using the Benjamini–Hochberg multiple testing correction method. All the GO terms with FDR corrected p-value ≤ 0.1 and at least 3 mRNA genes in their geneset in common with the cluster were defined as significant GO terms. GO term enrichment analysis was performed for the linear clusters for the "biological process" GO term category in the same way as for the Hi-C gene clusters. Further, GO enrichment analysis was also performed for gene sets provided by the Broad Institute [62–64] as a resource for annotating the lncRNA Hi-C gene clusters. Analysis was performed for all the Hi-C gene clusters and linear genes clusters with at least three mRNA genes with an EntrezID.

Hi-C gene cluster expression correlation analysis

For each cluster, Spearman correlation was calculated for gene expression between the reference lncRNA and mRNA genes in the cluster across the 18 cell types. To determine if the clusters have a preference for genes that have positive or negative expression correlation with the reference lncRNA, a one-sample Student's t-test was performed. The Hi-C gene cluster with p-value ≤ 0.05, and t-statistic value positive or negative were categorised as clusters with positive expression correlation or negative expression correlation respectively, or otherwise with no preference. The analysis was performed in the same manner including GO-annotated genes only.

TFBSs enrichment analysis and motif correlation

Genome-wide TFBS predictions for SwissRegulon motifs [65–67] were downloaded for the GRCh38 human genome assembly. For each cell type, the predicted TFBSs were intersected with promoters and enhancers regions extended by +/- 250 bp. In cases where multiple TFBSs for the same motif overlap with a promoter or enhancer, the posterior probability scores of the predicted TFBSs were summed. The TFBSs with an aggregate score of at least 0.1 were assigned to the promoters and enhancers. Motifs with predicted TFBSs overlapping with less than 200 promoters and enhancers were excluded from the analysis. The significance of motif enrichment in the Hi-C gene cluster was calculated using one-sided Fisher's exact test. The contingency table used for analysis is described in **Table 13**. The p-values were corrected for multiple testing using the Benjamini–Hochberg multiple testing correction method. All clusters where the number of promoters + enhancers was at least 5 were tested for enrichment.

LncRNA lineage assignment

LncRNA lineage classification by Sarropoulos *et al.* [31] was used after lifting over to hg38. The obtained gene models were intersected with FANTOM CAT gene models using bedtools (v2.26.0) to identify the corresponding lncRNAs. Lineage was assigned based on corresponding overlapping gene models by classifying FANTOM CAT lncRNAs into human-specific and older lncRNAs, which included lncRNAs 25 mya, 90 mya, 180 mya, 300 mya old as well as multimember family lncRNAs. In cases where two gene models with different ages were mapped to the same FANTOM CAT lncRNA the older lineage was assigned to the lncRNA.

GO term cell type specificity for the lncRNA with promoter type and lncRNA lineages

GO term cell type specificity in the context of lncRNAs with different promoter types were calculated by determining the GO term entropy across the cell types. To calculate the entropy, a consensus promoter type based on promoter types across 18 cell types was assigned to each lncRNA. Next, in each cell type, the number of associated H3K4me1 enriched or H3K4me3 enriched lncRNA clusters for each GO term was calculated. The obtained counts were used to calculate the GO term entropy across the 18 cell types. The cell type specificity of a GO term decreases with increase in the GO term entropy. All GO terms with the FDR corrected p-value ≤ 0.1 were used for this calculation. Similar to promoter types, entropy for GO terms in the context of lncRNA lineage analysis was determined using the counts of human-specific and older lncRNA gene clusters that showed enrichment for a GO term.

Cell type enriched and broadly expressed lncRNAs

To identify cell type enriched lncRNAs, differential expression analysis was performed using the Bioconductor package edgeR. The raw expression table for all the 53 CAGE libraries for 18 cell types used for the analysis (CAGE data for pancreatic cell lines and tissue are used as representative for pancreatic cell type Islets and Beta-H1). Genes with expression ≥ 1 TPM in at least two CAGE libraries across all cell types were tested for differential expression (DE). To identify the differentially expressed genes in each cell type, the corresponding CAGE libraries were compared with the CAGE libraries of remaining cell types using “Exact test” function from Bioconductor package edgeR (ver. 3.12.1) [68–70]. Genes with log2 (Fold change) ≥ 1.5 and FDR corrected p-value ≤ 0.05 were defined as differentially expressed genes. The intergenic lncRNAs that were differentially expressed in upto two cell types were defined as cell type enriched lncRNAs. Intergenic lncRNAs that are expressed in three or more cell types with expression 1 tpm or more and were not defined as cell type enriched lncRNAs were defined as broadly expressed lncRNAs.

RBP enrichment analysis

Right-tailed Fisher's exact test (on genes or clusters, depending on the context) was used in the enrichment tests. Benjamini-Hochberg (FDR) correction for multiple tested clusters was applied when estimating enrichment of HiC-clusters for each RBP.

Overview and analysis of ENCODE eCLIP data: The ENCODE eCLIP data was available for two cell lines: HepG2 and K562. In total, there were 234 <RBP, cell type> pairs. The data were analyzed as follows: (1) reads were preprocessed as in the original eCLIP pipeline [71], (2) trimmed reads were mapped to the hg38 genome assembly with F6 CAT genome annotation using *hisat2* [72], (3) the aligned reads were deduplicated [71], (4) the uniquely and correctly paired reads were filtered with *samtools* [73], (5) gene-level read counts in exons were obtained with *plastid* [74], (6) differential expression analysis against matched controls was performed with *edgeR* [69], (7) to account for commonly bound RNAs, Z-scores were estimated for each gene across data for all RBPs [55] for the fold change values estimated in (6). Based on (6) and (7) reliable RNA targets of each RBP were defined as those passing 5% FDR, raw $\log_2$FC > 0.5, Z-Score($\log_2$FC) > 1.645. RBPs with less than 10 targets in HiC-clusters were excluded from the downstream analysis resulting in the final set of 208 <RBP, cell type> pairs.

Overview and analysis of ENCODE ChIP-Seq data: Optimally thresholded ChIP-Seq peaks of 12 and 22 RBPs (for HepG2 and K562, respectively) were downloaded from ENCODE. U2AF1 (in K562) was excluded from analysis as having less than 10 RNA targets in eCLIP data. To annotate promoters, gene bodies, and enhancers with ChIP-Seq peaks we required non-zero overlap between the respective genomic regions and peaks, which was estimated with *bedtools*.

Annotation of RBP RNA targets with ChIP-Seq data: The annotation of RNA targets with ChIP-Seq data was based on an iterative hierarchical procedure, from direct evidence (ChIP-Seq at the promoter) to indirect evidence (localization in eCLIP-enriched clusters). For example, the targets with a ChIP-Seq peak at the promoter were not checked for a peak within gene body, or for localization within ChIP-Seq-enriched HiC-cluster; the targets within clusters with 'promoter+enhancer'-level ChIP-Seq enrichment were not checked for gene-level enrichment of clusters; etc. This procedure allowed us to plot a cumulative number of targets that could be additionally explained by considering each additional layer of annotation.

Red-C enrichment analysis

The processed Red-C data mapped to hg38 genome assembly was downloaded from GEO (GSE136141). Replicates for K562 cells were pooled providing 64,453,011 RNA-DNA contacts in total. Each contact consisted of an RNA part (genomic coordinates and orientation of paired RNA reads), and a DNA part (unstranded genomic coordinates). The contacts were mapped to

the FANTOM CAT gene annotation controlling for proper orientation of the RNA parts of the contacts, and only considering the cases when an lncRNA contained both RNA reads of a read pair. Only intrachromosomal contacts were considered in the following analysis. We considered 291 HiC gene clusters contacting at least 10 genes within and 10 genes outside of the cluster.

For each RNA and each *i*-th gene we estimated the contact densities as contacts per kilobase of the gene length: $CPK_i = (S_i \times 10^3) / L$ where S_i is the number of contacts with a particular gene and L is the gene length. Log-CPK values were used in the analysis. For each RNA, the contacts at the zero distance (primarily the contacts with its own gene) were excluded. To perform detrending and normalization of the contact densities, we applied the following linear model with the log-density of the contacts (log-CPK) as the dependent variable, and the RNA expression, the total RNA length (including introns), the distance to the contacting gene, and the cluster membership of the contacting gene (within or outside of the cluster) as predictors:

$$log2(CPK) \sim Cluster\ membership + log_{10}(Distance) + log_{10}(Distance):Cluster\ membership + log_2(Expression) + log_2(Length)$$

The model explicitly included a term for the interaction between the distance to the contacting gene and the cluster membership (within or outside of the selected cluster). CAGE data (mean promoter expression) was used as an estimate of the RNA expression;all terms, including the interaction term, were highly significant ($P < 10^{-16}$).

With detrended contact densities for each lncRNA under consideration, we performed the Wilcoxon-Mann-Whitney test on genes within and outside of cluster by comparing detrended density of contacts per kilobase within the cluster versus contacts outside of cluster.

Heritability enrichment analysis

We used stratified linkage disequilibrium (LD) score regression (ldsc software v.1.0.0, https://github.com/bulik/ldsc) [75] to partition the common (minor allele frequency (MAF) > 5% in European 1000 Genomes Project Phase 3 data) SNP heritability for 47 UK Biobank traits and diseases (https://data.broadinstitute.org/alkesgroup/UKBB) and four diseases which were used by O'Connor et al. [76–80]. We tested partitioned heritability of each Hi-C annotation (converted to hg19 genome build using UCSC LiftOver tool) conditioning on the baselineLD model v2.2 (https://data.broadinstitute.org/alkesgroup/LDSCORE/1000G_Phase3_baselineLD_v2.2_ldscores.tgz). We calculated the significance of the regression coefficient for the Hi-C annotation using the z-score.

Funding

FANTOM6 was made possible by a Research Grant for RIKEN Center for Life ScienceTechnology, Division of Genomic Technologies (CLST DGT) and RIKEN Center for Integrative Medical Sciences (IMS) from MEXT, Japan. This work was supported by the Russian Science Foundation (20-74-10075 to IVK), and the Natural Sciences and Engineering Research Council of Canada (RGPIN-2015-03948 to M.M.H.).

Competing interests

The authors declare no competing financial interests.

Bibliography

1. Hon, C.-C. *et al*. An atlas of human long non-coding RNAs with accurate 5' ends. *Nature* **543,** 199–204 (2017).
2. Holmes, D. S., Mayfield, J. E., Sander, G. & Bonner, J. Chromosomal RNA: its properties. *Science* **177,** 72–74 (1972).
3. Carninci, P. *et al*. The transcriptional landscape of the mammalian genome. *Science* **309,** 1559–1563 (2005).
4. Kapranov, P. *et al*. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316,** 1484–1488 (2007).
5. de Hoon, M., Shin, J. W. & Carninci, P. Paradigm shifts in genomics through the FANTOM projects. *Mamm. Genome* **26,** 391–402 (2015).
6. Vance, K. W. & Ponting, C. P. Transcriptional regulatory functions of nuclear long noncoding RNAs. *Trends Genet.* **30,** 348–355 (2014).
7. Gil, N. & Ulitsky, I. Regulation of gene expression by cis-acting long non-coding RNAs. *Nat. Rev. Genet.* **21,** 102–117 (2020).

8. Böhmdorfer, G. & Wierzbicki, A. T. Control of chromatin structure by long noncoding RNA. *Trends Cell Biol.* **25,** 623–632 (2015).

9. Mishra, K. & Kanduri, C. Understanding long noncoding RNA and chromatin interactions: what we know so far. *Noncoding RNA* **5,** (2019).

10. He, R.-Z., Luo, D.-X. & Mo, Y.-Y. Emerging roles of lncRNAs in the post-transcriptional regulation in cancer. *Genes Dis.* **6,** 6–15 (2019).

11. FANTOM Consortium and the RIKEN PMI and CLST (DGT) *et al.* A promoter-level mammalian expression atlas. *Nature* **507,** 462–470 (2014).

12. Clark, M. B. *et al.* Genome-wide analysis of long noncoding RNA stability. *Genome Res.* **22,** 885–898 (2012).

13. Tani, H. *et al.* Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. *Genome Res.* **22,** 947–956 (2012).

14. Ulitsky, I. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nat. Rev. Genet.* **17,** 601–614 (2016).

15. Groff, A. F., Barutcu, A. R., Lewandowski, J. P. & Rinn, J. L. Enhancers in the Peril lincRNA locus regulate distant but not local genes. *Genome Biol.* **19,** 219 (2018).

16. Amaral, P. P. *et al.* Genomic positional conservation identifies topological anchor point RNAs linked to developmental loci. *Genome Biol.* **19,** 32 (2018).

17. Tan, J. Y. *et al.* cis-Acting Complex-Trait-Associated lincRNA Expression Correlates with Modulation of Chromosomal Architecture. *Cell Rep.* **18,** 2280–2288 (2017).

18. Belton, J.-M. *et al.* Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58,** 268–276 (2012).

19. Rao, S. S. P. *et al*. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159,** 1665–1680 (2014).

20. Huang, P. *et al*. Comparative analysis of three-dimensional chromosomal architecture identifies a novel fetal hemoglobin regulatory element. *Genes Dev*. **31,** 1704–1713 (2017).

21. Lawlor, N. *et al*. Multiomic Profiling Identifies cis-Regulatory Networks Underlying Human Pancreatic β Cell Identity and Function. *Cell Rep*. **26,** 788-801.e6 (2019).

22. Darrow, E. M. *et al*. Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture. *Proc Natl Acad Sci USA* **113,** E4504-12 (2016).

23. Jin, F. *et al*. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503,** 290–294 (2013).

24. Phanstiel, D. H. *et al*. Static and Dynamic DNA Loops form AP-1-Bound Activation Hubs during Macrophage Development. *Mol. Cell* **67,** 1037-1048.e6 (2017).

25. Wutz, G. *et al*. Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *EMBO J.* **36,** 3573–3599 (2017).

26. D'Ippolito, A. M. *et al*. Pre-established Chromatin Interactions Mediate the Genomic Response to Glucocorticoids. *Cell Syst*. **7,** 146-160.e7 (2018).

27. Murata, M. *et al*. Detecting expressed genes using CAGE. *Methods Mol. Biol*. **1164,** 67–85 (2014).

28. Andersson, R. *et al*. An atlas of active enhancers across human cell types and tissues. *Nature* **507,** 455–461 (2014).

29. Lieberman-Aiden, E. *et al*. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–293 (2009).

30. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9,** 215–216 (2012).

31. Sarropoulos, I., Marin, R., Cardoso-Moreira, M. & Kaessmann, H. Developmental dynamics of lncRNAs across mammalian organs and species. *Nature* **571,** 510–514 (2019).

32. Van Nostrand, E. L. *et al*. A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583,** 711–719 (2020).

33. Briata, P. & Gherzi, R. Long Non-Coding RNA-Ribonucleoprotein Networks in the Post-Transcriptional Control of Gene Expression. *Noncoding RNA* **6,** (2020).

34. McClory, S. P., Lynch, K. W. & Ling, J. P. HnRNP L represses cryptic exons. *RNA* **24,** 761–768 (2018).

35. Rau, J. C., Beaulieu, L. M., Huntington, J. A. & Church, F. C. Serpins in thrombosis, hemostasis and fibrinolysis. *J. Thromb. Haemost.* **5 Suppl 1,** 102–115 (2007).

36. Chen, C. *et al*. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS ONE* **6,** e17238 (2011).

37. Gao, C., McDowell, I. C., Zhao, S., Brown, C. D. & Engelhardt, B. E. Context Specific and Differential Gene Co-expression Networks via Bayesian Biclustering. *PLoS Comput. Biol.* **12,** e1004791 (2016).

38. Leek, J. T. *et al*. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11,** 733–739 (2010).

39. Uygun, S., Peng, C., Lehti-Shiu, M. D., Last, R. L. & Shiu, S.-H. Utility and limitations of using gene expression data to identify functional associations. *PLoS Comput. Biol.* **12,** e1005244 (2016).

40. Zhang, Y., Cuerdo, J., Halushka, M. K. & McCall, M. N. The effect of tissue composition on gene co-expression. *Brief. Bioinformatics* (2019). doi:10.1093/bib/bbz135

41. Parsana, P. *et al*. Addressing confounding artifacts in reconstruction of gene co-expression networks. *Genome Biol*. **20,** 94 (2019).

42. Benabdallah, N. S. *et al*. Decreased Enhancer-Promoter Proximity Accompanying Enhancer Activation. *Mol. Cell* **76,** 473-484.e7 (2019).

43. Yan, J. *et al*. Histone H3 lysine 4 monomethylation modulates long-range chromatin interactions at enhancers. *Cell Res*. **28,** 387 (2018).

44. Meng, H. & Bartholomew, B. Emerging roles of transcriptional enhancers in chromatin looping and promoter-proximal pausing of RNA polymerase II. *J. Biol. Chem*. **293,** 13786–13794 (2018).

45. Lee, N. & Steitz, J. A. Noncoding RNA-guided recruitment of transcription factors: A prevalent but undocumented mechanism? *Bioessays* **37,** 936–941 (2015).

46. Won, H. *et al*. Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538,** 523–527 (2016).

47. Cusanelli, E. & Chartrand, P. Telomeric repeat-containing RNA TERRA: a noncoding RNA connecting telomere biology to genome integrity. *Front. Genet*. **6,** 143 (2015).

48. Lee, S. *et al*. Noncoding RNA NORAD regulates genomic stability by sequestering PUMILIO proteins. *Cell* **164,** 69–80 (2016).

49. da Rocha, S. T. & Heard, E. Novel players in X inactivation: insights into Xist-mediated gene silencing and chromosome conformation. *Nat. Struct. Mol. Biol*. **24,** 197–204 (2017).

50. Bonetti, A. *et al*. RADICL-seq identifies general and cell type-specific principles of genome-

wide RNA-chromatin interactions. *Nat. Commun.* **11,** 1018 (2020).

51. Li, X. *et al.* GRID-seq reveals the global RNA-chromatin interactome. *Nat. Biotechnol.* **35,** 940–950 (2017).
52. Wu, W. *et al.* Mapping RNA-chromatin interactions by sequencing with iMARGI. *Nat. Protoc.* **14,** 3243–3272 (2019).
53. Cecere, G. & Grishok, A. RNA Chromatin Immunoprecipitation (RNA-ChIP) in Caenorhabditis elegans. *Bio Protoc* **4,** (2014).
54. Chu, C. & Chang, H. Y. Understanding RNA-Chromatin Interactions Using Chromatin Isolation by RNA Purification (ChIRP). *Methods Mol. Biol.* **1480,** 115–123 (2016).
55. Ramilowski, J. A. *et al.* Functional annotation of human long noncoding RNAs via molecular phenotyping. *Genome Res.* **30,** 1060–1072 (2020).
56. Wingett, S. *et al.* HiCUP: pipeline for mapping and processing Hi-C data. [version 1; peer review: 2 approved, 1 approved with reservations]. *F1000Res.* **4,** 1310 (2015).
57. Rennie, S., Dalby, M., van Duin, L. & Andersson, R. Transcriptional decomposition reveals active chromatin architectures and cell specific regulatory interactions. *Nat. Commun.* **9,** 487 (2018).
58. Khetan, S. *et al.* Type 2 Diabetes-Associated Genetic Variants Regulate Chromatin Accessibility in Human Islets. *Diabetes* **67,** 2466–2477 (2018).
59. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34,** 525–527 (2016).
60. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst.* **3,** 95–98 (2016).

61. Mifsud, B. *et al*. GOTHiC, a probabilistic model to resolve complex biases and to identify real interactions in Hi-C data. *PLoS ONE* **12,** e0174744 (2017).

62. Liberzon, A. *et al*. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27,** 1739–1740 (2011).

63. Liberzon, A. *et al*. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst*. **1,** 417–425 (2015).

64. Subramanian, A. *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* **102,** 15545–15550 (2005).

65. van Nimwegen, E. Finding regulatory elements and regulatory motifs: a general probabilistic framework. *BMC Bioinformatics* **8 Suppl 6,** S4 (2007).

66. Alam, T. *et al*. Comparative transcriptomics of primary cells in vertebrates. *Genome Res*. **30,** 951–961 (2020).

67. Pachkov, M., Balwierz, P. J., Arnold, P., Ozonov, E. & van Nimwegen, E. SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. *Nucleic Acids Res*. **41,** D214-20 (2013).

68. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. **11,** R25 (2010).

69. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).

70. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-

Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40,** 4288–4297 (2012).

71. Van Nostrand, E. L. *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods* **13,** 508–514 (2016).
72. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37,** 907–915 (2019).
73. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
74. Dunn, J. G. & Weissman, J. S. Plastid: nucleotide-resolution analysis of next-generation sequencing and genomics data. *BMC Genomics* **17,** 958 (2016).
75. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47,** 1228–1235 (2015).
76. O'Connor, L. J. *et al.* Extreme polygenicity of complex traits is explained by negative selection. *Am. J. Hum. Genet.* **105,** 456–476 (2019).
77. Lambert, J. C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45,** 1452–1458 (2013).
78. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47,** 979–986 (2015).
79. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506,** 376–381 (2014).
80. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights

from 108 schizophrenia-associated genetic loci. *Nature* **511,** 421–427 (2014).

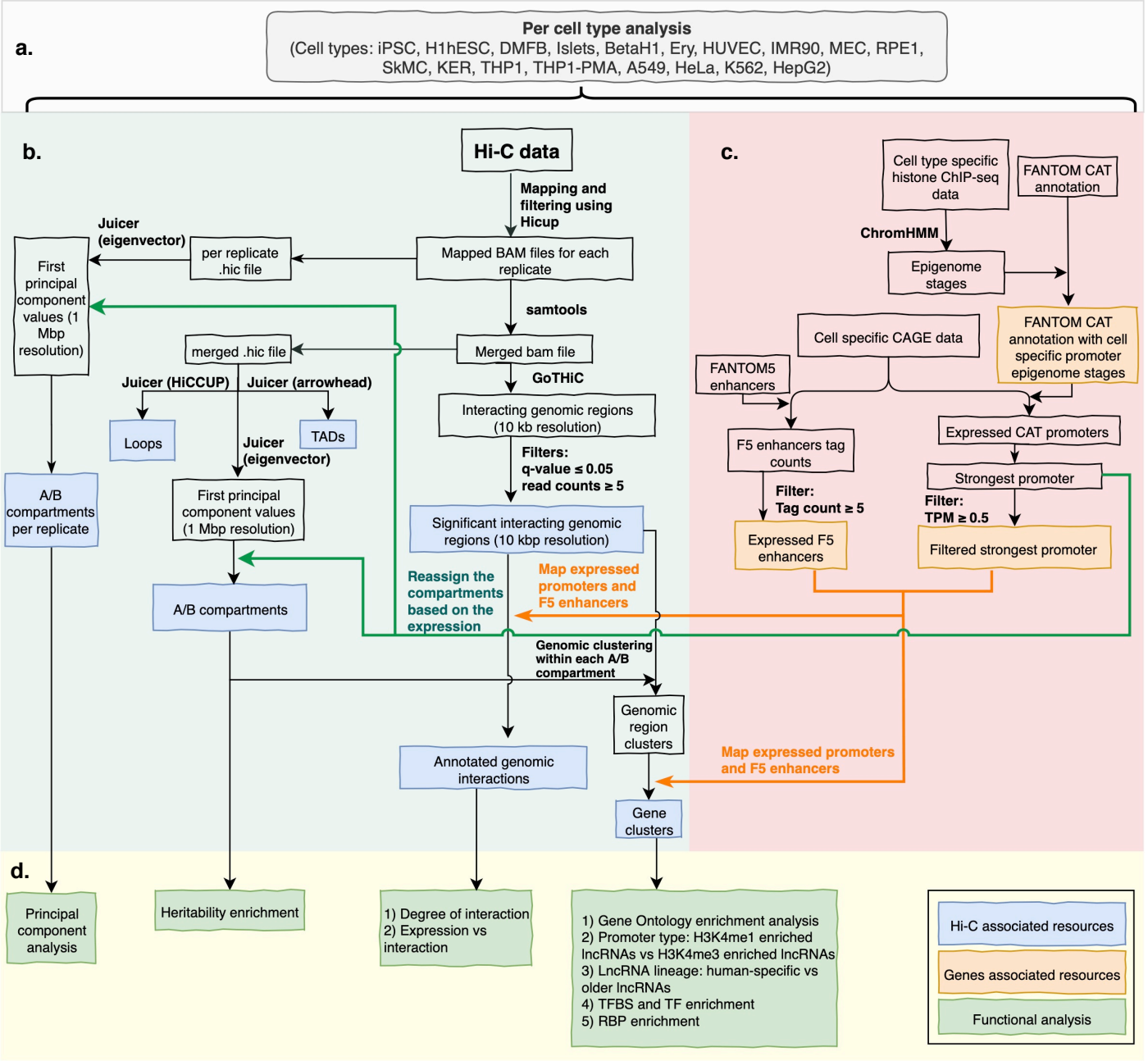


Figure 1: Overview of the project. a. Hi-C and CAGE data for 18 cell types were analyzed to characterize lncRNAs in these cells. **b.** For each cell type, Hi-C data was processed to identify the A/B compartments, chromatin loops, topologically associated domains (TADs), as well as significant interactions between genomic regions. **c.** CAGE data for each cell type were used to identify the cell type-specific expression of FANTOM CAT promoters and F5 enhancers. The expressed promoters were mapped to significant genomic interactions and genomic region clusters to obtain annotated genomic region interactions and gene clusters, respectively. **d.** Functional analyses were performed to explore the biological role of the long non-coding RNAs.

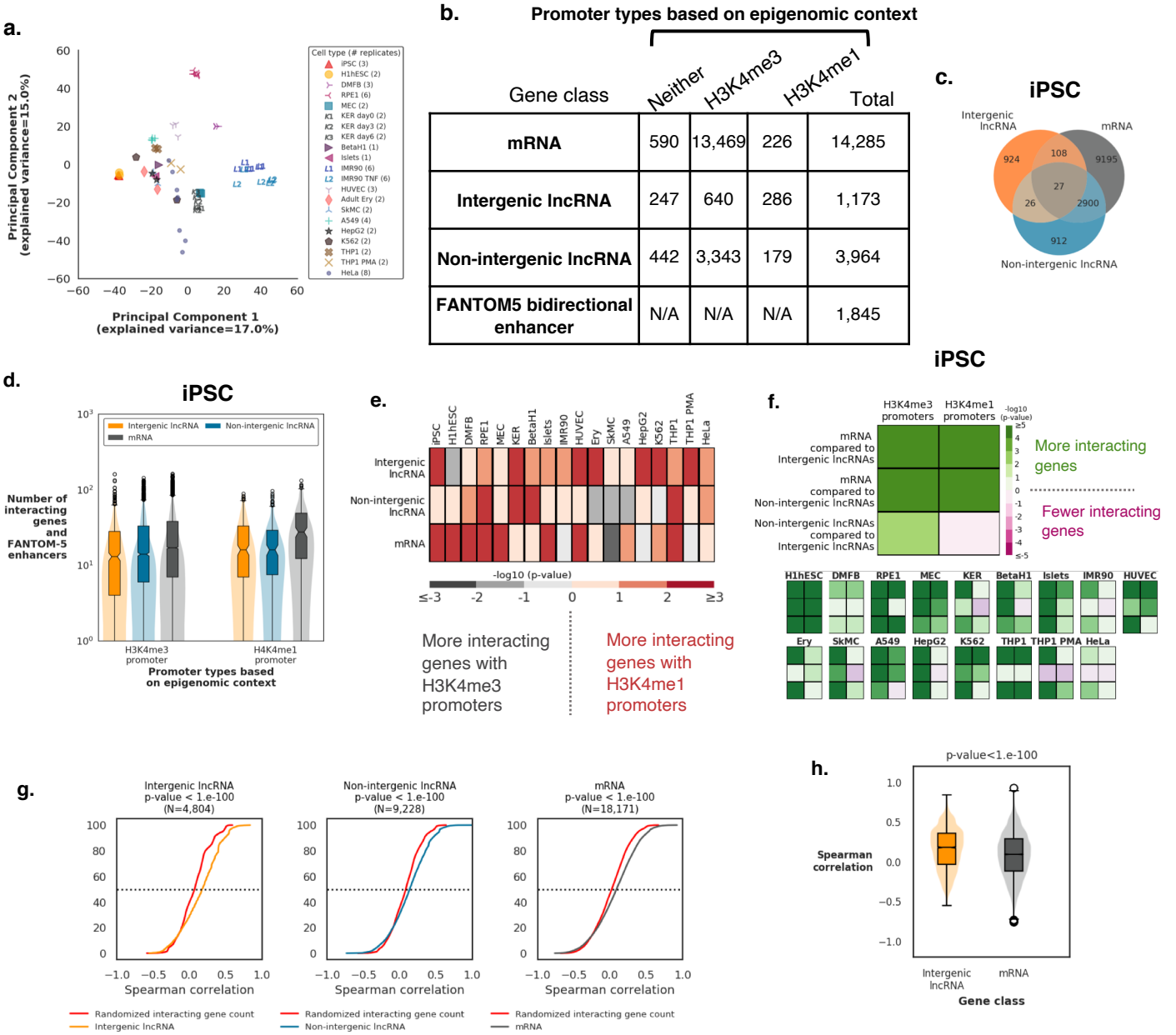


Figure 2: Characteristic features of pairwise interactions. a. Principal component analysis of A/B compartments per replicate of the cells used in the study. For time course (keratinocytes) and treated (IMR90) samples, each condition is shown separately. **b.** The number of expressed FANTOM CAT genes for each promoter type and gene class, and FANTOM5 bi-directionally expressed enhancers in iPSC based on CAGE data. In each of gene class, most genes have a H3K4me3 enriched promoter rather than a H4K4me1 enriched promoter. **c.** The number of promoters coinciding in a 10kb window for different gene classes. Intergenic lncRNA and mRNA coding promoters usually do not co-occur in the same window, while most other lncRNAs are within 10kb of at least one mRNA coding promoter. **d.** Comparison between promoter types and gene classes of the number of annotated interactions in iPSc. **e**. Statistical significance of the difference in number of interactions between the H3K4me1 and H3K4me3 promoter types for the same gene class, calculated using a one-tailed Mann-Whitney test. Red represents that H3K4me3 promoters have significantly more interaction than H3K4me1 promoters, while grey represents vice versa. **f.** Statistical significance of the difference in number of interactions between gene classes for the same promoter type, calculated using a one-tailed Mann-Whitney test. The top panel shows the comparison between gene classes in iPSc while the lower panel shows the comparison for the remaining 17 cell types arranged in the same way as iPSc heatmap. **g.** Cumulative distribution of Spearman's correlation between gene expression and number of interacting promoters across 18 cell types for each gene class. The number of genes (N) for each gene class used for correlation analysis is indicated on y-axis. The background distribution of correlation is calculated by randomizing the number of interacting promoters within each gene class 100 times and then taking the mean of correlation values across the 100 runs. The statistical significance of correlation for each gene class was calculated using Student's t-test for one sample. **h.** Comparison between intergenic lncRNA and mRNA coding region correlation value between gene expression and number of interacting promoters. Significance of correlation value difference between gene classes was determined using two-tailed Mann-Whitney test.

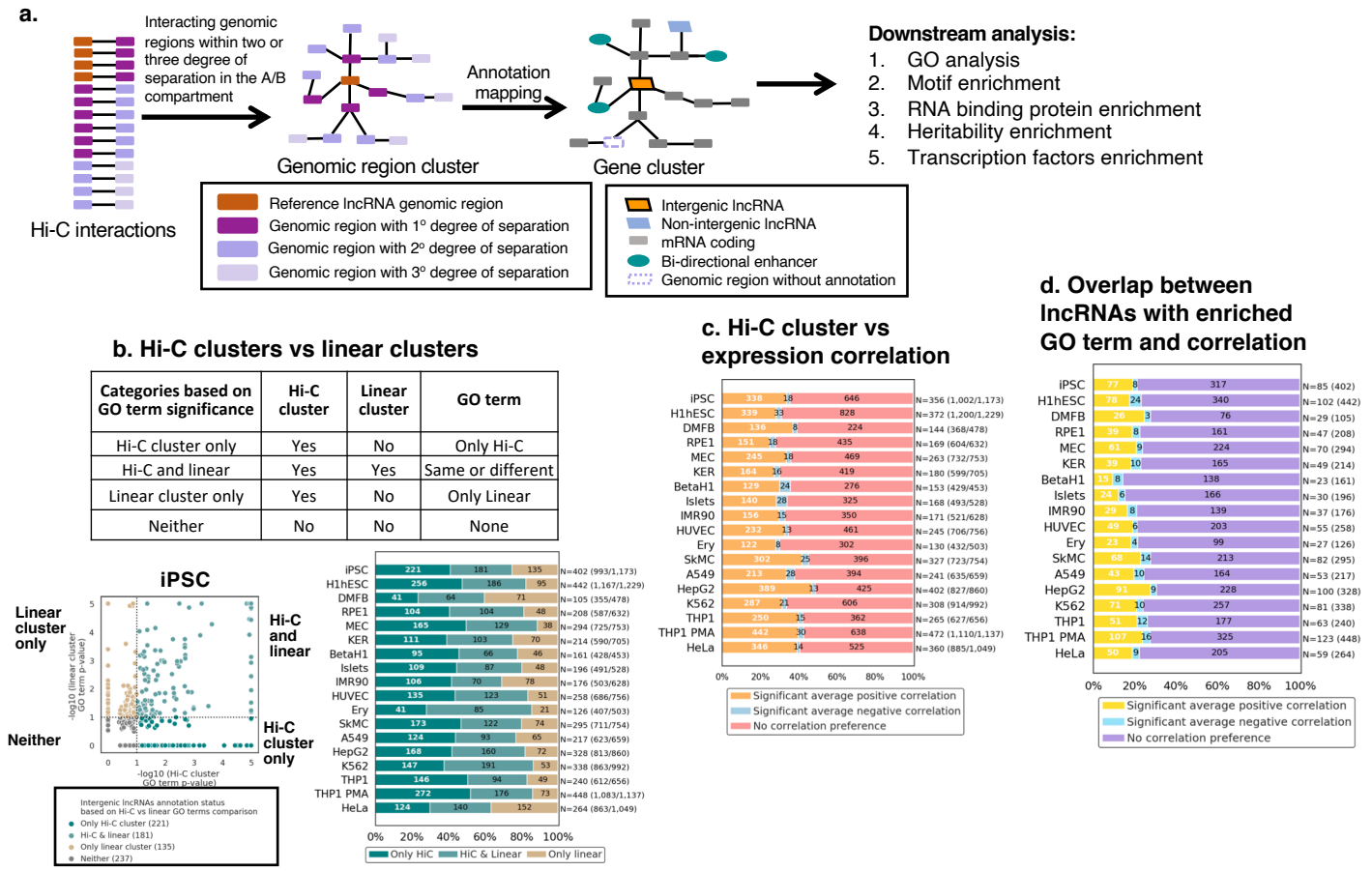


Figure 3: Intergenic lncRNAs gene clusters. a. Schematic diagram showing the workflow of lncRNA Hi-C gene cluster generation. **b.** Comparison between the Hi-C gene clusters and linear gene clusters (+/- 1MB region around the reference lncRNA) for intergenic lncRNAs using the most significant biological processes GO term (FDR corrected p-value <=0.1). The top panel shows the four possible annotation outcomes. The bottom left panel shows a detailed comparison for iPS cells, with each dot representing one lncRNA, and the color of dots denoting the annotation outcome. The number in parentheses in the legend is the number of lncRNAs. The dotted line shows the cutoff p-value of 0.1. The bottom right panel summarizes the comparison for all 18 cell types. Each row corresponds to one cell type, and the number of lncRNAs for each category is shown in the bar. The number on the right of each row is the number of intergenic lncRNAs with at least one GO term annotation supported by Hi-C data. In parentheses the number of intergenic lncRNAs tested for GO term enrichment / the number of intergenic lncRNAs expressed in that cell type. **c.** Number of clusters that have a significantly higher number of cluster genes with a positive expression correlation, with a negative expression correlation and with no significant preference for the sign of the expression correlation with the reference lncRNA in each cell type. The number on the right of each row is the number of intergenic lncRNAs significant positive or negative correlation preference with cluster mRNA genes. In parentheses the number of intergenic lncRNAs tested for correlation/ the number of intergenic lncRNAs expressed in that cell type. The clusters were classified into three categories by comparing the set of expression correlation values of cluster genes to 0 using one sample Student's t-test (p-value ≤ 0.05). **d.** Number of intergenic reference lncRNAs that have positive, negative or no correlation with GO genes. Each row is one cell type, and the number of lncRNAs for each category is shown in the bar. The number on the right of each row is the number of intergenic lncRNAs with at least one GO term annotation supported by Hi-C data.

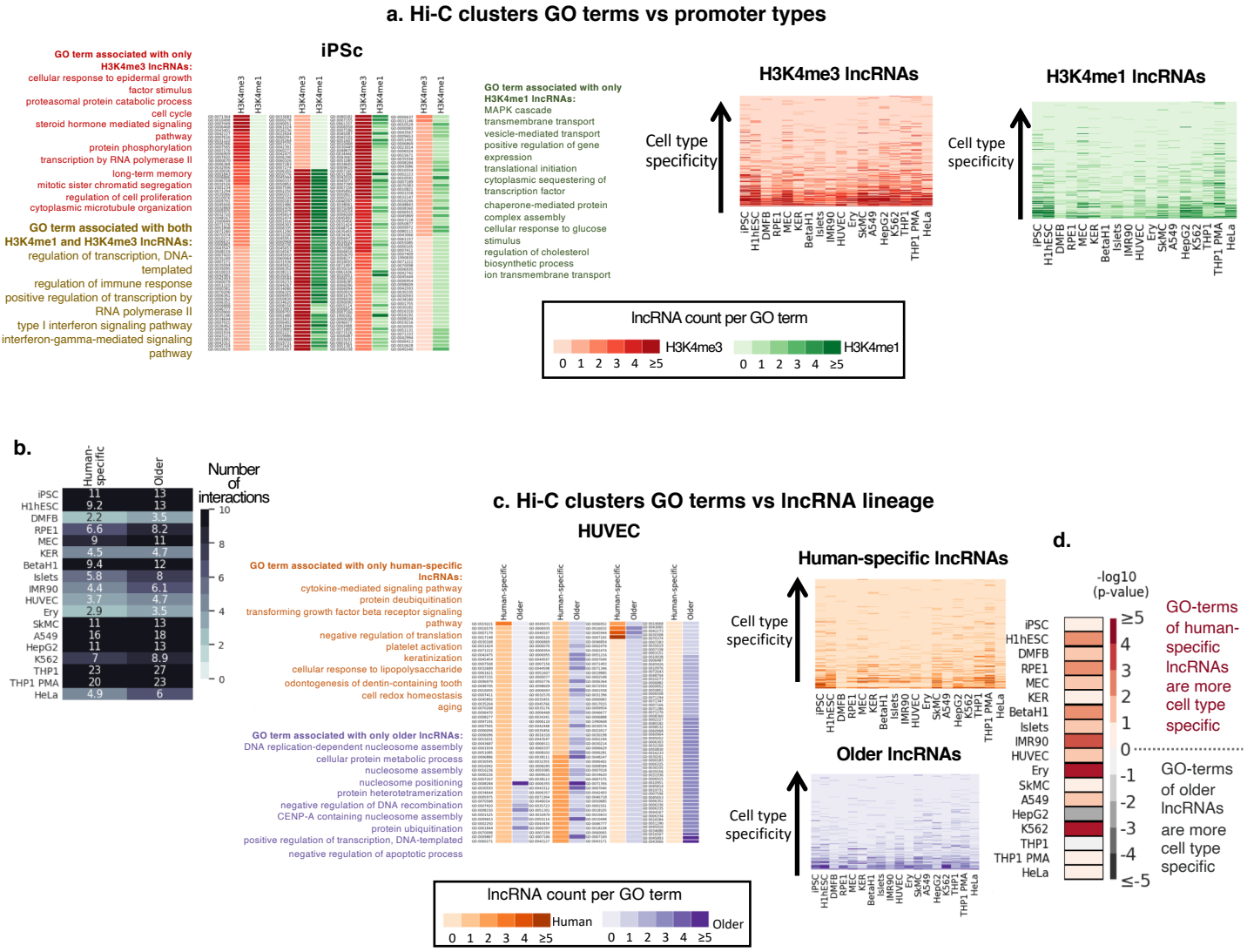


Figure 4: Properties of intergenic lncRNAs gene clusters. **a.** Association between lncRNA Hi-C cluster GO terms and promoter types. The top panel shows the number of lncRNAs in iPSC for each GO term separated by promoter type with the top 10 GO terms listed explicitly. Right panel shows the cell type specificity of the GO terms vs the number of lncRNAs for each promoter type. The heatmaps show the number of lncRNAs associated with each GO term, with each row representing one GO term and each column to a cell type. GO terms are sorted in descending order of GO term cell type specificity. Colorbar is the same as for the left panel. **b.** Number of mRNA coding promoters interacting with intergenic lncRNAs, separated by their evolutionary lineage, in all the 18 cell types. The value in each cell represents the mean number of interactions for each lineage for the cell type shown on the left. The lower number for interactions for human specific lncRNAs compared to older lncRNAs was calculated using the binomial test. **c.** Association between lncRNA Hi-C cluster GO terms and lncRNA lineage. The left panel shows the number of lncRNAs for each GO term, separated by the evolutionary age of the lncRNA. In the right panel, a heatmap showing the number of lncRNAs separated by evolutionary age as a function of the cell type specificity of the GO term (rows) in each cell type (columns) . GO terms are sorted in descending order of cell type specificity. **d.** Comparison of cell type specificity of GO terms associated with human-specific and older lncRNAs using one-tailed Mann-Whitney test.

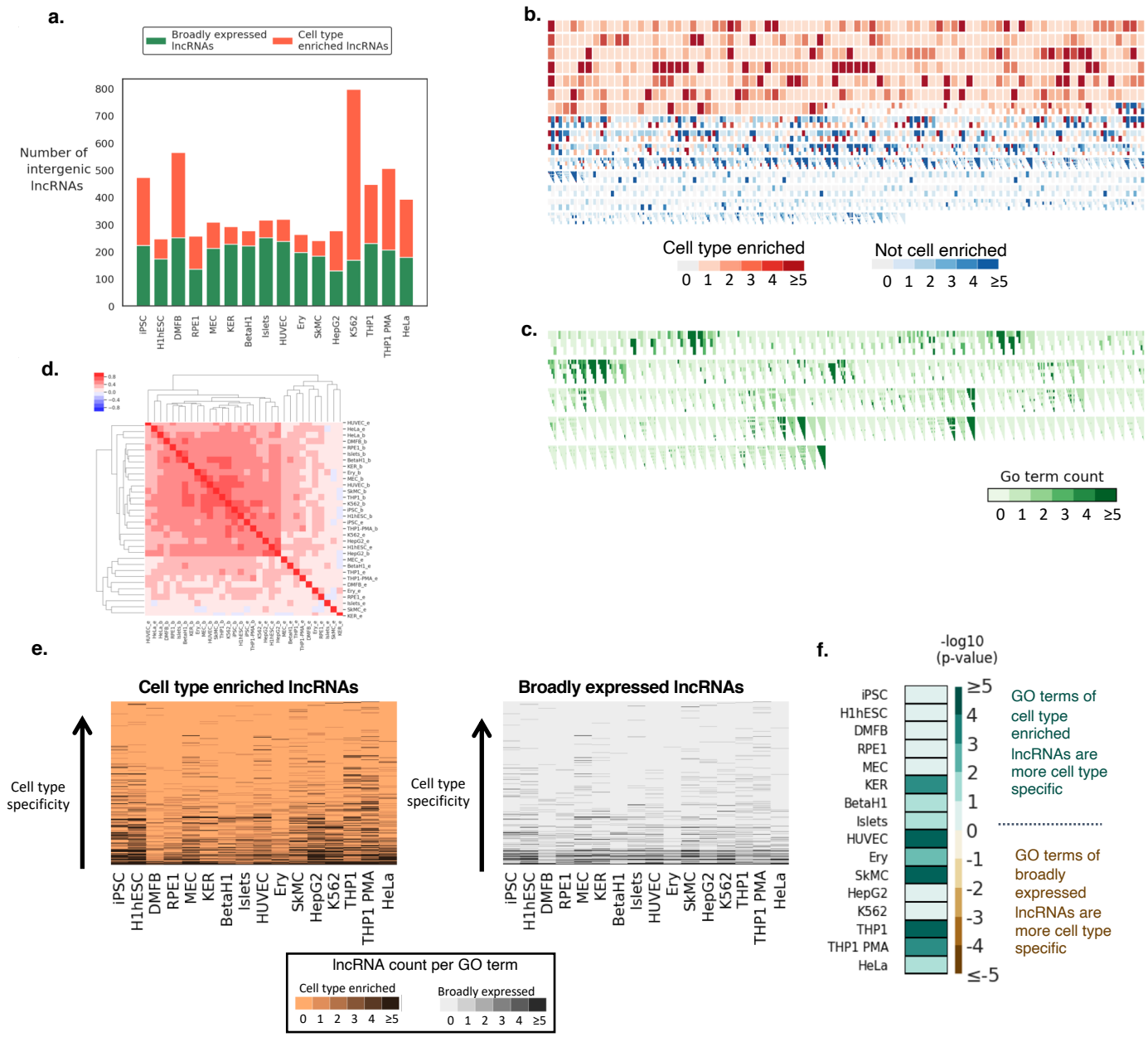


Figure 5: Overall patterns for the cell type-enriched and broadly expressed lncRNAs. a. The number of broadly expressed and cell type-enriched lncRNAs in each cell type. **b.** Count of significant biological processes GO terms associated with cell type enriched lncRNA clusters that overlap between the cell types. Each box represents a heatmap for one lncRNA. The count of significant GO term overlapped between two cell types if both are enriched for lncRNA are shown in red, while if one or both are relatively depleted, then overlapped GO term is shown in blue. The diagonal value represents the number of significant GO terms in each cell type **c.** Count of GO terms associated with broadly expressed lncRNA clusters that overlap between the cell types. Each box represents a eat map for one lncRNA. The diagonal value represents the number of significant GO terms in each cell type, while the rest of the triangle is an overlap between the cell types. **d.** Heatmap showing Spearman correlation between cluster counts for each motif. The clusters were separated based on broadly expressed (labeled _b) and cell type-enriched (labeled _e) lncRNA for each cell type before correlation calculation. **e.** Heatmap showing the cell type specificity of the GO term vs the number of lncRNA for cell type enriched and broadly expressed lncRNAs. Each row is a GO term, and each column is cell type, and values are lncRNA count for that GO term. GO terms are sorted in descending order of cell type specificity. **f.** Comparison of cell type specificity of GO terms associated with cell type enriched and broadly expressed lncRNAs using one-tailed Mann-Whitney test.

a.

K562

		DNA binding only	Overlap	RNA binding only	
FUS	41	172	147	208	79
HNRNPK	31	161	127	440	78
PCBP1	20	174	119	603	58
NONO	32	121	91	349	107
RBFOX2	9	51	76	302	26
RBM15	11	81	39	410	77
GTF2F1	11	156	23	416	76
TARDBP	6	309	14	82	4
SAFB	16	232	13	13	9
ILF3	27	212	13	287	61
YBX3	1	63	13	162	13
LARP7	0	289	2	51	1
PTBP1	8	234	2	337	43
TAF15	2	82	1	222	36
ZC3H8	1	327	0	0	0
HLTF	1	303	0	8	2
RBM22	0	286	0	0	0
HNRNPL	21	280	0	25	11
ZC3H11A	3	262	0	72	5
SRSF1	0	163	0	58	1
KHSRP	3	158	0	15	2
HNRNPUL1	4	124	0	152	20
SRSF7	1	49	0	21	1
EWSR1	0	32	0	462	104
SAFB2	4	27	0	7	3

HepG2

		DNA binding only	Overlap	RNA binding only	
HNRNPK	66	215	184	436	128
PCBP1	10	380	167	507	43
PCBP2	11	296	86	381	43
PTBP1	12	542	63	158	16
GTF2F1	41	509	54	353	68
PRPF4	16	255	50	615	71
RBM22	6	439	47	80	8
SRSF1	1	573	43	42	0
U2AF1	0	626	19	41	0
RBFOX2	3	7	18	360	30
SFPQ	7	143	11	249	39
FUS	2	405	7	0	0
U2AF2	2	516	3	11	1
HNRNPL	19	521	0	7	6
TAF15	5	467	0	0	0
HNRNPUL1	3	308	0	153	12

b.

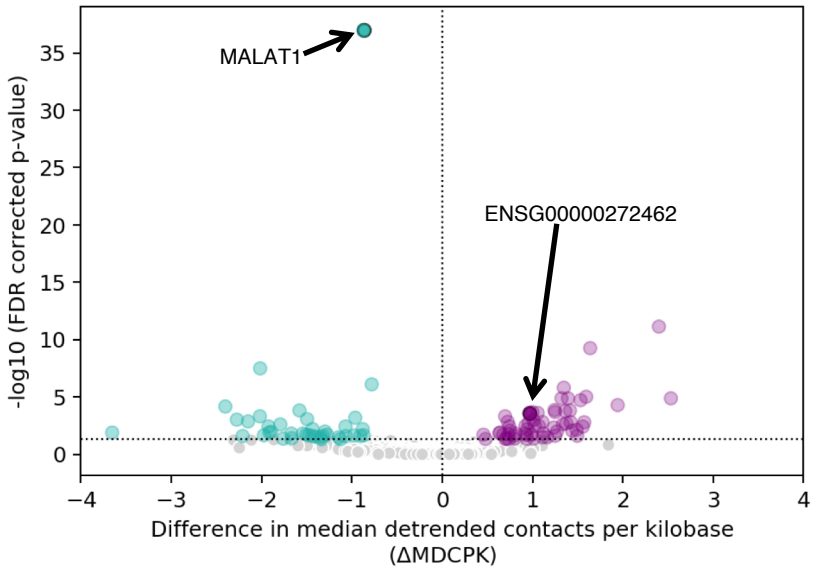
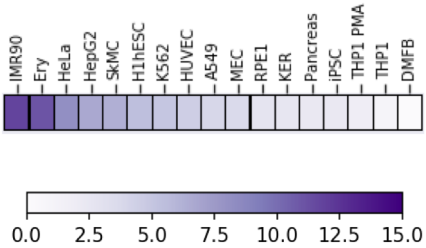
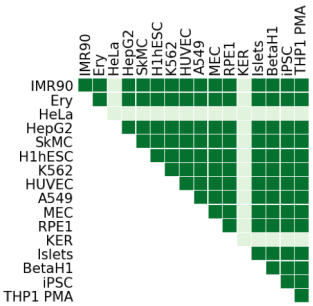
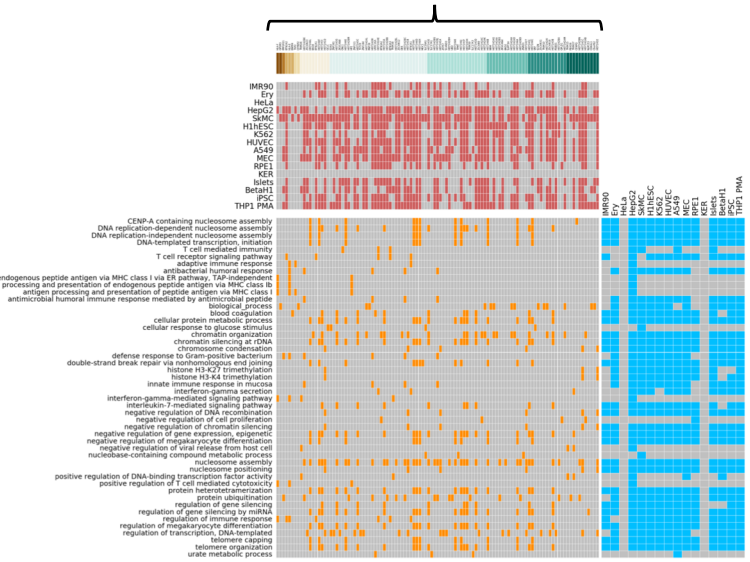


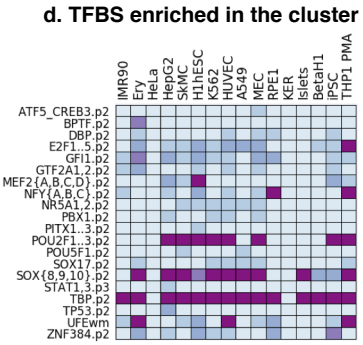
Figure 6: RBP and Red-C enrichment analysis for the selected cell types. a. Each row represents an RBP that has ChIP-seq and eCLIP or only ChIP-seq data. The numbers in each row are the number of clusters that have RBP enrichment for DNA binding only (enrichment for RBP DNA binding sites at promoter and enhancers), RNA binding only (enrichment for RBP eCLIP peak at RNA transcripts), or both (enrichment for RBP ChIP-Seq peaks and eCLIP peaks). The number of clusters for which the reference lncRNA has an RNA binding site for the RBP while Hi-C cluster genes show enrichment for DNA binding sites is shown in green (left). The number of clusters for which the reference lncRNA has an RNA binding site for the RBP while Hi-C cluster gene transcripts show enrichment for RNA binding sites in blue (right). **b.** Scatter plot illustrating results of the Wilcoxon-Mann-Whitney test comparing detrended RNA-DNA contact densities. Each point represents one reference lncRNA. The purple dots represent lncRNAs that have higher RNA-DNA contact density inside the Hi-C cluster compared to outside while green dots represent lncRNAs that have higher RNA-DNA contact density outside the Hi-C cluster compared to in the Hi-C cluster. The example ENSG00000272462 is highlighted by a black circle.

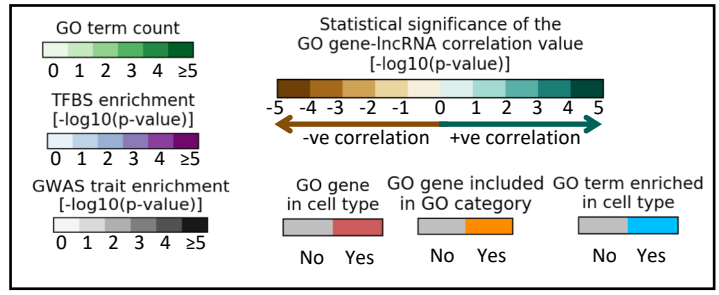
ENSG00000272462











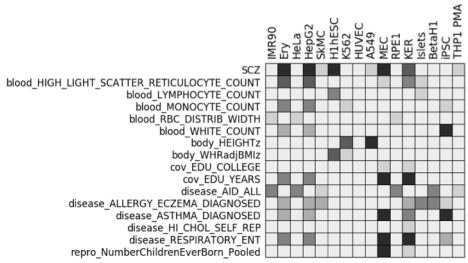
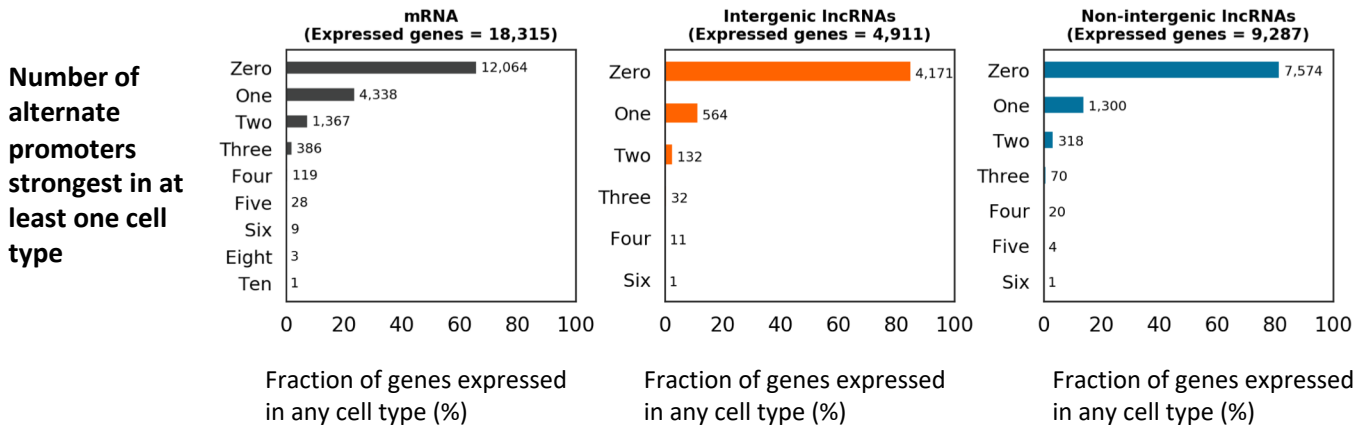


Figure 7: Annotation results for lncRNA ENSG00000272462. a. Expression of lncRNA in all 17 cell types, **b.** The number of significant GO terms overlapping between the cell types in which the lncRNA is expressed. The number of significant GO terms in each cell type is shown on the diagonal. **c.**GO annotation results. The topmost heatmap shows the expression correlation between the lncRNA and GO-annotated genes in the Hi-C gene cluster. For readability, the gene names are shown on the top of the heatmap in the same order as in the heatmap. The middle heatmap shows whether the Hi-C gene cluster in each cell contains each of the GO-annotated genes . The bottom heatmap shows the membership of each GO gene in each GO category, and the heatmap on the right shows whether each GO term is enriched in the each cell type. d**.** TFBS enriched (p-value <=0.1) in the Hi-C gene cluster. **e.** GWAS trait enriched (p-value <=0.05) in the A/B compartment overlaps with the Hi-C GO cluster.



Supplementary Figure 1: Number of genes with alternative strongest promoters. For each gene, we identified the strongest promoter in each cell type. For most genes, the same promoter was found in each cell type This graph shows the number of alternative strongest promoters across 18 cell types for the mRNA coding (grey), intergenic lncRNA (orange) and non-intergenic lncRNAs (blue) gene classes. The total number of expressed genes of each gene class across the cell types is shown in the title of each graph. The percentage of genes on the x-axis, with the absolute number of genes is shown next to each bar.

mRNA

Number of mRNA coding promoters in a 10kb window

	iPSC	H1hESC	DMFB	RPE1	MEC	KER	BetaH1	Islets	IMR90	HUVEC	Ery	SkMC	A549	HepG2	K562	THP1	THP1 PMA	HeLa
One	10266	10542	9278	10166	9491	9516	9646	9922	9341	9409	7819	9841	9420	8771	8350	9255	9383	8656
Two	1780	1674	1467	1530	1516	1463	1449	1486	1542	1577	1327	1637	1467	1604	1593	1516	1632	1641
Three	149	157	135	129	127	138	112	124	143	133	120	154	131	142	130	128	135	145
Four	23	21	18	18	17	16	16	15	18	18	12	21	15	18	21	17	21	20
Five	4	4	3	4	3	3	3	3	5	3	4	4	3	4	4	3	4	4
Six	2	3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Seven	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
Eight	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0

Gene count: 2500, 2000, 1500, 1000, 500, 0

Intergenic lncRNAs

Number of intergenic lncRNAs promoters in a 10kb window

	iPSC	H1hESC	DMFB	RPE1	MEC	KER	BetaH1	Islets	IMR90	HUVEC	Ery	SkMC	A549	HepG2	K562	THP1	THP1 PMA	HeLa
One	993	1042	410	577	655	630	411	476	554	642	429	653	609	728	842	569	941	849
Two	90	87	32	27	42	26	24	26	34	57	36	46	25	60	73	42	96	91
Three	1	5	1	1	5	5	0	0	2	0	0	2	1	3	2	1	2	5
Four	1	0	1	1	1	3	2	2	1	1	1	2	0	2	1	1	1	2
Five	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Gene count: 500, 400, 300, 200, 100, 0

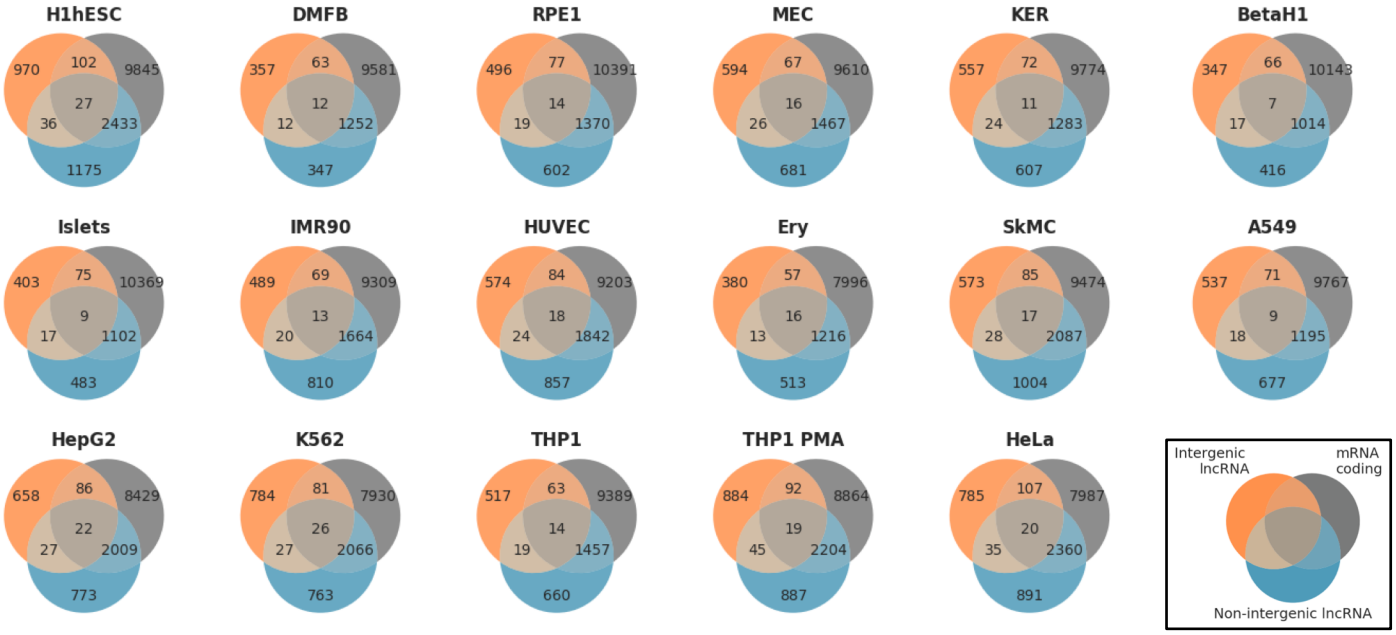
Non-intergenic lncRNAs

Number of other lncRNAs promoters in a 10kb window

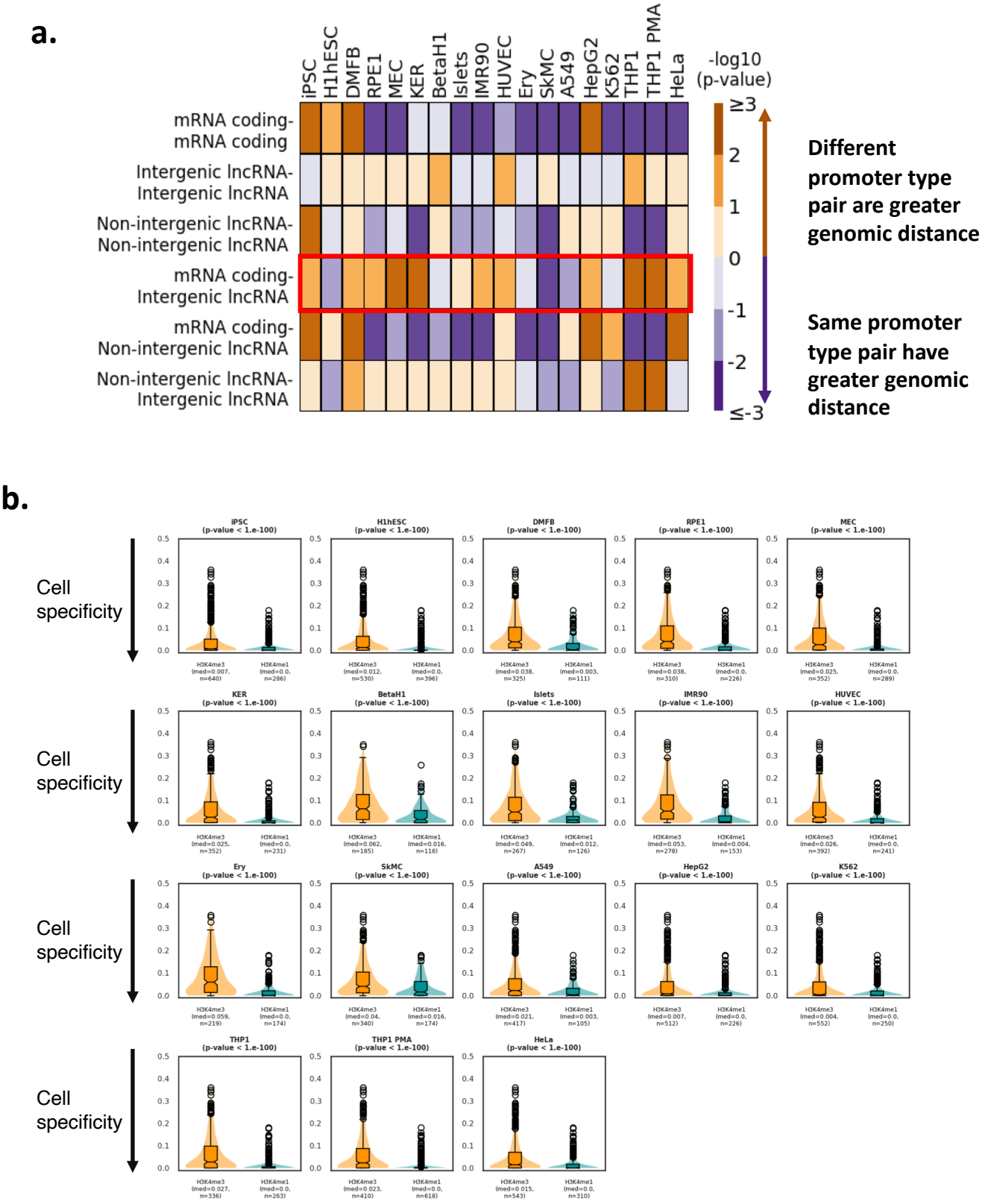
	iPSC	H1hESC	DMFB	RPE1	MEC	KER	BetaH1	Islets	IMR90	HUVEC	Ery	SkMC	A549	HepG2	K562	THP1	THP1 PMA	HeLa
One	3765	3554	1580	1955	2121	1867	1420	1571	2437	2663	1710	3024	1854	2735	2805	2085	3042	3189
Two	99	113	42	46	66	57	33	39	68	76	45	108	44	93	75	60	108	113
Three	1	4	1	4	3	1	1	1	2	1	3	3	1	3	2	5	5	4
Four	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0

Gene count: 2500, 2000, 1500, 1000, 500, 0

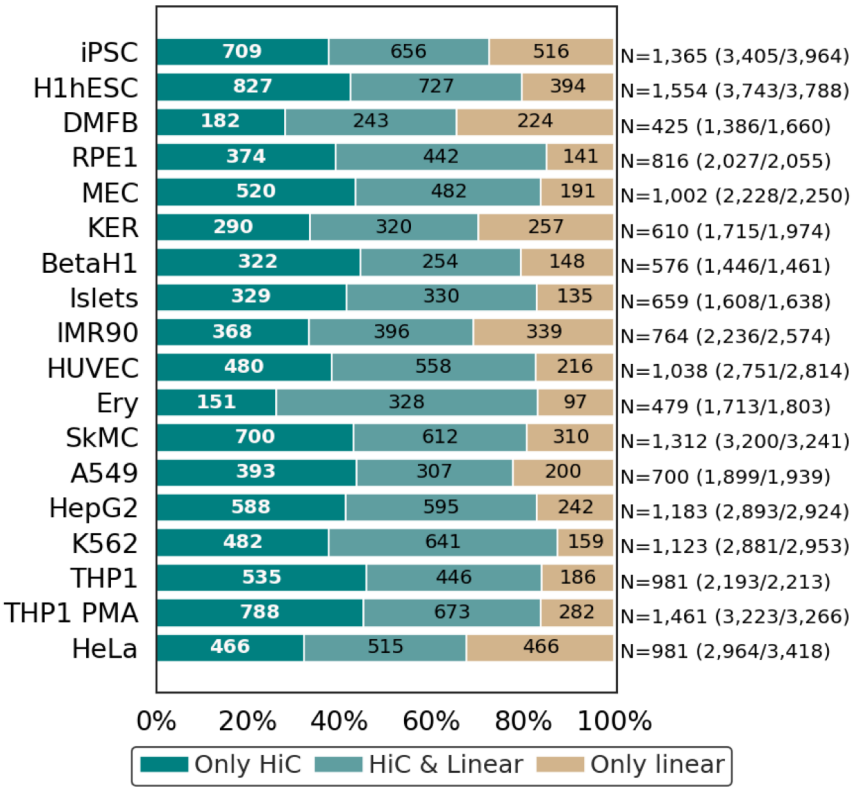
Supplementary Figure 2: Number of promoters of a geneclass in a 10 kb window. From top to bottom, heatmap showing number of promoters in a 10kb windows for mRNA (grey), intergenic lncRNA (orange) and non-intergenic lncRNA (blue) gene classes in selected 18 cell types. Each column represents a cell type and each row is the number of promoters in a 10kb window for that gene class. The number in the cells is the count of genes. Regardless of gene class, most 10kb windows contain only one promoter.



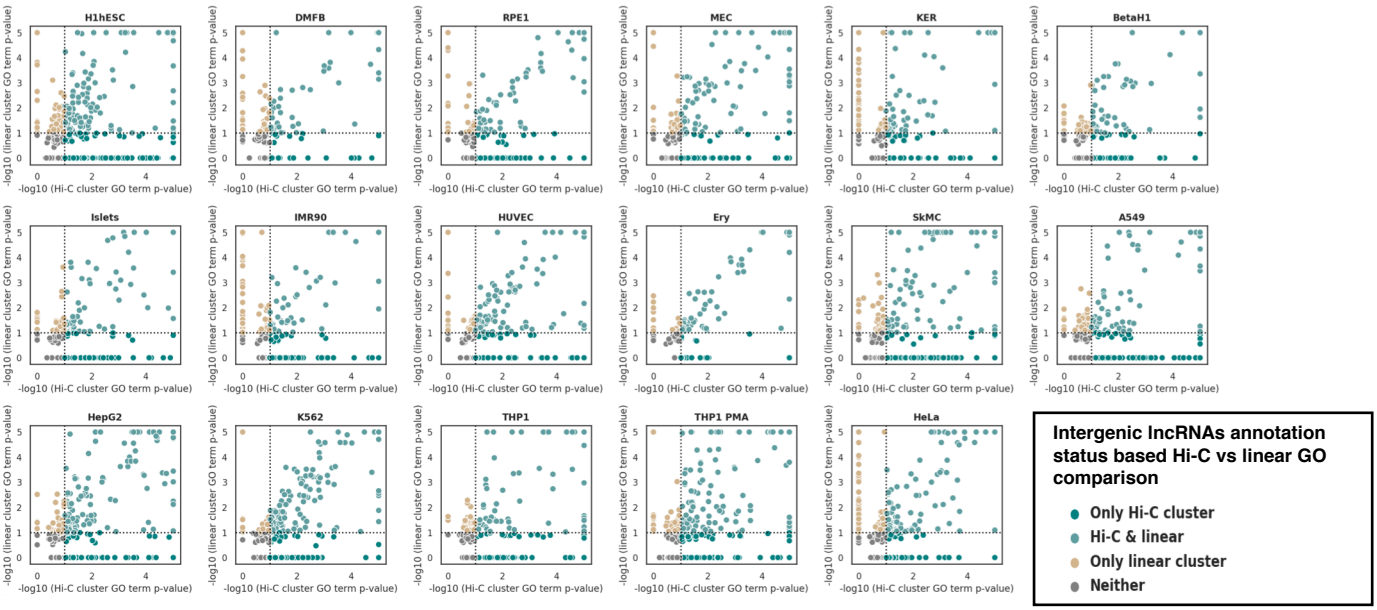
Supplementary Figure 3: Number of promoters of different gene classes occurring in the same 10 kb window. As in case of iPSc (Figure 2c), in all cell types intergenic lncRNA and mRNA coding promoters rarely occur in the same 10kb genomic window, while most non-intergenic lncRNAs have at least one mRNA coding promoter in the same window.



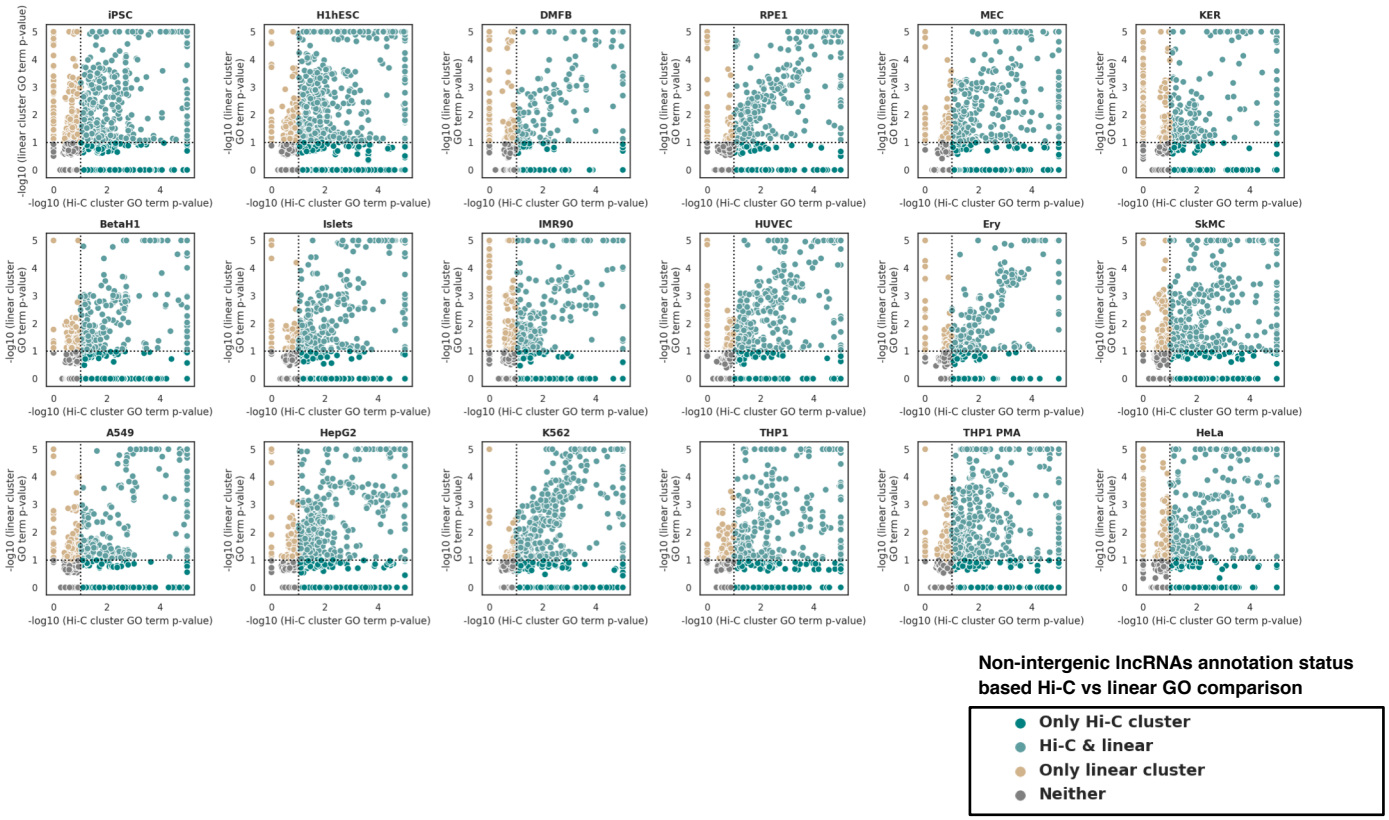
Supplementary Figure 4: a) The heatmap showing the significance of the difference between the same and different promoter types pair distance for all the gene class pairs in 18 cell types (p-value determined using one-tailed Mann-Whitney test). **b)** Cell specificity for the of intergenic lncRNA promoters in selected cell types. The number of promoters for each type and median cell specificity is given in parenthesis.



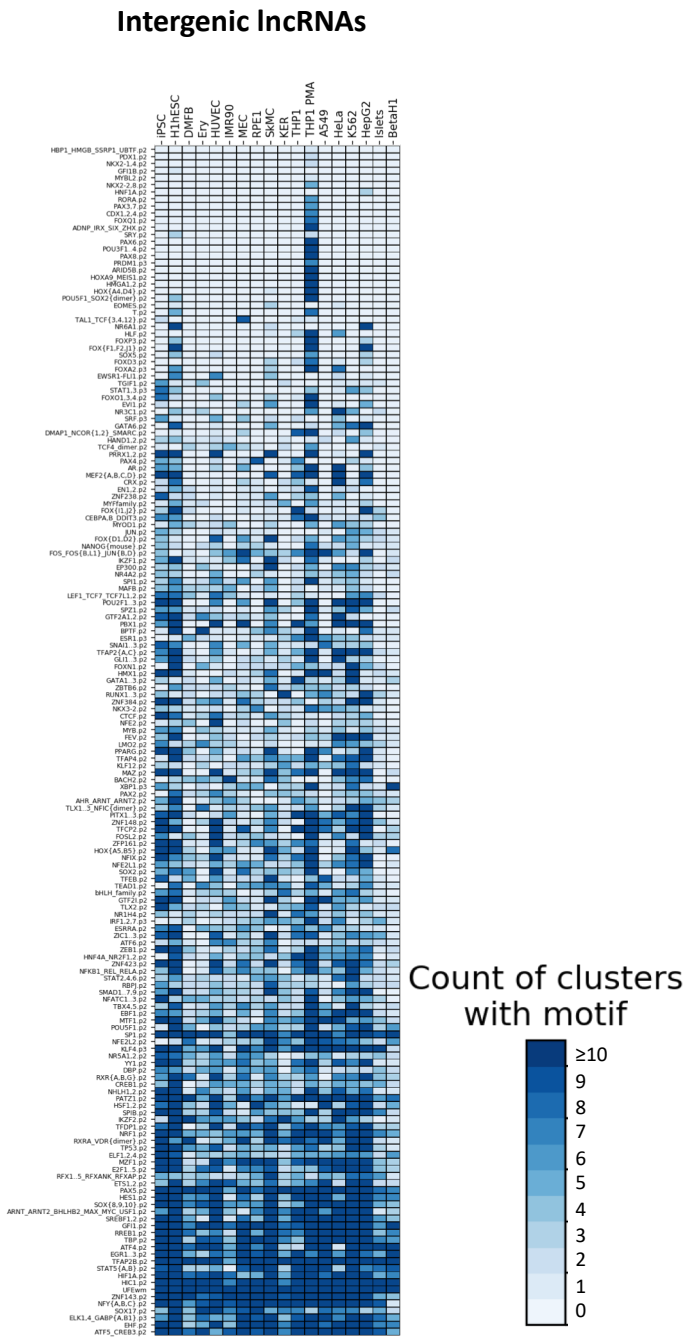
Supplementary Figure 5: GO annotation of non-intergenic lncRNAs with Hi-C gene clusters or linear gene clusters. Comparison using the most significant biological processes GO term (FDR corrected p-value <=0.1), Each row corresponds to one cell type, with the number of lncRNAs for each category shown in the bars. The number of non-intergenic lncRNAs annotated with at least one GO term supported by Hi-C clusters is shown on the right. In parentheses the number of non-intergenic lncRNAs tested for GO term enrichment / the number of non-intergenic lncRNAs expressed in that cell type.



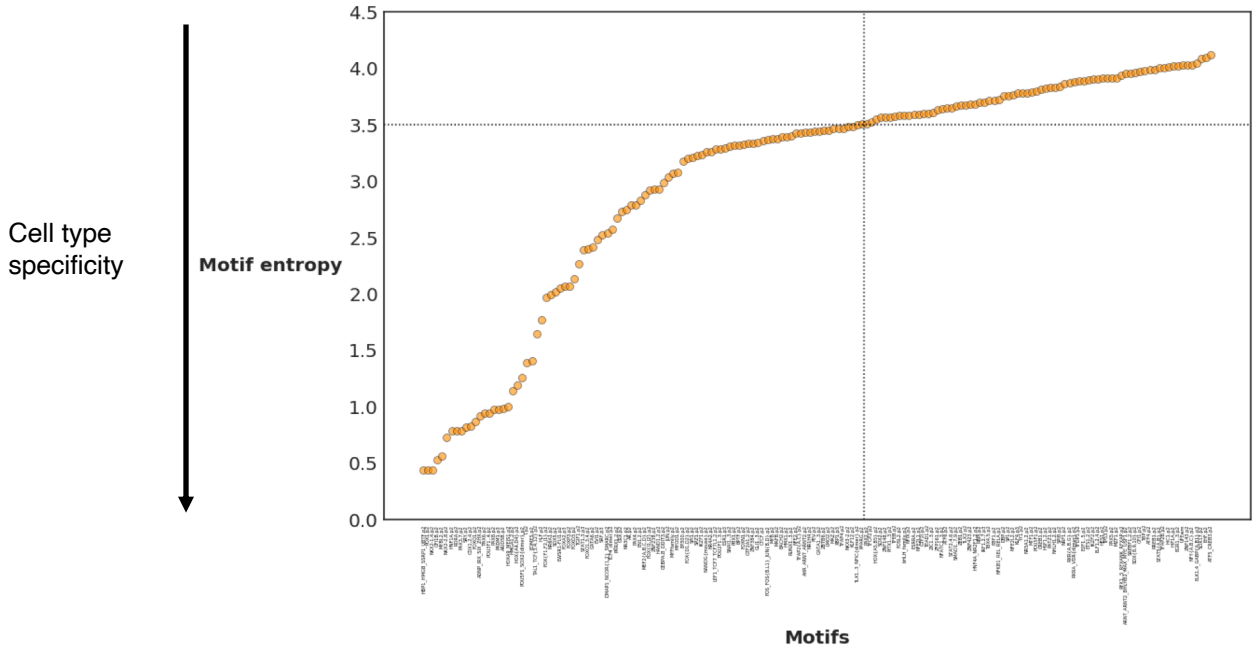
Supplementary Figure 6: Comparison of GO annotation of intergenic lncRNAs based on Hi-C gene clusters or linear gene clusters. The horizontal and vertical axis show the statistical significance for the most significant biological processes GO term using Hi-C and linear gene clusters, respectively, using a FDR corrected p-value threshold of 0.1 indicated by the dotted line. Each panel shows one cell type, with a dot representing one lncRNA, colored based on its annotation status as defined in the legend. The number of lncRNAs is shown in parentheses in the legend



Supplementary Figure 7: Comparison of GO annotation of non-intergenic lncRNAs based on Hi-C gene clusters or linear gene clusters. The horizontal and vertical axis show the statistical significance for the most significant biological processes GO term using Hi-C and linear gene clusters, respectively, using a FDR corrected p-value threshold of 0.1 indicated by the dotted line. Each panel shows one cell type, with a dot representing one lncRNA, colored based on its annotation status as defined in the legend. The number of lncRNAs is shown in parentheses in the legend



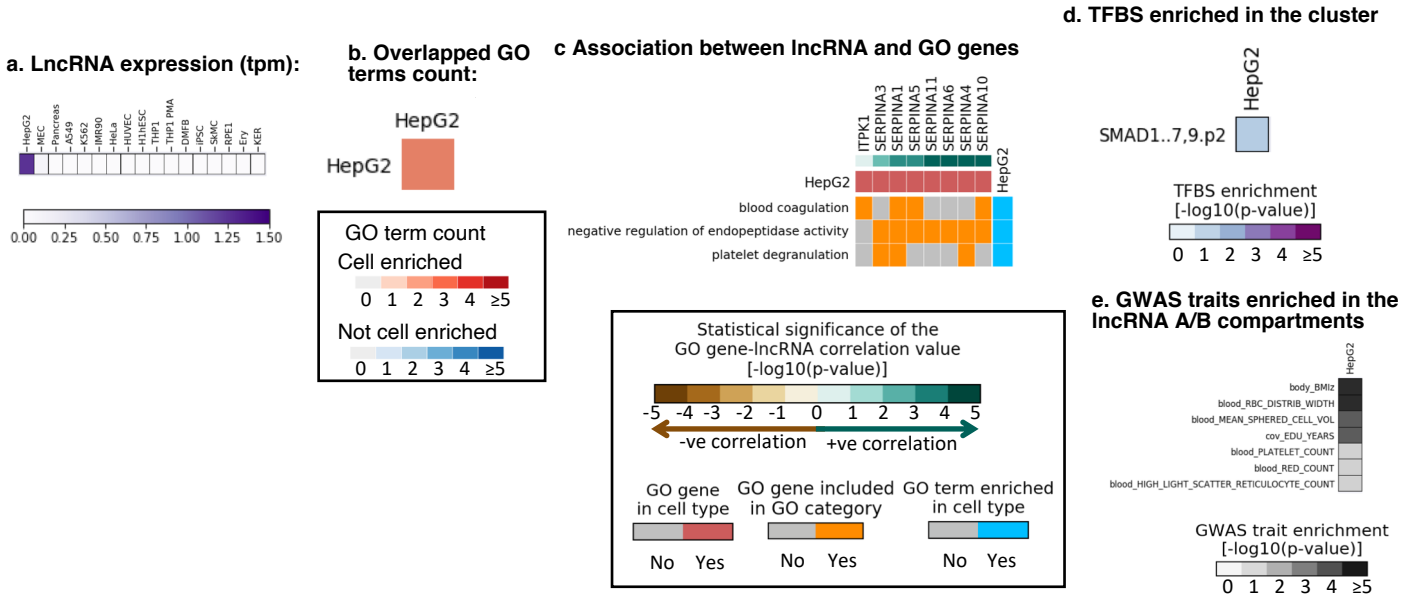
Supplementary Figure 8: Count of lncRNA clusters that show enrichment for each motif in individual cell type. Each row represent one motif and each column is one cell type. Each box represent cluster count for the motif. The example of motifs motifs specific cell type is highlighted by red boxes.



1. HBP1_HMGB_SSRP1_UBTF.p2
2. PDX1.p2
3. NKX2-1,4.p2
4. GFI1B.p2
5. MYBL2.p2
6. NKX2-2,8.p2
7. HNF1A.p2
8. RORA.p2
9. PAX3,7.p2
10. SRY.p2
11. CDX1,2,4.p2
12. FOXQ1.p2
13. ADNP_IRX_SIX_ZHX.p2
14. PAX6.p2
15. POU3F1..4.p2
16. PAX8.p2
17. PRDM1.p3
18. ARID5B.p2
19. HOXA9_MEIS1.p2
20. HMGA1,2.p2
21. HOX{A4,D4}.p2
22. POU5F1_SOX2{dimer}.p2
23. T.p2
24. EOMES.p2
25. TAL1_TCF{3,4,12}.p2
26. HLF.p2
27. FOX{F1,F2,J1}.p2
28. NR6A1.p2
29. SOX5.p2
30. EWSR1-FLI1.p2
31. FOXA2.p3
32. FOXP3.p2
33. FOXD3.p2
34. TGIF1.p2
35. STAT1,3.p3
36. FOXO1,3,4.p2
37. GATA6.p2
38. EVI1.p2
39. SRF.p3
40. DMAP1_NCOR{1,2}_SMARC.p2
41. TCF4_dimer.p2
42. PRRX1,2.p2
43. CRX.p2
44. NR3C1.p2
45. AR.p2
46. PAX4.p2
47. EN1,2.p2
48. MEF2{A,B,C,D}.p2
49. FOX{I1,J2}.p2
50. ZNF238.p2
51. HAND1,2.p2
52. CEBPA,B_DDIT3.p2
53. JUN.p2
54. MYFfamily.p2
55. MYOD1.p2
56. EP300.p2
57. FOX{D1,D2}.p2
58. SPI1.p2
59. SPZ1.p2
60. IKZF1.p2
61. NANOG{mouse}.p2
62. NR4A2.p2
63. LEF1_TCF7_TCF7L1,2.p2
64. POU2F1..3.p2
65. ESR1.p3
66. SNAI1..3.p2
67. PBX1.p2
68. BPTF.p2
69. FOXN1.p2
70. GTF2A1,2.p2
71. ZNF384.p2
72. GLI1..3.p2
73. CTCF.p2
74. FOS_FOS{B,L1}_JUN{B,D}.p2
75. MYB.p2
76. MAFB.p2
77. BACH2.p2
78. HMX1.p2
79. RUNX1..3.p2
80. NFE2.p2
81. TFAP2{A,C}.p2
82. AHR_ARNT_ARNT2.p2
83. NR1H4.p2
84. FEV.p2
85. GATA1..3.p2
86. ZBTB6.p2
87. LMO2.p2
88. MAZ.p2
89. XBP1.p3
90. TFAP4.p2
91. NKX3-2.p2
92. KLF12.p2
93. TLX1..3_NFIC{dimer}.p2

1. PPARG.p2
2. PAX2.p2
3. TFCP2.p2
4. HOX{A5,B5}.p2
5. SOX2.p2
6. ZNF148.p2
7. PITX1..3.p2
8. TFEB.p2
9. FOSL2.p2
10. bHLH_family.p2
11. NFIX.p2
12. ESRRA.p2
13. NFE2L1.p2
14. GTF2I.p2
15. TEAD1.p2
16. ZIC1..3.p2
17. ZFP161.p2
18. NFATC1..3.p2
19. ATF6.p2
20. STAT2,4,6.p2
21. SMAD1..7,9.p2
22. ZEB1.p2
23. TLX2.p2
24. ZNF423.p2
25. HNF4A_NR2F1,2.p2
26. RBPJ.p2
27. IRF1,2,7.p3
28. TBX4,5.p2
29. EBF1.p2
30. NFKB1_REL_RELA.p2
31. DBP.p2
32. SP1.p2
33. NFE2L2.p2
34. KLF4.p3
35. YY1.p2
36. NR5A1,2.p2
37. MTF1.p2
38. POU5F1.p2
39. CREB1.p2
40. HSF1,2.p2
41. PATZ1.p2
42. NHLH1,2.p2
43. SPIB.p2
44. NRF1.p2
45. RXR{A,B,G}.p2
46. RXRA_VDR{dimer}.p2
47. TFDP1.p2
48. E2F1..5.p2
49. ETS1,2.p2
50. IKZF2.p2
51. ELF1,2,4.p2
52. HES1.p2
53. TP53.p2
54. PAX5.p2
55. MZF1.p2
56. RFX1..5_RFXANK_RFXAP.p2
57. ARNT_ARNT2_BHLHB2_MAX_MYC_USF1.p2
58. SREBF1,2.p2
59. SOX{8,9,10}.p2
60. GFI1.p2
61. TBP.p2
62. ATF4.p2
63. RREB1.p2
64. STAT5{A,B}.p2
65. TFAP2B.p2
66. HIC1.p2
67. HIF1A.p2
68. EGR1..3.p2
69. UFEwm
70. ZNF143.p2
71. NFY{A,B,C}.p2
72. ELK1,4_GABP{A,B1}.p3
73. SOX17.p2
74. EHF.p2
75. ATF5_CREB3.p2

Supplementary Figure 9: TFBS entropy. Scatter plot showing motif entropy. Each point represents one TFBS. The horizontal dotted line shows the entropy cutoff 3.5. The left box shows the TFBS with entropy ≤ 3.5 while right box show the TFBS with entropy > 3.5.



Supplementary Figure 10: Examples of lncRNA characterization based on their Hi-C gene clusters. LncRNA CATG00000020001 **a.** showing expression of lncRNA in all the 17 cell types, **b.** The number of significant GO terms in the cell type in which the lncRNA is expressed, i.e HepG2, **c.** The topmost heatmap shows the expression correlation between lncRNA and GO-annotated genes in the clusters (in this case all of them have positive correlation). The middle heatmap shows if GO genes are part of the Hi-C cluster GO term in each cell type. The bottom heatmap shows whether the GO gene is a member of each GO term category, and the heatmap on the right shows if the GO term is enriched in the Hi-C cluster in each cell. **d.** TFBS enriched (p-value <=0.1) in the Hi-C gene cluster. **e.** GWAS trait enriched (p-value <=0.05) in the A/B compartment overlapping with the Hi-C gene cluster.