HHS Public Access

Author manuscript

Cell Rep. Author manuscript; available in PMC 2020 May 02.

Published in final edited form as:

Cell Rep. 2020 March 10; 30(10): 3296–3311.e5. doi:10.1016/j.celrep.2020.02.048.

Genomic Repeats Categorize Genes with Distinct Functions for Orchestrated Regulation

J. Yuyang Lu[1], **Wen Shao**[1,6], **Lei Chang**[2,6], **Yafei Yin**[1,6], **Tong Li**[1,6], **Hui Zhang**[1,6], **Yantao Hong**[1,6], **Michelle Percharde**[3,4], **Lerui Guo**[1], **Zhongyang Wu**[1], **Lichao Liu**[1], **Wei Liu**[1], **Pixi Yan**[1], **Miguel Ramalho-Santos**[5], **Yujie Sun**[2], **Xiaohua Shen**[1,7,*]

[1]Tsinghua Center for Life Sciences, School of Medicine and School of Life Sciences, Tsinghua University, Beijing 100084, China

[2]State Key Laboratory of Membrane Biology, School of Life Sciences, and Biomedical Pioneering Innovation Center (BIOPIC), Peking University, Beijing 100871, China

[3]MRC London Institute of Medical Sciences (LMS), London W120NN, UK

[4]Institute of Clinical Sciences (ICS), Faculty of Medicine, Imperial College London, London W120NN, UK

[5]Lunenfeld-Tanenbaum Research Institute and Department of Molecular Genetics, University of Toronto, Toronto, ON M5T 3H7, Canada

[6]These authors contributed equally

[7]Lead Contact

SUMMARY

Repetitive elements are abundantly distributed in mammalian genomes. Here, we reveal a striking association between repeat subtypes and gene function. SINE, L1, and low-complexity repeats demarcate distinct functional categories of genes and may dictate the time and level of gene expression by providing binding sites for different regulatory proteins. Importantly, imaging and sequencing analysis show that L1 repeats sequester a large set of genes with specialized functions in nucleolus- and lamina-associated inactive domains that are depleted of SINE repeats. In addition, L1 transcripts bind extensively to its DNA in embryonic stem cells (ESCs). Depletion of L1 RNA in ESCs leads to relocation of L1-enriched chromosomal segments from inactive domains to the nuclear interior and de-repression of L1-associated genes. These results demonstrate a role of L1 DNA and RNA in gene silencing and suggest a general theme of genomic repeats in orchestrating the function, regulation, and expression of their host genes.

*Correspondence: xshen@tsinghua.edu.cn.

AUTHOR CONTRIBUTIONS

X.S. supervised the study. X.S. and J.Y.L. conceived of and designed the experiments. J.Y.L. performed bioinformatics analysis. W.S. conducted nucleolar DNA sequencing (DNA-seq). L.C. performed DNA FISH and Oligopaint FISH. Y.Y. performed L1 ChIRP-seq. T.L. and L.L. analyzed L1 RNA half-life. H.Z. performed electron microscopy and *NCL* knockdown and RNA-seq with Z.W. Y.H., L.G., and J.Y.L. performed NCL ChIP. M.P. and M.R.-S. contributed to early experiments of L1 ASO. Y.S., P.Y., and W.L. contributed technical assistance and suggestions. X.S. and J.Y.L. wrote the manuscript with input from all authors.

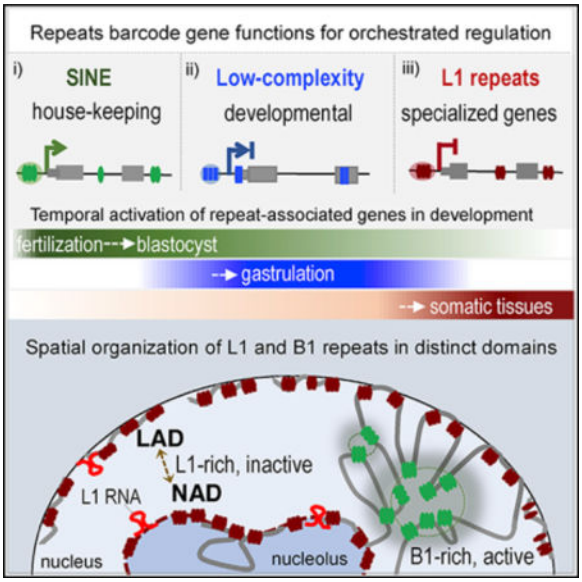
DECLARATION OF INTERESTS

The authors declare no competing interests.

In Brief

Lu et al. report a striking association between genomic repeats and gene regulation and demonstrate a key role of L1 repeat RNA in sequestering L1-rich sequences and associated genes in inactive domains for silencing, revealing a general theme of repeat sequences in shaping gene regulatory networks within their host genome.

Graphical Abstract



INTRODUCTION

Repetitive sequences, comprising transposable elements and simple repeats, constitute up to 45% of the genome in mouse and 50%–70% in human (Biémont, 2010; de Koning et al., 2011). On the basis of transposition mechanisms, transposable elements can be divided into DNA transposons and retrotransposons. The latter are predominant in most mammals and can be further divided into long terminal repeat (LTR)-containing endogenous retrovirus (ERV) transposons and non-LTR transposons (including short interspersed nuclear elements [SINEs] and long interspersed nuclear elements [LINEs]) (Rebollo et al., 2012). The most abundant subclass of SINEs comprises primate-specific Alu elements in human and the closely related B1 repeats in mouse, which are ~300 nt in length and more abundant in GC-rich DNA. Mice and humans have up to 0.6 million and 1.4 million copies, respectively, of these repeats, which constitute about 2.7% or 10.6% of the genomic DNA (Lander et al., 2001; Waterston et al., 2002). Long interspersed element-1 (LINE1 or L1), which are 6–7 kb in length and abundant in AT-rich DNA, constitute 19% and 17% (0.9 million to 1.0 million copies) of the genome in mouse and human, respectively, and make up the largest proportion of transposable element-derived sequences (Taylor et al., 2013).

Repetitive elements were once regarded as junk or “parasite” DNA (Doolittle and Sapienza, 1980; Orgel and Crick, 1980), but increasing lines of evidence have gradually revised and expanded our understanding of genomic repeats and how they influence mammalian

genomes. Genomic repeats may influence host gene expression at both transcriptional and post-transcriptional levels through cis and trans mechanisms and participate in the regulation of diverse biological and pathological processes (Boeva, 2016; Bourque et al., 2008; Carrieri et al., 2012; Chuong et al., 2016; Durruthy-Durruthy et al., 2016; Grow et al., 2015; Kunarso et al., 2010; Lynch et al., 2011; Muotri et al., 2010). For example, short tandem repeats contribute to gene expression variations and the genetic architecture of quantitative human traits (Gymrek et al., 2016, 2017). ERV1 and HERVH harbor DNA binding sites for the transcription factors POU5F1, NANOG, and STAT1 and have been implicated in stem cell pluripotency and innate immunity (Chuong et al., 2016; Kunarso et al., 2010). SINE repeats carry new binding sites for CTCF and may serve as boundary elements to influence chromatin structure and transcription (Lunyak et al., 2007; Schmidt et al., 2012). L1 repeats regulate global chromatin accessibility at the beginning of development, and embryos are arrested at the two-cell stage if L1 activation and silencing are disrupted (Jachowicz et al., 2017). In mouse embryonic stem cells (ESCs), L1 RNA facilitates the binding of nucleolin (NCL) and the nuclear corepressor KRAB-associated protein-1 (KAP1 or TRIM28) to ribosomal DNA (rDNA) and *DUX* gene loci to promote rRNA transcription or to repress a transcriptional program specific to the two-cell embryo, respectively (Percharde et al., 2018). Because knockout of *DUX* causes minor defects in zygotic genome activation (ZGA) and is compatible with mouse development (Chen and Zhang, 2019), we speculate a more extensive role of L1 repeats beyond regulation of *DUX* gene.

Despite these initial findings, our current knowledge of how repetitive sequences shape the structure and function of the genome is still limited. The extent to which the function of genomic repeats can be generalized regardless of biological context is poorly understood. Delineation of the roles of individual repeat subfamilies in gene regulation is still lacking. Here, we conducted a comprehensive and quantitative analysis of diverse repeat subclasses in mouse and human genomes and revealed a striking association of genic repeats with the function, regulation, and expression of their host genes. Importantly, we demonstrate a key role of L1 RNA in transacting L1 DNA information and in sequestering a large set of genes that are specialized in functions associated with terminally differentiated cells, in the heterochromatic nucleolar and nuclear peripheries for transcriptional silencing in ESCs. These results reveal a general theme of repeat sequences in shaping gene regulatory networks within their host genome.

RESULTS

Non-random Distribution of Repetitive Elements

We surveyed the mouse and human genomes and found that significant proportions (~21%–73%) of individual repeat subclasses reside near a gene, which often harbors several subclasses of repeats at different frequencies. More than 72% of B1/Alu and 42%–59% of L1 repeats are positioned within ±10 kb of a gene in human and mouse and 59% of B1/Alu and 32%–48% of L1 within ±2 kb of a gene (Figures 1A, S1A, and S1B). In the example of the four genes encoding the proteins RPS15 (a ribosomal protein), OLFR441 (an olfactory receptor), FGF5 (a transcription factor), and ZFP72 (a zinc finger protein), they have drastically different repeat compositions (Figure 1B). Although they all contain SINEs in

their promoters, it is clear that these four genes harboring fewer or a cluster of SINE elements belong to distinct functional categories. This observation suggests that the simple definition of a gene as “repeat containing” would obscure potential regulatory differences that are conferred by different repeat subclasses.

Heatmap plotting of repeat content in genic regions (±2 kb of a gene) revealed differential distributions of various repeats in mouse and human genomes (Figures 1C, S1C, and S1D). First, SINEs are highly enriched in mRNA genes compared with long non-coding RNAs (lncRNAs) in regulatory regions, including “promoter,” “intron,” and “downstream” (highlighted with the pink box). Second, L1 appears to be more enriched in the genome background than genic regions. Third, lncRNAs have a higher content of ERVs in exons and downstream regions compared with mRNA genes (highlighted with the orange box), consistent with a previous report (Kelley and Rinn, 2012). Fourth, low-complexity and simple repeats are strongly enriched in the promoters of mRNA and lncRNA genes, while low-complexity sequences are also enriched in the 5′ UTRs of mRNAs (highlighted with the green box). Fifth, satellite repeats are the only repeat subclass that is strongly enriched in the CDS of mouse mRNAs. These observations indicate a non-random distribution of genic repeats in mouse and human genomes and suggest a potential link between repeat features and functions of repeat-containing host genes.

Genic Repeats Categorize Gene Function

To further explore how repeat composition and distribution are related to gene function, we calculated the DNA content for each of 14 repeat subclasses that fall into six genic regions for 22,432 protein-coding genes in mouse (Figure S2). Genes that were grouped by their repeat features are more likely to be enriched in particular functions than random groups (Figures 1D and S1D). Hierarchical clustering revealed four prominent clusters of genes that harbor distinct repeat subtypes (Figures 1E and S3A): (1) a set of 2,041 genes that are enriched in SINE repeats (B1, B2, and B4) in regulatory regions (designated as “SINE-enriched genes”); (2) a set of 1,480 genes enriched in L1 in regulatory regions (designated as “L1-enriched genes”); (3) a set of 2,439 genes enriched in low-complexity sequences in the promoters, 5′ UTR, and CDS regions and enriched in simple repeats in the 5′ UTR and CDS (designated as “low-complexity repeat-enriched genes”); and (4) a set of 383 genes enriched in satellite repeats in the CDS and 3′ UTR regions (designated as “satellite repeat-enriched genes”) (Table S1).

Interestingly, Gene Ontology (GO) analysis revealed that distinct functional terms are enriched in these four clusters of genes (Figures 1F and S3B). SINE-enriched genes are significantly enriched in RNA-related “housekeeping” functions, including ribosome, translation, nucleolus, and RNA binding and processing. By contrast, L1-enriched genes are strongly enriched in specialized functions, including olfactory, vomeronasal, and pheromone receptor activities, immunoglobulin function, and retinol metabolism, which tend to be expressed in terminally differentiated cells. In comparison, low-complexity repeat-enriched genes are highly enriched in transcription regulation and developmental processes. Developmental and tissue-specific transcription factors are known to harbor CpG islands or GC-rich low-complexity sequences in their promoters (Mendenhall et al., 2010). Notably,

satellite repeat-enriched genes mainly encode KRAB-containing and zinc finger transcription factors, which have been implicated in suppressing newly emerged retrotransposons (Jacobs et al., 2014; Thomas and Schneider, 2011). The position of satellite repeats in these genes largely overlaps with DNA sequences that encode the zinc-finger domain (Figures S3C and S3D), implying that endogenous satellite repeats might have been evolved to defense exogenous repetitive elements like retrotransposon.

Repetitive Elements Are Targeted by Different Classes of Regulators

The strong correlations between repeat content and gene function prompted us to ask whether and how repetitive elements coordinate the expression of genes with similar functions and repeat contents. A previous study of a handful of transcription factors has suggested that transposable elements carry DNA binding sites (Sundaram et al., 2014). To achieve a comprehensive view of repeat-mediated transcription and chromatin control, we collected a large set of 1,000 published ChIP-seq (chromatin immunoprecipitation followed by sequencing) datasets, covering 218 factors in 85 human cell lines and tissues (Figure 2A; Table S2). Our analysis showed that genomic repeats contribute extensively to the DNA binding sites for many regulatory proteins.

Many transcription and chromatin regulators exhibit specific binding activities toward particular subclasses of repeats, although a subset of general factors, such as CTCF, RAD21, MYC, and KDM5A, bind to multiple repeat subfamilies. In particular, Alu repeats show highly enriched occupancy by proteins involved in active transcription, including the RNA polymerase (Pol) II and III subunits POLR2A and POLR3A and the general transcription factors GTF3C2 and CEBPB (Figures 2A and 2B). In contrast, L1 repeats associate with chromatin regulators involved in gene silencing, heterochromatin formation, and maintenance. This list includes KAP1 (TRIM28); SETDB1, a histone methyltransferase that catalyzes histone H3 lysine 9 trimethylation (H3K9me3); the heterochromatin protein HP1α, which recognizes H3K9me3 and mediates silencing; and ZNF84, a member of the family of KRAB zinc-finger transcriptional repressors. In comparison, ERV repeats are specifically targeted by the pluripotency regulators POU5F1 and NANOG, and low-complexity and simple repeats exhibit enriched binding of the polycomb repressive complex 2 (PRC2) components EZH2 and SUZ12, which are consistent with previous reports (Kunarso et al., 2010; Mendenhall et al., 2010). Notably, significant proportions of the ChIP targets of individual proteins fall into repetitive elements. For example, 48% of the DNA binding sites of SETDB1 (in HEK293) and 47% of KAP1 and 23% of HP1a sites (in K562 cells) overlap with L1 repeats, and 65% of the GTF3C2 sites (in K562 cells) and ~39% of the POLR2A sites (in GM12891) overlap with Alu repeats. Among the above repeat-associated ChIP target sites, large portions reside in genic regions, for example, 42%–49% for L1 and >50% for Alu (including 9%–19% in gene promoters and 36%–48% in the introns).

In mouse ESCs, SINE-enriched, L1-enriched, and low-complexity repeat-enriched genes also exhibit specific enrichment for the binding of total Pol II, KAP1, and EZH2, respectively (Figures 2C and S4A); vice versa, KAP1-targeted genes show significantly higher levels of L1 repeats than genes targeted by Pol II and EZH2, which contain more

SINE B1 and low-complexity repeats, respectively (Figure S4B). These observations indicate that different subfamilies of repeats harbor extensive yet distinct *cis*-regulatory DNA elements that can be targeted by different sets of transcription and chromatin regulators with either active or repressive roles on chromatin, which further suggests a potential role of repeat DNA in coordinating the transcription of genes that harbor similar repeat features.

Orchestrated Activation of Repeats and Repeat-Associated Genes in Early Embryogenesis

During early embryogenesis, extensive transcriptional and epigenetic reprogramming occurs after fertilization, providing an ideal developmental paradigm to study repeat-associated gene activation and epigenetic changes (Zhang et al., 2016). We first analyzed chromatin accessibility in embryonic cells from early two-cell embryos, through the pluripotent inner cell mass (ICM) and ESCs that are derived from the ICM of blastocysts, to lineage-committed progenitor cells, on the basis of published ATAC-seq (assay for transposase-accessible chromatin followed by sequencing) and DNA hypersensitivity profiles (ENCODE Project Consortium, 2012; Wu et al., 2016). Chromatin regions that harbor SINEs are most accessible at the late two-cell stage and least accessible in lineage-committed cells; however, in lineage-committed cells, SINEs remain notably more open than the L1 and ERV subfamilies (Figure 3A). Following the initial activation of SINE repeats, chromatin regions with embedded low-complexity sequences become accessible and exhibit the highest level of opening in lineage-committed cells. In contrast, L1-containing genomic sequences remain inaccessible in these embryonic cells.

Next, we performed expression analysis of early embryonic cells (Wu et al., 2016; Yin et al., 2015). Consistent with the sequential chromatin opening of SINEs and low-complexity repeats, their associated genes are sequentially activated during early embryonic development (Figure 3B). For example, a SINE-enriched ribosomal gene *RPL39* and a low-complexity repeat-enriched transcription factor gene *SOX9* show coordinated chromatin opening and transcription activation at different embryonic stages (Figure 3C). We defined a cluster of 2,358 genes that are first activated during ZGA at the two-cell stage and show peak expression at the eight-cell stage, as ZGA genes. These genes are highly enriched in SINE-related functional terms, including RNA processing and ribosomal biogenesis. In comparison, a cluster of 1,184 genes that are enriched in developmental regulators and transcription factors (designated as developmental genes) are not activated until later in lineage-committed cells, including mesoderm and neural progenitor cells. Notably, the set of ZGA genes exhibit a significantly higher SINE content than the set of developmental genes, which contain more low-complexity repeats (Figure 3D). The ZGA and developmental gene sets are both depleted of L1 repeats (Figure S5A).

L1-Enriched Genes Are Sequestered in Nucleolus- and Lamina-Associated Inactive Domains

In ESCs, genes associated with different subfamilies of repeats are also expressed at different levels, congruent with coordinated chromatin opening and gene expression in early embryogenesis (Figure S5B). As the nuclear localization of a gene often influences its expression (Dekker et al., 2017), we then sought to visualize the nuclear location of L1 and

SINE B1 repeats, the two predominant subfamilies of retrotransposons in most mammals (Mandal and Kazazian, 2008). We performed DNA fluorescence *in situ* hybridization (FISH) using fluorescence-tagged oligonucleotide probes that specifically target the consensus sequences of B1 and L1 elements in ESCs. Interestingly, L1 repeats show highly organized, strong signals surrounding the nucleolus and in the nuclear periphery (Figure 4A). In contrast, B1 DNA signals are present mainly in the nuclear interior and show no overlap with L1 DNA signals. The opposing yet complementary localizations of B1 and L1 DNA in the nuclear space mirror the spatial organization of active and repressive nuclear compartments (Buchwalter et al., 2019) and support distinct expression of SINE- and L1-enriched genes that are either highly expressed or mostly silenced in ESCs, respectively (Figure S5B).

The nucleolus is the site in which rRNA synthesis and ribosomal biogenesis occur (O'Sullivan et al., 2013). Except for rDNA, the DNA content in the nucleolus is largely unknown in ESCs. We isolated the nucleoli from ESCs and performed DNA sequencing (Figures 4B and 4C). As expected, strong nucleolar DNA signals were detected at the 45S rDNA locus but not in ESCs treated with transcription inhibitors (Figures 4B and S6A), which effectively disrupted the nucleolar structure and served as the negative controls for nucleolar isolation. We identified a total of 424 nucleolus-associated domains (NADs) with sizes in the range of 0.1–5 Mb, which together constitute ~7.5% of the mouse genome, in ESCs. Notably, 60% of chromosome 19 (chr19), 34% of chr18, and 29% of chr12 were detected in the NADs (Figure S6B). Consistently, these three chromosomes were previously reported to assemble around the nucleolus, on the basis of genome-wide mapping of higher order chromosomal interaction hubs (Quinodoz et al., 2018).

We also analyzed the published sequencing data of DNA regions that are associated with the nuclear lamina, which lies underneath the inner nuclear membrane, representing a major structural element for genome organization inside the nucleus (Peric-Hupkes et al., 2010). ESCs contain ~1,180 lamina-associated domains (LADs), ranging in size from 40 kb to 15 Mb and covering ~40% of the genome. Both LADs and NADs contain significantly higher levels of L1 (1.6- to 2.3-fold, $p < 3.8E-09$) compared with random genomic regions and are depleted of SINEs (Figure 4D). This result corroborates the peri-nuclear and nucleolar staining pattern of L1 as revealed by DNA FISH. In addition, 59% of L1-enriched genes (875 of 1,480) were detected in NADs and/or LADs, in sharp contrast to only 2.5% of SINE-enriched genes (52 of 2,041) (Figure 4E).

About 42% of DNA sequences in NADs overlap with those in LADs. In total, 4,686 non-redundant mRNA genes were detected in NADs and/or LADs (defined as "NAD/LAD-associated genes"), and 528 genes were present in both domains (Figure 4F; Table S3). This result is consistent with previously reported observations that some NADs are located near the nuclear lamina and that nucleolus- and lamina-associated loci switch positions after mitosis (Kind et al., 2013; van Steensel and Belmont, 2017). Similar to L1-enriched genes, NAD- and LAD-associated genes are also enriched in highly specialized functions (Figures 4F and S6C). Interestingly, these specialized genes with similar functions (e.g., olfactory genes) are often in genomic juxtaposition to form large gene clusters that extensively overlap with L1 but not B1 repeats; examples are shown in Figure 4G and Figure S6D. Furthermore,

compared with the genome background, the NADs and LADs are significantly depleted (4- to 60-fold, p < E-10) of Pol II and active histone modifications (Figure 4H) but are enriched in silencing H3K9me2 and H4K20me3 marks that are associated with heterochromatin (Kidder et al., 2017). Accordingly, NAD- and LAD-associated genes are transcriptionally repressed in ESCs (Figure S6E). Together, imaging and sequencing-based analyses demonstrate that L1-enriched genes are silenced and sequestered in the inactive NADs and LADs in ESCs.

L1 RNA Binds to L1 DNA Sequences

The strikingly high genomic content and distinct localization of L1 repeats in heterochromatin domains led us to further explore a functional link of L1 RNA in mediating L1 DNA function in gene regulation. To reveal genome-wide DNA targets of L1 RNA, we performed chromatin isolation by RNA purification followed sequencing (ChIRP-seq) (Figure 5A). L1 ChIRP specifically captured L1 subfamilies of repeats at both the RNA and DNA levels, including *L1Md-A*, *L1-Gf*, and *L1-Tf*, but not the lncRNAs *Malat1* and *Neat1* (Figures S6F and S6G), confirming the efficiency and specificity of L1 pull-down.

L1 ChIRP-seq revealed 24,666 L1-binding sites, of which 92.0% fall into L1 DNA loci and 68.3% are present in NADs or/ and LADs (Figures S6H and S6I). L1-binding sites are positioned closer to L1-enriched genes, with a median distance of ~7 kb, but far away from SINE-enriched genes (378 kb median distance) and low-complexity repeat-enriched genes (166 kb median distance) (Figure 5B). We defined the set of ~2,397 genes that are located within 2 kb of L1 RNA-binding sites as “L1-ChIRP genes” (Table S4). L1-ChIRP genes show significant overlap (observed/expected = 2.6–3.5, p < 1.1E-172; Figure 5C) with the sets of NAD/LAD-associated genes (Figure 4F) and L1-enriched genes (Figure 1E). The strong binding preference of L1 transcripts to L1 DNA sequences suggests a role of L1 RNA to facilitate the function of its own DNA in regulating gene expression. Intriguingly, L1 transcripts are extremely unstable, with a half-life of approximately 40 min in ESCs (Figure 5D), implying that L1 RNA may act at the sites where it is transcribed.

Depletion of L1 RNA Alters Nuclear Localizations of L1-Rich DNA Sequences

Next, we sought to ask whether L1 RNA regulates the nuclear localization of its targets. We chose a 1.4 Mb L1-rich genomic region (chr17: 54.3–55.7 Mb) that show high levels of L1 ChIRP-seq signals for direct visualization by Oligopaint DNA FISH (Figure 5E). To mark the nucleolus, we co-stained ESCs with an antibody against NCL. Indeed, most cells show two strong signal spots of this L1-rich DNA segment at the nuclear and nucleolar peripheries (Figure 5F). We then depleted L1 transcripts by using the same morpholino antisense oligonucleotides (ASO) as we reported previously (Percharde et al., 2018). In ESCs treated with L1 ASO, this L1-rich region appears to be relocated from the nuclear membrane and the nucleolus into the nuclear interior (Figure 5F). Significantly fewer L1 ASO-treated ESCs showed proper LAD/NAD localizations (43% versus 80% of cells), compared with control ESCs transfected with a control ASO that is reverse complementary to the L1 ASO (Figure 5G). Thus, depletion of L1 RNA led to detachment of a representative L1-rich DNA segment from LADs and NADs into the nuclear interior, indicating an active role of L1 RNA in sequestering its target sequences into inactive domains.

L1 RNA Represses L1-Associated Genes with Interacting Proteins

L1-associated genes are transcriptionally silenced in ESCs (Figures S5B and S6E). Our analysis of the published RNA sequencing (RNA-seq) data from ESCs treated with L1 ASO (Percharde et al., 2018) revealed de-repression of ~1,414 L1-associated genes compared with control ESCs (Figures 6A, 6B, and S6J). For example, upon L1 ASO, the immunoglobulin gene *IGHV3–3* and the olfactory gene *OLFR376*, two L1-ChIRP targets, were upregulated 30- and 8.5-fold ($p < 0.01$), respectively (Figure 6C). These are still low levels of expression of these genes relative to the differentiated cells in which they function, perhaps due to the lack of specific transcriptional activators; however, their coordinate de-repression in L1 knockdown ESCs is notable. Most of ~1,414 upregulated L1-associated genes are the DNA targets of L1 RNA and exhibit higher L1-repeat contents (Figures 6B and S6K). In contrast, L1 depletion led to globally decreased expression of pluripotency (Figures 6A and 6B) and ribosomal genes in ESCs (Percharde et al., 2018). Considering predominant binding of L1 RNA at L1-associated genes but not at pluripotency and ribosomal genes (Figures 5A–5C and data not shown), we conjecture that the pluripotency-related defects likely result from indirect consequences of decreased rRNA transcription due to depletion of L1 RNA.

L1 DNAs exhibit enriched binding of KAP1, SETDB1, and HP1α proteins (Figures 2 and S4). It has been reported that L1 RNA binds NCL and promotes chromatin associations of NCL and KAP1 at *DUX* and rDNA genes (Percharde et al., 2018). In addition, NCL and KAP1 interact with each other, and both were also detected to be present in the protein interactome of L1 RNA (Peddigari et al., 2013; Percharde et al., 2018). To explore a role of NCL in regulating L1-associated genes, we performed low-input *in situ* ChIP of NCL in ESCs. NCL binds strongly to rDNA, which was attenuated upon depletion of L1 RNA (Figure S7A). Interestingly, NCL preferentially binds to L1 DNA sequences compared with other subclasses of repeats such as B1 and satellite repeats. Depletion of L1 RNA significantly reduced the binding of NCL to L1 DNA (Figure 6D). This result is consistent with the specific enrichment of L1 DNA in NADs and LADs and suggests that L1 RNA facilitates the binding of its interacting protein to targeted chromatin.

Next, we asked whether L1 repeat-associated proteins play a repressive role similar to L1 RNA. We first knocked down *NCL* by two short hairpin RNAs (shRNAs) (Figures S7B and S7C). Interestingly, the expression level of L1-associated genes was significantly upregulated in *NCL*-depleted ESCs (Figures 6E and S7D). Collectively, these data indicate that L1 RNA and NCL cooperatively regulate transcription silencing of L1-associated genes. We then analyzed published RNA-seq datasets in ESCs that were depleted of *SETDB1* or lacked *KAP1* or genes encoding HP1 protein isoforms (Deniz et al., 2018; Ostapcuk et al., 2018; Rowe et al., 2013). Knockout of *KAP1*, knockdown of SETDB1, and triple knockouts of HP1α, β, and γ all led to global de-repression of L1-associated genes (Figures S7E–S7G). Knockout of *HP1α* alone failed to have an effect (Figure S7G), which is consistent with the reported redundant role of HP1 subfamily members in gene silencing and heterochromatin formation (Ostapcuk et al., 2018; Yi et al., 2018). Together, these results illustrate that L1 RNA represses the expression of L1-associated genes, likely acting together with its interacting proteins at RNA and/or DNA levels.

DISCUSSION

The effect of genome colonization by repeats on mammalian gene organization and regulation remains a matter of speculation and controversy. Considering the widespread and diverse nature of repeats, treating them as an aggregate class without defining their subtypes and content would underestimate their potential regulatory differences and functions. In this study, our quantitative and systematic survey of repeat compositions for each gene across the genome reveals extensive associations of repeats with the function, regulation, and expression of their host genes. In particular, SINE, L1, and low-complexity repeats demarcate their associated genes into three major functional categories that are differentially expressed in distinct developmental stages, probably by recruiting distinct sets of regulators to their genomic sequences and/or sequestering their associated genes into distinct nuclear domains (Figure 7). The combinatory effects of protein targeting and nuclear sequestering coordinate the genome-wide expression of genes harboring similar repeats and also dictate different levels of gene activity across different repeat subclasses. Genes enriched in SINEs are more likely to encode housekeeping proteins related to RNA processing, ribosomal biogenesis, and nucleolar function; these genes show enrichment of binding sites for factors involved in active transcription, and they are first activated during ZGA and are highly expressed in ESCs. L1-enriched genes tend to produce proteins with specialized functions in terminally differentiated cells; they are preferentially targeted by heterochromatin proteins and epigenetic repressors and are sequestered in repressive nuclear domains for gene silencing in ESCs. Low-complexity repeat-enriched genes tend to encode developmental and tissue-specific transcription factors that are preferentially targeted by PRC2 for transcriptional poise. Dynamic and orchestrated chromatin opening of SINE- and low-complexity repeat-associated sequences, and sequential activation of their associated genes in early embryonic development, further support the role of genic repeats in wiring transcription regulatory networks to achieve stage-specific activation or silencing of genes with distinct functions.

Despite drastic differences of L1 and Alu/B1 distributions in the genome, they both depend on the reverse transcriptase encoded by L1 ORFs for retrotransposition and share a common AT-rich insertion site specificity during integration (Gilbert et al., 2002; Jurka et al., 2004; Wagstaff et al., 2012). Two recent studies of L1 retrotransposon insertions in cultured cells revealed that the landscape of endogenous L1 elements differs significantly from that of new insertions, which appear to broadly target all regions of the human genome, being insensitive to chromatin state (Flasch et al., 2019; Sultana et al., 2019). It has been proposed that purifying selection, rather than biased insertions, reshapes the genomic distributions of L1 and Alu/B1 after their integration (Graham and Boissinot, 2006; Pavlícek et al., 2001; Sultana et al., 2019). We speculate that the specific association of gene and repeat families is so important that during evolution it has imposed selective pressures on different classes of repeats to accumulate in specific sets of genes, depicting the co-adaptive trajectories of transposable elements with their host. Despite detectable expression of L1 ORF1 protein in ESCs, treatments with antiretroviral drugs that inhibit L1 retrotransposition did not phenocopy the effects of L1 RNA depletion (Percharde et al., 2018). Inhibition of L1 retrotransposition also failed to rescue the phenotypes of two-cell arrest and aberrant

chromatin accessibility due to prolonged activation of L1 (Jachowicz et al., 2017). These reports suggest that the functional role of L1 in gene and chromatin regulation is likely independent of its retrotransposition activity.

Imaging and sequencing analyses of ESCs illustrate a marked effect of spatial nuclear segregation of L1 and B1 repeats in sequestering their associated genes in distinct nuclear compartments. The predominant localization of L1 repeats and L1-associated genes in NADs and LADs that are depleted of B1 likely contributes in large part to the genome-wide silencing of L1-associated genes. This notion is supported by the results that depletion of L1 RNA in ESCs led to detachment of L1 repeat DNA from NADs and LADs and global upregulation of L1-associated genes (Figures 5 and 6). In terminally differentiated pro-B cells or sensory neurons, the activation of L1-associated immunoglobulin or olfactory genes, respectively, is accompanied by their relocation from the nuclear periphery to the nuclear interior (Kosak et al., 2002; Rother et al., 2016; Yoon et al., 2015). Reductions in levels of HP1 family proteins, H3K9me3 and heterochromatin contents, recurrent L1 retrotransposition, and abnormal immunoglobulin expression in non-lymphoid neoplastic cells have been reported to correlate with human cancers (Chen et al., 2009; Ehrlich, 2009; Gurrion et al., 2017; Iskow et al., 2010; Kirschmann et al., 2000; Lee et al., 2012; Qiu et al., 2003; Ranzani et al., 2017; Shukla et al., 2013; Solyom et al., 2012; Tubio et al., 2014). These findings indicate that dynamic regulation of the nuclear positioning of L1-rich DNA and associated genes by epigenetic and transcriptional mechanisms is critical for proper gene expression and cellular function. Given the essential role of L1 RNA in regulating the nuclear localization and repression of L1-associated genes and considering the short half-life and extensive binding of L1 RNA to its DNA sequences, we propose that L1 transcripts may act in chromatin neighborhoods of their transcription sites to anchor L1-rich genomic segments to the nuclear and nucleolar peripheries, in part through L1-interacting protein partners. In such a way, L1 RNA transacts the repressive and structural information encoded in L1 DNA repeats, contributing to the macroscopic structure and regulation of host genome (Lu et al., 2019).

In summary, individual repeat subclasses endow the genome with hundreds of thousands of similar sequences, which may provide an efficient and powerful way to coordinate diverse genomic sequences into one regulatory network. Analogous to the complex mixtures of microorganisms that are believed to have co-evolved with their human hosts (Dethlefsen et al., 2007), transposable elements, another stably embedded “parasite,” have greatly defined, shaped, and influenced their host genomes.

STAR★METHODS

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Dr. Xiaohua Shen (xshen@tsinghua.edu.cn)

All unique/stable reagents generated in this study are available from the Lead Contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

ESCs (CJ9 and 46C) were maintained in DMEM (Dulbecco's modified Eagle's medium) supplemented with 15% heat-inactivated FCS (fetal calf serum), 2 mM Glutamax (100 × Life Technology), 1% nucleoside mix (100 × stock, Millipore), 1 × Penicillin-Streptomycin Solution (100 × stock, Life Technologies), 0.1 mM non-essential amino acids, 0.1 mM 2-mercaptoethanol and supplied with 1000 U/ml recombinant leukemia inhibitory factor (LIF, Millipore). HEK293T cells and MEF cells were cultured in DMEM medium containing 10% FCS and 1 × Penicillin-Streptomycin Solution.

METHOD DETAILS

Repetitive elements catalog—The reference catalog of repetitive elements was built from RepeatMasker annotations (Tempel, 2012). We removed repetitive elements marked by ""? at the end of "Family" or "Class" name (for example, Alu?) which represent unclear classification. We also subtracted non-coding RNAs and other repetitive element classes with low copy numbers. Finally, 96.1% of whole annotated repetitive element were retained for the following analysis.

Hierarchical clustering—The raw repetitive element matrix containing repeat percentage in genic regions for each gene was calculated according to the flow chart in Figure S2A. Briefly, we divided a protein-coding gene into six regions (Promoter, Intron, Downstream, 5′UTR, CDS, 3′UTR) and intersected them with the catalog of processed repetitive element annotations (14 different repeat subclasses). We calculated the precise repeat compositions by normalizing the raw repeat content the with genomic length of each genic region. The repetitive element content of each genic region in every protein-coding gene can be quantitatively represented by a combination of 84 (14 × 6) dimensional vectors. In total, we collected the 84-dimensional feature vectors for the 40,005 annotated protein-coding transcripts (22,432 genes) in mouse. The raw repetitive element matrix (40,005 × 84) was normalized with "normalize.quantiles" function in "preprocessCore" package in R software (Bolstad et al., 2003). Hierarchical clustering was performed using the 'hclust' function, and 'average' method in R software. Pearson correlation coefficients (c.c) were calculated for any two transcript pair and (1-P.C.C.) was used as distance matrix for clustering. Corresponding heatmaps were plotted in R software. Cluster number: we used stepwise strategy to decide the final cluster number. The clustering number kept increasing as the height used for cutting the clustering tree decreased at the very beginning and then the clustering number became relative stable when the height decreased to 0.9. Therefore, we chose 0.9 as the height to cut the clustering tree and 77 gene clusters were generated. In Figures 1D and S1D, gene ontology (GO) analysis of 77 gene clusters generated by hierarchical clustering of genic repeat features (in red) was compared to that of random gene sets (in blue). For each cluster, GO analysis was performed and the p values of three most significant terms were kept and the corresponding cumulative distribution curve (CDC) was plotted (red line). Similar analyses were performed for 77 random gene sets (blue). A significant difference between the two CDCs was observed (Wilcoxon p = 2.7E-71), indicating that genes with similar repeat patterns tend to share similar biological function. Should we choose a *p-value* of 1E-3 as the cutoff (indicated by the dotted vertical line), ~70% of the gene clusters generated by their genic repeat patterns are enriched in specific

functional terms, in sharp contrast to < 10% of the random gene sets. This observation indicates that genes with similar repeat patterns tend to share similar biological function.

GO analysis—GO analysis, except in Figure 1D and Figure S1C, were performed using DAVID bioinformatics tools (Huang et al., 2009). P value of Fisher's exact test was used to evaluate the enrichment of certain GO term. For Figure 1D and Figure S1C, we used Python-based library to process over- and under-representation of certain GO terms, based on Fisher's exact test (Klopfenstein et al., 2018). The goatools version number is v0.5.9. The python version is "Python 2.7.9:: Anaconda 2.2.0 (64-bit)."

ChIP-seq analysis—Raw reads were mapped to the mouse genome (mm9) using Bowtie2 (version 2.2.2) (Langmead and Salzberg, 2012) with the parameter –gbar 200 allowing no gaps. Multiple alignments were allowed and we only reported the best alignments based MAPping Quality (MAPQ) values. Positive peaks were identified with the Model-based Analysis for ChIP-Seq (MACS) program (Zhang et al., 2008) by compared to the input samples. ChIP-seq signal analysis around gene-body regions was performed with ngs.plot with the Version: 2.61 (Shen et al., 2014) with the parameter –RB 0.01, indicating that 1% of extreme values will be trimmed on both ends.

RNA-Seq and Gene Set Enrichment Analysis—Alignments of RNA-Seq data to mouse genome assembly mm10 were performed using Tophat v2.0.10 (Kim et al., 2013). Fragments Per Kilobase of exon model per Million mapped reads (FPKM) were calculated by Cufflink 2.1.1 to represent expression levels of transcripts (Trapnell et al., 2012). Gencode v19 and vM9 were used as human and mouse gene annotation, respectively (Derrien et al., 2012). Gene set enrichment analysis (GSEA) was performed as described previously (Luo et al., 2016). The sets of genes highly expressed in ESCs (95 genes) was selected as previously described (Luo et al., 2016).

Cell culture—ESCs were cultured on gelatin-coated plates in standard ESC medium consisting of DMEM (Cellgro) supplemented with 15% heat-inactivated fetal bovine serum (Hyclone), 1% Glutamax (GIBCO), 1% Penicillin/Streptomycin (Cellgro), 1% nucleoside (Millipore), 0.1mM 2-mercaptoethanol (GIBCO), 1% MEM nonessential amino acids (Cellgro), and 1000U/ml recombinant leukemia inhibitory factor (Millipore).

shRNA knock down—The shRNA expressing lentivirus were packed in 293T cells. At 24 hours after infection, ESCs were treated with puromycin and then harvested for RNA-seq analysis at the 72-hour time point.

DNA FISH—DNA FISH in ESCs was performed as previously described (Wang et al., 2016). Fluorescence-tagged oligonucleotide probes were used to specifically target the consensus sequences of L1 and B1. First, ESCs cultured on 35 mm glass-bottom dishes was fixed with 4% formaldehyde (PFA) diluted in phosphate-buffered saline (PBS) for 10 minutes at room temperature (RT), and then washed with PBS for 2 minutes, followed by permeabilizing with 0.5% Triton X-100 in PBS. After washing three times with PBS, mESCs were treated with 0.1M HCl for 5 minutes and then incubated with 0.1 mg/mL RNase A diluted in PBS for 45 minutes at 37°C. Before prehybridization, cells were washed

three times in 2x saline-sodium citrate (SSC) buffer. Prehybridization was conducted by incubating cells in 50% formamide diluted in 2x SSCT (2x SSC + 0.1% Tween-20) for 5 minutes at room temperature, then for 20 minutes at 47°C. Samples were denatured for 2.5 minutes at 78°C on the top of a water-immersed heat block. For hybridization, 200 μL hybridization buffer containing 2x SSC, 50% formamide, 20% dextran sulfate, and 0.5 μM of each probes of L1 and B1 could cover the glass bottom and then samples were incubated in a humidified chamber at 37°C for about 16 hours. Finally, ESCs were washed with 2X SSCT for 15 minutes at 50°C twice, and then washed twice with 2X SSCT for 1 hour at room temperature. The labeled mESCs were rinsed briefly in 2xSSC and mounted with ProLong® Diamond Antifade Mountant with DAPI (P36962, Thermo Fisher Scientific).

DNA-FISH Probes—L1-mouse-1 AGGACACATGCTCCACTATGTTCATAGCAG

L1-mouse-2 AGATGCCCATGAACATACAAGAAGCCTACAGAACT

B1-mouse-1 gcctggtctacagagtgagttccaggacag

B1-mouse-2 cagcacttgggaggcagaggcaggcggatt

Oligopaint FISH—A set of 4,500 DNA probes at a density of ~300 bp per probe were designed using standard procedures to target the genomic regions defined in Table S5. For the primary probe pool, we purchased the Oligoarray pool (Synbio Technologies), and prepared FISH probes via limited cycle PCR, *in vitro* transcription, and reverse transcription as described previously (Wang et al., 2016). The primary probes were freshly mixed with corresponding secondary probes in hybridization buffer at 1 μM final concentration for each secondary probe. The hybridization buffer with the primary and secondary probes was heated to 86°C for 3 minutes, and then placed on ice immediately. Cell samples were prepared for DNA FISH as described above for ESCs, except that the cells were incubated in hybridization buffer at 86°C for 3 minutes before hybridization overnight.

Transmission Electron Microscopy—ESCs were fixed with 2.5% glutaraldehyde, dehydrated in an ethanol series, transferred to methanol, and immersed into a freshly prepared mixture of methanol and acetic anhydride (5:1, v/v) at 25°C for 24 hours in the dark (Tandler and Solari, 1982; Testillano et al., 1991). Cells were then washed in pure methanol for 20 minutes, transferred in ethanol and embedded in Epon (Sigma). Ultrathin (50 nm) sections were contrasted with 5% aqueous uranyl acetate for 60 minutes at room temperature and examined with a transmission electron microscope (HITACHI, H-7650-B).

Nucleolar DNA-seq—ESCs (4×15 cm dishes) were treated with DMSO or with transcription inhibitors Actinomycin D (ActD, 1 μg/ml) or 5,6-dichloro-1-beta-D-ribofuranosylbenzimidazole riboside (DRB, 100 μM) for 2 hours before nucleolar isolation. Cells were treated by 10 mL of Buffer A (10 mM HEPES, pH 7.9, 10 mM KCl, 1.5 mM MgCl2, 0.5 mM DTT) on ice for 5 minutes, followed by Dounce homogenization for 10 strokes. Centrifuge cells at 218 g for 5 minutes at 4°C. Then resuspend the pellet with 3 mL S1 solution (0.25 M sucrose, 10 mM MgCl2) and layer over 3 mL S2 solution (0.35 M sucrose, 0.5 mM MgCl2) and centrifuge at 1,430 g for 5 minutes at 4°C. Resuspend the

pellet with 3 mL of S2 solution and sonicate briefly to break the nuclei. Layer the sonicated sample over 3 mL of S3 solution (0.88 M Sucrose, 0.5 mM MgCl2) and centrifuge at 3,000 g for 10 minutes at 4°C to pellet down the nucleoli. Resuspend the nucleolar pellet with 500 mL of S2 solution and centrifuge for 5 minutes. The nucleolar pellet was subjected to DNA purification by phenol-chloroform extraction. Nucleolar DNA libraries were constructed by following Illumina library preparation protocols. Raw reads were mapped to the mouse genome (mm9) using Bowtie2 (version 2.2.2) (Langmead and Salzberg, 2012) with the parameter –gbar 200 allowing no gaps. Multiple alignments were allowed and we only reported the best alignments based MAPping Quality (MAPQ) values. Positive peaks were identified with the MACS program (Zhang et al., 2008) by compared the sample treated with DMSO to sample treated with ActD or DRB.

Chromatin Isolation by RNA Purification (ChIRP)—ChIRP was performed as previously described with some modifications (Chu et al., 2011; Percharde et al., 2018; Yin et al., 2015). A set of 32 antisense oligos with 59-nt long probes tiled along the L1 consensus sequence (6.5 kb) were biotinylated through terminal transferase (NEB) with Bio-N6-ddATP (ENZO) as substrate. ESCs were harvested by trypsin digestion and first crosslinked with 2 mM dithiobis succinimidyl propionate (DSP, Thermo Scientific) in PBS for 30 minutes with gentle end-to-end rotation at room temperature. Then, a final concentration of 3.7% formaldehyde was added for 10 minutes, then quenched with 250 mM Glycine for 5 minutes at room temperature. Crosslinked ESCs were centrifuged and washed with ice-cold PBS for 3 times, then snapped frozen in liquid nitrogen and stored at −80°C.

Crosslinked cells (1×10^7) were resuspended with 500 mL DNase I digestion solution (20 mM Tris-HCl, pH7.5, 5 mM MgCl2, 0.5 mM CaCl2, 0.5% Triton X-100) with 1/20 volume of vanadyl ribonucleoside complex (VRC, NEB), 2.5 mL protease inhibitors and 2.5 mL of 200 mM PMSF. DNase I was added to a final concentration of 12 U/ml; the reaction was rotated at 37°C for 10 minutes and stopped with 20mM EDTA. Chromatin was pelleted, washed once with nuclear lysis buffer (NLB, 50mM Tris-HCl, pH7.5, 10mM EDTA, 1% SDS, inhibitors) and sonicated in NLB (5 cycles of: 25% amplitude, 6 s on, 15 s off, Vibra-Cell Ultrasonic Liquid Processors). Insoluble material was removed by centrifugation and the supernatant used for ChIRP experiments. For the RNase treatment control, samples were treated with 10 mg/ml Rnase A/T1 for 20 minutes at 37C. For hybridization, samples were incubated with 20 pmol probes per 200ml lysate, supplemented with one-fourth volume of 5x hybridization buffer (50mM Tris-HCl, pH7.5, 10mM EDTA, 1.5M NaCl, 50% formamide). The hybridization was conducted at 39°C rotating for 3 hours. 50 mL prebalanced streptavidin M280 beads were then added and the incubation continued for additional 3 hours. The beads were washed 5 times with 0.2x SSC wash buffer (1% SDS) at 42°C. For the RNase treatment control, after 3 washes, the beads were treated once with 10 mg/ml RNase A/T1 at 37°C in RNase digestion buffer (50mM Tris-HCl, pH7.5, 75mM NaCl, 1mM DTT), before washing two more times. To elute, the beads were washed once with SDS elution buffer (50mM Tris-HCl, 5mM MgCl2, 75mM NaCl, 1% SDS) at 39°C for 20 minutes, and once with elution buffer (50mM Tris-HCl, 5mM MgCl2, 75mM NaCl, 0.1% Triton X-100) at 39°C for 5 minutes. DNA was eluted from the beads by RNase H treatment in two sequential incubations with RNase H (NEB) at 37°C for 20 minutes, and with SDS

elution buffer at room temperature, 2 minutes, combining all eluents. Crosslinking was reversed by treatment with 0.1 mg /ml protease K, 150mM NaCl, and 10mM EDTA at 65°C overnight and the DNA was purified using the MinElute PCR Purification Kit (QIAGEN). ChIRP enrichments were analyzed by qPCR of the purified DNA. For RNA-ChIRP analysis, beads were boiled in NLB after the 0.1x SSC washes, then further reverse-crosslinked by boiling at 95°C for 30 minutes in the presence of 1mM DTT. Reverse crosslinked RNA was purified using Trizol and processed for RT-qPCR analysis.

ChIRP-Seq DNA libraries were constructed by following Illumina library preparation protocols. Raw reads were mapped to the mouse genome (mm9) using Bowtie2 (version 2.2.2) (Langmead and Salzberg, 2012) with the parameters –gbar 200 allowing no gaps. Multiple alignments were allowed and we only reported the best alignments based MAPping Quality (MAPQ) values. Aligned files were further converted to bedgraph files with BEDTools (Quinlan, 2014). Positive peaks were identified with the MACS program (Zhang et al., 2008) by input sample with a *p-value* cutoff of 0.01 and false discovery rate (fdr) < 0.05.

Low-input in situ ChIP—This method utilizes Tn5 transposase-targeted chromatin release to reveal genomic distributions of a protein of interest in limited numbers of cells (Wang et al., 2019). The experiment was performed as described previously with modifications (Wang et al., 2019). Briefly, ESCs were fixed with 1% formaldehyde for 10 minutes, then quenched with 0.125 M glycine at room temperature for 5 minutes, and washed 3 times by cold PBS and resuspended in 200 μL hypotonic buffer (0.3% SDS, 20 mM HEPES pH 7.9, 10 mM KCl, 10% Glycerol, 0.2% NP40, protease inhibitor and 10 mM sodium butyrate). To open up the chromatin prior to the ChIP experiment, cells were first treated at 62°C for 10 minutes and then followed by addition of 20 mL of 20% Triton X-100 at 37°C for one hour in a ThermoMixer (Eppendorf) with gentle shaking at 600 rpm. After two washes with the wash buffer (20 mM HEPES pH 7.5, 150 mM NaCl, 10mM sodium butyrate, supplemented with 1X Protease inhibitor, 1X PMSF, and 0.5 mM spermidine), ~10,000 treated cells were incubated with the activated Con-A magnetic beads (Bangs Laboratories) at room temperature for 10 minutes. Cells were captured by a magnet stand and washed once with the Dig-wash buffer (wash buffer with 0.01% digitonin and 0.1% triton). The cell-beads mixture was resuspended in 100 μL of the antibody buffer (Dig-wash buffer supplemented with 2.5 mM EDTA) into an RNase-free PCR tube. About 0.5 mg antibody (NCl or IgG) was added then incubated at 4°C for 2~4 hours. Antibody-conjugated cells (coupled with Con-A beads) were washed twice with the Dig-wash buffer, and suspended in 100 μL Dig-wash buffer containing 1.2 mM of the PAT enzyme, a fusion protein of Protein A and Tn5 transposase. The assembly of PAT with barcoding primers was described as previously (Wang et al., 2019). After one-hour incubation at 4°C followed by three washes, the complex of PAT enzyme and antibody-conjugated cells (coupled with Con-A beads) was resuspended in 20 μL of reaction buffer (10 mM TAPS-NaOH pH 8.3, 5 mM MgCl2, 10% DMF and supplemented with protease inhibitor and 10 mM sodium butyrate). Onehour incubation at 37°C allows Tn5-targeted tagmentation to occur. To quench the reaction, 20 μL of 40 mM EDTA was added for 15 minutes. After one wash with 1% BSA in PBS, the cell-antibody-PAT complexes were resuspended with 5 μL lysis buffer (10 mM

Tris-HCl pH 7.5, 0.05% SDS and 0.2 mg/ml Proteinase K) and incubated at 65°C for 8 hours to reverse the crosslinking and then at 85°C for 15 minutes to deactivate proteinase K. Prior to PCR amplification, 1 μL 1.8% Triton X-100 was added to the lysates to quench SDS and incubated at 37°C for 1 hour. Released DNA fragments were amplified with Q5 enzyme (NEB) for 15 cycles of PCR reactions by the primers (5′-TCGTCGGCAGCGTCAGAT-3′ and 5′-GTCTCGTGGGCTCGGAGA-3′). DNA was purified using MinElute PCR Purification Kit (QIAGEN) and subjected to quantitative real-time PCR analysis.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analyses were carried out using Excel or R (version 3.4.3). All of the statistical details can be found in the relevant figure legends.

DATA AND CODE AVAILABILITY

All sequencing data are available through the Gene Expression Omnibus (GEO) via accession GEO: GSE103610 and GEO: GSE125766.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank the Shen Laboratory members for insightful discussion. We thank Aibin He for sharing low-input ChIP protocol. This work was supported in part by the National Basic Research Program of China (2018YFA0107604 and 2017YFA0504204 to X.S.), the National Natural Science Foundation of China (31630095 and 31925015 to X.S., 21825401 to Y.S.), the National Key R&D Program of China (2017YFA0505300 to Y.S.), the Center for Life Sciences (CLS) at Tsinghua University (to X.S), the National Institutes of Health (R01GM123556 to M.R.-S.), the Canada 150 Research Chair in Developmental Epigenetics and the University of Toronto Medicine by Design Program (to M.R.-S.), and the Medical Research Council and the UK Research and Innovation (UKRI) Future Leaders Fellowship (to M.P.).

REFERENCES

Biémont C (2010). A brief history of the status of transposable elements: from junk DNA to major players in evolution. Genetics 186, 1085–1093. [PubMed: 21156958]

Boeva V (2016). Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in eukaryotic cells. Front. Genet 7, 24. [PubMed: 26941778]

Bolstad BM, Irizarry RA, Astrand M, and Speed TP (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19, 185–193. [PubMed: 12538238]

Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew JL, Ruan Y, Wei CL, Ng HH, and Liu ET (2008). Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res. 18, 1752–1762. [PubMed: 18682548]

Buchwalter A, Kaneshiro JM, and Hetzer MW (2019). Coaching from the sidelines: the nuclear periphery in genome regulation. Nat. Rev. Genet 20, 39–50. [PubMed: 30356165]

Carrieri C, Cimatti L, Biagioli M, Beugnet A, Zucchelli S, Fedele S, Pesce E, Ferrer I, Collavin L, Santoro C, et al. (2012). Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. Nature 491, 454–457. [PubMed: 23064229]

Chen Z, and Zhang Y (2019). Loss of DUX causes minor defects in zygotic genome activation and is compatible with mouse development. Nat. Genet 51, 947–951. [PubMed: 31133747]

Chen Z, Qiu X, and Gu J (2009). Immunoglobulin expression in non-lymphoid lineage and neoplastic cells. Am. J. Pathol 174, 1139–1148. [PubMed: 19246641]

Chu C, Qu K, Zhong FL, Artandi SE, and Chang HY (2011). Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. Mol. Cell 44, 667–678. [PubMed: 21963238]

Chuong EB, Elde NC, and Feschotte C (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science 351, 1083–1087. [PubMed: 26941318]

de Koning AP, Gu W, Castoe TA, Batzer MA, and Pollock DD (2011). Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet. 7, e1002384. [PubMed: 22144907]

Dekker J, Belmont AS, Guttman M, Leshyk VO, Lis JT, Lomvardas S, Mirny LA, O'Shea CC, Park PJ, Ren B, et al.; 4D Nucleome Network (2017). The 4D Nucleome project. Nature 549, 219–226. [PubMed: 28905911]

Deniz Ö., de la Rica L, Cheng KCL, Spensberger D, and Branco MR (2018). SETDB1 prevents TET2-dependent activation of IAP retroelements in naïve embryonic stem cells. Genome Biol. 19, 6. [PubMed: 29351814]

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, 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. [PubMed: 22955988]

Dethlefsen L, McFall-Ngai M, and Relman DA (2007). An ecological and evolutionary perspective on human-microbe mutualism and disease. Nature 449, 811–818. [PubMed: 17943117]

Doolittle WF, and Sapienza C (1980). Selfish genes, the phenotype paradigm and genome evolution. Nature 284, 601–603. [PubMed: 6245369]

Durruthy-Durruthy J, Sebastiano V, Wossidlo M, Cepeda D, Cui J, Grow EJ, Davila J, Mall M, Wong WH, Wysocka J, et al. (2016). The primate-specific noncoding RNA HPAT5 regulates pluripotency during human preimplantation development and nuclear reprogramming. Nat. Genet 48, 44–52. [PubMed: 26595768]

Ehrlich M (2009). DNA hypomethylation in cancer cells. Epigenomics 1, 239–259. [PubMed: 20495664]

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74. [PubMed: 22955616]

Flasch DA, Macia A, Sanchez L, Ljungman M, Heras SR, Garcia-Perez JL, Wilson TE, and Moran JV (2019). Genome-wide de novo L1 retrotransposition connects endonuclease activity with replication. Cell 177, 837–851.e28. [PubMed: 30955886]

Gilbert N, Lutz-Prigge S, and Moran JV (2002). Genomic deletions created upon LINE-1 retrotransposition. Cell 110, 315–325. [PubMed: 12176319]

Graham T, and Boissinot S (2006). The genomic distribution of L1 elements: the role of insertion bias and natural selection. J. Biomed. Biotechnol 2006, 75327. [PubMed: 16877820]

Grow EJ, Flynn RA, Chavez SL, Bayless NL, Wossidlo M, Wesche DJ, Martin L, Ware CB, Blish CA, Chang HY, et al. (2015). Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. Nature 522, 221–225. [PubMed: 25896322]

Gurrion C, Uriostegui M, and Zurita M (2017). Heterochromatin reduction correlates with the increase of the KDM4B and KDM6A demethylases and the expression of pericentromeric DNA during the acquisition of a transformed phenotype. J. Cancer 8, 2866–2875. [PubMed: 28928876]

Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, Daly MJ, Price AL, Pritchard JK, Sharp AJ, and Erlich Y (2016). Abundant contribution of short tandem repeats to gene expression variation in humans. Nat. Genet 48, 22–29. [PubMed: 26642241]

Gymrek M, Willems T, Reich D, and Erlich Y (2017). Interpreting short tandem repeat variations in humans using mutational constraint. Nat. Genet 49, 1495–1501. [PubMed: 28892063]

Huang DW, Sherman BT, and Lempicki RA (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc 4, 44–57. [PubMed: 19131956]

Iskow RC, McCabe MT, Mills RE, Torene S, Pittard WS, Neuwald AF, Van Meir EG, Vertino PM, and Devine SE (2010). Natural mutagenesis of human genomes by endogenous retrotransposons. Cell 141, 1253–1261. [PubMed: 20603005]

Jachowicz JW, Bing X, Pontabry J, Bošković A, Rando OJ, and Torres-Padilla ME (2017). LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. Nat. Genet 49, 1502–1510. [PubMed: 28846101]

Jacobs FM, Greenberg D, Nguyen N, Haeussler M, Ewing AD, Katzman S, Paten B, Salama SR, and Haussler D (2014). An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. Nature 516, 242–245. [PubMed: 25274305]

Jurka J, Kohany O, Pavlicek A, Kapitonov VV, and Jurka MV (2004). Duplication, coclustering, and selection of human Alu retrotransposons. Proc. Natl. Acad. Sci. U S A 101, 1268–1272. [PubMed: 14736919]

Kelley D, and Rinn J (2012). Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol. 13, R107. [PubMed: 23181609]

Kidder BL, Hu G, Cui K, and Zhao K (2017). SMYD5 regulates H4K20me3-marked heterochromatin to safeguard ES cell self-renewal and prevent spurious differentiation. Epigenetics Chromatin 10, 8. [PubMed: 28250819]

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, and Salzberg SL (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14, R36. [PubMed: 23618408]

Kind J, Pagie L, Ortabozkoyun H, Boyle S, de Vries SS, Janssen H, Amendola M, Nolen LD, Bickmore WA, and van Steensel B (2013). Single-cell dynamics of genome-nuclear lamina interactions. Cell 153, 178–192. [PubMed: 23523135]

Kirschmann DA, Lininger RA, Gardner LM, Seftor EA, Odero VA, Ainsztein AM, Earnshaw WC, Wallrath LL, and Hendrix MJ (2000). Down-regulation of HP1Hsalpha expression is associated with the metastatic phenotype in breast cancer. Cancer Res. 60, 3359–3363. [PubMed: 10910038]

Klopfenstein DV, Zhang L, Pedersen BS, Ramírez F, Warwick Vesztrocy A, Naldi A, Mungall CJ, Yunes JM, Botvinnik O, Weigel M, et al. (2018). GOATOOLS: a Python library for Gene Ontology analyses. Sci. Rep 8, 10872. [PubMed: 30022098]

Kosak ST, Skok JA, Medina KL, Riblet R, Le Beau MM, Fisher AG, and Singh H (2002). Subnuclear compartmentalization of immunoglobulin loci during lymphocyte development. Science 296, 158–162. [PubMed: 11935030]

Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng HH, and Bourque G (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat. Genet 42, 631–634. [PubMed: 20526341]

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al.; International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature 409, 860–921. [PubMed: 11237011]

Langmead B, and Salzberg SL (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359. [PubMed: 22388286]

Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ 3rd, Lohr JG, Harris CC, Ding L, Wilson RK, et al.; Cancer Genome Atlas Research Network (2012). Landscape of somatic retrotransposition in human cancers. Science 337, 967–971. [PubMed: 22745252]

Lu JY, Chang L, Li T, Wang T, Yin Y, Zhan G, Zhang K, Percharde M, Wang L, Peng Q, et al. (2019). L1 and B1 repeats blueprint the spatial organization of chromatin. bioRxiv. 10.1101/802173.

Lunyak VV, Prefontaine GG, Núñez E, Cramer T, Ju BG, Ohgi KA, Hutt K, Roy R, García-Díaz A, Zhu X, et al. (2007). Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. Science 317, 248–251. [PubMed: 17626886]

Luo S, Lu JY, Liu L, Yin Y, Chen C, Han X, Wu B, Xu R, Liu W, Yan P, et al. (2016). Divergent lncRNAs regulate gene expression and lineage differentiation in pluripotent cells. Cell Stem Cell 18, 637–652. [PubMed: 26996597]

Lynch VJ, Leclerc RD, May G, and Wagner GP (2011). Transposonmediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. Nat. Genet 43, 1154–1159. [PubMed: 21946353]

Mandal PK, and Kazazian HH Jr. (2008). SnapShot: vertebrate transposons. Cell 135, 192–192.e1. [PubMed: 18854165]

Mendenhall EM, Koche RP, Truong T, Zhou VW, Issac B, Chi AS, Ku M, and Bernstein BE (2010). GC-rich sequence elements recruit PRC2 in mammalian ES cells. PLoS Genet. 6, e1001244. [PubMed: 21170310]

Muotri AR, Marchetto MCN, Coufal NG, Oefner R, Yeo G, Nakashima K, and Gage FH (2010). L1 retrotransposition in neurons is modulated by MeCP2. Nature 468, 443–446. [PubMed: 21085180]

O'Sullivan JM, Pai DA, Cridge AG, Engelke DR, and Ganley AR (2013). The nucleolus: a raft adrift in the nuclear sea or the keystone in nuclear structure? Biomol. Concepts 4, 277–286. [PubMed: 25436580]

Orgel LE, and Crick FH (1980). Selfish DNA: the ultimate parasite. Nature 284, 604–607. [PubMed: 7366731]

Ostapcuk V, Mohn F, Carl SH, Basters A, Hess D, Iesmantavicius V, Lampersberger L, Flemr M, Pandey A, Thomä NH, et al. (2018). Activity-dependent neuroprotective protein recruits HP1 and CHD4 to control lineage-specifying genes. Nature 557, 739–743. [PubMed: 29795351]

Pavlícek A, Jabbari K, Paces J, Paces V, Hejnar JV, and Bernardi G (2001). Similar integration but different stability of Alus and LINEs in the human genome. Gene 276, 39–45. [PubMed: 11591470]

Peddigari S, Li PW, Rabe JL, and Martin SL (2013). hnRNPL and nucleolin bind LINE-1 RNA and function as host factors to modulate retrotransposition. Nucleic Acids Res. 41, 575–585. [PubMed: 23161687]

Percharde M, Lin CJ, Yin Y, Guan J, Peixoto GA, Bulut-Karslioglu A, Biechele S, Huang B, Shen X, and Ramalho-Santos M (2018). A LINE1-nucleolin partnership regulates early development and ESC identity. Cell 174, 391–405.e19. [PubMed: 29937225]

Peric-Hupkes D, Meuleman W, Pagie L, Bruggeman SW, Solovei I, Brugman W, Grä f S, Flicek P, Kerkhoven RM, van Lohuizen M, et al. (2010). Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. Mol. Cell 38, 603–613. [PubMed: 20513434]

Qiu X, Zhu X, Zhang L, Mao Y, Zhang J, Hao P, Li G, Lv P, Li Z, Sun X, et al. (2003). Human epithelial cancers secrete immunoglobulin g with unidentified specificity to promote growth and survival of tumor cells. Cancer Res. 63, 6488–6495. [PubMed: 14559841]

Quinlan AR (2014). BEDTools: the Swiss-army tool for genome feature analysis. Curr. Protoc. Bioinform 47, 11.12.1–11.12.34.

Quinodoz SA, Ollikainen N, Tabak B, Palla A, Schmidt JM, Detmar E, Lai MM, Shishkin AA, Bhat P, Takei Y, et al. (2018). Higher-order interchromosomal hubs shape 3D genome organization in the nucleus. Cell 174, 744–757.e24. [PubMed: 29887377]

Ranzani M, Iyer V, Ibarra-Soria X, Del Castillo Velasco-Herrera M, Garnett M, Logan D, and Adams DJ (2017). Revisiting olfactory receptors as putative drivers of cancer. Wellcome Open Res. 2, 9. [PubMed: 28492065]

Rebollo R, Romanish MT, and Mager DL (2012). Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annu. Rev. Genet 46, 21–42. [PubMed: 22905872]

Rother MB, Palstra RJ, Jhunjhunwala S, van Kester KA, van IJcken WF, Hendriks RW, van Dongen JJ, Murre C, and van Zelm MC (2016). Nuclear positioning rather than contraction controls ordered rearrangements of immunoglobulin loci. Nucleic Acids Res. 44, 175–186. [PubMed: 26384565]

Rowe HM, Kapopoulou A, Corsinotti A, Fasching L, Macfarlan TS, Tarabay Y, Viville S, Jakobsson J, Pfaff SL, and Trono D (2013). TRIM28 repression of retrotransposon-based enhancers is necessary to preserve transcriptional dynamics in embryonic stem cells. Genome Res. 23, 452–461. [PubMed: 23233547]

Schmidt D, Schwalie PC, Wilson MD, Ballester B, Goncalves A, Kutter C, Brown GD, Marshall A, Flicek P, and Odom DT (2012). Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell 148, 335–348. [PubMed: 22244452]

Shen L, Shao N, Liu X, and Nestler E (2014). ngs.plot: quick mining and visualization of next-generation sequencing data by integrating genomic databases. BMC Genomics 15, 284. [PubMed: 24735413]

Shukla R, Upton KR, Muñoz-Lopez M, Gerhardt DJ, Fisher ME, Nguyen T, Brennan PM, Baillie JK, Collino A, Ghisletti S, et al. (2013). Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. Cell 153, 101–111. [PubMed: 23540693]

Solyom S, Ewing AD, Rahrmann EP, Doucet T, Nelson HH, Burns MB, Harris RS, Sigmon DF, Casella A, Erlanger B, et al. (2012). Extensive somatic L1 retrotransposition in colorectal tumors. Genome Res. 22, 2328–2338. [PubMed: 22968929]

Sultana T, van Essen D, Siol O, Bailly-Bechet M, Philippe C, Zine El Aabidine A, Pioger L, Nigumann P, Saccani S, Andrau JC, et al. (2019). The landscape of L1 retrotransposons in the human genome is shaped by preinsertion sequence biases and post-insertion selection. Mol. Cell 74, 555–570.e7. [PubMed: 30956044]

Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, Snyder MP, and Wang T (2014). Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome Res. 24, 1963–1976. [PubMed: 25319995]

Tandler CJ, and Solari AJ (1982). Methanol-acetic anhydride: an efficient blocking agent for electron microscope cytochemistry. Its application to mouse testis and other tissues. Histochemistry 76, 351–361. [PubMed: 6186647]

Taylor MS, LaCava J, Mita P, Molloy KR, Huang CR, Li D, Adney EM, Jiang H, Burns KH, Chait BT, et al. (2013). Affinity proteomics reveals human host factors implicated in discrete stages of LINE-1 retrotransposition. Cell 155, 1034–1048. [PubMed: 24267889]

Tempel S (2012). Using and understanding RepeatMasker. Methods Mol. Biol 859, 29–51. [PubMed: 22367864]

Testillano PS, Sanchez-Pina MA, Olmedilla A, Ollacarizqueta MA, Tandler CJ, and Risueño MC (1991). A specific ultrastructural method to reveal DNA: the NAMA-Ur. J. Histochem. Cytochem 39, 1427–1438. [PubMed: 1719069]

Thomas JH, and Schneider S (2011). Coevolution of retroelements and tandem zinc finger genes. Genome Res 21, 1800–1812. [PubMed: 21784874]

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, and Pachter L (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc 7, 562–578. [PubMed: 22383036]

Tubio JMC, Li Y, Ju YS, Martincorena I, Cooke SL, Tojo M, Gundem G, Pipinikas CP, Zamora J, Raine K, et al.; ICGC Breast Cancer Group; ICGC Bone Cancer Group; ICGC Prostate Cancer Group (2014). Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. Science 345, 1251343. [PubMed: 25082706]

van Steensel B, and Belmont AS (2017). Lamina-associated domains: links with chromosome architecture, heterochromatin, and gene repression. Cell 169, 780–791. [PubMed: 28525751]

Wagstaff BJ, Hedges DJ, Derbes RS, Campos Sanchez R, Chiaromonte F, Makova KD, and Roy-Engel AM (2012). Rescuing Alu: recovery of new inserts shows LINE-1 preserves Alu activity through A-tail expansion. PLoS Genet. 8, e1002842. [PubMed: 22912586]

Wang S, Su JH, Beliveau BJ, Bintu B, Moffitt JR, Wu CT, and Zhuang X (2016). Spatial organization of chromatin domains and compartments in single chromosomes. Science 353, 598–602. [PubMed: 27445307]

Wang Q, Xiong H, Ai S, Yu X, Liu Y, Zhang J, and He A (2019). Co-BATCH for high-throughput single-cell epigenomic profiling. Mol. Cell 76, 206–216.e7. [PubMed: 31471188]

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al.; Mouse Genome Sequencing Consortium (2002). Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562. [PubMed: 12466850]

Wu J, Huang B, Chen H, Yin Q, Liu Y, Xiang Y, Zhang B, Liu B, Wang Q, Xia W, et al. (2016). The landscape of accessible chromatin in mammalian preimplantation embryos. Nature 534, 652–657. [PubMed: 27309802]

Yi Q, Chen Q, Liang C, Yan H, Zhang Z, Xiang X, Zhang M, Qi F, Zhou L, and Wang F (2018). HP1 links centromeric heterochromatin to centromere cohesion in mammals. EMBO Rep. 19, e45484. [PubMed: 29491004]

Yin Y, Yan P, Lu J, Song G, Zhu Y, Li Z, Zhao Y, Shen B, Huang X, Zhu H, et al. (2015). Opposing roles for the lncRNA haunt and its genomic locus in regulating HOXA gene activation during embryonic stem cell differentiation. Cell Stem Cell 16, 504–516. [PubMed: 25891907]

Ying QL, Stavridis M, Griffiths D, Li M, and Smith A (2003). Conversion of embryonic stem cells into neuroectodermal precursors in adherent monoculture. Nat. Biotechnol 21, 183–186. [PubMed: 12524553]

Yoon KH, Ragoczy T, Lu Z, Kondoh K, Kuang D, Groudine M, and Buck LB (2015). Olfactory receptor genes expressed in distinct lineages are sequestered in different nuclear compartments. Proc. Natl. Acad. Sci. U S A 112, E2403–E2409. [PubMed: 25897022]

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, and Liu XS (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9, R137. [PubMed: 18798982]

Zhang B, Zheng H, Huang B, Li W, Xiang Y, Peng X, Ming J, Wu X, Zhang Y, Xu Q, et al. (2016). Allelic reprogramming of the histone modification H3K4me3 in early mammalian development. Nature 537, 553–557. [PubMed: 27626382]

Highlights

- SINE, L1, and low-complexity repeats barcode genes with distinct functions
- Genomic repeats dictate the time and level of gene expression during development
- L1-enriched genes are sequestered in the inactive NAD/LAD domains for silencing
- L1 RNA promotes the nuclear localization and repression of L1-enriched genes

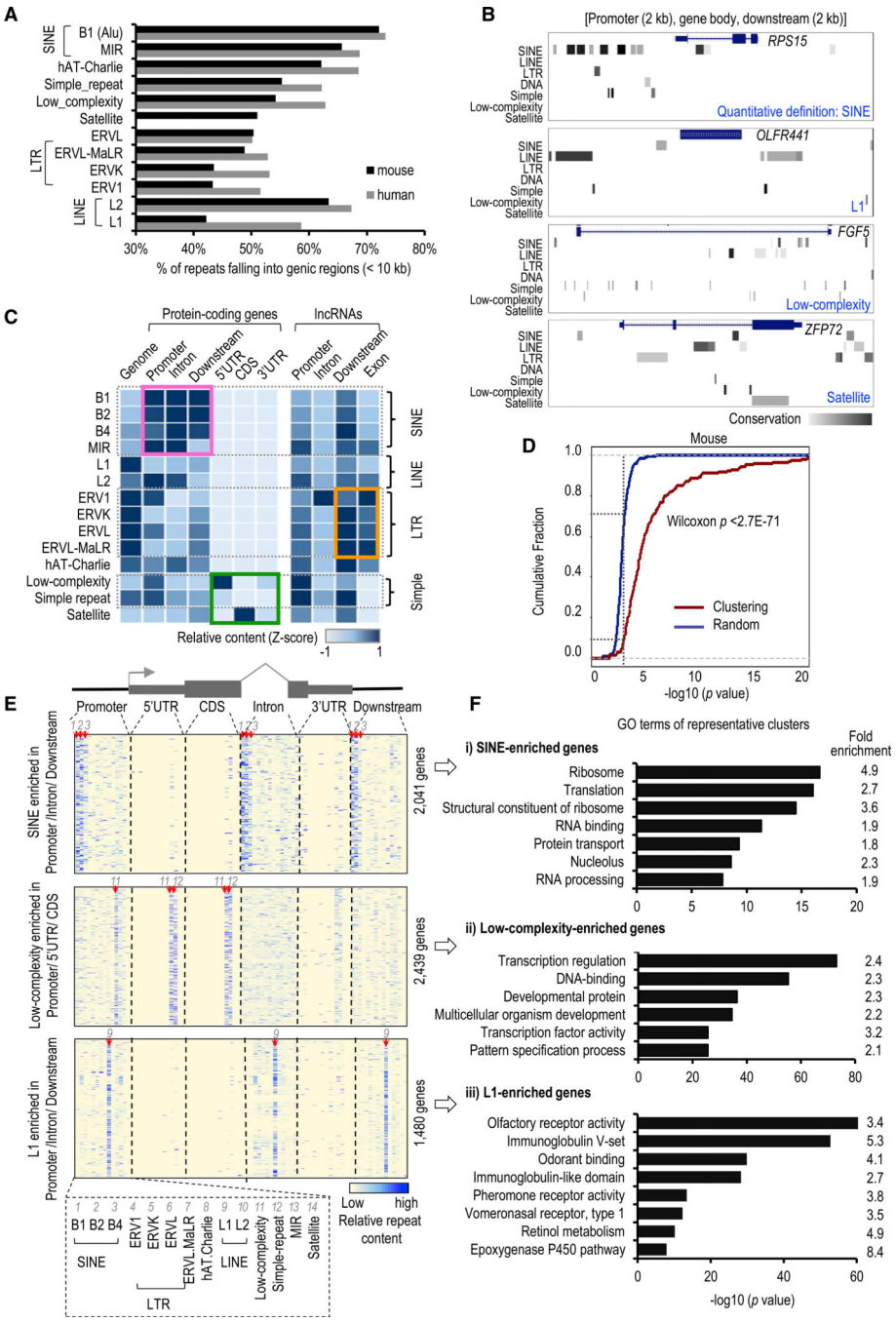


Figure 1. Genic Repeats Categorize Gene Function

(A) Percentage of repeats located within 10 kb of GENCODE-annotated genes in human and mouse.

(B) Genome Browser tracks of mouse RepeatMasker at representative loci. The SINE content in *RPS15* (46%) is ~5- to 7-fold higher than that of *OLFR441* (8%), *FGF5* (9%), and *ZFP72* (7%). *RPS15* is depleted of LINE repeats, while *OLFR441* is flanked by large stretches (~30%) of LINEs in the promoter and downstream regions. Simple and low-complexity repeats are sprinkled across the *FGF5* gene. The last exon of *ZFP72* largely

overlaps with satellite repeats. We defined *RPS15*, *OLFR441*, *FGF5*, and *ZFP72* as SINE-, L1-, low-complexity repeat-, and satellite-enriched genes, respectively, on the basis of quantitative classification of genic repeats in (E).

(C) The distribution of genic repeats in mouse. Each row presents a repeat subfamily and each column presents specific genic region. B1 shows 1.6-fold enrichment ($p < 2.2E-130$) in promoters and 1.9-fold ($p < 2.3E-200$) in introns of mRNA genes compared with lncRNAs. ERV1 shows 14.9-fold enrichment in lncRNA than mRNA genes ($p < 3.5E-202$). L1 is ~2.4-fold depleted from genic regions compared with genome background ($p < 1.6E-287$).

(D) Cumulative distribution curves (CDC) for GO analysis of mouse gene clusters generated by hierarchical clustering of genic repeat features (in red) and for that of random gene sets (in blue). With a p value of 1E-3 as the cutoff (indicated by the dotted vertical line), ~70% of the gene clusters generated by their genic repeat patterns are enriched in specific functional terms, in sharp contrast to <10% of the random gene sets. Differential cumulative distributions (Wilcoxon $p = 2.7E-71$) indicate that genes with similar repeat patterns tend to share similar biological function.

(E) Heatmaps depicting relative repeat content for genes (rows) across different genic regions (columns) in specific clusters. Top panel: SINE-enriched genes; middle panel: low-complexity repeat-enriched genes; bottom panel: L1-enriched genes. Red arrows indicate the dominant repeat subfamilies in the cluster. The cluster of satellite repeat-enriched genes is shown in Figure S3A.

(F) GO analysis of genes with specific repeat patterns defined in (E).

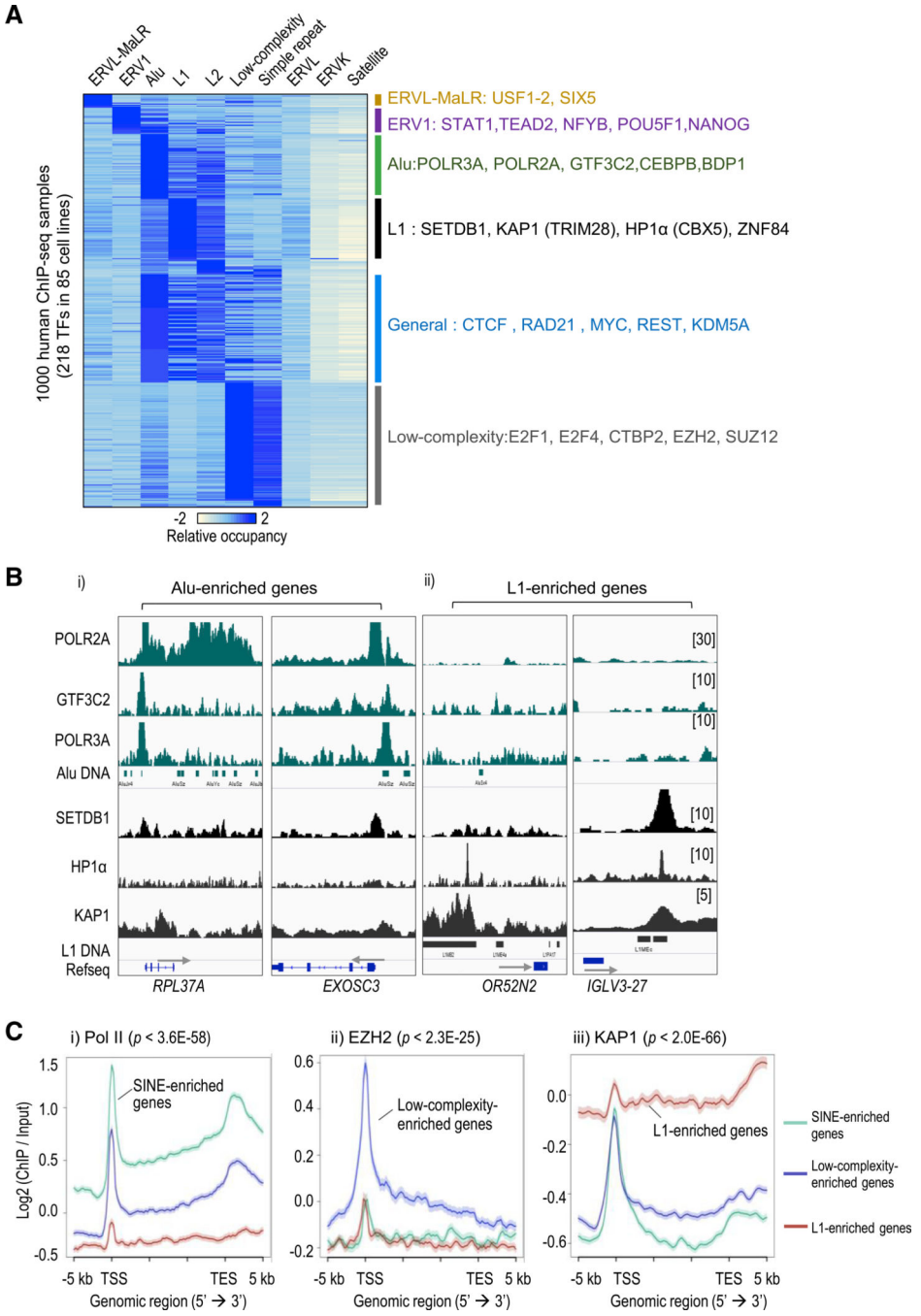


Figure 2. Repetitive Elements Are Targeted by Different Classes of Chromatin and Transcription Regulators

(A) Heatmap of chromatin binding preference for transcription factors and chromatin remodelers (rows) across different repeat subfamilies (columns). Representative proteins enriched in each repeat subfamily are shown at the right. Relative occupancy: *Z* score normalization of peak percentages in different repeat subfamilies.

(B) ChIP-seq tracks of representative loci. (i) ChIP-seq signals of POLR2A (GM12891), GTF3C2 (HeLa), and POLR3A (HeLa) at two Alu-rich gene loci with RNA metabolism-related functions (*RPL37A* and *EXOSC3*). (ii) ChIP-seq signals of SETDB1 (HEK293),

HP1α (K562), and KAP1 (K562) at two gene loci with specialized functions (*OR52N2* and *IGLV3–27*). Transcription direction is indicated by gray arrows.

(C) Metagene analysis showing differential binding activities of total Pol II (1), EZH2 (2), and KAP1 (3) at SINE-, low-complexity-, and L1-enriched genes in mouse ESCs. We show the largest p value between the gene set with the highest ChIP-seq signals and either of the other two gene sets by two-tailed Student's t test in each panel. The shadow around each line represents the standard error.

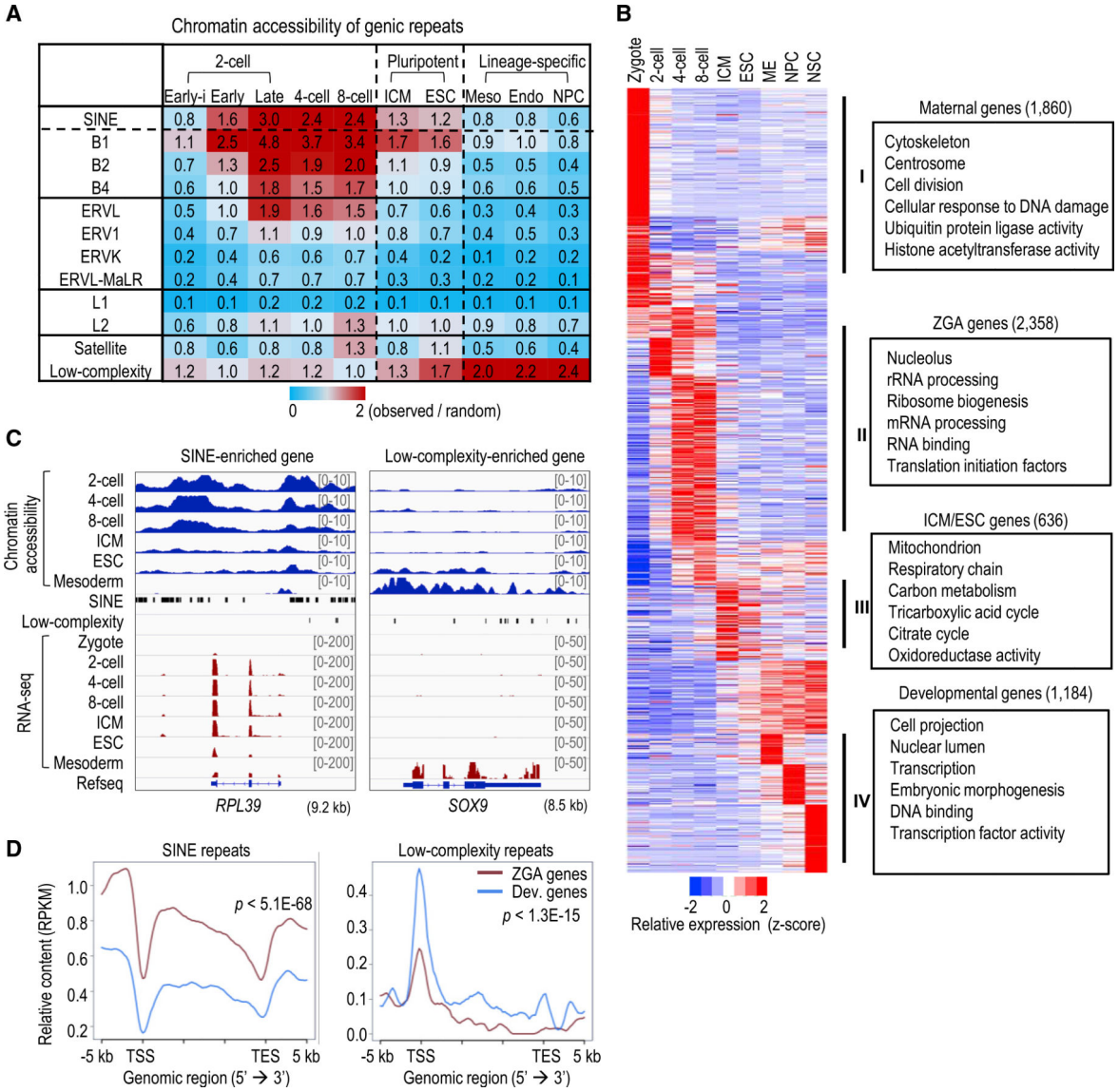


Figure 3. Orchestrated Activation of SINE and Low-Complexity Repeats Together with Their Enriched Genes during Embryonic Development

(A) Heatmap showing the enrichment of repeat subfamilies in accessible chromatin sites during early embryonic development stages or in cell lines. Enrichments are shown as ratios, calculated by dividing the number of observed peaks that overlap with repeats by the number for random genomic regions. Early-i, early two-cell embryos developed from zygotes treated with the transcription inhibitor α-amanitin. endo, endoderm; meso, mesoderm; NPC, neural progenitor cell.

(B) Heatmap showing relative gene expression patterns during early embryonic development and ESC differentiation. Genes are divided into four sets according to their expression pattern. Enriched GO terms are shown at the right. Relative expression: *Z* score normalization of fragments per kilobase million reads (FPKM). ME, mesendoderm; NPC, neural progenitor cell; NSC, neural stem cell.

(C) Signals of RNA-seq and chromatin accessibility at the SINE-enriched ribosomal gene *RPL39* and the low-complexity repeat-enriched transcription factor gene *SOX9*.
(D) The densities of SINE and low-complexity repeats around the ZGA and developmental gene sets identified in (B). p values are calculated using two-tailed Student's t test.

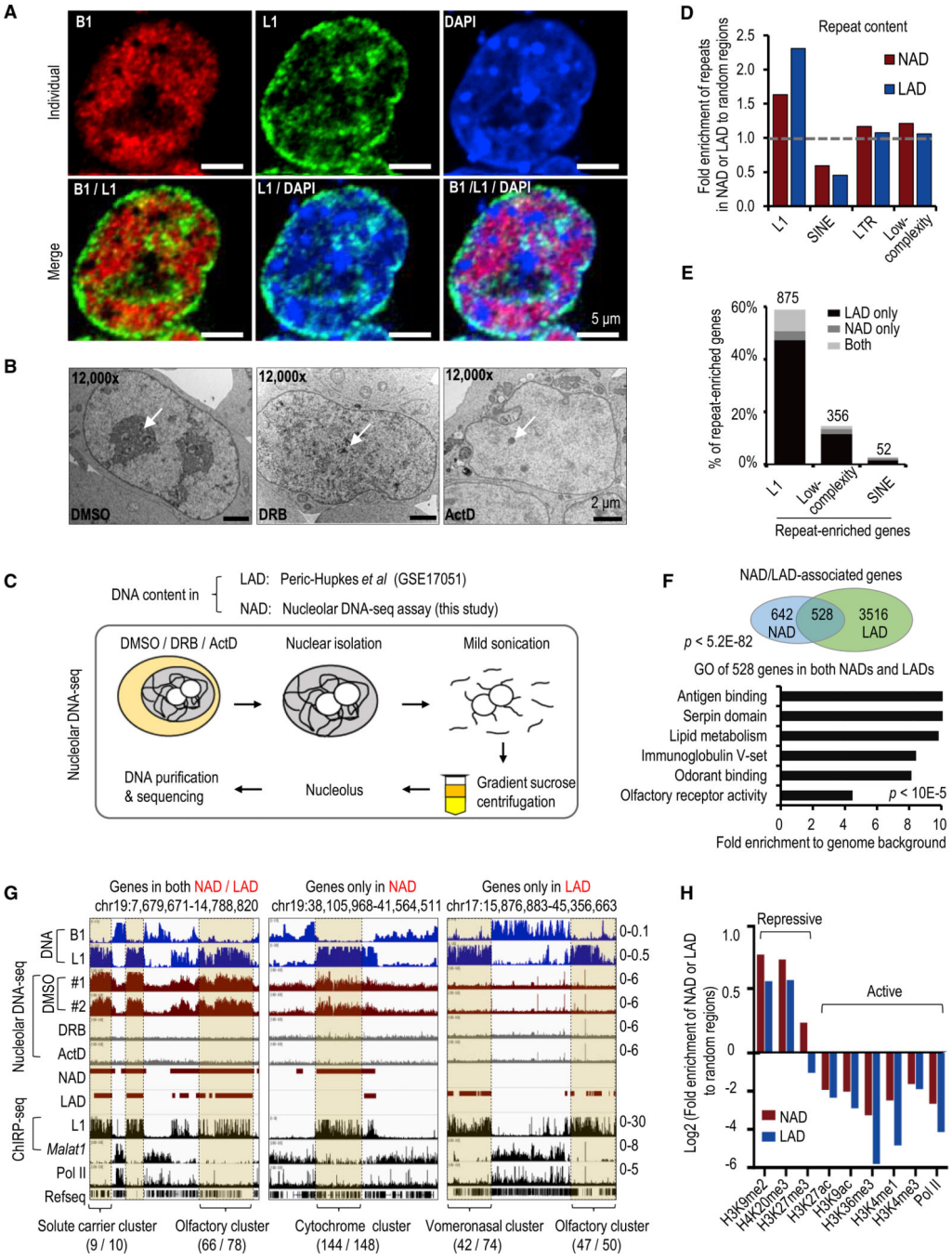


Figure 4. L1-Enriched Genes Are Sequestered in Inactive Domains in ESCs

(A) Representative images of endogenous SINE B1 (red) and L1 (green) repeats in ESCs by DNA FISH. DAPI (blue); all scale bars, 5 μm.

(B) Representative electron microscopy (EM) images of ESCs treated with DMSO (mock) and the transcriptional inhibitors DRB and ActD. Nucleoli (indicated by arrows) were shrunk and fragmented in DRB or ActD-treated cells. Scale bars, 2 μm. DRB (5,6-dichloro-1-beta-D-ribofuranosylbenzimidazole, 25 μM, 2 h) inhibits the release and elongation of Pol II; ActD (actinomycin D, 1 μg/mL, 2 h) inhibits both Pol I and II.

(C) Summary diagram of nucleolar DNA-seq. See STAR Methods for more detail.
(D) Genomic contents of repeat elements in NADs and LADs. Fold enrichments of individual repeats were compared with random genomic regions.
(E) Bar chart showing the percentage of repeat-enriched genes (defined in Figure 1E) that are located in NADs, LADs, or both domains in ESCs.
(F) NAD/LAD-associated genes. The top Venn diagram shows genes that were detected in NADs and/or LADs. $p < 5.2E-82$ by exact hypergeometric probability. Bottom panel: GO analysis of the 528 genes that overlap between NADs and LADs. All p values for each enriched term shown are <10E-5.
(G) Tracks of genomic regions containing clusters of genes located in both NADs/LADs (left), NADs only (middle), and LADs only (right). The genomic density of B1 and L1 repeats, nucleolar DNA-seq signals, annotated NAD/LAD, ChIRP-seq of L1 RNA and *Malat1*, and Pol II ChIP-seq in ESCs are shown. For gene clusters shown at the bottom, the number of genes involved in the indicated function and the total number of genes in each cluster are indicated before and after the forward slash, respectively.
(H) Chromatin features of NADs and LADs.

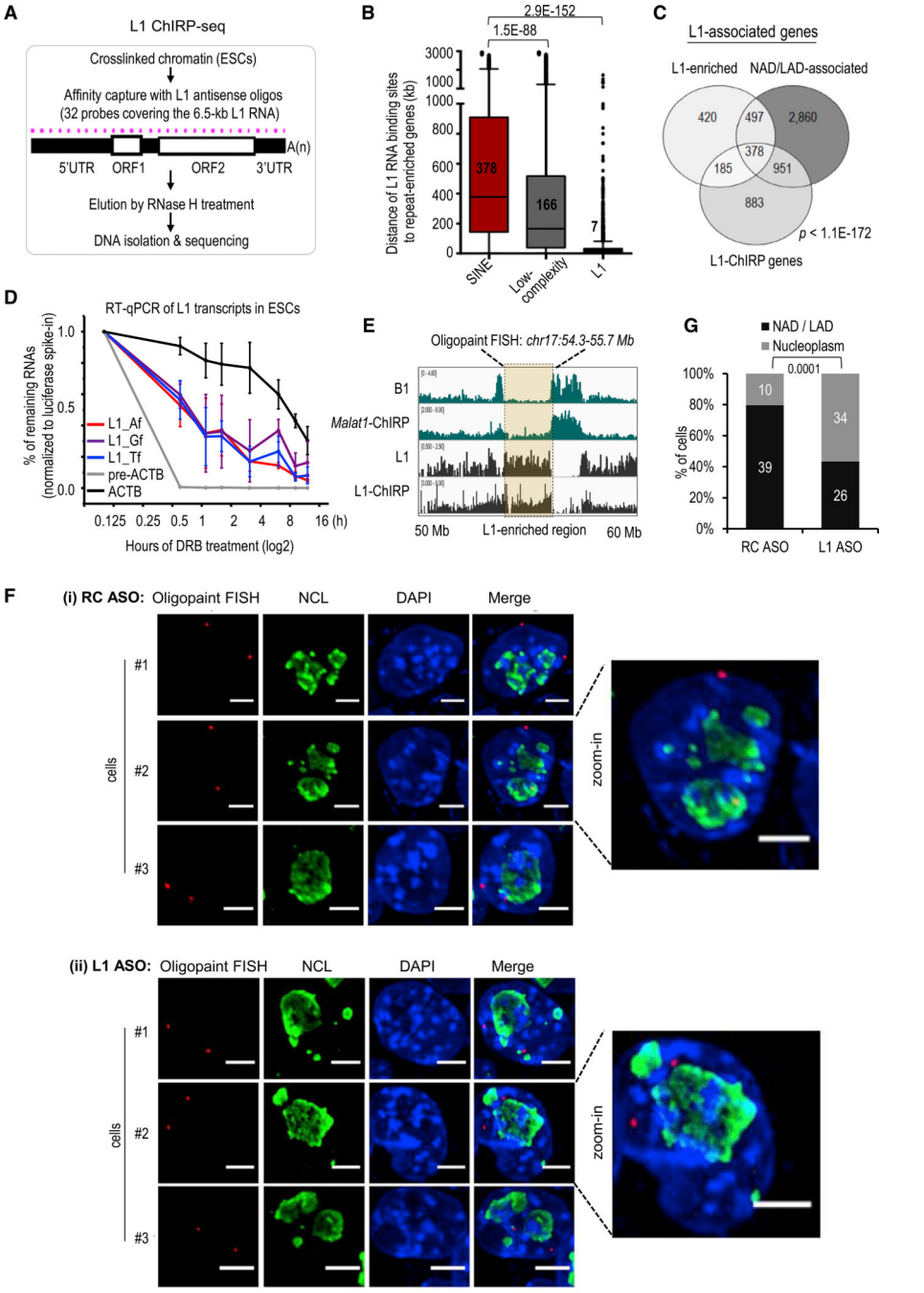


Figure 5. L1 RNA Binds L1 DNA and Sequesters L1-Rich DNA Region in NADs and LADs

(A) Schematic diagram of L1 ChIRP-seq. See STAR Methods for more detail.

(B) Boxplot showing the distances of L1 RNA binding sites to the nearest repeat-enriched gene. The median distances (kb) are indicated. p values are calculated using two-tailed Student's t test.

(C) Venn diagram showing the overlap between L1-enriched genes, NAD/LAD-associated genes, and L1-ChIRP genes. Genes in these three sets are termed as L1-associated genes. p value was calculated by exact hypergeometric probability.

(D) Half-life analysis of L1 RNA by quantitative RT-qPCR in ESCs. The x axis shows hours in $\log_2$ scale after inhibition of transcription by DRB (100 μM). The y axis shows relative RNA levels of three subtypes of L1 repeats and *ACTB* to spike-in mRNA and to the level of corresponding transcripts prior to DRB treatment. Error bars represent ± standard deviation of the mean in two independent replicates of DRB treatments.
(E) Genomic tracks of a representative locus for Oligopaint FISH (F and G). Genomic contents for B1 and L1 DNA and ChIRP-seq signals of L1 and *Malat1* are shown.
(F and G) Oligopaint DNA FISH of a L1-enriched region (red signals) shown in (E) combined with immunofluorescence for NCL (green). Representative images and quantification of its nuclear localization in mESC transfected with RC or L1 ASO are shown in (F) and (G), respectively. DAPI (blue); all scale bars, 5 μm.

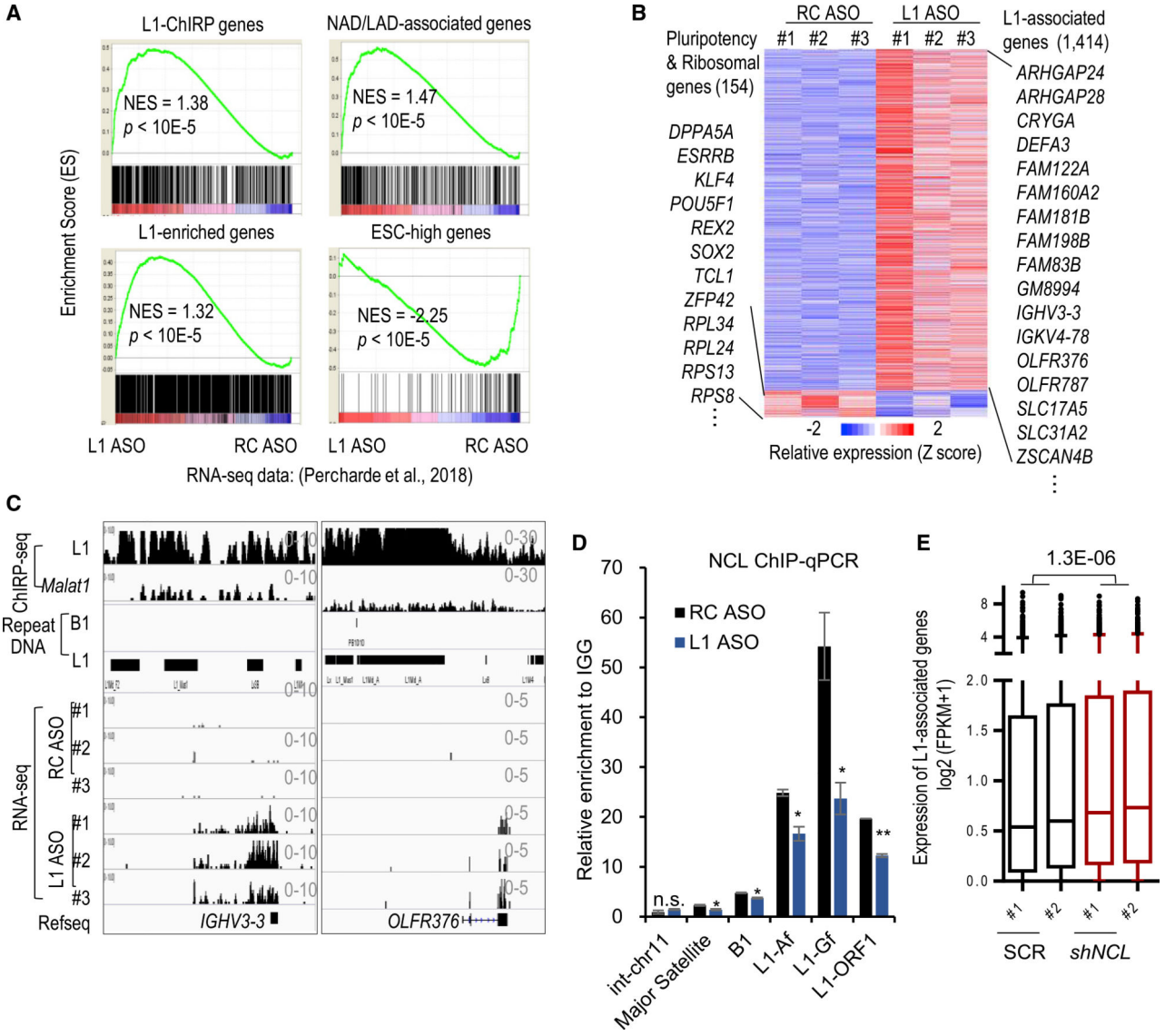


Figure 6. L1 RNA Binds NCL and Represses L1-Associated Genes

(A) Gene set enrichment analysis (GSEA) showing global upregulation of L1-ChIRP genes, NAD/LAD-associated genes, and L1-enriched genes but downregulation of ESC-high genes in mESCs transfected with L1 ASO compared with the control RC ASO-treated cells. Normalized enrichment scores (NES) and nominal p values are shown.

(B) Heatmap of 1,414 upregulated L1-associated genes that were defined as core enrichment genes by GSEA in (A). Pluripotency and ribosomal genes (154) are included for comparison. Names of representative genes are shown. Relative expression: *Z* score normalization of FPKM.

(C) Tracks of the *IGHV3–3* and *OLFR376* regions. L1 and *Malat1* ChIRP-seq, RepeatMasker annotations of B1 and L1, and RNA-seq of ESCs treated with L1 or RC ASO (three biological replicates shown).

(D) Low-input *in situ* ChIP-qPCR analysis of NCL in ESCs treated with L1 or RC ASO. Data are shown as fold enrichments to the IgG control. Error bars represent standard

deviation in two independent experiments. p values are calculated using two-tailed Student's t test. *$p < 0.05$ and **$p < 0.01$.

(E) Boxplot showing the expression level of L1-associated genes in *NCL*-deleted ESCs. p values are calculated using two-tailed Student's t test.

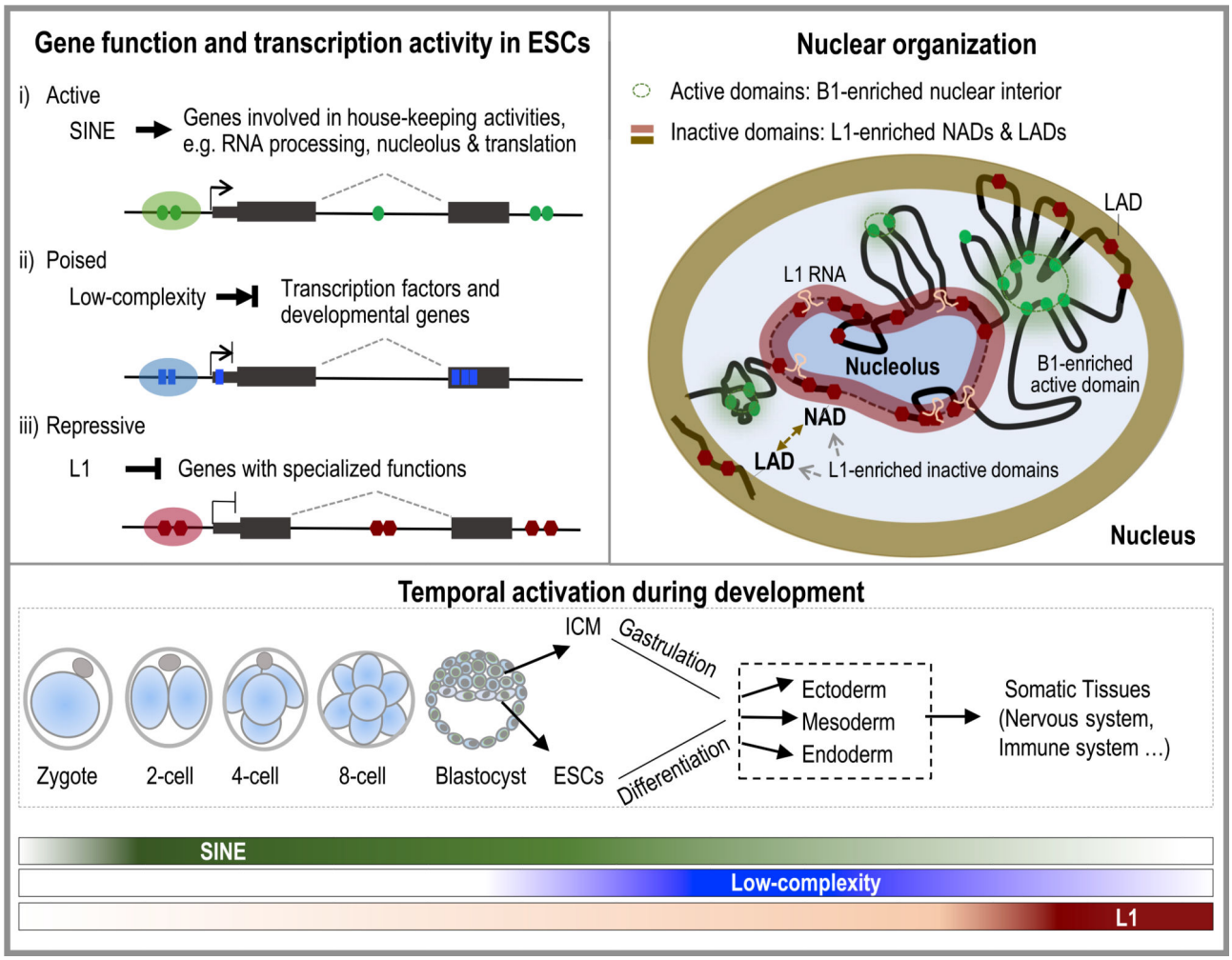


Figure 7. A Model Depicting Three Aspects of Repetitive Elements in Categorizing Genes with Distinct Functions for Orchestrated Regulation and Expression

First, SINE, L1, and low-complexity repeats classify genes with distinct functions that are associated with different levels of transcription activity (left panel). Second, SINE and L1 repeats sequester their enriched genes in distinct active and inactive nuclear domains for coordinated activation or silencing, respectively (right panel). In particular, L1 RNA binds L1 DNA to facilitate its function in silencing L1-enriched genes that are associated with the inactive NADs and LADs in the peripheries of the nucleolus and nucleus. Third, temporal activation of repeats and repeat-enriched genes during development and differentiation (bottom panel). After fertilization and before the blastocyst stage, SINE repeats become active and housekeeping genes related to RNA processing, ribosome biogenesis, and nucleolus function are highly expressed. When the pluripotent cells in the ICM of the blastocyst or in cultured ESCs differentiate into three embryonic germ layers, low-complexity repeat-enriched genes, which typically encode developmental transcription factors, are highly expressed. In terminally differentiated cells, L1 repeats and L1-enriched genes become activated. We propose that genomic repeats shape transcription regulatory

networks to achieve orchestrated activation or silencing of genes with distinct functions at specific stages. The activation levels are shown as the degree of color darkness.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit polyclonal anti-NCL	Santa Cruz Biotech	Cat# sc-13057; RRID:AB_2229696
IgG	Cell Signaling Technology	Control #3900; RRID:AB_1550038
Reagents, Chemicals and Peptides		
Actinomycin D (ActD)	Sigma Aldrich	Cat# A4262
DRB	Abcam	Cat#: ab120939
Proteinase K	Merck Millipore	Cat# 539480
Protease inhibitor cocktail	Selleck	Cat# K4000
β-mercaptoethanol	Sigma Aldrich	Cat# M7522
Non-essential amino acids	Life Technologies	Cat#11140050
L-Glutamine	Life Technologies	Cat# 25030081
Penicillin-streptomycin	Life Technologies	Cat#15140122
Polybrene	Sigma	Cat# 107689
Puromycin	Life Technologies	Cat#A1113802
Trizol	Thermo Fisher	Cat#15596018
Lipofectamine 3000	Thermo Fisher	Cat# L3000015
8% Paraformaldehyde	EMS	Cat# 157-8
Critical Commercial Assays		
RevertAid First Strand cDNA Synthesis Kit	Thermo Fisher	Cat# K1622
Dynabeads mRNA purification kit	Thermo Fisher	Cat# 61006
MinElute PCR Purification Kit	QIAGEN	Cat# 28004
NBT/BCIP stock solution	Roche	Cat# 11681451001
NEBNext Ultra II First Strand Synthesis Module	NEB	Cat# E7771L
NEBNext Ultra II Second Strand Synthesis Module	NEB	Cat# E7550L
NEBNext Ultra II DNA Library Prep Kit	NEB	Cat# E7645
Strep-Tactin XT purification system	IBA Lifesciences	Cat# 2-1201/2-1000
Colloidal Blue Staining Kit	Thermo Fisher	Cat# LC6025
Experimental Models: Cell Lines		
Human: HEK293T	ATCC	CRL-3216
Mouse: 46C ES	Austin Smith Lab	Ying et al., 2003
Mouse: CJ9 ES	Shen X lab	Luo et al., 2016
Deposited Data		
Nucleolar DNA	This study	GEO: GSE103610
ChIRP-seq of L1 RNA	This study	GEO: GSE125766
Oligonucleotides		
A full list of oligos is provided in Table S6		
Software and Algorithms		
ImageJ (1.51h)	NIH, Univ. of Wisc. Madison	https://imagej.nih.gov/ij/

REAGENT or RESOURCE	SOURCE	IDENTIFIER
IGV (2.4.14)	Broad Institute	https://software.broadinstitute.org/software/igv/
Graphpad Prism (6.0)	Graphpad	https://www.graphpad.com/scientific-software/prism/
Bowtie	Langmead and Salzberg, 2012	http://bowtie-bio.sourceforge.net/index.shtml
Tophat	Trapnell et al., 2012	http://ccb.jhu.edu/software/tophat/index.shtml
Cufflinks	Trapnell et al., 2012	http://cole-trapnell-lab.github.io/cufflinks/
Gene set enrichment analysis (GSEA)	Broad Institute	RRID:SCR_003199
R language	R Core Team	R x64 3.4.3