Systematic identification of functional elements derived from human endogenous retroviruses

Ito, Jumpei

Doctor of Philosophy

Department of Genetics

School of Life Science

SOKENDAI (The Graduate University of Advanced Studies)

SUMMARY

Human endogenous retrovirus (HERV) belongs to a class of transposable elements (TEs) and occupies approximately 8% of the entire human genome. Although HERVs were initially thought to be non-functional and merely parasitic sequences in the genome, evidences have been accumulated that the human genome carries enormous HERV-derived functional elements, which affect host physiologies and diseases. These elements play a role in the DNA level (i.e., transcriptional regulatory elements), RNA level (i.e., non-coding RNA), or protein level. Recent efforts using high-throughput sequencing have generated a massive amount of genomic, epigenomic, and transcriptomic data. In this doctoral thesis, I aimed to identify HERV-derived functional elements (particularly regulatory elements and transcripts) by reanalyzing epigenomic and transcriptomic sequencing data accumulated in the public databases.

In Chapter 1 (general introduction), I describe basic knowledge about HERVs and recent progresses that have showed associations of HERV-derived functional elements with host physiologies and diseases.

In Chapter 2, I investigated HERV-derived regulatory elements using 519 ChIP-Seq data for 97 transcription factors (TFs) provided by ENCODE and Roadmap Epigenomics. I identified 794,972 TF-binding events on HERVs and 2,201 specific HERV-TF associations. Using unsupervised clustering analysis, I demonstrated that HERVs could be grouped according to TF binding patterns: HERV groups bound by pluripotent TFs (e.g., SOX2, POU5F1, and NANOG), embryonic endoderm/mesendoderm TFs (e.g., GATA4/6, SOX17, and FOXA1/2), hematopoietic TFs (e.g., SPI1 (PU1), GATA1/2, and TAL1), and CTCF were identified. By analyzing the three-dimensional chromosomal interactions, I demonstrated that HERV-derived regulatory elements tend to interact with host genes relating to the innate immune response. This suggests that the HERV-derived regulatory elements play a role in the modulation of this biological pathway. We further demonstrated heterogeneities of regulatory elements within LTR7 group: SOX2, POU5F1, and KLF4-binding sites were highly enriched in the youngest subgroup of LTR7, which had the highest transcriptional activity in pluripotent cells. This suggests that the subgroup acquired those regulatory activities for efficient replication in the host germ cells. Furthermore, my colleagues and I

constructed dbHERV-REs (http://herv-tfbs.com/), a database of HERV-derived regulatory elements.

In Chapter 3, I investigated HERV-derived transcripts in tumors by reanalyzing RNA-Seq data of 5,550 patients across 12 solid tumors provided by TCGA. I identified 10,060 transcribed HERV loci in tumors and the corresponding normal tissues. In nine out of 12 tumor types, the overall transcription levels of HERVs significantly increased. Particularly, transcription levels of HERVH group were highly up-regulated in a broad range of tumors. In unsupervised clustering analysis based on the HERV transcriptome, RNA-Seq samples clustered according to tumor and tissue types, and even molecular subtypes within a type of tumors. This indicates that HERVs had unique transcription profiles among tumor/tissue types and the subtypes. The transcriptionally up-regulated HERVs in tumors were associated with TFs that were overexpressed in the tumors. In case of breast cancer, the up-regulated HERVs tended to be bound by ESR1 (estrogen receptor 1), PGR (progesterone receptor), GATA3, and FOXA1, which were overexpressed in the cancer type. Furthermore, transcription levels of HERVs in tumors were positively correlated with those of genes targeted by ZNF274, TRIM28, and/or SETDB1. These are known to form a protein complex and suppress the transcription of TEs in mouse embryos. This result suggests that these genes work on the transcriptional silencing of HERVs also in human tumors. Some HERVs were transcribed as parts of mRNA of genes and contributed to produce non-canonical transcripts of those genes. For example, the fused transcript of ERVL-B4-int and TMPRSS4, a major causal gene of prostate cancer, was highly up-regulated in prostate adenocarcinoma. The ERVL-B4-int locus worked as an alternative transcription start site of TMPRSS4 and contributed to the overexpression of this gene in the tumors.

This doctoral thesis depicts the landscape of HERV-derived functional elements providing insights into effects of these elements on host physiologies and diseases.

Table of Contents

Chapter 1: General Introduction

Transposable element (TE), known as mobile element, is a DNA sequence that can change its position within a genome [1]. TEs constitute the majority of the eukaryotic genome and play an essential role in expanding its genome size [1-3]. TEs are classified into two major groups according to the mechanism of the transposition; retrotransposon and DNA transposon (Fig. 1) [1]. Retrotransposon is transcribed from DNA to RNA, and then the RNA is reversely transcribed to DNA (complementary DNA (cDNA)) by its reverse transcriptase [1, 3]. The cDNA is inserted into a new genomic position [1, 3]. This manner of the transposition is referred to as *copy and paste* because the original insert is retained after the transposition [1, 3] (Fig. 1). DNA transposon is cut out from the host chromosome by its transposase, and then the cut DNA is inserted into a new genomic position [1, 4]. This manner of the transposition is referred to *cut and paste* because the original insert is lost after the transposition (Fig. 1) [1, 4]. Due to the difference in the transposition manner, retrotransposons tend to increase their copy number in the host genome more than DNA transposons [2, 3]. Retrotransposon is classified into three major groups; long terminal repeat (LTR)-type retrotransposon, long interspersed element (LINE), and short interspersed element (SINE) [1, 3]. LTR-type retrotransposon is further classified into endogenous retrovirus (ERV) and others [5, 6]. ERVs highly resemble exogenous retroviruses (i.e., retroviruses horizontally transmitting such as human immunodeficiency virus type 1 (HIV-1)) probably because ERVs were arisen from exogenous retroviruses (described in the later paragraph) [6]. In this doctoral thesis, I refer to LTR-type retrotransposons simply as ERVs because almost all LTR-type transposons in mammalian genomes are ERVs. In the human genome, repetitive sequences derived from TEs occupy almost half of the genome [7]; LINEs (20% of the entire genome), SINEs (13%), DNA transposon (3%), and human ERVs (HERVs) (8%).

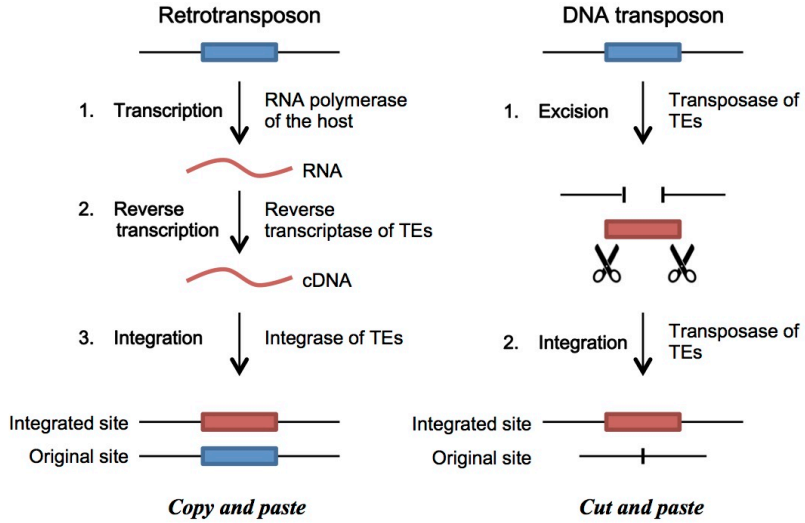


Figure 1. Transposition manners of retrotransposon and DNA transposon.

HERVs are composed of 5′- and 3′-LTR sequences and an internal sequence (Fig 2A) [6]. The LTRs contain various regulatory elements and modulate viral transcription [6]. HERVs are transcribed as mRNA species by the host machinery including RNA polymerase II (Pol II) [6]. The internal sequence contains three viral genes: *gag* (encoding structural proteins of the viral core), *pol* (encoding reverse transcriptase, integrase, and protease), and *env* (encoding envelope protein) [6]. Although most HERVs have lost the protein-coding capacity due to the accumulation of mutations, some still retain the capacity [5, 8]. In the host chromosome, HERVs are present either as a complete structure (referred to as provirus) or as a single LTR structure (referred to as solo-LTR) [5, 6] (Fig. 2A). The solo-LTR was generated through the homologous recombination between 5′- and 3′-LTRs (Fig. 2B) [5, 6].

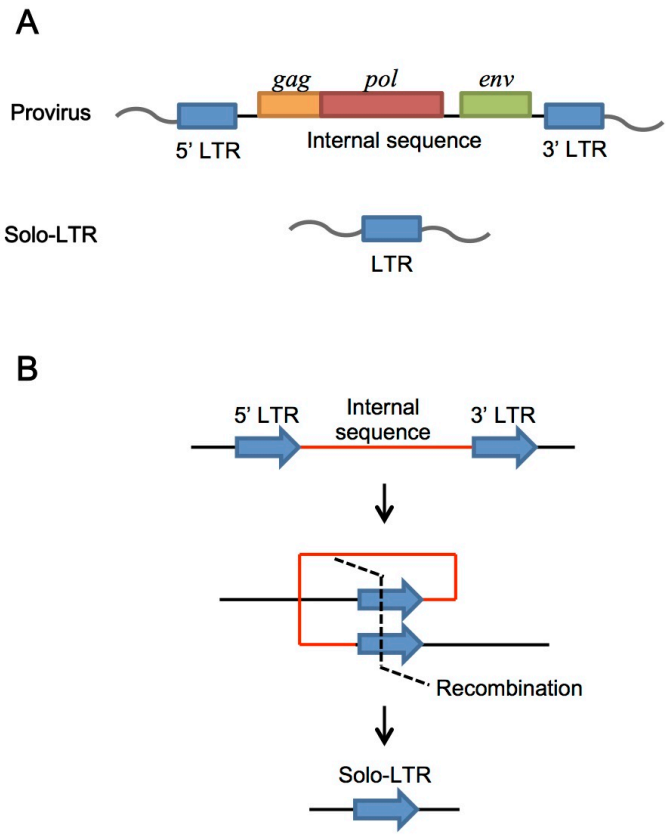


Figure 2. Genomic structures of HERVs. A) Genomic structures of provirus and solo-LTR. B) A mechanism of solo-LTR formation. Recombination between the two LTRs of a provirus results in deletion of the internal sequence and formation of the solo-LTR.

HERVs were generated from ancient exogenous retroviruses, which were horizontally transmitted among individuals of the host (the ancestors of humans), through the infection to host germ cells [6]. This step is referred to as retroviral endogenization [9]. The endogenization could have repeatedly occurred because >100 of phylogenetically distinct groups of HERVs are present in the human genome [5]. After the endogenization, HERVs have vertically transmitted to the host offspring as a part of host chromosomes [6]. While transmitting vertically, HERVs increased their copy number in the host genome by replicating (i.e., re-infecting or retrotransposing) in germ cells [6]. Through this replicating process, HERVs could have gradually lost an "identity" as exogenous retroviruses (e.g., high pathogenesis, horizontal transmission,

and replicating in somatic cells) and gained that as TEs (e.g., low pathogenesis and replicating in germ cells). Present-day HERVs have lost their replication and transposition activities in germ cells owing to the accumulation of mutations [6]. Because the mutation rate of host chromosomes is much slower than those of exogenous retroviruses and replicating TEs, present-day HERVs can be considered as genomic "fossils" of ancient exogenous retroviruses and past HERVs that replicated in germ cells [9]. Therefore, the traits and evolutionary dynamics of ancient retroviruses and their descendants can be inferred by scrutinizing present-day HERVs [9].

TEs were initially thought to be non-functional and parasitic sequences of the genome [10]. However, evidences have been accumulated that functional elements derived from TEs, particularly HERVs, have diverse effects on host physiologies [9, 11, 12]. The phenomenon is called as "exaptation", "co-option", or "domestication" of TEs because functional elements of selfish TEs have came to play a role in biological processes of the host [9]. TEs can work at three biochemical levels; DNA, RNA, and proteins. At the DNA level, TEs work as regulatory sequences (i.e., promoter, enhancer, and insulator) that modulate transcriptions of host genes [13, 14]. This is because TEs harbor various regulatory elements (transcription factor-binding sites (TFBSs)) that originally modulated TE's transcriptions [13, 14]. For example, LTR7 (a group of HERVs) harbors POU5F1- (OCT4-), SOX2-, KLF4-, and NANOG-binding sites [15-18]. The insertions of LTR7 modulate transcriptions of protein-coding/non-coding genes that are essential for maintaining the cellular pluripotency [15-18]. At the RNA level, TEs work as non-coding RNA, especially as long non-coding RNA (lncRNA) [19]. Indeed, most (>70%) of lncRNA contain sequences derived from TEs [19]. As mentioned above, LTR7-derived lncRNAs work in maintaining the cellular pluripotency [15-18]. Several studies showed that the over-expression of HERV RNAs trigger innate immune responses via sensors to exogenous RNA (i.e., viral RNA) [20-22]. At the protein level, a retroviral envelope protein encoded by HERVW,

Syncytin-1, works in placentation [11, 23]. Thus, several lines of evidences support that TEs have served as functional elements in the host. However, these examples seem to be a tip of the iceberg because 1) enormous TEs (>4.6 million of TE loci; >1,000 of TE groups) have been identified in the human genome, and 2) functions of a limited number of TEs have been experimentally examined. In *Initial sequencing and analysis of the human genome*, Lander ES et al. mentioned that TEs are extraordinary trove of information about biological process [7]. Exploration of the trove, that is, comprehensive investigation of TE-derived functional elements is needed.

Some TE-derived functional elements are associated with specific diseases, particularly cancers [24, 25]. In tumors particularly treated with DNA methyltransferase inhibitor, a set of HERV loci works as alternative promoters of host genes and alters global transcriptome patterns [26, 27]. In breast cancer, the overexpression of the envelope protein of HERVK activates the proliferation and migration of the cancer cells via the activation of Ras/ERK pathway [28-30]. The overexpression of protein-coding HERVs is associated with high infiltration of cytotoxic T cells in several tumors [31, 32]. This is probably because HERVs work as cancer-germ cell antigens, which are expressed only in germ cells and tumors and associated with antigen-specific responses to tumors [31, 32]. Thus, TE-derived functional elements can affect both on host physiologies and diseases.

Recent efforts using high-throughput sequencing techniques have produced a massive amount of genomic, epigenomic, and transcriptomic data of various cells. Particularly, several international consortiums have played a central role in the data production: Encyclopedia of DNA Elements (ENCODE) [33] and Roadmap Epigenomics (Roadmap) [34, 35] projects have decoded epigenomic and transcriptomic states of hundreds of primary and cultured cells in order to characterize functional elements in the human genome. The Cancer Genome Atlas (TCGA) [36] has produced multi-dimensional omics data of >10,000 tumors across >30 tumor types in order to

identify cancer-causal mutations, therapeutic targets, and molecular subtypes within a tumor type. The subtype classification is clinically important because these subtypes show distinct phenotypes (e.g., drug response and diagnose) [36]. Importantly, the data is publicly available, and researcher can reuse the data for their professional interests. Since TE is a part of the genome, we can extract epigenomic and transcriptomic information about TEs from these data.

In this doctoral thesis, I would like to discuss human functional elements (particularly regulatory elements and transcripts) derived from TEs. I have particularly focused on HERVs among TEs because HERVs showed the strongest statistical associations with functional elements among TEs in previous studies [19, 37, 38]. The aims of this doctoral thesis are: 1) to identify HERV-derived functional elements based on publicly available epigenomic and transcriptomic datasets, and 2) to make a catalog of the HERV-derived functional elements. First, I systematically identified HERV-derived regulatory elements using ENCODE [33] and Roadmap [35] datasets (CHAPTER 2). Based on the results, my colleague and I developed dbHERV-REs (http://herv-tfbs.com/), a database of HERV-derived regulatory elements. Second, I comprehensively identified unannotated transcripts derived from HERVs in solid tumors using TCGA dataset [36] (CHAPTER 3). Additionally, by examining regulatory elements of present-day HERVs, I attempted to illustrate evolutionary dynamics of the transcriptional regulation system of HERVs that had occurred probably for the adaptation to germ cells (CHPTER 3). Through this doctoral thesis, I would like to discuss about various effects of HERV-derived functional elements on host physiologies and diseases.

Chapter 2: Systematic identification of regulatory elements derived from human endogenous retroviruses

The contents of this chapter are also described in the below paper: Systematic identification and characterization of regulatory elements derived from human endogenous retroviruses. Ito J. et al. (2017) PLoS Genet.

Introduction

Transposable elements (TEs) were initially thought to be parasitic, selfish, and junk DNA [7, 10]. However, evidences have been accumulated that some TEs are co-opted by the host and acquire new physiological functions as protein-coding/-non-coding genes and regulatory elements for host genes [9, 16-19, 23, 37, 39-44]. TEs have their own regulatory elements for transcription and replication [16-18, 37, 42-53]. Such TE-derived regulatory elements are abundant in the human genome and have various effects on transcriptional modulations of host genes as promoters, enhancers, and insulators [14, 16-18, 37, 42-44, 54-62]. Notably, numerous TE insertions sharing the same regulatory elements can affect multiple genes in a coordinate manner. Several studies have suggested that TE insertions have contributed to the rewiring and evolution of regulatory networks by recruiting multiple genes into the same regulatory circuit [13, 14, 16-18, 37, 38, 42-44, 62, 63].

Human endogenous retroviruses (HERVs) are a class of TEs that developed through the infection of host germ cells by ancient retroviruses, followed by their transmission to the offspring (referred to as endogenization) [6]. HERVs occupy approximately 8% of the human genome [7]. HERVs have lost their replication and transposition activities in germ cells owing to the accumulation of mutations [6]. According to RepeatMasker (20-Mar-2009) (http://www.repeatmasker.org/), 375 and 130 groups of LTRs and internal sequences of HERVs, respectively, have been discovered in the human genome. This indicates that HERVs show the greatest diversity for all classes of human TEs.

HERVs are transcribed as mRNA by RNA polymerase II (Pol II), and many regulatory elements bounded to Pol II-associated transcription factors (TFs) are present in LTR sequences [6, 64]. HERVs show the highest enrichment in regulatory sequences such as open chromatin regions among all classes of human TEs [37, 38, 65]. Reflecting the considerable diversity of HERVs, each group of HERVs has various regulatory

elements involved in modulating diverse host genes [16-18, 37, 42-44, 49-53]. For instance, LTR7 insertions provide POU5F1- (OCT4-), SOX2-, KLF4-, and NANOG-binding sites for protein-coding/non-coding genes, which are essential for maintaining pluripotency in embryonic stem (ES) and induced pluripotent stem (iPS) cells [15-18, 42, 64, 65]. As a further example, MER41 insertions harboring STAT1- and IRF1-binding sites in several genes contribute to primate-specific interferon responses [43]. Clarifying the properties of HERVs regulatory elements provides a better understanding of their impact on host transcriptional regulation.

I systematically identified and characterized regulatory elements derived from HERVs based on publicly available datasets of chromatin immunoprecipitation followed by sequencing (ChIP-Seq) of sequence-specific TFs. The ChIP-Seq datasets were provided by ENCODE [33] and Roadmap Epigenomics (Roadmap) (Tsankov *et al.* [35]) projects. Previous studies have comprehensively investigated regulatory elements of TEs (including HERVs) based on the ENCODE dataset [33, 37, 38, 64-66]. Jacques *et al*. demonstrated that the majority of primate-specific regulatory sequences are derived from TEs [37]. Because this particular study was mainly focused on the dataset of DNase I hypersensitive sites (DHSs), it provided a limited insight into the specific associations of TEs and TFs [37]. Sundaram *et al*. showed specific associations of TEs and TFs using a dataset of ChIP-Seq for TFs [38]. However, the number of sequence-specific TFs investigated in that study was restricted (15 sequence-specific TFs) owing to the focus on TFs for which ChIP-Seq was performed in both human and mouse cells to compare the binding profiles [38]. In the present study, I performed a more comprehensive study than earlier of regulatory elements on HERVs by evaluating 519 ChIP-Seq datasets of 97 sequence-specific TFs (Table 1). Furthermore, my colleagues and I constructed dbHERV-REs, a database of HERV regulatory elements with an interactive interface (http://herv-tfbs.com/). This study provides fundamental information to understand the impact of HERVs on host transcription.

Table 1. TFs for which ChIP-Seq data was used in the present study.

Dataset	TFs
ENCODE	ATF3, BATF, BCL11A, BCL3, BCLAF1, BHLHE40, BRCA1, CEBPB, **CTCF**, CTCFL, E2F4, E2F6, EBF1, EGR1, ELF1, ELK4, ESR1, ETS1, FOS, FOSL1, FOSL2, **FOXA1**, **FOXA2**, FOXP2, GABPA, GATA1, GATA2, GATA3, **HNF4A**, HNF4G, IRF1, IRF3, JUN, JUNB, JUND, MAFF, MAFK, MAX, MEF2A, MEF2C, MXI1, **MYC**, **NANOG**, NFE2, NFKB1, NFYA, NFYB, NR2C2, NR3C1, NRF1, PAX5, PBX3, POU2F2, **POU5F1**, **PRDM1**, REST, RFX5, RXRA, SIX5, **SP1**, SP2, SPI1, SREBF1, SRF, STAT1, STAT2, STAT3, TAL1, TCF12, TCF7L2, THAP1, USF1, USF2, YY1, ZBTB33, ZBTB7A, ZNF143, ZNF263, ZNF274
Roadmap	**CTCF**, EOMES, **FOXA1**, **FOXA2**, GATA4, GATA6, HAND1, HAND2, HEY1, HNF1B, **HNF4A**, KLF5, LEF1, **MYC**, **NANOG**, OTX2, PAX6, **POU5F1**, **PRDM1**, SMAD1, SMAD2/3, SMAD4, SNAI2, SOX17, SOX2, **SP1**, TCF4

Bold TFs were used for ChIP-Seq by ENCODE and Roadmap.

Results

Detection of transcription factor-binding sites (TFBSs) using ChIP-Seq datasets

I analyzed 519 ChIP-Seq datasets provided by ENCODE and Roadmap. The datasets included ChIP-Seq analysis of 97 sequence-specific and Pol II-associated TFs (Table 1). The ChIP-Seq experiments were performed using 94 cell types. Although ENCODE and Roadmap provided datasets of pre-determined ChIP-Seq peaks (pre-determined TFBSs), there are substantial differences in analytical pipelines between the two projects (Table 2). Therefore, I determined ChIP-Seq peaks using a uniform analytical pipeline (Fig. 3B). In the next generation sequence (NGS) analysis focusing on repetitive sequences such as HERVs, it is necessary to handle carefully multiple mapped reads, which are NGS reads that can be mapped to two or more genomic

regions [38, 67]. If multiple mapped reads are not excluded, false positive peaks may be detected at regions that have sequences similar to those authentically bounded by the TF. If they are excluded, it is unfeasible to identify ChIP-Seq peaks on recently integrated HERVs that show low sequence divergence among the copies. Some studies on TEs excluded multiple mapped reads [37], while others did not [16]. Therefore, I generated two types of ChIP-Seq peak datasets: all-read and unique-read TFBSs (Fig. 3B). All-read TFBSs are ChIP-Seq peaks that were determined with all reads mapped to the human reference genome. The unique-read TFBSs are ChIP-Seq peaks that were determined with only the reads uniquely mapped to the reference genome; in other words, multiple mapped reads were excluded before the peak calling of ChIP-Seq. Consequently, I identified 7,262,985 and 6,833,767 of all- and unique-read TFBSs, respectively (Fig. 4A); for estimating the numbers, overlapped TFBSs of the same TF were merged among cell types.

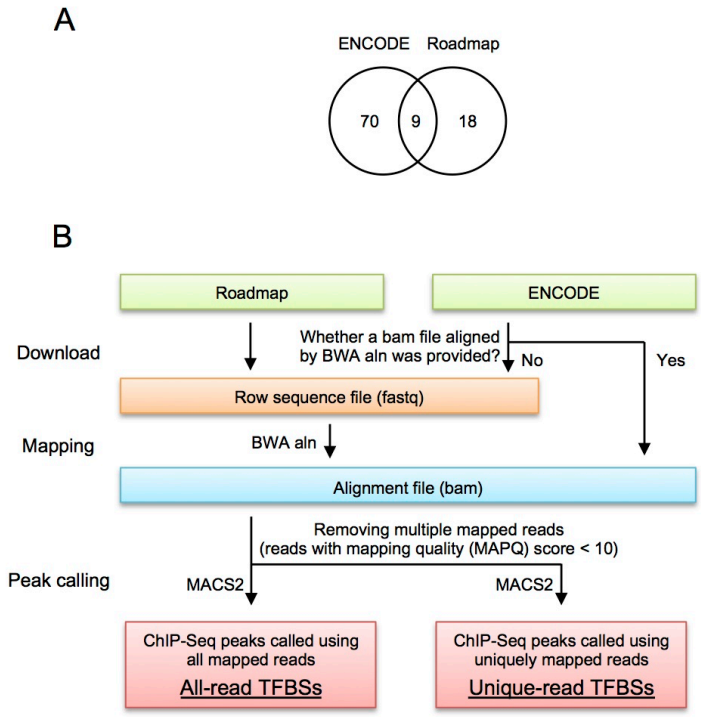


Figure 3. An analytical pipeline for peak calling of ChIP-Seq. A) TFs for which ChIP-Seq was performed in this study. ChIP-Seq data for MYC, CTCF, FOXA1, FOXA2, HNF4A, NANOG, POU5F1, PRDM1, and SP1 were provided by ENCODE and Roadmap. ChIP-Seq data for other TFs were provided by either ENCODE or Roadmap. Detailed information is summarized in Table 1. B) An analytical pipeline for peak calling of ChIP-Seq. I generated two types of TFBS datasets: all- and unique-read TFBSs. All-read TFBSs are ChIP-Seq peaks called with all reads mapped to the reference human genome. Unique-read TFBSs are ChIP-Seq peaks called with only reads that were uniquely mapped to the reference human genome.

Table 2. Sequencing and analytical pipelines of ChIP-Seq used in ENCODE and Roadmap.

	ENCODE (HAIB lab)	**Roadmap**
Read length	25-50 bp	25-36 bp
Layout	Single	Single/Paired
Platform	Genome Analyzer	HiSeq2000
Mapping	Eland/Bowtie	MAQ/Bowtie2
Filtering multiple mapped reads	Yes	No
Peak calling	SPP peak caller with calculation of IDR (Irreproducible Discovery Rate)	MACS
Using input control for peak calling	Yes	No

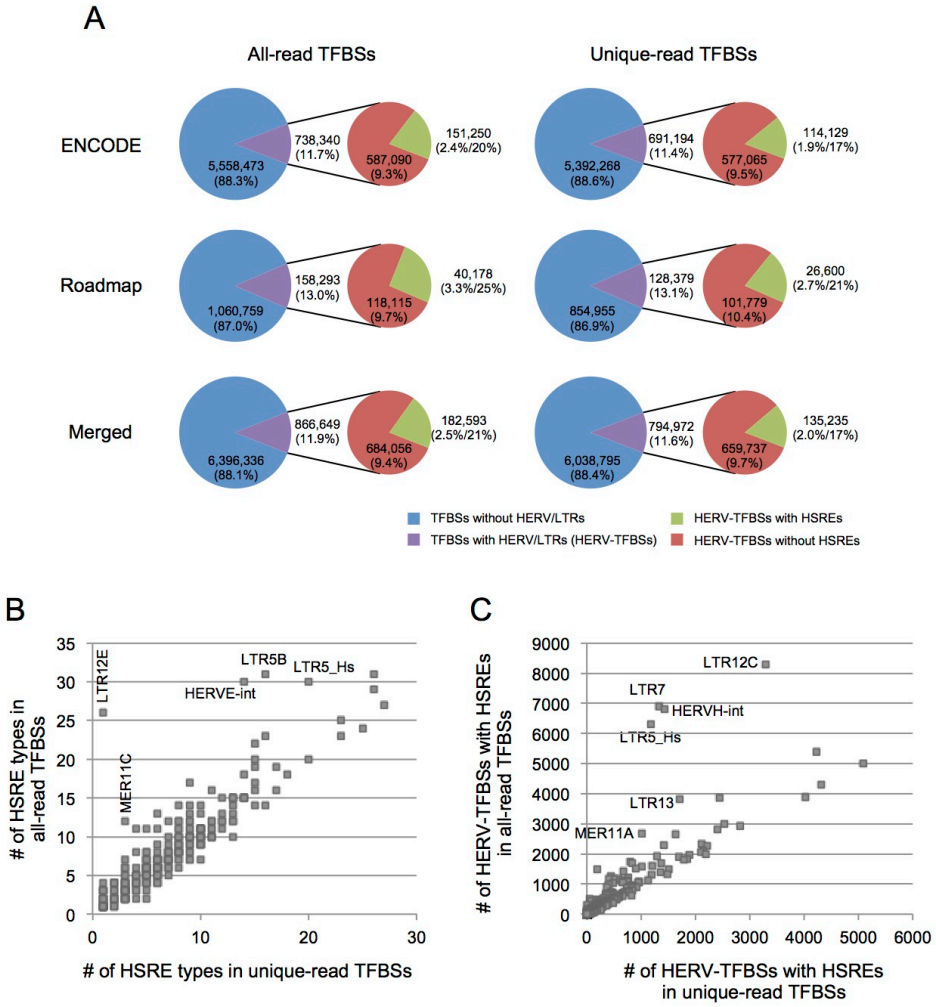


Figure 4. TFBSs, HERV-TFBSs, and HSREs identified from all- and unique-read TFBSs. A) Proportions of HERV-TFBSs and HERV-TFBSs with HSREs. The left and right panels show results of all- and unique-read TFBSs, respectively. Proportions of HERV-TFBSs harboring HSREs in entire TFBSs (left value) and in HERV-TFBSs (right value) are shown. In the "merged" dataset, TFBSs of the same TF were merged between ENCODE and Roadmap, and were then counted. B) Comparison between the numbers of HSRE types identified from all- and unique-read TFBSs. A dot indicates a HERV group. C) Comparison between the numbers of HERV-TFBSs harboring HSREs from all- and unique-read TFBSs. A dot indicates a HERV group.

Detection of TFBSs on HERVs (HERV-TFBSs)

I identified TFBSs observed on HERV sequences (HERV-TFBS overlaps (HERV-TFBSs)) belonging to the all- and unique-read TFBSs (Fig. 5A). I first identified HERV-TFBSs in each cell type, and then merged HERV-TFBSs of the same TF in all cell types (merged HERV-TFBSs). Thus, I identified 866,649 merged HERV-TFBSs from all-read TFBSs and 794,972 from unique-read TFBSs (Fig. 4A). HERV-TFBSs respectively occupied 11.9% and 11.6% of entire TFBSs in all- and unique-read TFBSs (Fig. 4A).

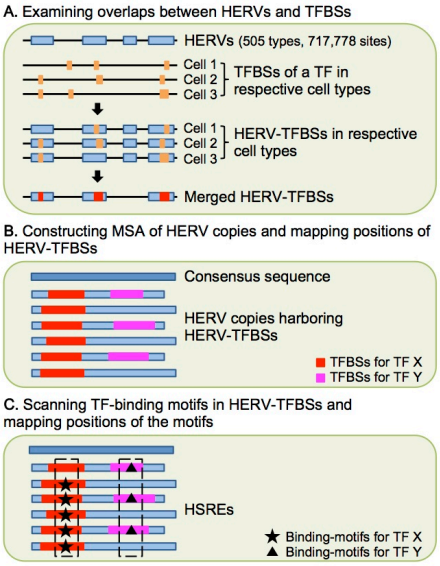


Figure 5. Scheme of identification of HERV-TFBSs and HSREs. HERV-TFBSs and HSREs were identified separately using ENCODE and Roadmap datasets. HERV-TFBSs and HSREs were identified for all- and unique-read TFBSs. A) HERV-TFBSs were identified in respective cell types by examining overlaps between HERVs and TFBSs. HERV-TFBSs of each TF were merged among cell types (merged HERV-TFBSs). B) In each HERV group, MSA of HERV copies was constructed with the consensus sequence, and then the position of the merged HERV-TFBS was mapped on each HERV sequence in the MSA. Red and pink regions indicate HERV-TFBSs for TF X and Y, respectively. C) TF-binding motif was scanned in HERV-TFBS and mapped on each HERV sequence in the MSA. Star and triangle marks indicate TF-binding motifs for TF X and Y, respectively. A set of TF-binding motifs was regarded as HSRE if the TF-binding motifs were shared among greater than 60% of HERV-TFBSs at the same position in MSA. Boxed TF-binding motifs are HSREs for TF X and Y, respectively.

To evaluate the differences between all- and unique-read TFBSs, I compared the number of HERV-TFBSs for both the TFBS datasets. In most HERV groups, the numbers of HERV-TFBSs were approximately the same for all- and unique-read TFBSs; however, the difference was quite large for some HERV groups such as LTR7 and LTR5_Hs (Figs. 6A, 6C, and 6D). These HERV groups were recently inserted [see dbHERV-REs (http://herv-tfbs.com)] and showed low 'genomic mappability' (sequence uniqueness) (Figs. 6B and 6E). Therefore, a substantial number of sequence reads was not uniquely mapped on the HERVs and was discarded. Based on these results, I generally used unique-read TFBSs for further analyses. When I individually focused on HERV groups with low genomic mappability, such as LTR7 and LTR5_Hs, I used all-read TFBSs.

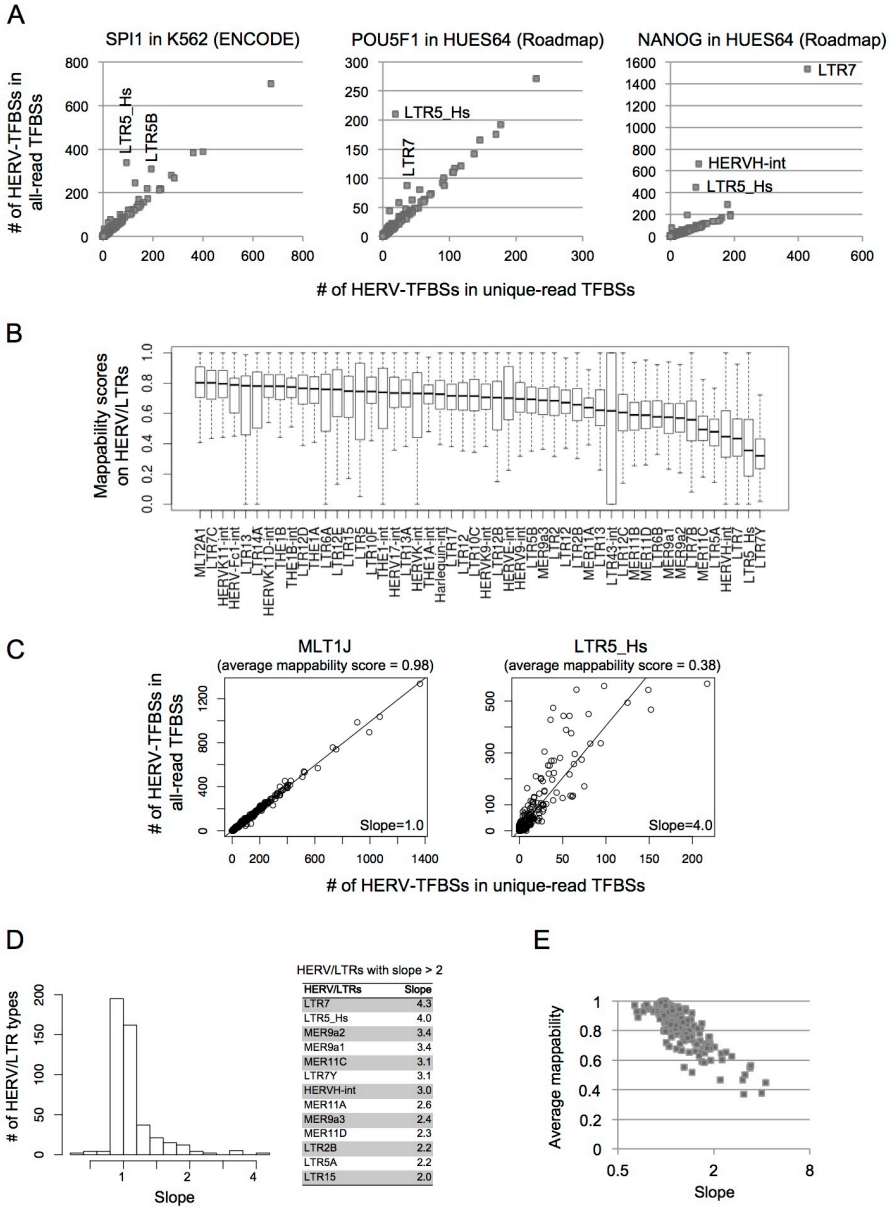


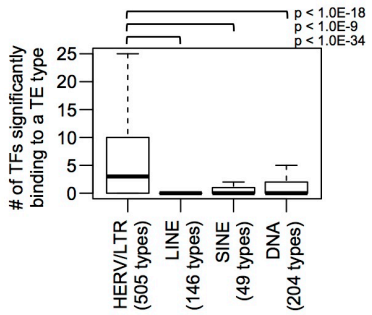
Figure 6. Comparison of all- and unique-read TFBSs. A) Comparison between the numbers of HERV-TFBSs of all- and unique-read TFBSs. The comparison was performed in respective ChIP-Seq experiments, and the results for SPI1 in K562 cells from the ENCODE dataset, POU5F1 in HUES64 cells from the Roadmap dataset, and NANOG in HUES64 cells from the Roadmap dataset are shown. In all the three ChIP-Seq experiments, 36-bp single-end sequencing was performed. A dot indicates a HERV group. In most HERVs, numbers of HERV-TFBSs were approximately the same. However, in some HERVs such as LTR5_Hs and LTR7, numbers of HERV-TFBSs was higher for all-read TFBSs than for unique-read TFBSs. B) Distribution of genomic mappability (uniqueness) scores on HERV sequences. Scores are normalized between 0 and 1, with 1 representing a unique sequence and 0 representing a sequence that occurs more than 4 times in the genome (see http://genome.ucsc.edu/). Mappability score of 36-bp single-end sequencing was calculated with gem-mappability. Average mappability scores of HERV copies were calculated, and the distribution was shown separately in respective HERV groups. With respect to median value of the mappability score, the worst 50 of HERV groups are shown. C) Comparison between the numbers of HERV-TFBSs of all- and unique-read TFBSs. The comparison was performed in respective HERV groups. Results for MLT1J and LTR5_Hs are shown. A dot indicates a ChIP-Seq experiment. Linear regression was performed, and the slope was indicated. In MLT1J with high genomic mappability (average score = 0.98), numbers of HERV-TFBSs in respective ChIP-Seq experiments are approximately the same for all- and unique-read TFBSs (slope = 1.0). In LTR5_Hs with low genomic mappability (average score = 0.38), numbers of HERV-TFBSs in respective ChIP-Seq experiments tended to be approximately four times higher for all-read TFBSs than for unique-read TFBSs (slope = 4.0). D) Distribution of slopes of linear regressions (mentioned in (C)) in respective HERVs. The X-axis is $\log_2$ scale. HERVs with slopes >2 are listed in the right table. E) Association between the slopes and average values of genomic mappability scores. A dot indicates a HERV group. The X-axis is $\log_2$ scale.

I compared HERVs with other classes of TEs with respect to the TF binding profiles. In the unique-read TFBSs, LINE, SINE, and DNA transposons were respectively overlapped to 15%, 16%, and 6% of the entire TFBSs (Fig. 7A). It is important to check whether a TF binds to a group of TE significantly more than expected, because TEs occupy a large fraction of the genome, and therefore, TF binding would be partially observed on the TEs regardless of the absence of a special association between the TEs and TFs. Therefore, I evaluated statistical enrichment of binding of a TF in respective groups of TEs to random expectation. The enrichment of TF binding was measured using a randomization test shuffling genomic positions of TFBSs (see Materials and Methods). Subsequently, I counted the number of TFs bounded significantly to a group of TE, and then the distribution was compared among the TE classes (Fig. 7B). I demonstrated that the number of TFs binding significantly to a TE group tended to be substantially higher in the HERV class than the other TE classes (Fig. 7B). In the other TE classes, a few TEs were bounded by a large number of TFs (Fig. 7C). Thus, HERVs were distinguished from the other TEs with respect to numbers of TF bindings. Previous studies reported the same tendency that HERVs have more regulatory sequences (e.g., DHSs and TFBSs) than the other TEs [37, 38].

A

TE class	Fraction in genome	TFBS with TEs	Proportion
HERV/LTR	8%	794,972	12%
LINE	21%	1,024,907	15%
SINE	13%	1,092,098	16%
DNA	3%	379,614	6%

B



C

HERV/LTR

Type	# of TFs
LTR4	52
LTR35A	42
LTR13	40
LTR13_	38
MER61E	35
LTR21A	34
LTR10A	34
MER61F	33
MER4C	33
LTR22C	32

LINE

Type	# of TFs
Plat_L3	10
X3_LINE	8
HAL1-2a_MD	7
L1MCb	4
X6B_LINE	4
HAL1-3A_ME	4
X9_LINE	3
L1M3f	3
L3b	3
L1MC3	2

SINE

Type	# of TFs
AmnSINE1	55
LFSINE_Vert	19
MamSINE1	15
AmnSINE2	5
MIR3	3
FRAM	2
AluJb	1
AluJo	1
AluJr	1
AluYc5	1

DNA

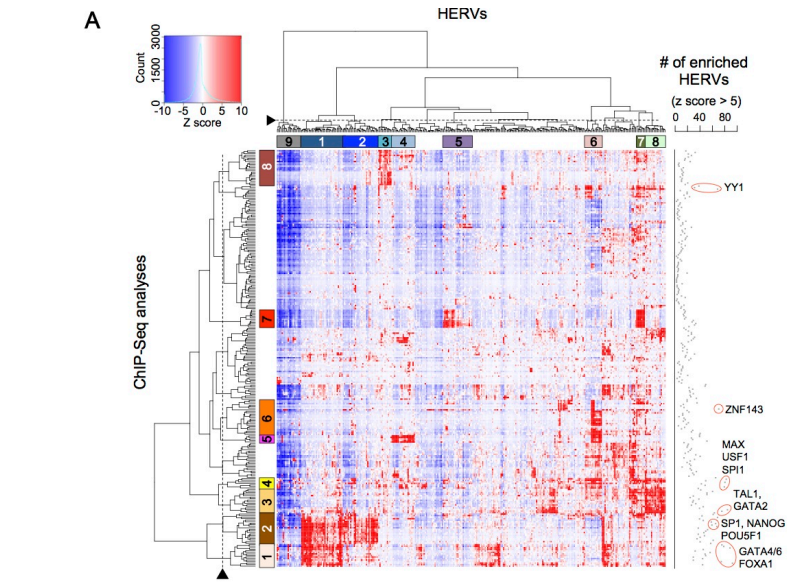
Type	# of TFs
MER91C	31
MER44B	22
MER44C	21
Tigger7	18
MER44D	17
MER91B	17
MamRep434	9
MER126	9
Eulor8	8
MER81	8

Figure 7. Comparison of HERVs with other TE classes with respect to TF binding. Results of unique-read TFBSs are shown. A) Number of TFBSs overlapping with respective TE classes. B) Distribution of the number of TFs significantly binding to respective TE groups. Out of 106 TFs (79 ENCODE TFs + 27 Roadmap TFs), the number of TFs that are significantly bounded to a TE group was counted. The distribution is separately shown in respective TE classes. Outliers of TE groups are not shown. Enrichment significance values were measured using a randomization test shuffling genomic position of TFBSs. TFs with z score >5 and fold enrichment score >2 were considered as significantly binding to the TE group. To statistically compare HERV with other TEs with respect to the numbers of TFs, Mann-Whitney U test was performed. C) TE groups bounded by many TFs.

Classification of HERVs based on TF binding patterns

To understand the characteristic patterns of TF binding to HERVs, I performed hierarchical clustering analysis based on statistical enrichments of TF binding to random expectation (Fig. 8A). Enrichment significance was measured for each combination between HERVs and TFBSs in respective cell types to consider the cell type-specific binding of TFs to HERVs. Fourteen HERV and TFBS clusters were identified (Fig. 8A), of which, I characterized 8 TFBS clusters (TF_1–8) (Fig. 8B) [33, 35, 68-70]: TF_1 contained TFBSs for FOXA1/2, GATA4/6, and SOX17, which are critical for the differentiation of embryonic mesendoderm or endoderm. TF_2 contained TFBSs for POU5F1, SOX2, and NANOG, essential for pluripotency of ES and iPS cells. TF_3 contained TFBSs for GATA1/2 and TAL1, essential in hematopoietic and leukemia cells. TF_4 contained SPI1, which is critical for the differentiation of hematopoietic cells. TF_5 and TF_6 contained TFBSs for NFYA/B, USF1/2, and other TFs expressed in a broad-range of cell types. TF_8 contained TFBSs for PAX5 and PBX3, essential for the differentiation of B lymphocytes. TF_7 contained CTCF-binding sites found in all the cell types, which function as insulators and regulate chromatin architecture. I also characterized 9 HERV clusters (HERV_1–9) (Figs. 8A-C). HERV_1 was enriched in TF_1 (endoderm TF cluster) and TF_2 (pluripotent TF cluster). HERV_2 was enriched in TF_2 (pluripotent TF cluster). HERV_3 was enriched in TF_8 (B-lymphocyte TF cluster). HERV_4 cluster was enriched in TF_5 cluster. HERV_5 and HERV_7 were enriched in TF_7 (CTCF cluster). HERV_6 was enriched in TF_5 and TF_6 clusters. HERV_8 was enriched in TF_3 and TF_4 (hematopoietic TF clusters). Lastly, HERV_9 was not enriched in most TFBSs. Taken together, I identified the characteristic clusters of HERVs by the hierarchical clustering analysis, indicating that HERV groups can be classified based on their TFBSs. Each HERV cluster typically contained several HERV groups belonging to different HERV families (Fig. 8C). This indicates that the pattern of HERV regulatory elements do not

match their phylogenic classifications. TFBSs for FOXA1/2, GATA4/6, NANOG, POU5F1, SP1, GATA2, TAL1, MAX, USF1, SPI1, ZNF143, and YY1 were enriched in various groups of HERVs (Fig. 8A right).



B

Cluster	Dataset	Cell	TFs
TF_1	Roadmap	HUES64	EOMES, GATA4
	Roadmap	Endoderm	EOMES, FOXA1, FOXA2, GATA4, GATA6, OTX2, SOX17
	Roadmap	Mesoderm	GATA4, GATA6, SMAD1
	Roadmap	Mesendoderm	EOMES, GATA6, OTX2
TF_2	ENCODE	H1-hESC	BCL11A, MAX, NANOG, POU5F1, SP1, TCF12, YY1
	Roadmap	HUES64	NANOG, OTX2, POU5F1, SMAD1, SOX2, TCF4
	Roadmap	Endoderm	NANOG, POU5F1, SOX2
	Roadmap	Mesendoderm	NANOG, POU5F1, SMAD1, SOX2
TF_3	ENCODE	K562	BHLHE40, CEBPB, E2F6, FOSL1, FOS, GABPA, GATA1, GATA2, JUNB, JUND, JUN, MYC, NFE2, REST, TAL1
	ENCODE	PBDE	GATA1
TF_4	ENCODE	GM12878	SPI1
	ENCODE	GM12891	SPI1
	ENCODE	K562	ELF1, MAX, SPI1, USF1
TF_5	ENCODE	GM12878	NFYB
	ENCODE	HeLa-S3	NFYA, NFYB
	ENCODE	K562	NFYA, NFYB
TF_6	ENCODE	A549	USF1
	ENCODE	GM12878	ATF3, SIX5, USF1, USF2, ZNF143
	ENCODE	H1-hESC	ATF3, SIX5, USF2, ZNF143
	ENCODE	HUVEC	MAX
	ENCODE	HeLa-S3	USF2, ZNF143
	ENCODE	HepG2	ATF3, USF1, USF2
	ENCODE	K562	ATF3, SIX5, USF2, ZNF143
	ENCODE	NB4	MAX, MYC
	ENCODE	SK-N-SH.RA	USF1
TF_7	Both	All	CTCF
	ENCODE	K562	CTCFL
TF_8	ENCODE	GM12878 and other B cells	BATF, BCL11A, BCL3, CEBPB, EBF1, GABPA, MEF2A, MEF2C, NFKB1, PAX5, PBX3, POU2F2, SP1, TCF12

C

Cluster	Family	HERVs
HERV_1	ERV1	LTR23, LTR24C, LTR29, LTR34, LTR36, LTR39, LTR48, LTR48B, LTR49, LTR54, LTR56, LTR68, LTR6A, MER34A, MER34A1, MER34B, MER34C2, MER4A, MER4A1_, MER4B, MER4C, MER4D, MER4D1, MER4E, MER4E1, MER67B, MER67C, MER67D, MER84, PABL_A, PRIMA4_LTR
	ERVK	LTR5B, LTR5_Hs
	ERVL	LTR40A1
HERV_2	ERV1	HERV4_I-int, HERVIP10FH-int, LTR23-int, LTR26B, LTR26E, LTR37A, LTR7, LTR9, LTR9B, MER34C, MER34C_, MER4-int, MER4A1, MER4B-int, MER4D0, MER50, MER51-int, MER51B
	ERVK	MER11A, MER11B, MER11C
	ERVL	MER21B, MER21C, MER74A, MER74B
	ERVL-MaLR	MSTB-int, MSTB1
	Gypsy	LTR81AB, MamGypLTR2c
HERV_3	ERV1	LTR2B, LTR35B, LTR43, LTR43B, LTR77, MER51D, MER57E1, MER57E3, MER57F
	ERVL	LTR69, MER73
HERV_4	ERV1	LTR12, LTR12B, LTR12D, LTR12E, LTR12F, LTR12_, LTR45B, LTR71A, LTR71B, MER101, MER101B, MER51A, MER51E, MER57A1, MER57B1, MER57B2, MER57C1, MER57C2, MER57D
HERV_5	ERV1	HERVE-int, HERVH48-int, HERVIP10F-int, HUERS-P2-int, HUERS-P3-int, HUERS-P3b-int, LTR1, LTR1B, LTR1D, LTR27, LTR27B, LTR28, LTR43-int, MER52A, MER52C, MER52D, MER61A
	ERVK	HERVK-int
	ERVL	LTR16A, LTR16A1, LTR16A2, LTR16C, LTR16E1, LTR33
HERV_6	ERVL-MaLR	MLT1F, MLT1F1, MLT1G, MLT1G1, MLT1G3, MLT1H, MLT1H2, MLT1I, MLT1J, MLT1J1, MLT1J2, MLT1K, MLT1L, MLT1M
HERV_7	ERV	LTR55
	ERV1	MER66C
	ERVK	LTR13
	ERVL	LTR41, LTR41B, LTR50
	Gypsy	MamGypLTR3
HERV_8	ERV1	LTR19B, LTR19C, LTR4, LTR45C, LTR46, LTR73, LTR76, LTR7C, MER41B, MER41E, MER48, MER61E
	ERVK	LTR13A
	ERVL	LTR53, LTR62, MER54A, MER54B
HERV_9	ERV1	HERVH-int, LTR12C
	ERVL	HERVL-int
	ERVL-MaLR	MLT1A, MLT1A0, MLT1A1, MLT1B, MLT1C, MLT1D, MSTA, MSTA-int, MSTB, MSTD, THE1A, THE1B, THE1B-int, THE1C, THE1C-int, THE1D, THE1D-int

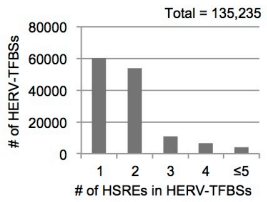
Figure 8. Statistical enrichment of respective TFBSs in each group of HERVs. Results from unique-read TFBSs are shown. A) The heatmap with hierarchical clustering, which shows statistical enrichment of respective TFBSs in each group of HERVs. Color in heatmap (from blue to red) indicates enrichment significance (z score) to random expectation. The row indicates TFBSs from a ChIP-Seq analysis. The column indicates a HERV group. The dendrograms were cut at heights denoted by broken lines. Fourteen clusters were identified for HERVs and TFBSs. Of these, characteristic clusters of TFBSs (TF_1–8) and HERVs (HERV_1–9) are shown. The cut heights and the characteristic clusters were manually chosen according to dendrograms and color patterns in heatmap. The number of HERV groups highly enriched in each TFBS dataset (z score >5) is shown on the right side of the heatmap. B) Characteristic clusters of TFBSs (TF_1–8). Ectoderm, endoderm, mesoderm, and mesendoderm were differentiated from HUES64 cells. C) Characteristic clusters of HERVs (HERV_1–9). Classification of the HERV family is based on RepeatMasker (20-Mar-2009) (http://www.repeatmasker.org/).

Identification of HERV-shared regulatory elements (HSREs)

HERV-shared regulatory element (HSRE) was defined as a TF-binding motif identified in a substantial fraction of HERV-TFBSs at the same consensus position (Fig. 5). HSREs can indicate that the regulatory elements of HERVs are present before their insertions into the respective genomic loci [71]. I identified HSREs according to a scheme shown in Fig. 5. HSREs were identified separately from ENCODE and Roadmap dataset. In total, 2,525 and 2,201 types of HSREs were respectively identified from all- and unique-read TFBSs. Regarding all-read TFBSs, HSREs comprised specific associations of 370 HERVs and 85 TFs. These HSREs were composed of 255,225 genomic loci and present in 21% of the total HERV-TFBSs and in 2.5% of the entire TFBSs (Fig. 4A). For unique-read TFBSs, HSREs comprised specific associations between 354 HERVs and 84 TFs. These HSREs were composed of 178,121 genomic loci and present in 17% of the total HERV-TFBSs and in 2.0% of the entire TFBSs (Fig. 4A). In most HERV groups, the numbers of identified HSREs were approximately the same between unique- and all-read TFBSs; however, in HERV groups with low genomic mappability (e.g., LTR7 and LTR5_Hs), more HSREs were identified from all-read TFBSs than unique-read TFBSs (Figs. 4B and 4C). This was consistent with the comparison of the number of HERV-TFBSs between the two datasets (Fig. 6). Concerning HERV-TFBSs harboring HSREs, approximately half of HERV-TFBSs had more than one of TF-binding motif corresponding to HSRE (Fig. 9A). Most of the HSREs were identified in LTR sequences (87%; 1,935/2,201 combinations in unique-read TFBSs), and the others were identified in the internal sequences of HERVs (13%; 266/2,201 combinations). Large proportions of copies of LTR12, LTR22, LTR13 groups and LTR6B contained HSREs (with respect to proportions of copies harboring HSREs, top 15 of HERVs are shown in Table 3). Regarding TFs, MER41B, LTR13/13A, LTR8/8A, LTR10A/10F, LTR9/9B, and LTR5B/5_Hs contained various HSREs (Fig. 9B). HSREs were identified in both

recently and anciently inserted HERVs, the latter of which was inserted into the genome of the common ancestor of the clade *Eutheria* (Figs. 9C, 9D, and 9E). As degrees of divergences (or 'ages') of HERVs increased, proportions of copies harboring HSREs decreased (Fig. 9D), indicating regulatory elements of ancient HERVs were more divergent than those of young HERVs. As in the case of HERV-TFBSs, HSREs bounded by TFs essential for pluripotent, embryonic endoderm, and hematopoietic cells were frequently identified in addition to CTCF (Figs. 9F and 9G). HSREs bounded by CTCF were frequently observed in internal sequences rather than LTR sequences (Figs. 9H and 9I). Regarding LTR2B, LTR5B, MER41B, and MLT1J, HSREs identified from unique-read TFBSs are shown in Fig. 10.

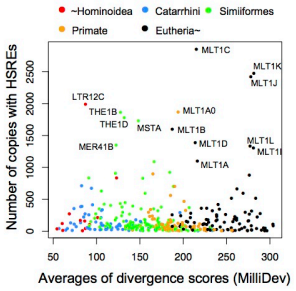
A



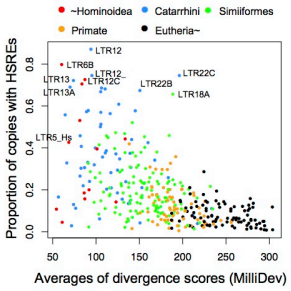
B

Family	Type	# of HSRE types
ERV1	MER41B	27
ERVK	LTR13A	26
ERVK	LTR13	26
ERV1	LTR8	25
ERV1	LTR8A	23
ERV1	LTR10A	23
ERV1	LTR9	20
ERVK	LTR5_Hs	20
ERV1	LTR10F	18
ERV1	MER4C	17
ERV1	LTR9B	17
ERV1	MER48	16
ERVK	LTR5B	16
ERVL	HERVL-int	16
ERV1	MER51A	15

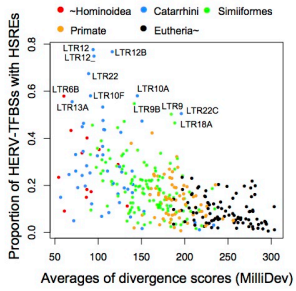
C



D



E



F

TF	# of HERV/LTR copies
USF1	14,671
YY1	9,609
NFYB	8,095
MAX	7,292
SPI1	6,830
JUND	5,236
CTCF	4,631
FOXA1	4,604
OTX2	4,597
NANOG	4,522
GATA4	4,306
USF2	3,186
CEBPB	3,050
MAFK	2,935
GATA2	2,582

G

TF	# of HERV/LTR types
SPI1	97
GATA4	95
USF1	93
FOXA1	88
YY1	76
GATA2	75
NANOG	68
TAL1	66
CTCF	66
GATA6	64
GATA1	62
FOXA2	59
MAFK	58
CEBPB	58
FOS	55

H

TF	# of HERV/LTR types (LTR)
GATA4	89
SPI1	86
USF1	84
GATA2	72
YY1	68
FOXA1	68
TAL1	63
GATA6	59
GATA1	59
NANOG	57
NFYB	49
FOXA2	48
MAX	45
MAFK	45
POU5F1	43

I

TF	# of HERV/LTR types (internal)
CTCF	24
FOXA1	20
CEBPB	15
MAFK	13
FOS	13
SPI1	11
NANOG	11
MAFF	11
FOXA2	11
JUND	10
HNF4A	10
USF1	9
POU5F1	9
YY1	8
OTX2	8

Figure 9. Characteristics of HSREs. Results of unique-read TFBSs are shown. A) Distribution of HSREs present in HERV-TFBSs. The Y-axis indicates the number of HERV-TFBSs containing 1, 2, 3, 4, and greater than or equal to 5 HSREs. B) HERVs that contained many types of HSREs (TFs). C) and D) average divergence of each HERV group from the consensus sequence and absolute numbers (C) or proportions (D) of copies containing HSREs. Color of a dot indicates insertion period of the HERV group judged by distribution of orthologous copies in the mammalian genome. E) average divergence of each HERV group from the consensus sequence and proportions of HERV-TFBSs containing HSREs. Please note the difference in Y-axis between (D) and (E). F) HSREs (TFs) observed in many HERV copies. G) HSREs (TFs) observed in many groups of HERVs. H) and I) HSREs (TFs) observed in many groups of HERVs classified into LTR (H) and internal sequence (I).

Table 3. Absolute numbers and proportions of HERV copies harboring HSREs.

Family	Group	# of copies with HSREs	Proportion
ERV1	LTR12	675	0.87
ERV1	LTR6B	123	0.80
ERVK	LTR22C	294	0.75
ERV1	LTR12_	414	0.75
ERV1	LTR12C	1,993	0.73
ERVK	LTR13	355	0.72
ERVK	LTR13A	130	0.69
ERV1	LTR12D	336	0.69
ERVK	LTR22B	157	0.67
ERV1	MER48	129	0.67
ERVL	LTR18A	170	0.66
ERVK	LTR22A	115	0.61
ERV1	LTR10F	259	0.58
ERV1	LTR10A	181	0.58
ERV1	LTR12B	119	0.56

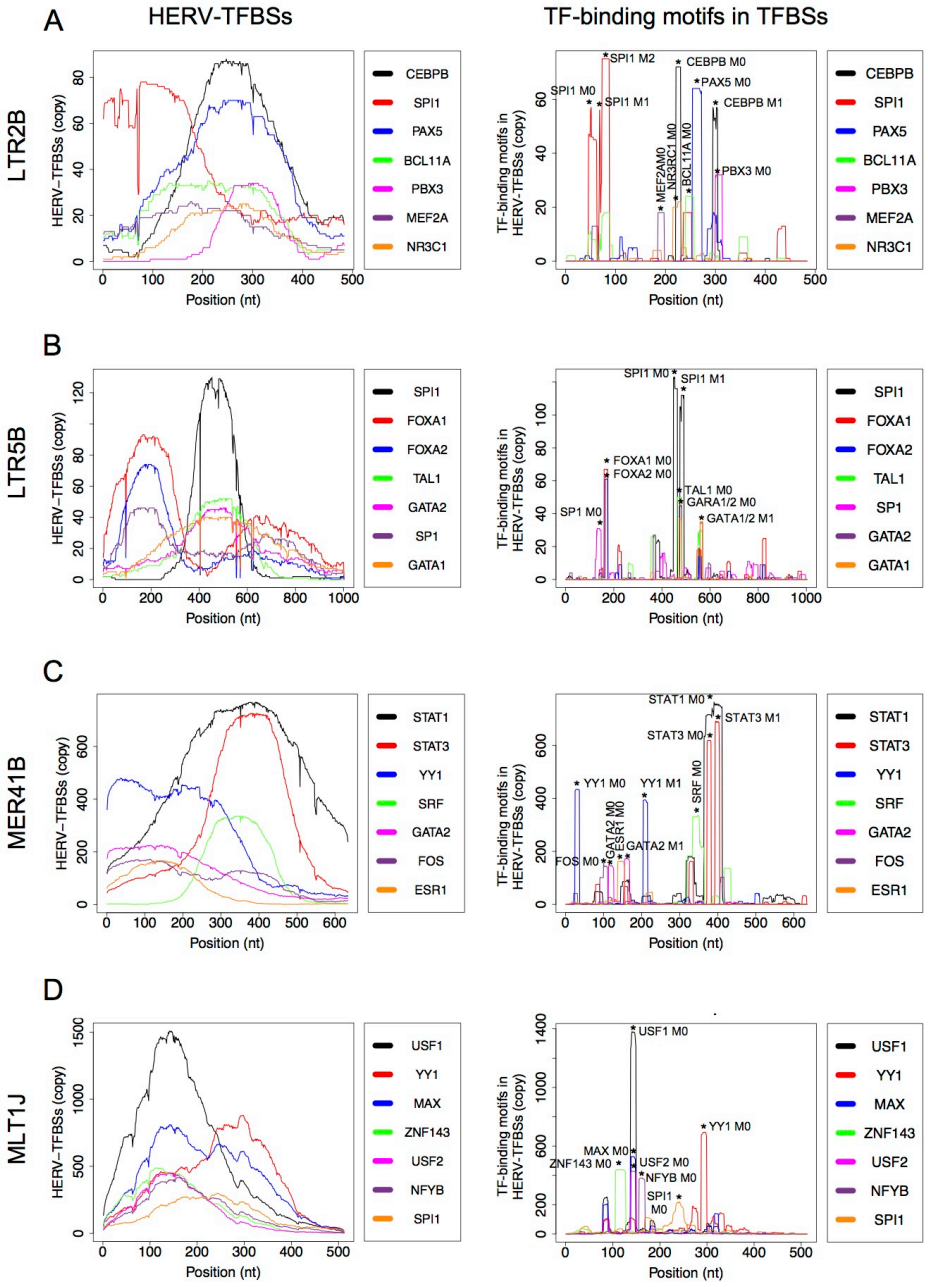


Figure 10. HSREs identified from unique-read TFBSs. Left panel: number of HERV-TFBSs mapped on each consensus position of LTR2B (A), LTR5B (B), MER41B (C), and MLI1J (D). The X-axis indicates the nucleotide position on the consensus sequence of the corresponding HERV group. The Y-axis indicates the number of HERV copies harboring HERV-TFBSs at each position. Right panel: number of TF-binding motifs in HERV-TFBSs mapped on each consensus position of LTR2B (A), LTR5B (B), MER41B (C), and MLI1J (D). The X-axis indicates the nucleotide position of the consensus sequence. The Y-axis indicates the number of HERV copies harboring the TF-binding motifs at each position. Peaks of the motifs corresponding to HSREs are indicated with an asterisk (*) with motif names.

Characteristics of HSREs in LTR7

Characteristics of HSREs in LTR7 identified from the Roadmap dataset are shown in Fig. 11. LTR7 showed low genomic mappability (Fig. 6B), and, therefore, the results of all-read TFBSs were considered (those of unique-read TFBSs are shown in Fig. 12). LTR7 is an LTR sequence of the HERVH provirus belonging to the ERV1 family. In our clustering analysis, LTR7 belonged to the HERV_2 cluster, whose members were highly bounded by SOX2, POU5F1, and NANOG (Fig. 8). These TFBSs were observed at approximately the same consensus positions of LTR7 among those copies (Figs. 11A and 11B). For example, a peak of SOX2 binding was observed at around the 150th nucleotide position on the consensus sequence of LTR7 (Fig. 11B). Splits of HERV-TFBS peaks were observed in NANOG, EOMES, and FOXA1/2 due to an insertion/deletion in multiple sequence alignment of LTR7. TF-binding motifs in HERV-TFBSs were observed at approximately the same consensus position of LTR7 among those copies (Figs. 11C and 11D). I identified HSREs according to the scheme described in Fig. 5 (and Materials and Methods). To identify HSREs, I compared heights of the peaks between HERV-TFBSs and TF-binding motifs (Fig. 13). If the peak of TF-binding motifs (Figs. 11C and 11D) was higher than 60% of that of HERV-TFBSs (Figs. 11A and 11B), the set of TF-binding motifs was regarded as HSRE. I identified novel HSREs in LTR7, such as EOMES, FOXA1/2, and GATA6, and confirmed the previous reports showing that NANOG-, SOX2-, and POU5F1-binding sites were shared across the LTR7 copies [16-18, 42]. Although the HSREs of NANOG, EOMES, FOXA1, and SOX2 were recaptured from unique-read TFBSs, the peaks of HERV-TFBSs in unique-read TFBSs were substantially lower than those in all-read TFBSs (Fig. 12). Chromatin accessibilities evaluated by DHSs and chromatin states [64-66] showed that the regulatory elements of LTR7 were specifically active in ES cells (Figs. 11E and 11F), consistent with the results of previous studies [16-18, 37, 72].

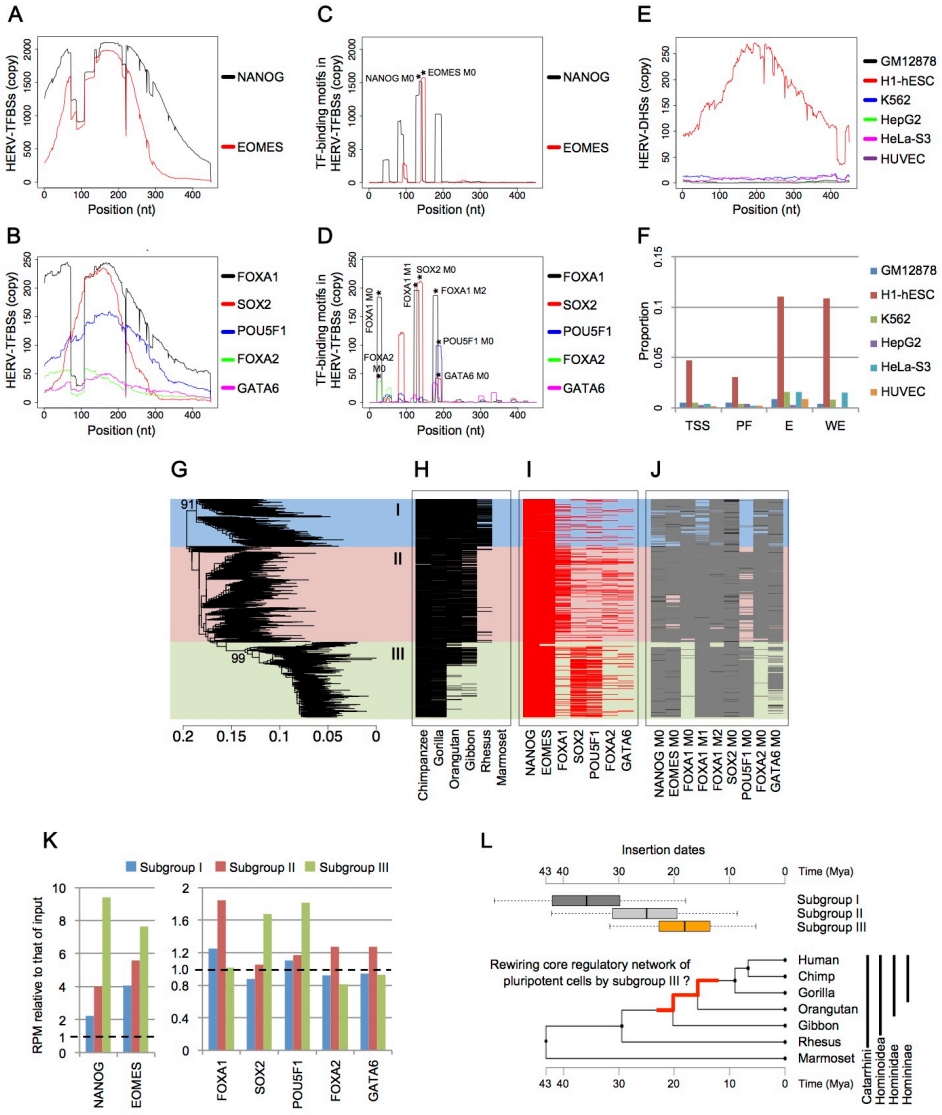


Figure 11. Characteristics of HSREs identified in LTR7 from the Roadmap dataset. Results from all-read TFBSs are shown. A) and B) Number of HERV-TFBSs mapped on each consensus position of LTR7. Results for NANOG and EOMES are shown in (A), and those for FOXA1, SOX2, POU5F1, FOXA2, and GATA6 are sown in (B). The X-axis indicates nucleotide position of the consensus sequence of LTR7. The Y-axis indicates the number of HERV copies harboring HERV-TFBSs at each position. C) and D) Number of TF-binding motifs in HERV-TFBSs mapped on each consensus position of LTR7. Results for NANOG and EOMES are shown in (C), and those for FOXA1, SOX2, POU5F1, FOXA2, and GATA6 are shown in (D). Peaks of the motifs corresponding to HSREs are denoted by an asterisk (*) with motif names (e.g., SOX2 M0). E) The number of HERV-DHSs (DHSs on HERVs) mapped on each consensus position of LTR7. F) Proportion of LTR7 copies overlapped with each chromatin state predicted by genome segmentation method. TSS, promoter region including TSS; PF, predicted promoter flanking region; E, enhancer; WE, weak enhancer or open chromatin cis regulatory element. G) The unrooted phylogenetic tree of LTR7 copies reconstructed using the maximum likelihood method with RAxML. Representative supporting values calculated by Shimodaira-Hasegawa (SH)-like test are shown on the corresponding branches. Identified phylogenetic subgroups (subgroups I, II, and III) are shown. H) Orthologous copies of LTR7 in the reference genomes of primates. The order of LTR7 copies is the same to (G). I) TFBSs on each LTR7 copy. J) TF-binding motifs at positions corresponding to HSREs on each LTR7 copy. Black and gray colors respectively indicate the presences of motifs with p values of <0.0001 and <0.001, identified by FIMO. K) Enrichment of sequence reads mapped to LTR7 copies belonging to respective subgroups. The Y-axis shows reads per million (RPM) relative to that of input control. L) Insertion dates of proviruses of HERVH/LTR7 along with the species tree of primates. Upper panel: The boxplot showing insertion dates of the respective proviruses estimated by sequence comparison between 5′- and 3′-LTRs. Insertion dates of the proviruses are separately shown in the respective subgroups. Categories of subgroups I, II, and III contained 66, 248, and 227 copies of proviruses, respectively. Lower panel: Phylogenetic tree of primates with time scale. The tree was obtained from TIMETREE. Red branch in the tree indicates the period when the rewiring of the core regulatory network of pluripotent cells seems to have occurred.

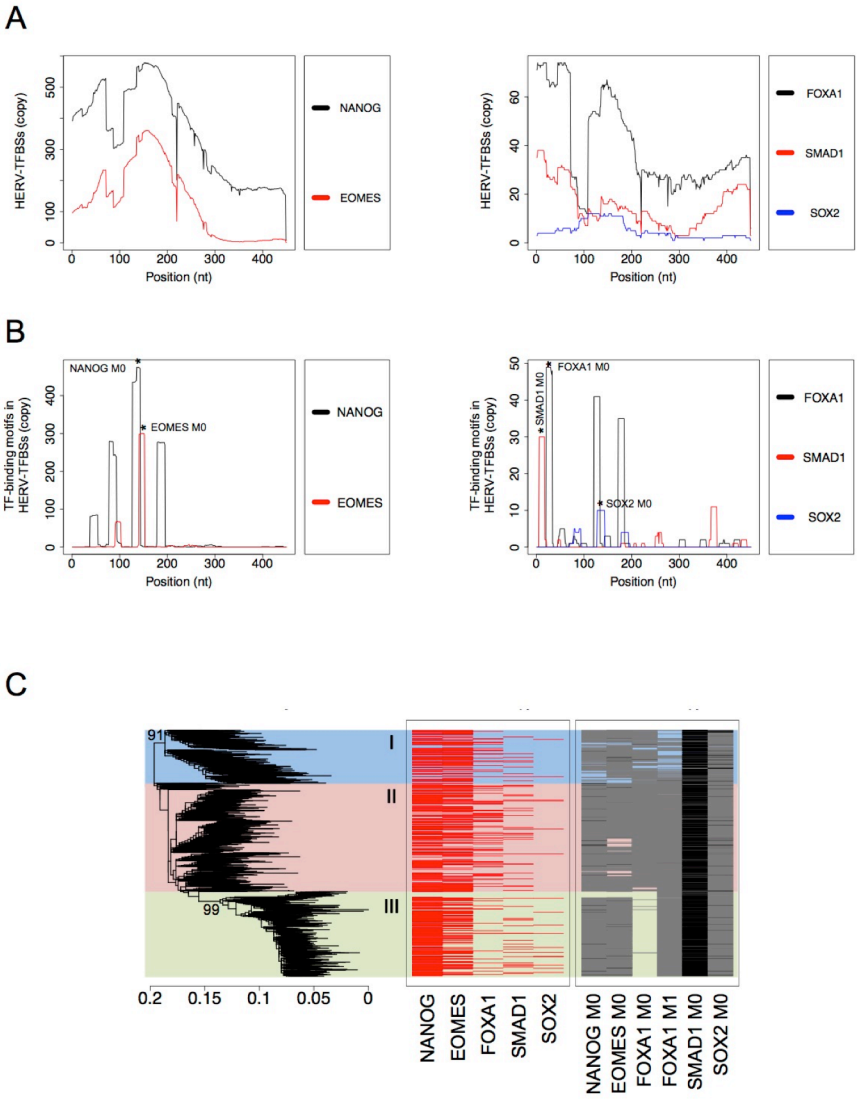


Figure 12. Characteristics of HSREs of LTR7 identified from unique-read TFBSs. A) Number of HERV-TFBSs mapped on each consensus position of LTR7. Results of NANOG and EOMES are shown in the left panel, and those of FOXA1, SMAD1, and SOX2 are shown in the right panel. The X-axis indicates nucleotide position of the consensus sequence of LTR7. The Y-axis indicates the number of HERV copies harboring HERV-TFBSs at each position. B) Number of TF-binding motifs in HERV-TFBSs mapped on each consensus position of LTR7. Results of NANOG and EOMES are shown in the left panel, and those of FOXA1, SMAD1, and SOX2 are shown in the right panel. The X-axis indicates a consensus position of LTR7. The Y-axis indicates the number of HERV copies harboring the TF-binding motifs in TFBSs at each position. Peaks of the motifs corresponding to HSREs are indicated by an asterisk (*) with motif names (e.g., SOX2 M0). C) Left, phylogenetic tree of LTR7 copies as seen in Fig 3G. Middle, TFBSs on each LTR7 copy. The order of LTR7 copies is the same to the left tree. Right, TF-binding motifs at positions corresponding to HSREs on each LTR7 copy. The order of LTR7 copies is the same to the left tree. Black and gray colors respectively indicate the presence of motifs with p values of <0.0001 and <0.001.

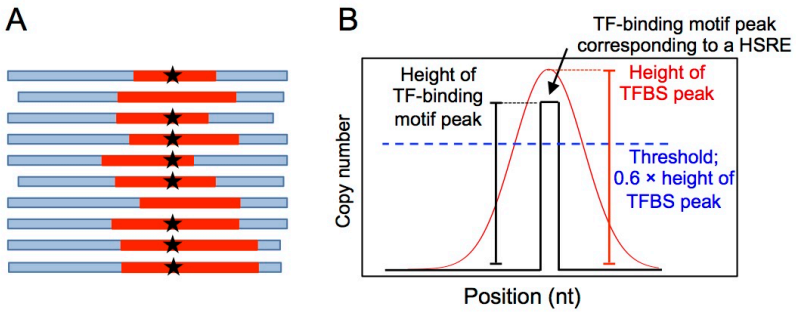


Figure 13. The method to identify HSREs. A) MSA of HERV copies (blue) harboring HERV-TFBSs (red). TF-binding motifs in HERV-TFBSs are indicated as star marks. B) Number of HERV-TFBSs (red) and TF-binding motifs (black) mapped on each consensus position of the HERVs. To identify HSREs, peak heights are compared between HERV-TFBSs and TF-binding motifs. If the height of the TF-binding motif peak is greater than 60% of the height of the HERV-TFBS peak, I regard the set of TF-binding motifs as HSRE.

Heterogeneity of regulatory elements in LTR7

To approach the evolutionary dynamics of HERV regulatory elements, I investigated heterogeneity of the regulatory elements. I focused on HSREs that was disproportionately present in a specific subgroup of a HERV group. LTR7 copies were classified into three main subgroups (subgroups I, II, and III) by phylogenetic analysis based on the sequences (Fig. 11G). Examining orthologous copies of LTR7 in primates indicated that these subgroups were inserted at different time points (Fig. 11H). NANOG- and EOMES-binding sites were uniformly present among the three subgroups (Fig. 11I). SOX2- and POU5F1-binding sites were found to be enriched in subgroup III, and FOXA1-binding sites (and, to a certain extent, FOXA2- and GATA6-binding sites) were enriched in subgroup II (Fig. 11I). I referred to the ChIP-Seq dataset provided by Ohnuki *et al.* [16] (Fig. 14) because this dataset contained ChIP-Seq of SOX2, POU5F1, and KLF4 in iPS cells, and the sequence read lengths (75-bp) were much longer than those of ENCODE/Roadmap dataset (25- or 36-bp). I also referred to the ChIP-Seq data of NANOG in ES cells provided by Durruthy-Durruthy *et al.* [44], performing 100-bp pair-ended sequencing (Fig. 14). Genomic mappability of LTR7 substantially improved in the 75-bp sequencing compared with 36-bp (Fig. 14A). In this dataset, I demonstrated

that binding of SOX2, KLF4, and POU5F1 were enriched in subgroup III (Fig. 14D). In particular, the enrichments were observed in both all- and unique-read TFBSs. POU5F1-binding motifs at positions corresponding to HSREs were enriched in subgroup III, while FOXA1/A2-binding motifs were excluded (Fig. 11J). To quantitatively compare TF binding among the subgroups, I counted the number of reads mapped on LTR7 copies and summed them in respective subgroups, and then I estimated the enrichment of the reads to input control in respective subgroups (Fig. 11K). In NANOG and EOMES, the enrichment was relatively higher in subgroup III although the reads were enriched in all the three subgroups. In SOX2 and POU5F1, the reads were enriched in subgroup III. In FOXA1 (and, to a certain extent, in FOXA2- and GATA6-binding sites), the reads were enriched in subgroup II. Thus, I demonstrated subgroup-specific TF binding in LTR7. In a previous study, LTR7 copies were divided into transcriptionally active and inactive groups based on RNA-Seq using pluripotent cells [17]. I further demonstrated that the active LTR7 copies were enriched in the subgroup III (Fig. 15). Some LTR7 copies fuse with host coding/noncoding genes and play an essential role in maintenance of cell pluripotency [15-18]. I demonstrated that most of the LTR7 copies comprising the chimeric transcripts belonged to the subgroup III (Fig. 15). Finally, I attempted to estimate insertion dates (i.e., ages) of proviruses of HERVH/LTR7 based on sequence comparison between 5′- and 3′-LTRs (see Materials and Methods). As shown in Fig. 11L, majority of the subgroup I, II, and, III seem to have been inserted in branch of the genera *Catarrhini* and *Hominoidea* and the span from the end of *Hominoidea* to the beginning of *Homininae* (interquartile range of insertion dates; 29.7–42.0, 19.4–31.1, and 13.5–22.7 million years ago (Mya), respectively). This is consistent with the insertion dates estimated by presence of orthologous copies in primates (Fig. 11H).

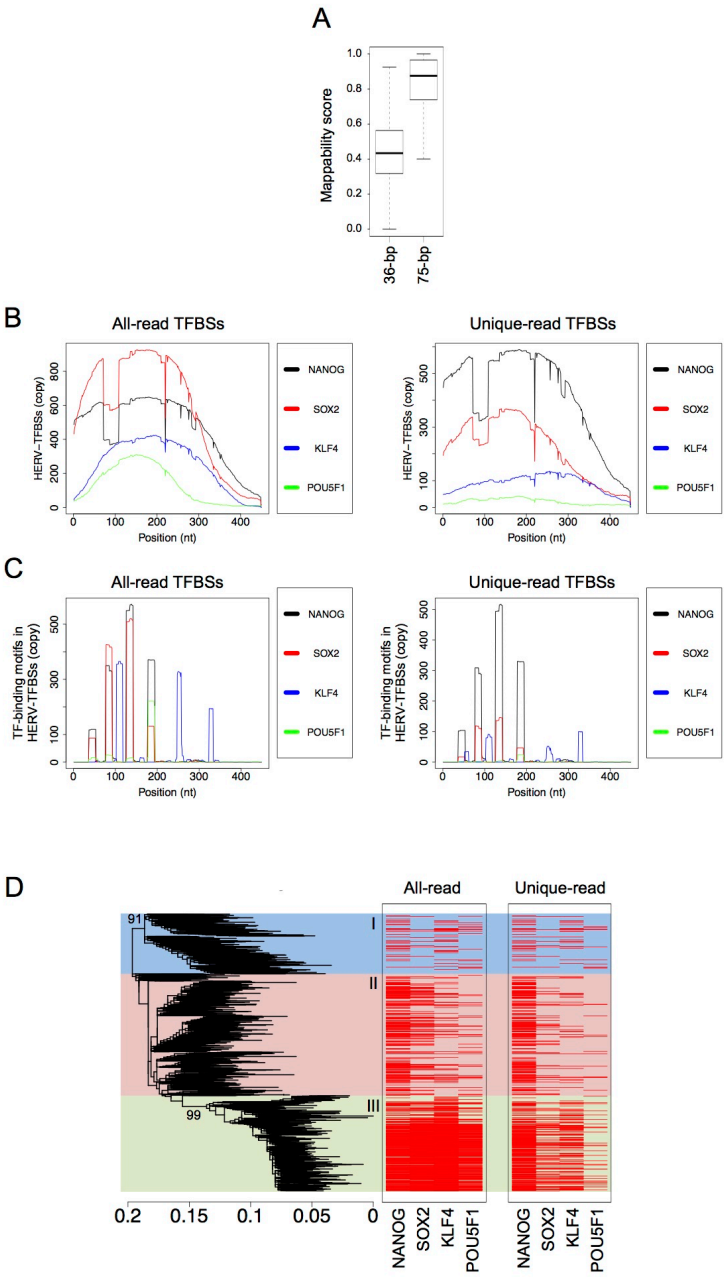


Figure 14. TFBSs of LTR7 identified in ChIP-Seq with 75-bp single-end or 100-bp paired-end sequencing. ChIP-Seq data on SOX2, KLF4, and POU5F1 (75-bp single-end) was provided by Ohnuki *et al.*. ChIP-Seq data on NANOG (100-bp paired-end) was provided by Durruthy-Durruthy *et al.*. A) Comparison between genomic mappability scores of LTR7 for 36-bp and 75-bp sequencing. B) Number of HERV-TFBSs mapped on each consensus position of LTR7. Results of all- and unique-read TFBSs are shown in the left and right panels, respectively. C) Number of TF-binding motifs in HERV-TFBSs mapped on each consensus position of LTR7. Results of all- and unique-read TFBSs are shown in the left and right panel, respectively. D) Left, phylogenetic tree of LTR7 copies as seen in Fig 3G. Middle and right, TFBSs on each LTR7 copy in all-read (middle) and unique-read (right) TFBSs. The order of LTR7 copies is the same to the left tree.

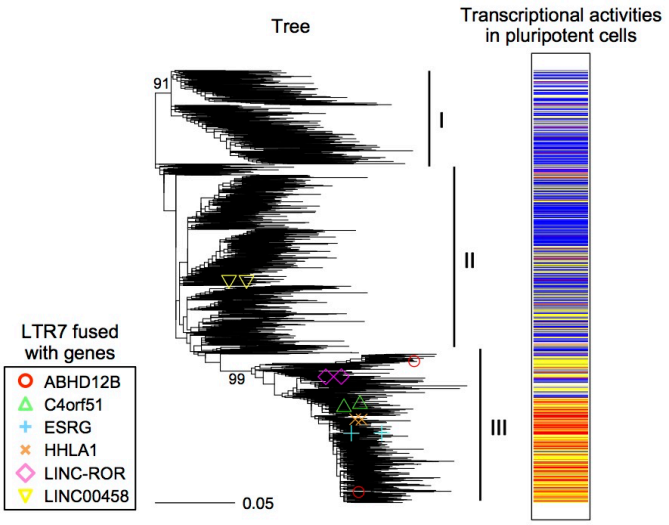


Figure 15. LTR7-chmeric transcripts and transcriptional activities of LTR7. Left, the unrooted tree of LTR7 copies as seen in Fig 3G. LTR7 copies fused with ABHD12B, C4orf51, ESRG, HHLA1, LINC-ROR, and LINC00458 are shown with markers. Right, transcriptional activities of LTR7 copies in pluripotent cells as defined by Wang *et al.*. Red, highly active; yellow, moderately active; blue, inactive.

Changes in regulatory elements during LTR5 evolution

I showed that regulatory elements of HERVs were different within the same HERV group (Figs. 11G-K). In order to approach evolutionary dynamics of regulatory elements in HERVs, I examined changes in the regulatory elements in the LTR5 (HERVK/HML-2) group. LTR5 is composed of LTR5A, LTR5B, and LTR5_Hs. LTR5_Hs is the youngest HERV group, and a previous study reported that LTR5_Hs has regulatory elements for POU5F1, SOX2, and NANOG [50]. Also consistent with the results of a previous study [73], phylogenetic analysis and examination of orthologous copies indicated that LTR5B was the oldest ancestral group, and LTR5A and LTR5_Hs were independently generated from LTR5B-like viruses (Figs. 16A and 16B). Here, I divided LTR5 into five groups (groups I–V) based on their phylogenetic relationship and the TFs binding to them (Figs. 16A, 16C, and 16E). Group I was rarely bounded by TFs (Figs. 16C and 16E). Group II was bounded by SPI1, TAL1, and GATA1/2, which are vital in hematopoietic cells. Group III was bounded by GATA4/6, SOX17, and FOXA1/2, essential in embryonic endoderm cells, together with the

hematopoietic TFs. Group IV was bounded by NANOG, MYC, POU5F1, and SOX2, which are critical in pluripotent cells, in addition to the hematopoietic and the endoderm TFs. In group V, which is the youngest group, binding levels of some hematopoietic TFs (SPI1 and GATA1/2) and endoderm TFs (GATA4/6 and SOX17) were low. These differences in TF binding correlated with the differences in TF-binding motifs at positions corresponding to the HSREs (Fig. 16D). Chromatin accessibilities evaluated by DHSs indicate that the cell specificity of LTR5 members shifted along with their gain/loss of TFBSs (Fig. 16F). Group I was not active in any cell types, as expected owing to the absence of the regulatory elements. Group II was active in K562 (leukemia) cells. Group III was active in HepG2 (hepatoblastoma) and A549 (lung epithelial cancer) cells, in addition to K562 cells. Group IV was active in H1-hESC (ES) cells, in addition to the above cells; group V was not active in K562 cells.

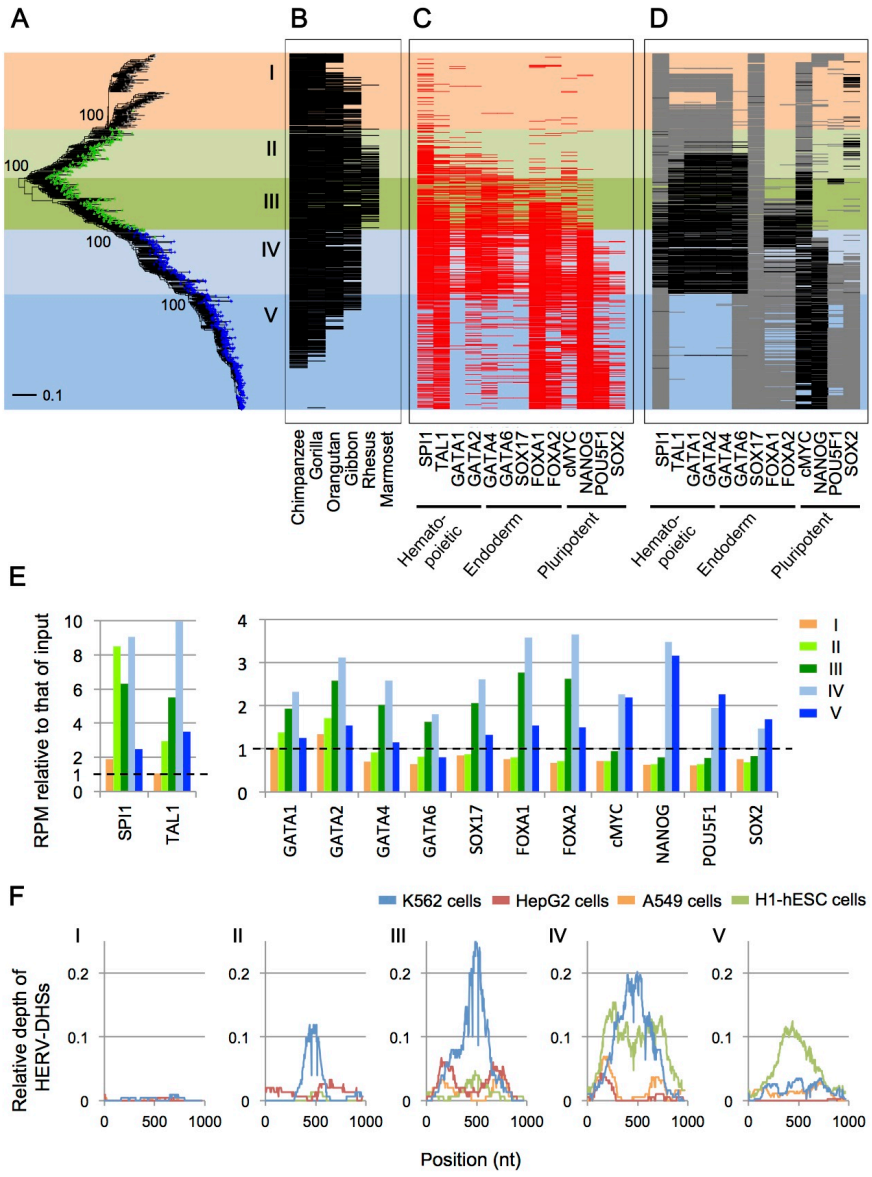


Figure 16. Changes in regulatory elements in LTR5 group. Results from all-read TFBSs are shown. A) The unrooted phylogenetic tree of LTR5A (red), LTR5B (green), and LTR5_Hs (blue) copies constructed using the maximum likelihood method. LTR5 was divided into five groups (I–V) based on the tree and their TFBSs (shown in (C)). Fragmented and outlier copies were excluded from the analysis. Copies of 233, 300, and 532 respectively belonging to LTR5A, LTR5B, and LTR5_Hs were included in the tree (out of 265, 431, and 645, respectively). Representative bootstrap values are shown at the corresponding nodes. B) Orthologous copies in the reference genomes of primates. The order of LTR5 copies is the same to (A). C) TFBSs present on each copy; representative TFBSs are shown. TFBSs of SPI1, TAL1, and GATA1/2 were from the ENCODE dataset, and others were from the Roadmap dataset. The order of LTR5 copies is the same to (A). D) TF-binding motifs at positions corresponding to HSREs on each LTR5 copy. The order of LTR5 copies is the same to (A). Black and gray colors respectively indicate the presence of motifs with p values of <0.0001 and <0.001, as identified by FIMO. E) Enrichment of sequence reads mapped to LTR5 copies belonging to respective subgroups. The Y-axis shows RPM relative to that of the input control. F) Relative number of HERV-DHSs mapped on each consensus position. The X-axis indicates nucleotide position in the consensus sequence of LTR5_Hs. The Y-axis indicates proportion of HERV copies harboring HERV-DHSs at each position.

Signatures of the HERV regulatory elements

I examined chromatin states [64-66] of HERVs with and without TFBSs/HSREs. Compared with the entire population of HERVs, HERVs harboring HERV-TFBSs or HSREs were enriched in promoter [transcription start site (TSS) and promoter flanking regions (PF)], enhancer (E), weak enhancer (WE), and CTCF-binding regions (CTCF), but not in transcribed (T) and repressed (R) regions (Fig. 17A). The HERV groups enriched in enhancer regions were different across different cell types (Fig. 17B). These differences seem to reflect the differences of their HSREs; LTR2B [37], LTR7 [17, 37, 72], MER41B, and LTR5B, which were respectively enriched in the enhancer regions of GM12878, H1-hESC, K562/HeLa-S3, and HepG2 cells, had HSREs bounded by TFs essential in the corresponding cell types (Figs. 10A, 11A-B, 10C, 10B, respectively). Unlike enhancers, HERVs enriched in CTCF-binding regions remained unchanged among the cell types (Fig 17C), which is consistent with previous findings [35].

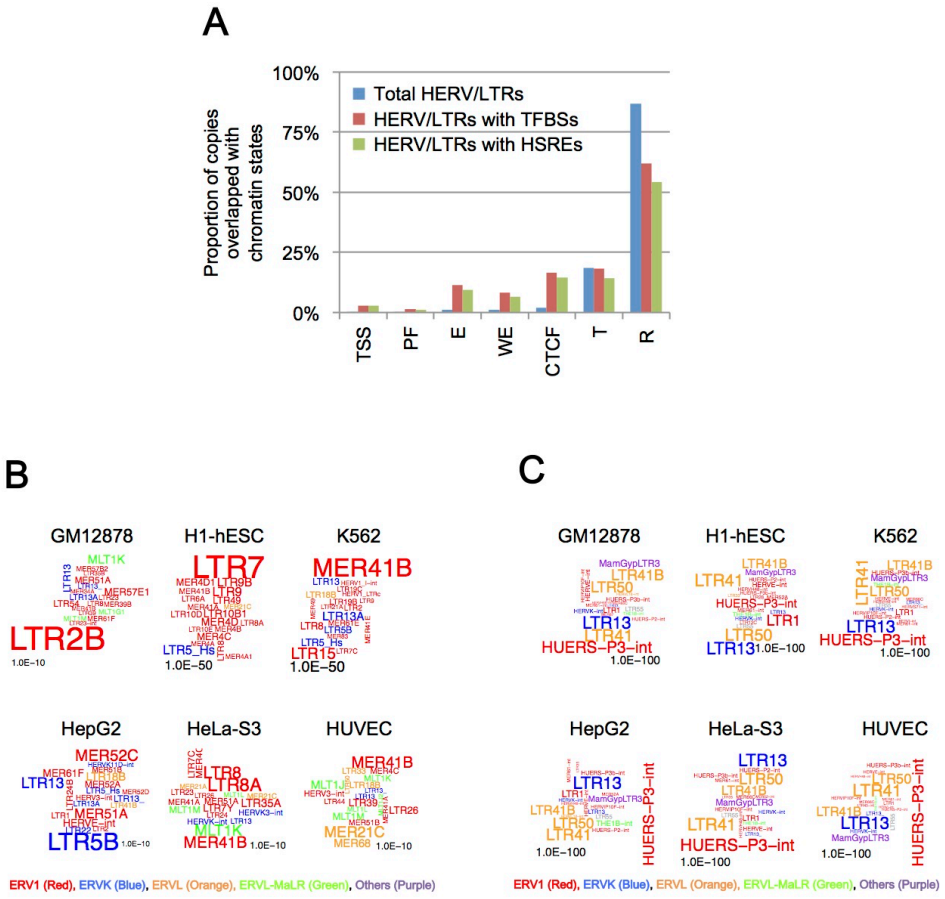


Figure 17. HERVs enriched in regions with various chromatin signatures. A) Proportion of HERV copies overlapped with each chromatin state. Chromatin states were predicted by genome segmentation method. Proportions in total HERVs, HERVs with HERV-TFBSs, and HERVs with HSREs are separately shown. Results of unique-read TFBSs are shown. Averages of the proportions among six cells (GM12878, H1-hESC, K562, HepG2, HeLa-S3, and HUVEC) are shown. TSS, promoter region including TSS; PF, predicted promoter flanking region; E, enhancer; WE, weak enhancer or open chromatin cis regulatory element; CTCF, CTCF enriched element; T, transcribed region; R, repressed or low activity region. B) Word clouds showing HERVs enriched in enhancer regions of each cell type. The word sizes are proportional to $-\log_{10}$ (p values) calculated with Fisher's exact test. The word colors indicate HERV families. Word clouds were created by wordcloud package implemented in R. C) Word clouds showing HERV groups enriched in CTCF-binding regions of each cell type.

I examined TFs in which large fractions of TFBSs were occupied by HERV-TFBSs (Table 4). Binding sites of NFYA/B, USF1/2, GATA4/6, TAL1, SOX2, SOX17, and TCF4 were highly overlapped with HERVs. Nearly half of NFYB-binding sites were observed on HERVs [74]. NFYA/B frequently bound to members of the HERV_4 cluster in Fig. 8 (e.g., LTR12, MER51, and MER57 groups) and members of the HERV_6 cluster (MLT1 group) (Fig. 8). These HERVs contained HSREs for

NFYA/B [see dbHERV-REs (http://herv-tfbs.com/)].

Table 4. Proportions of HERV-TFBSs in the entire TFBSs in respective TFs.

TF	TFBSs	HERV-TFBSs	Proportion
NFYB	20,930	9,805	46.8%
NFYA	8,786	1,999	22.8%
GATA6	32,230	7,073	21.9%
USF1	138,147	29,524	21.4%
GATA4	81,738	16,359	20.0%
TAL1	35,321	6,951	19.7%
SOX2	9,018	1,710	19.0%
SOX17	13,348	2,470	18.5%
USF2	33,600	6,205	18.5%
TCF4	12,764	2,209	17.3%
EOMES	32,963	5,662	17.2%
STAT1	21,727	3,515	16.2%
YY1	193,183	30,506	15.8%
GATA1	49,133	7,714	15.7%
OTX2	126,138	19,576	15.5%
MAX	268,057	41,571	15.5%
SPI1	131,487	19,731	15.0%
ZNF143	72,347	10,813	14.9%
GATA2	112,602	16,309	14.5%
SRF	32,776	4,623	14.1%
NANOG	102,008	13,903	13.6%
NFE2	56,155	7,528	13.4%
JUNB	31,088	4,086	13.1%
JUND	220,440	28,583	13.0%
STAT3	114,240	14,509	12.7%

Then, I investigated specific associations between the insertion dates of HERVs and TFs that bound to the HERVs (Fig. 18). HERVs integrated after the

divergence of primates were highly bounded by members of TF_2 (pluripotent cluster) shown in Fig. 8, such as POU5F1, SOX2, SMAD1, TCF4, and NANOG (Figs. 18 and 8). This is consistent with the results of a previous study showing that SOX2- and POU5F1-binding sites were amplified after the divergence of primates by insertions of HERVs harboring the binding sites [42]. HERVs integrated before the divergence of primates were highly bounded by members of the TF_6 cluster, such as SIX5, USF1/2, and ATF3 (Figs. 18 and 8). This is because these TFs frequently bound to the MLT1 group (Fig. 8), which inserted before the divergence of primates. HERVs that inserted at the span from *Catarrhini* to *Hominoidea* were highly bounded by NFYA/B and LEF1 (Fig. 18). This is because these TFs bound to the LTR12 group, which inserted at the span from *Catarrhini* to *Hominoidea* [see dbHERV-REs (http://herv-tfbs.com/)].

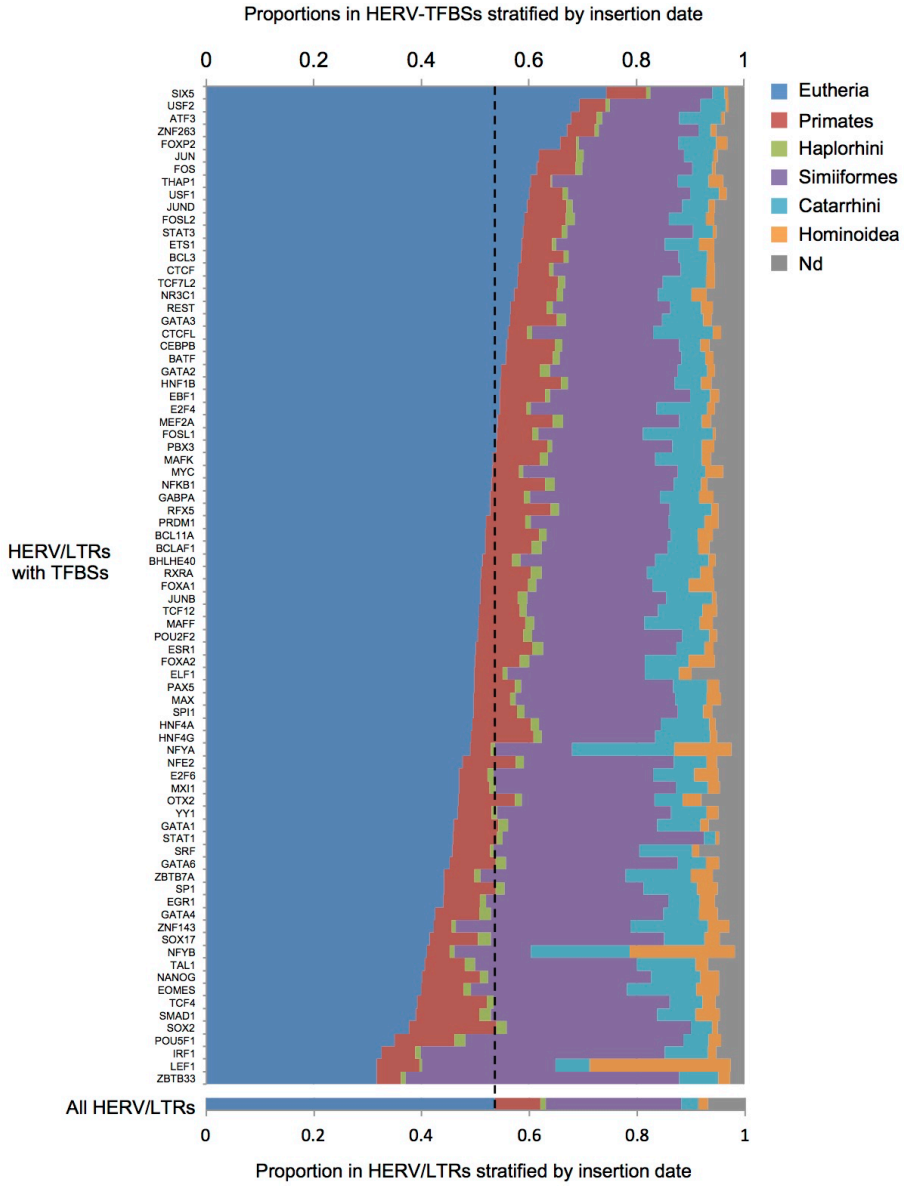


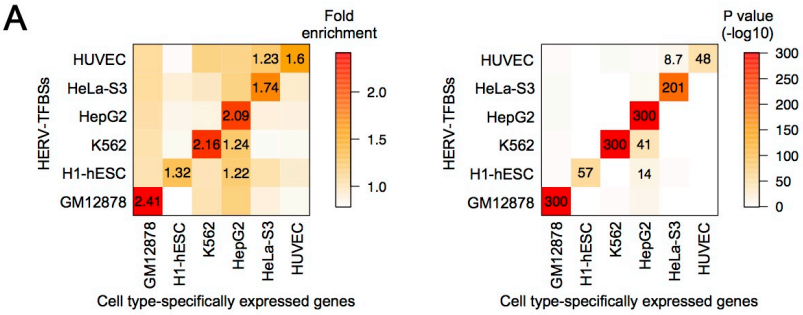
Figure 18. Proportions in HERV-TFBSs stratified by insertion date. Results of unique-read TFBSs are shown. In respective TFs, HERVs with TFBSs were stratified by insertion date. TFs in which HERV-TFBSs overlapped with HERVs at least 1,000 times are shown. The integration date of HERV groups was judged by distribution of orthologous of HERVs among the mammalian genome (see Materials and Methods). Proportions in all HERVs are shown at bottom of the figure.

Characteristics of host genes in the vicinity of HERV regulatory elements

It is important to clarify whether HERV-TFBSs contribute to the regulation of host genes, especially in a cell type-specific manner. I examined the association between

HERV-TFBSs and genes specifically expressed in a particular cell type. In six cell types (GM12878, H1-hESC, K562, HepG2, HeLa-S3, and HUVEC cells), I identified 200 genes that specifically expressed in each cell type. Subsequently, I examined the enrichment of HERV-TFBSs according to the cell types in regions nearby the genes that were specifically expressed. I demonstrated that HERV-TFBSs in each cell type were enriched in region nearby the specifically expressed genes in the corresponding cell type (Fig. 19A). This suggests that HERV-TFBSs are involved in cell type-specific regulation of host genes.

A



B

Cell	GO term (biological process)	P value (-log10)	Fold enrichment
GM12878	response to interferon-gamma	300.0	2.3
	negative regulation of transposition	141.3	15.5
	negative regulation of multi-organism process	131.5	2.2
	cytidine deamination	131.0	8.2
	positive regulation of cytokine biosynthetic process	130.0	2.2
H1-hESC	quinolinate catabolic process	67.9	23.2
	negative regulation of NF-kappaB import into nucleus	47.2	2.3
	regulation of cellular ketone metabolic process by negative regulation of transcription from RNA polymerase II promoter	42.1	4.2
	cholesterol catabolic process	42.0	3.1
	cerebellar molecular layer development	39.8	5.1
K562	regulation of triglyceride metabolic process	300.0	3.2
	type I interferon signaling pathway	284.7	2.5
	androgen metabolic process	280.3	2.5
	steroid catabolic process	278.9	3.4
	regulation of lipoprotein particle clearance	262.0	4.7
HepG2	C21-steroid hormone metabolic process	282.4	4.2
	type I interferon signaling pathway	159.9	2.4
	cellular ketone metabolic process	139.6	2.1
	glucuronate metabolic process	138.1	3.7
	arachidonic acid metabolic process	123.8	2.1
HeLa-S3	cholesterol catabolic process	133.9	6.4
	lipopolysaccharide transport	99.9	10.5
	type I interferon signaling pathway	85.9	2.6
	regulation of hepatocyte growth factor receptor signaling pathway	80.6	7.7
	regulation of phosphatidylcholine catabolic process	72.1	8.1
HUVEC	apoptotic cell clearance	23.3	2.9
	regulation of chemokine secretion	21.7	5.0
	lipopolysaccharide transport	21.2	7.8
	cholesterol catabolic process	18.1	4.0

Figure 19. Characteristics of genes in the vicinity of HERV-TFBSs. Results from unique-read TFBSs are shown. A) Enrichment of HERV-TFBSs as seen in regions near cell type-specific genes. In respective cell types, 200 of the specifically expressed genes according to the cell type were identified. Then I measured enrichments of HERV-TFBSs of respective cell types in regions near the cell type-specific genes using the GREAT. Fold enrichment scores (left) and p values (right) are shown as heatmaps. Fold enrichment scores of >1.2 are shown with the corresponding p values. B) Distance-based GO enrichment analysis. GO terms in the category of biological process were examined. The GREAT analyses were performed using sets of all HERV-TFBSs in respective cell types. HERV-TFBSs identified in cells treated with special conditions (e.g., supplement of interferon) were excluded. GO terms were summarized by REVIGO. GO terms with hold enrichment scores of >2 are shown.

To ascertain which biological functions are associated with HERV-TFBSs/HSREs, I performed Gene Ontology (GO) enrichment analysis with GREAT [75]. First, I performed the analysis using a set of all HERV-TFBSs in one cell type (Fig. 19B). HERV-TFBSs in cells such as GM12878 and K562 were highly enriched in regions nearby the genes associated with innate immunity-related pathways such as “response to interferon-gamma” and “type I interferon signal pathway” (Fig.

19B). The MER41 and MLT1 groups occupied significant fractions of HERV-TFBSs nearby the genes associated with the above biological processes (Fig. 20; left panel). Regarding TFBSs, binding sites of SPI1, POU2F2, ZNF263, and USF1 were found to be enriched (Fig. 20 right panel). Next, I ascertained biological processes in GO term with which HERV-TFBSs were more enriched compared to the other TFBSs (i.e., TFBSs did not overlap with HERVs). HERV-TFBSs showed significantly stronger associations with biological processes relevant to immune responses compared to the other TFBSs (Table 5). I also performed GO enrichment analysis to examine biological functions in which HERV-TFBSs were enriched compared to the entire population of HERVs, and I obtained similar results (Table 6). Finally, I performed the GO enrichment analyses to infer biological functions with which each type of HSRE is associated. In this analysis, I used sets of HERV-TFBSs harboring each type of HSRE in respective cell types. In total, 39,946 significant associations for combinations of cell types, HSREs, and GO terms were identified [summary data is deposited in dbHERV-REs (http://herv-tfbs.com/)]. Consistent with the above analyses, GO terms associated with the immune response were frequently observed (Table 7), and the associations between HSREs and various biological processes were identified [see dbHERV-REs (http://herv-tfbs.com/)].

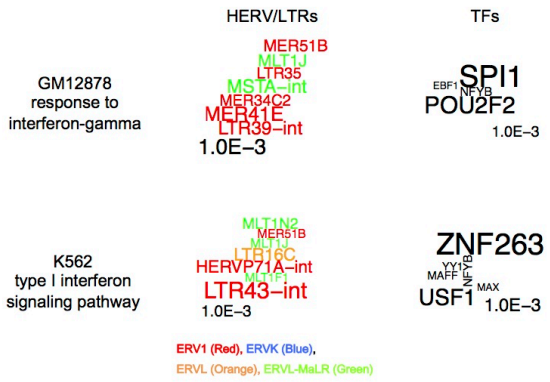


Figure 20. HERVs (left) and TFs (right) occupying significantly large fractions in HERV-TFBSs associated with interferon-related biological processes. Regarding biological processes identified in Fig 5B, enrichment significance values of HERVs and TFs are shown. The word sizes are proportional to $-\log_{10}$ (p value) calculated with Fisher's exact test. The word colors indicate HERV families.

Table 5. Distance-based GO enrichment analysis to ascertain biological processes in which HERV-TFBSs were more enriched compared to the other TFBSs.

Cell	GO term (biological process)	P value (-log10)	Fold enrichment
GM12878	negative regulation of viral process	62.4	4.5
	cytidine deamination	51.5	3.5
	positive regulation of type 2 immune response	43.6	2.6
	glutamate receptor signaling pathway	42.9	2.1
	DNA cytosine deamination	42.2	3.9
H1-hESC	heparan sulfate proteoglycan metabolic process	32.3	2.2
	negative regulation of viral process	27.9	3.1
	serotonin receptor signaling pathway	27.9	2.0
	presynaptic membrane assembly	24.4	2.0
	positive regulation of fever generation	23.8	2.7
K562	cholesterol catabolic process	100.5	2.4
	cellular response to estrogen stimulus	73.2	2.1

	heparan sulfate proteoglycan metabolic process	51.7	2.3
	negative regulation of triglyceride catabolic process	41.7	2.1
	positive regulation of type 2 immune response	41.2	2.5
HepG2	flavonoid biosynthetic process	61.4	2.9
	cellular glucuronidation	46.7	2.3
	thyroid hormone metabolic process	39.7	2.3
	doxorubicin metabolic process	39.5	2.3
	cellular response to prostaglandin D stimulus	39.2	3.3
HeLa-S3	cholesterol catabolic process	44.9	2.8
	positive regulation of chemokine secretion	22.9	2.6
	stabilization of membrane potential	20.5	2.1
	androgen biosynthetic process	19.0	2.1
	opioid receptor signaling pathway	18.2	3.0
HUVEC	flavonoid biosynthetic process	8.2	3.6
	neuroligin clustering involved in postsynaptic membrane assembly	8.1	2.1
	oligosaccharide biosynthetic process	7.8	2.0
	cholesterol catabolic process	7.1	2.3
	protection from natural killer cell mediated cytotoxicity	7.0	3.4

Table 6. Distance-based GO enrichment analysis to ascertain biological processes in which HERV/LTRs harboring TFBSs were more enriched compared to entire HERV/LTRs.

Cell	GO term (biological process)	P value (-log10)	Fold enrichment
GM12878	type I interferon signaling pathway	55.0	2.6
	liver regeneration	28.9	2.5
	blood vessel endothelial cell migration	18.1	2.7
	negative regulation of viral genome replication	16.7	2.1
	entrainment of circadian clock by photoperiod	16.5	2.3
H1-hESC	response to purine-containing compound	8.4	2.1
	positive regulation of keratinocyte differentiation	8.2	2.2
	positive regulation of meiotic nuclear division	8.1	2.0
	negative regulation of protein dephosphorylation	6.4	2.2
	high-density lipoprotein particle clearance	6.3	2.7
K562	platelet aggregation	37.5	2.0
	hepatocyte apoptotic process	31.6	2.5
	liver regeneration	30.9	2.1
	cholesterol catabolic process	26.3	2.5
	regulation of cytokine production	26.1	2.2
HepG2	cellular response to estrogen stimulus	15.2	2.0
	platelet-derived growth factor receptor-beta signaling pathway	13.8	2.8
	cholesterol catabolic process	13.4	2.1
	L-serine transport	12.7	2.3
	cellular response to nutrient levels	10.9	2.3
HeLa-S3	cholesterol catabolic process	17.7	2.9
	embryonic placenta development	16.0	2.3
	liver regeneration	15.4	2.1
	regulation of interferon-gamma-mediated signaling pathway	14.5	2.3

	positive regulation of transcription from RNA polymerase II promoter in response to endoplasmic reticulum stress	13.9	3.5
HUVEC	programmed necrotic cell death	11.1	3.0
	positive regulation of nitric-oxide synthase activity	8.1	2.2
	protection from natural killer cell mediated cytotoxicity	7.2	3.9
	high-density lipoprotein particle clearance	6.7	4.3
	androgen biosynthetic process	6.6	2.6

Table 7. Biological processes in which many types of HSREs were enriched.

Cell	GO term (biological process)	# of HSRE types associated with the GO term
GM12878	negative regulation of transcription from RNA polymerase II promoter	17
	interferon-gamma-mediated signaling pathway	16
	transcription, DNA-templated	14
	small molecule metabolic process	14
	blood coagulation	14
	response to estradiol	13
	regulation of transcription, DNA-templated	13
	protein phosphorylation	12
	liver regeneration	12
	inflammatory response	12
	transforming growth factor beta receptor signaling pathway	11
	positive regulation of defense response to virus by host	11
H1-hESC	transcription, DNA-templated	16
	small molecule metabolic process	14
	response to wounding	14
	liver regeneration	14
	defense response to bacterium	13
	positive regulation of canonical Wnt signaling pathway	12
	positive regulation of Wnt signaling pathway	12

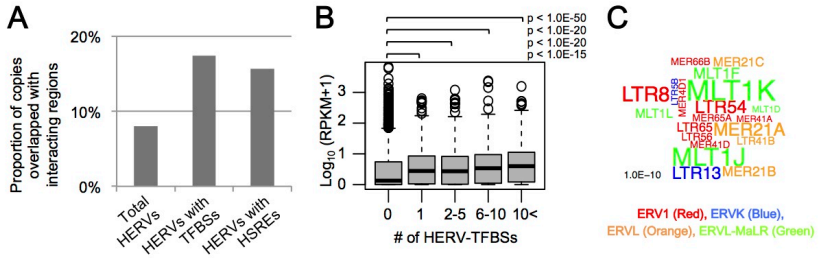
	negative regulation of transcription, DNA-templated	12
	negative regulation of transcription from RNA polymerase II promoter	12
	response to endoplasmic reticulum stress	11
	positive regulation of transcription, DNA-templated	11
	negative regulation of cell proliferation	11
	negative regulation of NF-kappaB transcription factor activity	11
K562	small molecule metabolic process	79
	transcription, DNA-templated	52
	protein phosphorylation	49
	immune response	41
	blood coagulation	39
	innate immune response	38
	negative regulation of transcription from RNA polymerase II promoter	37
	oxidation-reduction process	32
	negative regulation of transcription, DNA-templated	31
	inflammatory response	31
	viral process	30
	cellular lipid metabolic process	30
	xenobiotic metabolic process	29
	transforming growth factor beta receptor signaling pathway	29
	regulation of transcription, DNA-templated	29
	gene expression	29
HepG2	small molecule metabolic process	23
	xenobiotic metabolic process	20
	defense response to bacterium	18
	viral process	17
	transcription, DNA-templated	14
	response to wounding	14
	negative regulation of apoptotic process	14
	kidney development	14
	cell differentiation	14
	blood coagulation	14

	angiogenesis	14
	ubiquitin-dependent protein catabolic process	13
	positive regulation of transcription from RNA polymerase II promoter	13
	chondroitin sulfate metabolic process	13
	response to ethanol	12
HeLa-S3	small molecule metabolic process	14
	immune response	14
	liver regeneration	11
	interferon-gamma-mediated signaling pathway	11
	cytokine-mediated signaling pathway	11
	cellular response to lipopolysaccharide	11
	cellular nitrogen compound metabolic process	11

Long-range interactions between promoters and HERV regulatory elements

Some regulatory elements affect the remote genes via three-dimensional (3D) interactions by forming chromatin loops [70]. I attempted to extract such 3D interactions between HERV-TFBSs/HSREs and promoters of host genes from the data on promoter-captured Hi-C (pcHi-C) in GM12878 cells [76, 77]. pcHi-C is a modified "chromosome conformation capture" method for a comprehensive identification of the 3D interaction between promoters and other genomic regions [76]. I first examined HERV-TFBSs or HSREs present in promoter-interacting regions (interacting regions). In total, 26,194 and 3,860 of HERV-TFBSs and HSREs-containing HERV-TFBSs, respectively, were present in the interacting regions. Some interacting regions were associated with several genes, and 81,536 or 12,452 of interactions between promoters of genes and HERV-TFBSs or HSREs-containing HERV-TFBSs were identified, respectively. The average interval of interactions between promoters and interacting regions containing HERV-TFBSs was 392 kb (average interval of interactions between promoters and all interacting regions was 411 kb in this dataset). HERVs harboring TFBSs or HSREs were enriched two-fold in interacting regions compared with the

population of the entire HERVs (Fig. 21A). Transcription levels (reads per kilobase per million mapped reads; RPKM) of genes tended to be higher as the number of HERV-TFBSs interacting with the genes increased (Fig. 21B). Thus, the HERV regulatory elements in interacting regions seem to work as transcriptional modulators of host genes via long-range interactions. Members of the MLT1, MER21, and MER41 groups were enriched in interacting regions, together with LTR8, LTR54, and LTR13 (Fig. 21C). Next, I developed and performed a "Hi-C-based" GO enrichment analysis by modifying a statistical method used in GREAT [75] (see Materials and Methods). As shown in Fig. 21D, HERV-TFBSs were highly enriched in GO terms associated with immune response such as "positive regulation of interleukin-2 production" and "dendritic cell chemotaxis," consistent with the result of "distance-based" GO enrichment analysis as shown in Fig. 19B. Furthermore, using the Hi-C-based GO enrichment analysis, I ascertained biological processes in GO term with which HERV-TFBSs were more enriched compared to the other TFBSs. Consistent with the above results, HERV-TFBSs showed significantly stronger associations with biological processes relevant to immune responses compared to the other TFBSs (Table 8).



D

GO term (biological process)	P value (-log10)	Fold enrichment
positive regulation of interleukin-2 production	68.3	6.2
dendritic cell chemotaxis	63.9	4.7
intraciliary transport	60.2	5.9
N-glycan processing	59.1	4.6
replication fork processing	48.8	3.3
positive regulation of alpha-beta T cell proliferation	47.5	4.7
response to X-ray	46.4	3.2
regulation of phagocytosis	44.7	5.6
regulation of cysteine-type endopeptidase activity involved in apoptotic process	40.4	4.7
cellular response to estradiol stimulus	39.4	2.6
response to antibiotic	34.7	2.6
positive regulation of protein oligomerization	32.6	4.3
phosphate ion transmembrane transport	29.7	2.6
developmental process	29.7	4.3
JAK-STAT cascade	29.5	2.3

Figure 21. Long-range interactions between HERV-TFBSs/HSREs and promoters of host genes. The interactions were extracted using pcHi-C dataset in GM12878 cells. Results from unique-read TFBSs are shown. A) Proportion of HERV copies overlapped with promoter-interacting regions. Proportions of total HERVs, HERVs with HERV-TFBSs, and HERVs with HSREs are separately shown. B) Transcription levels ($\log_{10}$ (RPKM+1)) of protein-coding genes and number of HERV-TFBSs interacting with the genes. Genes were divided into five categories based on the number of HERV-TFBSs interacting with the genes (0, 1, 2–5, 6–10, and 10<). Categories of the 0, 1, 2–5, 6–10, and 10< respectively contained 13,265, 1,179, 1,946, 822, and 1,639 of genes. P values were calculated using the Mann-Whitney U test with adjustment for multiple tests using the BH method. C) The word cloud indicating HERV groups enriched in the interacting regions. Word sizes are proportional to the $-\log_{10}$ (p value) calculated using the Fisher's exact test. The word colors indicate HERV families. D) Hi-C-based GO enrichment analysis. A set of all HERV-TFBSs in GM12878 cells was used. HERV-TFBSs identified in cells treated with special conditions (e.g., supplement of interferon) were excluded. GO terms were summarized by REVIGO. GO terms with hold enrichment scores of >2 are shown.

Table 8. Hi-C-based GO enrichment analysis to ascertain biological processes in which HERV-TFBSs were more enriched than the other TFBSs.

GO term (biological process)	P value (-log10)	Fold enrichment
N-glycan processing	50.0	4.3
intraciliary transport	48.6	5.3
positive regulation of interleukin-2 production	35.4	3.6
dendritic cell chemotaxis	33.6	3.0
positive regulation of alpha-beta T cell proliferation	32.6	3.7
positive regulation of keratinocyte differentiation	32.4	2.7
positive regulation of synapse maturation	28.4	3.5
detection of chemical stimulus involved in sensory perception of smell	28.4	2.1
phosphate ion transmembrane transport	27.1	2.6
cellular response to exogenous dsRNA	25.7	3.0
regulation of cysteine-type endopeptidase activity involved in apoptotic process	25.7	3.5
positive regulation of dendrite morphogenesis	24.3	2.5
regulation of phagocytosis	23.8	3.4
positive regulation of protein oligomerization	23.7	3.5
response to tumor necrosis factor	23.1	2.1

Construction of dbHERV-REs

My colleagues and I constructed dbHERV-REs, a database of HERV regulatory elements with an interactive user interface (http://herv-tfbs.com/) (Fig. 22). The database provides (i) general information on HERVs such as family classification, copy number, and insertion date judged by distribution of orthologous copies among mammalian genome; (ii) positions of HERV-TFBSs, HSREs, and HERV-DHSs in the consensus sequence of HERVs and in the human reference genome; and (iii) results of GO enrichment analyses with GREAT [75] using sets of respective HSREs. The

database also can compare phylogenetic relationship of HERV copies with the presence of orthologous copies across the mammalian genome, TFBSs, and TF-binding motifs. Results of all- and unique-read TFBSs are available in the database. Additionally, the database provides results on pre-determined TFBSs provided by ENCODE and Roadmap, which were based on their analytical pipelines of ChIP-Seq peak calling (Table 2). As of May 2017, TFBSs for 97 TFs and DHSs for 125 cell types were deposited. A user can focus on significant associations between HERVs and TFs by setting statistical and other thresholds.

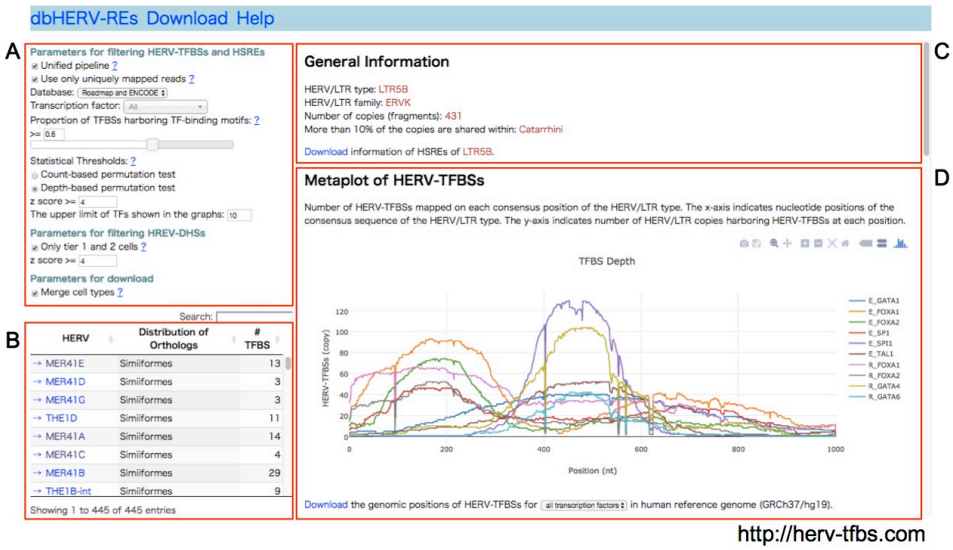


Figure 22. A screenshot of dbHERV-REs (http://herv-tfbs.com/). The screenshot when LTR5B was selected is shown. A) Statistical and other parameters filtering HERV-TFBSs, HSREs, and HERV-DHSs. B) The list of HERVs that can be selected under the parameters. C) General information of the selected HERVs. D) Visualized data. In this figure, the graph shows number of HERV-TFBSs mapped on each consensus position.

Discussion

I showed that HERVs frequently contained HERV-TFBSs/HSREs for TFs essential in hematopoietic (e.g., SPI1, TAL1, and GATA1/2), pluripotent (e.g., SOX2, POU5F1, and NANOG), and embryonic endoderm/mesendoderm cells (e.g., GATA4/6, SOX17, and FOXA1/2). Hematopoietic regulatory elements of HERVs seem to descend from ancestral exogenous retroviruses, which would have replicated in the hematopoietic (or blood) cells, considering that modern exogenous retroviruses frequently contain such regulatory elements [6]. Pluripotent regulatory elements seem to have been crucial for efficient replication of HERVs in germ cells, as with other TEs such as LINE1, because transcriptional environments are similar between pluripotent and early embryonic cells [50, 78]. Endoderm/mesendoderm regulatory elements also seem to be important for HERVs, possibly for their replication in the host germ cells immediately after the endogenization, as these TFs highly expressed in both somatic and germ cells [35]. A previous study showed that the regulatory elements of HERVs are active in various cells and tissues by evaluating enrichment of active histone modifications on HERVs [72]. Therefore, as the number of available ChIP-Seq datasets increase, a greater number of regulatory elements of HERVs will be identified.

Although the role of retroviral internal sequences in transcription remains unclear, it is known that an internal sequence in Human T-cell Leukemia Virus Type 1 (HTLV-1) contains a CTCF-binding site functioning as an insulator [79]. In the present study, I found that a substantial fraction of HSREs was present in the internal sequences, and the most frequently observed HSRE in the internal sequences was the CTCF-binding site (Fig. 9I). These findings suggest that regulatory elements, particularly CTCF-binding sites, would be present in the internal sequences of retroviruses, including HERVs, more than previously considered [6, 79]. Further investigation is needed for clarifying the role of retroviral internal sequences in transcriptional modulation.

Pluripotent regulatory elements seem to be essential for HERVs and other TEs to replicate efficiently in the host germ cells and to expand in the host genome. However, the pluripotent regulatory elements are rarely observed in exogenous retroviruses, even though HERVs descended from ancient exogenous retroviruses [6]. In this study, I demonstrated the heterogeneity of regulatory elements among subgroups in LTR7 (Figs. 11G-K), LTR5 group (Fig. 16), LTR6A (Fig. 23), LTR9 (Fig. 24), MER11C (Fig. 25), and MER11B (Fig. 26). Such heterogeneity of regulatory elements was also observed in endogenous retroviruses (ERVs) of other mammals [80, 81]. These indicate that gains or losses of the regulatory elements occurred during genomic expansions of the HERVs (or the ERVs). I observed a tendency that younger subgroup of HERVs had more regulatory elements for pluripotent TFs (e.g., NANOG, POU5F1, and SOX2) in LTR7, LTR5_Hs, LTR6A, and MER11C (Figs. 11G-K, 16, 23, and 25, respectively) although I observed an opposite tendency in MER11B (Fig. 26). Thus, HERVs seem to have frequently acquired pluripotent regulatory elements. I hypothesize that these HERVs acquired the pluripotent regulatory elements after endogenization for efficient replication and genomic expansion in the host germ cells. Thus, investigation of heterogeneity of regulatory elements of HERVs can illuminate the evolutionary dynamics of transcriptional modulation system of HERVs.

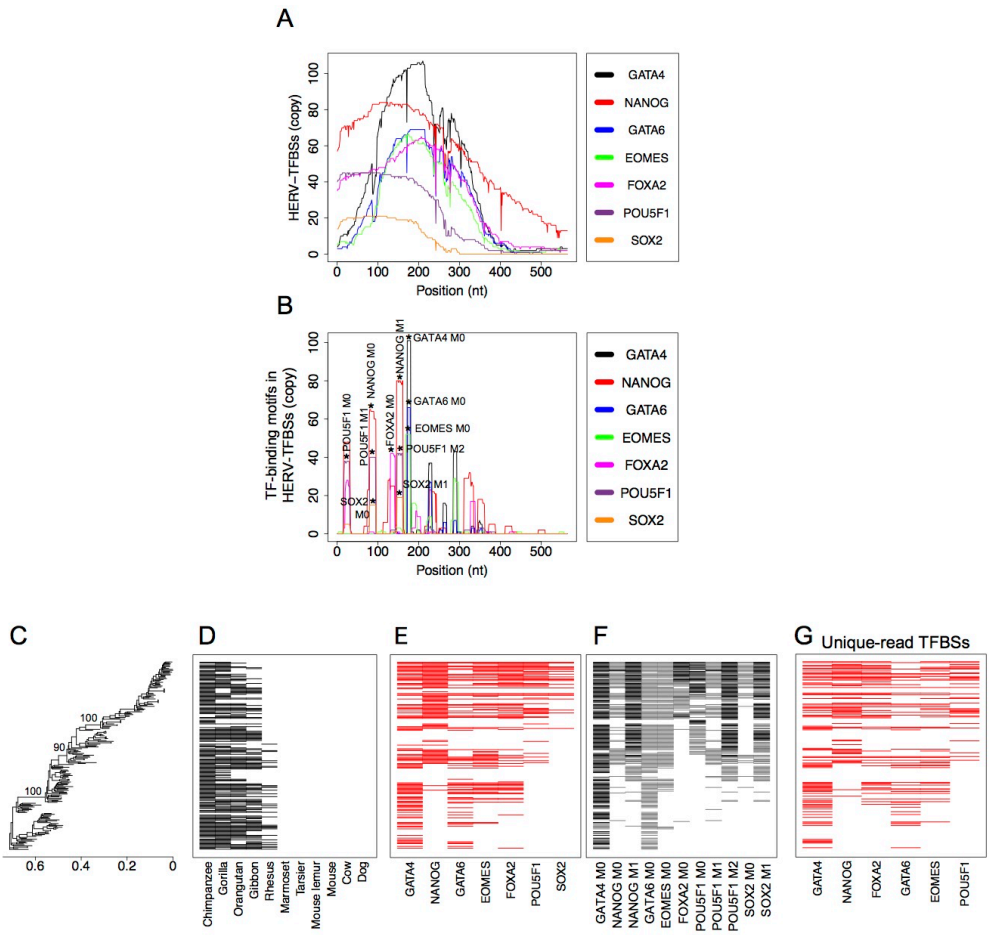


Figure 23. Characteristics of HSREs identified in LTR6A from Roadmap dataset. Results of all-read TFBSs are shown except for (G). A) Number of HERV-TFBSs mapped on each consensus position of LTR6A. The X-axis indicates nucleotide position of the consensus sequence. The Y-axis indicates number of HERV copies harboring HERV-TFBSs at each position. B) Number of TF-binding motifs in HERV-TFBSs mapped on each consensus position of LTR6A. The X-axis indicates nucleotide position of the consensus sequence. The Y-axis indicates number of HERV copies harboring the TF-binding motifs at each position. Peaks of the motifs corresponding to HSREs are indicated by an asterisk (*) with motif names. C) The unrooted phylogenetic tree of LTR6A copies constructed by maximum likelihood method. Fragmented and outlier copies were excluded from the analysis. In total, 204 (out of 288) of LTR6A copies were included in the tree. Representative supporting values calculated by SH-like test are shown on the corresponding branches. D) Orthologous copies of LTR6A in the reference genomes of other mammals. E) TFBSs on each LTR6A copy. F) TF-binding motifs on each copy at positions corresponding to HSREs. Black and gray colors respectively indicate presence of motifs with p values of <0.0001 and <0.001. G) TFBSs on each LTR6A copy. Results of unique-read TFBSs are shown.

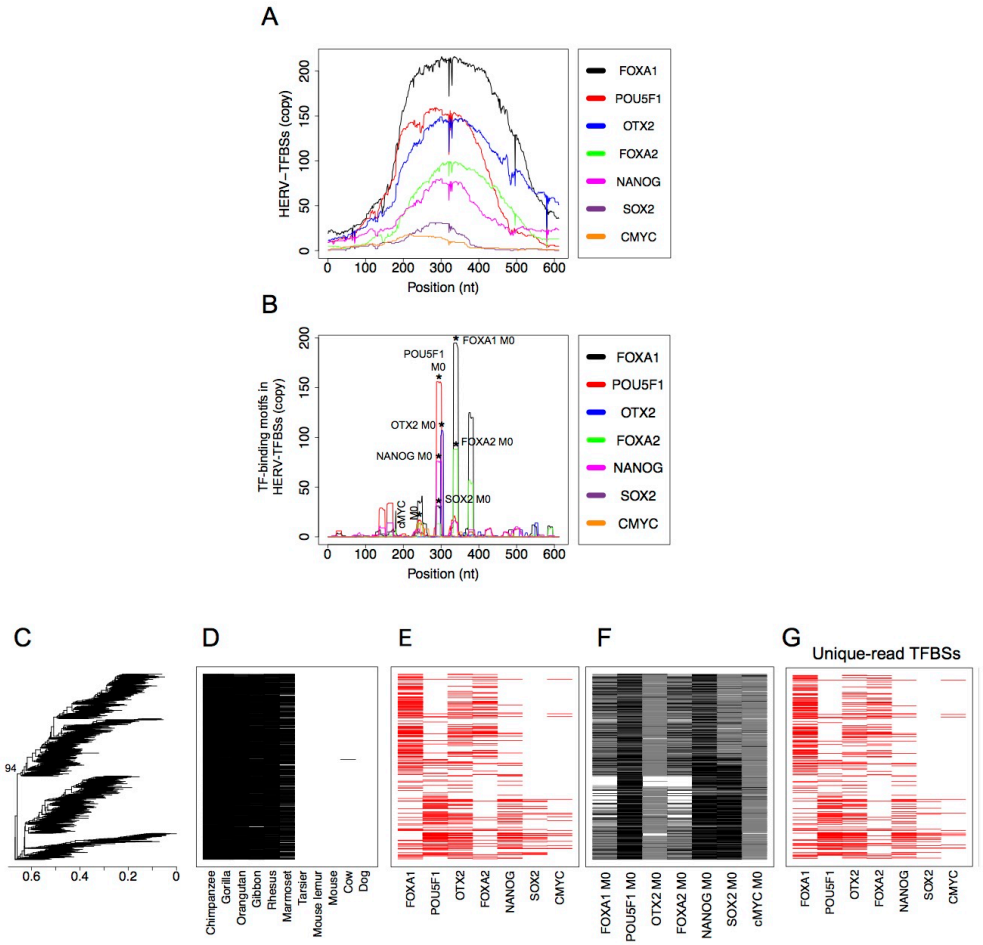


Figure 24. Characteristics of HSREs identified in LTR9 from Roadmap dataset. Results of all-read TFBSs are shown except for (G). A) Number of HERV-TFBSs mapped on each consensus position of LTR9. The X-axis indicates nucleotide position of the consensus sequence. The Y-axis indicates number of HERV copies harboring HERV-TFBSs at each position. B) Number of TF-binding motifs in HERV-TFBSs mapped on each consensus position of LTR9. The X-axis indicates nucleotide position of the consensus sequence. The Y-axis indicates number of HERV copies harboring the TF-binding motifs at each position. Peaks of the motifs corresponding to HSREs are indicated by an asterisk (*) with motif names. C) An unrooted phylogenetic tree of LTR9 copies constructed using the maximum likelihood method. Fragmented and outlier copies were excluded from the analysis. In total, 1,077 (out of 2,011) of LTR9 copies were included in the tree. Representative supporting values calculated by SH-like test are shown on the corresponding branches. D) Orthologous copies of LTR9 in reference genomes of other mammals. E) TFBSs on each LTR9 copy. F) TF-binding motifs on each copy at positions corresponding to HSREs. The black and gray colors respectively indicate the presence of motifs with p values of <0.0001 and <0.001. G) TFBSs on each LTR9 copy. Results of unique-read TFBSs are shown.

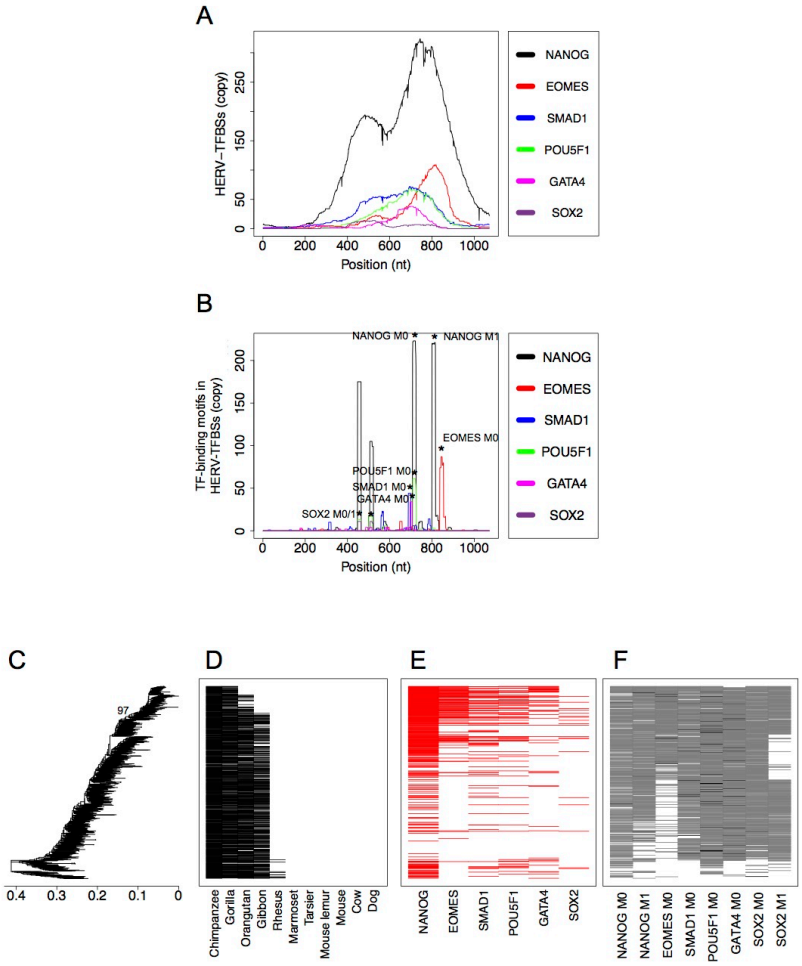


Figure 25. Characteristics of HSREs identified in MER11C from Roadmap dataset. Results of all-read TFBSs are shown. A) Number of HERV-TFBSs mapped on each consensus position of MER11C. The X-axis indicates nucleotide position of the consensus sequence. The Y-axis indicates number of HERV copies harboring HERV-TFBSs at each position. B) Number of TF-binding motifs in HERV-TFBSs mapped on each consensus position of MER11C. The X-axis indicates nucleotide position of the consensus sequence. The Y-axis indicates number of HERV copies harboring the TF-binding motifs at each position. Peaks of the motifs corresponding to HSREs are indicated by an asterisk (*) with motif names. C) An unrooted phylogenetic tree of MER11C copies constructed using the maximum likelihood method. Fragmented and outlier copies were excluded from the analysis. In total, 748 (out of 866) of MER11C copies were included in the tree. Representative supporting values calculated by SH-like test are shown on the corresponding branches. D) Orthologous copies of MER11C in reference genomes of other mammals. E) TFBSs on each MER11C copy. F) TF-binding motifs on each copy at positions corresponding to HSREs. The black and gray colors respectively indicate the presence of motifs with p values of <0.0001 and <0.001.

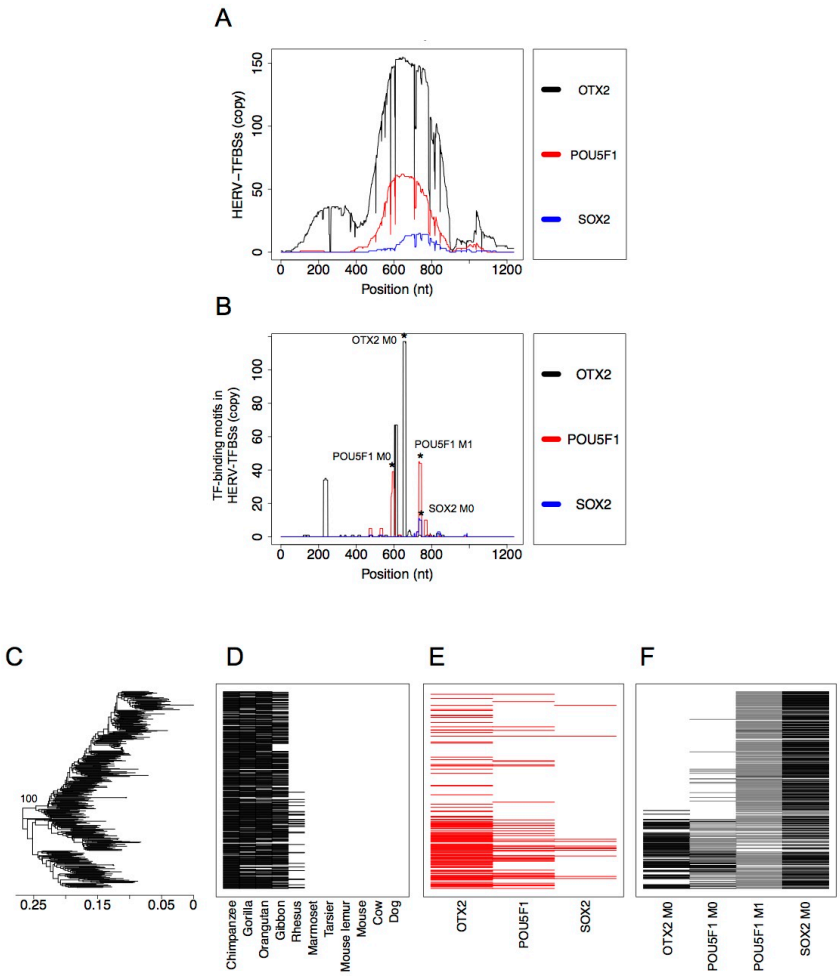


Figure 26. Characteristics of HSREs identified in MER11B from Roadmap dataset. Results of all-read TFBSs are shown. A) Number of HERV-TFBSs mapped on each consensus position of MER11B. The X-axis indicates nucleotide position of the consensus sequence. The Y-axis indicates number of HERV copies harboring HERV-TFBSs at each position. B) Number of TF-binding motifs in HERV-TFBSs mapped on each consensus position of MER11B. The X-axis indicates nucleotide position of the consensus sequence. The Y-axis indicates number of HERV copies harboring the TF-binding motifs at each position. Peaks of the motifs corresponding to HSREs are indicated by an asterisk (*) with motif names. C) An unrooted phylogenetic tree of MER11B copies constructed using the maximum likelihood method. Fragmented and outlier copies were excluded from the analysis. In total, 377 (out of 548) of MER11B copies were included in the tree. Representative supporting values calculated by SH-like test are shown on the corresponding branches. D) Orthologous copies of MER11B in reference genomes of other mammals. E) TFBSs on each MER11B copy. F) TF-binding motifs on each copy at positions corresponding to HSREs. The black and gray colors respectively indicate the presence of motifs with p values of <0.0001 and <0.001.

LTR7 is essential for the maintenance of pluripotency in ES and iPS cells, and it has been hypothesized that LTR7 insertions rewired the core regulatory network of the pluripotent cells [16-18]. I further clarified the heterogeneity among subgroups of LTR7 with respect to insertion dates, TF binding profiles, and transcriptional activities. Subgroup III, the youngest subgroup of LTR7, was most frequently bounded by SOX2, POU5F1, and KLF4 (Figs. 11G-K and 14). Subgroup III also showed the highest enrichment of ChIP-Seq reads of NANOG (Fig. 11K). Subgroup III showed the highest transcriptional activity in pluripotent cells (Fig. 15). Most LTR7-chimeric transcripts, which are vital in maintaining pluripotency [15-18], were composed of LTR7 belonging to the subgroup III (Fig. 15). These findings suggest that the evolutionary rewiring of the core regulatory network of pluripotent cells was caused by a specific population of LTR7, i.e., members of the subgroup III, rather than by the entire population of LTR7 (Fig. 11L). Moreover, this rewiring seems to have occurred more recently than previously thought [82], the branch from the end of *Hominoidea* to *Homininae.* This is because the rewiring should have occurred during the period when subgroup III was inserted (Figs. 11G, 11H, and 11L). Further investigation is needed to elucidate the evolution of pluripotent cells due to LTR7 insertions.

The GO enrichment analysis based on genomic positions of HERV-TFBSs/HSREs demonstrated that HERV-TFBSs/HSREs tend to be located near the genes involved in innate immune responses such as cytokine-mediated signaling (Figs. 19B, Tables 5, 6, 7). This tendency was recaptured by Hi-C-based GO analysis, which used information on 3D interactions between HERV-TFBSs and promoters of host genes in B-lymphocytes (GM12878 cells) (Fig. 21). In those GO enrichment analyses, HERV-TFBSs showed significantly stronger associations with biological processes relevant to innate immune responses compared to the other TFBSs (Tables 5 and 8). This suggests that HERV regulatory elements were likely to be associated with regulatory networks controlling innate immune responses. Furthermore, this tendency

seems to be more attributable to natural selection of HERVs after the insertions than preferential insertions in specific genomic regions, because HERV copies with TFBSs were more enriched in regions near the genes related to innate immune response than HERVs without TFBSs (Table 6). The tendency of regulatory elements of HERVs being associated with innate immune response seemed to be affected by cell types (e.g., B-lymphocytes) in which ChIP-Seq was performed. Therefore, as the number of cell types in which ChIP-Seq are performed increase, more associations between HERVs with TFBSs and specific biological functions will be identified. Finally, GO enrichment analyses showed that each type of HSRE was statistically associated with various biological processes in addition to the immune response [deposited in dbHERV-REs (http://herv-tfbs.com)]. Further research, especially knockout-based studies such as the one by Chong *et al.* [43], is necessary to prove the causal relationship between regulatory elements of HERVs and regulatory networks controlling specific biological processes.

To summarize, I identified various HERV regulatory elements involved in several host regulatory networks. Our study provides the foundation to understand the impact of HERVs on host transcription, and provides insights into transcriptional modulation systems that HERVs and ancestral retroviruses of HERVs originally used.

Materials and Methods

Datasets

Information on the ChIP-Seq dataset is summarized in the "peak calling of ChIP-Seq" section. RepeatMasker output file (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/chromOut.tar.gz) was downloaded from the UCSC genome browser (https://genome.ucsc.edu/). This is an annotation file of repetitive elements on the human reference genome (GRCh37/hg19) used in RepeatMasker track in the genome browser. Consensus sequences of HERVs were obtained from the RepeatMasker library (20140131 release) and Repbase Update (1.1.3 release) in Repbase (http://www.girinst.org/server/RepBase/). DHS datasets were obtained from ENCODE. Genome segmentations in six cell types (combined between ChromHMM and Segway) [64-66] were obtained from ENCODE. Datasets of Cold Spring Harbor Laboratory (CSHL) LongPolyA RNA-Seq were obtained from ENCODE in the GTF format. Ontology file (go-basic.obo, date; 3/16/2016) and GO association file (gene_association.goa_human, submission date; 3/16/2016) were downloaded from the GO Consortium (http://geneontology.org/). The UCSC known genes were downloaded from UCSC (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/knownGene.txt.gz). pcHi-C dataset in GM12878 cells [76, 77] (GSE81503_GM12878_PCHiC_merge_final_seqmonk.txt.gz and GSE81503_GM12878_PCHiC_merge_final_washU_text.txt.gz, accession GSE81503) were obtained from the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/).

Peak calling of ChIP-Seq

An analytical pipeline used in this study is summarized in Fig. 3B. For the Roadmap dataset, I obtained a sequence read file (fastq format) from the Sequence Read Archive (SRA) using the SRA Toolkit fastq-dump

(http://www.ncbi.nlm.nih.gov/books/NBK158900/). For the ENCODE dataset, I downloaded an unfiltered alignment file, if available, for GRCh37/hg19 (bam format) from the ENCODE database (http://www.encodeproject.org/). The unfiltered alignment file was generated using the ENCODE Processing Pipeline with BWA 0.7.10 (aln and samse). If the unfiltered alignment file was not available, I downloaded a fastq file from the ENCODE database. Fastq or bam files of biological replicates were then concatenated. Sequence reads in the fastq files were mapped to human reference genome (GRCh37/hg19) using BWA 0.7.12 (aln and samse/sampe). In the default setting of BWA aln, a multiple mapped read is randomly assigned to a particular genomic position chosen from candidate positions. For the all-read TFBSs, ChIP-Seq peaks were called using MACS2 with default setting. For unique-read TFBSs, multiple mapped reads or reads with low mapping quality (reads with MAPQ score of <10) were removed using samtools view [83], and then ChIP-Seq peaks were called. In peak calling, input control file was used with ChIP-treated file.

Identification of HERV-TFBSs and HSREs

HERV-TFBSs and HSREs were identified separately in ENCODE and Roadmap datasets. HERV-TFBSs and HSREs were identified both in all- and unique-read TFBSs.

I identified HERV-TFBSs in respective cell types by examining the overlaps between HERVs and TFBSs with bedtools intersect [84]. In the respective TFs, TFBSs or HERV-TFBSs among all cell types or conditions were merged with bedtools merge [84] (referred to as the merged TFBSs or HERV-TFBSs). For counting TFBSs and HERV-TFBSs, the merged TFBSs and HERV-TFBSs were used.

For the identification of HSREs, the merged HERV-TFBSs were used. First, sequences of HERV copies were extracted from human reference genome (GRCh37/hg19) using bedtools getfasta [84]. Multiple sequence alignment (MSA) of HERV copies was constructed with a consensus sequence of the corresponding HERV

group. MAFFT v7.239 [85] was used for the construction of MSA with the options --addfragments, --keeplength, and --retree 2. In this setting, the consensus sequence was used as input, and sequences of HERV copies were used as fragment sequence. In MSA, the position of HERV-TFBSs was mapped on each HERV sequence, and then the number of the mapped HERV-TFBSs was counted at every consensus position (referred to as "depth" of HERV-TFBSs). For setting the threshold to identify peaks of HERV-TFBSs, randomized (shuffled) TFBS datasets were generated with bedtools shuffle [84] for 500 times. In the respective randomized datasets, the depth of HERV-TFBSs was counted for each consensus position with the above-mentioned procedures. For every consensus position, average and standard deviation of the depth of HERV-TFBSs among randomized datasets was calculated. Standardized score (z score) of HERV-TFBS depth was calculated for every consensus position with the average and standard deviation in randomized datasets (termed as base-wise z score). If base-wise z score of a given region (>50-bp) in the consensus sequence was higher than four, the region was defined as a peak of HERV-TFBSs. Finally, known TF-binding motifs of the corresponding TF were scanned in original HERV-TFBS sequences. For motif scanning, FIMO [86] and known TF-binding motifs recorded in JASPAR [87] and HOCOMOCO [88] were used. The threshold (p value) of the motif scanning was set at 0.001. In MSA, position of the TF-binding motif was mapped on each HERV sequence, and then the number of the mapped motifs was counted at every consensus position (referred to as "depth" of TF-binding motifs). To identify HSREs, heights of peaks of depths were compared between HERV-TFBSs and TF-binding motifs. If the height of the TF-binding motif peak is (i) greater than or equal 10 and (ii) greater than 60% of the height of the HERV-TFBS peak, I regard the set of TF-binding motifs as HSRE (Fig. 13). For counting the number of genomic positions of HSREs, overlapping HSREs of the same TF were merged for avoiding double counts. This is because some TF-binding motifs were present in both strands at approximately the same positions due

to their palindrome signatures.

After identifying HSREs, overlaps between HSREs and HERV-TFBSs in respective cell types were examined, and the cell specificities of HSREs were determined.

Randomization test shuffling genomic positions of TFBSs

HERV-TFBS overlaps were counted for all combinations. In each dataset of TFBS, I generated 100 times of randomized TFBS datasets using bedtools shuffle [84] and counted the number of HERV-TFBS overlaps in the randomized datasets. Among the randomized datasets, average and standard deviation of numbers of HERV-TFBS overlaps were calculated. In each HERV-TFBS combination, I calculated z score (count-based z score) using the number of HERV-TFBS overlaps in an observed dataset and the average and standard deviation among randomized datasets.

For TEs other than HERVs, z scores for all combinations of respective TE groups and the merged TFBSs were calculated using the same procedures.

Hierarchical clustering

I used unique-read TFBSs, and separately dealt with TFBSs of the same TF in distinct cell types. If there were several TFBS files for the same ChIP-Seq condition, the TFBS files were merged using bedtools merge [84]. All TFBSs (e.g., SOX2-binding sites in HUES64 cells from Roadmap) were used for the analysis, except for CTCF-binding sites; I used CTCF-binding sites that were determined in tier 1 and 2 cells of ENCODE (GM12878, H1-hESC, K562, HepG2, HeLa-S3, and HUVEC), HUES64 cells, and germ layer (ectoderm, endoderm, mesoderm, and mesendoderm) cells that were differentiated from the HUES64 cells. Z scores were calculated using the method in the "randomization test shuffling genomic positions of TFBSs" section. A matrix containing the z scores was created. HERV group whose copy number was less than

100 was excluded from the matrix. Rows (TFBSs) and columns (HERVs) were excluded if they did not contain any elements whose z scores were greater than or equal to 10. Distance matrix was constructed using the Euclid method based on the z score matrix. I performed hierarchical clustering with the distance matrix using Ward's method. All analyses were performed by packages of amap and ReorderCluster implemented in R.

Phylogenetic analyses

Phylogenetic trees were constructed for HERV groups satisfying the following criteria: (i) after removal of the fragmented copies (described below), the number of copies fell within the range of 10–2,500; and (ii) greater than 30% of their copies remained after the removal of fragmented copies. Fragmented and outlier copies were excluded from the analysis. For defining the fragmented copies, I constructed preliminary MSA of HERV copies with the consensus sequence using MAFFT v7.239 [85] with options of --addfragments, --keeplength, and --retree 2 (in this setting, the consensus sequence was used as input, and sequences of HERV copies were used as fragment sequence). HERV copies were defined as fragmented if less than 80% of their sequences were only aligned to the consensus sequences in the preliminary MSA. After the removal of fragmented copies, I constructed MSA of HERV copies using MAFFT v7.239 with --auto options. Sites in the MSA containing gaps were excluded if site coverages of those positions were less than 30%. For defining the outlier copies, a preliminary tree was reconstructed with RAxML v8.2.0 [89]. GTRCAT was used as a nucleotide substitution model. Z score of the length of external branch was calculated for the preliminary tree. Outlier copy, whose z score of the branch length was greater than three, was excluded from the MSA. I constructed the final tree using the same procedures with the preliminary tree. Supporting values were calculated using the SH-like test [90]. In addition to the SH-like test, rapid bootstrap analysis [89] (100 times) was performed for

the phylogenetic tree of the LTR5 group.

Estimation of the insertion dates of HERVH/LTR7 copies

The age of a provirus of ERVs can be estimated by sequence comparison between 5′- and 3′-LTRs of the ERVs, as sequences of both LTRs were identical at the time of insertion, and after the insertion, both LTRs independently accumulated mutations as a part of the host genome [91]. In this analysis, I used the annotation of a provirus of HERVH/LTR7 as reported previously [17]. I only analyzed proviruses of HERVH/LTR7 harboring two LTR7 sequences that were categorized in the same subgroup in the tree (Fig. 11G). For each provirus, a pairwise sequence alignment of 5′- and 3′-LTRs was constructed using the EMBOSS Stretcher program [92]. After removal of all gapped sites in the alignment, p-distance of the paired LTRs was calculated, and then the genetic distance of the paired LTRs was computed using the Jukes-Cantor 69 model. A substitution rate of HERVs of 1.0×10^{-9} per site per year was used as described previously [93]. Insertion date of the provirus was calculated with the formula, D/2R (D, genetic distance of the paired LTRs; R, substitution rate of HERVs).

Insertion date (i.e., age) judged by distribution of orthologous HERV copies in the mammalian genome

For judging whether an orthologous copy of a HERV copy was present in a certain reference genome, liftOver (http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/liftOver) was used. If liftOver successfully converted the genomic position of a particular HERV copy in human reference genome to that of a reference genome of other species, I judged an orthologous copy of the HERV copy was present in the genome of the corresponding species. A minimum match parameter was set at 0.5. Reference genomes of PanTro4 (chimpanzee), GorGor3 (gorilla), PonAbe2 (orangutan), Nomleu3 (gibbon), RheMac3

(rhesus macaque), CalJac3 (marmoset), TarSyr1 (tarsier), MicMur1 (mouse lemur), Mm9 (mouse), Bostau7 (cow), and CanFam3 (dog) were used.

Classification of insertion date of HERVs was defined as follows: ~*Hominoidea*; greater than 10% of orthologous copies of the HERV group present in any of the chimpanzee, gorilla, orangutan, and gibbon genomes but absent in that of the rhesus macaque. *Catarrhini*; greater than 10% of orthologous copies of the HERV group present in the chimpanzee, gorilla, orangutan, gibbon, and rhesus macaque genomes but absent in that of the marmoset. *Simiiformes*; greater than 10% of orthologous copies of the HERV group present in the chimpanzee, gorilla, orangutan, gibbon, rhesus, and marmoset genomes but absent in those of the tarsier and mouse lemur. *Primates*; greater than 10% of orthologous copies of the HERV group present in the chimpanzee, gorilla, orangutan, gibbon, rhesus, marmoset, tarsier, and mouse lemur genomes but absent in those of the mouse, cow, and dog. *Eutheria*~; greater than 10% of orthologous copies of the HERV group present in the chimpanzee, gorilla, orangutan, gibbon, rhesus, marmoset, tarsier, mouse lemur, mouse, cow, and dog genomes. I only analyzed HERV groups whose copy numbers were greater than or equal to 100.

Gene ontology enrichments analysis

Unique-read TFBSs were used in GO enrichment analyses. GO associations described in gene_association.goa_human were used. GO term associated with greater than or equal to five genes was used in the analyses.

In distance-based GO enrichment analysis, the createRegulatoryDomains command in the local version of GREAT [75] was used for defining regulatory domains of respective GO terms with the option of basal (five kb upstream and one kb downstream of the TSS) plus extension (up to one Mb). I used the TSS annotation based on the UCSC known genes. Enrichment score and p values with binomial test were calculated by the original R script.

To determine the GO term in which TFBSs with HERVs were more enriched than the other TFBSs (TFBSs not on HERVs), I counted the number of TFBSs with HERVs and the entire TFBSs in regulatory domains associated with a certain GO term. Then, the enrichment significance was calculated by Fisher's exact test.

In order to examine the GO term in which HERVs harboring TFBSs were more enriched than the entire HERVs (all HERVs regardless of overlaps with TFBSs), I estimated the number of HERVs harboring TFBSs and the entire HERVs overlapped to regulatory domains associated with a certain GO term. The enrichment significance was calculated by Fisher's exact test.

To ascertain the GO term in which each type of HSRE was enriched, I performed the GREAT analysis [75] using a set of HERV-TFBSs harboring a HSRE in each cell type. The threshold for statistical significance was set at 0.1, with false discovery rates calculated using the Benjamini–Hochberg (BH) method.

I thus developed the "Hi-C-based" GO enrichment analysis by modifying the GREAT algorithm [75]. Interacting regions in pcHi-C [76, 77] of all genes were merged using bedtools merge [84] and were defined as "total region". Interacting regions of genes associated with a particular GO term were merged and were defined as "regulatory domain" for the corresponding GO term. The lengths of the total region and regulatory domain were calculated (termed total_length and regdom_length, respectively). HERV-TFBSs overlapping with the total region and regulatory domain were also counted (termed total_count and regdom_count, respectively). For calculating the enrichment significance, I performed a binomial test using the above total_count and regdom_count in addition to the ratio of regdom_length and total_length (regdom_length/total_length).

In Hi-C-based GO enrichment analysis, I performed GO enrichment analysis to determine the GO term in which TFBSs with HERVs were more enriched than the other TFBSs. I counted the number of TFBSs with HERVs and the other TFBSs in

regulatory domains associated with a certain GO term. Then, the enrichment significance was calculated by Fisher's exact test.

Enrichment of HERV-TFBSs near the cell type-specifically expressed genes

In CSHL LongPolyA RNA-Seq, protein-coding genes with RPKM >3 in any cell type were included in the analysis. For every gene, z score of RPKM was calculated for each cell type by using the average and standard deviation of the six cell types (GM12878, H1-hESC, K562, HepG2, HeLa-S3, and HUVEC cells). Regarding the z scores, top 200 genes in each cell type were defined as those expressed specifically in the corresponding cell type. Regulatory domain for genes specifically expressed in a certain cell type was created by using the createRegulatoryDomains command in GREAT [53] with a setting of basal (5 kb upstream and 1kb downstream of TSS) plus extension (up to 1 Mb). Enrichment scores and p values with binomial test were calculated by original R scripts.

Construction of dbHERV-REs

The system is running on Amazon Web Service (http://aws.amazon.com/). The relational database was constructed with MySQL. The server program was written in Python using Twisted (http://twistedmatrix.com/), an event-driven networking framework. The user interface was designed upon AJAX (Asynchronous JavaScript + XML) philosophy. plotly.js (http://plot.ly/javascript/) is used for data visualizations. jQuery (http://jquery.com/) was used for the browser scripting.

Chapter 3: Systematic identification of unannotated transcripts derived from human endogenous retroviruses in solid tumors

Introduction

Decades of studies have reported the reactivation of human endogenous retroviruses (HERVs) in tumors: The increases of mRNAs, proteins, and even viral-like particles (VLPs) (i.e., non-infectious viral particles) of HERVs have been observed in various tumors (see review: [94, 95]). Particularly, studies using next generation sequencing (NGS) techniques showed the presence of a large number of unannotated transcripts derived from HERVs in tumors [27, 96]. Although effects of HERV-derived transcript/proteins on tumor characteristics have been controversial, some protein/RNAs of HERVs are likely to have the capacity to promote the tumorigenesis. Envelope (Env) protein of HERVK activates the proliferation of cancer cells via the activation of Ras/ERK pathway [28-30, 97]. The lncRNAs derived from HERVH, which play critical roles in the cellular reprograming [16-18], are highly expressed in bladder and colorectal tumors and promotes the invasion and metastasis of the tumors [98, 99].

Mechanisms underlying the up-regulation of HERV transcriptions in human tumors are mostly unknown [100]. Although the global epigenetic change such as the DNA demethylation in tumors is thought to be a major cause of the up-regulation of HERV transcriptions, there are only a few evidences supporting this hypothesis [100, 101]. Studies based on mouse model have demonstrated that a protein complex comprising KRAB zinc finger proteins (e.g., ZNF274), TRIM28, and SETDB1 plays a central role in the transcriptional suppression of ERVs and other retrotransposons during early embryonic development [9, 102-106]. However, roles of these genes on the suppression of HERVs in human tumor tissues are little known. Furthermore, HERVs possess specific regulatory elements for TFs such as ESR1 (estrogen receptor 1) that are overexpressed and rewire the gene regulatory network in tumors (CHPTER 2) [107, 108]. Therefore, these TFs seem to play an essential role in the up-regulation of HERV transcriptions in tumors; however, the associations of TFs and the HERV up-regulations in tumors are poorly understood.

HERVs can contribute to the production of non-canonical and unannotated transcripts of host genes in tumors transcribed as parts of mRNA of these genes [109]. In HERV-fused gene transcripts, HERVs work as alternative transcription start sites (TSSs), exons, and transcriptional terminal sites (TTSs). Alternative TSSs originated from HERVs are activated in tumors particularly under the treatments with DNA methyltransferase inhibitor (DNMTi) or histone deacetylase inhibitor (HADCi) [27, 110]. Since non-canonical transcripts of genes are likely to encode non-canonical forms of proteins, the production of these transcripts possibly have various effects on tumors, such as the disruption of the protein-protein network in cancer cells [111, 112]. For a better understanding of the landscape of transcriptome in tumors, it is needed to investigate the HERV-fused gene transcripts.

To improve our understanding of tumor pathogenesis, The Cancer Genome Atlas (TCGA) has generated multi-dimensional omics data for >11,000 patients across 33 types of tumors [36]. One of the major purposes of TCGA is to stratify a type of tumors (e.g., breast carcinoma) into subtypes based on molecular profiles of the tumors [36]. The subtype classification is clinically important because these subtypes show distinct phenotypes (e.g., prognosis or drug resistance). Another purpose of TCGA study is to identify pathogenic mechanisms that are shared across distinct types of tumors. This effort is referred to as pan-cancer analysis [36]. TCGA is a publicly available dataset. Therefore, using this dataset, TCGA group and others have carried out pan-cancer analyses from specific viewpoints, such as the alternative splicing [111, 112], immune cell inflation [32, 113], and mutation signature [114, 115].

Although importance of the HERV reactivation in tumors has been emphasized [94, 95], previous TCGA studies have not approached information on HERV transcriptome in tumor tissues. Transcriptome information of HERVs can be extracted from RNA-Seq, particularly ploy A-enriched mRNA-Seq. Previous studies using TCGA data examined only a limited number of HERV loci and tumor types [29,

32, 101]. Therefore, the landscape of the HERV transcriptions in tumors is still unclear. In the present study, I mined the HERV transcriptome data from the TCGA RNA-Seq dataset for 5,550 patients across 12 solid tumors and performed pan-cancer analysis focusing on transcripts derived from HERVs. This study will provide a better understanding of tumor characteristics from the pan-cancer view of HERV transcriptions.

Results

Extracting HERV transcriptome information from TCGA RNA-Seq data

Because HERV annotations are not included in transcript models such as GENCODE [116] and RefSeq [117], I first constructed "gene-HERV" transcript model based on GENCODE and RepeatMasker, an annotation dataset of repetitive sequences including HERVs (http://www.repeatmasker.org/) (Fig. 27A). A few HERVs are annotated as genes (e.g., syncytin-1/2 [23, 39]) and included in the gene transcript model. To focus on unannotated HERV-derived transcripts, I removed HERV loci overlapping with known gene transcripts from the gene-HERV model. This gene-HERV model contains 60,483 protein coding/non-coding genes and 138,124 HERV loci. The total sequence length of HERVs in the model corresponds to 3.5% of the human genome. In order to extract transcriptome information of genes and HERVs, I counted RNA-Seq fragments mapped on genes and HERVs, respectively, using TCGA RNA-Seq data (BAM file) with the gene-HERV model (Fig. 27B). HERVs contribute to the production of non-canonical and unannotated transcripts of genes in tumors transcribed as parts of mRNA of these genes [109]. To identify such fused transcripts of genes and HERVs, I counted RNA-Seq fragments mapped both on genes and HERVs (Fig. 27B).

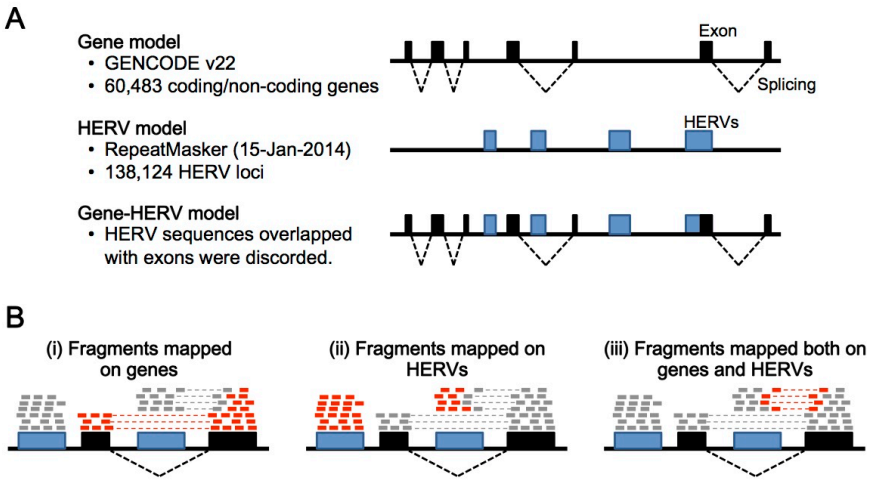


Figure 27. Mining of transcriptome data using the gene-HERV transcript model. A) The gene-HERV transcript model. Black boxes and dot lines respectively indicate exons and intronic regions of genes. Blue boxes indicate HERV sequences. As a gene model, GENCODE Version 22 was used, which contains 19,814 protein-coding genes and 40,669 non-coding genes. As a HERV model, RepeatMakser out file (15-Jan-2014) was used with filtering out unreliable HERV loci (Smith–Waterman (SW) score <2,500). The HERV model contains 138,124 HERV loci. The gene-HERV model was constructed by merging the gene and HERV models. HERV sequences overlapped with exons were excluded from the gene-HERV model. B) Three types of counting methods of RNA-Seq fragments. Fragments mapped on genes and HERVs were counted ((i) and (ii), respectively). To identify gene-HERV fused transcripts, fragments mapped both on genes and HERVs were counted (iii).

In total, approximately 70 terabyte of RNA-Seq BAM files were analyzed. These RNA-Seq data were produced using the pair-ended and 48–50 bp sequencing. I mined transcriptome data of host genes, HERVs, and the gene-HERV fused transcripts for 5,550 patients across 12 TCGA projects analyzing solid tumors: bladder urothelial carcinoma (BLCA) [118, 119], breast invasive carcinoma (BRCA) [107, 108], colon adenocarcinoma (COAD) [120], head and neck squamous cell carcinoma (HNSC) [121], kidney chromophobe renal cell carcinoma (KICH) [122], kidney clear renal cell carcinoma (KIRC) [123], kidney renal papillary cell carcinoma (KIRP) [124], liver hepatocellular carcinoma (LIHC) [125], lung adenocarcinoma (LUAD) [126], lung squamous cell carcinoma (LUSC) [127], prostate adenocarcinoma (PRAD) [128], and thyroid carcinoma (THCA) [129]. Sample information is summarized in Table 9. Of these 5,550 patients, 590 patients had transcriptome data both for tumors and normal

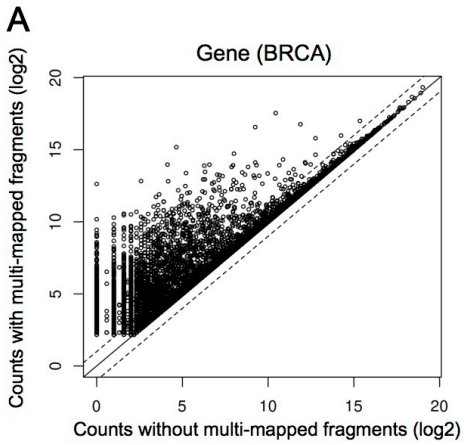
tissue controls (i.e., non-tumor tissues adjacent to tumors). The other patients only had the tumor data.

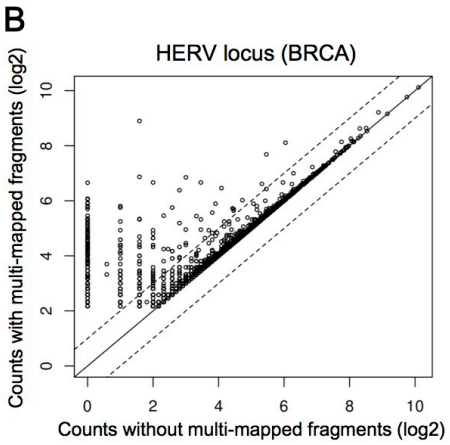
Table 9. Summary of RNA-Seq samples.

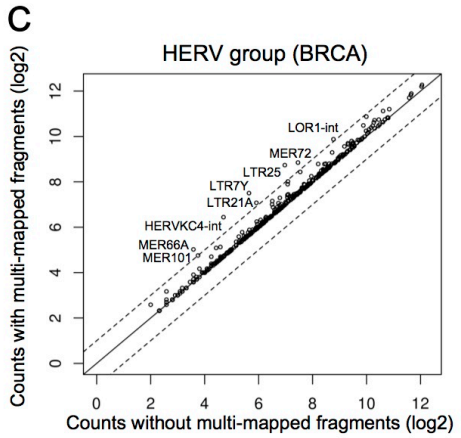
TCGA project (tumor type)	Abbreviation	Patients	Patients with normal tissue control
Breast invasive carcinoma	BRCA	1,091	112
Kidney renal clear cell carcinoma	KIRC	523	72
Lung adenocarcinoma	LUAD	515	57
Thyroid carcinoma	THCA	502	58
Lung squamous cell carcinoma	LUSC	501	49
Head and Neck squamous cell carcinoma	HNSC	501	43
Prostate adenocarcinoma	PRAD	496	52
Bladder urothelial carcinoma	BLCA	408	19
Liver hepatocellular carcinoma	LIHC	371	50
Kidney renal papillary cell carcinoma	KIRP	289	31
Colon adenocarcinoma	COAD	287	24
Kidney chromophobe renal cell carcinoma	KICH	66	23

In the NGS analysis focusing on repetitive sequences such as HERVs, it is necessary to handle carefully multi-mapped fragments, which are NGS fragments that can be mapped to two or more genomic regions [67, 130]. To examine effects of the multi-mapped fragments, I generated transcriptome data in the presence or absence of the multi-mapped fragments and then compared these two data (Fig. 28). Although the effect of the multi-mapped fragments on the HERV transcriptome was recognized, the effect was comparable with that of the gene transcriptome (Fig. 28). Thus, the effect of multi-mapped fragments on the HERV transcriptome is not so problematic in this

dataset. Therefore, I showed results only for transcriptome data excluding multi-mapped reads in the main figures of this paper.







D

Tumor type	Gene	HERV locus	HERV group
BLCA	12.0%	8.4%	1.0%
BRCA	**11.5%**	**8.6%**	**1.7%**
COAD	12.1%	7.7%	1.5%
HNSC	12.6%	14.6%	1.5%
KICH	11.6%	7.0%	1.2%
KIRC	11.4%	8.4%	1.7%
KIRP	11.9%	8.6%	1.9%
LIHC	11.0%	6.0%	1.1%
LUAD	11.4%	8.5%	1.0%
LUSC	11.8%	8.6%	1.0%
PRAD	12.1%	8.1%	1.7%
THCA	11.8%	8.1%	0.9%

Figure 28. Effects of including/excluding multi-mapped fragments in transcriptome analysis. A)–C) Comparisons of transcriptome data in the presence or absence of multi-mapped fragments. Results in BRCA are shown for genes (A), HERV loci (B), and HERV groups (C). The X-axis indicates fragment counts in the absence of multi-mapped fragments, and the Y-axis indicates the counts in the presence of multi-mapped fragments. The median values of the log2-transformed counts (with +1 pseudocounts) are shown. The lines with slope 1 and intercept 0 or ±1 are shown. Genes/HERVs with >2 fold differences in the counts are shown in the outside of the two dashed lines. D) Proportions of genes/HERVs in which fragment counts were >2 fold different in the presence or absence of multi-mapped fragments.

The transcriptome landscape of HERVs in tumors and normal tissues

To examine global transcription levels of HERVs in tumors and normal tissues, I calculated proportions of RNA-Seq fragments mapped on HERVs in each sample (Fig. 29A). In most tumor and normal tissue samples, proportions of the fragments mapped on HERVs were less than 0.5 % (Fig. 29A). In tumor samples of BLCA, COAD, HNSC,

KIRC, KIRP, LIHC, LUAD, LUSC, and PRAD, proportions of HERVs in total transcripts significantly increased compared to the normal tissues (Fig. 29A). Next, I identified transcribed HERV loci in tumor and normal tissues. In this study, I defined transcribed HERVs as HERV loci that were transcribed with >0.2 fragments per million (FPM) in >10% of samples in either of 12 TCGA projects. In this criterion, I identified 10,060 of transcribed HERV loci in 12 tumors and the normal tissue controls. Of these transcribed HERVs, an approximately half of them were present in upstream regions of transcription sites (TSSs), downstream regions of transcriptional terminal sites (TTSs), or intronic regions (Fig. 29B).

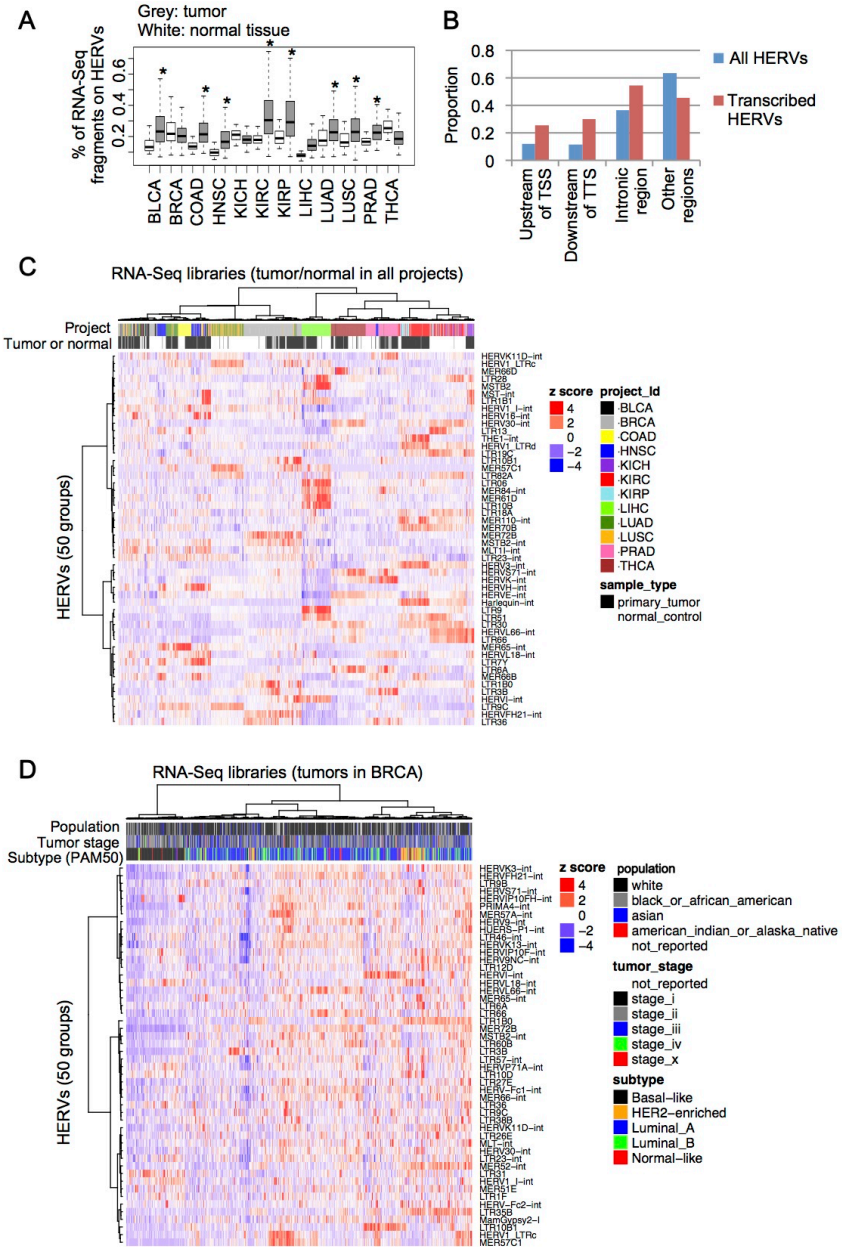


Figure 29. The landscape of HERV transcriptions in tumors and normal tissues. A) Global transcription levels of HERVs in tumors and normal tissues. Proportions of RNA-Seq fragments mapped on HERVs are shown. Results of tumors and normal tissues are respectively shown in grey and white. Outliers are not shown. Asterisks (*) denote significant up-regulation in tumors (adjusted $p < 0.05$). The adjusted p value was calculated using the Wilcoxon signed-rank test with Benjamini–Hochberg (BH) method. B) Proportions of all and transcribed HERVs in each category of genomic region. Results are shown for categories of upstream (<5kb) regions of transcription start sites (TSSs), downstream (<5kb) regions of transcription terminal sites (TTSs), and intronic regions. C) A heatmap showing relative transcription levels of HERVs in tumors and normal tissues. The heatmap was dealt with hierarchical clustering. A column indicates a RNA-Seq library. A row indicates a HERV group. Transcription levels were normalized as standardized (z) score in each row (HERV group). Fifty HERV groups with the highest variances were used in the analysis. Samples of patients having both of tumor and normal samples were only used. C) A heatmap showing relative transcription levels of HERVs in BRCA tumor samples.

To examine whether transcriptome of HERVs shows unique patterns depending on the tumor and tissue types, I performed unsupervised clustering analysis based on the HERV transcriptome information (Fig. 29C). In this analysis, I used the HERV group level transcriptome data, which was obtained by summing fragment counts of HERV loci in respective HERV group. Generally, samples clustered according to tumor and tissue types. For example, three kidney tumors (KICH, KIRC, and KIRP) separately clustered, whereas normal tissues of them (i.e., samples of normal kidney tissues) clustered together (Fig. 29C). TCGA classified a tumor type (e.g., BRCA) into molecular subtypes (e.g., Basal-like, HER2-enrichd, Luminal A/B, and Normal-like subtypes for BRCA) based on transcriptome and other information [36, 107, 108]. I investigated whether the HERV transcriptome shows unique patterns in each molecular subtype (Fig. 29D). In BRCA, samples of Basal-like, HERV2-enriched, and Luminal A/B were separated in unsupervised clustering based on the HERV transcriptome information (Fig. 29D). Thus, the HERV transcriptome showed unique patterns depending on the tumor and tissue types, and even the subtypes. This indicates that transcriptome of HERVs is informative for knowing cellular features as with that of host genes.

Identification of up/down-regulated HERVs in tumors

To identify specific HERVs that are up/down-regulated in tumors compared to the corresponding normal tissues (referred to as differentially expressed HERVs (deHERVs)), I performed differential expression analysis of HERVs with DESeq2 [131] (Fig. 30). I examined deHERVs at two levels; HERV locus and group levels. At the HERV locus level, 7,387 deHERVs were identified across 12 tumor types, which

corresponded to 73% of the transcribed HERV loci. At the HERV group level, 408 deHERV groups were identified. Although previous studies have emphasized the overexpression of HERVs in tumors [94], both of up/down-regulated HERVs in tumors were identified. Although expression levels of deHERVs were likely to be lower than those of differentially expressed genes (deGenes), some deHERVs such as HERVH-int/LTR7Y in HNSC showed relatively high expression levels (Figs. 30A and 30B). HERVH-int is the internal sequence of HERVH, and LTR7Y is a type of the LTR sequence of HERVH (there are four types of HERVH having distinct LTR sequences: LTR7, LTR7B, LTR7C, and LTR7Y). To compare deHERVs across tumors, I performed unsupervised clustering analysis based on fold change values in the differential expression analysis (Figs. 30C and 30D). Although deHERVs varied among tumors, some deHERVs were commonly identified in several tumors. Particularly, HERVH-int, LTR7Y, and HERVL18-int were up-regulated in broad ranges of tumors (Fig. 30D).

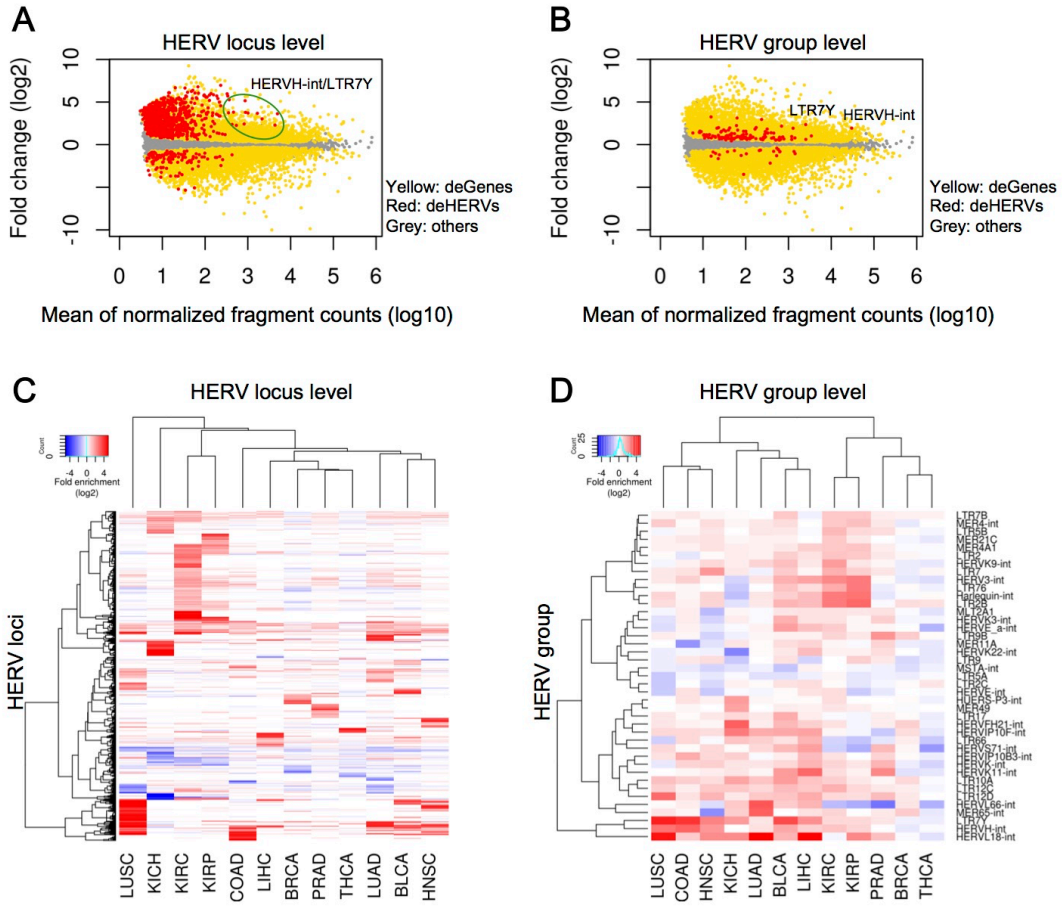
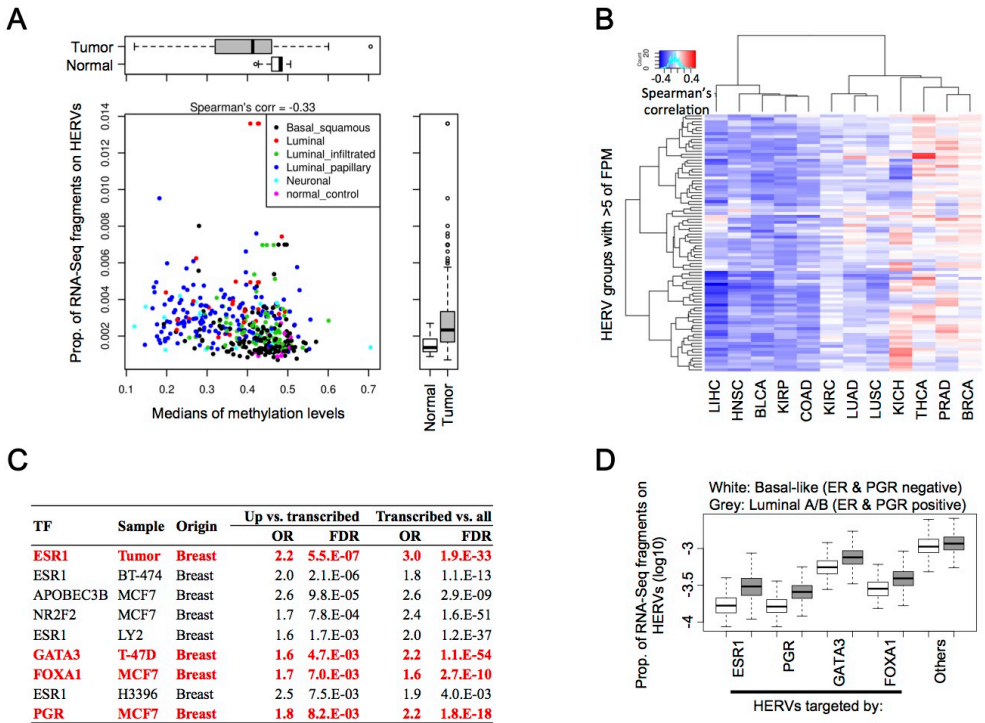


Figure 30. Differentially expressed HERVs (deHERVs) on tumors. A) and B) MA-plots showing genes and HERVs that were differentially expressed between tumor and normal tissue samples in HNSC. Results at the HERV locus and group levels are respectively shown in (A) and (B). The X-axis indicates mean transcription levels between tumors and normal tissues. The Y-axis indicates fold changes of transcription levels (tumors on normal tissues). Differentially expressed genes and HERVs are respectively shown in yellow and red (family-wise error rates (FWER) <0.05), and the others are shown in grey. C) and D) Heatmaps showing fold changes of deHERVs. Results at the HERV locus and group levels are respectively shown in (C) and (D). The heatmaps contain HERVs that were differentially expressed in ≥1 tumor types. In the heatmap (D), deHERV groups with high transcription levels (>20 FPM) in ≥1 tumor types are only shown.

Associations of HERV transcriptions and epigenetic signatures

DNA methylation particularly at CpG sites is essential to suppress transcriptions of HERVs and other retrotransposons in early embryos and matured tissues [100]. However, effects of the DNA methylation status on HERV transcription in tumors are poorly understood [100]. To clarify this point, I examined association of the overall transcription level of HERVs and the global DNA methylation level in each tumor type.

In case of HNSC, the HERV transcription level was negatively correlated with the global DNA methylation level (Spearman's correlation coefficient: -0.33) (Fig. 31A). Next, I examined associations of transcription levels of respective HERV groups and the global DNA methylation level in each tumor type (Fig. 31B). In case of HNSC, transcription levels of most HERV groups were negatively correlated with the global methylation level (Fig. 31B). Similar negative correlations were observed in LIHC, BLCA, KIRP, and COAD (Fig. 31B), suggesting that DNA methylation plays a role in the regulation of HERV transcriptions in these tumors. On the other hand, such negative correlations were not observed in KICH, THCA, PRAD, and BRCA (Fig. 31B), suggesting that DNA methylation does not play a major role in HERV regulations in these tumors.



TF	Sample	Origin	Up vs. transcribed		Transcribed vs. all	
			OR	FDR	OR	FDR
ESR1	**Tumor**	**Breast**	**2.2**	**5.5.E-07**	**3.0**	**1.9.E-33**
ESR1	BT-474	Breast	2.0	2.1.E-06	1.8	1.1.E-13
APOBEC3B	MCF7	Breast	2.6	9.8.E-05	2.6	2.9.E-09
NR2F2	MCF7	Breast	1.7	7.8.E-04	2.4	1.6.E-51
ESR1	LY2	Breast	1.6	1.7.E-03	2.0	1.2.E-37
GATA3	**T-47D**	**Breast**	**1.6**	**4.7.E-03**	**2.2**	**1.1.E-54**
FOXA1	**MCF7**	**Breast**	**1.7**	**7.0.E-03**	**1.6**	**2.7.E-10**
ESR1	H3396	Breast	2.5	7.5.E-03	1.9	4.0.E-03
PGR	**MCF7**	**Breast**	**1.8**	**8.2.E-03**	**2.2**	**1.8.E-18**

Figure 31. Associations of HERV transcriptions with epigenetic signatures. A) Association of the overall transcription level of HERVs and the global DNA methylation level in HNSC. A dot indicates a particular patient. The X-axis indicates the median value of methylation levels (beta values) among probes in the methylation array HumanMethylation450 (Illumina). The Y-axis indicates proportion of RNA-Seq fragments mapped on HERVs. B) Associations of transcription levels of respective HERV groups and the global methylation level in each tumor type. Colors in the heatmap indicate Spearman's correlation coefficients. HERV groups with >5 of FPM are only shown. C) TFBSs that were statistically enriched in the up-regulated/transcribed HERVs. Results for BRCA are shown. Results are shown for comparisons of up-regulated HERVs vs. transcribed HERVs and transcribed HERVs vs. all HERVs. Results are shown only for TFBSs with >1.5 of odds ratio (OR) and <0.01 of FDR in the both comparisons. D) Comparisons of transcription levels of HERVs targeted by ESR1, PGR, GATA3, and FOXA1 between BRCA subtypes with or without expressing ER and PGR. Basal-like subtype is ER and PGR-negative, whereas Luminal A/B subtypes are ER and PGR-positive. TF-binding statuses of HERVs were extracted from TFBS datasets shown in red in (C).

Transcription of HERVs is also affected by availability of specific transcription factors (TFs) [78, 80, 100]. To identify TFs responsible for the up-regulation of HERV transcriptions in tumors, I searched TFs whose binding sites were statistically enriched in up-regulated HERVs compared to transcribed HERVs (Fig. 31C and Table 10). In this analysis, I investigated 2,644 datasets of TF-binding sites (TFBSs) provided by ChIP-Atlas (http://chip-atlas.org/). In case of BRCA, TFBSs for ESR1 (estrogen receptor 1), PGR (progesterone receptor), GATA3, FOXA1, and others

were statistically enriched in up-regulated HERVs (Fig. 31C). All of these were TFBSs for ChIP-Seq whose experiments were performed in breast tumor/cell lines. ESR1 and PGR play a critical role in tumorigenesis of breast cancer [107, 108]. GATA3 and FOXA1 modulate ESR1 expression in breast cancer, and gain-of-function mutations of these genes are frequently observed in breast cancer [108].

Table 10. TFBSs enriched in enriched in up-regulated HERVs.

Project Id	TF	Sample	Sample class	Up vs. transcribed		Transcribed vs. al	
				OR	FDR	OR	FDR
BLCA	BRD4	K-562	Blood	3.9	2.6.E-03	4.4	4.6.E-1
BLCA	EP300	hESC H9	Pluripotent stem cell	1.9	8.0.E-04	1.6	3.0.E-1
BLCA	FOXA1	MDA-MB-453	Breast	1.7	5.5.E-03	2.7	1.3.E-10
BLCA	FOXA1	T-47D	Breast	1.8	4.0.E-04	2.1	1.0.E-5
BLCA	NR3C1	Unclassified	Unclassified	2.7	5.4.E-03	2.5	3.6.E-1
BLCA	TBP	K-562	Blood	2.2	1.7.E-04	3.3	1.8.E-6
BRCA	APOBEC3B	MCF-7	Breast	2.6	9.8.E-05	2.6	2.9.E-0
BRCA	ESR1	BT-474	Breast	2.0	2.1.E-06	1.8	1.1.E-1
BRCA	ESR1	H3396	Breast	2.5	7.5.E-03	1.9	4.0.E-0
BRCA	ESR1	LY2	Breast	1.6	1.7.E-03	2.0	1.2.E-3
BRCA	ESR1	Tumour tissues	Others	2.2	5.5.E-07	3.0	1.9.E-3
BRCA	FOXA1	MCF-7	Breast	1.7	7.0.E-03	1.6	2.7.E-1
BRCA	GATA3	T-47D	Breast	1.6	4.7.E-03	2.2	1.1.E-5
BRCA	NR2F2	MCF-7	Breast	1.7	7.8.E-04	2.4	1.6.E-5
BRCA	PGR	MCF-7	Breast	1.8	8.2.E-03	2.2	1.8.E-1
COAD	ATF2	LoVo	Digestive tract	1.7	2.6.E-03	5.4	3.8.E-6
COAD	CDX2	LS-180	Digestive tract	2.2	1.7.E-08	3.2	5.2.E-4
COAD	CEBPB	LS-180	Digestive tract	1.5	2.0.E-03	2.5	6.2.E-5
COAD	FOXA1	A549	Lung	1.6	7.6.E-03	1.9	6.6.E-1
COAD	FOXA1	hESC derived mesendodermal cells	Pluripotent stem cell	1.5	2.1.E-03	1.7	1.2.E-2
COAD	FOXA1	MCF-7	Breast	2.0	5.9.E-08	1.9	4.8.E-1
COAD	FOXA1	VCaP	Prostate	1.5	7.4.E-04	1.7	7.9.E-2
COAD	FOXA2	A549	Lung	1.6	7.0.E-04	2.2	1.1.E-3
COAD	FOXA2	hESC derived mesendodermal cells	Pluripotent stem cell	2.0	6.0.E-06	2.0	1.1.E-1
COAD	FOXP1	LoVo	Digestive tract	2.2	3.3.E-05	3.8	1.0.E-2
COAD	KLF5	GP5d	Digestive tract	2.5	1.9.E-03	4.7	3.2.E-1
COAD	KLF5	KATOIII	Digestive tract	1.8	7.5.E-05	5.2	1.2.E-7
COAD	NR1H3	HT-29	Digestive tract	1.6	3.4.E-03	2.9	6.0.E-4

COAD	RUNX1	LoVo	Digestive tract	Inf	5.9.E-03	17.1	7.3.E-03
COAD	RUNX2	C4-2	Blood	1.6	3.2.E-03	2.4	2.4.E-37
COAD	RXRA	LS-180	Digestive tract	2.9	7.6.E-06	5.7	3.7.E-34
HNSC	5-mC	MCF-10A	Breast	2.2	1.9.E-03	2.4	4.8.E-08
HNSC	5-mC	MCF-7	Breast	1.7	3.0.E-03	2.0	3.0.E-14
HNSC	BRD4	MCF-10A	Breast	6.4	9.3.E-12	11.3	1.0.E-24
HNSC	CDX2	LS-180	Digestive tract	1.9	2.7.E-03	2.6	1.2.E-20
HNSC	EP300	hESC H9	Pluripotent stem cell	1.6	1.6.E-03	2.0	1.5.E-29
HNSC	FOXA1		Pluripotent stem cell	1.6	7.5.E-04	2.0	3.8.E-31
HNSC	FOXA2	A549	Lung	1.6	3.4.E-03	2.7	1.2.E-54
HNSC	FOXA2	hESC derived mesendodermal cells	Pluripotent stem cell	2.4	6.1.E-10	2.4	3.2.E-26
HNSC	PAX5	OCI-LY-7	Blood	2.4	2.3.E-03	3.3	6.4.E-10
HNSC	RXRA	LS-180	Digestive tract	3.4	6.9.E-09	8.7	2.0.E-56
HNSC	SFPQ	MCF-7	Breast	1.6	5.6.E-03	1.9	2.0.E-49
KIRC	ARNT	786-O	Kidney	5.3	4.9.E-04	12.4	1.9.E-22
KIRC	ARNTL	786-O	Kidney	3.2	2.5.E-04	7.5	2.9.E-32
KIRC	EPAS1	786-O	Kidney	2.7	1.4.E-05	5.6	4.3.E-36
KIRC	RBPJ	B-Lymphocytes	Blood	3.3	5.6.E-08	3.8	1.0.E-25
KIRP	ARNT	786-O	Kidney	5.9	3.2.E-07	10.8	2.4.E-17
KIRP	ARNTL	786-O	Kidney	3.7	1.5.E-06	6.5	1.4.E-23
KIRP	BATF	GM12878	Blood	1.7	8.2.E-06	2.0	3.0.E-27
KIRP	EPAS1	786-O	Kidney	3.3	3.6.E-09	5.1	2.5.E-28
KIRP	FOXA1	CFPAC-1	Adipocyte	1.5	9.5.E-03	1.8	3.6.E-18
KIRP	NR3C1	ECC-1	Uterus	1.7	5.5.E-03	3.4	8.5.E-45
KIRP	RBPJ	B-Lymphocytes	Blood	3.6	9.3.E-12	4.1	7.9.E-28
KIRP	SPI1	DOHH-2	Blood	1.6	1.5.E-06	2.0	4.9.E-34
KIRP	SPI1	GM12878	Blood	1.5	1.7.E-04	1.9	3.0.E-32
LUAD	ETV5	LoVo	Digestive tract	1.8	2.1.E-03	2.8	4.2.E-22
LUAD	FOSL2	A549	Lung	1.6	2.0.E-04	2.1	9.2.E-26
LUAD	FOXA1	A549	Lung	1.6	2.3.E-04	1.9	5.9.E-26
LUAD	FOXA1	CFPAC-1	Adipocyte	1.8	1.3.E-09	2.0	2.6.E-30
LUAD	JUND	Calu-3	Lung	1.9	2.9.E-03	1.8	2.4.E-04
LUAD	JUND	HT-29	Digestive tract	1.7	2.6.E-04	2.0	1.2.E-14
LUAD	KLF5	CFPAC-1	Adipocyte	1.7	1.1.E-05	3.4	1.9.E-64
LUAD	KLF5	KATOIII	Digestive tract	1.6	2.7.E-03	3.8	2.9.E-53
LUAD	NKX2-1	NCI-H441	Lung	1.6	4.4.E-07	2.6	2.6.E-66
LUAD	NR1H3	HT-29	Digestive tract	1.6	4.4.E-04	2.6	6.0.E-41
LUAD	PPARG	HT-29	Digestive tract	1.6	6.8.E-03	2.3	2.3.E-24
LUAD	SMAD3	NCI-H441	Lung	1.7	1.2.E-06	2.4	2.0.E-44
LUAD	STAT3	HCC1143	Breast	1.8	3.8.E-03	2.1	9.2.E-12
LUAD	USF2	HeLa	Uterus	1.8	8.7.E-05	2.2	1.4.E-18
LUSC	5-mC	MCF-7	Breast	1.5	1.2.E-03	2.0	8.9.E-27
LUSC	CDX2	LS-180	Digestive tract	1.6	6.5.E-03	2.0	9.5.E-17
LUSC	RXRA	LS-180	Digestive tract	3.0	1.8.E-08	5.3	5.5.E-40
PRAD	AR	Prostate cancer	Prostate	2.4	1.0.E-03	2.9	9.0.E-12

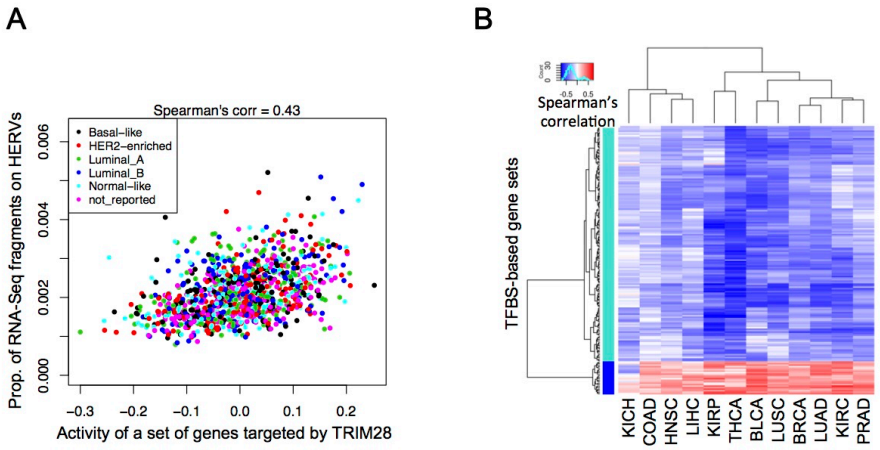
PRAD	AR	VCaP	Prostate	1.7	1.9.E-03	2.3	1.5.E-2
PRAD	GATA2	LNCAP	Prostate	1.7	2.1.E-03	2.0	1.0.E-2
PRAD	PIAS1	VCaP	Prostate	2.0	1.4.E-03	2.5	5.0.E-1
PRAD	RXRA	HepG2	Liver	1.7	3.2.E-03	3.3	7.0.E-4

Breast cancer is composed of several molecular subtypes showing distinct expression statuses of estrogen receptor (ER) and PGR: Basal-like (a.k.a., Triple-Negative) subtype expresses ER and PGR, whereas Luminal A/B subtypes do not [107, 108]. Subtypes expressing ER and PGR (Luminal A/B) are major group in BRCA (Fig. 29D) [107, 108], and these subtypes also highly expressed GATA3 and FOXA1 [108]. To examine effects of the availability of these four TFs on HERV transcriptions in BRCA, I compared transcription levels of HERVs targeted by these four TFs between Basal-like and Luminal A/B subtypes (Fig. 31D). Transcription levels of HERVs targeted by any of these four TFs in Luminal A/B subtypes were higher than those in Basal-like subtype (Fig. 31D). Meanwhile, transcription levels of HERVs that were not targeted by any of these four TFs were comparable between the two groups (Fig. 31D). Thus, HERVs targeted by these four TFs showed transcription patterns consistent with the availability of these TFs. Taken together, ESR1, PGR, GATA3, and FOXA1 were likely to be critical for up-regulation of HERV transcriptions in BRCA Luminal A/B subtypes. In many tumor types including BRCA (BLCA, BRCA, COAD, HNSC, KIRP, and LUAD), TFBSs for FOXA1 were statistically enriched in up-regulated HERVs (Table 10). This indicates importance of this TF for the up-regulation of HERVs in a broad range of tumor types.

Regulatory axes associated with HERV transcriptions

To gain further insights into regulatory mechanisms of HERV transcriptions in tumors, I searched TFs whose activities were associated with HERV transcriptions. Activities of TFs can be estimated from transcriptome data by examining transcription levels of sets of genes targeted by the TFs [132]. To perform this estimation, TFBS-based gene set,

which is a set of genes targeted by a particular TF in a certain condition, was defined using TFBS datasets in ChIP-Atlas (http://chip-atlas.org/). Transcriptional activities of the TFBS-based gene sets in each sample were measured with gene set variation analysis (GSVA) [133], and then associations of the activities and the overall HERV transcription level in each tumor type were examined. As a representation, result for the HERV transcription level and the activity of the set of genes targeted by TRIM28 in BRCA is shown in Fig. 32A. In this case, positive correlation was observed (Spearman's correlation coefficients: 0.43) (Fig. 32A). Similarly, comparisons were performed for all combinations of TFBS-based gene sets and HERV transcription levels in respective tumor types. To identify TFBS-based gene sets associated with HERV transcriptions across tumor types, TFBS-based gene sets that showed correlations with HERV transcriptions in most tumor types were extracted (Figs. 32B and 32C). Transcriptional activities of genes targeted by TRIM28, SETDB1, and ZNF274 were positively correlated with transcription levels of HERVs in most tumor types (Fig. 32C). These three are known to form a complex and suppress transcriptions of HERVs and other retrotransposons in early embryos [9, 102-106]. Taken together, these three were likely to play a role in HERV regulation in tumors as well as early embryos. Additionally, several epigenetic modifiers such as MDB2, BRD3/4, and FANCD2 were identified in the analysis (Fig. 32C). BRD4 was also identified in a search of TFs whose binding sites were enriched in up-regulated HERVs in BLCA and HNSC (Table 10), suggesting importance of this epigenetic modifier for the regulation of HERV transcriptions in tumors.



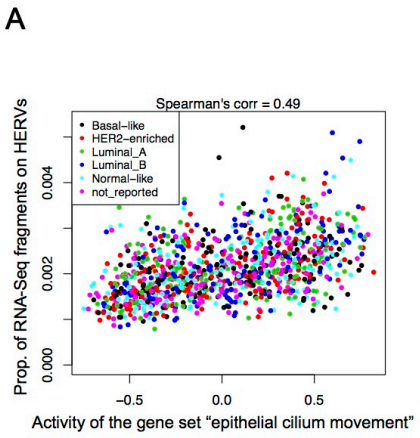
C

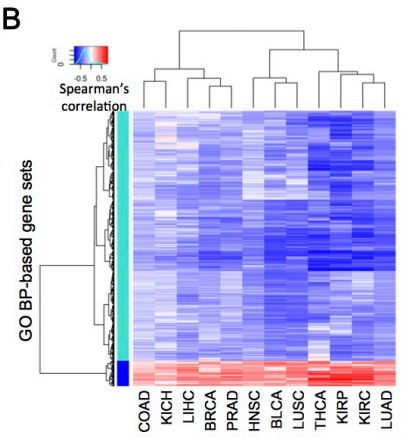
TF	Sample	Origin	Median of Spearman's correlation
MBD2	HMEC	Breast	0.41
TRIM28	H1-hESC	ES cells	0.40
BRD4	GM12878	Blood	0.38
BRD3	293	Kidney	0.32
ZNF274	H1-hESC	ES cells	0.32
CNOT3	Tumor tissue	Digestive tract	0.32
PARK7	SH-SY5Y	Neural	0.32
BRD4	T-47D	Breast	0.32
FANCD2	293	Kidney	0.30
DYRK1A	HeLa	Uterus	0.30
SETDB1	WM853	Epidermis	0.29
SPDEF	MCF-7	Breast	0.29
BRD4	MCF-7	Breast	0.28
RELA	Lymphoblast	Blood	0.27
SMAD1	H1-hESC	ES cells	0.27

Figure 32. Estimation of regulatory axes associated with HERV transcriptions. A) Association of the overall transcription level of HERVs and the transcriptional activity of a set of genes targeted by TRIM28 in H1-hESC. Results for BRCA are shown. The X-axis indicates transcriptional activity (normalized enrichment score (NES)) of the gene set calculated with gene set variation analysis (GSVA). B) Associations of the overall HERV transcription level and transcriptional activities of TFBS-based gene sets in each tumor type. Colors in the heatmap shows Spearman's correlation coefficients. A row indicates a TFBS-based gene set. Results are shown only for TFBS-based gene sets with >0.25 of absolute values of medians of Spearman's correlation coefficients among tumor types. C) TFBS-based gene sets whose transcriptional activities were positively correlated with HERV transcription levels in most tumor types. Top 15 TFBSs are shown with respect to the median value of Spearman's correlation coefficients among tumor types.

Biological pathways associated with HERV transcriptions

To find tumor characteristics associated with HERV transcriptions, I investigated biological processes whose transcriptional activities were correlated with HERV transcriptions (Fig. 33). I used an approach similar to the one described in the last section: For all gene sets in the category of biological process (BP) in Gene Ontology (GO), transcriptional activities of the gene sets were calculated, and then correlations between the activities and the overall transcription level of HERVs were examined in each tumor type (for example, Fig. 33A). Gene sets that were correlated with HERV transcriptions in several tumor types were extracted (Figs. 33B, 33C, and 33D). GO terms relating to the epigenetic regulation were positively correlated with HERV transcriptions in most tumor types, supporting importance of the epigenetic regulation on HERV transcriptions (Fig. 33C). In addition to these, GO terms relating to the cilium development, centriole assembly, inner dynein arm assembly, and mitochondrion distribution were positively correlated with HERV transcriptions in most tumor types (Fig. 33C). All of these processes are microtubule/tubulin-mediated biological processes. Meanwhile, GO terms relating to the apoptosis, immune response (e.g., antigen processing and presentation) were negatively correlated with HERV transcriptions in most tumor types (Fig. 33D).





C

GO (BP)	Median of Spearman' s correlation
Cilium or flagellum dependent cell motility	0.49
Centriole assembly	0.47
Epithelial cilium movement	0.47
Sulfate transport	0.46
Inner dynein arm assembly	0.45
Photoreceptor cell development	0.44
Eye photoreceptor cell development	0.44
Cilium movement	0.44
Peptidyl lysine methylation	0.43
Photoreceptor cell maintenance	0.43
Histone deubiquitination	0.42
Mitochondrion distribution	0.41
DNA alkylation	0.41
Oxalate transport	0.40
DNA methylation or demethylation	0.40

D

GO (BP)	Median of Spearman' s correlation
Regulation of apoptotic signaling pathway	-0.64
Regulation of proteolysis	-0.61
Negative regulation of apoptotic signaling pathway	-0.59
Protein maturation	-0.58
Non-canonical WNT signaling pathway	-0.57
Negative regulation of transferase activity	-0.57
Antigen processing and presentation	-0.57
Negative regulation of canonical WNT signaling pathway	-0.56
Cell death	-0.55
Negative regulation of protein modification process	-0.55
Regulation of establishment of planar polarity	-0.55
Positive regulation of proteolysis	-0.55
Fc receptor signaling pathway	-0.54
Negative regulation of catalytic activity	-0.54
Positive regulation of apoptotic signaling pathway	-0.54

Figure 33. Biological functions associations with HERV transcriptions. A) Association of the overall transcription level of HERVs and the transcriptional activity of the gene set "epithelial cilium movement". This gene set is collected in the category of biological process (BP) in Gene Ontology (GO). B) Associations of overall HERV transcription level and transcriptional activities of gene sets in each tumor type. Colors in the heatmap show Spearman's correlation coefficients. A row indicates a GO BP-based gene set. Results are shown only for gene sets with >0.3 of absolute values of medians of Spearman's correlation coefficients among tumor types. C) and D) Gene sets whose transcriptional activities showed positive (C) or negative (D) correlations with HERV transcription levels in most tumor types. Top (C) or worst (D) 15 GO terms are shown with respect to the median value of Spearman's correlation coefficients among tumor types.

Gene-HERV fused transcripts in tumors

Some HERVs are transcribed as a part of mRNA of genes and formed non-canonical (and unannotated) transcripts in tumors (and even in normal tissues) [109]. As described in Fig. 27B, I identified such gene-HERV fused transcripts by counting RNA-Seq fragments mapped both on genes and HERVs. To identify the fused transcripts that were up/down-regulated in tumors, I performed differential expression analysis using

count data of the fragments mapped on junctions of the fused transcripts. In total, 1,410 gene-HERV fused transcripts were differentially expressed in ≥1 tumor types. To examine whether genes forming the fused transcripts are statistically enriched in known gene sets, I performed GO enrichment analysis. Genes forming the fused transcripts were statistically enriched in genes relating to metabolic processes of lipid or other small molecules (Table 11).

Table 11. GO enrichment analysis using genes forming the gene-HERV fused transcripts.

Project	GO term (biological process)	Odds ratio	Adjusted p value
COAD	Organonitrogen compound metabolic process	3.9	7.E-03
KICH	Cofactor metabolic process	6.2	3.E-02
KIRC	Fatty acid metabolic process	6.6	5.E-02
LUSC	Lipid metabolic process	4.8	4.E-03
PRAD	Lipid metabolic process	6.7	2.E-04
PRAD	Small molecule metabolic process	5.3	1.E-03
PRAD	Single organism biosynthetic process	6.1	1.E-03
PRAD	Lipid biosynthetic process	7.8	3.E-03
PRAD	Cellular lipid metabolic process	5.4	4.E-03
PRAD	Organonitrogen compound metabolic process	4.7	4.E-03
PRAD	Cellular catabolic process	5.3	6.E-03
PRAD	Small molecule biosynthetic process	6.5	9.E-03
PRAD	Organic acid metabolic process	4.6	9.E-03
PRAD	Monocarboxylic acid metabolic process	4.9	2.E-02
PRAD	Catabolic process	3.4	2.E-02
PRAD	Organic hydroxy compound metabolic process	4.3	2.E-02
PRAD	Single organism catabolic process	3.6	2.E-02
THCA	Establishment of protein localization	4.6	2.E-02
THCA	Protein localization	3.9	2.E-02
THCA	Establishment of localization in cell	4.0	2.E-02

GO terms with >2 of odds ratio and <0.05 of adjusted p value are shown.

Next, I examined whether cancer-associated genes are included in the list of genes forming the fused transcripts that were differentially expressed in tumors. I particularly focused on 567 cancer causal genes annotated by Cancer Gene Census [134]. A total of 11 gene-HERV fused transcripts was identified across 12 tumor types (Fig. 34A). Of these, the fused transcript of ERVL-B4-int (chr21:41,511,218-41,512,860) and TMPRSS2 were approximately 10 times up-regulated in tumor samples of PRAD (Fig. 34A). TMPRSS2 is a major causal gene of prostate cancer [128]. In PRAD, FPM values (relative transcriptions) of the ERVL-B4-int locus and the junction of the fused transcript were highly up-regulated in tumors, whereas that of TMPRSS2 was moderately up-regulated (Fig. 34B). The ERVL-B4-int locus was located on the upstream region of the constitutive TSS of TMPRSS2 (Fig. 34C). The unannotated exon on the ERVL-B4-int locus had 3′-splice junctions much more than 5′-splice junctions. The transcription level of this exon was up-regulated in the tumor sample (Fig. 34C). Taken together, the ERVL-B4-int locus worked as an alternative TSS of TMPRSS2, and the activity of this TSS was up-regulated in tumors.

A

Tumor type	Gene	HERV	Fold change (log2)	Adjusted p value
LUAD	ATF1	MER41-int	-2.71	4.E-07
LUSC	ATF1	MER41-int	-3.79	1.E-22
HNSC	CBLB	THE1B-int	1.45	4.E-06
BRCA	CBLB	THE1B-int	1.06	9.E-10
LUSC	CBLB	THE1B-int	1.08	1.E-05
KIRP	CBLB	THE1B-int	-1.11	2.E-02
BRCA	CDH11	MER41-int	1.01	4.E-04
LUSC	CDH11	MER41-int	-1.16	6.E-05
LIHC	FANCC	MER21C	-3.84	2.E-25
BRCA	FHIT	LTR26E	-1.02	6.E-08
HNSC	FHIT	LTR26E	-1.09	5.E-02
LIHC	FHIT	LTR26E	1.00	3.E-03
LUAD	FLT4	LTR13A	-1.66	4.E-26
BRCA	FLT4	LTR13A	-1.62	3.E-31
LUSC	FLT4	LTR13A	-2.35	2.E-35
KIRP	FLT4	LTR13A	-1.41	2.E-06
COAD	HLA-A	ERV3-16A3_I-int	-1.36	3.E-02
LUSC	HLA-A	ERV3-16A3_I-int	-1.29	3.E-04
HNSC	POU5F1	Harlequin-int	1.69	8.E-04
KIRC	POU5F1	Harlequin-int	3.48	2.E-17
KIRC	POU5F1	Harlequin-int	1.70	2.E-29
BLCA	POU5F1	Harlequin-int	1.51	4.E-02
KICH	POU5F1	Harlequin-int	-2.62	6.E-06
HNSC	PTK6	MER31-int	-2.10	5.E-05
LUSC	ROS1	LTR12C	-2.87	1.E-19
PRAD	**TMPRSS2**	**ERVL-B4-int**	**3.33**	**1.E-15**

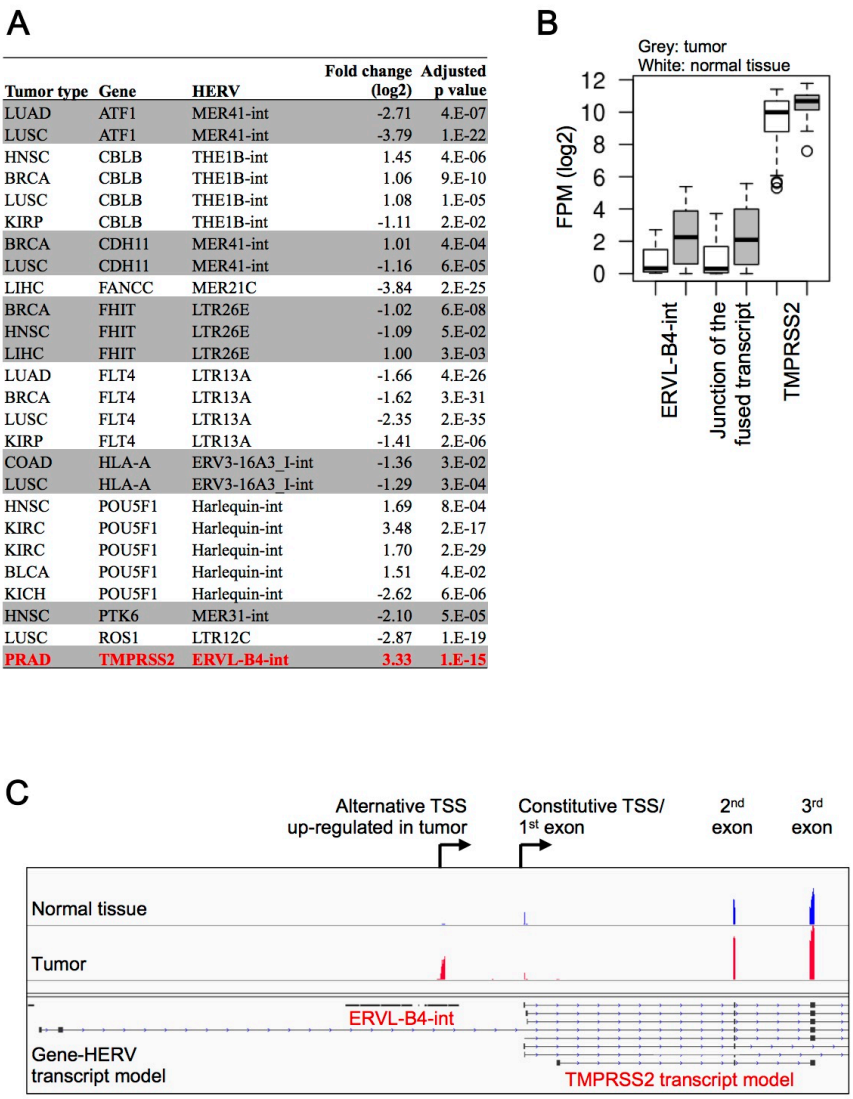


Figure 34. The HERV-derived alternative TSS of TMPRSS2 in prostate adenocarcinoma. A) Gene-HERV fused transcripts differentially expressed in tumors. Results are shown only for transcripts fused with cancer causal genes annotated by Cancer Gene Census. The fused transcript of ERVL-B4-int (chr21: 41,511,218-41,512,860) and TMPRSS2 in PRAD is highlighted in red. B) Relative transcription levels of the fused transcript of ERVL-B4-int and TMPRSS2. Results in PRAD are shown. FPM values of the ERVL-B4-int locus, the junction of the fused transcript, and TMPRSS2 are shown. Results of tumors and normal tissues are respectively shown in grey and white. C) A genome browser snapshot of the fused transcript. Results in PRAD are shown. Integrated Genome Viewer (IGV) shows read depths of RNA-Seq with the gene-HERV transcript model. Tumor and normal tissue data of a particular patient (case ID: TCGA-EJ-7331) is shown. The ERVL-B4-int locus was located on the upstream region of the constitutive TSS of TMPRSS2 and worked as an alternative TSS in tumor.

Discussion

Although importance of the HERV reactivation in tumors has been emphasized [94, 95], previous TCGA studies have not approached information on HERV transcriptome in tumor tissues. This is because the widely used gene transcript models do not contain the HERV information. In the present study, I constructed the transcript model containing the HERV information and produced HERV transcriptome data by reanalyzing the TCGA RNA-Seq dataset (Fig. 27). I identified a large number of transcripts derived from HERVs in tumors and normal tissues. Importantly, these were unannotated transcripts because I only focused on HERVs that were not overlapped with known transcripts (Fig. 27A). Based on the HERV transcriptome information, I carried out a pan-cancer analysis focusing on HERV transcriptions.

Previous studies examining transcriptions of HERVs in tumors investigated only a limited number of HERV loci or tumor types (or cell lines) [32, 96, 101]. In the present study, I comprehensively identified the HERV-derived transcripts and depicted the landscape of HERV transcriptions in 12 solid tumors and the corresponding normal tissues (Fig. 29). The HERV transcriptome showed unique profiles according to tumor and tissue types, and even the molecular subtypes (Figs. 29C and 29D). Thus, transcriptome of HERVs is informative as with that of genes.

Mechanisms of the up-regulation of HERV transcriptions in tumors have not been clarified. In the present study, I investigated associations of HERV transcriptions and epigenetic signatures in tumors (Fig. 31). In some tumor types such as HNSC and LIHC, HERV transcriptions were negatively correlated with the global DNA methylation levels (Figs. 31A and 31B). Meanwhile, in other tumor types, such negative correlations were not observed (Fig 31B). Thus, global DNA methylation levels are likely to be not necessary associated with HERV transcriptions in tumors. I showed that several TFs were associated with up-regulation of HERV transcriptions in tumors (Figs. 31C and 31D). In case of BRCA Luminal A/B subtypes, the transcriptional

up-regulation of HERVs was associated with ESR1, PGR, GATA3, and FOXA1 (Figs. 31C and 31D), which are overexpressed in the subtypes [108]. Thus, the overexpression of TFs was likely to play an essential role in the transcriptional up-regulation of HERVs in tumors. The reason why up-regulated HERVs were different among tumor types (Figs. 29 and 30) is probably that overexpressed TFs differ among tumor types. Furthermore, I showed that overall transcription levels of HERVs were positively correlated with transcriptional activities of sets of genes targeted by ZNF274, TRIM28, and SETDB1 (Fig. 32). Although these three are known transcriptional suppressors against HERVs working in early embryos [9, 102-106], it has not been clarified whether these suppressors work in tumors. My result suggests that these suppressors work in tumors as well as early embryos. Similarly, other epigenetic modifiers such as MBD2 and BRD3/4 also showed positive correlations with HERV transcriptions (Fig. 32G). These epigenetic modifiers are candidates of a novel transcriptional regulator of HERVs working in tumors.

I identified a large number of gene-HERV fused transcripts in tumors (and even in normal tissues). This indicates the presence of abundant non-canonical and unannotated gene transcripts containing HERV sequences. Of these, I particularly focused on the fused transcript of ERVL-B4-int and TMPRSS2 identified in PRAD (Fig. 34). TMPRSS2 is a causal gene of prostate cancer [128]. In prostate tumorigenesis, TMPRSS2 causes the overexpression of ERG, a ETS transcription factor, through gene fusion [128]. Indeed, the TMPRSS2-ERG fusion is the most frequently observed gene alteration in prostate cancer [128]. In TMPRSS2, the ERVL-B4-int locus worked as an alternative TSS that was highly up-regulated in tumors (Fig. 34B and 34C). Together, I consider that the ERVL-B4-int locus plays an essential role in the tumorigenesis in prostate cancer through the up-regulation of expressions of TMPRSS2 and ERG. Further investigation based on molecular works (i.e., the knock out experiment with CRISPR-Cas9 system) is needed to prove the essentiality of the ERVL-B4-int locus in

the tumorigenesis.

In summary, I identified unannotated HERV-derived transcripts in tumors and normal tissues. These transcripts seemed to be modulated as a part of the gene regulatory network. Furthermore, I generated a resource of HERV transcriptome for thousands of tumors. Thus, this study provides fundamental information to understand impacts of transcripts derived from HERVs in tumors.

Materials and Methods

Ethical approval

The utilization of TCGA RNA-Seq data was authorized by dbGaP (http://dbgap.ncbi.nlm.nih.gov) as the following project: Systematic identification of reactivated human endogenous retroviruses in cancers (#15126).

The construction of the gene-HERV transcript model

As a gene transcript model, GENCODE Version 22 was downloaded from the GENCODE website (http://www.gencodegenes.org/). The model is for GRCh38/hg38. From the gene transcript model, "retained intron"-type transcripts were excluded. As a HERV transcript model, the RepeatMasker output file (15-Jan-2014) was downloaded from the UCSC genome browser (http://genome.ucsc.edu/). This is an annotation of repetitive sequences including HERVs in the human reference genome (GRCh38/hg38). From the HERV transcript model, unreliable HERV loci (Smith-Waterman (SW) score <2,500) were excluded. HERV sequences overlapping with known transcripts in GENCODE Version 22 were also excluded. The gene-HERV transcript model was constructed by merging the gene and HERV transcript models. The gene-HERV transcript model includes 60,483 protein-coding/non-coding genes and 138,124 HERV loci.

Extracting transcriptome information of HERVs from TCGA RNA-Seq data

Poly A-enriched RNA-Seq (mRNA-Seq) data was provided by TCGA. Of the RNA-Seq data, I only analyzed the data produced by the paired-ended and 48–50 bp sequencing. Of 33 TCGA projects (tumor types), 12 solid tumors were analyzed. In these tumor types, RNA-Seq data of tumors and normal tissue controls was provided for >20 patients (Table 9).

NGS read alignment (BAM) file of TCGA RNA-Seq was downloaded from GDC data portal (http://portal.gdc.cancer.gov/) using GDC Data Transfer Tool

(http://gdc.cancer.gov/access-data/gdc-data-transfer-tool/). This BAM file is for GRCh38/hg38. This BAM file was generated using the GDC mRNA analysis pipeline (http://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/). This pipeline used STAR [135] with the 2-pass mapping mode, which is the most sensitive setting to align RNA-Seq reads on unannotated transcripts.

To obtain transcriptome data of HERVs and genes, RNA-Seq fragments mapped on HERVs and known transcripts of genes were counted using featureCounts in subread package [136] with the BAM file and the gene-HERV transcript model. The "fracOverlap" argument, which is the minimum fraction of overlapping bases in a fragment that is required for read assignment, was set at 0.25. Fragments assigned to more than one features were discarded (the "allowMultiOverlap" option was not used). In case of generating transcriptome data including multi-mapped fragments, the "primaryOnly" option was added. In this setting, a multi-mapped fragment is assigned to one genomic position with the best alignment score (in case of a tie, one genomic position is randomly chosen from the positions with the best score). The mean and standard deviation of assigned RNA-Seq fragments were 57.0 and 15.9 million, respectively.

To obtain transcriptome data of the gene-HERV fused transcripts, RNA-Seq fragments mapped both on HERVs and known transcripts of genes were identified with featureCount [136]. To this aim, the "allowMultiOverlap" and "reportReads SAM" options were added to the setting mentioned in the above paragraph. In this setting, a NGS read alignment (SAM) file is generated with an extra column reporting assigned features for each NGS read. In the outputted SAM file, fragments assigned both to HERVs and genes were counted using the original python script.

The quality control of each RNA-Seq library was performed. Regarding the proportion of non-assigned fragments (fragments mapped not on HERVs nor transcripts of genes), outliers were recursively detected using Smirnov-Grubbs test (the threshold

was set at 0.05). The outlier RNA-Seq libraries were not used in the downstream analysis.

Unsupervised clustering based on the HERV transcriptome information

The analysis was performed in R. Transcriptome information at the HERV group level was used. The count data of RNA-Seq fragments was normalized using variance-stabilizing transformation in DESeq2 [131]. In this step, the size factor of each RNA-Seq library was also normalized. In the analysis, 50 HERV groups with the highest variance were used. Euclid distances among RNA-Seq libraries or HERV groups were calculated. Hierarchical clustering was performed with Ward method. The heatmap was created using the ComplexHeatmap package [137]. Clinical and other information of each patient (gender, population, tumor stage) was downloaded from GDC data portal (http://portal.gdc.cancer.gov/). The molecular subtype information of each tumor was obtained using the TCGAbiolinks package [138].

Differential expression analysis of genes and HERVs

The analysis was performed in R. The analysis was conducted in each tumor type. RNA-Seq data of patients having both of tumor and the normal tissue control data was used. The paired comparison of tumor and the normal tissue control data performed using DESeq2 [131]. Genes or HERVs with >2 of fold change and <0.05 of family-wise error rate (FWER) were regarded as differentially expressed genes or HERVs. The FWER was calculated by Bonferroni correlation.

Differential expression analysis of the gene-HERV fused transcripts was performed by the procedure similar to the above. As a difference, I used count data of RNA-Seq fragments mapped both on HERVs and known transcripts of genes.

DNA methylation data analysis

DNA methylation data for tumors and normal tissue controls were obtained from GDC data portal (http://portal.gdc.cancer.gov/) using GDC Data Transfer Tool (http://gdc.cancer.gov/access-data/gdc-data-transfer-tool/). These data are text files describing the methylation level (beta value; proportion of methylated CpG at a CpG site) of each probe in the methylation micro array HumanMethylation450 (Illumina). After removing probes overlapping with single nucleotide polymorphisms (SNPs) that have >0.05 of minor allele frequency, the median value of beta values of probes were calculated for using the downstream analysis.

Identification of TFBSs enriched in up-regulated/transcribed HERVs

The 2,644 TFBS datasets were downloaded from ChIP-Atlas (http://chip-atlas.org/). ChIP-Seq peaks (TFBSs) with <1.0E-5 of MAC2 Q-value were used. Since TFBS datasets provided by ChIP-Atlas are for GRCh37/hg19, genome coordinates of these TFBSs were converted for GRCh38/hg38 using UCSC liftOver (http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/liftOver). In each category of up-regulated HERVs, transcribed HERVs, and all HERVs, number of TFBSs present within 5kb from HERVs was counted with bedtools slop [84]. Using these counts, comparisons of up-regulated HERVs vs. transcribed HERVs and transcribed HERVs vs. all HERVs were performed with Fisher's exact test. FDR was calculated with Benjamini–Hochberg (BH) method.

Gene set variation analysis

Gene Set Variation Analysis (GSVA) [133] is a modified method of Gene Set Enrichment Analysis (GSEA) [139]. GSVA calculates sample-wise enrichment score (or activity) of a gene set in an unsupervised manner [133]. To perform GSVA, two kinds of gene sets were prepared: Gene Ontology biological process (GO BP)-based and

TFBS-based gene sets. GO BP-based gene sets were obtained from in MSigDB Version 6.1 [139]. TFBS-based gene set, which is a set of genes targeted by a particular TF in a certain condition, was defined for each TFBS dataset in ChIP-Atlas (http://chip-atlas.org/) as described below: Genes having GO annotations in the BP category were extracted from GENCODE Version 22 (most of these genes were protein-coding genes). Of these genes, genes in which TFBSs were present within 5 kb upstream or 3 kb downstream from the TSSs were selected. Of these selected genes, up to 1,000 genes with potent TFBS signals were chosen, and then these chosen genes were defined as a gene set for the TFBS dataset. Using the two kinds of gene sets, GSVA was performed with the default setting.

Reference

1. Bruce Alberts AJ, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. Molecular Biology of the Cell. 5th edition. New York: Garland Science; 2007.
2. Liu G, Zhao S, Bailey JA, Sahinalp SC, Alkan C, Tuzun E, et al. Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. Genome Res. 2003;13: 358-368.
3. Cordaux R, Batzer MA. The impact of retrotransposons on human genome evolution. Nat Rev Genet. 2009;10: 691-703.
4. Feschotte C, Pritham EJ. DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet. 2007;41: 331-368.
5. Blikstad V, Benachenhou F, Sperber GO, Blomberg J. Evolution of human endogenous retroviral sequences: a conceptual account. Cell Mol Life Sci. 2008;65: 3348-3365.
6. Coffin JM, Hughes SH, Varmus HE, editors. Retroviruses. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 1997.
7. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al.

Initial sequencing and analysis of the human genome. Nature. 2001;409: 860-921.
8. Nakagawa S, Takahashi MU. gEVE: a genome-based endogenous viral element database provides comprehensive viral protein-coding sequences in mammalian genomes. Database (Oxford). 2016;2016.
9. Feschotte C, Gilbert C. Endogenous viruses: insights into viral evolution and impact on host biology. Nat Rev Genet. 2012;13: 283-296.
10. Hurst GD, Werren JH. The role of selfish genetic elements in eukaryotic evolution. Nat Rev Genet. 2001;2: 597-606.
11. Dupressoir A, Lavialle C, Heidmann T. From ancestral infectious retroviruses to bona fide cellular genes: role of the captured syncytins in placentation. Placenta. 2012;33: 663-671.
12. Malfavon-Borja R, Feschotte C. Fighting fire with fire: endogenous retrovirus envelopes as restriction factors. J Virol. 2015;89: 4047-4050.
13. Feschotte C. Transposable elements and the evolution of regulatory networks. Nat Rev Genet. 2008;9: 397-405.
14. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet. 2017;18: 71-86.
15. Koyanagi-Aoi M, Ohnuki M, Takahashi K, Okita K, Noma H, Sawamura Y, et al. Differentiation-defective phenotypes revealed by large-scale analyses of human pluripotent stem cells. Proc Natl Acad Sci U S A. 2013;110: 20569-20574.
16. Ohnuki M, Tanabe K, Sutou K, Teramoto I, Sawamura Y, Narita M, et al. Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. Proc Natl Acad Sci U S A. 2014;111: 12426-12431.
17. Wang J, Xie G, Singh M, Ghanbarian AT, Rasko T, Szvetnik A, et al. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. Nature. 2014;516: 405-409.
18. Lu X, Sachs F, Ramsay L, Jacques PE, Goke J, Bourque G, et al. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. Nat Struct Mol Biol. 2014;21: 423-425.
19. Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, et al. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet. 2013;9: e1003470.

20. Chiappinelli KB, Strissel PL, Desrichard A, Li H, Henke C, Akman B, et al. Inhibiting DNA Methylation Causes an Interferon Response in Cancer via dsRNA Including Endogenous Retroviruses. Cell. 2015;162: 974-986.

21. Cuellar TL, Herzner AM, Zhang X, Goyal Y, Watanabe C, Friedman BA, et al. Silencing of retrotransposons by SETDB1 inhibits the interferon response in acute myeloid leukemia. J Cell Biol. 2017;216: 3535-3549.

22. Izquierdo-Bouldstridge A, Bustillos A, Bonet-Costa C, Aribau-Miralbes P, Garcia-Gomis D, Dabad M, et al. Histone H1 depletion triggers an interferon response in cancer cells via activation of heterochromatic repeats. Nucleic Acids Res. 2017;45: 11622-11642.

23. Mi S, Lee X, Li X, Veldman GM, Finnerty H, Racie L, et al. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature. 2000;403: 785-789.

24. Burns KH. Transposable elements in cancer. Nat Rev Cancer. 2017;17: 415-424.

25. Kassiotis G. Endogenous retroviruses and the development of cancer. J Immunol. 2014;192: 1343-1349.

26. Babaian A, Romanish MT, Gagnier L, Kuo LY, Karimi MM, Steidl C, et al. Onco-exaptation of an endogenous retroviral LTR drives IRF5 expression in Hodgkin lymphoma. Oncogene. 2016;35: 2542-2546.

27. Brocks D, Schmidt CR, Daskalakis M, Jang HS, Shah NM, Li D, et al. DNMT and HDAC inhibitors induce cryptic transcription start sites encoded in long terminal repeats. Nat Genet. 2017;49: 1052-1060.

28. Zhou F, Li M, Wei Y, Lin K, Lu Y, Shen J, et al. Activation of HERV-K Env protein is essential for tumorigenesis and metastasis of breast cancer cells. Oncotarget. 2016;7: 84093-84117.

29. Johanning GL, Malouf GG, Zheng X, Esteva FJ, Weinstein JN, Wang-Johanning F, et al. Expression of human endogenous retrovirus-K is strongly associated with the basal-like breast cancer phenotype. Sci Rep. 2017;7: 41960.

30. Lemaitre C, Tsang J, Bireau C, Heidmann T, Dewannieux M. A human endogenous retrovirus-derived gene that can contribute to oncogenesis by activating the ERK pathway and inducing migration and invasion. PLoS Pathog. 2017;13: e1006451.

31. Cherkasova E, Weisman Q, Childs RW. Endogenous retroviruses as targets

for antitumor immunity in renal cell cancer and other tumors. Front Oncol. 2013;3: 243.
32. Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell. 2015;160: 48-61.
33. Consortium. EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489: 57-74.
34. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518: 317-330.
35. Tsankov AM, Gu H, Akopian V, Ziller MJ, Donaghey J, Amit I, et al. Transcription factor binding dynamics during human ES cell differentiation. Nature. 2015;518: 344-349.
36. Tomczak K, Czerwinska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn). 2015;19: A68-77.
37. Jacques PE, Jeyakani J, Bourque G. The majority of primate-specific regulatory sequences are derived from transposable elements. PLoS Genet. 2013;9: e1003504.
38. Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, et al. Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome Res. 2014;24: 1963-1976.
39. Blaise S, de Parseval N, Benit L, Heidmann T. Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution. Proc Natl Acad Sci U S A. 2003;100: 13013-13018.
40. Best S, Le Tissier P, Towers G, Stoye JP. Positional cloning of the mouse retrovirus restriction gene Fv1. Nature. 1996;382: 826-829.
41. Ikeda H, Laigret F, Martin MA, Repaske R. Characterization of a molecularly cloned retroviral sequence associated with Fv-4 resistance. J Virol. 1985;55: 768-777.
42. Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet. 2010;42: 631-634.
43. Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science. 2016;351: 1083-1087.

44. Durruthy-Durruthy J, Sebastiano V, Wossidlo M, Cepeda D, Cui J, Grow EJ, et al. The primate-specific noncoding RNA HPAT5 regulates pluripotency during human preimplantation development and nuclear reprogramming. Nat Genet. 2016;48: 44-52.

45. Becker KG, Swergold GD, Ozato K, Thayer RE. Binding of the ubiquitous nuclear transcription factor YY1 to a cis regulatory sequence in the human LINE-1 transposable element. Hum Mol Genet. 1993;2: 1697-1702.

46. Minakami R, Kurose K, Etoh K, Furuhata Y, Hattori M, Sakaki Y. Identification of an internal cis-element essential for the human L1 transcription and a nuclear factor(s) binding to the element. Nucleic Acids Res. 1992;20: 3139-3145.

47. Tchenio T, Casella JF, Heidmann T. Members of the SRY family regulate the human LINE retrotransposons. Nucleic Acids Res. 2000;28: 411-415.

48. Mathias SL, Scott AF. Promoter binding proteins of an active human L1 retrotransposon. Biochem Biophys Res Commun. 1993;191: 625-632.

49. Fuchs NV, Kraft M, Tondera C, Hanschmann KM, Lower J, Lower R. Expression of the human endogenous retrovirus (HERV) group HML-2/HERV-K does not depend on canonical promoter elements but is regulated by transcription factors Sp1 and Sp3. J Virol. 2011;85: 3436-3448.

50. Grow EJ, Flynn RA, Chavez SL, Bayless NL, Wossidlo M, Wesche DJ, et al. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. Nature. 2015;522: 221-225.

51. Manghera M, Douville RN. Endogenous retrovirus-K promoter: a landing strip for inflammatory transcription factors? Retrovirology. 2013;10: 16.

52. Sjottem E, Anderssen S, Johansen T. The promoter activity of long terminal repeats of the HERV-H family of human retrovirus-like elements is critically dependent on Sp1 family proteins interacting with a GC/GT box located immediately 3' to the TATA box. J Virol. 1996;70: 188-198.

53. Yu X, Zhu X, Pi W, Ling J, Ko L, Takeda Y, et al. The long terminal repeat (LTR) of ERV-9 human endogenous retrovirus binds to NF-Y in the assembly of an active LTR enhancer complex NF-Y/MZF1/GATA-2. J Biol Chem. 2005;280: 35184-35194.

54. Gerlo S, Davis JR, Mager DL, Kooijman R. Prolactin in man: a tale of two promoters. Bioessays. 2006;28: 1051-1055.

55. Jordan IK, Rogozin IB, Glazko GV, Koonin EV. Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet. 2003;19: 68-72.

56. van de Lagemaat LN, Landry JR, Mager DL, Medstrand P. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet. 2003;19: 530-536.

57. Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, et al. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature. 2006;441: 87-90.

58. Pi W, Zhu X, Wu M, Wang Y, Fulzele S, Eroglu A, et al. Long-range function of an intergenic retrotransposon. Proc Natl Acad Sci U S A. 2010;107: 12992-12997.

59. Suntsova M, Gogvadze EV, Salozhin S, Gaifullin N, Eroshkin F, Dmitriev SE, et al. Human-specific endogenous retroviral insert serves as an enhancer for the schizophrenia-linked gene PRODH. Proc Natl Acad Sci U S A. 2013;110: 19472-19477.

60. Chuong EB, Rumi MA, Soares MJ, Baker JC. Endogenous retroviruses function as species-specific enhancer elements in the placenta. Nat Genet. 2013;45: 325-329.

61. Roman AC, Gonzalez-Rico FJ, Molto E, Hernando H, Neto A, Vicente-Garcia C, et al. Dioxin receptor and SLUG transcription factors regulate the insulator activity of B1 SINE retrotransposons via an RNA polymerase switch. Genome Res. 2011;21: 422-432.

62. Wang J, Vicente-Garcia C, Seruggia D, Molto E, Fernandez-Minan A, Neto A, et al. MIR retrotransposon sequences provide insulators to the human genome. Proc Natl Acad Sci U S A. 2015;112: E4428-4437.

63. Lynch VJ, Nnamani MC, Kapusta A, Brayer K, Plaza SL, Mazur EC, et al. Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. Cell Rep. 2015;10: 551-561.

64. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods. 2012;9: 473-476.

65. Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, et al. Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. 2013;41: 827-841.

66. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012;9: 215-216.

67. Jin Y, Tam OH, Paniagua E, Hammell M. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. Bioinformatics. 2015;31: 3593-3599.

68. Goode DK, Obier N, Vijayabaskar MS, Lie ALM, Lilly AJ, Hannah R, et al. Dynamic Gene Regulatory Networks Drive Hematopoietic Specification and Differentiation. Dev Cell. 2016;36: 572-587.

69. Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res. 2012;22: 1798-1812.

70. Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet. 2013;14: 390-403.

71. Cohen CJ, Lock WM, Mager DL. Endogenous retroviral LTRs as promoters for human genes: a critical assessment. Gene. 2009;448: 105-114.

72. Leung D, Jung I, Rajagopal N, Schmitt A, Selvaraj S, Lee AY, et al. Integrative analysis of haplotype-resolved epigenomes across human tissues. Nature. 2015;518: 350-354.

73. Subramanian RP, Wildschutte JH, Russo C, Coffin JM. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. Retrovirology. 2011;8: 90.

74. Fleming JD, Pavesi G, Benatti P, Imbriano C, Mantovani R, Struhl K. NF-Y coassociates with FOS at promoters, enhancers, repetitive elements, and inactive chromatin regions, and is stereo-positioned with growth-controlling transcription factors. Genome Res. 2013;23: 1195-1209.

75. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 2010;28: 495-501.

76. Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L,

et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat Genet. 2015;47: 598-606.

77. Cairns J, Freire-Pritchett P, Wingett SW, Varnai C, Dimond A, Plagnol V, et al. CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. Genome Biol. 2016;17: 127.

78. Goke J, Lu X, Chan YS, Ng HH, Ly LH, Sachs F, et al. Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. Cell Stem Cell. 2015;16: 135-141.

79. Satou Y, Miyazato P, Ishihara K, Yaguchi H, Melamed A, Miura M, et al. The retrovirus HTLV-1 inserts an ectopic CTCF-binding site into the human genome. Proc Natl Acad Sci U S A. 2016.

80. Collins PL, Kyle KE, Egawa T, Shinkai Y, Oltz EM. The histone methyltransferase SETDB1 represses endogenous and exogenous retroviruses in B lymphocytes. Proc Natl Acad Sci U S A. 2015;112: 8367-8372.

81. Kuse K, Ito J, Miyake A, Kawasaki J, Watanabe S, Makundi I, et al. Existence of Two Distinct Infectious Endogenous Retroviruses in Domestic Cats and Their Different Strategies for Adaptation to Transcriptional Regulation. J Virol. 2016;90: 9029-9045.

82. Izsvak Z, Wang J, Singh M, Mager DL, Hurst LD. Pluripotency and the endogenous retrovirus HERVH: Conflict or serendipity? Bioessays. 2016;38: 109-117.

83. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078-2079.

84. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinformatics. 2014;47: 11.12.11-34.

85. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30: 772-780.

86. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011;27: 1017-1018.

87. Mathelier A, Fornes O, Arenillas DJ, Chen CY, Denay G, Lee J, et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2016;44: D110-115.

88. Kulakovskiy IV, Vorontsov IE, Yevshin IS, Soboleva AV, Kasianov AS,

Ashoor H, et al. HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. Nucleic Acids Res. 2016;44: D116-125.

89. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30: 1312-1313.

90. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59: 307-321.

91. Johnson WE, Coffin JM. Constructing primate phylogenies from ancient retrovirus sequences. Proc Natl Acad Sci U S A. 1999;96: 10254-10260.

92. Myers EW, Miller W. Optimal alignments in linear space. Comput Appl Biosci. 1988;4: 11-17.

93. Magiorkinis G, Blanco-Melo D, Belshaw R. The decline of human endogenous retroviruses: extinction and survival. Retrovirology. 2015;12: 8.

94. Gonzalez-Cao M, Iduma P, Karachaliou N, Santarpia M, Blanco J, Rosell R. Human endogenous retroviruses and cancer. Cancer Biol Med. 2016;13: 483-488.

95. Downey RF, Sullivan FJ, Wang-Johanning F, Ambs S, Giles FJ, Glynn SA. Human endogenous retrovirus K and cancer: Innocent bystander or tumorigenic accomplice? Int J Cancer. 2015;137: 1249-1257.

96. Haase K, Mosch A, Frishman D. Differential expression analysis of human endogenous retroviruses based on ENCODE RNA-seq data. BMC Med Genomics. 2015;8: 71.

97. Li M, Radvanyi L, Yin B, Li J, Chivukula R, Lin K, et al. Downregulation of Human Endogenous Retrovirus Type K (HERV-K) Viral env RNA in Pancreatic Cancer Cells Decreases Cell Proliferation and Tumor Growth. Clin Cancer Res. 2017;23: 5892-5911.

98. Chen Y, Peng Y, Xu Z, Ge B, Xiang X, Zhang T, et al. LncROR Promotes Bladder Cancer Cell Proliferation, Migration, and Epithelial-Mesenchymal Transition. Cell Physiol Biochem. 2017;41: 2399-2410.

99. Li H, Jiang X, Niu X. Long Non-Coding RNA Reprogramming (ROR) Promotes Cell Proliferation in Colorectal Cancer via Affecting P53. Med Sci Monit. 2017;23: 919-928.

100. Hurst TP, Magiorkinis G. Epigenetic Control of Human Endogenous Retrovirus Expression: Focus on Regulation of Long-Terminal Repeats (LTRs). Viruses.

2017;9.
101. Desai N, Sajed D, Arora KS, Solovyov A, Rajurkar M, Bledsoe JR, et al. Diverse repetitive element RNA expression defines epigenetic and immunologic features of colon cancer. JCI Insight. 2017;2: e91078.
102. Liu S, Brind'Amour J, Karimi MM, Shirane K, Bogutz A, Lefebvre L, et al. Setdb1 is required for germline development and silencing of H3K9me3-marked endogenous retroviruses in primordial germ cells. Genes Dev. 2014;28: 2041-2055.
103. Matsui T, Leung D, Miyashita H, Maksakova IA, Miyachi H, Kimura H, et al. Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. Nature. 2010;464: 927-931.
104. Wolf G, Yang P, Fuchtbauer AC, Fuchtbauer EM, Silva AM, Park C, et al. The KRAB zinc finger protein ZFP809 is required to initiate epigenetic silencing of endogenous retroviruses. Genes Dev. 2015;29: 538-554.
105. Rowe HM, Jakobsson J, Mesnard D, Rougemont J, Reynard S, Aktas T, et al. KAP1 controls endogenous retroviruses in embryonic stem cells. Nature. 2010;463: 237-240.
106. Hutchins AP, Pei D. Transposable elements at the center of the crossroads between embryogenesis, embryonic stem cells, reprogramming, and long non-coding RNAs. Sci Bull (Beijing). 2015;60: 1722-1733.
107. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490: 61-70.
108. Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, et al. Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. Cell. 2015;163: 506-519.
109. Ayarpadikannan S, Lee HE, Han K, Kim HS. Transposable element-driven transcript diversification and its relevance to genetic disorders. Gene. 2015;558: 187-194.
110. Babaian A, Mager DL. Endogenous retroviral promoter exaptation in human cancer. Mob DNA. 2016;7: 24.
111. Sebestyen E, Singh B, Minana B, Pages A, Mateo F, Pujana MA, et al. Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. Genome Res. 2016;26: 732-744.
112. Climente-Gonzalez H, Porta-Pardo E, Godzik A, Eyras E. The Functional

Impact of Alternative Splicing in Cancer. Cell Rep. 2017;20: 2215-2226.
113. Charoentong P, Finotello F, Angelova M, Mayer C, Efremova M, Rieder D, et al. Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. Cell Rep. 2017;18: 248-262.
114. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, et al. Mutational landscape and significance across 12 major cancer types. Nature. 2013;502: 333-339.
115. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013;500: 415-421.
116. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22: 1760-1774.
117. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44: D733-745.
118. Comprehensive molecular characterization of urothelial bladder carcinoma. Nature. 2014;507: 315-322.
119. Robertson AG, Kim J, Al-Ahmadie H, Bellmunt J, Guo G, Cherniack AD, et al. Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. Cell. 2017;171: 540-556.e525.
120. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487: 330-337.
121. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature. 2015;517: 576-582.
122. Davis CF, Ricketts CJ, Wang M, Yang L, Cherniack AD, Shen H, et al. The somatic genomic landscape of chromophobe renal cell carcinoma. Cancer Cell. 2014;26: 319-330.
123. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature. 2013;499: 43-49.
124. Linehan WM, Spellman PT, Ricketts CJ, Creighton CJ, Fei SS, Davis C, et al. Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. N Engl J Med. 2016;374: 135-145.

125. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell. 2017;169: 1327-1341.e1323.
126. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014;511: 543-550.
127. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012;489: 519-525.
128. The Molecular Taxonomy of Primary Prostate Cancer. Cell. 2015;163: 1011-1025.
129. Integrated genomic characterization of papillary thyroid carcinoma. Cell. 2014;159: 676-690.
130. Ito J, Sugimoto R, Nakaoka H, Yamada S, Kimura T, Hayano T, et al. Systematic identification and characterization of regulatory elements derived from human endogenous retroviruses. PLoS Genet. 2017;13: e1006883.
131. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15: 550.
132. Kawakami E, Nakaoka S, Ohta T, Kitano H. Weighted enrichment method for prediction of transcription regulators from transcriptome and global chromatin immunoprecipitation data. Nucleic Acids Res. 2016;44: 5010-5021.
133. Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14: 7.
134. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nat Rev Cancer. 2004;4: 177-183.
135. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29: 15-21.
136. Liao Y, Smyth GK, Shi W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res. 2013;41: e108.
137. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016;32: 2847-2849.
138. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016;44: e71.
139. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting

genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102: 15545-15550.