Epigenetic characterization of housekeeping core promoters and their importance in tumor suppression

Martin Loza[1*], Alexis Vandenbon[2], and Kenta Nakai[1*]
[1]The Institute of Medical Science, The University of Tokyo, Japan.
[2]Institute for Life and Medical Sciences, Kyoto University, Japan.

Abstract

The precise control of gene transcription in humans requires the interaction of different cis-regulatory elements (CREs), as enhancers and promoters, to ensure the accurate transcription of genes. There has been extensive research on the description of cell type-specific (CTS) regulatory interactions, especially between enhancers and promoters. However, constitutively active interactions between CREs have been less studied. In this research, we elucidate the presence of around 11,000 housekeeping CREs (HK-CREs) and describe their main characteristics. Contrary to expectations, our findings indicate that the majority of the ubiquitously active CREs (around 83%) are not the core promoters of housekeeping genes and are involved in a broader role beyond housekeeping gene regulation. HK-CREs are conserved regions rich in unmethylated CpG sites. Their distribution across chromosomes highly correlates with the number of protein-coding genes and they interact with a large number of target genes in long-distance interactions, mostly through promoter-promoter interactions. In the context of cancer, we observed a reduction in the activity of a subset of HK-CREs, particularly those located near the telomere at the end of chromosome 19 and associated with zinc finger genes, that implicates them as important elements to maintain the integrity of the cell. Overall, our work unravels the presence of ubiquitously active CREs essential for the maintenance and stability of cells.

Introduction

The characterization of the different components within the complex transcriptional regulatory mechanism has become a long-standing topic of research [1, 2]. Thanks to the advances in sequencing techniques and experimental protocols, it is possible to investigate different cis-regulatory elements (CREs), e.g. enhancers and promoters, and their differences and similitudes in numerous cell types with unprecedented resolution [3, 4]. These advances have also made clear that classical discrete definitions of CREs do not necessarily fit the intricate interplay between them. For example, several studies have described CREs that may have both enhancer and promoter functions with similar epigenetic features to either classical marks of enhancers or promoters [2, 5]. In this way, promoters with enhancer capabilities could drive the activation of neighboring genes through promoter-promoter interactions [5-7]. In a recent study, contrary to the classical enhancer-promoter interactions on the regulation of cell type-specific genes, it was demonstrated that housekeeping genes (HKG) are mainly regulated through promoter-promoter interactions [8], delineating the importance of this kind of interplay on the maintenance of cells. These new insights into the regulation of HKG along with the intrinsic specificity of the core

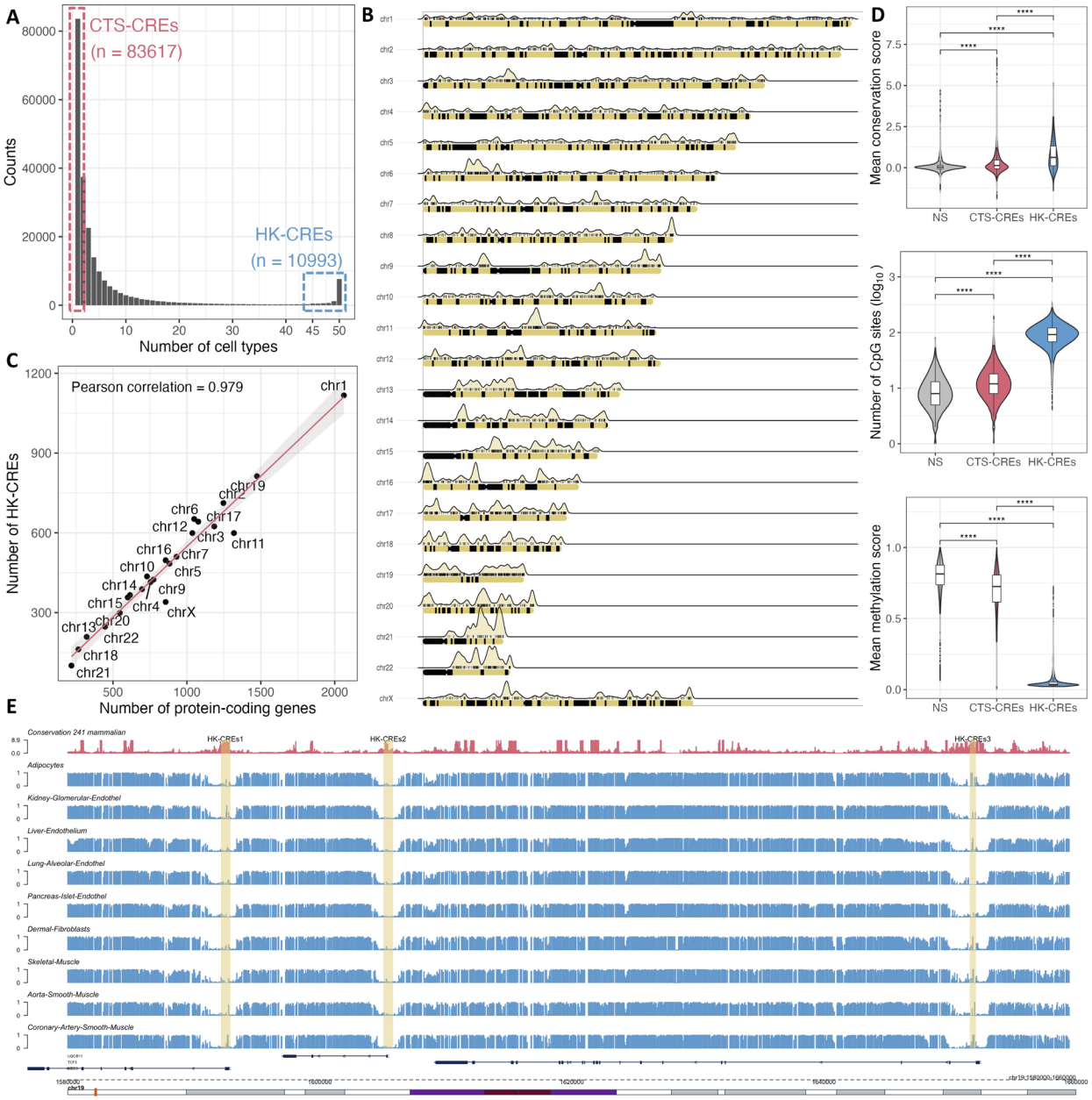

Figure 1. Housekeeping cis-regulatory elements in the human genome. A) After merging overlapping regions from 50 cell types, we observed a significant number of housekeeping cis-regulatory elements (HK-CREs) found to be active in at least 90% of the cells. **B,C)** The distribution of the HK-CREs tends to create clusters **(B)** and highly correlates with the number of protein-coding genes across chromosomes **(C)**. **D,E)** HK-CREs reside in conserved regions dense on unmethylated CpG sites.

promoters of HKG to certain gene regulators [9] suggest the existence of ubiquitously active CREs besides the promoters of HKG. However, to our knowledge, the existence of these CREs that are active in all cell types has not been described. In this study, we leveraged the activity by contact (ABC) dataset [10], which provides a list of active CREs and their predicted target genes from 131 cell types and tissues, to investigate ubiquitously active CREs in healthy cell types. We elucidate the existence of such housekeeping CREs (HK-CREs), e.g. CREs active in at least 90% of the healthy

cell types, and describe their main features. Through several bioinformatic analyses of multiomics data, we show that HK-CREs resided in highly conserved regions that are rich in unmethylated CpG sites, and, similar to HKG [8], they reside in specific loci within the genome as compared with cell type-specific CREs. In conformity with current knowledge on the regulation of HKG [8, 9], we show that most of these ubiquitously active CREs localize in core promoter regions, and they interact with neighboring genes in complex long-distance promoter-promoter interactions. However, contrary to expectations, our analysis revealed that the core promoters of HKG, which are inherently assumed to be active in every cell type, were not the only ubiquitously active CREs; the core promoters of HKG only accounted for around 18% of the total number of HK-CREs. We then compared the biological processes of HKG and the rest of the genes with ubiquitously active core promoters revealing an intrinsic association between these sets of genes, e.g. both sets are enriched in similar or same housekeeping functions, which suggests their complementarity to ensure the proper functioning and maintenance of cells. Finally, we investigated the HK-CREs in diverse cancer cell lines, discovering a reduction in the activity of a subset of these elements. Interestingly, we found that the most affected HK-CREs in cancer cells reside close to the telomere region of chromosome 19 and are inherently related to zinc finger genes. Moreover, we show that while most of these HK-CREs with reduced activity were cancer-specific, there were a set of elements found to be inactive in numerous cancer cell lines, suggesting them as housekeeping CREs of pan-cancer tumor suppressor genes. Overall, our work unveils key cis-regulatory elements active in multiple healthy cell types that, if lost, could strongly affect the housekeeping biological processes of cells and possibly culminate in the development of cancer.

Results

Section 1. Active housekeeping cis-regulatory elements in the human genome.

We investigated the cis-regulatory elements (CREs) from the ABC public dataset [10]. This dataset contains a genome-wide list of active CREs from 131 cell types (Supplementary Table 1), identified from the combination of DNase I hypersensitivity (DNase) or assay for transposase-accessible chromatin sequencing (ATAC-seq) data, with chromatin immunoprecipitation sequencing (ChIP-seq) data for the acetylation of the lysine residue at the N-terminal position 27 of the histone H3 protein (H3K27ac). One asset of this dataset is that it provides a list of predicted interactions between CREs and their surrounding genes within 5 Mb; interactions are quantified from 3D chromatin conformation (Hi-C) data. After selecting significant CRE-gene interactions and filtering out CREs from cell types or tissues related to cancer (n = 26) or various stimuli (n = 34), we obtained around 8.5 million predicted interactions from around 2 million CREs in 71 healthy cell types (Supplementary Table 1). After merging overlapping regions from different cell types (see *Materials and Methods*), we noticed a set of housekeeping CREs (HK-CREs) active in all cell types, which number was stable provided that a sufficient number of cell types were included (see Supplementary Figure 1 and *Materials and Methods*). Then, we randomly selected 50 healthy cell types for further analysis and left the remaining 21 healthy cell types for validation (see *Section 4* and Supplementary Table 1). Finally, we obtained a merged dataset with around one million predicted interactions from around 225,000 CREs. Figure 1A shows the distribution of the resulting CREs across the number of cell types they are active in. While most of the CREs are active in one cell type (cell type-specific; CTS-CREs), we found around 11,000 HK-CREs active

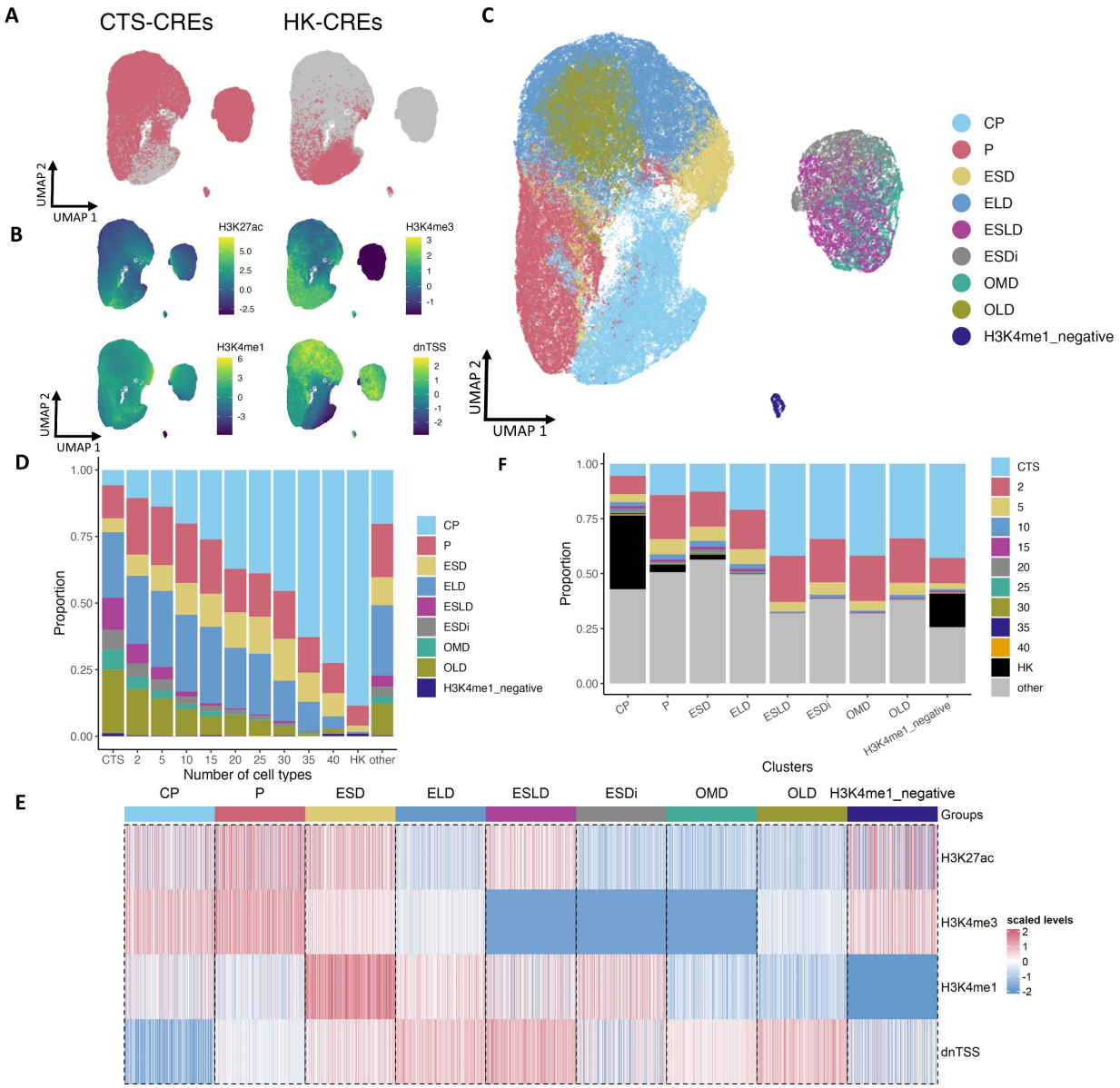

Figure 2. HK-CREs resided in core promoter regions. CREs were annotated with signal levels from H3K27ac, H3K4me3, H3K4me1, and their distance to the nearest TSS (dnTSS). These annotations were used as features in an unsupervised analysis including clustering and visualization. **A,B)** UMAP plot shows a distinct distribution of HK-CREs and cell type-specific-CREs (CTS-CREs) **(A)** with distinct signal levels of histone marks and dnTSS **(B)**. **C,D)** Clustering analysis showed similar results where CTS and HK-CREs were assigned to distinct clusters. **E,F)**. Annotation of the clusters using histone signal levels and dnTSS showed that HK-CREs reside mostly in core promoter regions.

in at least 90% of the 50 cell types. Interestingly, the HK-CREs tend to localize together creating clear peaks on their density distribution across chromosomes (Figure 1B), and are predicted to interact with a considerably larger number of neighboring genes over long distances as compared with CTS-CREs (Supplementary Figures 2 and 3), suggesting their importance in local processes within cells. Moreover, the distribution of HK-CREs across chromosomes highly correlates with

Table 1. Characterization of clusters. The annotations of clusters were manually assigned using a combination of their epigenetics signals and their distance to the nearest TSS.

Group	H3K27ac	H3K4me	H3K4me1	Distance to TSS	Counts	Percentage of total
Core promoters (CP)	+	+	-	Very short	28,968	21.20%
Promoters (P)	++	++	-	Short	24,233	17.73%
Enhancers short distance (ESD)	+	+/-	++	Intermediate	11,445	8.37%
Enhancers long distance (ELD)	+/-	-	+	Long	32,564	23.83%
Enhancers super long distance (ESLD)	+	--	+	Very long	8,063	5.90%
Enhancer short distance inactive (ESDi)	-	--	+	Intermediate	5,975	4.37%
Other medium distance (OMD)	-	--	-	Intermediate	5,109	3.73%
Other long distance (OLD)	-	--	--	Long	19,563	14.31%
H3K4me1 negative (H3K4me1_negative)	+/-	+/-	Zero	Long	745	0.56%

the distribution of protein-coding genes (Figure 1C), implying their importance in the gene regulation process. As expected, because of their presumed housekeeping capacity [11, 12], the loci of HK-CREs are significantly conserved and enriched in unmethylated CpG sites as compared with CTS-CREs and a set of 10,000 randomly selected negative sample (NS) regions (Figure 1D and Supplementary Figures 4 and 5); NS regions were defined as regions not overlapping with any CRE from the ABC and the SCREEN projects [4, 10] (see *Materials and Methods*). Figure 1E shows an example of the conservations and methylation scores of three HK-CREs across 9 tissues (tracks across 30 cell types and tissues can be seen in Supplementary Figure 6), where it can be observed that the loci of the HK-CREs overlap with highly conserved and unmethylated regions within a locus that is highly dense on CpG sites across samples. In contrast, a randomly selected CTS-CRE from T cells presents different methylation patterns across cell types within a not-dense CpG locus (Supplementary Figure 6).

Section 2. Most of the active housekeeping cis-regulatory elements reside in core promoters.

To further characterize and investigate the functions of housekeeping cis-regulatory elements (HK-CREs) we performed an unsupervised classification analysis. We annotated the CREs with their main epigenetic marks (H3K27ac, H3K4me1, and H3K4me3) and their distance to the nearest transcription start site (dnTSS), using a set of ChIP-seq datasets from the ENCODE project [3](Supplementary Table 2) and the TSS locations from the ENSEMBL project [13], respectively. We used these annotations in an unsupervised classification and visualization process, where we obtained the principal components (PCs) [14] coordinates of CREs and used them as the input data to the Louvain clustering [15] and to the Uniform Manifold Approximation and Projection (UMAP) method [16]. To improve the quality of the groups, we filtered out suspicious CREs (n = 88,775) that showed similar attributes to the set of negative samples (NS) defined in Section 1

(see *Materials and Methods*). While most of the filtered CREs were cell type-specific (n = 60,880 filtered CTS-CREs), only a few HK-CREs were filtered (n = 28 filtered HK-CREs), demonstrating the robust epigenetic features of these elements (Supplementary Figure 7). We noticed that the HK-CREs were located at specific parts of the UMAP plot and that they didn't present a strong overlap with cell type-specific CREs (CTS-CREs), suggesting their distinctive epigenetic and dnTSS features (Figure 2A-B). The results from clustering confirmed these differences (Figure 2C-D), where CTS-CREs appeared enriched in groups where the number of HK-CREs is scarce, and vice versa. Figure 2E and Table 1 show the comparison of the feature levels used to manually label nine groups of CREs found by clustering; 5 groups of active CREs: core promoters (CP), promoters (P), enhancer short distance (ESD), enhancers long distance (ELD), and enhancers super long distance (ESLD); one group of inactive enhancers: enhancer short distance inactive (ESDi); two groups of CREs with low signals of epigenetic histone marks: other medium distance (OMD) and other long distance (OLD); and one small group of CREs negative for H3K4me1 (H3K4me1_negative). Interestingly, the majority of the HK-CREs were classified as CPs (Figure 2D and 2F) with a "very short" mean dnTSS of 56 bps.

Section 3. Genes whose core promoter is a HK-CRE complement the biological processes of housekeeping genes.

Housekeeping genes (HKGs) have been widely investigated as constitutively expressed genes in any cell type regardless of their function or developmental stage [17]. In contrast, to our knowledge, the existence of ubiquitously active cis-regulatory elements (CREs) is not clearly defined. We hypothesize that these two elements are closely related, especially for the set of housekeeping CREs identified as core promoters, e.g. HK-CREs active in more than 90% of the 50 cell types and very close to a transcription start site (TSS). To get an insight into the relationship between our identified housekeeping core promoters (HK-CPs) and housekeeping genes, we obtained the list of genes whose TSS is within 300 bp from the center of an HK-CP and compared them with a set of known 2,131 HKGs [17]. As expected, most of the core promoters of HKGs were included in our set of HK-CPs (Figure 3A). However, our list included the core promoters of a considerable number of other genes (n = 8,681) that haven't been identified as HKGs, e.g. non-housekeeping genes (nonHKGs) whose core promoter is constitutively active in all cells (an HK-CP) and predicted to interact with surrounding genes by the ABC method. This suggests that, regardless of the activity of their core promoter, the gene expression levels of nonHKG would be different from those of the HKGs. To verify this, we collected RNA sequencing (RNA-seq) data from 37 cell types from the ENCODE project (Supplementary Table 2) and compared the expression levels of HKGs and nonHKGs. Figures 3B and 3C show the mean gene expression of these sets of genes across cell types and the number of cell types expressing them. As expected, the mean expression and the number of cell types expressing nonHKGs were significantly lower than those of HKGs. However, we argued that nonHKGs, whose core promoter is ubiquitously active, could still play an important role in housekeeping processes within cells regardless of their low RNA expression levels. Therefore, we analyzed the biological process (BP) of HKGs and nonHKG from the Gene Ontology project [18]. Interestingly, the analysis showed a high overlap between the BPs of HKGs and nonHKGs with different significance and enrichment levels (Supplementary Figures 8 and 9), suggesting the complementarity of these two sets of ubiquitously active genes. We then performed a differential analysis of the significance and

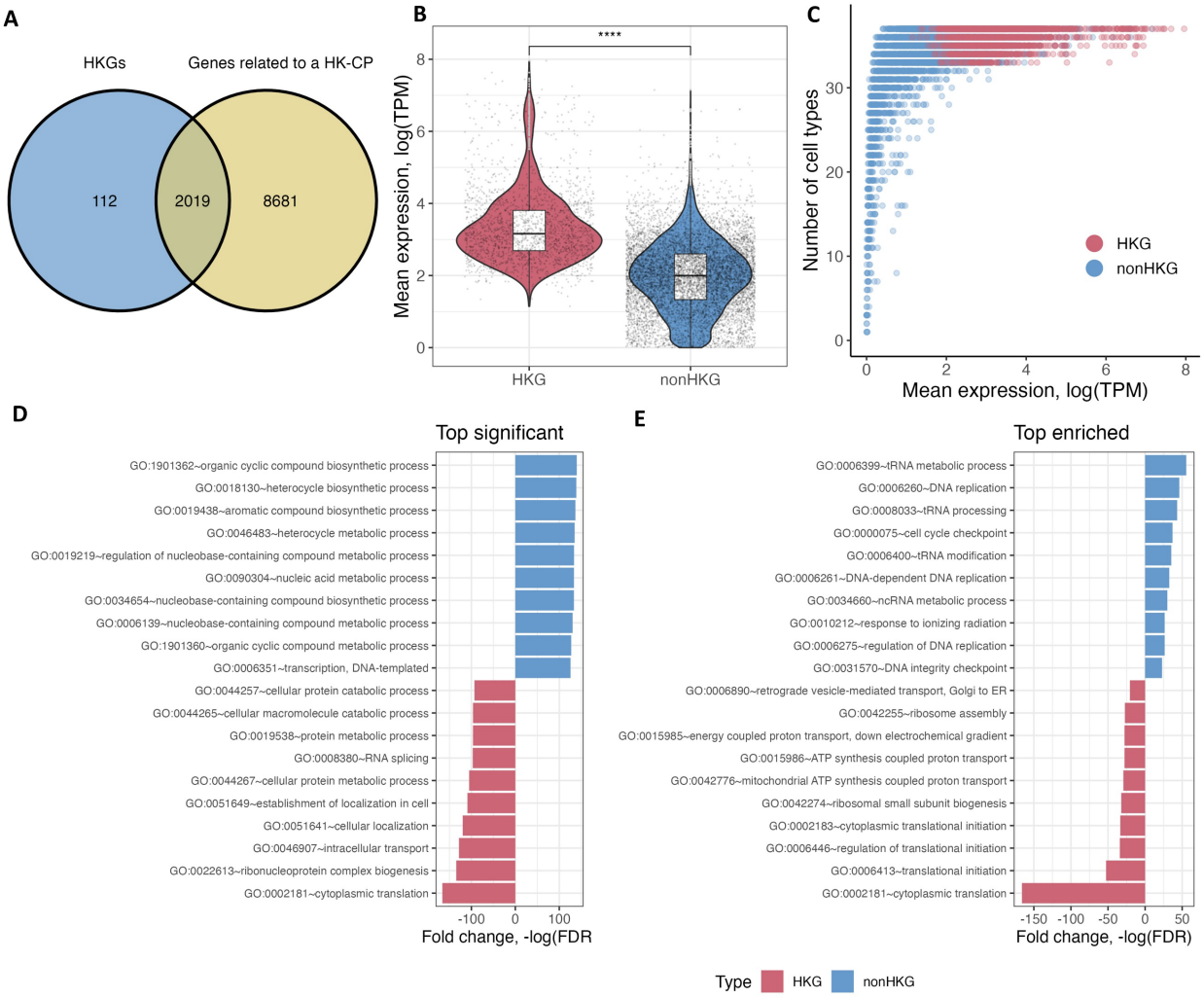

Figure 3. Genes whose core promoter is an HK-CP are significantly enriched in housekeeping biological processes. The genes whose TSS is less than 300 bps from a housekeeping core promoter (HK-CP) were listed and compared with housekeeping genes (HKG). **A)** As expected, most of the core promoters of HKG are HK-CP, however, the core promoters of another 8,681 genes were listed as HK-CPs, e.g. non-housekeeping genes (nonHKG). **B)** Mean gene expression levels of HKGs and nonHKGs over a set of 37 healthy cell types; HKGs present a significantly higher expression than nonHKG. **C)** Mean gene expression levels and the number of cells expressing HKGs and nonHKG; HKGs are ubiquitously expressed in most of the cells, in contrast, nonHKGs present a wider distribution. Besides the differences in gene expression levels and the number of cells expressing HKG and non HKGs, the top biological processes (BPs) sorted by significancy **D)**, e.g. enrichment of genes in general BPs, and by enrichment **E)**, e.g. enrichment of genes in specific BPs, suggested the complementarity of these two sets of genes in important housekeeping biological processes.

enrichment levels between overlapping BPs of HKGs and nonHKG (max FDR = 0.05). Figures 3D and Figure 3E show the top 10 resulting BPs by significance and enrichment, respectively. Interestingly, the significance analysis (Figure 3D and Supplementary Figure 8) showed that, while HKGs were highly associated with general metabolic, localization, translation, and transport BPs within the cell, nonHKG were related to more specific metabolic and transcription processes, supporting the hypothesized complementarity between these two sets of genes. Surprisingly, the enrichment analysis showed that BPs related to tRNA and regulation of RNA polymerase II,

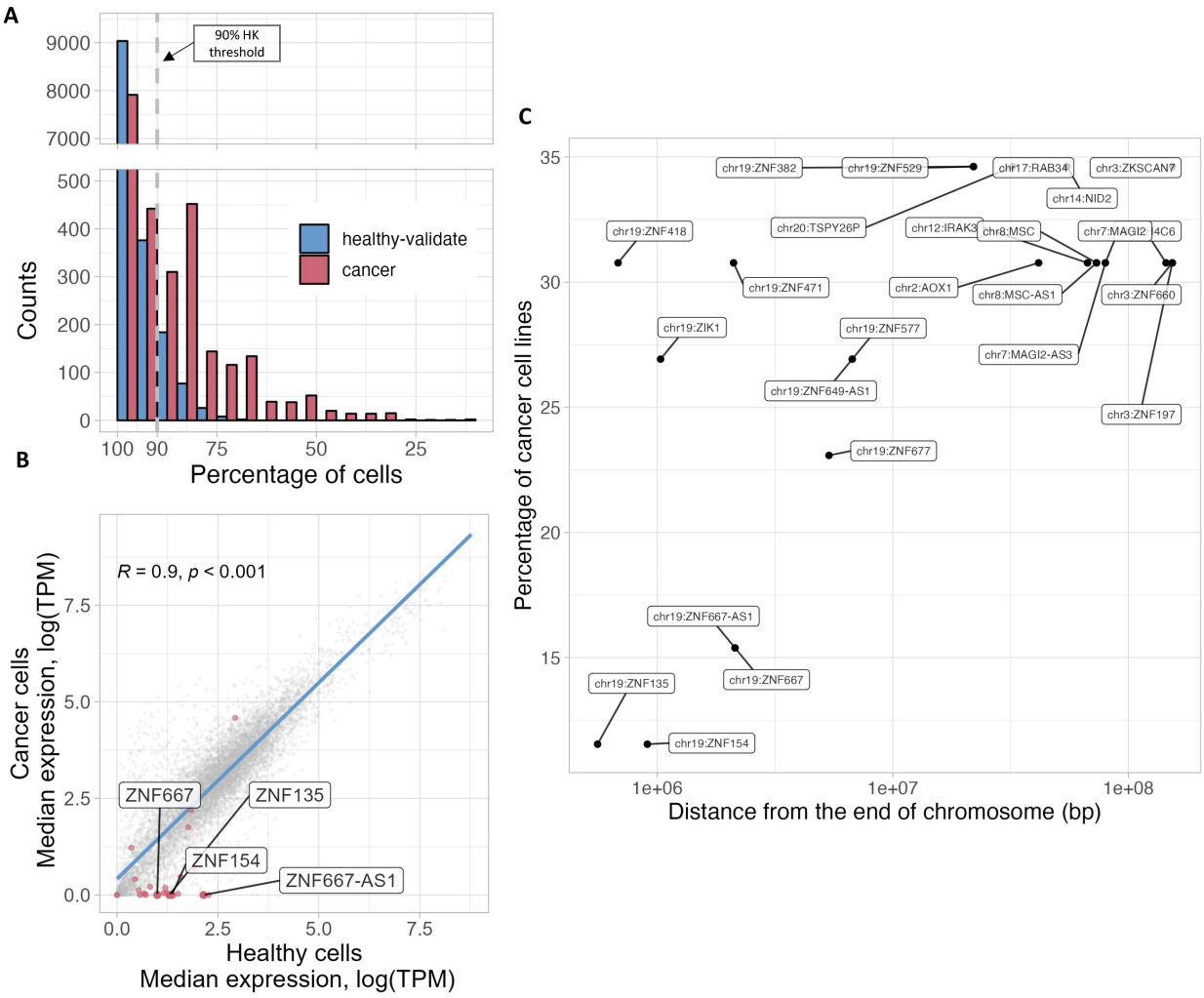

Figure 4. Perturbation of HK-CPs in cancer cell lines. HK-CP regions were compared with a set of 21 healthy cell types (healthy-validate) and a set of 26 cancer cell lines**. A)** While most of the HK-CPs were found active in more than 90% of the healthy cell types (indicated by the gray dashed line), cancer cells presented a considerable number of inactive HK-CPs; 1,354 HK-CPs were found in less than 90% of the cancer cells. **B)** The expression levels of genes which core promoter is a HK-CP were compared between cancer and healthy samples. In general, there is a highly significant correlation between expression levels in healthy and cancer samples. However, the set of genes whose HK-CPs were found active in less than 35% of cancer cells (red-colored dots) presented lower levels of expression on cancer samples. **C)** The genes whose HK-CPs is active in less than 35% of cancer cells were compare in terms of their distance to the end of the chromosome they reside in. The most affected chromosome was chromosome 19 with 9 HK-CPs active in less than 35% of cancer cells. Interestingly, the top inactive HK-CPs in this chromosome resides in loci close to its telomere, e.g. the regions reside within 2Mbps from the end of chromosome 19.

important BPs within the transcription process, were enriched almost exclusively for nonHKGs (Figure 3E and Supplementary Figure 9). Overall, these results confirm the importance of housekeeping core promoters in the gene regulation process and could explain, to some degree, the high correlation between these cis-regulatory elements and the protein-coding genes (Figure 1C and Supplementary Figure 10).

Section 4. The inactivity of housekeeping core promoters could result in cancer development.

In our definition of housekeeping cis-regulatory elements (HK-CREs) in *Section 1*, we labeled putative HK regions as those loci active in at least 90% of 50 randomly selected healthy cell types from the ABC dataset. To further validate our list of HK-CREs, we evaluated them using the rest of the 21 healthy cell types and 26 cancer cell lines from the same ABC dataset (Supplementary Table 1). Thus, for each HK-CRE we calculated the percentage of cells in which it was found to be active in the healthy-validation and cancer datasets. Figure 4A shows the results from this analysis for the set of housekeeping core promoters (HK-CPs; similar results for other annotated HK-CREs are shown in Supplementary Figure 11), where the gray dashed line represents our 90% threshold used to define putative housekeeping elements in *Section 1*. Most of the HK-CPs fulfilled this threshold in healthy cell types (96.94% of HK-CPs, n = 9,411), validating them as housekeeping elements. However, we observed a significant reduction in the distribution of HK-CPs in cancer cell lines (86.05% of HK-CPs, n = 8,354), where 1,354 HK-CPs didn't fulfill the given housekeeping threshold; Supplementary Table 3 contains the list of HK-CPs that didn't fulfill the threshold ranked by their ascending percentage of cancer cells they are found active on. To further explore these findings, similar to the analysis performed in *Section 3,* we obtained the list of genes whose TSS is within 300 bps from the center of a HK-CP, and compared the gene expression levels of these genes between a set of 13 cancer cell lines from the ENCODE project [3] and the healthy set used in Section 3 (n = 37) (Supplementary Table 2). While most of the genes showed a significant correlation between their expression levels in cancer and healthy cells (Figure 4B), the genes whose core promoter was found active in less than 35% of cancer cells (core promoters, n = 19; genes, n = 25) presented a considerably lower median gene expression in cancer cells as compared with healthy cell types (red points in Figure 4B; Supplementary Figure 12). Chromosome 19 was the most affected chromosome with 9 active HK-CPs in less than 35% of cancer cells (Figure 4C, Supplementary Figure 13 and Table 2). Interestingly, most of the HK-CPs with reduced activity in this chromosome tend to localize towards telomere regions, with the top 3 inactive HK-CPs residing in the region less than 2Mbps away from the end of this chromosome, e.g. near the telomere region, suggesting the importance of this locus for maintaining the integrity of cells. We then investigated the expression levels of the genes with the top 3 inactive core promoters: ZNF667-AS1, ZNF667, ZNF154, and ZNF135, which presented significantly reduced expression levels in cancer cells, where the median expression in cancer cells was almost zero (Figure 4B and Supplementary Figure 12). We further validated these results by performing a pan-cancer analysis using the TNMplot web tool [19] for differential gene expression analysis of tumor versus normal samples, obtaining similar results, where the expression levels of these four genes were lower in tumor samples in a variety of cancer subtypes (Supplementary Figure 14 and 15). Because of the ubiquitous activity of their core promoters in healthy cell types and their downregulation in cancer cell lines, we then hypothesized that these four genes could be related to housekeeping tumor suppression functions. Surprisingly, the tumor suppression capacity of these genes has been already observed for different cancer subtypes in diverse tissues: non-small cell lung cancer, renal cancer carcinoma, thyroid cancer, nasopharyngeal carcinoma, squamous cell carcinoma, uveal carcinoma, etc. [20-44]. Moreover, the attenuation of the anti-tumor capacity of these genes has been related to the aberrant methylation of their core promoters in individual [35, 38, 44] and pan-cancer analyses [39, 42,

Table 2. Top inactive HK-CPs in cancer cell lines. Inactive HK-CPs found in less than 35% of cancer cells ranked by their distance to the end of the chromosome they reside in, e.g. telomere regions.

HK-CP Region	Gene name	TSS	% cancer cell lines	% healthy cell types	Distance to the end of the chromosome
chr19:58058989-58059488	**ZNF135**	58,059,239	11.53	95.23	558,377
chr19:57935138-57935637	**ZNF418**	57,935,387	30.76	100.00	682,229
chr19:57708962-57709461	**ZNF154**	57,709,204	11.53	95.23	908,412
chr19:57583726-57584309	**ZIK1**	57,584,140	26.92	100.00	1,033,476
chr19:56507593-56508092	**ZNF471**	56,507,850	30.76	95.23	2,109,766
chr19:56477152-56478177	**ZNF667-AS1**	56,477,590	15.38	95.23	2,140,026
chr19:56477152-56478177	**ZNF667**	56,477,374	15.38	95.23	2,140,242
chr19:53254646-53255153	**ZNF677**	53,254,898	23.07	95.23	5,362,718
chr19:51887687-51888274	**ZNF649-AS1**	51,888,025	26.92	95.23	6,729,591
chr19:51887687-51888274	**ZNF577**	51,887,980	26.92	95.23	6,729,636
chr19:36605027-36605554	**ZNF382**	36,605,313	34.61	90.47	22,012,303
chr19:36605027-36605554	**ZNF529**	36,605,276	34.61	90.47	22,012,340

43]. Together, these results suggest the existence of active housekeeping core promoters related to tumor-suppression processes, that when lost, may culminate in the development of cancer regardless of their malignancy phenotype variant.

Discussion

Our analysis unveils a key housekeeping element of the gene regulatory mechanism that, to our knowledge, hasn't been described before. Through several bioinformatics analyses leveraging diverse public epigenetic and gene expression datasets, we reveal the existence of housekeeping cis-regulatory elements (HK-CREs) that are mostly located within core promoter regions and are not necessarily related to housekeeping genes (HKG), e.g. they are not the core promoters or promoters of HKG. To minimize biases in our analysis, we established our definition of HK-CREs by randomly selecting a substantial number (n = 50) of healthy cell types, ensuring a representative sample. Moreover, besides using a validation dataset with 21 healthy cell types and a set of 10,000 negative samples to assess our findings, we proved in distinct ways the robust location within the genome of the HK-CREs. For instance, we argue that the probability that most of the overlapping regions from 50 cell types intersect within the loci of core promoters (median distance to TSS = 56bps) is exceptionally low. In addition, our analysis using a methylation atlas from healthy cells [45] showed that HK-CREs locate in dense unmethylated CpG loci for multiple

cell types, which can't just occur by chance. Therefore, we feel confident about the existence and robust location of the HK-CREs.

When we first observed that most of the HK-CREs overlapped with core promoter regions, we first thought that we captured the core promoters of HKG. However, the HKG only represented less than 20% of the genes in our list, e.g. genes which core promoter is a HK-CRE. Our immediate conjecture was that the rest of the genes in our list would be genes that have not been correctly annotated as HKG. However, our comparison of gene expression levels showed significant differences between both groups. These results do not contradict our finding of ubiquitously active CREs, as their definition is not based on their transcription levels, but we recognize that there is a bias towards the list of HKG and datasets used in the analysis. Further experimental validations with higher sequencing depth may clarify the distinction between these two sets of genes that seems to complement each other in housekeeping biological processes in the cell.

In this manuscript, we focused on exploring and characterizing HK-CREs instead of the interactions predicted by the ABC method. We observed intricate cooperative interactions between HK-CPs, where various HK-CPs exhibit interactions with one another as well as with the genes in their proximity. This suggests the presence of cis-regulatory hubs with promoter-promoter interactions in their core forming complex networks of regulation. We are aware of the importance of these intriguing promoter-promoter interactions not only for the regulation of HKGs but also for cell type-specific genes. Nonetheless, these interactions come from predictions and must be verified experimentally, but we believe that these two results can be treated separately without conflicting with one other.

In terms of transcription factor binding motifs found in HK-CREs, we observed an enrichment of common GC-rich motifs, e.g. SP1 and SP2. As this information somehow overlaps with the conservation and enrichment of CpG sites shown in Figure 1D, we did not further explore the enriched motifs. However, together with appropriate experimental validations, differential analyses of enriched motifs across annotations of HK and CTS-CREs, e.g. core promoters vs promoter regions, could reveal important motifs for the complex enhancer grammar at different levels.

Finally, in our analysis of inactive housekeeping core promoters (HK-CPs) in cancer cell lines (see Section 4), we tried to avoid possible bias toward the 26 cell lines included in the ABC dataset by focusing on the top inactive HK-CPs from a list of around 1,300 elements. Moreover, our investigation revealed that the most affected elements group together in the telomere of chromosome 19, which proves, to some degree, the robustness of our results. In addition, we verified the reduced gene expression levels of the genes related to these inactive HK-CPs through analysis using data from ENCODE and through the web tool TNMplot, obtaining similar results. Thus, even though our results should be verified experimentally, we are confident that they continue to hold significant relevance for pan-cancer studies concerning the characterization of housekeeping elements involved in tumor suppression.

Materials and Methods

Quality control and pre-processing of regions.

First, we filtered out not significant regions with an ABC score lower than 0.015 as in the original publication [46]. Then, we lifted up the regions' coordinates to the hg38 human genome using the *CrossMap* assembly converter [47] implemented in the Ensemble website [13].

We manually filtered out cancer cell lines or stimulated cell types from the original 131 cells included in the ABC dataset [10], obtaining a healthy dataset of 71 cell types and tissues (Supplementary Table 1). To investigate the dependency of the housekeeping (HK) regions on the number of included cell types and on how strictly we define a region as to be HK, we merge the regions (see *Merging process*) with an increasing number of randomly selected cell types using a 70, 80, 90, or 100 % definition of HK, e.g. a given region is an HK region if it is found in at least 70, 80, 90, or 100% of the cell types. We observed that the number and length of the merged HK regions stabilize after we included around 30 to 40 cell types with a 90 or 100% definition (Supplementary Figure 1). Therefore, we randomly selected 50 healthy cell types (Supplementary Table 1) and merged their regions for further analyses. The regions from the remaining 21 healthy cells were kept for validation. To show the genomic regions of CREs in Figures 1B and Supplementary Figure 13, we used the Chromomap R package [48].

After annotating the main histone marks of the merged regions (see *Annotation of histone marks*), as a quality control step, we filter out regions with histone marks similar to a set of 10,000 negative samples (see *Definition of negative samples*). For this, we clustered regions using the Louvain algorithm [15] implemented in the *igraph* R package [49] using the histone marks and the distance to the nearest TSS as features. Finally, we retained 135,917 regions as putative cis-regulatory elements (see Table 1).

Merging process.

For a given set of cell types, we merged overlapping regions following the next iterative process until no regions were left in the raw dataset:

- <u>Step 1: Merge raw overlapping regions across cell types.</u> We used the function *bedr.merge.region* from the bedr R package [50] with a distance of 20bps.
- <u>Step 2. Select merged regions that overlap on the maximum number of cell types.</u> In the case of housekeeping regions, we merged and selected regions overlapping with a defined 90% percentage of the total number of cells, e.g. regions that overlap in at least 90% of the cell types.
- <u>Step 3. Filter out raw regions overlapping with the selected merged regions.</u> We used the *findOverlaps* function from the GenomicRanges R package [51].
- <u>Step 4. Repeat Step 1 using the filtered raw regions.</u>

Definition of negative samples

We defined a set of 10,000 negative samples (NSs) as non-overlapping regions with cis-regulatory elements (CREs) from the ABC database (131 cell types) nor the SCREEN project [4] as follows:

- Merge regions from the ABC and SCREEN datasets using the *reduce* function from the GenomicRanges R package [51] with a minimum gap width of 100bps.

- Calculate the gaps between the merged regions using the *gaps* function from the GenomicRanges R package.
- Filter out gaps located in centromeres or at the end of telomeres.
- Filter out gaps less than 3Kbps long. This length intends to keep regions of 1.5Kbp which are at least 500bps away from any extended CRE (see *Annotation of histone marks*).
- Select 10,000 random gaps and define NS as 500bps regions located at the center of the selected gaps.

Annotation of histone marks

We annotated the H3K27ac, H3K4me1, and H3K4me3 histone marks of an extended version of the cis-regulatory elements and negative samples. The extended version includes the original regions plus 250bps on each flank to account for adjacent histones.
We downloaded ChIP-seq histone marks datasets from the ENCODE project [3] (Supplementary Table 2) of samples selected as being similar to the main 50 healthy cells/tissues (see *Quality control and pre-processing of regions*). Following the pre-processing used on the SCREEN project [4], we performed independent log(signal+1) normalization and scaling of the samples. Finally, we annotated the histone marks of the extended regions as the average signal of overlapping peaks across pre-processed samples; overlapping peaks were selected as those peaks with at least 250bps overlap with the extended regions.

Annotation of the distance to the nearest TSS

We annotated the distance to the nearest TSS of cis-regulatory elements and negative samples using the ENSEMBL transcripts annotation version GRCh38.107.

Annotation of the conservation scores

We annotated the mean conservation scores on a fixed-length version of the cis-regulatory elements (CREs). The fixed-length CREs were defined to be 500 bps at both flanking loci from the center of each CRE. Then, we overlapped each fixed-length CRE with the conservation scores of 241 mammals from the Zoonomia project [52] and calculated the mean conservation score as the mean score among overlapping regions.

RNA-seq datasets

All the healthy and cancer RNA-seq data were obtained from the ENCODE project [3] (Supplementary Table 2). For pre-processing and quality control of samples we followed the next pipeline:

- Load the TPM data using the *tximport* Bioconductor package [53].
- Filter out genes with zero TPM in all the samples.
- Map the gene labels to match the gene names used in the ABC dataset. We used the *org.Hk.eg.db* and *AnnotationDbi* Bioconductor packages [54, 55] using a combination of ALIAS, ENSEMBL, and ENTREZID to match the gene names.
- Natural log-normalize the TPM data.

Methylation Atlas

We downloaded the methylation scores from 34 randomly picked samples from the DNA methylation atlas of normal human cell types [45] (Supplementary Table 2). Then, for each fixed-length CRE and NS, we counted the number of overlapping CpG sites and calculated the mean methylation score across cell types; the fixed-length regions were defined as regions of 1kbps concentric with every CRs and NS. To produce the methylation tracks from Figure 1E and Supplementary Figure 6, we used the *track_plot* function from the trackplot R package [56].

Gene Ontology Analysis

Gene ontology analysis of genes was performed using the DAVID web tool [57]. Significant processes were selected to have max FDR <= 0.05. The differential enrichment analysis was performed by calculating the fold change of -log(FDR) between housekeeping and non-housekeeping genes.

Statistical information

For the statistical comparison used in the manuscript, we used the *stat_compare_means* function from the ggpbur R package [58] with the next convention for symbols indicating statistical significance:

- ns: p > 0.05
- *: p <= 0.05
- **: p <= 0.01
- ***: p <= 0.001
- ****: p <= 0.0001

In Figure 1D and Figure 3B, we used an unpaired Wilcoxon test with a two-sided alternative hypothesis. In Supplementary Figure 12, we used a paired Wilcoxon test with a two-sided alternative hypothesis.

Data Availability

The ABC dataset was downloaded from https://www.engreitzlab.org/resources.
Histone modification data and RNA data were downloaded from the ENCODE project. The sources and accession numbers are summarized in Supplementary Table 2.
Conservation scores were downloaded from the Zoonomia project from https://cglgenomics.ucsc.edu/data/cactus/.
Methylation scores were downloaded from the GEO (accession no. GSE186458). The accession numbers of cell types are summarized in Supplementary Table 2.
Annotations of genes and transcripts were downloaded from the ENSEMBL project http://asia.ensembl.org/Homo_sapiens/Info/Index.
Any additional information generated in this paper is available from the corresponding author upon reasonable request.

Bibliography

1. Nakai, K. and A. Vandenbon, *Higher-order chromatin structure and gene regulation*, in *Epigenetics in Organ Specific Disorders*. 2023, Elsevier. p. 11-32.

2. Andersson, R. and A. Sandelin, *Determinants of enhancer and promoter activities of regulatory elements.* Nat Rev Genet, 2020. **21**(2): p. 71-87.
3. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
4. Consortium, E.P., et al., *Expanded encyclopaedias of DNA elements in the human and mouse genomes.* Nature, 2020. **583**(7818): p. 699-710.
5. Dao, L.T.M., et al., *Genome-wide characterization of mammalian promoters with distal enhancer functions.* Nat Genet, 2017. **49**(7): p. 1073-1081.
6. Li, G., et al., *Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation.* Cell, 2012. **148**(1-2): p. 84-98.
7. Santiago-Algarra, D., et al., *Epromoters function as a hub to recruit key transcription factors required for the inflammatory response.* Nat Commun, 2021. **12**(1): p. 6660.
8. Dejosez, M., et al., *Regulatory architecture of housekeeping genes is driven by promoter assemblies.* Cell Rep, 2023. **42**(5): p. 112505.
9. Zabidi, M.A., et al., *Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation.* Nature, 2015. **518**(7540): p. 556-9.
10. Nasser, J., et al., *Genome-wide enhancer maps link risk variants to disease genes.* Nature, 2021. **593**(7858): p. 238-243.
11. Zhu, J., et al., *On the nature of human housekeeping genes.* Trends Genet, 2008. **24**(10): p. 481-4.
12. Deaton, A.M. and A. Bird, *CpG islands and the regulation of transcription.* Genes Dev, 2011. **25**(10): p. 1010-22.
13. Cunningham, F., et al., *Ensembl 2022.* Nucleic Acids Res, 2022. **50**(D1): p. D988-D995.
14. Pearson, K., *LIII. On lines and planes of closest fit to systems of points in space.* The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 1901. **2**(11): p. 559-572.
15. Subelj, L. and M. Bajec, *Unfolding communities in large complex networks: combining defensive and offensive label propagation for core extraction.* Phys Rev E Stat Nonlin Soft Matter Phys, 2011. **83**(3 Pt 2): p. 036103.
16. McInnes, L., J. Healy, and J. Melville, *Umap: Uniform manifold approximation and projection for dimension reduction.* arXiv preprint arXiv:1802.03426, 2018.
17. Hounkpe, B.W., et al., *HRT Atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets.* Nucleic Acids Res, 2021. **49**(D1): p. D947-D955.
18. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
19. Bartha, A. and B. Gyorffy, *TNMplot.com: A Web Tool for the Comparison of Gene Expression in Normal, Tumor and Metastatic Tissues.* Int J Mol Sci, 2021. **22**(5).
20. Ma, X., et al., *Development and Validation of a Novel Ferroptosis-Related LncRNA Signature for Predicting Prognosis and the Immune Landscape Features in Uveal Melanoma.* Front Immunol, 2022. **13**: p. 922315.
21. Wei, L., et al., *Regulation of lncRNA ZNF667-AS1 in Proliferation and Invasion of Esophageal Squamous Cell Carcinoma Cells via Mediating ceRNA Network.* Crit Rev Eukaryot Gene Expr, 2022. **32**(6): p. 57-68.

22. Yu, C., et al., *The lncRNA ZNF667-AS1 Inhibits Propagation, Invasion, and Angiogenesis of Gastric Cancer by Silencing the Expression of N-Cadherin and VEGFA.* J Oncol, 2022. **2022**: p. 3579547.
23. Zheng, Y.J., et al., *Long non-coding RNA ZNF667-AS1 retards the development of esophageal squamous cell carcinoma via modulation of microRNA-1290-mediated PRUNE2.* Transl Oncol, 2022. **21**: p. 101371.
24. Yang, H., et al., *ZNF667-AS1, a positively regulating MEGF10, inhibits the progression of uveal melanoma by modulating cellular aggressiveness.* J Biochem Mol Toxicol, 2021. **35**(5): p. e22732.
25. Zhao, J., et al., *lncRNA ZNF667-AS1 Suppresses Epithelial Mesenchymal Transformation by Targeting TGF-β1 in Oral Squamous Cell Carcinoma.* Clin Lab, 2021. **67**(7).
26. Zhuang, L., et al., *lncRNA ZNF667-AS1 (NR_036521.1) inhibits the progression of colorectal cancer via regulating ANK2/JAK2 expression.* J Cell Physiol, 2021. **236**(3): p. 2178-2193.
27. Wang, N., et al., *Long Non-Coding RNA ZNF667-AS1 Knockdown Curbs Liver Metastasis in Acute Myeloid Leukemia by Regulating the microRNA-206/AKAP13 Axis.* Cancer Manag Res, 2020. **12**: p. 13285-13300.
28. Yuan, Q., et al., *Analysis of Long Noncoding RNA ZNF667-AS1 as a Potential Biomarker for Diagnosis and Prognosis of Glioma Patients.* Dis Markers, 2020. **2020**: p. 8895968.
29. Dong, Z., et al., *Aberrant hypermethylation-mediated downregulation of antisense lncRNA ZNF667-AS1 and its sense gene ZNF667 correlate with progression and prognosis of esophageal squamous cell carcinoma.* Cell Death Dis, 2019. **10**(12): p. 930.
30. Li, Y.J., et al., *Long noncoding RNA ZNF667-AS1 reduces tumor invasion and metastasis in cervical cancer by counteracting microRNA-93-3p-dependent PEG3 downregulation.* Mol Oncol, 2019. **13**(11): p. 2375-2392.
31. Meng, W., et al., *Aberrant methylation and downregulation of ZNF667-AS1 and ZNF667 promote the malignant progression of laryngeal squamous cell carcinoma.* J Biomed Sci, 2019. **26**(1): p. 13.
32. Zhao, L.P., et al., *Independent prognostic Factor of low-expressed LncRNA ZNF667-AS1 for cervical cancer and inhibitory function on the proliferation of cervical cancer.* Eur Rev Med Pharmacol Sci, 2017. **21**(23): p. 5353-5360.
33. Muthamilselvan, S., A. Raghavendran, and A. Palaniappan, *Stage-differentiated ensemble modeling of DNA methylation landscapes uncovers salient biomarkers and prognostic signatures in colorectal cancer progression.* PLoS One, 2022. **17**(2): p. e0249151.
34. Xi, T. and G. Zhang, *Epigenetic regulation on the gene expression signature in esophagus adenocarcinoma.* Pathol Res Pract, 2017. **213**(2): p. 83-88.
35. Ito, Y., et al., *Association of frequent hypermethylation with high grade histological subtype in lung adenocarcinoma.* Cancer Sci, 2023.
36. Pearson, P., et al., *Kruppel-family zinc finger proteins as emerging epigenetic biomarkers in head and neck squamous cell carcinoma.* J Otolaryngol Head Neck Surg, 2023. **52**(1): p. 41.
37. He, J., et al., *Low Expression of ZNF154 is Related to Poor Prognosis in Gastric Cancer.* Cancer Manag Res, 2022. **14**: p. 659-672.

38. He, W., et al., *Targeted demethylation at ZNF154 promotor upregulates ZNF154 expression and inhibits the proliferation and migration of Esophageal Squamous Carcinoma cells.* Oncogene, 2022. **41**(40): p. 4537-4546.
39. Miller, B.F., H.M. Petrykowska, and L. Elnitski, *Assessing ZNF154 methylation in patient plasma as a multicancer marker in liquid biopsies from colon, liver, ovarian and pancreatic cancer patients.* Sci Rep, 2021. **11**(1): p. 221.
40. Noruzinia, M., et al., *Young Breast Cancer: Novel Gene Methylation in WBC.* Asian Pac J Cancer Prev, 2021. **22**(8): p. 2371-2375.
41. Hu, Y., et al., *Candidate tumor suppressor ZNF154 suppresses invasion and metastasis in NPC by inhibiting the EMT via Wnt/β-catenin signalling.* Oncotarget, 2017. **8**(49): p. 85749-85758.
42. Margolin, G., et al., *Robust Detection of DNA Hypermethylation of ZNF154 as a Pan-Cancer Locus with in Silico Modeling for Blood-Based Diagnostic Development.* J Mol Diagn, 2016. **18**(2): p. 283-98.
43. Sánchez-Vega, F., et al., *Recurrent patterns of DNA methylation in the ZNF154, CASP8, and VHL promoters across a wide spectrum of human solid epithelial tumors and cancer cell lines.* Epigenetics, 2013. **8**(12): p. 1355-72.
44. Reinert, T., et al., *Diagnosis of bladder cancer recurrence based on urinary levels of EOMES, HOXA9, POU4F2, TWIST1, VIM, and ZNF154 hypermethylation.* PLoS One, 2012. **7**(10): p. e46297.
45. Loyfer, N., et al., *A DNA methylation atlas of normal human cell types.* Nature, 2023. **613**(7943): p. 355-364.
46. Fulco, C.P., et al., *Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations.* Nat Genet, 2019. **51**(12): p. 1664-1669.
47. Zhao, H., et al., *CrossMap: a versatile tool for coordinate conversion between genome assemblies.* Bioinformatics, 2014. **30**(7): p. 1006-7.
48. Anand, L. and C.M. Rodriguez Lopez, *ChromoMap: an R package for interactive visualization of multi-omics data and annotation of chromosomes.* BMC Bioinformatics, 2022. **23**(1): p. 33.
49. Csardi, G. and T. Nepusz, *The igraph software package for complex network research.* InterJournal, complex systems, 2006. **1695**(5): p. 1-9.
50. Haider, S., et al., *A bedr way of genomic interval processing.* Source Code Biol Med, 2016. **11**: p. 14.
51. Lawrence, M., et al., *Software for computing and annotating genomic ranges.* PLoS Comput Biol, 2013. **9**(8): p. e1003118.
52. Zoonomia, C., *A comparative genomics multitool for scientific discovery and conservation.* Nature, 2020. **587**(7833): p. 240-245.
53. Soneson, C., M.I. Love, and M.D. Robinson, *Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences.* F1000Res, 2015. **4**: p. 1521.
54. Carlson, M., et al., *org. Hs. eg. db: Genome wide annotation for Human.* R package version, 2019. **3**(2): p. 3.
55. Pages, H., et al., *Package 'AnnotationDbi'.* Bioconductor Packag. Maint, 2017.
56. Pohl, A. and M. Beato, *bwtool: a tool for bigWig files.* Bioinformatics, 2014. **30**(11): p. 1618-9.

57. Sherman, B.T., et al., *DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update).* Nucleic Acids Res, 2022. **50**(W1): p. W216-W221.
58. Kassambara, A., *ggpubr:"ggplot2" based publication ready plots.* R package version 0.4. 0, 2020. **438**.

Acknowledgments

The authors thank Prof. Katsuhiko Shirahige and Dr. Takashi Fukaya for the amazing discussion and their invaluable comments.
This work was supported by Grants-in-Aid for Scientific Research (grant number 22K06189 for K. N. and JP22K21301 for M. L.) from the Japan Society for the Promotion of Science.

Authors' contributions

M. L. contributed to the conception, design, data acquisition, data analysis and interpretation, drafted and critically revised the manuscript. A. V. and K. N. contributed to the conception, design, interpretation of results, edited and critically revised the manuscript. K. N. supervised the project.

Competing interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Corresponding author

Name: Kenta Nakai
Email: knakai@ims.u-tokyo.ac.jp