Clustered and diverse transcription factor binding underlies cell type specificity of enhancers for housekeeping genes

Iris Zhu and David Landsman

Genome Res. 2023 33: 1662-1672 originally published online October 26, 2023
Access the most recent version at doi:10.1101/gr.278130.123

Clustered and diverse transcription factor binding underlies cell type specificity of enhancers for housekeeping genes

Iris Zhu and David Landsman

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA

Housekeeping genes are considered to be regulated by common enhancers across different tissues. Here we report that most of the commonly expressed mouse or human genes across different cell types, including more than half of the previously identified housekeeping genes, are associated with cell type–specific enhancers. Furthermore, the binding of most transcription factors (TFs) is cell type–specific. We reason that these cell type specificities are causally related to the collective TF recruitment at regulatory sites, as TFs tend to bind to regions associated with many other TFs and each cell type has a unique repertoire of expressed TFs. Based on binding profiles of hundreds of TFs from HepG2, K562, and GM12878 cells, we show that 80% of all TF peaks overlapping H3K27ac signals are in the top 20,000–23,000 most TF-enriched H3K27ac peak regions, and approximately 12,000–15,000 of these peaks are enhancers (nonpromoters). Those enhancers are mainly cell type–specific and include those linked to the majority of commonly expressed genes. Moreover, we show that the top 15,000 most TF-enriched regulatory sites in HepG2 cells, associated with about 200 TFs, can be predicted largely from the binding profile of as few as 30 TFs. Through motif analysis, we show that major enhancers harbor diverse and clustered motifs from a combination of available TFs uniquely present in each cell type. We propose a mechanism that explains how the highly focused TF binding at regulatory sites results in cell type specificity of enhancers for housekeeping and commonly expressed genes.

[Supplemental material is available for this article.]

Enhancers are key regulators of differential gene expression in different tissue or cell types (Heinz et al. 2015; Field and Adelman 2020; Panigrahi and O'Malley 2021). Research on enhancers have been largely focused on those that are critically involved in development or cell identity. Many enhancers linked with early development or genes establishing cell lineage have been well studied (Banerji et al. 1983; Gillies et al. 1983; Palstra et al. 2003; Whyte et al. 2013; Zhou et al. 2014; Shin et al. 2016; Osterwalder et al. 2018). Enhancers for housekeeping genes, however, have not received much attention.

There are about 20,000–25,000 genes in the human and mouse genomes. Comprehensive analysis of RNA-seq data from the BodyMap Project identified 3804 and 4781 housekeeping genes in the human and mouse genomes, respectively (Eisenberg and Levanon 2013; Li et al. 2017). Regarding the regulation of housekeeping genes, it seems to be a generally accepted concept that they tend to use common enhancers across different cell types. An important study that established this view is the work by Zabidi et al. (2015), which used STARR-seq to test enhancer functions and showed that enhancers for housekeeping promoters are mostly shared across two different cell types. This work has been cited extensively as evidence for housekeeping genes using common enhancers. The conclusion, however, was based on the work conducted on a few promoters in two *Drosophila* cell lines (Zabidi et al. 2015) and the STARR-seq system using a reporter gene to observe properties of the sequences out of their native context (Gasperini et al. 2020). Another study by Borsari et al. (2021) reported that housekeeping genes are often controlled by intergenic enhancers, common to many tissues. In their study, the number of the common enhancers is only 555 without mentioning their target housekeeping genes.

Housekeeping genes also are known to need fewer enhancers than developmental genes. Osterwalder et al. (2018) reported that each housekeeping gene has less than one enhancer on average, calculated on 1287 mouse genes, far less than the 4781 housekeeping genes reported in the study by Li et al. (2017), which comprehensively analyzed expression data from 17 mouse tissues. Previously, we found that there are few enhancers at regions of densely packed active genes. We identified 120 such active gene clusters that include 1050 genes, which are highly enriched in housekeeping genes (Zhu et al. 2021). Therefore, housekeeping genes that require few enhancers represent a small fraction (about one-quarter) of them.

Our previous work on active transcription hubs found that the "isolated" active promoters, which are distant from other active promoters, tend to have multiple strong enhancers nearby (within 100 kb), including many housekeeping genes (Zhu et al. 2021). In the current study, we report a common phenomenon: that genes expressed at similar levels between multiple cell types in mouse or human, including more than half of previously identified housekeeping genes (Eisenberg and Levanon 2013; Li et al. 2017), are associated with cell type–specific enhancers. This finding expands our knowledge of gene regulation and draws attention to a generally assumed view that universally expressed genes use common enhancers across different cell types. We further

Corresponding author: zhuz2@nih.gov
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.278130.123. Freely available online through the *Genome Research* Open Access option.

33:1662–1672 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/23; www.genome.org

explored the underlying mechanisms from the perspectives of collective recruitment of transcription factors (TFs) and TFBS motif compositions at the major enhancer sites and proposed a model that explains how highly focused TF binding at regulatory sites results in the cell type specificity of enhancers for housekeeping and commonly expressed genes.

Results

Commonly expressed genes across multiple cell types, including most housekeeping genes, are linked with cell type–specific enhancers

Previously, we showed that sparsely located active promoters tend to associate with multiple enhancers regardless of the gene's cell type specificity (Zhu et al. 2021). Next, we found it common that the same gene whose expression level is similar between mouse embryonic stem (ES) cells and MEF cells is associated with different enhancers in the two cell types, including some housekeeping genes. An example is shown in Figure 1A: *Tgif1* is highly expressed in both ES and MEF cells (~25 RPKM in both cells) and is the only active gene in the flanking 300-kb region. The associated enhancers of *Tgif1*, however, are located at different sites in the two cell types. Additional examples of the same gene associated with cell type–specific enhancers in ES and MEF cells are shown in Supplemental Figure S1: Genes *Rbpj* (housekeeping), *Fr2l1*, *Zc3h15* (housekeeping), and *Pdpn* are highly expressed at approximately the same level in ES and MEF cells and associated with different enhancers in the two cell types.

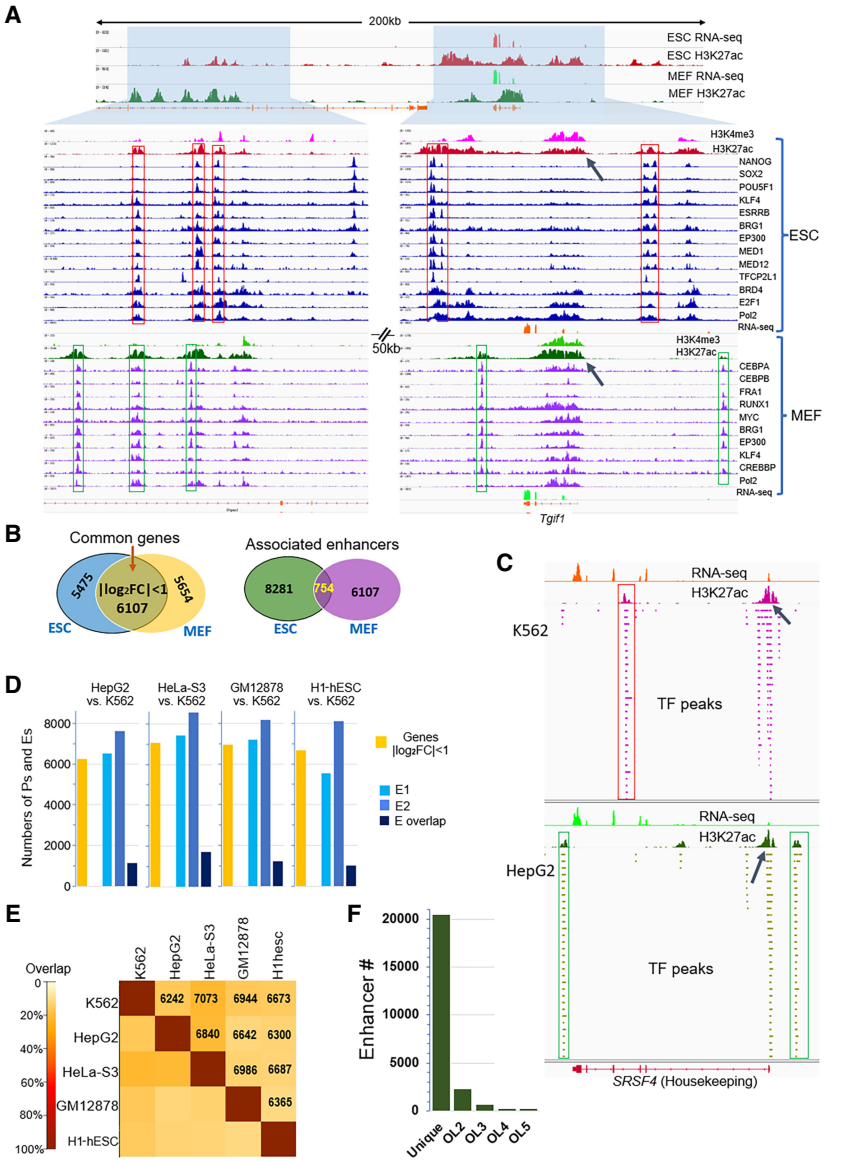

Figure 1. Commonly expressed genes across multiple cell types associated with cell type–specific enhancers. (*A*) RNA-seq and ChIP-seq signals of H3K4me3, H3K27ac, and multiple TFs around a highly expressed gene *Tgif1* in mouse ESCs and MEF cells. The gene promoters are marked with arrows; enhancers, with frames. *Tgif1* is the nearest active promoter to the shown enhancers. (*B*) Number of commonly expressed genes whose expression levels vary within twofold between ESC and MEF, and numbers of their associated strong enhancers and their overlaps between the two cells. (*C*) An example of cell type–specific enhancers for a housekeeping gene *SRSF4* in K562 and HepG2 cells. (*D*) Analysis similar to what is presented in *B*, between K562 and another cell line. (*E*) Pairwise analysis of common genes (numbers in the grid) and overlap rates of associated strong enhancers among five human cell lines: K562, HepG2, GM12878, HeLa-S3, and H1-hESC. (*F*) For the 6272 commonly expressed genes among the five cell lines, the number of associated strong enhancers that are unique in one cell type or are shared by two, three, four, and five cell types. (OL) Overlap.

These observations appear to contradict the generally accepted view that housekeeping genes use common enhancers, which is based mainly on a study conducted in *Drosophila* cell lines with STARR-seq, in which DNA sequences do not function in their natural contexts (Zabidi et al. 2015). We further examined all the commonly expressed genes in mouse ESCs and MEF cells and their associated enhancers. We selected the genes whose expression level is >1 RPKM and differs less than twofold between ES and MEF cells, which we define as commonly expressed genes. We then identified the enhancers associated with the promoters of these genes. There are approximately 37,000 and 56,000 non-promoter H3K27ac peaks in ESCs and MEF cells, respectively. Although H3K27ac modification is generally considered the marker of active enhancers, previous studies have shown that most of the H3K27ac peak sites are poor indicators of enhancer activities (Barakat et al. 2018), whereas collective TF recruitment is a necessary condition for effective enhancer activity (Smith et al. 2013; Long et al. 2016; Lambert et al. 2018; Field and Adelman 2020; Ray-Jones and Spivakov 2021; Singh et al. 2021). Therefore, we selected the top 50% (about 18,000) strongest H3K27ac peaks in ESCs based on TF enrichment and about the same number of strong enhancers in MEF cells (see Methods). We assigned the

nearest TSS (RPKM > 1) to each enhancer to establish its gene target (Hnisz et al. 2013). Of approximately 13,000 genes expressed at RPKM > 1 in each cell type, 6107 are commonly expressed genes in MEFs and ESs (only genes with NCBI RefSeq transcript ID that start with NM_ or NR_ are considered), which include 3297 out of 4781 (69%) housekeeping genes previously reported (Li et al. 2017). The number of strong enhancers associated with the 6107 genes are 8281 and 6756 in ESCs and MEF cells, respectively. The number of common enhancers is only 754 (Fig. 1B). Therefore, we posit that most enhancers associated with commonly expressed genes are cell type–specific. Varying the number of the strongest enhancers in each cell type (e.g., the top most TF-enriched 10,000, 12,000, and 15,000 H3K27ac sites) led to similar results. Below, we will show the analysis that suggests the top 12,000–15,000 most TF-enriched nonpromoter H3K27ac peak sites represent the major effective enhancers.

We observed a similar enhancer specificity for common genes in human cell lines. An example is shown in Figure 1C: The housekeeping gene *SRSF4* is linked to cell type–specific enhancers in K562 and HepG2 cells. We analyzed the histone modification, TF binding, and expression data of several human cell lines from the ENCODE Project (The ENCODE Project Consortium 2012). Each cell line has ChIP-seq data of tens to 200 TFs available, so we can identify strong enhancers based on TF enrichment. The results are very similar to those from the analysis of mouse ES and MEF cells. For the comparison between K562 and another cell line, HepG2, HeLa-S3, GM12878, and H1-hESC, Figure 1D shows the numbers of the active genes (RPKM > 1) whose expression differs less than twofold and the associated strong enhancers in each cell type and their overlaps between the two cells. Clearly, these common genes are associated mainly with cell type–specific enhancers. Some examples of cell type–specific enhancers linked to commonly expressed or housekeeping human genes are shown in Supplemental Figure S2. The results are similar when we compare any other two cell types (Fig. 1E). We also examined DNA accessibility at the selected strong enhancer sites with ATAC-seq data. In each of the seven cell types, 92%–98% of the selected strong enhancer sites overlap with ATAC-seq signals, which indicates an open chromatin state and further supports their potential regulatory functions.

We next examined the common enhancers for commonly expressed genes across the above five human cell lines. In the studies of Eisenberg and Levanon (2013) and Li et al. (2017), the housekeeping genes were identified by three criteria: (1) expression observed in all tissues; (2) low variance over tissues, standard deviation [$\log_2$(RPKM)] < 1; and (3) no exceptional expression in any single tissue, meaning that the log-expression value differs from the averaged $\log_2$(RPKM) by two (fourfold) or more. For the RNA-seq data from the above five cell lines, in addition to the above three criteria, we applied a stricter rule: max[$\log_2$(RPKM)] − min [$\log_2$(RPKM)] < 2, (i.e., the variation of expression level is within fourfold between the highest and lowest expressed cell type). We identified 6272 genes (RPKM > 1) commonly expressed among the five cell lines, with an overlap of 2937 of the housekeeping genes from the study by Eisenberg and Levanon (2013). The most enriched biological process GO terms in the identified 6272 genes are those essential to cellular functions, including mRNA splicing, translation, cell division, etc. (Supplemental Table S2). We then examined the overlaps of the associated strong enhancers from the five cell types. There is a total of 29,734 strong enhancers from the five cell types associated with 4419 (71%) of the above-identified commonly expressed genes. As shown in Figure 1F, most of the associated enhancers to these common genes are unique to one cell type. There are only 231 enhancers common to all five cell types. These results reveal the cell type specificities of enhancers for commonly expressed genes.

Cell type–specific TF binding

These results challenge the previous view that housekeeping genes are regulated by common enhancers across different cell types (Lorberbaum and Barolo 2015; Zabidi et al. 2015). We show that more than half of the housekeeping promoters are linked to cell type–specific enhancers. To explore the mechanism behind this phenomenon, we examined the TF binding patterns at enhancer regions, as enhancers are sites of collective TF recruitment. By comparing the ChIP-seq results of two TFs, KLF4 and MYC, in mouse ESCs and MEF cells, we found that both TFs bind to widely different genomic sites in the two cell types. A snapshot of ChIP-seq signals of the two TFs is shown in Figure 2A. In other words, the same TFs show a cell type–specific binding pattern. There is only a small percentage of overlap of the KLF4 and MYC binding sites between ESCs and MEF cells (Fig. 2B). We examined the binding sites of a few TFs whose ChIP-seq data are available across multiple human cell lines used in the above analysis: the TFs JUND, MAX, CEBPB, ZNF143, YY1, and SP1 in the GM12878, K562, HepG2, H1-hESC, and HeLa-S3 cell lines. For each TF, we examined the number of overlapping binding sites among these five cell lines. The results show that the binding locations of these TFs are mostly cell type–specific (Fig. 2C). Among the six TFs, JUND and CEBPB binding sites have the greatest cell type specificities, whereas ZNF143 binding sites are more shared among the five cell lines. If the promoter regions are excluded, the percentages of shared sites are even lower (Supplemental Fig. S3).

Most TFs have sequence-specific binding motifs. A TF can potentially bind to hundreds of thousands to millions of sites in a mammalian genome. For example, at a *P*-value cutoff of 0.0001, there are 2.1 million KLF4 and 0.68 million ESRRB binding sites identified in the mouse genome based on motif match using the program FIMO (Grant et al. 2011). There is only a small percentage of overlap between the sites that match the motif and the experimentally determined TF binding sites (e.g., ChIP-seq). Cell type–specific TF bindings have been studied before (Arvey et al. 2012; Gertz et al. 2013; Srivastava and Mahony 2020), but it is not fully understood yet what factors determine which subset of genomic sites containing the motifs that a TF would physically bind in a given cell type.

Cell type–specific TF binding results from collective TF binding at regulatory sites

A well-established concept regarding TF binding is the cooperativity and synergy between TFs: Most TFs need to work together to make an effective regulatory site (Smith et al. 2013; Lambert et al. 2018; Ray-Jones and Spivakov 2021; Singh et al. 2021), which we believe plays a critical role in cell type–specific TF binding and, consequently, the cell type specificity of enhancers, as where a TF binds depends on where other TFs are bound, and each cell type has a unique repertoire of TFs. For example, KLF4 and MYC bind to different sites in ESCs and MEF cells. On the other hand, they bind to the same places where other TFs bind in each cell type (Fig. 2D). Using the ChIP-seq data, we examined the cobinding of the 14 ES TFs and nine MEF TFs as follows: Starting with one TF, we added more TFs one by one and compared the sum of the peaks of all TFs and the number of merged peaks (Fig. 2E,F). As new TFs are added, the

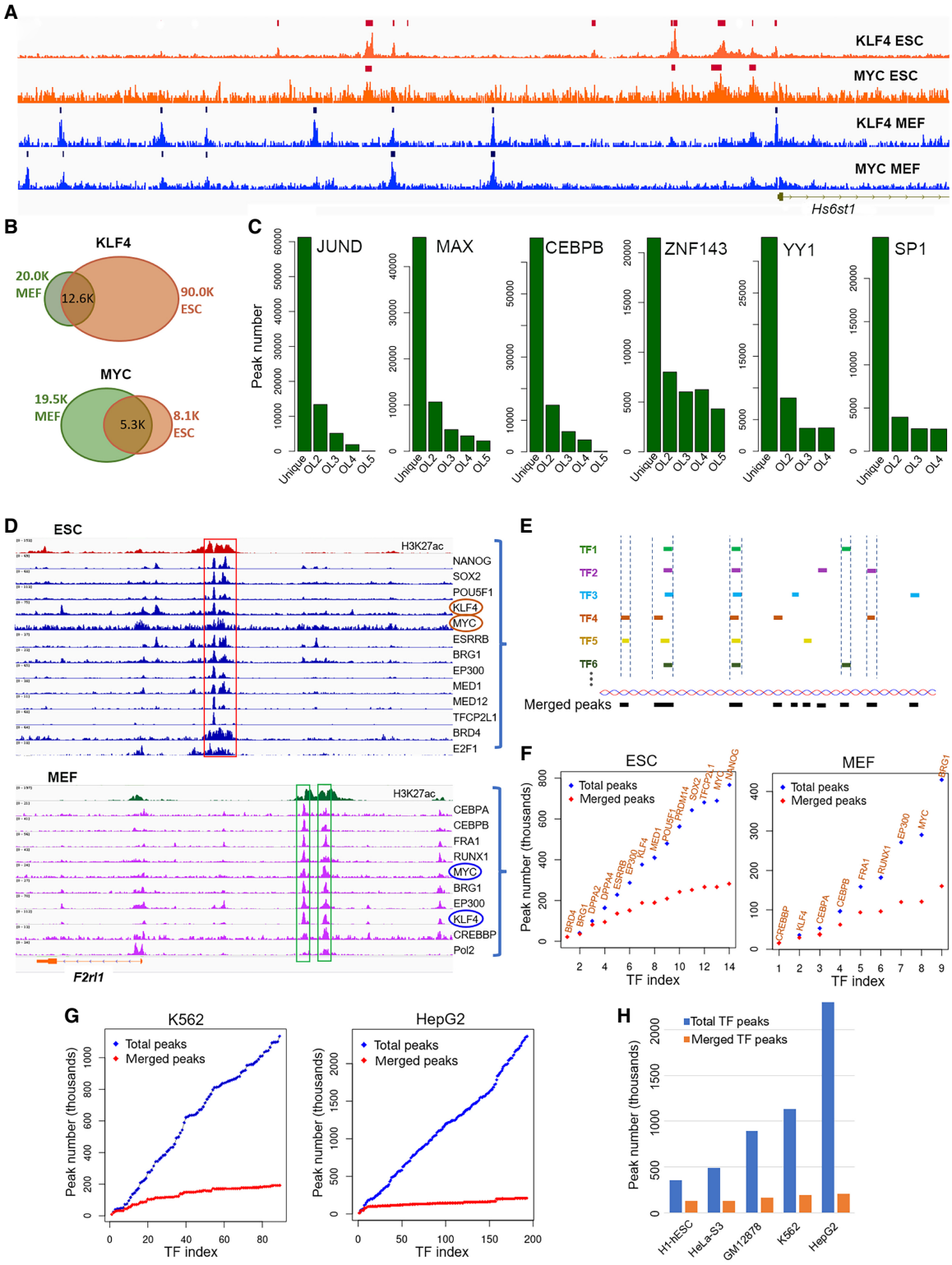

Figure 2. Cell type–specific TF binding and collective TF binding. (*A*) An example of ChIP-seq signals of the TF KLF4 and MYC in mouse ESCs and MEF cells, showing that the same TF binds to different genomic sites in different cell types. The bars in the genome tracks are the called peaks. (*B*) Numbers of total ChIP-seq peaks of KLF4 and MYC in ESCs and MEF cells and their overlaps between the two cells. (*C*) The number of TF ChIP-seq peaks that are unique to one cell type or are shared by two, three, four, and five different cell types, which are K562, HepG2, GM12878, H1-hESC, and HeLa-S3. (OL) Overlap or shared. There is no YY1 or SP1 ChIP-seq data in HeLa-S3, so the comparisons of YY1 and SP1 are only among four cell lines. (*D*) An example of the collective TF binding and cell type–specific binding of KLF4 and MYC in mouse ESCs and MEF cells. (*E*) When peaks from different TFs overlap with each other, we merged them into one peak. (*F*) The total ChIP-seq peaks of all the TFs and the merged peaks as the number of TFs increases, in mouse ESCs and MEF cells. (*G*) An analysis similar to what is presented in *F*, in K562 and HepG2 cell lines. (*H*) The number of total TF peaks and merged peaks in the five different cell lines.

increase of the merged peaks is not at the same rate as that of the total peaks, as different TFs bind together.

We did the same analysis of TF cobinding with data from several human cell lines from the ENCODE Project (https://genome.ucsc.edu/ENCODE/downloads.html), among which K562 and HepG2 cells have the largest number of TFs with ChIP-seq data. Because this study focused on enhancers, we excluded the TFs known to associate with the repression of genes or those that bind to only promoters (e.g., EZH2, HDAC1, HDAC2, TAF1). The sources of all TF ChIP-seq data sets of K562, HepG2, GM12878, HeLa-S3, and H1-hESC cell lines that we used are listed in Supplemental Table S1. The numbers of TFs with available ChIP-seq data we used for the analysis were 39 TFs in H1-hESC, 54 TFs in HeLa-S3, 66 TFs in GM12878, 89 TFs in K562, and 193 TFs in HepG2 cell lines. Figure 2G shows the increase of total TF peaks and the merged peaks in K562 and HepG2 cells. As new TFs were added to the calculation, the number of merged peaks increased at a much slower rate than that of total peaks, meaning that TFs bind to the same sites that are bound by other TFs. A comparison of the number of the total TF peaks and the merged TF peaks across all the five cell lines (Fig. 2H) led to the same conclusion.

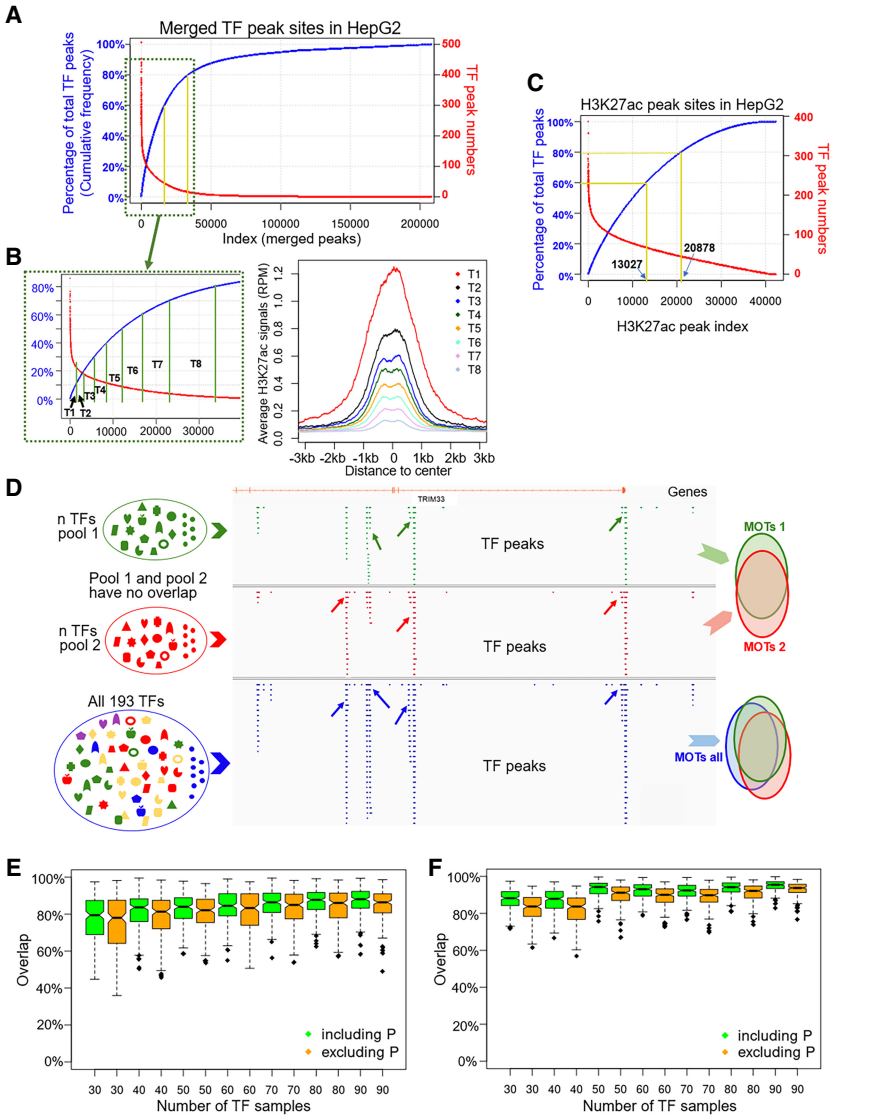

Figure 3. The majority of all TF binding events occur at <10% of all TF sites and can be largely predicted from as few as 30 TFs. (*A*) Merged TF binding sites in HepG2 cells ranked on decreasing TF enrichment (number of the overlapping TF peaks; red line) and the corresponding cumulative frequency curve (blue line). (*B*) From zero to 80% cumulative frequency increased at 10% intervals, the average H3K27ac signal intensity is plotted for the merged TF binding sites at each interval. Green lines mark the points at 10%, 20%, 30%, 40%, 50%, 60%, 70%, and 80% of cumulative levels. (*C*) H3K27ac peak sites ranked on decreasing TF enrichment (red line) and the corresponding accumulative frequency curve (blue line). Peak index at 60% and 80% of accumulative levels is marked. (*D*) From two randomly selected nonoverlapping pools of *n* TFs (*n* varies from 30–90), we identified the top 15,000 most TF-occupied targets (MOTs), as well as the top 15,000 MOTs obtained from all 195 TFs in HepG2 cells. The MOTs are marked with arrows. (*E*) The overlaps of the top 15,000 MOTs between two randomly selected nonoverlapping pools of *n* TFs ($n = 30, 40, 50, 60, 70, 80, 90$). Each box in the plot represents results from 100 times of random TF selection. (*F*) The overlaps of these MOTs and the top 15,000 MOTs obtained from all 193 TFs.

Estimation of the number of the major effective enhancers in a cell

We ranked the merged TF binding sites based on the number of TF peaks at each site (Fig. 3A). Out of 220,000 merged sites in HepG2 cell, 60% of all TF peaks are located at the top 16,556 (8%) most highly occupied merged sites. In addition, 86% of the 16,556 sites are marked by H3K27ac signals, indicating regulatory functions of these sites as promoters or enhancers. Eighty percent of all TF peaks are at the top 33,484 (16%) most highly occupied merged sites. The average H3K27ac signals at these merged sites decrease rapidly as the TF occupancy decreases (Fig. 3B), which suggests that low TF occupied sites likely represent background TF binding. Because effective regulatory elements are places of collective TF recruitment, we made the same plot for the peak sites of H3K27ac, which marks both active enhancers and promoters (Fig. 3C). Eighty percent of all TF peaks that overlap H3K27ac signals are located at the top 20,878 most TF-enriched H3K27ac peak sites, and 12,028 of these do not overlap promoters of any active genes, indicating the function of an enhancer. We did the same analysis in K562 and GM12878 cells and obtained similar results. In K562 cells, 80% of all TF peaks overlapping H3K27ac signal are located at the top 22,449 most TF-enriched H3K27ac sites and 11,595 are enhancers. In GM12878 cells, 80% of the TF peaks are at the top 23,379 most TF-enriched H3K27ac sites and 14,198 are enhancers (Supplemental Fig. S4).

From these results, we estimate that the top approximately 12,000–15,000 most TF-enriched nonpromoter H3K27ac peak

sites represent the majority of effective enhancers in a cell. This is the reason that the number of enhancer sites we chose for the analysis is within this range. Among the major enhancers identified this way (n = 12,028 in HepG2, n = 11,595 in K562, and n = 14,198 in GM12878), only 791 enhancers are shared among all three cell types. The rest of them (>92% in each cell type) are cell type–specific and include the enhancers associated with 57% of the commonly expressed genes identified in the results from Figure 1F.

Binding profiles of a small number of TFs largely predict highly occupied targets

In the above analysis, the most TF-occupied targets (MOTs) were obtained with ChIP-seq data of hundreds of TFs. For most of the cell types used in research, we do not have binding data for this many TFs. We then asked the following question: If we have only a small number of TF ChIP-seq data sets, to what extent can we predict the MOTs to be obtained when more TF binding data are available? We did an experiment with the data from HepG2 cells. From the set of 193 TFs, we randomly selected two nonoverlapping pools of n TFs, for which n varies from 30–90. We then identified the top approximately 15,000 MOTs obtained from the two pools and calculated their overlap (Fig. 3D). There was remarkable overlap between the two MOT sets obtained from two nonoverlapping TF pools. The overlap rates ranged from 65% to 90% with an average overlap rate >80%. When the MOTs overlapping promoter regions were removed, the overlap of the two MOT pools was slightly lower (Fig. 3E). We then calculated how many of these same sets of MOTs from random pools of n TFs overlap with the top 15,000 MOTs obtained from all 193 TFs. The overlap rates ranged from 80% to 95% (Fig. 3F). The results represent 100 times of random TF selection. Thus, the MOTs obtained from a small number of TFs represent, to a great extent, the major regulatory sites in a cell, which are associated with many other TFs. We analyzed 193 TFs in the HepG2 cell line, which is a small fraction of the total about 1000 TFs expressed in the cell (based on RNA-seq analysis). It is a reasonable assumption that if the binding data (ChIP-seq data) of more TFs were available, we would observe the same number of highly occupied sites associated with many more different TFs.

Here we use the term "most occupied targets" (MOTs) to distinguish from the "highly occupied targets" (HOTs) that have been studied from different aspects previously (Yip et al. 2012; Wreczycka et al. 2019; Partridge et al. 2020; Ramaker et al. 2020). We define MOTs as the top n (e.g., n = 15,000) most TF-occupied targets, even if the number of TFs can be relatively small. HOTs are the site of "extremely high TF occupancy," which has been defined inconsistently, and the number of HOT sites varies from about 2000 to about 14,000 in previous studies.

Motif enrichment analysis in mouse ESC and MEF regulatory sites suggests the binding of other highly expressed TFs

Figure 3 shows that the highest occupied sites obtained from the binding profiles of a small number of TFs are very likely to be bound by many other TFs. We further tested this conclusion with motif enrichment analysis in mouse ESCs and MEF cells.

Based on the ChIP-seq data of NANOG, POU5F1, SOX2, KLF4, ESRRB, MED1, MED12, TFCP2L1, and BRG1 in mouse ESCs and of CEBPA, CEBPB, KLF4, MYC, RUNX1, FRA1, BRG1, and CREBBP in MEF cells, we selected the top 15,000 MOTs out of 218,553 merged sites in ESCs and 160,377 in MEFs. We removed sites that overlap promoter regions owing to the potential enrichment of false-positive ChIP-seq signals associated with promoter regions and lack of sequence motifs (Wreczycka et al. 2019), which resulted in 10,560 sites in ESCs and 11,456 sites in MEF cells. Using the JASPAR TFBS enrichment analysis package (https://jaspar.genereg.net/enrichment/) that is based on the LOLA tool (Sheffield and Bock 2016) and the JASPAR 2022 motif database (Castro-Mondragon et al. 2022), we examined the TF motifs enriched in the middle 700-bp window of the above selected sites (Singh et al. 2021). With randomly selected 20,000 700-bp-long regions (excluding promoters) as the control set, the 10 most enriched motif families in ESCs are zinc fingers (including three–zinc finger Kruppel-related, more than three adjacent zinc fingers, and multiple dispersed zinc fingers), AP-2, E2F, NR2, E2A, ETS-related, CP2-related, and steroid hormone receptors. In MEFs, the 10 most enriched families are JUN-related, MAF-related, FOS-related, zinc fingers, nuclear factor 1, E2F, and Hairy-related factors (Fig. 4A). The enriched motif family is associated with highly expressed TFs in each cell type, including SP1, KLF4, TFAP2C, MAZ, ZIC2, TFDP1, PATZ1, ZFP57, RXRB, TCF3, ETV5, TFCP2L1, and ESRRB in ESCs and JUNB, JUND, MAFG, MAFK, BACH1, JDP2, FOSL1, FOSL2, ZFX, ZBTB14, BNC2, NFIX, NFIC, KLF3, TFDP1, and HES1 in MEF cells. Most of these TFs have expression levels >30 RPKM (Fig. 4B).

The above-identified TFs whose motifs are most significantly enriched in these regulatory sites are not the TFs based on which these sites were selected, except for KLF4, TFCP2L1, and ESRRB in ESCs. The results, however, are not completely surprising, as we have shown in Figures 2 and 3 that many different TFs in a cell associate with the same set of highly occupied regulatory sites. The strongest binding sites selected based on ChIP-seq data of fewer than 10 TFs are most enriched in the motifs of some other TFs that are highly expressed in the cell, strongly suggesting the binding of these other TFs at those regulatory sites.

In addition to the above 10 most enriched TF families, the motifs of many other TF families or TFs are significantly enriched (P-value < 1×10^{-5}) in the regulatory sites, including the TFs based on which the regulatory sites were selected (Supplemental Table S3). The expression levels of TFs that have enriched motifs are significantly higher than those that do not have enriched motifs (P-value = 1.117×10^{-8} in ESCs and 8.565×10^{-9} in MEFs, Wilcoxon test) (Fig. 4C).

The regulatory sites for the motif enrichment analysis are selected based only on enrichment of binding signals (ChIP-seq data) of no more than 10 TFs in ESCs or MEF cells, without considering other chromatin signatures or signals. Approximately 80% of these sites, however, overlap with H3K27ac signals (n = 8142 in ESCs; n = 9697 in MEF cells; with <5% overlap), indicating enhancer functions. In ESCs, 8% of these enhancers are associated with genes specifically expressed in ESCs, compared with MEF and B cells (RPKM in ESCs 10 times higher than that of the other two cells); 31% are associated with housekeeping genes (Li et al. 2017); and 46% of them (n = 3713) are associated with commonly expressed genes between ESCs and MEF cells.

Motif composition reveals a feature of high motif flexibility of enhancer grammar

We further examined the motif composition for the cell type–specific enhancers associated with commonly expressed genes between ESCs and MEFs (n = 3713 for ESCs; n = 3941 for MEFs). Based on motif similarity calculated with Pearson correlation

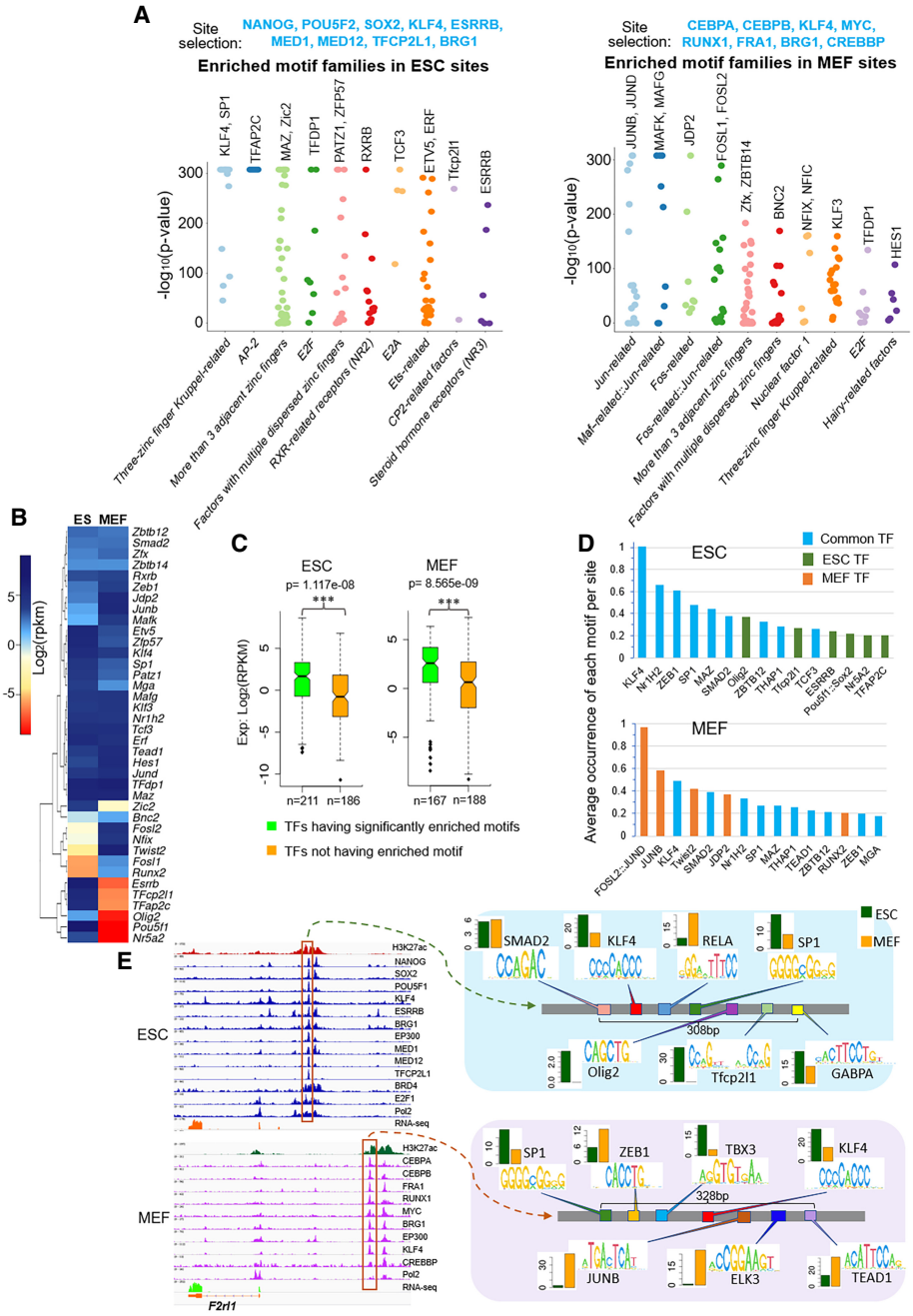

Figure 4. Motif compositions of enhancers in ESCs and MEFs show flexible enhancer grammar. (*A*) The 10 most significantly enriched TF motif families identified in the top TF-bound regulatory sites (nonoverlapping promoters) from mouse ESCs and MEF cells. For each of the 10 motif families, the motifs with the smallest *P*-values associated with highly expressed TFs are listed at the *top*. (*B*) The expression levels of the TFs presented in *A* and *D*, in ESCs and MEF cells. (*C*) The expression levels of TFs that have enriched motifs are significantly higher than those that do not have enriched motifs, with *p*-value = 1.117×10^{-8} in ESC and 8.565×10^{-9} in MEF (Wilcoxon test). (*D*) The average occurrence per site of the 15 most occurring motifs in the ESC- and MEF-specific enhancers (mid 700 bp) for common genes. (*E*) The TF motif composition of an ESC and an MEF enhancer linked to a commonly expressed gene, *F2rl1*. The expression levels (in RPKM) of each TF in the two cells (ESC as green and MEF as orange) are shown as a bar plot *beside* their motifs.

coefficient between the position weight matrixes (PWMs), we reduced the motifs in each enriched TF family into a compact set so that any two motifs in a compact set were significantly different from each other (Pearson correlation coefficient < 0.65). This resulted in 72 and 57 representative motifs in ESCs and MEF cells, respectively (Supplemental Table S4), and in total 97 motifs, as some motifs were enriched in both cell types. We identified these representative motifs in the cell type–specific enhancers (the mid 700-bp window) using the Bioconductor TFBSTools package (Tan and Lenhard 2016), at a strict threshold: similarity of 90% and threshold score of 11. On average, there are 9.8 and 8.6 motifs at each ESC and MEF site, respectively, and the mean distance between adjacent motifs is 48 bp. The motifs that occurred the most are all from highly expressed TFs, from both cell type–specific TFs,

including POU5F1, SOX2, ESRRB, TFCP2L1, NR5A2, TFAP2C, and OLIG2 in ESCs and FOSL2, JUNB, JUND, TWIST2, JDP2, and RUNX2 in MEF cells, and commonly expressed TFs, including KLF4, SP1, NR1H2, MAZ, SMAD2, ZBTB12, TCF3, ZEB1, and MGA (Fig. 4D). Figure 4E shows an example of the motif compositions of cell type–specific enhancers for common genes. The gene *F2rl1* is expressed at similar levels in ESCs and MEF cells and linked to ESC- and MEF-specific enhancers. In both enhancers, there is a cluster of motifs closely located within 330 bp, from both cell-specific TFs (OLIG2, TFCP2L1 in ESCs; JUNB and ELK3 in MEFs) and commonly expressed TFs. Two common motifs are Klf4 and Sp1.

The same analysis for HepG2, K562, and GM12878 cells yielded similar results. The enriched motif families are associated with highly expressed TFs in each cell type, and the cell type–specific enhancer sites for commonly expressed genes contain clusters of motifs from commonly expressed TFs and cell type–specific TFs (Supplemental Fig. S5).

A hypothesis for the mechanism underlying cell type specificity of enhancers for commonly expressed genes

Based on these results, we propose a mechanism underlying the cell type specificity of TF binding and, consequently, cell type specificity of enhancers: An effective enhancer requires many different TFs and other DNA-associated proteins (DAPs) to work cooperatively. A TF tends to bind to locations where many other TFs are bound, even though there can be hundreds of thousands of sites in the genome that harbor the binding motif for the TF. Therefore, the genomic sites where there are optimal combinations of closely located motifs for TFs highly expressed, or relatively highly expressed, in each cell type are likely to "attract" more TFs, and these sites become effective enhancers. Because each cell type has a specific repertoire and expression pattern of TFs, the enhancers occur at cell type–specific loci, even if they target the same genes.

A simplified schematic explanation is presented in Figure 5. Cell types A and B express different collections of TFs. TF X and Y are expressed in both cells but bind to different locations in the two cell types. Although there are many genomic sites that harbor the TF motifs, the actual bindings of each TF are more likely to occur at places where other TFs bind. In this example, the two locations, denoted as enhancer A and enhancer B in the two cell types, have an optimal combination of a diversity of motifs located in proximity and, thus, best facilitate the collective recruitment of TFs expressed in each cell type. This results in cell type–specific binding of TF X and Y, which directly relates to the cell type specificity of the enhancers, even if they target the same genes.

There are up to approximately 1000 different TFs expressed in each cell type (based on our RNA-seq analysis). Every cell type has a unique repertoire and expression pattern of TFs. The association of TFs to a regulatory site is driven by multiple driving forces, including direct DNA binding through motif recognition by TFs, protein–protein interactions between different TFs and cofactors, and chromatin folding and accessibilities. The formation of an effective regulatory site is the result of complicated interplays among different types of interactions from hundreds of different TFs and cofactors. Here we propose a mechanism that underlies part of that complexity: how highly focused TF binding results in cell type–specific usage of enhancers for commonly expressed genes.

Functionally preserved enhancers that share no obvious sequence conservation across different species in the animal kingdom have been previously studied (Blow et al. 2010; Arnold et al. 2014; Villar et al. 2015; Wong et al. 2020). Here we show another type of sequence flexibility but with functional conservation of enhancers: Consistently expressed genes across different cell types are associated with cell type–specific enhancers. Enhancer function is delivered technically by collective TF recruitment. Our results suggest that an essential feature of enhancer grammar is to contain diverse and clustered TFBS motifs while allowing extensive flexibility in motif composition, including their position, order, and spacing. This results in cell type–specific enhancer sites formed through recruitment of a flexible repertoire of TFs uniquely present in each cell type while preserving the function to facilitate the transcription of a nearby gene.

Discussion

In this study, we examined the concept that housekeeping genes associate with common enhancers, which is in contrast to developmental or cell type–specific genes linked to cell type–specific enhancers. Our results expand our knowledge of housekeeping gene regulation. We show that it is common for housekeeping genes and the consistently expressed genes across multiple cell types to be associated with cell type–specific enhancers. As such, there are different types, or different levels, of regulation for the transcription of housekeeping genes: (1) the constitutive or core promoter that is able to drive transcription by its own (Kadonaga 2012; Curina et al. 2017), (2) transcription factories or clustered housekeeping genes that require few enhancers (Sutherland and Bickmore 2009; Zhu et al. 2021), and (3) regulation by cell type–specific enhancers. Previous work by Zabidi et al. (2015) showed that housekeeping genes mostly share enhancers across different cell types (Lorberbaum and Barolo 2015). Our work presents a different result owing to, we think, the following: The work by Zabidi et al. (2015) was conducted

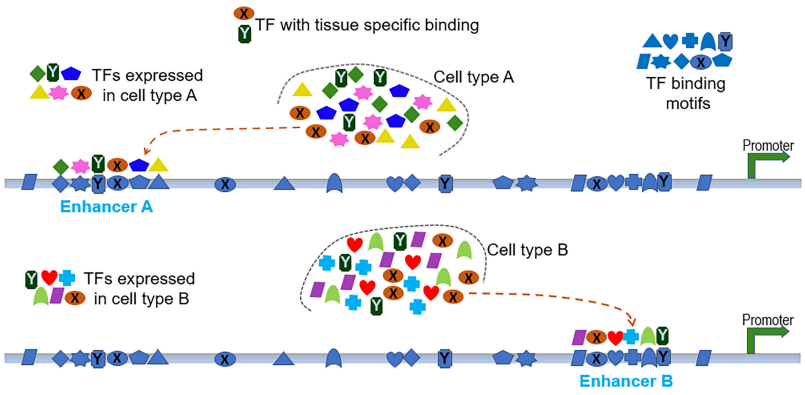

Figure 5. A hypothesis for the mechanism underlying cell type–specific enhancers associated with commonly expressed genes. Schematic representation of the principle: Cell types A and B express different repertoires of TFs. TF X and Y are expressed in both cells but bind to different places, denoted as enhancer A and B in the two cell types, where there are optimal combinations of motifs in proximity that facilitate the collective binding of TFs expressed in each cell type.

on only a few promoters in two *Drosophila* cell lines with the STARR-seq system. In a massively parallel reporter assay like STARR-seq, enhancer functions are observed out of their native context, and the functions may be irrelevant once that context is restored. In the current work, the enhancer identification is based on H3K27ac and TF binding signals, reflecting the chromatin state and function in its native microenvironment. In addition, we examined all the enhancers and their associated promoters genome-widely in multiple mouse and human cell types. On the other hand, there are a small portion of housekeeping promoters using common enhancers across different cell types (Borsari et al. 2021), but there are few such promoters, and they are not the focus of the current work.

Regarding the assignment of target gene to an enhancer, although an enhancer can be far distant from the gene it regulates, we posit that enhancers mostly target their nearest active genes in the majority of cases based on the following evidence: Direct correlation between the signal of an enhancer and the transcription level of its nearby active gene(s) or the nearest active gene has been shown across the mouse or human genome by many studies (Heintzman et al. 2007, 2009; Creyghton et al. 2010; Kim et al. 2010; Andersson et al. 2014; Zhu et al. 2021). There are only a handful of experimentally validated enhancer–promoter pairs in the human or mouse (Wall et al. 1988; Lettice et al. 2003; Li et al. 2014; Zhou et al. 2014; Blinka et al. 2016; Hay et al. 2016; Shin et al. 2016; Moorthy et al. 2017; Gasperini et al. 2020). The target gene of nearly all the in vivo validated enhancers (an exception is the *Shh* gene mediated by an enhancer 1 Mb away) (Lettice et al. 2003) is the nearest active gene, although there might be other annotated genes closer to the enhancer, but these genes are not expressed. In the case of *Nanog* (Blinka et al. 2016), another nearby gene, *Dppa3*, is also regulated by a *Nanog*-linked enhancer, which does not negate the fact that the enhancer is associated with *Nanog*, its nearest gene. By assigning the nearest genes, we did not include in our analysis that there might be other nearby gene(s) in addition to the nearest active gene also regulated by the enhancer, but this does not change the conclusions of this study.

Grammar rules that define how enhancers regulate gene expression have been an active research area for more than two decades. Developmental and cell lineage–linked enhancers have been the focus of the investigations, and three models of how enhancers interact with TFs have been proposed, which are the "enhanceosome" model, the "billboard" model, and the "TF-collective" model (Long et al. 2016; Jindal and Farley 2021). For most of the studied enhancers, the flexibility in arrangement order, spacing, and orientation of the TFs seems to be an essential feature of enhancer grammar. In this work, we present the flexible usage of enhancers that occurs across different cell types in the same species, which is proof at another level for the essential feature of enhancer grammar: flexibility in motif composition and arrangement with preservation of enhancer activities.

For most of cell types that have been studied, the number of TFs with ChIP-seq data available is at single or low-double digits. Currently, there are only a few cell lines for which the ChIP-seq of more than 100 TFs is available, which is an insignificant number compared with about 1000 TFs expressed in each cell type. Methods such as DNase I foot-printing can produce a comprehensive mapping of TF binding events in a cell (Neph et al. 2012; Vierstra et al. 2020), which, however, does not provide substantial information about collective TF binding. In addition, a large portion of the millions of sites in a cell identified through DNase I foot-printing could possibly be background binding with no functional significance. Our results strongly suggest that, based on the binding profiles of a small number of TFs, MOTs are very likely to be bound by many other TFs whose binding data are not available yet. In other words, a study of a small number of TFs reveals, to a large extent, the major regulatory sites associated with hundreds of other TFs. Further proof is required when more data on TF binding become available. Currently, we believe it is a useful guideline to consider for the study of TF binding across genomes.

Methods

ChIP-seq data selection for mouse ESCs and MEF cells

We first examined enhancers from mouse ESCs and MEF cells for two reasons: (1) The samples from which we used the ChIP-seq data were prepared from early embryonic mouse tissue, not transformed or exogenously immortalized cell lines that could differ greatly from their tissue of origin genetically and phenotypically (Alge et al. 2006; Pan et al. 2009; American Type Culture Collection Standards Development Organization Workgroup ASN 2010), and (2) there were multiple high-quality ChIP-seq data (eight to 14 proteins for each) of TFs or other DAPs available for these two cell types. We selected ChIP-seq data based on the quantitative metrics of FRiP >0.1 and NRF >0.8 on a minimum of 10 million uniquely mapped reads per sample (Landt et al. 2012).

We aligned the ChIP-seq and ATAC-seq data to mm10 or hg19 genome using Bowtie 2 (Langmead and Salzberg 2012). For peak calling, we used MACS2 with a cutoff *Q*-value of 0.01 (Zhang et al. 2008; https://github.com/taoliu/MACS/). We mapped RNA-seq data using STAR (Dobin et al. 2013) with following parameters: —outFilterMultimapNmax 20—alignSJoverhangMin 8—alignSJDBoverhangMin 1—outFilterMismatchNoverReadLmax 0.04—alignIntronMin 20—alignIntronMax 1000000—alignMatesGapMax 1000000.

Strong enhancer identification

We identified enhancer and strong enhancers based on H3K27ac and TF enrichment. Enhancers were H3K27ac peaks sites that had no overlap with promoter regions (2 kb upstream of to 1 kb downstream from the TSS). For mouse ES and MEF cells, we used ChIP-seq data of NANOG, POU5F1, SOX2, KLF4, ESRRB, TFCP2L1, DPPA2, E2F1, MYC, EP300, MED1, and BRG1 in ESCs and CEBPA, CEBPB, FRA1, RUNX1, KLF4, MYC, CREBBP, EP300, and BRG1 in MEF cells to further identify strong enhancers. EP300 and BRG1 were not considered as TFs, but for convenience, we refer to all the proteins as TFs in this study. We quantified the strength of enhancers in ESCs and MEF cells based on their TF enrichment: the number of total overlapping TF peaks and the total ChIP-seq signal strength (RPM) from all of the above TFs. The enhancers were ranked first on the total number of overlapping TF peaks. For the site with the same overlapping TF peak numbers, total TF signal strength (reads in peaks per mapped million reads [RPM]) was used as the secondary ranking criteria. The top n (e.g., $n=$ 15,000) strongest enhancers were selected based on the above two criteria.

For the human cell lines (GM12878, K562, HepG2, H1-hESC, HeLa-S3) from the ENCODE Project (https://genome.ucsc.edu/ENCODE/downloads.html), we used the .narrowPeak BED files (the coordinates of called ChIP-seq peaks) for all the TF ChIP-seq data sets (Supplemental Table S1). Enhancers were H3K27ac peak sites ranked by the number of overlapping TF peaks. Because it is common for multiple enhancers to overlap with the same number of TF peaks, which means many enhancers can have the same "strength," it was possible that we could not select exactly the

top n (e.g., $n = 15{,}000$) strongest enhancers. We selected a threshold TF number t so that the number of enhancers overlapping at least t TF peaks were closest to n (e.g., when $t = 15$, $n = 15{,}676$). That is why for the analysis in the human cells, we use this term, the top approximately 15,000 strongest enhancers. In this way, the number of strong enhancers in the five cell lines were 16,708, 15,432, 15,194, 15,206, and 12,214 in HepG2, K562, GM12878, HeLa-S3, and H1-hESC, which were used for the analysis for Figure 1, D through F.

The sources and accession numbers of all the experimental data sets, including ChIP-seq, RNA-seq, and ATAC-seq of human and mouse cells, are listed in Supplemental Table S1.

Except for the ATAC-seq data, the mapping results of all the human cell line data, including the ChIP-seq and RNA-seq data, were downloaded from ENCODE that was mapped to human genome assembly hg19. The mouse cell data are from our previous publication (Zhu et al. 2021) that was mapped to mouse genome assembly mm10. By reperforming the target gene assignment using enhancer elements in hg38 (i.e., GRCh38) or mm39 obtained through liftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver), we confirmed that using the most updated genome assemblies, hg38 and mm39, would not change the conclusions in this study (see Supplemental Methods).

Calculation of peak overlapping rates

Because the numbers of enhancers identified based on a given criteria or a threshold from two different cell types (analysis in Fig. 1B–E) or two pools (analysis for Fig. 3D–F) were typically different, we calculated the overlapping rates between two site sets in the following way: $OL_{\mathbf{rate}} = (N_{\mathbf{OL}}/N_{\mathbf{1}} + N_{\mathbf{OL}}/N_{\mathbf{2}})/2$, in which N_{OL}, N_1, and N_2 were the number of overlap sites, sites in data set 1, and sites in data set 2, respectively.

TFBS motif enrichment and composition analysis

TFBS enrichment analysis was run with the JASPAR tools (https://bitbucket.org/CBGR/jaspar_enrichment/src/master/) using the "twoSets" subcommand. Similar to the analysis conducted in the study of Singh et al. (2021), we chose the center 700-bp windows of the selected regulatory sites and compared them with 20,000 randomly selected 700-bp windows that did not overlap any promoter regions (−2000 bp to +1000 bp of TSS). The result files included 842 TF matrices, their enrichment scores (*P*-value), and the TFBS family to which they belong. TFs not expressed in each cell were removed from further analysis, which resulted in 530–560 TFs in each cell type. Within each identified significantly enriched TF family, it was common that multiple TF motifs were similar to each other. Pairwise comparison of motifs within each TF family was performed, and each family was reduced to a compact set. Within a compact set, the Pearson correlation coefficient of any two motifs was <0.65. The motifs in the compact sets were further used to perform motif composition analysis. TFBS motif comparison and homologous motif identification were performed in Bioconductor with the TFBSTools package (Tan and Lenhard 2016) and JASPAR 2022 motif database (Castro-Mondragon et al. 2022). The Pearson correlation coefficient was calculated with the PWMSimilarity() function. The motif search was performed with searchSeq() function with the parameter "min.score" set to 90%, and only the sites with the score >11 for a JASPAR PWM in the output were identified as a strong match.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This work was supported by the Intramural Research Program at the National Library of Medicine, National Institutes of Health.

References

Alge CS, Hauck SM, Priglinger SG, Kampik A, Ueffing M. 2006. Differential protein profiling of primary versus immortalized human RPE cells identifies expression patterns associated with cytoskeletal remodeling and cell survival. *J Proteome Res* **5:** 862–878. doi:10.1021/pr050420t

American Type Culture Collection Standards Development Organization Workgroup ASN. 2010. Cell line misidentification: the beginning of the end. *Nat Rev Cancer* **10:** 441–448. doi:10.1038/nrc2852

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* **507:** 455–461. doi:10.1038/nature12787

Arnold CD, Gerlach D, Spies D, Matts JA, Sytnikova YA, Pagani M, Lau NC, Stark A. 2014. Quantitative genome-wide enhancer activity maps for five *Drosophila* species show functional enhancer conservation and turnover during *cis*-regulatory evolution. *Nat Genet* **46:** 685–692. doi:10.1038/ng.3009

Arvey A, Agius P, Noble WS, Leslie C. 2012. Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Res* **22:** 1723–1734. doi:10.1101/gr.127712.111

Banerji J, Olson L, Schaffner W. 1983. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell* **33:** 729–740. doi:10.1016/0092-8674(83)90015-6

Barakat TS, Halbritter F, Zhang M, Rendeiro AF, Perenthaler E, Bock C, Chambers I. 2018. Functional dissection of the enhancer repertoire in human embryonic stem cells. *Cell Stem Cell* **23:** 276–288.e8. doi:10.1016/j.stem.2018.06.014

Blinka S, Reimer MH Jr, Pulakanti K, Rao S. 2016. Super-enhancers at the Nanog locus differentially regulate neighboring pluripotency-associated genes. *Cell Rep* **17:** 19–28. doi:10.1016/j.celrep.2016.09.002

Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al. 2010. ChIP-seq identification of weakly conserved heart enhancers. *Nat Genet* **42:** 806–810. doi:10.1038/ng.650

Borsari B, Villegas-Mirón P, Pérez-Lluch S, Turpin I, Laayouni H, Segarra-Casas A, Bertranpetit J, Guigó R, Acosta S. 2021. Enhancers with tissue-specific activity are enriched in intronic regions. *Genome Res* **31:** 1325–1336. doi:10.1101/gr.270371.120

Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Lemma RB, Turchi L, Blanc-Mathieu R, Lucas J, Boddie P, Khan A, Manosalva Pérez N, et al. 2022. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **50:** D165–D173. doi:10.1093/nar/gkab1113

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci* **107:** 21931–21936. doi:10.1073/pnas.1016071107

Curina A, Termanini A, Barozzi I, Prosperini E, Simonatto M, Polletti S, Silvola A, Soldi M, Austenaa L, Bonaldi T, et al. 2017. High constitutive activity of a broad panel of housekeeping and tissue-specific *cis*-regulatory elements depends on a subset of ETS proteins. *Genes Dev* **31:** 399–412. doi:10.1101/gad.293134.116

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29:** 15–21. doi:10.1093/bioinformatics/bts635

Eisenberg E, Levanon EY. 2013. Human housekeeping genes, revisited. *Trends Genet* **29:** 569–574. doi:10.1016/j.tig.2013.05.010

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74. doi:10.1038/nature11247

Field A, Adelman K. 2020. Evaluating enhancer function and transcription. *Annu Rev Biochem* **89:** 213–234. doi:10.1146/annurev-biochem-011420-095916

Gasperini M, Tome JM, Shendure J. 2020. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat Rev Genet* **21:** 292–310. doi:10.1038/s41576-019-0209-0

Gertz J, Savic D, Varley KE, Partridge EC, Safi A, Jain P, Cooper GM, Reddy TE, Crawford GE, Myers RM. 2013. Distinct properties of cell-type-specific and shared transcription factor binding sites. *Mol Cell* **52:** 25–36. doi:10.1016/j.molcel.2013.08.037

Gillies SD, Morrison SL, Oi VT, Tonegawa S. 1983. A tissue-specific transcription enhancer element is located in the major intron of a rearranged

immunoglobulin heavy chain gene. *Cell* **33:** 717–728. doi:10.1016/0092-8674(83)90014-4
Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27:** 1017–1018. doi:10.1093/bioinformatics/btr064
Hay D, Hughes JR, Babbs C, Davies JOJ, Graham BJ, Hanssen L, Kassouf MT, Marieke Oudelaar AM, Sharpe JA, Suciu MC, et al. 2016. Genetic dissection of the α-globin super-enhancer in vivo. *Nat Genet* **48:** 895–903. doi:10.1038/ng.3605
Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39:** 311–318. doi:10.1038/ng1966
Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, et al. 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459:** 108–112. doi:10.1038/nature07829
Heinz S, Romanoski CE, Benner C, Glass CK. 2015. The selection and function of cell type-specific enhancers. *Nat Rev Mol Cell Biol* **16:** 144–154. doi:10.1038/nrm3949
Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, Hoke HA, Young RA. 2013. Super-enhancers in the control of cell identity and disease. *Cell* **155:** 934–947. doi:10.1016/j.cell.2013.09.053
Jindal GA, Farley EK. 2021. Enhancer grammar in development, evolution, and disease: dependencies and interplay. *Dev Cell* **56:** 575–587. doi:10.1016/j.devcel.2021.02.016
Kadonaga JT. 2012. Perspectives on the RNA polymerase II core promoter. *Wiley Interdiscip Rev Dev Biol* **1:** 40–51. doi:10.1002/wdev.21
Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, et al. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465:** 182–187. doi:10.1038/nature09033
Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT. 2018. The human transcription factors. *Cell* **172:** 650–665. doi:10.1016/j.cell.2018.01.029
Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, et al. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* **22:** 1813–1831. doi:10.1101/gr.136184.111
Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9:** 357–359. doi:10.1038/nmeth.1923
Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de Graaff E. 2003. A long-range *Shh* enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* **12:** 1725–1735. doi:10.1093/hmg/ddg180
Li Y, Rivera CM, Ishii H, Jin F, Selvaraj S, Lee AY, Dixon JR, Ren B. 2014. CRISPR reveals a distal super-enhancer required for Sox2 expression in mouse embryonic stem cells. *PLoS One* **9:** e114485. doi:10.1371/journal.pone.0114485
Li B, Qing T, Zhu J, Wen Z, Yu Y, Fukumura R, Zheng Y, Gondo Y, Shi L. 2017. A comprehensive mouse transcriptomic BodyMap across 17 tissues by RNA-seq. *Sci Rep* **7:** 4200. doi:10.1038/s41598-017-04520-z
Long HK, Prescott SL, Wysocka J. 2016. Ever-changing landscapes: transcriptional enhancers in development and evolution. *Cell* **167:** 1170–1187. doi:10.1016/j.cell.2016.09.018
Lorberbaum DS, Barolo S. 2015. Enhancers: holding out for the right promoter. *Curr Biol* **25:** R290–R293. doi:10.1016/j.cub.2015.01.039
Moorthy SD, Davidson S, Shchuka VM, Singh G, Malek-Gilani N, Langroudi L, Martchenko A, So V, Macpherson NN, Mitchell JA. 2017. Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. *Genome Res* **27:** 246–258. doi:10.1101/gr.210930.116
Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK, et al. 2012. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489:** 83–90. doi:10.1038/nature11212
Osterwalder M, Barozzi I, Tissières V, Fukuda-Yuzawa Y, Mannion BJ, Afzal SY, Lee EA, Zhu Y, Plajzer-Frick I, Pickle CS, et al. 2018. Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554:** 239–243. doi:10.1038/nature25461
Palstra RJ, Tolhuis B, Splinter E, Nijmeijer R, Grosveld F, de Laat W. 2003. The β-globin nuclear compartment in development and erythroid differentiation. *Nat Genet* **35:** 190–194. doi:10.1038/ng1244
Pan C, Kumar C, Bohl S, Klingmueller U, Mann M. 2009. Comparative proteomic phenotyping of cell lines and primary cells to assess preservation of cell type-specific functions. *Mol Cell Proteomics* **8:** 443–450. doi:10.1074/mcp.M800258-MCP200
Panigrahi A, O'Malley BW. 2021. Mechanisms of enhancer action: the known and the unknown. *Genome Biol* **22:** 108. doi:10.1186/s13059-021-02322-1
Partridge EC, Chhetri SB, Prokop JW, Ramaker RC, Jansen CS, Goh ST, Mackiewicz M, Newberry KM, Brandsmeier LA, Meadows SK, et al. 2020. Occupancy maps of 208 chromatin-associated proteins in one human cell type. *Nature* **583:** 720–728. doi:10.1038/s41586-020-2023-4
Ramaker RC, Hardigan AA, Goh ST, Partridge EC, Wold B, Cooper SJ, Myers RM. 2020. Dissecting the regulatory activity and sequence content of loci with exceptional numbers of transcription factor associations. *Genome Res* **30:** 939–950. doi:10.1101/gr.260463.119
Ray-Jones H, Spivakov M. 2021. Transcriptional enhancers and their communication with gene promoters. *Cell Mol Life Sci* **78:** 6453–6485. doi:10.1007/s00018-021-03903-w
Sheffield NC, Bock C. 2016. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* **32:** 587–589. doi:10.1093/bioinformatics/btv612
Shin HY, Willi M, HyunYoo K, Zeng X, Wang C, Metser G, Hennighausen L. 2016. Hierarchy within the mammary STAT5-driven Wap super-enhancer. *Nat Genet* **48:** 904–911. doi:10.1038/ng.3606
Singh G, Mullany S, Moorthy SD, Zhang R, Mehdi T, Tian R, Duncan AG, Moses AM, Mitchell JA. 2021. A flexible repertoire of transcription factor binding sites and a diversity threshold determines enhancer activity in embryonic stem cells. *Genome Res* **31:** 564–575. doi:10.1101/gr.272468.120
Smith RP, Taher L, Patwardhan RP, Kim MJ, Inoue F, Shendure J, Ovcharenko I, Ahituv N. 2013. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat Genet* **45:** 1021–1028. doi:10.1038/ng.2713
Srivastava D, Mahony S. 2020. Sequence and chromatin determinants of transcription factor binding and the establishment of cell type-specific binding patterns. *Biochim Biophys Acta Gene Regul Mech* **1863:** 194443. doi:10.1016/j.bbagrm.2019.194443
Sutherland H, Bickmore WA. 2009. Transcription factories: gene expression in unions? *Nat Rev Genet* **10:** 457–466. doi:10.1038/nrg2592
Tan G, Lenhard B. 2016. TFBSTools: an R/Bioconductor package for transcription factor binding site analysis. *Bioinformatics* **32:** 1555–1556. doi:10.1093/bioinformatics/btw024
Vierstra J, Lazar J, Sandstrom R, Halow J, Lee K, Bates D, Diegel M, Dunn D, Neri F, Haugen E, et al. 2020. Global reference mapping of human transcription factor footprints. *Nature* **583:** 729–736. doi:10.1038/s41586-020-2528-x
Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, Park TJ, Deaville R, Erichsen JT, Jasinska AJ, et al. 2015. Enhancer evolution across 20 mammalian species. *Cell* **160:** 554–566. doi:10.1016/j.cell.2015.01.006
Wall L, deBoer E, Grosveld F. 1988. The human β-globin gene 3′ enhancer contains multiple binding sites for an erythroid-specific protein. *Genes Dev* **2:** 1089–1100. doi:10.1101/gad.2.9.1089
Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young RA. 2013. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153:** 307–319. doi:10.1016/j.cell.2013.03.035
Wong ES, Zheng D, Tan SZ, Bower NL, Garside V, Vanwalleghem G, Gaiti F, Scott E, Hogan BM, Kikuchi K, et al. 2020. Deep conservation of the enhancer regulatory code in animals. *Science* **370:** eaax8137. doi:10.1126/science.aax8137
Wreczycka K, Franke V, Uyar B, Wurmus R, Bulut S, Tursun B, Akalin A. 2019. HOT or not: examining the basis of high-occupancy target regions. *Nucleic Acids Res* **47:** 5735–5745. doi:10.1093/nar/gkz460
Yip KY, Cheng C, Bhardwaj N, Brown JB, Leng J, Kundaje A, Rozowsky J, Birney E, Bickel P, Snyder M, et al. 2012. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol* **13:** R48. doi:10.1186/gb-2012-13-9-r48
Zabidi MA, Arnold CD, Schernhuber K, Pagani M, Rath M, Frank O, Stark A. 2015. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518:** 556–559. doi:10.1038/nature13994
Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based Analysis of ChIP-seq (MACS). *Genome Biol* **9:** R137. doi:10.1186/gb-2008-9-9-r137
Zhou HY, Katsman Y, Dhaliwal NK, Davidson S, Macpherson NN, Sakthidevi M, Collura F, Mitchell JA. 2014. A *Sox2* distal enhancer cluster regulates embryonic stem cell differentiation potential. *Genes Dev* **28:** 2699–2711. doi:10.1101/gad.248526.114
Zhu I, Song W, Ovcharenko I, Landsman D. 2021. A model of active transcription hubs that unifies the roles of active promoters and enhancers. *Nucleic Acids Res* **49:** 4493–4505. doi:10.1093/nar/gkab235

Received May 26, 2023; accepted in revised form September 12, 2023.