Positional specificity of different transcription factor classes within enhancers

Sharon R. Grossman,a,b,e Jesse Engreitz,a John P. Ray,a Tung H. Nguyen,a Nir Hacohen,a,d and Eric S. Lander,a,b,e,1

*Broad Institute of MIT and Harvard, Cambridge, MA 02142; bDepartment of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139; cProgram in Health Sciences and Technology, Harvard Medical School, Boston, MA 02215; dCancer Research, Massachusetts General Hospital, Boston, MA 02114; and eDepartment of Systems Biology, Harvard Medical School, Boston, MA 02215

Contributed by Eric S. Lander, June 19, 2018 (sent for review March 26, 2018; reviewed by Gioacchino Natoli and Alexander Stark)

Gene expression is controlled by sequence-specific transcription factors (TFs), which bind to regulatory sequences in DNA. TF binding occurs in nucleosome-depleted regions of DNA (NDRs), which generally encompass regions with lengths similar to those protected by nucleosomes. However, less is known about where within these regions specific TFs tend to be found. Here, we characterize the positional bias of inferred binding sites for 103 TFs within ~500,000 NDRs across 47 cell types. We find that distinct classes of TFs display different binding preferences: Some tend to have binding sites toward the edges, some toward the center, and some at other positions within the NDR. These patterns are highly consistent across cell types, suggesting that they may reflect TF-specific intrinsic structural or functional characteristics. In particular, TF classes with binding sites at NDR edges are enriched for those known to interact with histones and chromatin remodelers, whereas TFs with central enrichment interact with other TFs and cofactors such as p300. Our results suggest distinct region-specific binding patterns and functions of TF classes within enhancers.

Significance

Gene expression is controlled by sequence-specific transcription factors (TFs), which bind to regulatory sequences in DNA. The degree to which the arrangement of motif sites within regulatory elements determines their function remains unclear. Here, we show that the positional distribution of TF motif sites within nucleosome-depleted regions of DNA fall into six distinct classes. These patterns are highly consistent across cell types and bring together factors that have similar functional and binding properties. Furthermore, the position of motif sites appears to be related to their known functions. Our results suggest that TFs play distinct roles in forming a functional enhancer, facilitated by their position within a regulatory sequence.
enriched motifs in the cell type (after removing highly similar motifs) (Materials and Methods and SI Appendix, Fig. S3A). Overall, these enriched motifs corresponded to 103 different TFs across the 47 cell types. As expected, the motif-enrichment profiles were correlated among related cell types (SI Appendix, Fig. S3B).

We then studied the positions of inferred binding sites for each of the 103 TFs relative to the peak of the DHS/ATAC-seq signal in the active regulatory elements (SI Appendix, Fig. S4). Different TFs show strikingly different positional binding-site patterns (Fig. 2 and SI Appendix, Figs. S5 and S6). Some are strongly concentrated at the peak of the DNase/ATAC-seq signal (e.g., CTCF); some are enriched over a more widely distributed central region (e.g., ELF1); some are clustered near the edges of the region (e.g., FOXP1 and ARID3A); and some tend to bind at a specific distance from the center of the NDRs (e.g., EPAS1 and RREB1).

To classify these patterns, we calculated the density profiles in ±200-bp regions around the peak and clustered them using k-medoid clustering (Materials and Methods). The analysis identified six clusters of distinct position patterns (Fig. 3A). The clusters are clearly significant: The mean Kullback–Leibler divergence between density profiles within the same cluster is one to two orders of magnitude smaller than the mean divergence between density profiles in different clusters (Fig. 3B), and the density profiles cannot be explained by local sequence composition (SI Appendix, Fig. S7 A and B). Three of these clusters represent motifs that occur most frequently near the center of NDRs, while the other three clusters tend to occur nearer to the edges (Fig. 3C).

Cluster 1 contains 10 TFs with inferred binding sites that are strongly biased toward the peak of highest DNA accessibility at the middle of the NDR, suggesting that their binding directly shapes local chromatin architecture. For six of these TFs (CTCF, NF-I, C/EBPβ, KLF7, GRHL1, and TFAP2) there is clear functional evidence to support this notion: (i) CTCF induces stably positioned arrays of nucleosomes around its genomic binding sites (12); (ii) NF-I, C/EBPβ, KLF7, and GRHL1 can function as pioneer factors that can establish and maintain chromatin accessibility (13–18); (iii) a recent systematic analysis of the TF-dependent changes in chromatin accessibility induced by the binding of 733 TFs identified CTCF, KLF7, and TFAP2 as having some of the strongest effects on local chromatin accessibility during ES cell differentiation (19); (iv) CTCF, NF-I, C/EBPβ, and GRHL1 show unusually stable binding to DNA and long residence times (14, 20–22); and (v) motifs in cluster 1 have especially strong DNase footprinting signals (SI Appendix, Fig. S8), a feature associated with a slow DNA-binding off-rate (17, 23). The properties of the six TFs may enable them to serve as central anchor points for displacing the central nucleosome, adapting the surrounding chromatin, and stabilizing the NDR and flanking nucleosomes (14).

The remaining three TFs in cluster 1 are nuclear receptors (ESRRB, HNF4A, and PPAR). Unlike the other TFs in the cluster, nuclear receptors are characterized by transient binding to DNA with short residence times (24, 25) and localize almost exclusively to preaccessible chromatin (16, 25–27). Nuclear receptor binding to genomic motif sites is often aided by assisted...
loading by a partner factor, which binds to a site overlapping or adjacent to the nuclear receptor motif site and opens the chromatin (28). Notably, two of the pioneer TFs in cluster 1 (C/EBPβ and NF-1) have been shown to catalyze the assisted loading of several nuclear receptors (13, 16, 29–31). The central location of the nuclear receptor motifs may be related to the assisted loading by pioneer TFs in cluster 1.

Cluster 2 contains 31 TFs whose binding sites also are peaked at the center of the NDR but with a wider distribution than for cluster 1. The cluster is strongly enriched for transcriptional activators [Gene Ontology (GO) category enrichment, \( P_{\text{Benjamini}} = 3.2 \times 10^{-15} \)], such as the activator protein 1 (AP-1) subunits (JUN, FOS, ATF, and MAF factors) and activating factors from the TCF, TEA, RUNX, IRF, and KLF families. Based on known interactions reported in the bioGRID and IntAct databases (32, 33), these TFs are enriched for interactions with numerous transcriptional coactivators, including p300, CREB-binding protein (CBP), YAP1, KDM1A, KAT2B, and WWTR1 (Fig. 4 and SI Appendix, Table S1). Furthermore, the TFs in this cluster interact frequently with each other [average of 1.8 pairwise interactions among the 32 TFs vs. 0.7–1.4 (mean = 1.0) interactions among the TFs in other clusters], suggesting they could cooperatively activate transcription. For example, studies of the IFNβ enhancer have shown that two TFs from this cluster (ATF2 and Jun) bind overlapping motif sites to form a scaffold that recruits CBP/p300 through multidentate interactions (34), leading to synergistic transcriptional activation in response to viral infection (35, 36). Interestingly, TFs in cluster 2 are twice as likely to participate in signaling pathways as the TFs in other clusters [52% of TFs in cluster 2 vs. 16–30% of TFs in other clusters, based on the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (37)] (Fig. 4 and SI Appendix, Table S1), and AP-1 factors are required to maintain accessible chromatin to facilitate the binding of stimulus-regulated TFs (38). Therefore the tightly clustered pattern of these motifs in this cluster may promote cooperativity by both facilitating TF–TF interactions and positioning TFs to form complexes that contact multiple sites on cofactors, thereby allowing enhancers to link multiple signaling pathways and respond in a highly synergistic fashion to specific regulatory cues.

Cluster 3, which contains 25 TFs, also peaks at the center of the NDR with a broader distribution than cluster 2. These TFs are generally characterized by greater cell-type specificity in expression across the 57 cell types profiled in the Epigenomics Roadmap project and greater motif enrichment than the TFs in other clusters (SI Appendix, Table S1). Consistent with this observation, cluster 3 contains numerous TFs that play critical roles in development, including all the homeobox, POU, SOX, ETS, zinc finger, TCF, TEA, and NF-I factors and other TFs than the TFs in cluster 2 (average 5.2 vs. 10.2 interactions per TF). One possible explanation is that the TFs in cluster 3 have fewer annotated interactions with cofactors and other TFs than the TFs in cluster 2 (average 5.2 vs. 10.2 interactions per TF). One possible explanation is that the TFs in cluster 3 have fewer annotated interactions with cofactors and other TFs than the TFs in cluster 2 (average 5.2 vs. 10.2 interactions per TF). One possible explanation is that the TFs in cluster 3 have fewer annotated interactions with cofactors and other TFs than the TFs in cluster 2 (average 5.2 vs. 10.2 interactions per TF). One possible explanation is that the TFs in cluster 3 have fewer annotated interactions with cofactors and other TFs than the TFs in cluster 2 (average 5.2 vs. 10.2 interactions per TF).
Fig. 3. TF motif position patterns fall into six distinct clusters. (A) Motif-density profiles in 400-bp regions centered around the peak of the DHS signal (gray lines) were clustered using k-medoids clustering with k = 6. Density profiles were generated by calculating the frequency of motif occurrences in 20-bp bins tiled every 1 bp in the region. Blue lines depict the smoothed overall density profile of the cluster using the LOESS method. MNase-seq read density (indicating the position of the flanking nucleosomes) is shown by dashed dotted lines for context. (B) Average Kullback–Leibler divergence between the motif-density profiles of pairs of motifs in the same cluster (diagonal boxes) and different clusters (off-diagonal boxes). Motif-density profiles within the same cluster are substantially more similar than those in different clusters. (C) Schematic of NDR structure and motif positions. The arrows indicate the peak of transcriptional initiation estimated from CAGE data. The colored bars represent regions for each cluster with motif densities above the mean. Tick marks occur at 20-bp intervals.
Cluster 4, which contains 16 TFs, is unusual in several respects. The motif profiles show both a central peak and flanking peaks at ~70 bp upstream and downstream. Moreover, many of the motifs in this cluster are asymmetric. When the NDRs are oriented so that the motif occurrences for each TF all appear on the same strand (SI Appendix, Fig. S9A), the motif occurrences in the flanking peaks show a clear preferred orientation relative to the center of the NDRs—that is, one of the two reverse-complementary sequences defining the motifs preferentially points inward (SI Appendix, Figs. S9B and S10A). This bias indicates that one side of the TFs is generally positioned facing the edges of the NDR, while the other side faces the NDR core.

The motifs in cluster 4 are also strongly enriched in promoter-proximal regions (6% of such NDRs contain occurrences for motifs in cluster 4 vs. 1–3% for other clusters) (Fig. 5A). ENCODE ChIP-seq data for 39 TFs in our dataset show greater enrichment in promoter regions for TFs in cluster 4 than for TFs in other clusters (38% of reported peaks within <1 kb of a TSS vs. 8–28% for other clusters) (Fig. 5B). One of the TFs in this cluster, SP1, is a well-characterized promoter-proximal factor that binds GC-rich elements in a wide variety of cellular and viral promoters. Many of the other TFs in the cluster (including SP3, EGR1, EPAS1, ZBTB7B, E2F, KLF15, MEF2C, WT1, and PURA) also bind GC-rich motifs and are known to interact with SP1 at promoters (45–55). We compared the motif-density profiles in NDRs classified as promoter-proximal versus distal enhancers but found them to be indistinguishable (SI Appendix, Fig. S11).

TFs in this cluster are also enriched for interactions with p300 [false-discovery rate (FDR) = 2.6 × 10−6] and Dnmt1, a DNA methyltransferase that plays a key role in maintaining CpG island methylation (56) (FDR = 0.03). Notably, functional studies have demonstrated that SP1 stimulates transcription when bound close to the initiation site but not in distal positions (57, 58), unlike distal enhancer-binding factors from clusters 1–3. These results suggest that SP1 and other TFs in cluster 4 may belong to a distinct functional class of TFs with specialized promoter-associated functions.

Because a key function of promoters is transcription initiation, we hypothesized that the flanking peaks and orientation of TFs in cluster 4 might reflect a role in establishing or stabilizing TSSs at both promoters and enhancers. Recent studies have suggested that, in addition to such features as TATA boxes and INR elements, TF-binding sites also contribute to determining the position of the TSS (9, 59). To examine the relationship of TFs in each cluster with the TSS, we examined cap analysis of gene expression (CAGE) data for both enhancer and promoter-proximal NDRs for 14 of the cell lines in our dataset (60). Transcriptional initiation tends to peak at 50–60 bp from the center of the NDRs (as noted above) (Fig. 1C) and ~50 bp away for TF motif occurrences (Fig. 5C). However, 64% of TFs in cluster 4 and 42% of TFs in cluster 3 (vs. 0–8% in other clusters) show an additional peak of transcriptional initiation ~10 bp away from the location of motifs (EGR1, EGR4, MAZ, PURA, SP1, SP3, ZBTB7B, and ZNF281 from cluster 4 and ELF1, ELF2, ELF5, FLI1, SP1, and SPIC from cluster 3) (Fig. 5C). This observation suggests these TFs play unique roles in positioning the site of initiation.

Cluster 5 contains six TFs whose binding sites are not enriched at the center of NDRs but have peaks at ~60 bp upstream or
downstream. The TFs in this cluster all belong to the FOX family of TFs and include the two best-characterized pioneer factors, FOXA and FOXO. The DNA-binding domain (DBD) of FOX factors structurally resembles the DBD of linker histones H1 and H5 (61, 62), and FOXA factors can compete for binding to linker histone-binding sites, which are located near the edges of the core nucleosome, ~65 bp away from its center (61, 63–65). However, whereas linker histone binding leads to the compaction of nucleosomal arrays, FOXA binding destabilizes nucleosomes and opens the region for binding by other TFs (66–68).

Since enhancer activation typically entails the elimination of a well-positioned central nucleosome (69), motif sites for FOXA and other FOX factors in cluster 5 may be positioned ±60 bp to displace linker histones and destabilize the central nucleosome, helping other TFs bind their target sites.

Finally, cluster 6 contains 14 TFs with binding sites enriched near the edges of the accessible region (80–200 bp from the center), suggesting these TFs could interact with the surrounding chromatin. As with cluster 4, the TFs in cluster 6 have asymmetric motifs and mostly exhibit a preferred orientation relative to the center of the region (SI Appendix, Figs. S9 and S10B), allowing directional interactions with the surrounding nucleosomes and larger chromatin landscape. Consistent with this notion, 10 of the 14 TFs in cluster 6 are known to play roles in chromatin remodeling. These include BPTF, the DNA-binding subunits of nucleosome remodeling factor (NURF), which recognizes H3K4me3 and facilitates ATP-dependent nucleosome sliding (70–72), ARID3A, which facilitates the opening of the IgH enhancer (73–75), and several FOX factors, which interact with histones and mediate recruitment of chromatin remodeling complexes such as SWI/SNF (68, 76). Many of the motifs in this cluster are A/T-rich (SI Appendix, Fig. S10). It is possible that they also recruit additional members of the ARID (A+T-rich interaction domain) family that binds nonspecifically to A/T sequences and has been implicated in chromatin remodeling, including ARID1A/BAF250, the DNA-binding subunit of the BAF chromatin remodeling complex (77).
The TFs in cluster 6 also play roles in nuclear attachment, DNA bending, and DNA unwinding. These TFs are enriched for interactions with the chromatin organizers SATB1 and SATB2, which induce chromatin looping and tether DNA to the nuclear matrix (78, 79). For example, ARID3A binds to sites on the periphery of the IgH enhancer to mediate the attachment of the nuclear matrix (80). Several of the TFs (ARID3A, SRY, and YY1) induce significant DNA bending (74, 81, 82), facilitating TF binding and TF–TF interactions (83, 84). Some (SRY and FUBP1) unwind the DNA double helix, which can promote transcriptional initiation and attachment to the nuclear matrix (82, 85, 86).

To test directly whether TFs in cluster 6 interact with the surrounding nucleosomes, we used MNase-seq data from two cell types (GM12878 and K562) to infer the position of the flanking nucleosomes for each individual NDR and then aligned the TF motifs. For the TFs present in these cell types, we examined the motif distribution relative to the inferred edge of the flanking nucleosome (rather than to the peak of the DHS signal). The TFs in clusters 1–5 did not show peaks of motif sites adjacent the nucleosome edge, but five of the eight TFs in cluster 6 (FOXC1, FOXJ3, FOXO1, FOXP1, and ARID3A) showed a peak (SI Appendix, Fig. SI2). The remaining three TFs (FUBP1, IRF1, and IRES) are not known to play roles in chromatin remodeling.

Finally, we wondered whether certain classes of TFs tend to co-occur in enhancers. To investigate this, we examined whether the distribution of motif sites from each class in the NDRs varied with the presence or absence of motif sites from each of the other classes (Fig. 5D and E and SI Appendix, Table S3). We counted the number of nonoverlapping motif sites from each cluster in the NDRs and calculated the odds ratio (OR) for coenrichment between the motif from each pair of clusters. To control for motif similarities, we also calculated the baseline OR for each pair of clusters in shuffled sequences. Significantly coenriched or codepleted cluster pairs were defined as pairs for which the OR falls outside the 95% CI of the OR in shuffled sequences (SI Appendix, Table S3). We found that all six clusters showed significant preferences for coenrichment and codepletion with specific other clusters (SI Appendix, Table S3). For example, regulatory elements with TF motif sites in cluster 4 (associated with TSS-related functions) contain significantly more motif sites from cluster 3 (associated with cell type–specific activation) (Fig. 5D) and significantly fewer motif sites from clusters 5 and 6 (associated with nucleosome remodeling and chromatin architecture) (Fig. 5E) than regulatory elements without cluster 4 motifs. Importantly, these cluster associations are consistent across cell types, even though the specific set of TFs active in each cell type differs (SI Appendix, Fig. SI3). Thus, the TF clusters may constitute a general regulatory code, with different cell types substituting specific TFs to activate different sets of enhancers.

It has long been suggested that TFs may belong to different functional classes. In some cases, prior biological knowledge of certain TFs has been used to categorize TFs into classes, such as pioneer factors that have the capability to bind motif sites in closed chromatin versus nonpioneer factors that bind motif sites only in open chromatin and cell type–specific versus ubiquitous factors. However, there have been few systematic approaches to recognize distinct classes and properties independent of the known biological properties of the individual TFs. One such functional study was recently performed in Drosophila, in which investigators asked which TFs could substitute for each other across a variety of regulatory contexts (10).

Here, we show that, solely by looking at the positional distribution of motif sites within NDRs, we are able to recognize six distinct classes of TFs. These classes bring together factors that have a number of similar properties, such as binding stability, interactions with other TFs and cofactors, cell-type specificity, and pioneering ability. Furthermore, the position of motif sites appears to be related to their known functions—for example, localizing pioneer factors to the optimal positions to displace nucleosomes and targeting chromatin remodelers in close proximity to flanking nucleosome.

The degree to which the arrangement of motif sites within regulatory elements determines their function remains an open question. At one end of the spectrum, there are examples of enhancersomes, such as the IFNβ enhancer, that are exquisitely sensitive to the spacing and orientation of the motif sites (34, 87, 88). However, the activity of other regulatory elements, referred to as “billboard” enhancers, appears to be relatively insensitive to the arrangement of motif sites (89–91). Instead, our work suggests a different kind of constraint, whereby TFs play distinct roles in forming a functional enhancer, facilitated by their position within a regulatory sequence.

The classes identified here also help shed light on the properties of some less characterized TFs. For example, they suggest that several other FOX factors in cluster 5 may use a mechanism similar to that of FOXA1 to displace nucleosomes and that the uncharacterized zinc finger TFs in cluster 6 (ZNF148, ZNF202, and ZNF35) may have pioneering abilities. In addition, the positional preferences identified may prove useful in building predictors of enhancer activity and recognizing functional enhancers in genomic sequence.

Here, we focus on the classes of TFs, these results naturally raise the question of whether different functional classes of enhancers are formed based on these classes of TFs. Identifying such enhancer classes may shed light on the classes of TFs that must come together to accomplish all the functions necessary to build a functional enhancer. In addition, to helping us understand natural enhancers, better knowledge about the constraints and the functional implications of TF positions may aid in creating synthetic enhancers with specific properties that can be used in synthetic biology.

Materials and Methods
ATAC-Seq for Jurkat and U937 Cell Lines. Cells were washed with ice-cold FACS buffer and were kept on ice until cell sorting. Twenty-five thousand live cells from each condition were sorted into FACS buffer and were pelleted by centrifugation at 500 × g for 5 min at 4 °C in a precooled fixed-angle centrifuge. Cell lines then were digested according to the previously described Fast-ATAC protocol (92). Briefly, all supernatant was removed with care not to disturb the not-visible cell pellet. Transposase mixture (50 μL: 25 μL of 2× TD, 2.5 μL of TDE1, 0.5 μL of 1% digitonin, 22 μL of nuclease-free water) (catalog no. FC-121-1030, Illumina; catalog no. G9441, Promega) was added to the cells, and the cells were dissolved by pipetting. Transposition reactions were incubated at 37 °C for 30 min in an Eppendorf ThermoMixer with agitation at 300 rpm. Transposed DNA was purified using a Qiagen MinElute Reaction Cleanup kit (catalog no. 28204), and purified DNA was eluted in 12 μL of elution buffer (10 mM Tris-HCl, pH 8). Transposed fragments were amplified and purified as described previously (93) with modified primers (94). Libraries were quantified using qPCR before sequencing. All Fast-ATAC libraries were sequenced using paired-end, dual-index sequencing on a NextSeq sequencer (Illumina) with 76 × 8 × 8 × 76 cycle reads at an average read depth of 30 million reads per sample.

Definition of NDRs. To define NDRs for our analysis, we used DNASel-seq and H3K27ac ChIP-seq data for 45 cell types in the Epigenomics Roadmap and ENCODE Projects (1, 60). We supplemented this dataset with ATAC-seq data for Jurkat and U937 cells generated in the N.H. laboratory and H3K27ac ChIP-seq data for Jurkat and U937 cells from studies deposited in the National Center for Biotechnology Information Gene Expression Omnibus database (accession no. SRR1057274) (95) and the European Nucleotide Archive database (accession no. ERR671846), respectively. We aligned the ATAC-seq and H3K27ac data for Jurkat and U937 cells as described in ref. 96 and called peaks using MACS2 (97) with the standard parameters used by the Epigenomics Roadmap Project. To define our initial set of NDRs, DHS/ATAC-seq narrowPeaks regions and H3K27ac gappedPeaks regions. We then filtered out NDRs that were present in more than 24 (50%) of the cell types in our analysis and selected the top 7,500 cell type–restricted NDRs for
motif enrichment and positioning analysis. We defined the coordinates in the NDRs relative to the summit called by MACS2 (i.e., the position with the maximum DHS:ATAC-seq signal). For NDA-seq analysis, we used data from GM12878 and K562 cells generated by the ENCODE project. The center of the nucleosomes flanking the NDRs was identified by estimating the position with the highest MNase-seq read coverage in the 300 bp upstream and downstream of the peak of the DHS signal.

Motif Enrichment Analysis. We calculated motif counts for all vertebrate motifs in TRANSFAC (JASPAR) (99), and CIS-BP (100) in the genomic NDR sequences as well as scrambled genomic NDR sequences (holding dinucleotide frequencies constant). To identify enriched motifs in each cell type, we used AME (101) with the mhg method to calculate the enrichment of the total number of matches of each motif in the genomic sequences compared with the scrambled sequences. When the combined databases contained multiple position weight matrices (PWMs) corresponding to a single TF, we selected the most enriched motif in each cell type corresponding to each TF. To remove highly similar motifs, we calculated the pairwise similarity of the motifs using the R package PWMEnrich and removed motifs that had a similarity of >0.8 with a more highly enriched motif. We then selected the top 20 motifs from the filtered list in each cell type for positioning analysis. We called motif sites in the genomic and scrambled sequences by running FIMO (102) with a P-value threshold of 10^{-4}.

Motif-Position Profiles and Clustering. To analyze the positioning of the motifs with NDRs, we collapsed the motif matches to their central position and calculated the coverage of each motif in 20-bp windows tiled every 1 bp across the 400 bp centered around the position of maximum DHS/ATAC signal in each NDR. The motif-position profiles were then clustered using the pam function from the R package cluster with k = 6.

To assess how much each motif-position profile is due to the variation in dinucleotide content across the regions, we calculated the background motif frequency profiles by modeling DNA profile magnitude and shape. We then subtracted the dinucleotide content at each position constant, and normalized the genomic-density profiles by subtracting out the background motif frequencies (Supplementary Fig. S7B).

TF Cluster Feature Enrichment Analysis. Enrichment analysis was performed using DAVID (103) for each of the six clusters for four types of features: protein domains (PFAM, PIR, and SMART), functional annotations (GO and Enrichr gene), protein–protein interactions (BioGRID interaction and intact databases), and pathways (KEGG and BioCarta). P-values were calculated using the Benjamini correction for multiple testing.

TF Coenrichment Analysis. We tested for coenriched and codelocated motifs from the six TF motif clusters in genomic NDR sequences using a Fisher exact test. For each pair of cluster A and cluster B, we calculated the OR that a genomic sequence contains a motif from cluster B, conditional on the presence of a motif from cluster A. To control for motif similarities between motifs in different clusters, we also calculated the same OR in scrambled sequences (holding dinucleotide content constant). To identify significantly coenriched or codelocated pairs, we selected pairs for which the 95% CI of the genomic OR did not overlap the 95% CI of the shuffled OR.

ACKNOWLEDGMENTS. We thank Karen Adelman, Telmo Henriques, Bradley Bernstein, Aivis Regev, Cigall Kadoch, Seth Cassel, and Kaylyn Williamson for valuable comments and discussion and Ray Louis for help with DNase footprint analysis. This work was supported by National Human Genome Research Institute Grant 2ZU4HG003067-10 (to E.S.L.) and National Institute of General Medical Sciences Grant T32GM007753 (to S.R.G.).


