Dissecting the regulatory activity and sequence content of loci with exceptional numbers of transcription factor associations

Ryne C. Ramaker, Andrew A. Hardigan, Say-Tar Goh, et al.

Genome Res. 2020 30: 939-950 originally published online July 2, 2020
Access the most recent version at doi:10.1101/gr.260463.119

Research

Dissecting the regulatory activity and sequence content of loci with exceptional numbers of transcription factor associations

Ryne C. Ramaker,[1,2,4] Andrew A. Hardigan,[1,2,4] Say-Tar Goh,[3] E. Christopher Partridge,[1] Barbara Wold,[3] Sara J. Cooper,[1] and Richard M. Myers[1]

[1]*HudsonAlpha Institute for Biotechnology, Huntsville, Alabama 35806, USA;* [2]*Department of Genetics, University of Alabama at Birmingham, Birmingham, Alabama 35294, USA;* [3]*California Institute of Technology, Division of Biology and Biological Engineering, Pasadena, California 91125, USA*

DNA-associated proteins (DAPs) classically regulate gene expression by binding to regulatory loci such as enhancers or promoters. As expanding catalogs of genome-wide DAP binding maps reveal thousands of loci that, unlike the majority of conventional enhancers and promoters, associate with dozens of different DAPs with apparently little regard for motif preference, an understanding of DAP association and coordination at such regulatory loci is essential to deciphering how these regions contribute to normal development and disease. In this study, we aggregated publicly available ChIP-seq data from 469 human DAPs assayed in three cell lines and integrated these data with an orthogonal data set of 352 nonredundant, in vitro–derived motifs mapped to the genome within DNase I hypersensitivity footprints to characterize regions with high numbers of DAP associations. We establish a generalizable definition for high occupancy target (HOT) loci and identify putative driver DAP motifs in HepG2 cells, including HNF4A, SPI, SP5, and ETV4, that are highly prevalent and show sequence conservation at HOT loci. The number of different DAPs associated with an element is positively associated with evidence of regulatory activity, and by systematically mutating 245 HOT loci with a massively parallel mutagenesis assay, we localized regulatory activity to a central core region that depends on the motif sequences of our previously nominated driver DAPs. In sum, this work leverages the increasingly large number of DAP motif and ChIP-seq data publicly available to explore how DAP associations contribute to genome-wide transcriptional regulation.

[Supplemental material is available for this article.]

Gene expression networks underlie many cellular processes (Spitz and Furlong 2012). These expression networks are controlled in *cis* by DNA regulatory elements, such as promoters and enhancers, which can be proximal, can be distal, or can be within their target genes in a given expression network. Extensive mapping of epigenetic modifications and 3D chromatin structure has provided an increasingly rich set of clues to the locations and physical connections among such elements. Nevertheless, these biochemical signatures cannot yet accurately predict the presence or amount of regulatory activity encoded in underlying DNA. There are many known and suspected reasons for this difficulty, including the relative strength, number of interacting partners, and redundancy of each element, each of which may modulate a locus' contribution to the native expression level(s) of its respective target gene(s) in a manner difficult to predict without direct experimentation (The ENCODE Project Consortium 2007, 2012; Sanyal et al. 2012; Roadmap Epigenomics Consortium et al. 2015). In this paper, we present evidence that the total number of DNA-associated proteins (DAPs) that associate with a locus can act as a quantitative predictor of the locus' regulatory activity and that the activities of loci with large numbers of DAP associations can be disrupted in a predictable manner by altering subsets of putative "driver motifs."

Classically, regulatory loci are thought to be discriminately bound by a small subset of expressed transcription factors (i.e., fewer than 10) in a manner governed by each factor's DNA sequence preference, and additional proteins are recruited through specific protein–protein interactions (Mitchell and Tjian 1989). However, this model is becoming incongruent with observed DAP associations as catalogs of genome-wide DAP binding maps continue to expand (Foley and Sidow 2013). Specifically, the discriminatory nature by which regulatory regions recruit DAPs is unclear at thousands of loci that have been shown to associate with dozens of different DAPs with seemingly no regard for motif preferences (Teytelman et al. 2013; Jain et al. 2015; Ramaker et al. 2017; Wreczycka et al. 2019). These loci, which have associations with dozens of DAPs, have been inconsistently defined but are broadly referred to as high occupancy target (HOT) sites. This phenomenon has been at least partly attributed to technical artifacts of chromatin immunoprecipitation sequencing (ChIP-seq), a common assay used to map DNA–protein interactions in vivo, resulting in a small number of blacklisted loci (Johnson et al. 2007; Landt et al. 2012; Carroll et al. 2014). These artifacts have largely been localized to regions of the genome for which it is difficult

[4]**These authors contributed equally to this work.**
Corresponding author: rmyers@hudsonalpha.org
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.260463.119. Freely available online through the *Genome Research* Open Access option.

to confidently align sequencing reads, such as repetitive elements (Landt et al. 2012; Carroll et al. 2014). Others found potentially misleading or nonspecific ChIP-seq signal more broadly at GC-rich promoters of highly expressed genes (Wreczycka et al. 2019). These findings provide motivation for proceeding with caution when analyzing DAP coassociations, particularly at HOT loci. In our analysis of these genomic regions, we present an extensive examination of potentially confounding characteristics of HOT loci, used conservative peak calling thresholds standardized by The ENCODE Consortium, and rely heavily on orthogonal, non-ChIP-seq-based data sets to define DAP associations. Despite this conservative approach, we find complex DAP coassociations to be pervasive throughout the genome, and the increasing completeness of our catalog of DAP occupancy maps, generated by ChIP-seq and other orthogonal approaches, invites a systematic investigation of the prevalence and significance of DAP coassociations and of the classic model for how DAPs interact with regulatory elements.

Previous work has investigated HOT loci using a combination of genome-wide transcription factor motif scanning and ChIP-seq experiments (Foley and Sidow 2013; Li et al. 2015, 2016). These studies have found thousands of loci harboring dozens of motif or ChIP-seq peak-based DAP associations throughout the genome and have labeled these regions as HOT loci. HOT loci were found to be enriched for markers of regulatory activity, such as initiating POLR2 binding, DNase I hypersensitivity, active histone marks, and strong activity in enhancer reporter assays conducted in transgenic mouse embryos. These studies also showed that context-specific HOT loci are generated in association with cell differentiation and oncogenesis at locations enriched for disease-risk variants. However, these studies were largely limited to experimental data from fewer than 100 transcription factors derived from several different cell lines. Previous studies also have not incorporated genome-wide 3D chromatin structure data, such as chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) and promoter capture Hi-C experiments, which have been performed on an increasing number of cell lines. Massively parallel reporter assay (MPRA) data that probes the regulatory activity at thousands of loci across the genome are also now available to assess for quantitative correlation with DAP associations. Furthermore, few studies have performed experimental perturbations on HOT loci to quantify their vulnerability to single-base-pair mutations and to probe the key sequence features driving their activity.

In this paper, we aim to (1) detail the prevalence and cell type specificity of regulatory element DAP coassociations using the extensive amount of publicly available ChIP-seq and DAP motif data currently available, (2) assess the utility of coassociations as a marker of regulatory activity, (3) perform a high-resolution dissection of key sequences driving activity at regions with large numbers of DAPs by using a massively parallel mutagenesis assay, and (4) explore potential factors influencing observed DAP coassociations, such as 3D chromatin interactions, sequence content, and copy number variation (CNV).

Results

HOT loci are prevalent in the genome

We used two orthogonal methods to infer DAP associations across the genome. The first involved analysis of ENCODE ChIP-seq peaks (208, 129, 312 DAPs in the HepG2, GM12878, K562 cell lines) (Supplemental Table S1). A subset of DAPs was further classified into sequence-specific transcription factors (ssTFs; $N = 117$) and non-sequence-specific DAPs (nssDAPs; $N = 85$). ssTFs were conservatively defined as those that had an in vitro–derived motif in the Cis-BP database (Weirauch et al. 2014), and nssDAPs were defined as DAPs without in vitro–derived motifs that had previously been characterized as non-sequence-specific chromatin regulators or transcription cofactors (Lambert et al. 2018; Partridge et al. 2020). As a second method to assess transcription factor associations, we used the protein interaction quantitation (PIQ) algorithm and in vitro–derived (SELEX, protein binding microarray, or B1H) motifs from 555 TFs in the Cis-BP database to identify DAP footprints that were present in ENCODE DNase I hypersensitivity (DHS) footprints (Supplemental Table S2; Sherwood et al. 2014). To quantify DAP coassociations, we binned the genome into a minimal set of nonoverlapping 2-kb loci that encompassed either every ChIP-seq peak or every distinct DHS footprint and counted the number of unique DAP peaks or footprinted motifs contained within each locus (Supplemental Tables S3–S6). We focused on HepG2 as the primary cell line in our analysis, and the figures in this paper contain HepG2-derived data unless otherwise specified.

To ensure that our definition of a "HOT" locus was generalizable across cell lines and data sets, we defined HOT regions as those associated with at least 25% of DAPs assayed. This definition requires 52 of 208 DAPs assayed with ChIP-seq in the HepG2 cell line to have a peak at a given locus to reach the HOT threshold. Nearly 6% of loci (13,792 out of 244,904) met this HOT threshold in HepG2, and we found this result to be consistent after varying the number of DAPs incorporated into our analysis via random sampling (Fig. 1A; Supplemental Fig. S1A–C). We found our 25% threshold to be preferable to other HOT thresholds based on the stability of number of loci detected and recall performance of the full data set in a series of random down-samples (Supplemental Fig. S1B,C). This threshold also performed similarly in the K562 and GM12878 cell lines (Supplemental Fig. S1D–F). The distribution of observed DAP coassociations was different than that observed after randomly scrambling DAPs across all loci (K-S test $P < 5 \times 10^{-16}$), with no locus reaching our HOT threshold by random chance (Supplemental Fig. S1G). A subset of our HOT loci fall under a previously established definition of super enhancers (Whyte et al. 2013); however, no HOT promoters and the vast majority (97.1%) of HOT enhancers are not encompassed by this definition (Supplemental Table S7; Supplemental Fig. S1H,I). The observed pattern of DAP coassociations was relatively consistent when restricting to ssTFs or nssDAPs (Fig 1B; Supplemental Fig. S1A), although a slightly larger proportion of ssTF peaks were found at a locus alone (44.0% vs. 34.4%) or with a relatively small number of coassociated ssTFs. No locus had ≥25% of the 352 nonredundant HepG2 DHS-footprinted motifs (DFMs) analyzed, suggesting the number of possible motifs at a locus is constrained in a manner not observed for ChIP-seq peaks. However, the number of DFMs was positively correlated with the number of gross DAP peaks across loci (rho = 0.494, $P < 5 \times 10^{-16}$) despite a minority (~10%) of ssTFs with ChIP-seq peaks present at any given HOT site possessing a corresponding DFM at the same locus (Fig. 1C). These data suggest that, although the presence of DFMs is a strong indicator of HOT loci, a majority of DAP associations at HOT loci likely represent non-sequence-specific or indirect interactions.

Although HOT sites represent a small minority of DAP-associated loci, because of the massive number of DAPs that localize to

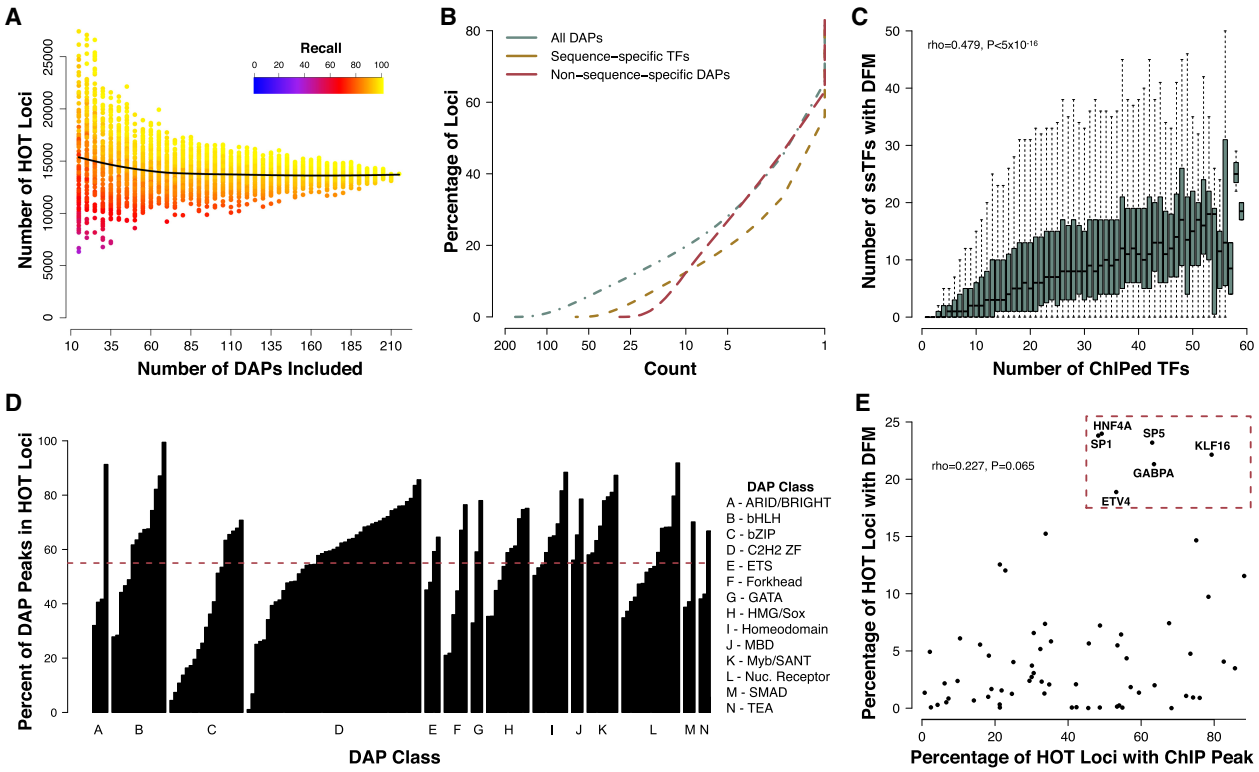

Figure 1. HOT loci are prevalent throughout the genome. (*A*) Number of loci reaching "HOT" threshold of 25% of unique ChIP-seq peaks after performing random down-sampling (from the original 208) of the number of DAPs included. Each data point represents the result of a random sampling of a specified number of DAPs. The color indicates the recall performance or the percentage of true HOT sites, as defined by >25% of DAPs bound in the full data set, detected with current sample of DAPs. The black line represents the median result of 100 random samples of each number of DAPs as specified by the *x*-axis. (*B*) Cumulative distribution function (CDF) showing the proportion of loci containing at least a given number of unique DAP ChIP-seq peaks in HepG2. The green line shows data for all 208 DAPs; the red dashed line, data for nssDAPs; and the yellow dashed line, data for ssTFs. (*C*) Boxplots showing the number of ChIP-defined DAPs with a corresponding DFM present at the same locus at various levels of DAP coassociation. (*D*) Barplots indicating the fraction of ChIP peaks for each DAP that fall within HOT loci. Bars are grouped by previously defined DAP classes. The dashed red line indicates the average fraction (55%) of ChIP peaks that fall within a HOT locus across all DAPs. (*E*) Scatter plot showing the fraction of HOT sites that contain a ssTF ChIP-seq peak and a DFM. ssTFs highlighted in the *top right* are putative driver TFs present at high proportion of HOT sites.

these sites, they account for 55% of any individual DAP's ChIP-seq peaks on average, potentially complicating the interpretation of any individual ChIP-seq data set. We observed a wide range in the rate of participation in HOT loci within previously defined DAP classes, but DAPs with a methyl-binding domain (MBD), a Myb/SANT domain, or a homeodomain show the highest rates of HOT site participation (Fig. 1D). These classes have been previously described as having an affinity for large multiprotein complex membership, such as the NuRD complex, and are plausible candidates to be indirectly recruited to HOT loci (Underhill et al. 2000; Basta and Rauchman 2015). A small number of ssTFs, including SP1 and SP5, which bind GC-rich sequences; HNF4A, a key driver of liver cell differentiation; GABPA and ETV4, which belong to the ETS family of ssTFs; and KLF16 had DFMs at an exceptional number of HOT sites (Fig. 1E; Tan and Khachigian 2009; Wei et al. 2010; DeLaForest et al. 2011). Many of these ssTFs have been implicated as drivers of liver expression programs and thus can be reasonably nominated as putative "drivers" of HOT sites in HepG2, a liver cancer–derived cell line (DeLaForest et al. 2011). Despite rampant coassociations of DFMs, we observed little evidence for specific cooperation among these driver ssTFs as HOT loci were roughly three times more likely to have only one of the HNF4A, GABPA, or SP1 DFMs present rather than any combination of the three (Supplemental Fig. S1J).

HOT loci are enriched for promoter and enhancer regions near highly expressed genes

After establishing the prevalence of HOT loci, we investigated the biological significance of loci with a large number of DAPs. By intersecting these loci with previously assigned HepG2 genomic annotations, we found a continuous relationship between the number of DAPs, identified as ChIP-seq peaks or DFMs, and enhancer or promoter designation from the IDEAS genome segmentation algorithm (Fig. 2A; Supplemental Fig S2A; Zhang et al. 2016). Loci containing a large number of DFMs were particularly enriched for promoters over other annotations (Supplemental Fig. S2A). Roughly half of all IDEAS promoters and likely enhancers in HepG2 met our HOT loci threshold, whereas genomic regions with other annotations rarely met this threshold (Supplemental Fig. S2B). Less than 2% of loci without an enhancer or promoter annotation met our HOT threshold (Supplemental Fig. S2B).

To assess the regulatory activity of loci as a function of the number of unique DAPs, we used a variety of publicly available

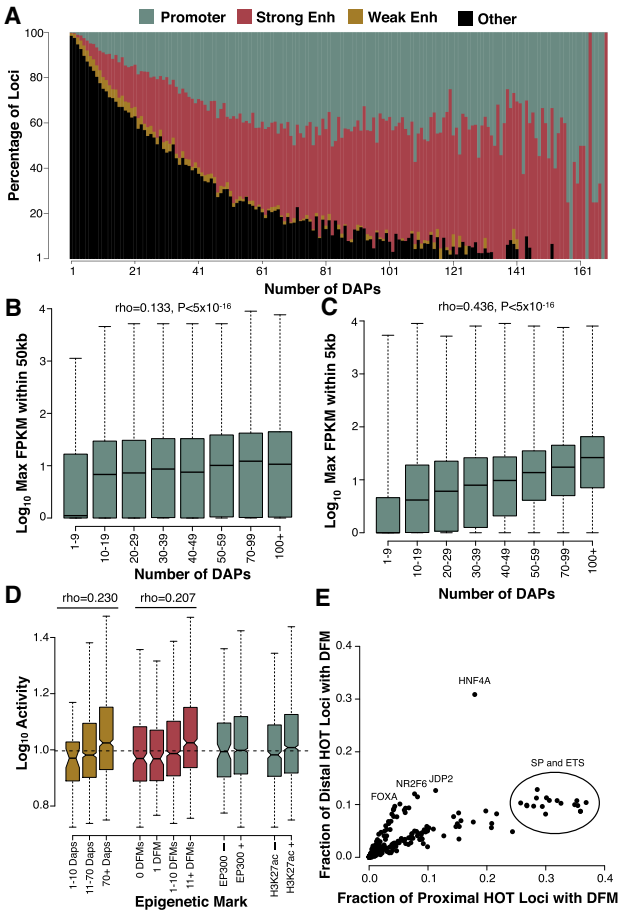

Figure 2. HOT loci are enriched for promoter and enhancer regions near highly expressed genes. (*A*) IDEAS annotations of loci binned by ChIP-defined DAP associations. Promoter, strong enhancer, and weak enhancer annotations represent 0.27%, 0.35%, and 0.22% of the HepG2 genome, whereas the remaining 99.16% of the genome (largely consisting of quiescent and repressed annotations) was used for the "other" annotation. (*B*,*C*) The expression level of the maximally expressed gene neighboring each locus binned by the number of ChIP-defined DAP associations. Plots show loci either distal (>5 kb; *B*) or proximal (<5 kb; *C*) to their nearest gene. The sample size of each bin is as follows: 1–9 (N = 194,028), 10–19 (N = 17,148), 20–29 (N = 8685), 30–39 (N = 5876), 40–49 (N = 4578), 50–69 (N = 6532), 70–99 (N = 5351), 100+ (N = 2706). (*D*) ChIP- and DFM-defined coassociation correlates with activity in a previous high-throughput reporter assay conducted on approximately 2000 selected enhancer regions in HepG2. (*E*) Scatter plots showing the fraction of distal (>5 kb from a TSS) and proximal (<5 kb from a TSS) HOT sites that contain a DFM for each ssTF in HepG2.

gene expression and reporter activity data sets. By using ENCODE HepG2 RNA-sequencing data, we found a positive association between the number of unique DAPs at a locus and the maximum expression level of nearby genes, particularly in loci proximal (<5 kb, rho = 0.436, $P < 5 \times 10^{-16}$) to a transcription start site (TSS) (Fig. 2B, C; Supplemental Fig. S2C–F). Specifically, 55% of genes whose TSS were <5 kb from a HOT locus were expressed at a level of ≥10 FPKM, whereas only 16% of genes near a locus with fewer than 10 DAPs bound showed a similar expression level. Highly expressed genes, with FPKMs greater than 100, were also three times more likely to have multiple HOT loci within 50 kb of their TSS than genes with FPKMs less than five (chi-square $P < 5 \times 10^{-16}$) (Supplemental Fig. S2G). Loci distal to a TSS showed a significantly weaker correlation (Fisher r-to-Z transformation = 67.87, $P < 5 \times 10^{-16}$) (Fig. 2B). Both ChIP-seq and DHS motif-defined (Fig. 2D) DAP associations positively correlated with activity in previous high-throughput reporter assays of approximately 2000 selected loci in HepG2 and in ATAC-seq fragments in GM12878 (rho = 0.230 and 0.207, $P < 5 \times 10^{-16}$) (Supplemental Fig. S2H; Inoue et al. 2017; Wang et al. 2018). For both reporter assay data sets, the number of DAPs represents a specific, quantitative marker of regulatory activity that compares favorably to commonly used markers of promoter or enhancer activity (Fig. 2D; Supplemental Fig. S2H).

A small number of DFMs showed a preference for loci distal (>5 kb) or proximal (<5 kb) to a TSS (Fig. 2E; Supplemental Fig. S3A–C). Specifically, for HepG2, HNF4A, NR2F6, JDP2, and FOX, family motifs showed a twofold preference for distal, enhancer HOT loci, and ETS and SP family motifs had a threefold bias for proximal, promoter HOT loci. These findings agree with previous studies that have found HNF4A occupancy at enhancers to be essential for activity in mouse hepatocytes (Thakur et al. 2019) and a strong promoter bias for the ETS family of motifs (Hollenhorst et al. 2007). The level of sequence conservation of driver TF motifs was higher in HOT loci (Supplemental Fig. S3D), and the degree of both TSS-distal and TSS-proximal motif conservation was correlated with total number of DAPs at a locus (Supplemental Fig. S3E,F). This correlation was not observed for the CTCF motif (Supplemental Fig. S3E, F). In sum, these data suggest a dose-dependent relationship between the number of DAPs and the regulatory activity of a locus. This relationship is relatively unchanged after restricting analyses to ssTFs or nssDAPs, although nssDAPs tended to be slightly more predictive of activity than did ssTFs (Supplemental Table S8).

High-throughput mutagenesis of HOT loci reveals motifs driving activity and possible mutational buffering

After establishing that HOT loci show strong regulatory activity in a variety of reporter assays, we next sought to explore the key sequence features driving this activity by performing experimental perturbations of the sequence content of several loci. A naive hypothesis for how sequence motifs contribute to activity at HOT loci is an additive one, in which the regulatory activity of a locus is simply the sum of each constituent motif's contribution. In this scenario, ablation of a motif would have a roughly equivalent effect across loci regardless of neighboring sequence content. A more sophisticated model allows for interactions between constituent motifs with synergistic or redundant relationships. If motif synergy is a prominent feature, one would expect individual motif disruptions to have a greater effect on activity in loci containing large numbers of motifs, whereas the opposite would be true if motif redundancy was the predominant relationship between motifs. Alternatively, it is possible regulatory activity at HOT loci is not wholly dependent upon individual motifs and is substantially derived from other features. In that situation, the most disruptive mutations would not map to known TF motifs.

To begin to resolve these competing models and to identify the sequence elements most important in controlling regulatory activity at HOT loci, we performed a self-transcribing active regulatory region sequencing (STARR-seq)–based mutagenesis assay on 245 genomic loci that had previously shown activity in massively parallel or single-locus reporter assays (Supplemental Table S9). Assayed loci contained a range of unique ChIP-seq peaks (one to 150 unique DAP peaks), although roughly two-thirds of the tested elements met our HOT loci threshold by containing called peaks

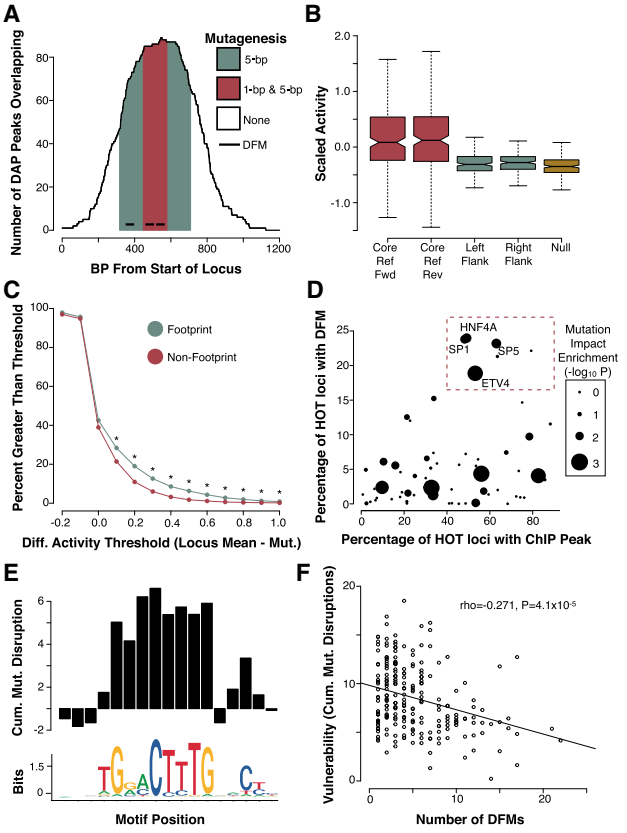

Figure 3. High-throughput mutagenesis of HOT loci reveals motifs driving activity. (*A*) Example locus depicting mutagenesis schema. The red region indicates a 130-bp core, centered upon the maximum number of unique ChIP-seq peaks and DFMs, in which we performed tiled single-bp and 5-bp mutagenesis in both the forward and reverse orientation. The flanking green regions represent 130-bp sequences flanking the core region in which we performed tiled 5-bp mutagenesis in the forward orientation only. (*B*) Boxplots indicating activity (as represented by the RNA/DNA ratio) was largely concentrated in the WT core loci in both the forward and reverse orientations and not in flanking regions or null regions (Wilcoxon $P < 510 \times ^{-16}$). (*C*) Plot indicating the proportion of mutations imposing a change of activity at a variety of thresholds for 5-bp mutations. Green points indicate data for mutations falling within DHS footprints. Red points indicate data for mutations falling outside of DHS footprints. An asterisk indicates Fisher's $P < 0.05$. (*D*) Scatter plot showing the fraction of HOT sites that contain an ssTF ChIP peak and DFM. TFs highlighted in the *top right* are putative direct binding TFs associated with a high proportion of HOT sites. The size of each point corresponds to an ssTFs DFM enrichment for high-impact mutations in our mutagenesis assay. (*E*) Barplot showing the cumulative differential activity (locus mean – mutation) across all positions in the HNF4A motif. (*F*) Scatterplot showing the number of nonredundant DFMs at a locus is inversely correlated with its vulnerability to mutation (expressed as the sum of all mutation delta activity scores).

for 52 or more DAPs. Within each 2-kb locus, we designed oligos centered around a 390-bp region of maximal ChIP-seq signal intensity across all DAP peaks (Fig. 3A). We found a majority of ChIP-seq peaks and DFMs localized to a few hundred base pairs within each HOT loci bin (Supplemental Fig. S4A), and thus, we reasoned this approach would capture a majority of active elements within each locus and allow us to assay several different loci. Each 130-bp oligo represented a left, right, or central window of the 390-bp core region. For the positive strand, we synthesized reference sequence for each window in addition to tiled 5-bp (AAAAA or TTTTT, depending on maximal disruption from reference sequence) mutations. For the central 130-bp window, we also included oligonucleotides with tiled single-base-pair mutations at each position in addition to the tiled 5-bp mutations for both the positive and reverse strand. Control sequences consisting of oligonucleotides matched for GC content and repeat length, and previously tested null sequences were also included in our library.

We cloned oligonucleotides into the STARR-seq reporter vector and transfected the plasmids into HepG2 cells. We subsequently collected RNA from transfected cells to assess the relative abundance (and thus activity) of each test element compared with DNA library input. We detected >90% of individual elements post-transfection (Supplemental Fig. S4B,C) and observed that poorly represented elements were evenly distributed in position across each locus and thus were likely not a product of alignment efficiency (Supplemental Fig. S4D). With the exception of a subset of mutated sequences, RNA and DNA counts were highly correlated across our element library (rho = 0.955, $P < 5 \times 10^{-16}$) (Supplemental Fig. S4E–H). RNA/DNA ratios were also highly correlated across sequencing replicates at our conservative minimum representation threshold of two DNA counts per million (CPM) (Supplemental Fig. S4I,J). As expected, elements from the central window were significantly more active (higher RNA/DNA ratio) than those on the border of regions of ChIP-seq signal (Fig. 3B; Supplemental Fig. S4E). Elements with single-base-pair mutations showed roughly equivalent activity to those with reference sequence on average but displayed a greater range in activity (Supplemental Fig. S5A). This suggests that, except for a small subset, most single-base mutations did not significantly affect activity. Elements with 5-bp mutations showed slightly less activity than reference sequence elements on average (Wilcoxon $P < 5 \times 10^{-16}$) (Supplemental Fig. S5A). We found the effects on activity of most mutations were highly correlated between strands (rho = 0.45, $P < 5 \times 10^{-16}$) (Supplemental Fig. S5B,C), and transversions tended to have more impact than transitions, as previously reported (Supplemental Fig. S5D; Guo et al. 2017). Furthermore, we successfully validated 14 high-impact mutations (including one gain-of-activity mutation) and 14 adjacent low-impact control mutations with individual luciferase reporter experiments using two different plasmids that place the test element either upstream of or downstream from the reporter (Supplemental Figs. S6, 7E; Supplemental Table S10).

Mutations that affect previously defined DFMs showed the greatest effect on test element activity (Fig. 3C; Supplemental Fig. S5E), and the magnitude of mutation effects was strongly correlated with that predicted by LS-GKM, an algorithm developed for predicting mutation effects on TF motifs (rho = 0.304, $P < 5 \times 10^{-16}$) (Supplemental Fig. S5F; Lee 2016). Thus, activity at loci with large numbers of DAPs associated seem to be controlled by conventional recognition motifs that can be disrupted in a predictable manner. A motif's predilection for impactful mutations was associated, albeit weakly, with its overall enrichment at HOT loci (rho = 0.320, $P = 0.022$) (Fig. 3D). Of particular interest are ETV4, SP1, SP5, and HNF4A, each of which is highly prevalent across all HOT loci and enriched for high impact SNVs, providing further evidence that these ssTFs may be important drivers of activity at HOT loci (Fig. 3D). Broadening our enrichment analysis to include all Cis-BP motifs of ssTFs expressed in HepG2, not just those assayed by ChIP-seq, reveals several additional ssTF motifs strongly enriched for high-impact mutations such as the AP-1 and FOXA sequences (Supplemental Table S11). The resolution of our mutagenesis assay allows us to identify the most important base pairs

governing activity in each of these motif sequences (Fig. 3E; Supplemental Fig. S5G,H). We also found evidence of partial motif redundancy, as loci with high numbers of motifs were generally less vulnerable to single-nucleotide variation (rho = –0.271, $P = 4.1 \times 10^{-5}$) (Fig. 3F). This suggests that some HOT loci are potentially buffered from motif-disrupting mutations that could completely ablate other loci with fewer motifs. Independent support for this hypothesis comes from the observation that the effect sizes of significant eQTL SNPs mapping to HOT loci tend to be significantly lower (rho = –0.175, $P < 5 \times 10^{-16}$) (Supplemental Fig. S8; Varshney et al. 2019).

HOT loci dichotomize into cell type–specific or ubiquitous groups

Integrating data from multiple cell lines allowed us to examine the cell type specificity of HOT loci and corresponding DAP associations. Loci containing an increasing number of unique ChIP-seq peaks in HepG2 were more likely to be present in both K562 and GM12878 than were loci with fewer ChIP-seq peaks (Supplemental Fig. S9A,B). HOT sites across each cell line tended to fall within two groups: one in which HOT sites were present in only one cell line, and a smaller group in which sites were present in all three cell lines (Fig. 4A). Relatively few loci were present in only two of three cell lines.

We found DFMs that were biased toward distal HOT loci in Figure 2E (HNF4A, NR2F6, and FOXA family) were also biased toward cell type–specific HOT loci and, conversely, that DFMs strongly associated with proximal HOT loci (ETS and SP family) were biased toward ubiquitously expressed genes (Fig. 4B; Supplemental Fig. S9C–E; Supplemental Table S12). The distal, cell type–specific class of DFMs differed among cell types with the GATA, NFE2, and TBX1 family of DFMs prominent in K562 cells and IRF8 and SPI1 DFMs prominent in GM12878 cells (Supplemental Fig. S9F,G). These distal, cell type–specific DFMs have nearly all been implicated in the regulation and differentiation of their corresponding cell lineage (Ferreira et al. 2005; Iwasaki et al. 2005; Wang et al. 2008; Davies 2013; Alder et al. 2014; Di Tullio et al. 2017). HOT loci that were common to all three cell lines were enriched for close proximity to housekeeping genes involved in cellular metabolism of organic compounds (Supplemental Table S13). Conversely, cell type–specific, HOT loci tended to neighbor corresponding cell type–specific genes (Supplemental Fig. S9H–J). In general, we found cell type–specific genes were more likely (49% vs. 16%, chi-square $P < 5 \times 10^{-16}$) to contain multiple, cell type–specific HOT loci within 50 kb of their

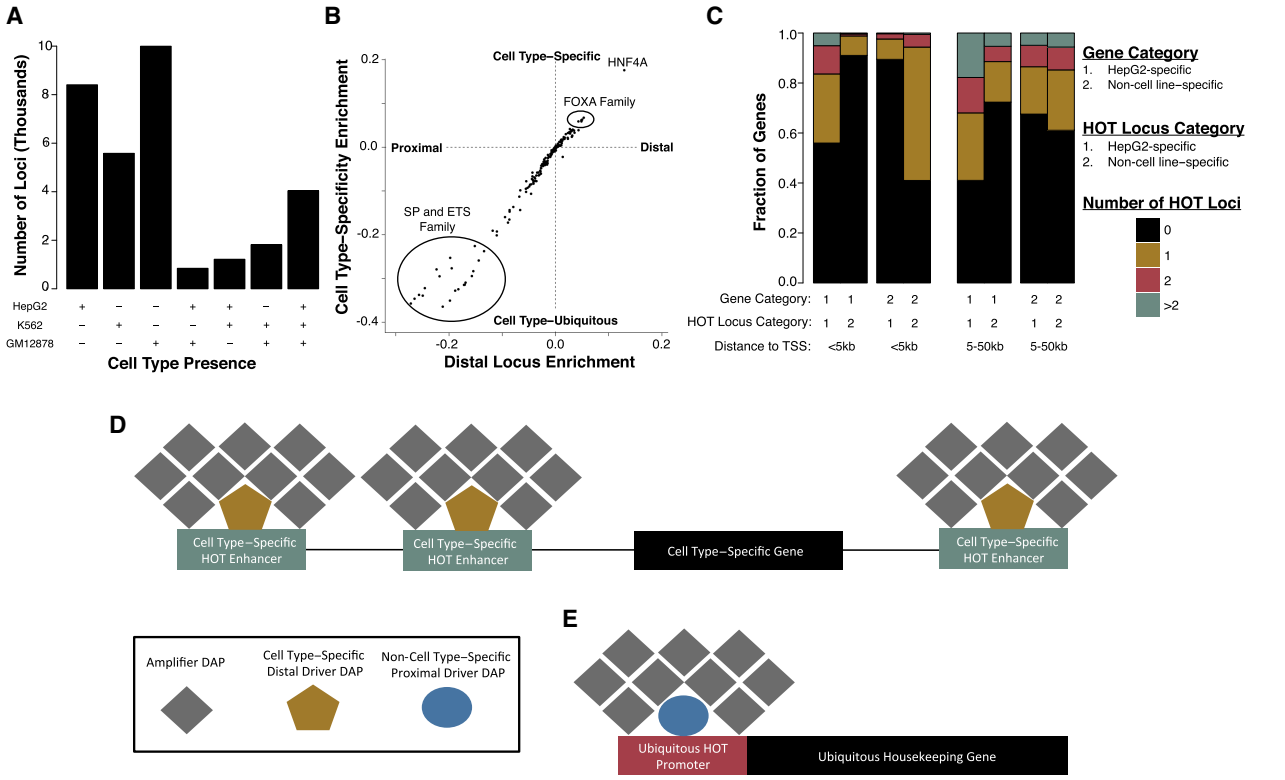

Figure 4. HOT loci dichotomize into cell type–specific or ubiquitous groups. (*A*) The number of HOT loci present in all possible combinations of each cell line. (*B*) Scatter plot showing the association between cell type–specific HOT loci enrichment and distal, HOT loci enrichment in HepG2. Cell type specificity enrichment value is computed by subtracting the fraction of HepG2-specific HOT loci ($N = 7692$) in which a DFM is present from the fraction of non-HepG2-specific HOT loci ($N = 6100$) in which a DFM is present. The distal locus enrichment is computed by subtracting the fraction of HOT loci >5 kb from the nearest TSS ($N = 6445$) in which a DFM is present from the fraction of HOT loci <5 kb from the nearest TSS in which a DFM is present ($N = 7347$). (*C*) Stacked bar plots displaying the proportion of cell type–specific or expression level–matched, non-cell type–specific genes that possess a specified number of neighboring cell type–specific or non-cell type–specific HOT loci at a specified distance to TSS threshold. Cell type–specific genes were computed by randomly sampling 500 genes that were expressed at least fourfold higher in the cell line of interest than the other two cell lines and had an FPKM of at least five in the cell line of interest. Non-cell type–specific genes were a cell type–specific gene expression level–matched sample of 500 genes with an FPKM of at least five in HepG2, K562, and GM12878. (*D*,*E*) Proposed model of how HOT loci relate to cell type–specific (*D*) and non-cell type–specific housekeeper gene (*E*) expression.

TSS, whereas ubiquitously expressed genes were more likely (59% vs. 9%, chi-square $P < 5 \times 10^{-16}$) to have a ubiquitously HOT promoter (Fig. 4C). These data support a model of multiple, cell type–specific HOT loci bound by cell type–specific, driver DAPs regulating cell type–specific gene expression (Fig. 4D), and ubiquitously expressed housekeeping genes regulated by an ubiquitously HOT promoter bound by common ETS or SP family DAPs (Fig. 4E).

CNV, 3D chromatin structure, and GC content associate with HOT loci

To further explore mechanisms underlying the formation of HOT loci, we examined a variety of genomic characteristics linked to sites with high densities of DAPs. In agreement with previous studies of TF motifs and flanking regions (Dror et al. 2015), we found HOT loci to be enriched for elevated GC content (rho = 0.387, $P < 5 \times 10^{-16}$) (Fig. 5A). In addition to being a byproduct of increased motif content, previous studies have proposed that elevated GC content, particularly in promoter regions, may lead to the formation of secondary DNA structures that induce indirect or nonspecific DAP associations (Wreczycka et al. 2019). This hypothesis was difficult to test directly as both DAP recruitment and promoter GC content were associated with neighboring gene expression (rho = 0.051, $P < 5 \times 10^{-16}$) (Supplemental Fig. S10). However, we found gene expression levels to be an independent predictor of the number of promoter DAPs after correcting for promoter GC content (regression F-statistic $P < 5 \times 10^{-16}$) and found little variation in the strength of the correlation between gene expression level and number of promoter DAPs based on promoter GC content (Fig. 5B; Supplemental Table S14), which disfavors the idea that elevated GC content artificially drives the number of DAPs beyond gene expression–based expectations. There was no association between total repeat masked sequence, repetitive element composition, or locus mappability and the number of DAPs (Supplemental Fig. S10B–D).

Another potential mechanism driving ChIP-seq signal inflation is chromosomal ploidy differences or smaller-scale CNV. Increasing the number of available DAP binding sites by copy number amplification could provide greater opportunities for DAP recruitment, resulting in a proportionally greater number of DNA fragments as input to the ChIP-seq assay and improved sensitivity for DAPs that may be incompletely accounted for with genomic background controls (Zhang et al. 2008). A gross assessment of the chromosomal distribution of HOT loci in HepG2 suggests this is an important variable to consider (Supplemental Fig. S11A–C). Increased ploidy of Chromosome 20 and partial chromosomal amplifications of Chromosomes 16 and 17 have been

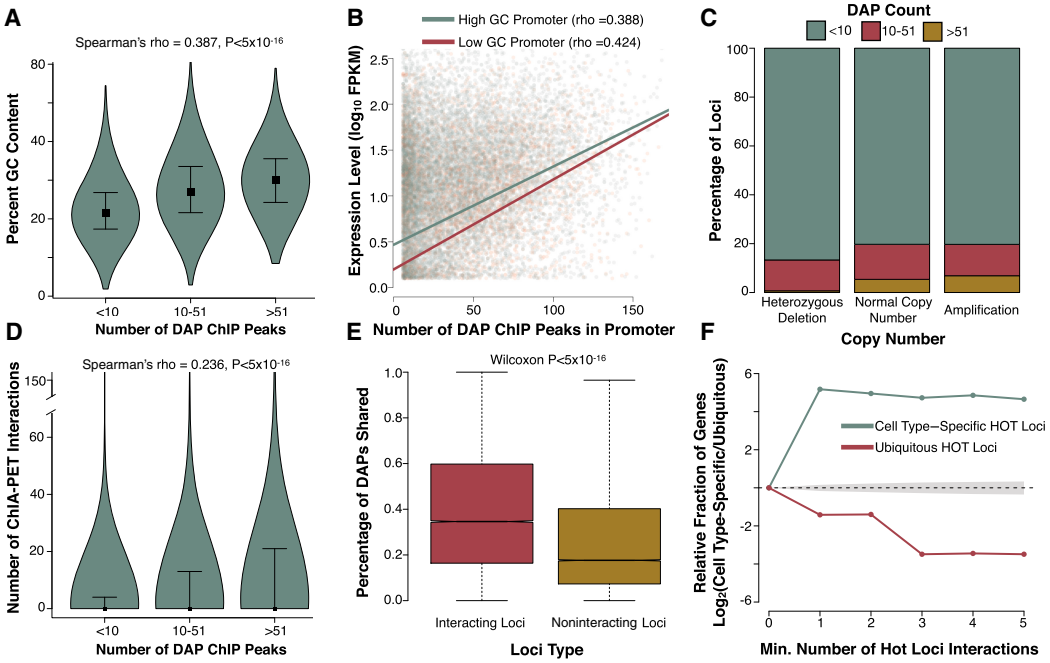

Figure 5. Copy number variation, 3D chromatin structure, and GC content associate with HOT loci. (*A*) Violin plots showing the GC content of loci with increasing numbers of DAP ChIP-seq peaks. Width of each violin indicates the relative fraction of data contained. Boxes represent the median of each bin and whiskers are drawn to the 25th and 75th percentiles. (*B*) Scatter plot showing the association between gene expression and DAP ChIP peaks in each genes promoter. Points and trend lines are colored based on promoter GC content. Promoters with GC content in the upper 50th percentile of GC content (high) are colored green, and those in the lower 50th percentile of GC content (low) are colored red. (*C*) Stacked bar plots showing proportion of loci with various levels of ChIP-derived DAP associations in genomic regions with heterozygous deletions, amplifications, or normal copy number. (*D*) Violin plots showing the correlation between the number of ChIP-defined DAP associations and the number of Promoter Capture-C interactions. Boxes represent the median of each bin, and whiskers are drawn to the 90th percentile. *P*-value reported is derived from Spearman's rho correlation of the entire data set. The sample size for each violin from *left* to *right* is 194,028, 37,084, and 13,792. (*E*) Boxplots showing the fraction of DAPs in common between interacting loci and matched noninteracting loci for HepG2 Promoter Capture-C. (*F*) Line plot indicating the relative fraction (cell type–specific/ubiquitously expressed) of gene promoters with at least the specified number of Promoter Capture-C interactions with other HOT loci. Interactions with cell type–specific loci are shown in green and interactions with loci that are HOT in all three cell lines are shown in red. The gray shaded area represents the 95% confidence, null interval of randomly shuffled loci interactions between cell type–specific and ubiquitously expressed promoters. The 500 cell type–specific and expression-matched ubiquitously expressed genes were identical to those selected in Figure 4.

previously described in HepG2, and we observed that Chromosomes 16, 17, and 20 harbored more HOT loci than expected based on their size and gene density (López-Terrada et al. 2009). However, we did not observe a higher rate of HOT loci on Chromosomes 2 and 14, which have also been described as having increased ploidy in HepG2, arguing that ploidy alone does not drive extreme numbers of DAP associations. Moreover, the K562 and GM12878 cell lines had much smaller chromosomal deviations in rates of HOT loci, despite having an equivalent number of total HOT loci (Supplemental Fig. S11B,C). Intersecting ENCODE CNV array data with our merged ChIP-seq peak loci, we found a significant depletion in DAP associations at loci with a heterozygous deletion compared to loci with diploid copy number (0.7% vs. 5.4%, chi-squared $P < 5 \times 10^{-16}$) (Fig. 5C). There was only a minor enrichment for HOT loci in amplified regions relative to normal copy number regions (5.4% vs. 6.9%, chi-squared $P < 5 \times 10^{-16}$), and ≤20% of loci at any DAP-association threshold were found in amplified regions (Supplemental Fig. S11D). Thus, locus copy number appears to be a statistically significant yet relatively minor contributor to the observed DAP association patterns.

3D chromatin structure might also contribute to the observed pattern of DAP coassociations. The importance of 3D chromatin structure is becoming increasingly recognized, and much of this structure is thought to be driven by large protein complex interactions with DNA (Quinodoz et al. 2018). Protein complexes that bring together multiple loci on a chromosome could give the appearance of indirect ChIP-seq binding at each locus involved in a given network. Analysis of Promoter Capture-C and chromatin interaction analysis with paired-end tag (ChIA-PET) data recently generated in HepG2 cells revealed a weak positive correlation between the number of DAPs and the number of 3D interactions detected across loci (rho = 0.236, $P < 5 \times 10^{-16}$) (Fig. 5D; Supplemental Fig. S12A; Chesi et al. 2019). Interacting loci did share a significantly higher proportion of DAPs than noninteracting loci (Fig. 5E). In agreement with the model proposed in Figure 4, D and E, cell type–specific promoters were significantly more likely to show distal Capture-C interactions with other cell type–specific HOT loci than were promoters of ubiquitously expressed housekeeping genes (Fig. 5F). These association trends were also found by ChIA-PET and chromatin capture available for the K562 and GM12878 cell lines (Supplemental Fig. S12B–E), and restricting these analyses to ssTFs or nssDAPs did not alter the strength of these correlations (Supplemental Table S8; Mifsud et al. 2015). Overall, we found the association between 3D chromatin interactions and DAP density to be weak but consistent across cell lines, and it is possible that 3D interactions are exceptionally abundant at a minority of HOT loci. However, HOT loci generally tend to cluster near each other relative to loci with low numbers of DAPs (Supplemental Fig. S12F), leading us to expect that these associations will likely strengthen as experimental approaches for examining 3D chromatin mature.

Discussion

We have performed an extensive analysis of DAPs across three cell lines. In each cell line, we found around 15,000 loci that harbored ChIP-seq peaks for >25% of DAPs assayed. The number of HOT loci defined by this criterion is consistent regardless of the number of DAPs incorporated into our analysis. Thus, we believe this result will be generalizable to future analyses that will incorporate increasingly comprehensive databases of genome-wide DAP associations. However, until all expressed DAPs have been assayed in a given cell line, it will be difficult to appreciate the total number of DAPs capable of associating with a single locus. As the prevalence of ChIP-seq peaks was only loosely correlated with their corresponding DFMs at HOT loci, a substantial proportion of signal at HOT loci is likely to be driven by indirect binding not constrained by the presence of specific motifs.

HOT loci identified in our analysis are distinct from previously blacklisted regions shown to be common high-signal artifacts in sequencing assays and are present at a majority of active enhancers and promoters in the cell lines we analyzed. Although it is extremely difficult to differentiate indirect DAP binding from nonspecific or artifactual ChIP-seq signal previously proposed to contribute to HOT loci (Wreczycka et al. 2019), the pervasiveness of complex DAP coassociations in non–ChIP-seq–dependent DFMs and the predictable nature of regulatory activity modulation by mutation of constituent motifs suggests these observations are likely not purely owing to ChIP artifacts. Furthermore, regardless of underlying mechanism, we find the number of DAP coassociations to be a useful marker of active regulatory elements. Rather than being a rare event capable of being filtered from future experiments, these loci appear to be a defining mark of neighboring transcription.

We do not yet know the mechanism(s) driving the HOT DAP association pattern, in part because of technical limitations of the ChIP-seq assay. Most critically, robust ChIP-seq requires a population of cells as input. Thus, it is impossible to conclude from these data what proportion of DAPs simultaneously coassociates in the same cell. Single-cell ChIP-seq is still in its infancy, but as it matures, it may provide important clues to assist in answering this question (Rotem et al. 2015). Furthermore, the allele specificity of DAPs was not considered by our analysis. Few allele-specific analyses have been conducted on a large number of DAPs in the same cell line or tissue, but some evidence exists that DAPs may favor a single allele in the context of allelic sequence variation (Reddy et al. 2012; Ramaker et al. 2017). We found the correlation between 3D chromatin interactions and observed DAP coassociations to be particularly intriguing. HOT loci are enriched for greater numbers of 3D interactions, and a greater number of shared DAPs are observed between equivalently bound interacting loci than noninteracting loci. These data coupled with the tendency of at least a subset of HOT loci to cluster near one another in the genome support a previously described long-range "flexible billboard" model of enhancer function (Arnosti and Kulkarni 2005; Vockley et al. 2017). This model proposes that enhancer output is largely dictated by the aggregate sum of interacting motif and "tethered," non-motif-driven DAPs, which have been shown to colocalize in high concentrations via phase-separated condensates (Shrinivas et al. 2019), rather than rigidly organized, enhanceosome structures.

HOT loci tend to naturally dichotomize into cell type–specific or cell type–ubiquitous groups. Cell type–specific genes tend to possess multiple, distal, neighboring cell type–specific HOT loci that typically contain cell type–specific "driver" motifs such as HNF4A or GATA. Conversely, universally expressed housekeeping genes generally had a ubiquitously HOT promoter containing SP or ETS family motifs. Thus, loci that play a role in regulating cell maintenance and differentiation seem to be readily identifiable by high densities of DAPs and can be readily segregated based on their constituent motifs. Because DAPs tend to aggregate at HOT loci in a partially cell type–specific manner, it may be difficult to fully impute the locations of HOT loci in other cell lines that have not been as extensively assayed as the core ENCODE cell lines

included in this study. However, we found DFM-defined HOT loci overlapped heavily with ChIP-defined HOT loci, which may obviate the need to perform extensive numbers of ChIP-seq experiments in every cell and tissue type to predict the presence of a HOT locus.

Lastly, our STARR-seq results provide high-resolution data on the most important sequence elements governing activity of hundreds of HOT loci. An important observation from our data is that a majority of regulatory activity can be localized to a central 130-bp region of maximal ChIP-seq peak signal at a given locus and that equivalently sized flanking regions showed activity roughly equivalent to our null sequences. Activity at HOT loci can be altered in a predictable manner by some single-base-pair mutations. HOT loci are most vulnerable to SNVs in previously identified, highly conserved portions of their constitutive motifs. In particular, a subset of ssTF motifs, including HNF4A, SP1, SP5, ETV4, FOXA, and JUN/AP-1 motifs, are highly prevalent at HepG2 HOT loci and are particularly enriched for high-impact SNVs in our mutagenesis assay. We believe that this provides sufficient evidence to nominate these ssTFs as putative drivers of regulatory activity at HOT loci in this cell line and that future experiments specifically modulating the activity of these ssTFs or their motifs at HOT loci will be informative. We also found evidence that the total number of DFMs at a locus can reduce its overall vulnerability to SNVs, suggesting that at least some HOT loci may be buffered from the effects of otherwise harmful mutations. This phenomenon is also apparent in the reduced effect size of GTEx eQTL SNPs that map to HOT loci. Similar mutagenesis experiments would need to be performed on roughly an order of magnitude greater number of loci to definitively test this hypothesis; however, our results justify further exploration as this buffering effect potentially complicates the interpretation of noncoding variation that is naïve to the presence of neighboring DAPs.

Future investigation and interpretation of ChIP-seq and related data types, especially when performed on a single DAP or a small number of DAPs, will hopefully benefit from the knowledge that extensive DAP coassociations at a significant number of functionally pertinent putative binding sites may be present. We intentionally structured our analysis within a framework that is generalizable and can act as a resource for nominating potentially interesting loci for future experiments.

Methods

All data analyzed in this study were aligned to hg19 genome to improve integration with pre-existing, publically available data; however, the conclusions made in this paper are not specific to a genome version. Detailed methods can be found in the Supplemental Methods

ChIP-seq data processing

BED files containing ChIP-seq peak information for the K562 and GM12878 cell lines were obtained directly from the ENCODE data portal (https://www.encodeproject.org) via the file accession number listed in Supplemental Table S1. BED files containing ChIP-seq peak information for the HepG2 cell line were generated by the Rick Myers and Eric Mendenhall laboratories under a consistent protocol in accordance with ENCODE standards and can be obtained from the NCBI Gene Expression Omnibus (GEO) database under the GSE104247 accession. We collapsed all neighboring peaks into a minimal set of nonoverlapping 2-kb loci and defined all peaks within a bin as "coassociated." The resulting set of 2-kb loci can be found in Supplemental Tables S3, S5, and S6 for HepG2, GM12878, and K562, respectively. DAPs were assigned to classes based on previous definitions (Lambert et al. 2018). This BED file binning method can be reproduced using the "SMART_BED_MERGE" repository available in the Supplemental Material and at GitHub (https://github.com/rramaker/GenomeTools2020/).

Motif footprint processing

All DAP motif position weight matrices (PWMs) were downloaded from the Cis-BP database (http://cisbp.ccbr.utoronto.ca/bulk.php) on 04/02/2018 (Weirauch et al. 2014). Only motifs derived from in vitro methods (SELEX, protein binding microarray, or B1H) were included in further analysis. Motifs assigned to DAPs that were unexpressed (zero reads aligned) in each cell line were excluded from further analysis. ENCODE DNase-seq raw FASTQs (accession numbers ENCFF002EQ-G,H,I,J,M,N,O,P) were downloaded from the ENCODE portal and processed using the Kundaje laboratory, ENCODE DNase-seq standard pipeline. High-confidence DHS footprints were binned into a minimal set of nonoverlapping 2-kb loci. The resultant set of 2-kb loci can be found in Supplemental Table S4. DAP motif pairs that possessed a significant (FDR< 0.05) TomTom similarity score or that shared significant similarity to another motif were treated as one motif capable of recruiting multiple DAPs as specified (Gupta et al. 2007).

Intersecting with annotations of interest

ChIP-seq peak and DHS footprint loci were intersected with a variety of other genome annotations using the BEDTools *intersect* and *map* functions (Quinlan 2014). A source BED file containing IDEAS regulatory annotations was obtained from https://main.genome-browser.bx.psu.edu (Zhang et al. 2016). Gene coordinates were obtained from the Ensembl genome browser (http://useast.ensembl.org/index.html) gene transfer format grch37.75 file. Gene expression data were obtained in the form of raw count data from the ENCODE data portal (HepG2 accession numbers ENCFF139ZPW, ENCFF255HPM, GM12878; accession numbers ENCFF790RDA, ENCFF809AKQ; K562 accession numbers ENCFF764ZIV, ENCFF489VUK). Cell type–specific genes were defined as those having a fourfold greater FPKM in a given cell line of interest than either of the other two cell lines and having a FPKM value of at least two in the cell line of interest. Cell type–ubiquitous genes were defined as those with an FPKM greater than five in HepG2, K562, and GM12878. HepG2 reporter assay data were obtained from previously published work hosted at the GEO accession GSE83894 in the file GSE83894_ActivityRatios.tsv (Inoue et al. 2017). GM12878 high-resolution dissection of regulatory assay (HiDRA) data were obtained from previously published work hosted at the GEO accession GSE104001 in the file GSE104001_HiDRA_counts_per_fragmentgroup.txt (Wang et al. 2018). Significant liver GTEx eQTL SNPs were downloaded with permission from GTEx download portal. Specifically, we obtained the "Liver_Analysis.snpgenes" file from the V6 data release that contains significant eQTL SNPs derived from liver tissue expression data. GERP scores were obtained from the Genome Browser under the "comparative genomics" group. CNV data were obtained from the ENCODE data portal under the file accession ENCFF074XLG. Deletions and amplifications were assigned as designated in the fourth column. Promoter Capture-C data for HepG2 was obtained from previously published work hosted in the Array Express database (https://www.ebi.ac.uk/arrayexpress/experiments/) under the accession E-MTAB-7144 (Chesi et al. 2019). POLR2A ChIA-PET BED files containing significant 3D

interactions for K562 were obtained from the ENCODE data portal under the file accessions ENCFF001THW and ENCFF001TIC. Promoter capture Hi-C BED files for GM12878 were obtained from previously published work hosted in the Array Express database (https://www.ebi.ac.uk/arrayexpress/experiments/) under the accession E-MTAB-2323 (Mifsud et al. 2015). BED files containing repetitive element alignment scores were obtained from the UCSC Table Browser "RepeatMasker" track under the "Repeats" group. BED files containing DUKE 35mer mappability scores were obtained from the UCSC Table Browser "mappability" track under the "mapping and sequencing" group. All *P*-values reported in the manuscript were capped at $P < 5 \times 10^{-16}$ to improve readability.

STARR-seq library design and cloning

STARR-seq library consisted of 90,581 sequences representing 390 bp within 245 unique loci in both the forward and reverse orientation with tiled single-base-pair or 5-mer mutations. We selected loci that had previously shown activity in the HepG2 cell line by Inoue et al. (2017), because we reasoned a baseline level of reporter assay activity is required to see differential activity upon mutation (Supplemental Table S9; Inoue et al. 2017). Alternate bases were randomly signed for single-base-pair mutations. 5-mer mutations were AAAAA or TTTTT, depending on which was most divergent from the reference sequence. Previously shown reporter activity in the top quartile of Inoue et al. (2017) or in-house data sets was the primary inclusion criteria (Inoue et al. 2017). GC-matched negative control sequences were generated using the nullseq_generate executable from the kmersvm website (http://beerlab.org/kmersvm/) on the provided hg19 genome indices (Fletez-Brant et al. 2013). Our complete oligonucleotide library is included in Supplemental_Table_S15.

Library oligonucleotides were synthesized by CustomArray as single-stranded 170-bp sequences corresponding to 130-bp test elements (from either the 130-bp activity core, 130-bp left or 130-bp right flanking sequence for each locus) with 20-bp Illumina sequencing primer binding site tails. This library was amplified and cloned into the hSTARR-seq (Addgene 99292) vector with InFusion cloning. InFusion products were transformed into Lucigen Endura electrocompetent cells pooled and grown overnight at 37°C in 2 L of LB ampicillin media at 200 RPM. The full plasmid library DNA was extracted from this culture using the Qiagen EndoFree gigaprep kit.

STARR-seq library transfection, RNA isolation, and library preparation

The STARR-seq library was transfected into HepG2 cells in 30-cm^2 plates (25 million cells per plate). Twenty-four hours after transfection, transfected cells were lysed on plate in RLT buffer (Qiagen) and stored at −80°C. Total RNA was then isolated using the Norgen total RNA purification kit using the manufacturer's instructions. STARR-seq libraries were prepared as previously described (Gaulton et al. 2013) and sequenced on an Illumina NextSeq with 150-bp paired-end reads using standard protocols. All primers used in array amplification, cloning, and library preparation are listed in Supplemental Table S16.

STARR-seq data processing and analysis

FASTQ files were adapter trimmed using cutadapt version 1.2.1 before alignment (Martin 2011). Trimmed reads were mapped to our oligo library using Bowtie 2 version 2.2.5 (Langmead and Salzberg 2013). A custom Bowtie index was generated with our oligo library (Supplemental Table S15) in FASTA format. Trimmed FASTQ files were subsequently aligned to our custom index in a manner that required a perfect sequence match only in the correct orientation. This alignment procedure can be reproduced with the STARR_SEQ_Mutagenesis folder at GitHub (https://github.com/rramaker/GenomeTools2020).

Oligo activity was defined as replicate median $\log_{10}$(RNA CPM/DNA CPM). The differential activity of a mutation containing oligo, or the effect of a mutation on a locus, was computed as the difference in the mean activity of all oligos associated with a locus from the activity of a given mutated oligo of interest. In all cases, the oligos containing forward-strand sequence were analyzed separately from oligos containing reverse-strand sequence for each locus. Raw count data and processed activity levels are available in Supplemental Tables S17 and S18. Predicted mutation effects were determined using the lsgkm analysis suite in a manner previously described (Lee 2016; Ramaker et al. 2017). Individual elements tested individually validate our results are included in Supplemental Table S19.

Data access

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under the accession number GSE142566.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank the Yijun Ruan and Struan Grant laboratories for uniform processing of the ENCODE ChIA-PET and Promoter Capture-C data, respectively. We also thank Eric Mendenhall and Surya Chhetri for their assistance with the alignment and quality control analysis of ChIP-seq experiments in HepG2, and particularly thank them and the Myers/Mendenhall ENCODE group members, including Mark Mackiewicz, Kim Newberry, Dianna Moore, Laurel Brandsmeier, Sarah Meadows, and Megan McEown, for generating the high-quality ChIP-seq data used in this paper. We thank Alessandra Chesi and the Struan F.A. Grant lab for generously providing their processed HepG2 Capture-C data. This work was supported by National Institutes of Health (NIH) grants U54 HG006998-0 (to R.M.M. and E. Mendenhall) and 5T32GM008361-21 (to R.C.R. and A.A.H.).

Author contributions: R.C.R., A.A.H., and E.C.P, conducted reporter assay experiments; R.C.R., A.A.H., and S.T.G. performed computational analysis of ChIP-seq, DFM, and 3D-chromatin interaction data; and R.C.R., A.A.H., E.C.P., S.T.G., S.J.C., B.W., and R.M.M. performed data interpretation and wrote the manuscript.

References

Alder O, Cullum R, Lee S, Kan AC, Wei W, Yi Y, Garside VC, Bilenky M, Griffith M, Morrissy AS, et al. 2014. Hippo signaling influences HNF4A and FOXA2 enhancer switching during hepatocyte differentiation. *Cell Rep* **9:** 261–271. doi:10.1016/j.celrep.2014.08.046

Arnosti DN, Kulkarni MM. 2005. Transcriptional enhancers: intelligent enhanceosomes or flexible billboards? *J Cell Biochem* **94:** 890–898. doi:10.1002/jcb.20352

Basta J, Rauchman M. 2015. The nucleosome remodeling and deacetylase (NuRD) complex in Development and Disease. *Transl Res* **165:** 36–47. doi:10.1016/j.trsl.2014.05.003

Carroll TS, Liang Z, Salama R, Stark R, de Santiago I. 2014. Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. *Front Genet* **5:** 75. doi:10.3389/fgene.2014.00075

Chesi A, Wagley Y, Johnson ME, Manduchi E, Su C, Lu S, Leonard ME, Hodge KM, Pippin JA, Hankenson KD, et al. 2019. Genome-scale Capture C promoter interactions implicate effector genes at GWAS loci for bone mineral density. *Nat Commun* **10:** 1260. doi:10.1038/s41467-019-09302-x

Davies EG. 2013. Immunodeficiency in DiGeorge syndrome and options for treating cases with complete athymia. *Front Immunol* **4**. doi:10.3389/fimmu.2013.00322

DeLaForest A, Nagaoka M, Si-Tayeb K, Noto FK, Konopka G, Battle Ma, Duncan Sa. 2011. HNF4A is essential for specification of hepatic progenitors from human pluripotent stem cells. *Development* **138:** 4143–4153. doi:10.1242/dev.062547

Di Tullio A, Passaro D, Rouault-Pierre K, Purewal S, Bonnet D. 2017. Nuclear factor erythroid 2 regulates human HSC self-renewal and T cell differentiation by preventing NOTCH1 activation. *Stem Cell Reports* **9:** 5–11. doi:10.1016/j.stemcr.2017.05.027

Dror I, Golan T, Levy C, Rohs R, Mandel-Gutfreund Y. 2015. A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Res* **25:** 1268–1280. doi:10.1101/gr.184671.114

The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816. doi:10.1038/nature05874

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

Ferreira R, Ohneda K, Yamamoto M, Philipsen S. 2005. GATA1 function, a paradigm for transcription factors in hematopoiesis. *Mol Cell Biol* **25:** 1215–1227. doi:10.1128/MCB.25.4.1215-1227.2005

Fletez-Brant C, Lee D, McCallion AS, Beer MA. 2013. kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets. *Nucleic Acids Res* **41:** W544–W556. doi:10.1093/nar/gkt519

Foley JW, Sidow A. 2013. Transcription-factor occupancy at HOT regions quantitatively predicts RNA polymerase recruitment in five human cell lines. *BMC Genomics* **14:** 720. doi:10.1186/1471-2164-14-720

Gaulton KJ, Nammo T, Pasquali L, Simon JM, Giresi PG, Fogarty MP, Panhuis TM, Mieczkowski P, Secchi A, Bosco D, et al. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339:** 1074–1077.

Guo C, McDowell IC, Nodzenski M, Scholtens DM, Allen AS, Lowe WL, Reddy TE. 2017. Transversions have larger regulatory effects than transitions. *BMC Genomics* **18:** 394. doi:10.1186/s12864-017-3785-4

Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. 2007. Quantifying similarity between motifs. *Genome Biol* **8:** R24. doi:10.1186/gb-2007-8-2-r24

Hollenhorst PC, Shah AA, Hopkins C, Graves BJ. 2007. Genome-wide analyses reveal properties of redundant and specific promoter occupancy within the ETS gene family. *Genes Dev* **21:** 1882–1894. doi:10.1101/gad.1561707

Inoue F, Kircher M, Martin B, Cooper GM, Witten DM, Mcmanus MT, Ahituv N, Shendure J. 2017. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res* **27:** 38–52. doi:10.1101/061606

Iwasaki H, Somoza C, Shigematsu H, Duprez EA, Iwasaki-Arai J, Mizuno SI, Arinobu Y, Geary K, Zhang P, Dayaram T, et al. 2005. Distinctive and indispensable roles of PU.1 in maintenance of hematopoietic stem cells and their differentiation. *Blood* **106:** 1590–1600. doi:10.1182/blood-2005-03-0860

Jain D, Baldi S, Zabel A, Straub T, Becker PB. 2015. Active promoters give rise to false positive "Phantom Peaks" in ChIP-seq experiments. *Nucleic Acids Res* **43:** 6959–6968. doi:10.1093/nar/gkv637

Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316:** 1497–1502. doi:10.1126/science.1141319

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 S, Marinov G, 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 S. 2013. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9:** 357–359. doi:10.1038/nmeth.1923

Lee D. 2016. Sequence analysis LS-GKM: a new gkm-SVM for large-scale datasets. *Bioinformatics* **32:** 2196–2198. doi:10.1093/bioinformatics/btw142

Li H, Chen H, Liu F, Ren C, Wang S, Bo X, Shu W. 2015. Functional annotation of HOT regions in the human genome: implications for human disease and cancer. *Sci Rep* **5:** 11633. doi:10.1038/srep11633

Li H, Liu F, Ren C, Bo X, Shu W. 2016. Genome-wide identification and characterisation of HOT regions in the human genome. *BMC Genomics* **17:** 733. doi:10.1186/s12864-016-3077-4

López-Terrada D, Cheung SW, Finegold MJ, Knowles BB. 2009. Hep G2 is a hepatoblastoma-derived cell line. *Hum Pathol* **40:** 1512–1515. doi:10.1016/j.humpath.2009.07.003

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17:** 10. doi:10.14806/ej.17.1.200

Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, Wingett SW, Andrews S, Grey W, Ewels PA, et al. 2015. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet* **47:** 598–606. doi:10.1038/ng.3286

Mitchell PJ, Tjian R. 1989. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. **245:** 371–378.

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* (in press) doi:10.1038/s41586-020-2023-4

Quinlan AR. 2014. BEDTools: the Swiss-army tool for genome feature analysis. *Curr Protoc Bioinformatics* **47:** 11.12.1–11.12.34. doi:10.1002/0471250953.bi1112s47

Quinodoz SA, Ollikainen N, Tabak B, Palla A, Schmidt JM, Detmar E, Lai MM, Shishkin AA, Bhat P, Takei Y, et al. 2018. Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell* **174:** 744–757.e24. doi:10.1016/j.cell.2018.05.024

Ramaker RC, Savic D, Hardigan AA, Newberry K, Cooper GM, Myers RM, Cooper SJ. 2017. A genome-wide interactome of DNA-associated proteins in the human liver. *Genome Res* **27:** 1950–1960. doi:10.1101/gr.222083.117

Reddy TE, Gertz J, Pauli F, Kucera KS, Varley KE, Newberry KM, Marinov GK, Mortazavi A, Williams BA, Song L, et al. 2012. Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Res* **22:** 860–869. doi:10.1101/gr.131201.111

Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518:** 317–330. doi:10.1038/nature14248

Rotem A, Ram O, Shoresh N, Sperling RA, Goren A, Weitz DA, Bernstein BE. 2015. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat Biotechnol* **33:** 1165–1172. doi:10.1038/nbt.3383

Sanyal A, Lajoie BR, Jain G, Dekker J. 2012. The long-range interaction landscape of gene promoters. *Nature* **489:** 109–113. doi:10.1038/nature11279

Sherwood RI, Hashimoto T, Donnell CWO, Lewis S, Barkal AA, Hoff JP Van, Karun V, Jaakkola T, Gifford David K. 2014. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat Biotechnol* **32:** 171–178. doi:10.1038/nbt.2798

Shrinivas K, Sabari BR, Coffey EL, Sharp PA, Young RA, Chakraborty AK, Shrinivas K, Sabari BR, Coffey EL, Klein IA, et al. 2019. Enhancer features that drive formation of transcriptional condensates. *Mol Cell* **75:** 549–561.e7. doi:10.1016/j.molcel.2019.07.009

Spitz F, Furlong EEM. 2012. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13:** 613–626. doi:10.1038/nrg3207

Tan NY, Khachigian LM. 2009. Sp1 phosphorylation and its regulation of gene transcription. *Mol Cell Biol* **29:** 2483–2488. doi:10.1128/MCB.01828-08

Teytelman L, Thurtle D, Rine J, van Oudenaarden A. 2013. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *PNAS* **110:** 18602–18607. doi:10.1073/pnas.1316064110

Thakur A, Wong JCH, Wang EY, Lotto J, Kim D, Cheng JC, Mingay M, Cullum R, Moudgil V, Ahmed N, et al. 2019. Hepatocyte nuclear factor 4-α is essential for the active epigenetic state at enhancers in mouse liver. *Hepatology* **70:** 1360–1376. doi:10.1002/hep.30631

Underhill C, Qutob MS, Yee SP, Torchia J. 2000. A novel nuclear receptor corepressor complex, N-CoR, contains components of the mammalian SWI/SNF complex and the corepressor KAP-1. *J Biol Chem* **275:** 40463–40470. doi:10.1074/jbc.M007864200

Varshney A, Vanrenterghem H, Orchard P, Boyle AP, Stitzel ML, Ucar D, Parker SCJ. 2019. Cell specificity of human regulatory annotations and their genetic effects on gene expression. *Genetics* **211:** 549–562. doi:10.1534/genetics.118.301525

Vockley CM, McDowell IC, D'Ippolito AM, Reddy TE. 2017. A long-range flexible billboard model of gene activation. *Transcription* **8:** 261–267. doi:10.1080/21541264.2017.1317694

Wang H, Lee CH, Qi C, Tailor P, Feng J, Abbasi S, Atsumi T, Morse HC III. 2008. IRF8 regulates B-cell lineage specification, commitment, and differentiation. *Blood* **112:** 4028–4038. doi:10.1182/blood-2008-01-129049

Wang X, He L, Goggin S, Saadat A, Wang L, Claussnitzer M, Kellis M. 2018. High-resolution genome-wide functional dissection of transcriptional regulatory regions in human. *Nat Commun* **9:** 5380.

Wei GH, Badis G, Berger MF, Kivioja T, Palin K, Enge M, Bonke M, Jolma A, Varjosalo M, Gehrke AR, et al. 2010. Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *EMBO J* **29:** 2147–2160. doi:10.1038/emboj.2010.106

Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, et al. 2014. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158:** 1431–1443. doi:10.1016/j.cell.2014.08.009

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

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

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

Zhang Y, An L, Yue F, Hardison RC. 2016. Jointly characterizing epigenetic dynamics across multiple human cell types. *Nucleic Acids Res* **44:** 6721–6731. doi:10.1093/nar/gkw278

Received December 26, 2019; accepted in revised form June 24, 2020.