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

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

***These authors contributed equally**

1. HudsonAlpha Institute For Biotechnology, Huntsville, AL

2. University of Alabama at Birmingham, Department of Genetics, Birmingham, AL

3. California Institute of Technology, Division of Biology and Biological Engineering, Pasadena, CA

Abstract

DNA associated proteins (DAPs) regulate gene expression by binding to regulatory loci such as enhancers or promoters. An understanding of how DAPs cooperate at 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 DNA-associated proteins assayed in three cell lines and integrated these data with an orthogonal dataset of 352 non-redundant, *in vitro*-derived motifs mapped to the genome within DNase hypersensitivity footprints in an effort to characterize regions of the genome that have exceptionally high numbers of DAP associations. We subsequently performed a massively parallel mutagenesis assay to discover the key sequence elements driving transcriptional activity at these loci and explored plausible biological mechanisms underlying their formation. We establish a generalizable definition for High Occupancy Target (HOT) loci and identify putative driver DAP motifs, including HNF4A, SP1, SP5, and ETV4, that are highly prevalent and exhibit sequence conservation at HOT loci. We also found the number of DAP associations is positively associated with evidence of regulatory activity and, by systematically mutating 245 HOT loci in our massively parallel reporter assay, localize regulatory activity in these loci to a central core region that is dependent on the motif sequences of our previously nominated driver DAPs. In sum, our 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.

Introduction

Gene expression networks underlie many cellular processes (Spitz and Furlong 2012). These expression networks are controlled by DNA regulatory elements, such as promoters or enhancers, which can be proximal, distal, or within their target genes in a given expression network. Extensive mapping of epigenetic modifications and 3D chromatin structure have

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 of the 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 (Roadmap Epigenomics Consortium 2015; The ENCODE Project Consortium 2007, 2012; Sanyal et al. 2012). In this manuscript, 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 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 genome 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 (Ramaker et al. 2017; Wreczycka et al. 2017; Teytelman et al. 2013; Jain et al. 2015). 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 and Marinov 2012; Carroll et al. 2014). However, we find this phenomenon 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 co-associations and of the classic model for how DAPs interact with regulatory elements.

In this manuscript, we aggregated ChIP-seq data from 469 DAPs assayed in three cell lines. We integrate these *in vivo* mapping data with an orthogonal dataset of 352 non-redundant, *in vitro*-derived motifs from 555 DAPs, which we have mapped to the genome within DNase hypersensitivity footprints of each cell line. Specifically, we aim to (1) detail the prevalence and cell type specificity of regulatory element DAP co-associations, (2) assess the utility of co-associations as a marker of regulatory activity, (3) perform a high-resolution dissection of key sequences driving activity at regions with large numbers of DAP associations by using a massively parallel mutagenesis assay, and (4) explore potential factors influencing observed DAP co-associations, such as 3D chromatin interactions and copy number variation.

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; Table S1). A subset of DAPs was further classified into sequence specific transcription factors (ssTFs) and non-sequence specific DAPs (nssDAPs). 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 (Partridge, 2018, Lambert, 2018). 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 (Table S2) (Sherwood et al. 2014). To quantify DAP co-associations, we binned the genome into a minimal set of non-overlapping 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 (Table S3-6). 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 datasets, 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 given locus to reach the HOT threshold. Nearly 6% (13,792) loci met this HOT threshold in HepG2, and we found this result to be consistent across cell lines and after varying the number DAPs incorporated into our analysis via random sampling (Figure 1A, Figure S1-S2). The distribution of observed DAP associations was dramatically different than that observed after randomly scrambling DAP associations across all loci (K-S test P< 5E-16) with no loci reaching our HOT threshold by random chance (Figure S3). The observed pattern of DAP co-associations was relatively consistent when restricting to ssTFs or nssDAPs (Figure 1B), 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 co-associated ssTFs. No locus had 25% or more of the 352 non-redundant 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 highly correlated with the number of unique DAP peaks across loci (rho=0.494) despite a minority (~10%) of ssTFs with ChIP-seq peaks present at any given HOT site possessing a corresponding DFM at the same loci (Figure 1C). These data suggest that, although the presence of DFMs is a strong indicator of HOT loci and the rate

of sequence-specific, direct TF binding increases significantly at HOT loci, a majority of DAP associations at HOT loci likely represent non-specific or indirect interactions.

While HOT sites represent a small minority of DAP-associated loci, due to the massive number of DAPs that localize to these sites, they account for 55% on average of any individual DAP's ChIP-seq peaks on average, potentially complicating the interpretation of any individual ChIP-seq dataset. 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 exhibit the highest rates of HOT site participation (Figure 1D). These classes have been previously described as having an affinity for large multi-protein complex membership, such as the NuRD complex, and are plausible candidates to be indirectly recruited to HOT loci (Basta and Rauchman 2009; Underhill et al. 2000). 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 HOTs sites (Figure 1E) (Wei et al. 2010; Tan and Khachigian 2009; 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).

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 DAP associations. Intersecting these loci with previously assigned HepG2 genomic annotations, we found a continuous relationship between the number of DAP associations, identified as ChIP-seq peaks or DFMs, and a strong enhancer or promoter designation (Figure 2A, S4) (Zhang et al. 2016). Loci containing a large number of DFMs were particularly enriched for promoters over other annotations (Figure S4). Roughly half of all promoters and strong enhancers in HepG2 met our HOT loci threshold, while genomic regions

with other annotations rarely met this threshold (Figure S5). Loci with lower numbers of DAP peaks largely lacked enhancer or promoter annotations.

To assess the regulatory activity of loci as a function of the number of unique DAP associations, we used a variety of publicly available gene expression and reporter activity datasets. Using ENCODE HepG2 RNA-sequencing data, we found a positive association between the number of unique DAP associations and the expression level of the nearest gene (Figure 2B-C, S6). This association was significantly stronger in loci proximal (<5kb) to transcription start sites than in loci located more distally. Both ChIP-seq and DHS motif-defined (Figure 2D) associations correlated strongly with activity in previous high-throughput reporter assays of ~2000 selected loci in HepG2 and in ATAC-seq fragments in GM12878 (Figure 2E) (Inoue et al. 2017; Wang et al. 2018). For both reporter assay datasets, the number of DAP associations represents a specific, quantitative marker of regulatory activity that compares favorably to commonly-used markers of promoter or enhancer activity (Figure 2D-E). Further supporting the biological relevance of HOT loci, we found the level of sequence conservation of our previously identified driver TF motifs (Figure 1E) was significantly higher in HOT loci and the degree of motif conservation was correlated with total number of DAP associations at a locus (Figure S7). This correlation was not observed for CCTF motif (Figure S7). In sum, these data suggest a dose-dependent relationship between the number of DAP associations and 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 (Figure S8, Table S7).

High-throughput mutagenesis of HOT loci reveals motifs driving activity and buffering against mutations

To elucidate the sequence elements that control 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 demonstrated activity in massively parallel or single locus reporter assays (Table S8). Assayed loci contained a range of unique ChIP-seq peaks (1-150 unique DAPs peaks), although the vast majority met our HOT loci threshold by containing called peaks 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 (Figure 3A). 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 to DNA library input. We detected more than 90% of individual elements post-transfection (Figure S9) and observed that poorly represented elements were evenly distributed in position across each locus, and thus were likely not a product of alignment efficiency (Figure S10). With the exception of a subset of mutated sequences, RNA and DNA counts were highly correlated across our element library (Figure S11). RNA/DNA ratios were also highly correlated across sequencing replicates at our conservative minimum representation threshold of two DNA counts per million (CPM) (Figure S12). 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 (Figure 3B). Elements with single base pair mutations exhibited roughly equivalent activity to those with reference sequence on average, but displayed a greater range in activity (Figure S13). This suggests that,

except for a small subset, most single base mutations did not significantly affect activity. Elements with 5-bp mutations exhibited slightly less activity than reference sequence elements on average (Figure S13). We found the effects on activity of most mutations were highly correlated between strands (Figure S14) and transversions tended to have more impact than transitions, as previously reported (Figure S15) (Guo et al. 2017). Furthermore, we successfully validated 14 high-impact mutations (including a 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 or downstream of the reporter (Figure S16-S17, Table S9).

Mutations that impacted previously defined DFMs showed the greatest effect on locus activity (Figure 3C, S18) 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 (Lee 2016) (Figure S19). Thus, activity at loci with large numbers of DAP associations seems to be controlled by motifs that can be disrupted in a predictable manner. A motif's predilection for impactful mutations was weakly associated with its overall enrichment at HOT loci (Figure 3D, rho = 0.320, P=0.02241). Of particular interest are ETV4, SP1, SP5, and HNF4A, each of which are 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 (Figure 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 (Table S10). The resolution of our mutagenesis assay allows us to identify the most important base pairs governing activity in each of these motif sequences (Figure 3E, Figure S20). We also found evidence of partial motif redundancy, as loci with high numbers of motifs were generally less vulnerable to single nucleotide variation (Figure 3F). This suggests that HOT loci are potentially buffered from motif disrupting mutations that could completely ablate other loci that have 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 have significantly lower effect sizes (Figure S21).

HOT loci dichotomize into cell-type specific or ubiquitous groups

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 (Figure 4A,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 (Figure 4C). Relatively few loci were present in only two of three cell lines. HOT loci that were common across all three cell lines were enriched for close proximity to housekeeping genes largely involved in cellular metabolism of organic compounds (Table S11). Conversely, cell type-specific HOT loci tended to fall in close proximity to corresponding cell type specific genes (Figure 4D-F). Thus, HOT loci are likely involved in regulating cell type-specific or core housekeeping expression programs.

Copy number variation and chromatin 3D structure may confound DAP associations

Because the extent of apparent DAP associations at HOT loci is only partially explained by DNA sequence specific recruitment (Figure 1C), we next explored other potential mechanisms of high-density DAP co-associations. One potential mechanism for ChIP-seq signal inflation is chromosomal ploidy differences or smaller scale copy number variation (CNV) at HOT loci. 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 DAP associations 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 (Figure 5A). Increased ploidy of chromosome

20 and partial chromosomal amplifications of chromosomes 16 and 17 have been previously described in HepG2, and we observed that chromosomes 16, 17, and 20 harbored more HOT loci than expected based on the 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 (Figure S22). The GM12878 cell line is notable for having a nearly completely diploid genome.

To more precisely assess the effects of copy number on DAP associations, we used publicly available CNV array data from each cell line. 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 normal copy number (0.7% vs. 5.4%, Chi-squared P=6.9E-50, Figure 5B). 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=2.9E-25) and 20% or less of loci at any DAP-association threshold were found in amplified regions (Figure S23). Thus, loci copy number appears to be a statistically significant but relatively minor driver of observed DAP association patterns.

Another potential explanation of the observed pattern of DAP co-associations is 3D chromatin structure. The importance of 3D chromatin structure is becoming increasingly recognized and is thought to be driven by large protein complex interactions with DNA (Quinodoz et al. 2018). Protein complexes that bring together multiple regions of DNA to form loop structures could give the appearance of dramatic levels of indirect binding at each locus involved in a given network. Recent POLR2A Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) conducted in HepG2 cells revealed a correlation between the number of DAP associations and the number of 3D interactions across loci (Figure 5C,D). POLR2A

ChIA-PET and chromatin capture (Hi-C) available for the GM12878 cell line reveals a similar correlation (Figure S24). Restricting these analyses to ssTFs or nssDAPs did not dramatically impact the strength of these correlations (Table S7). Further supporting the hypothesis of 3D structure driving a proportion of indirect binding, interacting loci share a significantly higher proportion of associated DAPs than non-interacting loci (Figure S25) and loci with higher numbers of DAP associations tend to cluster near each other relative to loci with low numbers of DAP associations (Figure S26).

Discussion

We have performed an extensive analysis of DAP associations across three cell lines using 647 ChIP-seq experiments and 941 *in vitro* derived motifs. In each cell line, we found ~15,000 loci that harbored ChIP-seq peaks for more than 25% of DAPs assayed. The number of HOT loci defined by this criterion appears to be 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 participating in 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. Thus, the true upper limit for the number of DAPs associated a given locus may exceed the number of DAPs currently assayed. Moreover, 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. Rather than being a rare event capable

of being filtered from future experiments, these loci appear to be a defining mark of neighboring transcription.

Unfortunately, we can still only speculate on the exact mechanisms driving DAP associations given the technical limitations of the ChIP-seq assay. Most critically, ChIP-seq requires a population of cells as input. Thus, it is impossible to delineate what proportion of DAPs are bound simultaneously in the same cell. Single-cell ChIP-seq is still in its infancy, but it will likely provide important clues to assist in answering this question (Rotem et al. 2015). Furthermore, the allele-specificity of DAP associations is 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 (Ramaker et al. 2017; Reddy et al. 2012).

Our analysis has found allele copy number to be a minor contributor to the number of observed DAP associations at a locus. Specifically, heterozygous deletions were depleted for high numbers of DAP associations and amplified loci showed a slight increase in HOT loci in terms of proportional representation. We hypothesize that this is largely due to DNA input being an important driver of sensitivity for the ChIP-seq assay. However, there may be selective pressures driving amplification of HOT loci in rapidly dividing cancer cell lines. Additionally, the 3D chromatin structure at a locus appears to play a role in observed DAP associations. In particular, HOT loci are enriched for greater numbers of 3D interactions and a greater number of shared DAP associations are observed between equivalently bound interacting loci than non-interacting loci. These data coupled with the tendency of at least a subset HOT loci to cluster near one another in the genome support a previously described long-range “flexible billboard” model of enhancer function (Vockley et al. 2017; Arnosti and Kulkarni 2005). This model proposes that enhancer output is largely dictated by the aggregate sum of interacting motif and “tethered,” non-motif driven DAP associations rather than rigidly organized enhanceosome structures.

High numbers of DAP associations, whether they represent direct or indirect DNA binding, strongly predict regulatory activity. We found that DAP co-associations were associated with independently-defined promoter and enhancer annotations, motif constraint, neighboring gene expression level, and reporter assay activity. HOT loci, as defined by ChIP-seq, tended be evenly distributed across enhancers and promoters; however, loci particularly enriched for DFMs were more likely to be promoters than enhancers. This finding suggests that direct, sequence-specific binding may be more prevalent in promoters than enhancers. The number of DAP associations exhibited a roughly continuous relationship with neighboring gene expression and the ability to drive expression of a minimal promoter in a reporter assay. We did not find a saturation point beyond which greater numbers of DAP associations were redundant; thus we hypothesize that DAP associations contribute to a locus' regulatory activity in a dose-dependent manner. Furthermore, HOT loci tended to occur near cell-type specific or cell-type ubiquitous metabolic pathways, indicating they may play a role in regulating cell maintenance and differentiation. Furthermore, DAPs tend to associate with HOT loci in a largely cell type-specific manner making it difficult to predict the presence of a HOT locus using only data derived from a core set ubiquitously expressed DAPs. However, we found DFM-defined HOT loci overlapped heavily with a subset of ChIP-defined HOT loci, particularly in promoter regions, 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. We also found that activity at HOT loci can be dramatically altered in a predictable manner by single base pair mutations. In agreement with previous dogma regarding enhancer activity, HOT loci also behaved relatively

independently to orientation such that SNV effects in the forward strand were strongly associated with those in the reverse strand, which provides us with internal replication of several high-impact SNVs. We found that HOT loci were 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, and ETV4, were highly prevalent at HepG2 HOT loci and particularly enriched for high-impact SNVs in our mutagenesis assay. We believe this provides sufficient evidence to nominate these ssTFs as putative drivers of regulatory activity at HOT loci, and 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 influence its overall vulnerability to SNVs, suggesting that HOT loci may act to buffer the effects of otherwise harmful mutations on highly expressed genes. This phenomenon is also apparent in the reduced effect size of GTEx eQTL SNPs that map to HOT loci. We believe this observation merits further exploration as this buffering effect potentially complicates the interpretation of non-coding variation that is naïve to the presence of neighboring DAP associations.

Overall, we hope our analysis provides a useful contribution to understanding of the prevalence and implications of DAP associations in the genome. Future investigations with ChIP-seq or related experiments on a single DAP or small number of DAPs could benefit from the knowledge that extensive DAP co-associations at a significant number of putative binding sites may be present. We structured our analysis within a framework that is generalizable and can act as a resource for nominating potentially interesting loci for future experiments.

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 (www.encodeproject.org) via the file accession number listed in Table S1. BED files containing ChIP-seq peak information for the HepG2 cell line were generated by the Myers and Eric Mendenhall labs under a consistent protocol in accordance with ENCODE standards and can be obtained from the GEO database under the GSE104247 accession. To define ChIP-seq derived DAP co-associations, we collapsed all neighboring peaks into a minimal set of non-overlapping 2-kb loci and defined all peaks within a bin as “co-associated”. This minimal set of 2-kb loci containing all ChIP-seq peaks was generated independently for each cell line in two steps. First, all peaks from each of a cell line’s BED files were merged into a single BED file with the bedtools package (https://bedtools.readthedocs.io/en/latest/) *merge* function and with the maximum distance required for merging (or –d flag) set to 2000 bp. Resulting merged peak loci that were smaller than 2kb were redefined by adding or subtracting 1 kb from midpoint to expand them to 2kb. Merged peak loci that grew larger than 2 kb were split into contiguous individual 2-kb bins using split points that intersected the fewest possible ChIP-seq peaks in the original, individual DAP BED files. The small number of individual DAP peaks that were split in this process were assigned to the bin to which >50% of the peak resided. The resultant set of 2 kb loci can be found in Table S3, Table S5, and Table S6 for HepG2, GM12878, and K562, respectively. DAPs were assigned to classes based on previous definitions (Lambert et al. 2018).

CIS-BP 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. 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 (0 reads aligned) in each cell line based on expression data available on the ENCODE portal (HepG2 accession numbers: ENCFF139ZPW,

ENCFF255HPM, GM12878 accession numbers: ENCFF790RDA, ENCFF809AKQ, K562 accession numbers: ENCFF764ZIV, ENCFF489VUK) were excluded from further analysis. ENCODE DNase-seq raw FASTQs (paired-end 36 bp) of roughly equivalent size (HepG2 accession numbers: ENCFF002EQ-G,H,I,J,M,N,O,P) were downloaded from the ENCODE portal and processed using ENCODE DNase-seq standard pipeline (https://github.com/kundajelab/atac_dnase_pipelines) with flags: -species hg19 -nth 32 -memory 250G -dnase_seq -auto_detect_adapter -nreads 15000000 -ENCODE3. Processed BAM files were merged and used as input for footprinting with PIQ under default settings (Sherwood et al. 2014). Only footprints called with a PIQ Purity (positive predictive value) greater than 0.9 were used for subsequent analysis. High confidence DHS footprints were binned into a minimal set of non-overlapping 2-kb loci as described for ChIP-seq peaks above. The resultant set of 2-kb loci can be found in Table S4. TomTom was used to identify related DAP motifs. Specifically, DAP motif pairs that possessed a significant (FDR<0.05) 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* function. In all cases, >50% of the locus was required to overlap with a given annotation to assign it to an annotation. A bed file containing IDEAS regulatory annotations was obtained from http://main.genome-browser.bx.psu.edu/. Strong enhancers were designated as “Enh”, weak enhancers as “EnhW”, Promoters as “Tss”, “TssW”, “TssF”, or “TssCtcf” in the source file (Zhang et al. 2016). All other annotations were grouped into an “other” class. Gene coordinates were obtained from the ensemble genome browser (http://useast.ensembl.org/index.html) gene transfer format grch37.75 file. Gene expression data was 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). Reads were normalized to counts per million (CPM) and averaged across replicates. Cell-type specific genes were defined as those having a 4-fold greater CPM in a given cell line of interest than either of the other two cell lines and having a CPM value of at least two in the cell line of interest. HepG2 reporter assay data was obtained from previously published work hosted at the GEO accession GSE83894 in the file GSE83894_ActivityRatios.tsv (Inoue et al. 2017). Replicate average activity from the “MT” and “WT” columns were used for our analysis. GM12878 High Resolution Dissection of Regulatory Assay (HiDRA) data was obtained from previously published work hosted at the GEO accession GSE104001 in the file GSE104001_HiDRA_counts_per_fragmentgroup.txt (Wang et al. 2018). Fragments with zero plasmid DNA reads in any replicate were removed prior to analysis. Subsequently fragment reads were normalized by counts per million and replicate median log10(RNA/plasmid DNA) ratios were used for our analysis. 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 UCSC genome browser under the “Comparative Genomics” group. Copy number variation data was obtained from the ENCODE data portal under the file accession ENCFF074XLG. Deletions and amplifications were assigned as designated in the fourth column. POLR2A ChIA-PET BED files for K562 were obtained from the ENCODE data portal under the file accessions ENCFF001THW and ENCFF001TIC. BED regions greater than 10 kb were removed and regions less than 10 kb were expanded at their midpoint to 10 kb prior to further analysis. Results were nearly identical using either ChIA-PET replicate. Data shown is derived from data in the replicate contained under the file accession ENCFF001THW. Promoter capture Hi-C bed files for GM12878 was 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). We used the TS5_GM12878_promoter-other_significant_interactions.txt file for analysis. Similar to POLR2A ChIA-PET data above, bed regions greater than 10kb were removed and regions less than 10kb were expanded at their midpoint to 10kb prior to further analysis.

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. 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 demonstrated reporter activity in the top quartile of Inoue et. al. or in house data sets was the primary inclusion criteria (Inoue et al. 2017). Additionally, 50 negative control loci with low reporter assay activity based on previous in-house experiments and 371 GC-content matched control loci were included as negative controls. GC matched 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 Table S12. 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. A first round of PCR was performed with the STARR-seq oligo amp F and R primers (Table S13) to amplify the library and generate double stranded DNA, complete the Illumina sequencing primer sequences and add 15 bp of sequence homologous to the hSTARR-seq (Addgene #99292) plasmid for InFusion cloning. PCR was performed with 20 reactions consisting of 1 uL CustomArray library input, 10 uM primers and the KAPA HiFi 2x PCR Master Mix (KAPA Biosytems) with the

following conditions : 98 ^{o}C for 30 s, 20 cycles of 98°C for 15 s, 65°C for 30 s, 70°C for 30 s, and a final extension of 72 ^{o}C for 2 min. PCR products were pooled and cleaned up with the Zymo PCR Cleanup Kit (Zymo) before performing 2% agarose gel separation and extraction with the Zymo Gel Extraction Kit following manufacturer's instructions. The cleaned up and amplified library was diluted to 10 ng/uL and 25 ng was used as insert in a 3:1 insert:vector InFusion reaction with 150 ng of hSTARR-seq plasmid (linearized with AgeI and SalI) in five replicate reactions following manufacturer's instructions. InFusion reactions were pooled and cleaned and concentrated with 1.8X Ampure beads (Agencourt) in DNA LoBind tubes (Manufacturer) and eluted in 16 uL dH20. Six transformation reactions consisting of 2 uL of the cleaned and concentrated InFusion product were transformed into Lucigen Endura electrocompetent cells pooled and grown overnight at 37 ^{o}C in 2 L of LB ampicillin media at 200 RPM. Serial dilution plating of the transformation yielded an estimated library complexity of 9.7x10^{7} colonies, or roughly 1000x representation of the 9 x 10^{4} library elements. The full 2 L overnight culture was centrifuged at 5,500 RPM yielding a total pellet weight of 7.5 g from which the full plasmid library DNA was extracted using the Qiagen EndoFree GigaPrep and eluted in 2.5 mL TE buffer at a final concentration of 2.4 ug/uL following manufacturer's instructions. Final library representation was determined by amplifying the insert library with the STARR-seq Sequencing Primers containing P5 and P7 Illumina sequences (Table S14) in 10 reactions consisting of 10 ng of plasmid library, 10 uM primers and the KAPA HiFi 2x PCR Master Mix(KAPA Biosytems) with the following conditions : 98°C for 45 s, 16 cycles of 98°C for 15 s, 65°C for 30 s, 70°C for 30 s, and a final extension of 72°C for 2 min. PCR products were pooled, ran on 2 % agarose gel and extracted using the Zymo Gel Extraction Kit (Zymo) with final elution in 16 uL. The final sequencing library was quantified using Qubit dsDNA Broad Range (ThermoFisher) and the KAPA Library Quantification Kit (KAPA Biosytems) and sequenced on the Illumina MiSeq with PE 150 bp reads following standard protocols.

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) with 532 ug DNA using FuGene reagents at 4:1 ratio, with 12 replicate transfections. 24 hours after transfection, transfected cells were lysed on plate in RLT buffer (Qiagen) and stored at -80°C. The 12 replicates were condensed to 6 total replicates by combining cell lysates from two transfections. Total RNA was then isolated using the Norgen Total RNA Purification Kit using manufacturer's instructions. STARR-seq libraries were prepared as previously described with any modifications included below (Gaulton et al. 2013). Poly-A RNA was isolated in triplicate for each replicate with 75 ug RNA input using Dynabead Oligo-dT_{25} beads (Life Technologies) with double selection and eluted in 40 uL 10 mM Tris-HCl. PolyA-RNA was then subjected to DNase digestion with TURBO DNase (Life Technologies) at 37°C for 30 min, cleaned up with the Zymo RNA Clean and Concentrate Kit (Zymo) and eluted in 50 uL RNA elution buffer. Reverse Transcription was performed with 45 uL of the cleaned and concentrated RNA for each replicate with 2 uM STARR-seq Gene Specific Primer (Table S14) and SuperScript III Reverse Transcriptase as previously described (Gaulton et al. 2013). cDNA from the RT was then treated with RNAse A for 1 hr at 37 ^{o}C, cleaned up with 1.8X Ampure beads and eluted in 100 uL Buffer EB. Junction PCR was performed in quintuplicate for each replicate with 20 uL cDNA and 10 uM input Forward and Reverse Junction Primers (Table S13) using KAPA HiFi 2x PCR Master Mix (KAPA Biosystems) with the following conditions: 98°C for 45 s, 15 cycles of 98°C for 15 s, 65°C for 30 s, 72°C for 70 s, and a final extension of 72°C for 1 min. Junction PCR cDNA reactions were pooled by replicate and cleaned up with Ampure beads and eluted in 100 uL dH20. After optimization, sequencing PCR was performed in quadruplicate for each replicate with 5 uL Junction PCR cDNA, 10 uM input STARR-Seq Sequencing Primer F and indexed Sequencing Primer R (Table S13) with KAPA HiFi 2x PCR Master Mix(KAPA Biosytems) with the following conditions : 98°C for 45 s, 5 cycles of 98°C for 15 s, 65°C for 30 s, 70°C for 30 s, and a final extension of 72°C for 1 min. Replicate sequencing

PCR products were pooled, gel extracted with the Zymo Gel Extraction Kit (Zymo), eluted in 20 uL elution buffer and quantified with the Qubit Broad Range dsDNA kit. The six STARR-Seq RNA replicate libraries and a STARR-Seq Plasmid DNA input library (amplified as before but with 20 ng DNA input and 15 PCR cycles) were normalized with the KAPA Library Kit (KAPA Biosystems) and sequenced on an Illumina NextSeq with 150-bp paired-end reads using standard protocols.

STARR-seq data processing and analysis

FASTQ files were adapter trimmed using cutadapt version 1.2.1 prior to alignment. Trimmed reads were mapped to our oligo library using bowtie 2 version 2.2.5. A custom bowtie index was generated with our oligo library (Table S12) in FASTA format using the build command under default settings. Trimmed FASTQ files were subsequently aligned to our custom index in a manner that required a perfect sequence match only in the correct orientation. Specifically, the --norc, --score-min 'C,0,-1', and -k 1 flags were used with the remainder under default settings. Aligned SAM files were converted to count tables using samtools version 1.2 indexing. Our plasmid DNA library resulted in 49739461 reads with a 48.0% alignment rate. Our six RNA sequencing replicates resulted in an average of 31,942,788 reads (min = 4,189,978, max = 52,296,433). To balance read depths across replicates, we collapsed our six initial replicates into three replicates with an average of 63,885,576 reads (min = 56,486,411, max = 68,063,096) with an average alignment rate of 49.66%. A total of 90.4% of synthesized oligos were detected in our plasmid DNA library and 99.4% of test sequences were detected in at least one orientation. Ultimately, we applied a relatively strict count threshold filter by excluding oligos with less than 2 counts per million reads in our plasmid DNA library (44.7% of total oligos, 15.9% in both orientations) from further analysis.

Oligo activity was defined as replicate median log10(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 Tables S14-15.

Predicted mutation effects were determined using the lsgkm analysis suite in a manner previously described (Lee 2016; Ramaker et al. 2017). Briefly, genome sequence was obtained in FASTA format for each HepG2 DAP narrow peak using the bedtools *getfasta* command. A GC content matched set of null peak sites 10 times greater in number than the number of peak observed for each factor was generated using the *nullseq_generate* executable from the kmersvm website on the provided hg19 genome indices as described above. SVMs were trained on narrow peak and matched background sequences using gapped 10-bp kmers and allowing for 3 non-informative bases using the "gkmtrain" executable obtained from the ls-gkm github webpage (https://github.com/Dongwon-Lee/lsgkm). All other settings were left at default. This resulted in 208 SVMs, one for each DAP analyzed in HepG2. Each mutant oligo sequence was scored with the SVM trained on each DAP and its resultant classifier value was subtracted from the reference sequence classifier value to determine a mutation's predicted effect.

DAP enrichment for high impact mutations was computed using high confidence DHS motifs identified as described above. Enrichment P-values were calculated using a Fisher's exact test comparing the ratio of high effect mutations (differential activity>0.25) to the total number of mutations falling within vs. outside of a given DAPs footprints.

STARR-seq high impact mutation validation

To validate 14 identified SNVs with high impact in our assay, for each SNV we ordered ssDNA ultramers (Integrated DNA Technologies) corresponding to that SNV's reference sequence, the high impact SNV sequence, and a neighboring low impact SNV, flanked by 15 bp primer binding tails (Table S16) for each SNV. To generate dsDNA suitable for inFusion cloning,

we amplified Ultramers with primers containing sequence homologous to either the STARR-seq luciferase validation vector_mP_empty (Addgene# 99298) or pGL4.23 (Promega) (Table S13). Sequences were cloned into both vectors using inFusion cloning according to manufacturer's instructions. Plasmid DNA was extracted from three separate colonies with the Spin Miniprep Kit (Qiagen) and sequence verified with Sanger sequencing (MCLAB, San Francisco, CA). Each colony was treated as a separate biological replicate for a given sequence as previously described (Whitfield et al. 2012). HepG2 cells were seeded at 40,000 cells per well in antibiotic free DMEM with 10% FBS in a 96-well plate. After 24 hours, 300ng of plasmid DNA for each biological replicate was transfected into HepG2 cells using FuGENE (Promega) in triplicate, resulting in 9 total replicates (3 biological X 3 technical) per sequence. Luciferase activity was measured 48-hours post-transfection with a 2-second integration time on a LMax II 384 Luminometer (Molecular Devices). Background subtracted luminescence values for each SNP were z-scored. Significance in expression was determined using a 2-tailed Student's T-test.

Data Availability

Raw FASTQs and count tables for our mutagenesis assay have been submitted to GEO. Final accession number is pending.

Acknowledgments

We thank Eric Mendenhall and Surya Chhetri for their assistance with the alignment and quality control analysis of ChIP-seq experiments in HepG2, and are particularly grateful to them and the Myers/Mendenhall ENCODE group members, including Mark Mackiewicz, Kim Newberry, Dianna Moore, Laurel Brandsmeier, Sarah Meadows, Megan McEown, for their generation of the high-quality ChIP-seq data used in this paper. This work was supported by

National Institute 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.).

References

Arnosti DN, Kulkarni MM. 2005. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J Cell Biochem* **94**: 890–898.

Basta J, Rauchman M. 2009. The nucleosome remodeling deacetylase (NuRD) complex in Development and Disease. *Transl Res* **165**: 36–47.

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**: 1–11.

DeLaForest A, Nagaoka M, Si-Tayeb K, Noto FK, Konopka G, Battle M a, Duncan S a. 2011. HNF4A is essential for specification of hepatic progenitors from human pluripotent stem cells. *Development* **138**: 4143–53.

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**: W544–56.

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

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 (80-)* **339**.

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**: 1.

Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. 2007. Quantifying similarity between motifs. *Genome Biol* **8**.

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 1 encoding of enhancer activity 2 Running title: Comparing chromosomal and episomal reporter assays 3 4. *Genome Res* 38–52.

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.

Johnson DS, Mortazavi A, Myers RM. 2007. Protein-DNA Interactions. *Science (80-)* 1497–1503.

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.

Landt S, Marinov G. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* **22**: 1813–1831.

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

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

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.

Mitchell PJ, Tjian R. 1989. Transcriptional Regulation in Mammalian Cells by Sequence-Specific DNA Binding Proteins. **245**: 371–378.

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.

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**.

Reddy TE, Gertz J, Pauli F, Kucera KS, Varley KE, Newberry KM, Marinov GK, Mortazavi A, Williams B a, Song L. 2012. The effects of genome sequence on differential allelic transcription factor occupancy and gene expression. *Genome Res* **22**: 1–11.

Roadmap Epigenomics Consortium. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**: 317–330.

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

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

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

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

Tan NY, Khachigian LM. 2009. Sp1 Phosphorylation and Its Regulation of Gene Transcription. *Mol Cell Biol* **29**: 2483–2488.

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.

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

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.

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.

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

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**.

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.

Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, et al. 2014. NIH Public Access. *Cell* **158**: 1431–1443.

Whitfield TW, Wang J, Collins PJ, Partridge EC, Aldred SF, Trinklein ND, Myers RM, Weng Z. 2012. Functional analysis of transcription factor binding sites in human promoters. *Genome Biol* **13**: R50.

Wreczycka K, Franke V, Uyar B, Wurmus R, AKALIN A. 2017. HOT or not: Examining the basis of high-occupancy target regions. Nucleic Acids Res **47**: 5735-5745.

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

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.

Figure Legends

Figure 1. (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 DAPs. The color indicates the

percentage of true HOT sites, as defined by >25% of DAPs bound in the full dataset, 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 shows data for nssDAPs, and the yellow dashed line shows data for ssTFs. (C) Boxplots demonstrating the number of ChIP'ed DAPs with a corresponding DFM present at the same locus at various levels of DAP co-association. (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 demonstrated the fraction of HOT sites that contain a ssTF ChIP peak and a DFM. ssTFs highlighted in the top right are putative driver TFs present at high proportion of HOT sites.

Figure 2. (A) IDEAS annotations of loci binned by ChIP-defined DAP associations. (B-C) The expression level of the nearest gene to each loci binned by ChIP-defined DAP associations. Plots show loci either distal (>5kB, B) or proximal (<5kB,C) to their nearest gene. (D) ChIP- and DFM-defined co-association correlates with activity in a previous high throughput reporter assay conducted on ~2000 selected enhancer regions in HepG2. (E) GM12878 ChIP-defined DAP associations correlate strongly with previously published ATAC-STARR-seq reporter assay activity. Green boxes for plots D and E show loci binned by the presence or absence of common markers of enhancer or promoter regions that delineated a subset of the data.

Figure 3. (A) Example loci describing mutagenesis schema. The red region indicates a 130bp core, centered upon the maximum number of unique ChIP-seq peaks and DFMs, in which we performed tiled single bp and five bp mutagenesis in both the forward and reverse orientation. The flanking green regions represent 130bp regions flanking the core region in which we performed tiled 5bp 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. (C) Plot indicating the proportion of mutations imposing a change of activity at a variety of thresholds for five bp mutations. Green points indicate data for mutations falling within DHS footprints. Red points indicate data for mutations falling outside of DHS footprints. (D) Scatter plot demonstrated the fraction of HOT sites that contain a ssTF ChIP peak and DFM. TFs highlighted in the top right putative direct binding TFs at a high proportion of HOT sites. The size of each point corresponds to a ssTFs DFM enrichment for high impact mutations in our mutagenesis assay. (E) These barplots show the cumulative differential activity (locus mean – mutation) across all positions in the HNF4 motif. (F) Scatter plot demonstrating the number of non-redundant DFMs at a locus is inversely correlated with its vulnerability to mutation (expressed as the sum of all mutation delta activity scores).

Figure 4. (A-B) The number of HepG2 loci bound by at least one DAP in GM12878 (A) and K562 (B). The red dot indicates the number of loci shared by an equivalent number of DAPs in both cell lines. (C) The number of HOT loci present in all possible combinations of each cell line. (D-F) Barplots indicating the fraction of each loci, stratified by number of ChIP-defined DAP associations in HepG2, that are present near HepG2 (D), GM12878 (E), or K562 (F) specific genes. Cell specific genes were computed by randomly sampling 500 genes that were expressed at least 4 fold higher in the cell line of interest than the other two cell lines and had an FPKM of at least two in the cell line of interest.

Figure 5. (A) Scatter plot indicating the observed rate of HOT loci as a function of the expected rate based on gene density in each HepG2 chromosome. The dashed line represents a linear fit of all data. Data points above the dashed line indicate a higher occurrence of HOT loci than expected. (B) Stacked bar plots showing proportion of loci with various levels of ChIP-derived DAP associations in genomic regions that have heterozygous deletions, amplifications, or normal copy number in HepG2. (C,D) Boxplots demonstrating the correlation between the

number of ChIP-defined DAP associations and the number of POLR2A ChIA-PET interactions in HepG2 (A) or K562 (B).

Supplemental Figures

Figure S1. Cumulative distribution function (CDF) showing the proportion of loci containing a given level of unique DAP or ssTF ChIP-seq peaks across the HepG2, K562 and GM12878 cell lines. Colors correspond to cell lines. Dashed lines indicate ssTF data only and solid lines represent data that includes all DAPs.

Figure S2. Number of loci reaching "HOT" threshold of 25% of DAPs associated via the ChIP-seq assay after performing random down sampling of the number of DAPs included. Each data point represents the result of a random sampling of a DAPs. The color indicates the percentage of true HOT sites, as defined by >25% of DAPs bound in the full dataset, 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. Data for HepG2 (A), K562 (B), and GM12878 (C) is shown.

Figure S3. Cumulative distribution function (CDF) showing the proportion of loci containing a given number of ChIP-seq defined DAP associations or randomly shuffled DAP associations. Random shuffling was performed in a manner that preserved the total number of DAP associations across all loci.

Figure S4. IDEAS annotations of loci binned by the number DFM-defined DAP associations.

Figure S5. Pie charts demonstrating the proportion of loci associated, based on ChIP-seq, with a specified number of DAPs for a variety of IDEAS annotations

Figure S6. The expression level of the nearest gene to each loci binned by DFM-define DAP occupancy. Plots show loci distal (>5 kb, A) or proximal (<5 kb, B) to their nearest gene.

Figure S7. (A) Base-wise conservation at each position of the HNF4A motif as defined by

GERP is plotted for each occurrence of the HNF4A motif with greater or less than 9 neighboring DFMs. (B-C) Spearman rho values representing the correlation between the median GERP score of each ssTF's DFM and the number of unique DAP ChIP-seq peaks (B) or neighboring DFMs (C) in its 2kB locus. ***, **, and * denote Wilcox P-values of <0.05, <0.005, and <0.0005 respectively.

Figure S8. IDEAS annotations of loci binned by the number ChIP-seq-defined ssTF associations.

Figure S9. Histograms indicating the distribution of reads assigned to oligonucleotides detected in DNA harvested from transfected cells. (A) represents data for all oligonucleotides including mutations. (B) represents only reference oligos from the "core" central 130-bp window of each assayed loci.

Figure S10. Boxplots demonstrating the number of counts for each single bp mutation oligo in the DNA input library. Each box contains 490 data points (one for each of our 245 loci in both orientations). A similar pattern was observed for our 5mer mutations suggesting no significant bias was introduced in our alignment methodology.

Figure S11. Scatter plot of the correlation between RNA counts per million (median across the 3 reps) and DNA counts per million. A strong correlation was observed as expected however significant residual variance was also observed with a subset of mutant sequences falling significantly short of their expected activity.

Figure S12. (A) Boxplots showing the fraction of oligonucleotides retained in our library at increasing DNA input representation thresholds. (B) Box plots showing replicate correlation of the RNA/DNA ratio at increasing DNA input representation thresholds.

Figure S13. Boxplots indicating that the distribution of activity in single mutant loci was roughly equivalent to that of WT regions with a slight decrease in activity observed in 5-bp mutations. The left group of boxes show data from the forward strand and the right group show data from reverse strand oligonucleotides.

Figure S14. Scatter plots show the correlation in the differential activity (mutant oligo – the locus mean) between the forward and reverse strands for single base pair mutations (A) and 5-bp mutations (B). The color indicates the number of points in contact with a given location on the plot.

Figure S15. Line plots showing the fraction of mutations that meet increasing thresholds of differential activity (loci mean – mutant oligo) for transitions versus transversions.

Figure S16. Validation data for the luc_mp_empty vector. Boxplots show inter-element z-scored data for reference (WT) oligos, oligos with loss of function (LOF) or gain of function (GOF) mutations, and oligos with predicted null mutations. Locus 40 is the only locus with a predicted GOF mutation. The remainder are predicted LOF.

Figure S17. Validation data for the PGL4.23 vector. Boxplots showing inter-element z-scored data for reference oligos, oligos with loss of function (LOF) or gain of function (GOF) mutations, and oligos with predicted null mutations. Locus 40 is the only locus with a predicted GOF mutation. The remainder are predicted LOF.

Figure S18. Plot indicating the proportion of mutations imposing a change of activity at a variety of thresholds for single base pair mutations. Green points indicate data for mutations falling within DHS footprints. Red points indicate data for mutations falling outside of DHS footprints.

Figure S19. Scatter plot describing the correlation between LSGKM predicted DAP binding disruptions and differential activity of mutant oligos. Color indicates the number of points in contact with a give region of the plot.

Figure S20. Bar plots showing the cumulative differential activity (loci mean – mutation) across all positions in the FOXA or AP1 motifs. Activity correlates with the motif consistency at each position.

Figure S21. Boxplots indicating the magnitude of the effect size (absolute beta value) for significant (FDR<0.05) liver GTEx eQTL SNPs as function of the number of unique DFMs (A) or DAP ChIP-seq peaks (B) detected at the locus to which they map. *** Denotes Wilcoxon P<5e-

16

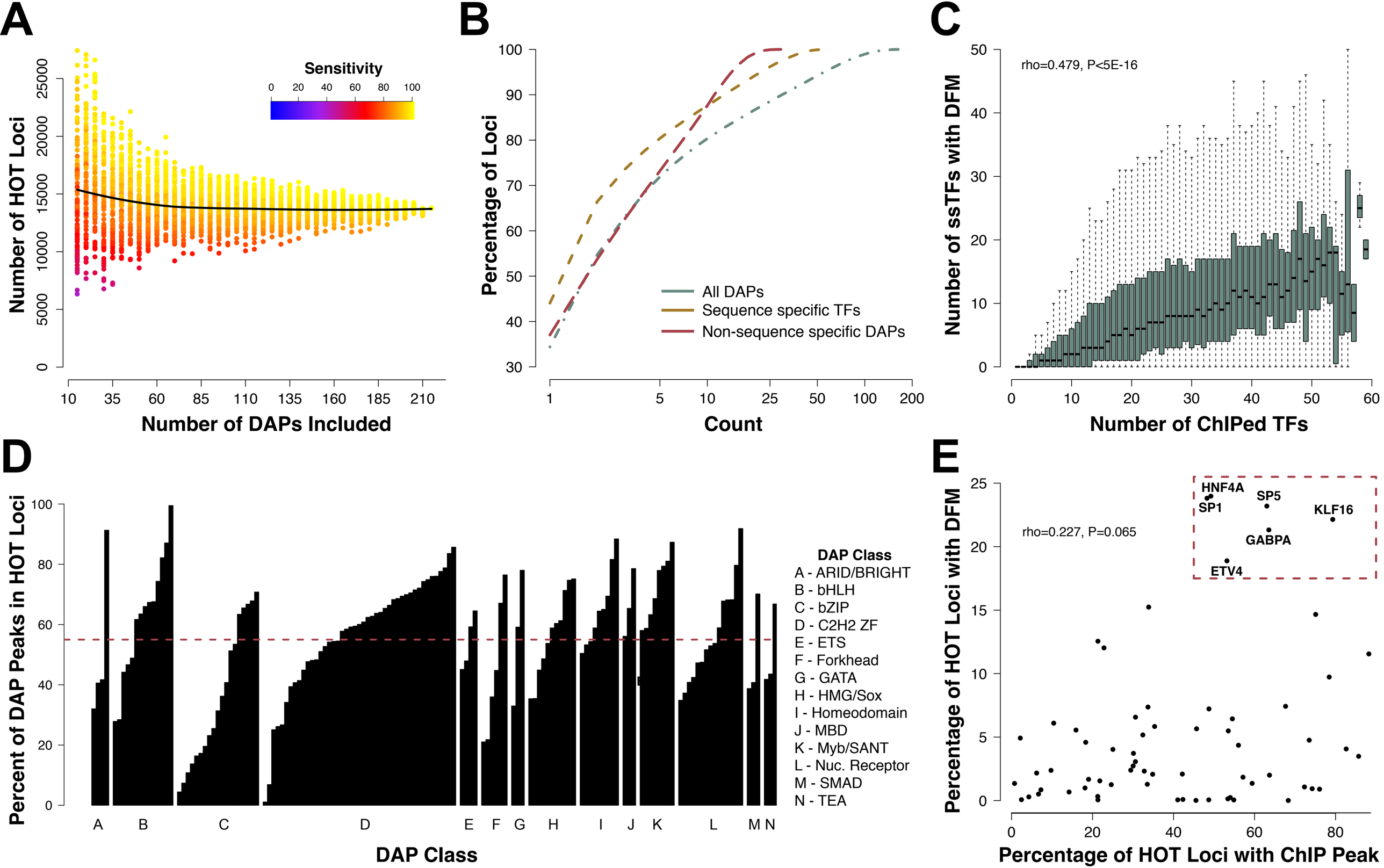
Figure S22. Scatter plot indicating the observed rate of HOT loci vs. the expected rate based on gene density in each K562 (A) or GM12878 (B) chromosome.

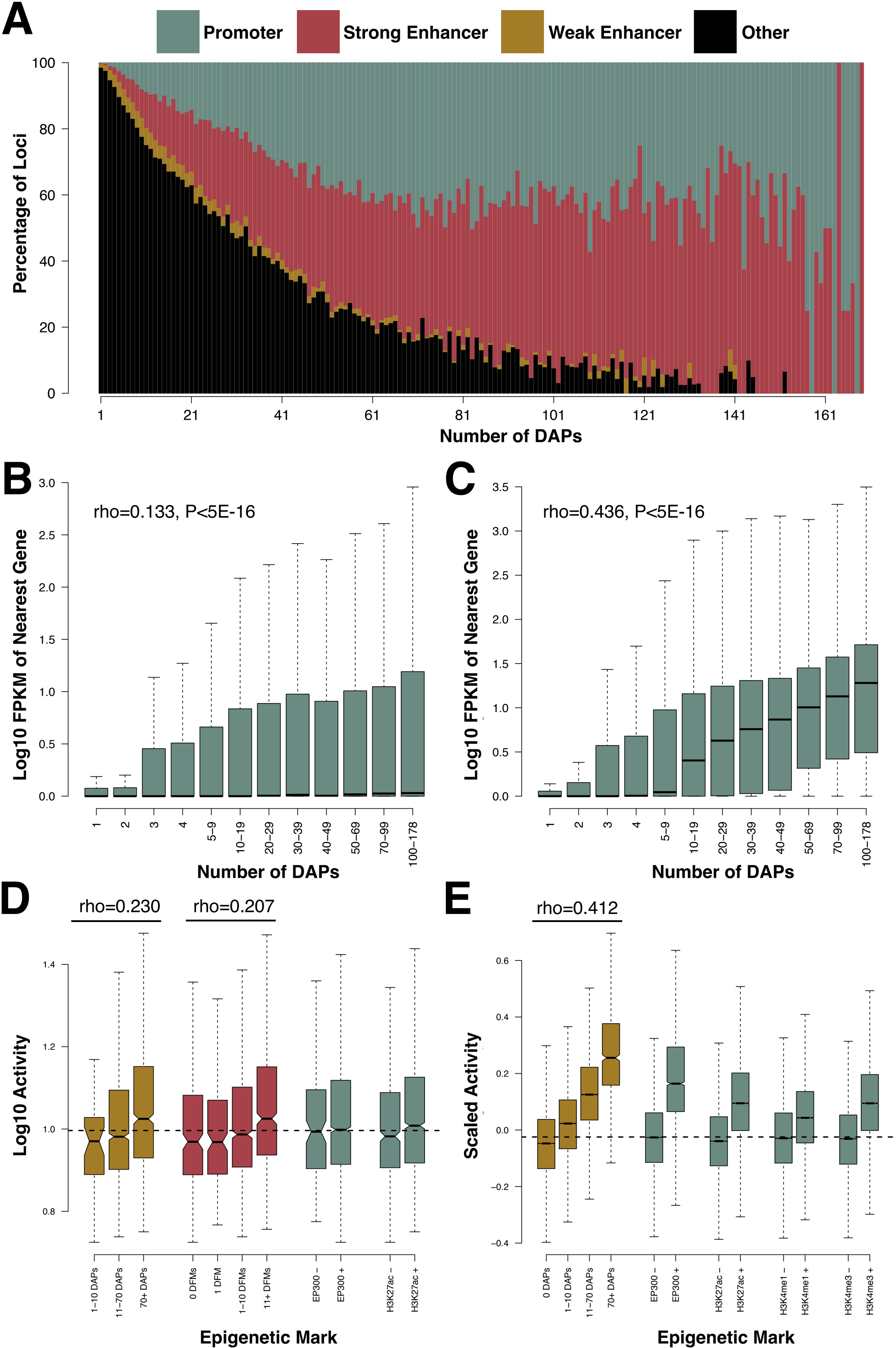
Figure S23. Bar plots showing the proportion of loci amplified in HepG2 at increasing numbers of ChIP-seq derived DAP associations.

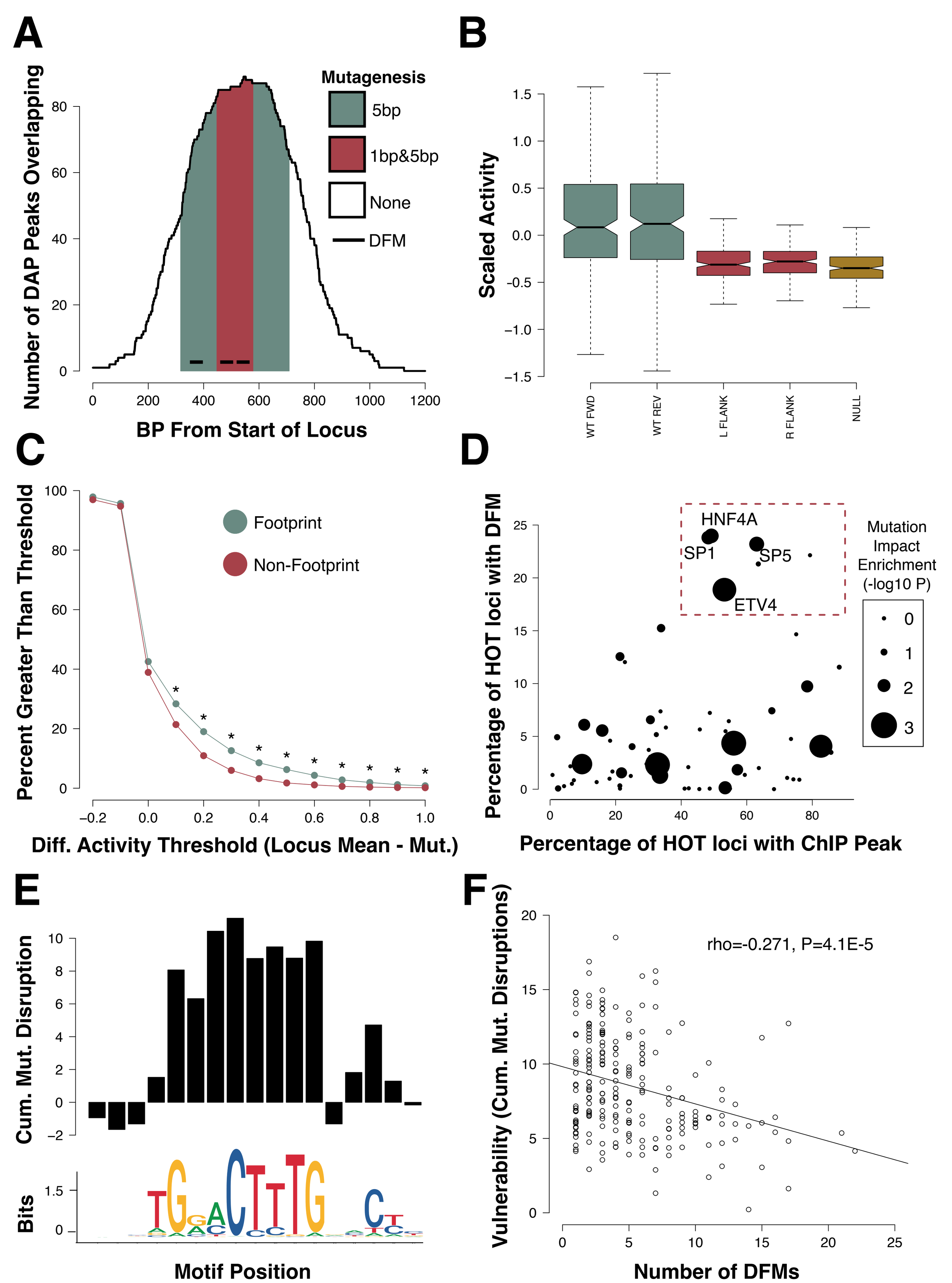
Figure S24. (A) Boxplot demonstrating the correlation between the number of DAPs bound (based on ChIP-seq peaks) and the number of ENCODE POLR2A ChIA-PET interactions observed in K562. (B) Boxplot describing the correlation between the number of DAPs bound (based on ChIP-seq peaks) and the number of observed promoter capture Hi-C interactions in GM12878. Y-axis capped at 30.

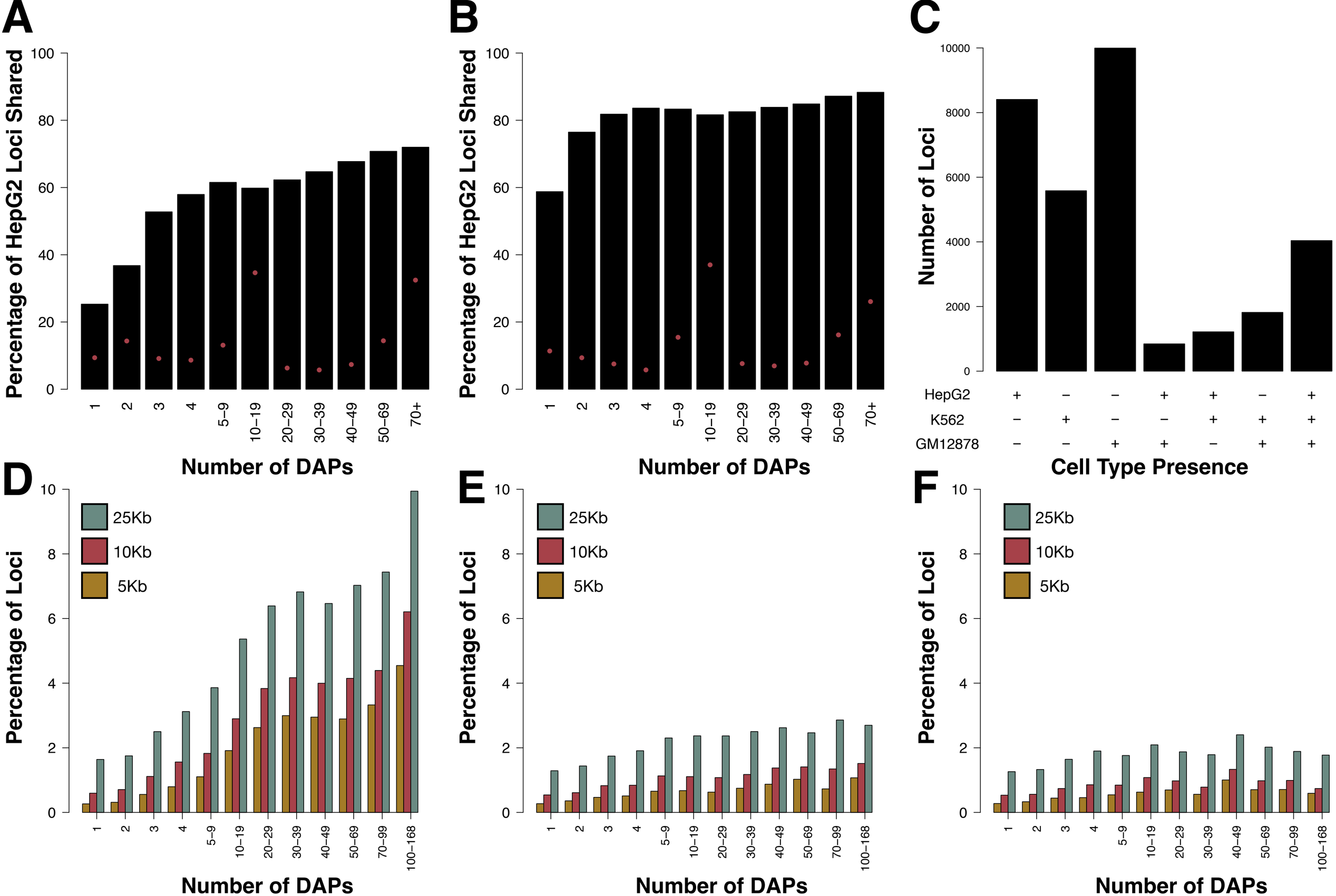
Figure S25. Boxplots demonstrating the fraction of DAPs in common between interacting loci matched non-interacting loci for K562 ChIA-PET data and GM12878 promoter capture Hi-C data. Non-interacting loci have the same total number of bound DAPs as interacting loci.

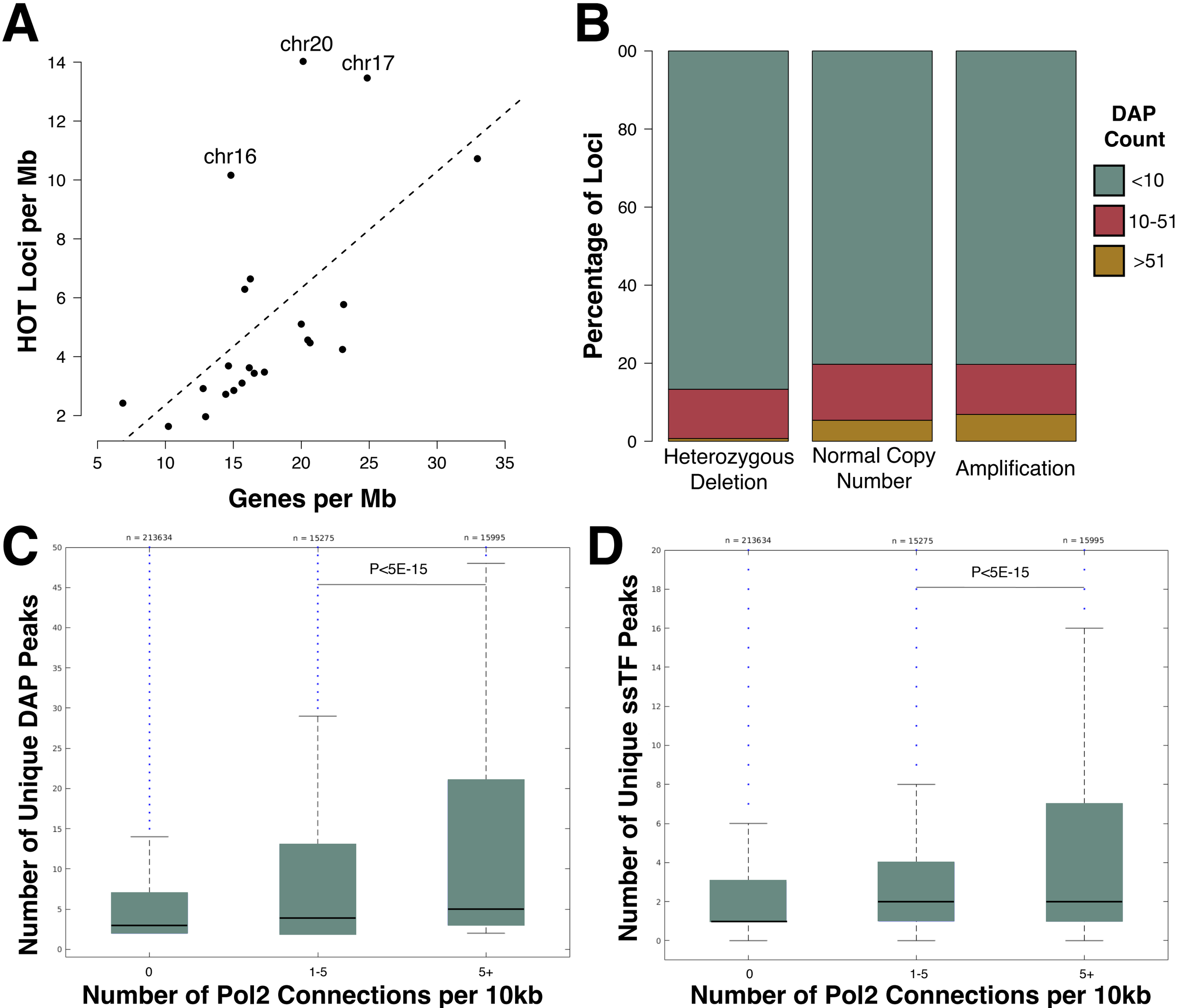
Figure S26. (A) Boxplots indicating the distance in base pairs of the nearest loci with an equivalent number of DAP associations as specified by the box colors (legend on right). (B) Cumulative distribution functions displaying data from the boxplots shown in (A). Each line indicates the cumulative fraction of loci that contain an neighboring loci with an equivalent number of DAP associations (as specified by the line color) within a given distance in base pairs.


A
Sensitivity
0 20 40 60 80 100
Number of HOT Loci
0 5000 10000 15000 20000 25000
10 35 60 85 110 135 160 185 210
Number of DAPs Included
B
Percentage of Loci
30 40 50 60 70 80 90 100
1 5 10 25 50 100 200
Count
All DAPs
Sequence specific TFs
Non-sequence specific DAPs
C
rho=0.479, P<5E-16
Number of ssTFs with DFM
0 10 20 30 40 50
0 10 20 30 40 50 60
Number of ChIPed TFs
D
Percent of DAP Peaks in HOT Loci
0 20 40 60 80 100
A B C D E F G H I J K L M N
DAP Class
DAP Class
A - ARID/BRIGHT
B - bHLH
C - bZIP
D - C2H2 ZF
E - ETS
F - Forkhead
G - GATA
H - HMG/Sox
I - Homeodomain
J - MBD
K - Myb/SANT
L - Nuc. Receptor
M - SMAD
N - TEA
E
rho=0.227, P=0.065
HNF4A
SP1
SP5
KLF16
GABPA
ETV4
Percentage of HOT Loci with DFM
0 5 10 15 20 25
0 20 40 60 80
Percentage of HOT Loci with ChIP Peak


A
Promoter
Strong Enhancer
Weak Enhancer
Other
Percentage of Loci
Number of DAPs
B
rho=0.133, P<5E-16
Log10 FPKM of Nearest Gene
Number of DAPs
C
rho=0.436, P<5E-16
Log10 FPKM of Nearest Gene
Number of DAPs
D
rho=0.230
rho=0.207
Log10 Activity
Epigenetic Mark
E
rho=0.412
Scaled Activity
Epigenetic Mark


A
Mutagenesis
5bp
1bp&5bp
None
DFM
Number of DAP Peaks Overlapping
BP From Start of Locus
B
Scaled Activity
WT FWD
WT REV
L FLANK
R FLANK
NULL
C
Footprint
Non-Footprint
Percent Greater Than Threshold
Diff. Activity Threshold (Locus Mean - Mut.)
D
HNF4A
SP1
SP5
ETV4
Mutation Impact Enrichment (-log10 P)
Percentage of HOT loci with DFM
Percentage of HOT loci with ChIP Peak
E
Cum. Mut. Disruption
Bits
Motif Position
F
rho=-0.271, P=4.1E-5
Vulnerability (Cum. Mut. Disruptions)
Number of DFMs


A
Percentage of HepG2 Loci Shared
100
80
60
40
20
0
1
2
3
4
5–9
10–19
20–29
30–39
40–49
50–69
70+
B
Percentage of HepG2 Loci Shared
C
Number of Loci
10000
8000
6000
4000
2000
0
HepG2 + − − + + − +
K562 − + − − + + +
GM12878 − − + + − + +
D
Number of DAPs
Percentage of Loci
25Kb
10Kb
5Kb
100–168
E
F
Cell Type Presence


A
chr20
chr17
chr16
HOT Loci per Mb
Genes per Mb
B
Percentage of Loci
Heterozygous Deletion
Normal Copy Number
Amplification
DAP Count
<10
10-51
>51
C
n = 213634
n = 15275
n = 15995
P<5E-15
Number of Unique DAP Peaks
Number of Pol2 Connections per 10kb
D
n = 213634
n = 15275
n = 15995
P<5E-15
Number of Unique ssTF Peaks
Number of Pol2 Connections per 10kb