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 (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 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 (Quinlan 2014). 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 (5.4% of the total) 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 2kb loci can be found in Supplemental_Table_S3, Supplemental_Table_S5, and Supplemental_Table_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 Supplementary Material and in the github repository: 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 (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 the Kundaje lab, 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 Supplemental_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* and *map* functions. In all cases, >50% of the locus was required to overlap with a given annotation to assign it to an annotation. A source BED file containing IDEAS regulatory annotations was obtained from http://main.genome-browser.bx.psu.edu/. Loci containing an "Enh" annotation were designated as "strong enhancers", those containing an "EnhW" annotation as "weak enhancers", those containing "Tss", "TssW", "TssF", or "TssCtcf" as "promoters" in the source file. All other annotations were grouped into an "other" class (Zhang et al. 2016). 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. 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 $\log_{10}$(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 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. 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). A processed BED file of significant interactions at 4 DpnII fragment resolution was graciously provided upon request of the Grant lab. BED regions greater than 10kb were removed and regions less than 10kb were expanded at their midpoint to 10kb prior to further analysis. 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. POLR2A ChIA-PET BED files containing significant 3D interactions for HepG2 were made available upon request from the Yijun Ruan lab. For both ChIA-PET analyses, BED regions greater than 10kb were removed and regions less than 10kb were expanded at their midpoint to 10 kb prior to further analysis. 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. BED files containing repetitive element alignment scores were obtained from the 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 $P<5\times10^{-16}$ 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 demonstrated activity in the HepG2 cell line in Inoue et. al, 2017, because we reasoned a baseline level of reporter assay activity is required to see differential activity upon mutation (Inoue et al. 2017). Roughly two thirds of our loci met our specified HOT threshold all details regarding the loci assayed are available in Supplemental_Table_S9. 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 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. A first round of PCR was performed with the STARR-seq oligo amp F and R primers (Supplemental_Table_S16) 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^{o}C for 15 s, 65^{o}C for 30 s, 70^{o}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 9.7×10^{7} colonies, or roughly 1000x representation of the 9×10^{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 (Supplemental_Table_S16) in 10 reactions consisting of 10 ng of plasmid library, 10 uM primers and the KAPA HiFi 2x PCR Master Mix (KAPA Biosystems) with the following conditions : 98^{o}C for 45 s, 16 cycles of 98^{o}C for 15 s, 65^{o}C for 30 s, 70^{o}C for 30 s, and a final extension of 72^{o}C for 2 min. PCR products were pooled, run 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 (Thermo Fisher Scientific) 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^{o}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^{o}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 (Supplemental_Table_S16) 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 (Supplemental_Table_S16) using KAPA HiFi 2x PCR Master Mix (KAPA Biosystems) with the following conditions: 98^{o}C for 45 s, 15 cycles of 98^{o}C for 15 s, 65^{o}C for 30 s, 72^{o}C for 70 s, and a final extension of 72^{o}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 (Supplemental_Table_S16) with KAPA HiFi 2x PCR Master Mix (KAPA Biosytems) with the following conditions : 98^{o}C for 45 s, 5 cycles of 98^{o}C for 15 s, 65^{o}C for 30 s, 70^{o}C for 30 s, and a final extension of 72^{o}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 (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 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. This alignment procedure can be reproduced with the STARR_SEQ_Mutagenesis folder in the github repository: https://github.com/rramaker/GenomeTools2020. 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 $\log_{10}$(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_Table_S17-18.

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 (Supplemental_Table_S19) 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) (Supplemental_Table_S16). 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 SNV were z-scored. Significance in expression was determined using a 2-tailed Student's *t*-test.

Identifying DFMs with TSS proximity or cell type-specific regulatory loci bias

We explored splitting HOT regions into those distal (>5kb from a TSS) and proximal (<5kb from a TSS) and found our driver ssTFs did not fully segregate into purely proximal or distal categories (Supplemental Fig S3A-C, Supplemental Table S12). However, we found HNF4A exhibited a ~2 fold preference for distal HOT sites while ssTFs with ETS or SP1 family motifs showed ~3 fold preference for proximal HOT sites. We did not immediately identify an ssTF with a similar enrichment for distal loci in K562 or GM12878,

Cell type-specific HOT loci were defined as loci that met our HOT threshold (25% of DAPs assayed with a ChIP-seq peak) in only one cell line. Cell type-ubiquitous HOT loci were defined as loci that met our HOT threshold in HepG2, K562 and GM12878. An examination of differential DFM enrichment between cell type-specific and cell type-ubiquitous HOT loci revealed as similar pattern to that observed for TSS proximity bias amongst DFMs (Fig 2E, Supplemental Fig S9C-E, Supplemental Table S12). In HepG2, HNF4A exhibited roughly 3-fold enrichment for HepG2-specific HOT loci relative to ubiquitously HOT loci while ETS and SP1 motifs had the opposite enrichment for ubiquitous HOT loci. K562 and GM12878 cell lines were relatively depleted for DAPs enriched for cell type specific HOT loci.

To gain further resolution on potential distal, cell type-specific driver DFMs, we plotted DFM TSS proximity bias against cell type-specificity bias (Fig 4B, Supplemental Fig 9F-G). This revealed the HN4A and FOX family of motifs as enriched distal, cell type specific HOT loci in HepG2 (Fig 4B). K562 and GM12878 had less prominent distal, cell type specific regulators, but the top motifs GATA, NFE2, TBX family, IRF8, and SPI1 (Supplemental Fig 9F-G).

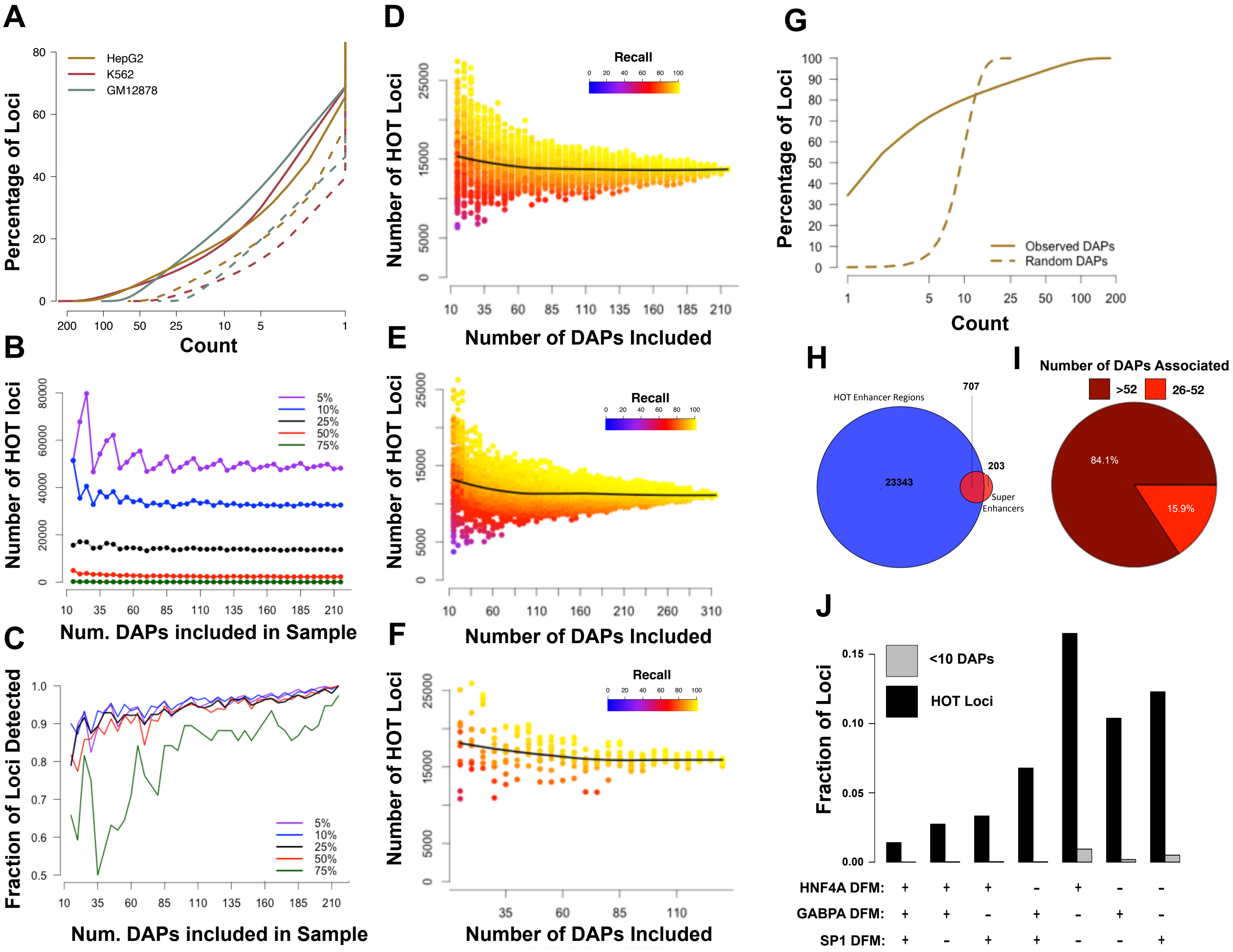


Figure S1. (A) Cumulative distribution function (CDF) showing the proportion of loci containing at least a given level of unique DAP or ssTF ChIP-seq peaks across the HepG2, K562 and GM12878 cell lines. Colors correspond to cell lines. Dashed lines indicate ssTF data only and solid lines represent data that includes all DAPs. (B) Number of loci reaching "HOT" threshold via the ChIP-seq assay after performing random down sampling of the number of DAPs included. Each data point for both plots represents the median result of a 100 random samples at a number indicated by the x-axis. The color of each line indicates the % of DAPs required to reach the 'HOT' threshold. (C) The recall performance or fraction of true HOT sites, as defined by the full dataset, detected with current sample of DAPs. Each data point for both plots represents the median result of a 100 random samples at a number indicated by the x-axis. The color of each line indicates the % of DAPs required to reach the 'HOT' threshold. (D-F) 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 specified number of DAPs. The color indicates recall performance or 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 at each number of DAPs as specified by the x-axis. Data for HepG2 (D), K562 (E), and GM12878 (F) is shown. (G) Cumulative distribution function (CDF) showing the proportion of loci included over an increasing exclusion threshold. The solid line represents observed HepG2 DAP ChIP-seq distributions and the dashed line represents a random shuffling of DAP assignments across all loci containing a peak. Random shuffling was performed in a manner that preserved the total number of DAP associations across all loci. (H) Venn diagram showing the overlap between HOT loci and super enhancers defined via the ROSE method. (I) Pie chart showing the number of DAPs associated with super enhancers. (J) Bar plot displaying the fraction of all HOT (black) or <10 DAP (gray) associated loci that have a specified combination of DFMs present.

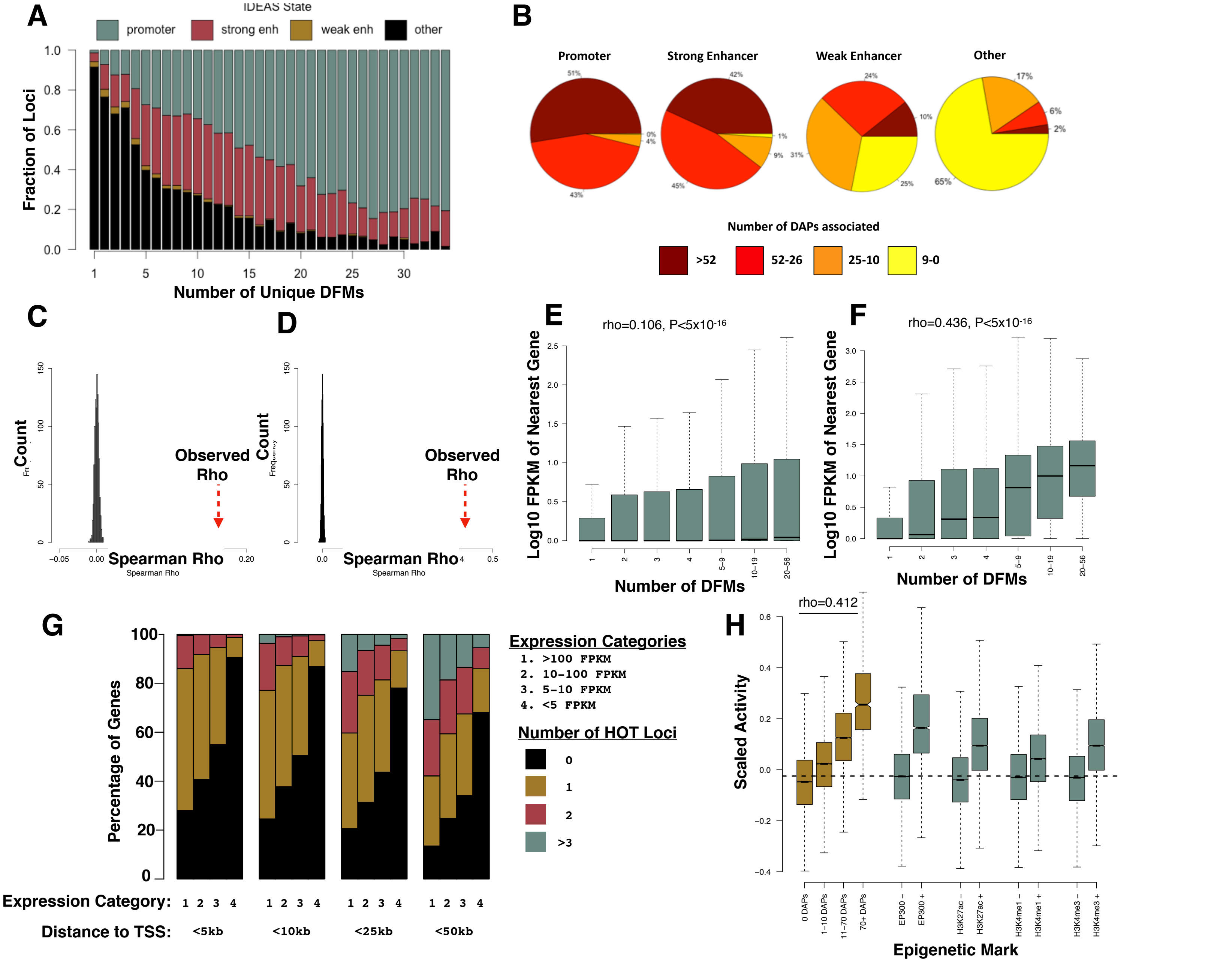


Figure S2. (A) IDEAS annotations of loci binned by the number of DFM-defined DAP associations. Promoter, strong enhancer, and weak enhancer annotation represent 0.27, 0.35, and 0.22% of the HepG2 genome, while the remaining 99.16% of the genome (largely consisting of quiescent and repressed annotations) was used for the "Other" annotation. (B) Pie charts demonstrating the proportion of loci associated, based on ChIP-seq, with a specified number of DAPs for a variety of IDEAS annotations. (C-D) Histograms displaying the distribution of correlations observed from 1000 random shuffles of gene-locus associations relative to that observed in the non-shuffled data sets shown in Figure 2B-C. (E-F) The expression level of the nearest gene to each loci binned by DFM-defined DAP occupancy. Plots show loci distal (>5kb, E) or proximal (<5kb, F) to their nearest gene. (G) Stacked bar plots displaying the proportion of genes meeting a specified expression level bin with a specified number of neighboring HOT loci at a specified distance to TSS threshold. (H) 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 th epresence or absence of common markers of enhancer or promoter regions.

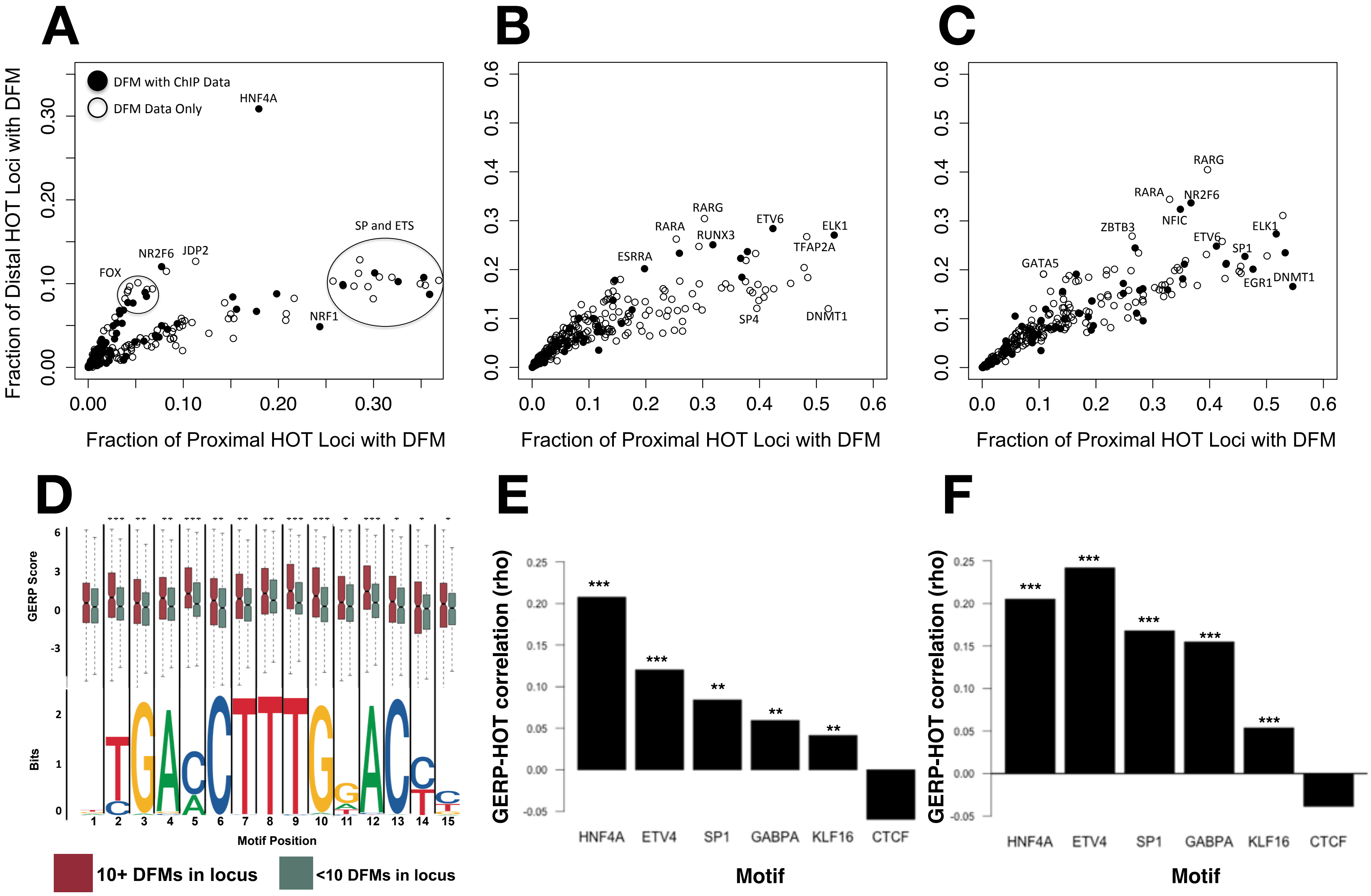


Figure S3. (A-C) Scatterplots demonstrating the fraction of distal (>5kb from a TSS) and proximal (<5kb from a TSS) HOT sites that contain a DFM for each ssTF in HepG2 (A) GM12878 (B) and K562 (C). Points in the top left of this plot indicate ssTFs enriched for distal HOT loci and points in the bottom right of this plot indicate ssTFs enriched for proximal HOT loci. (D) 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. (E-F) Spearman rho values representing the correlation between the median GERP score of each ssts's DFM and the number o unique DAP ChIP-seq peaks (E) or neighboring DFMs (F) in its 2kB locus. ***, **, and * denote Wilcox P-values of <0.0005, <0.005, and <0.05 respectively

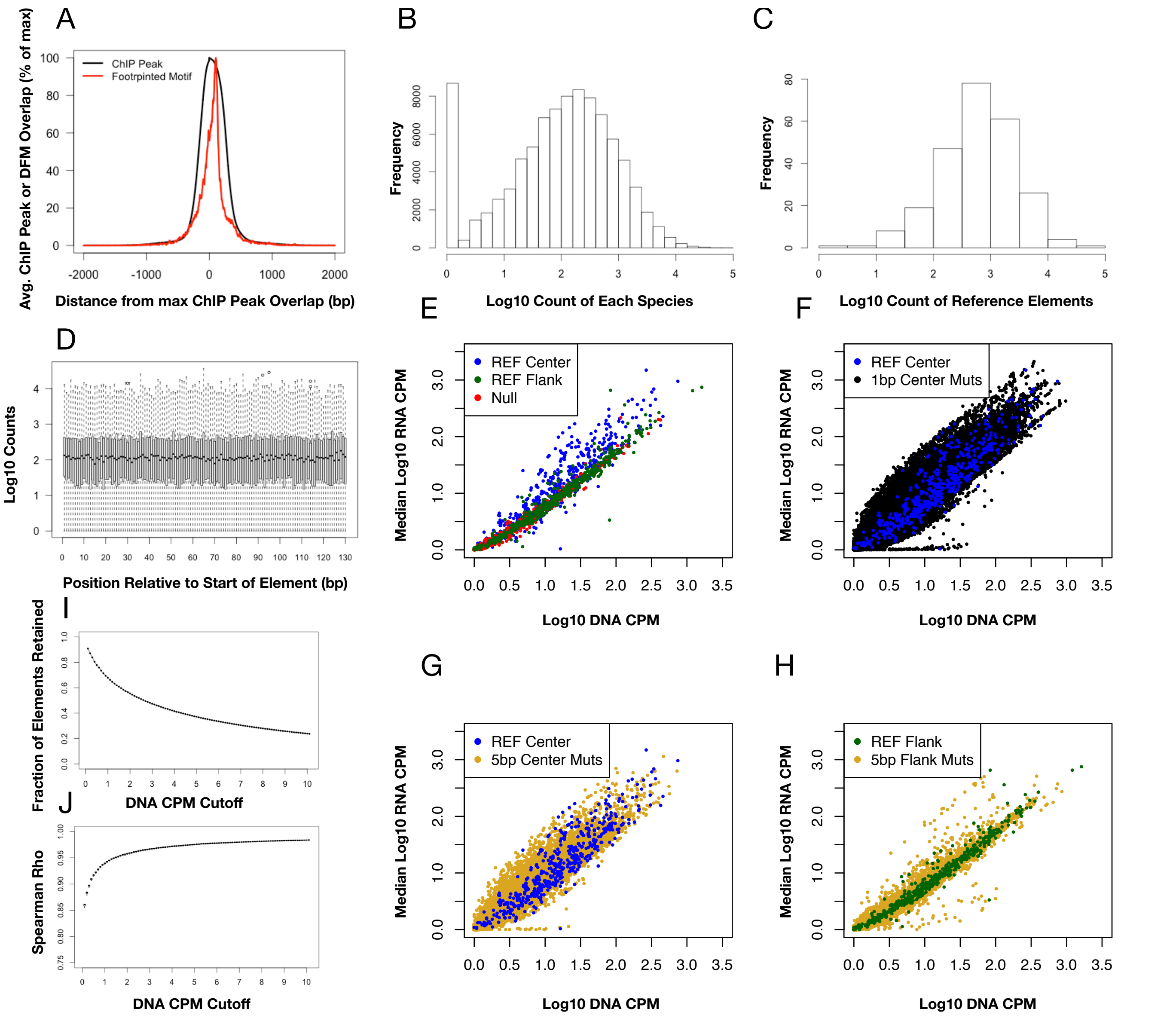


Figure S4. (A) Line plot indicating the positions of ChIP-seq peak (black) and DFM (red) pile up relative to point of maximum ChIP-seq peak pile up. Each point on the line indicates the average overlap, as a percentage of the maximum point of overlap, for all 13,792 HOT loci in HepG2. (B-C) Histograms indicating the distribution of reads assigned to oligonucleotides detected in DNA harvested from transfected cells. (B) represents data for all oligonucleotides including mutations. (C) represents only reference oligos from the “core” central 130-bp window of each assayed loci. (D) 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. (E-F) Scatter plot of the correlation between RNA counts per million (median across the 3 reps) and DNA counts per million. Plots are data subsets: (E) shows all reference HOT and null sequences, (F) shows central region reference elements and corresponding 1bp mutations, (G) shows central region reference elements and corresponding 5bp mutations, and (H) shows flanking reference elements and corresponding 5bp mutations. (I) Boxplots showing the fraction of oligonucleotides retained in our library at increasing DNA input representation thresholds. (J) Box plots showing replicate correlation of the RNA/DNA ratio at increasing DNA input representation thresholds.

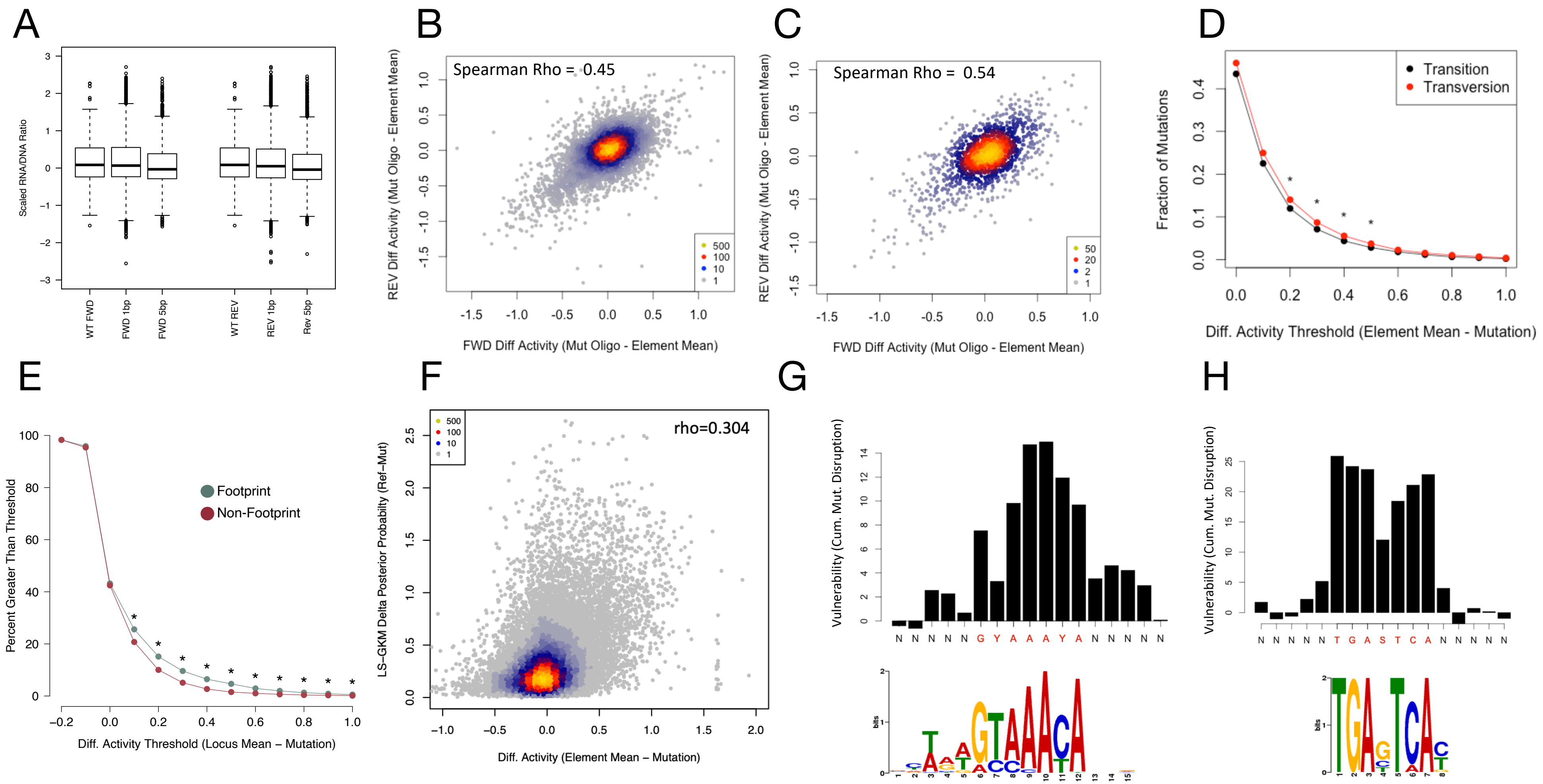


Figure S5. (A) Boxplots indicating the distribution of activity in single mutant loci was roughly equivalent to that of WT regions with a slight decrease (Wilcox $P<5x10^{-16}$) in activity observed in 5-bp mutations. The left group of boxes shows data from the forward strand and the right group shows data from the reverse strand oligonucleotides. (B-C) Scatter plots showing the correlation in the differential activity (mutant oligo - the locus mean) between the forward and reverse strands for single base pair mutations (B) and 5-bp mutations (C). The color indicates the number of points in contact with a given location on the plot. (D) Line plots showing the fraction of mutations that meet increasing thresholds of differential activity (loci mean - mutant oligo) for transitions vs. transversions. (*) indicates Fisher's P<0.05. (E) Line plot indicating the proportion of mutations imposing a change of activity at a variety of thresholds for a single base pair mutations. Green points indicate data for mutations fallen within DHS footprints. Red points indicate data for mutations falling outside of DHS footprints. (*) Indicates Fisher's P<0.05. (F) Scatter plot describing the correlation between LSGKM-predicted, DAP disruptions and differential activity of mutant oligos. Color indicates the number of points in contact with a given region of the plot. (G-H) Bar plots shown gthe cumulative differential activity (loci mean - mutation) across all positions in the FOXA (G) or AP1 (H) motifs. Activity correlates with the motif consistency at each position.

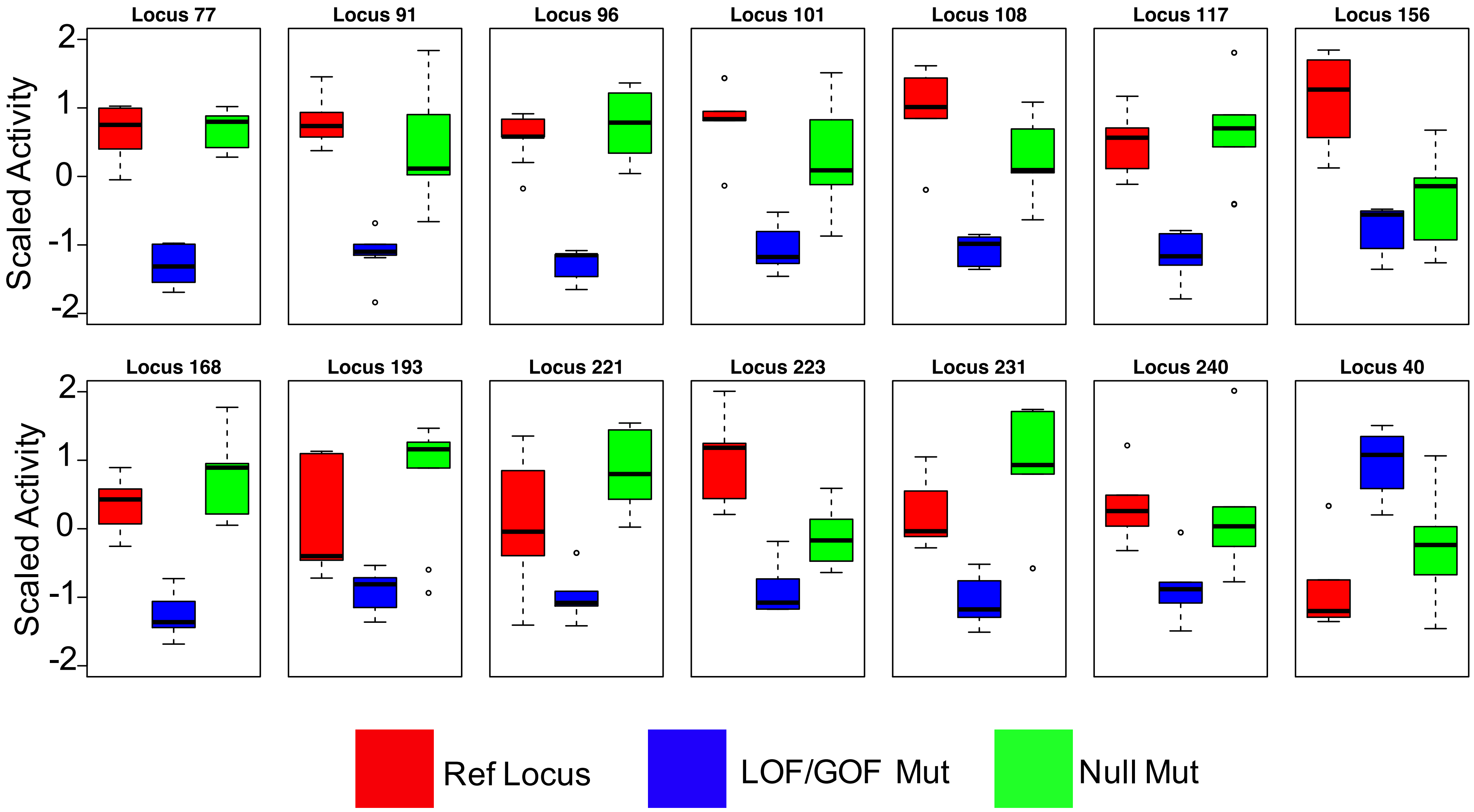


Figure S6. Validation data for the luc_mp_empty vector. Boxplots show inter-element z-scored data for reference (WT) oligos, oligos with loss of function (LOF) or gain of function (GOF) mutations, and oligos with predicted null mutations. Locus 40 is the only locus with a predicted GOF mutation. The remainder are predicted LOF.

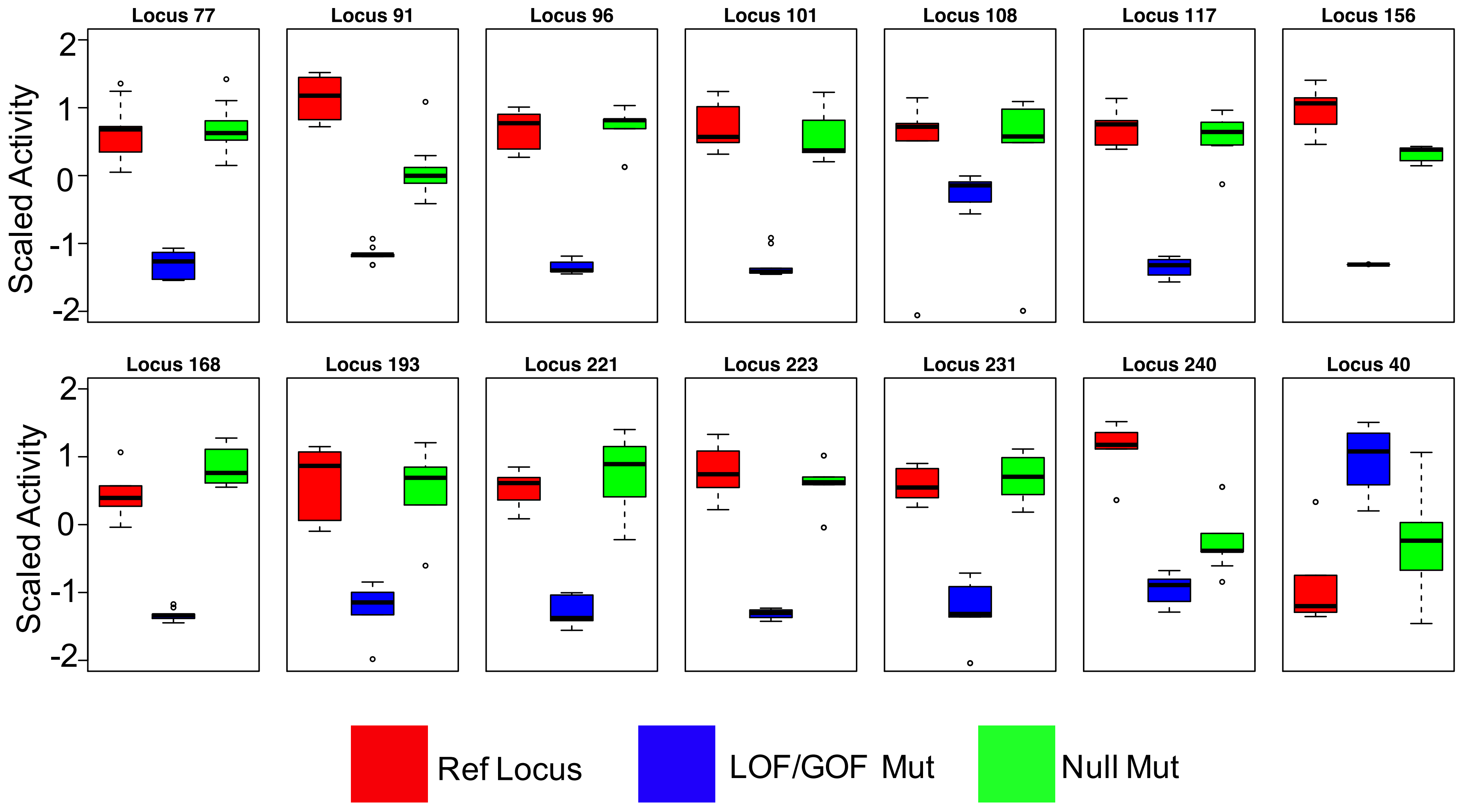
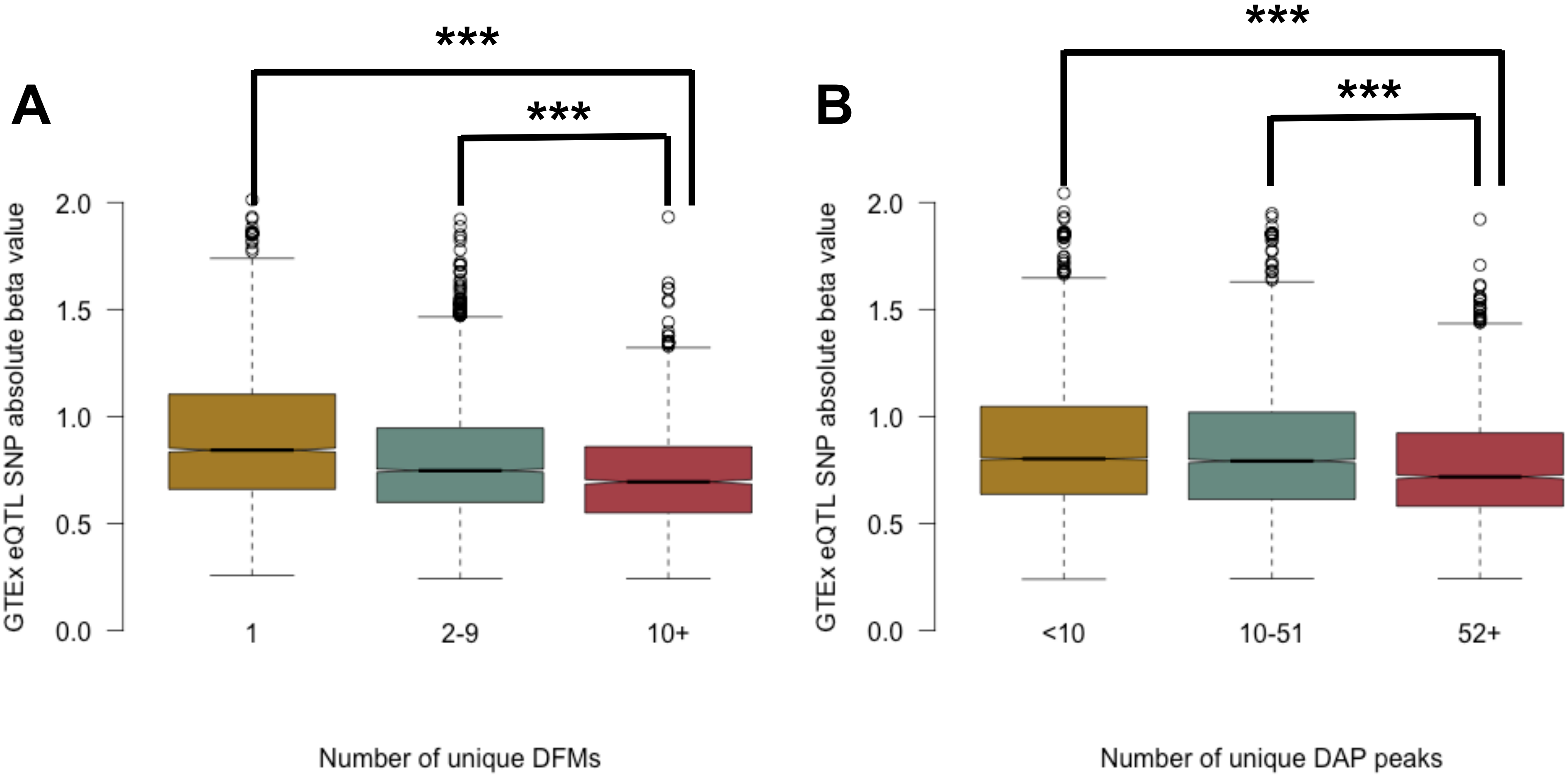


Figure S7. Validation data for the PGL4.23 vector. Boxplots showing inter-element z-scored data for reference oligos, oligos with loss of function (LOF) or gain of function (GOF) mutations, and oligos with predicted null mutations. Locus 40 is the only locus with a predicted GOF mutation. The remainder are predicted LOF.



Spearman rho -0.025, P=1.462e-06

Spearman rho -0.175, P<5e-16

Figure S8. 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 $P<5\times10^{-16}$

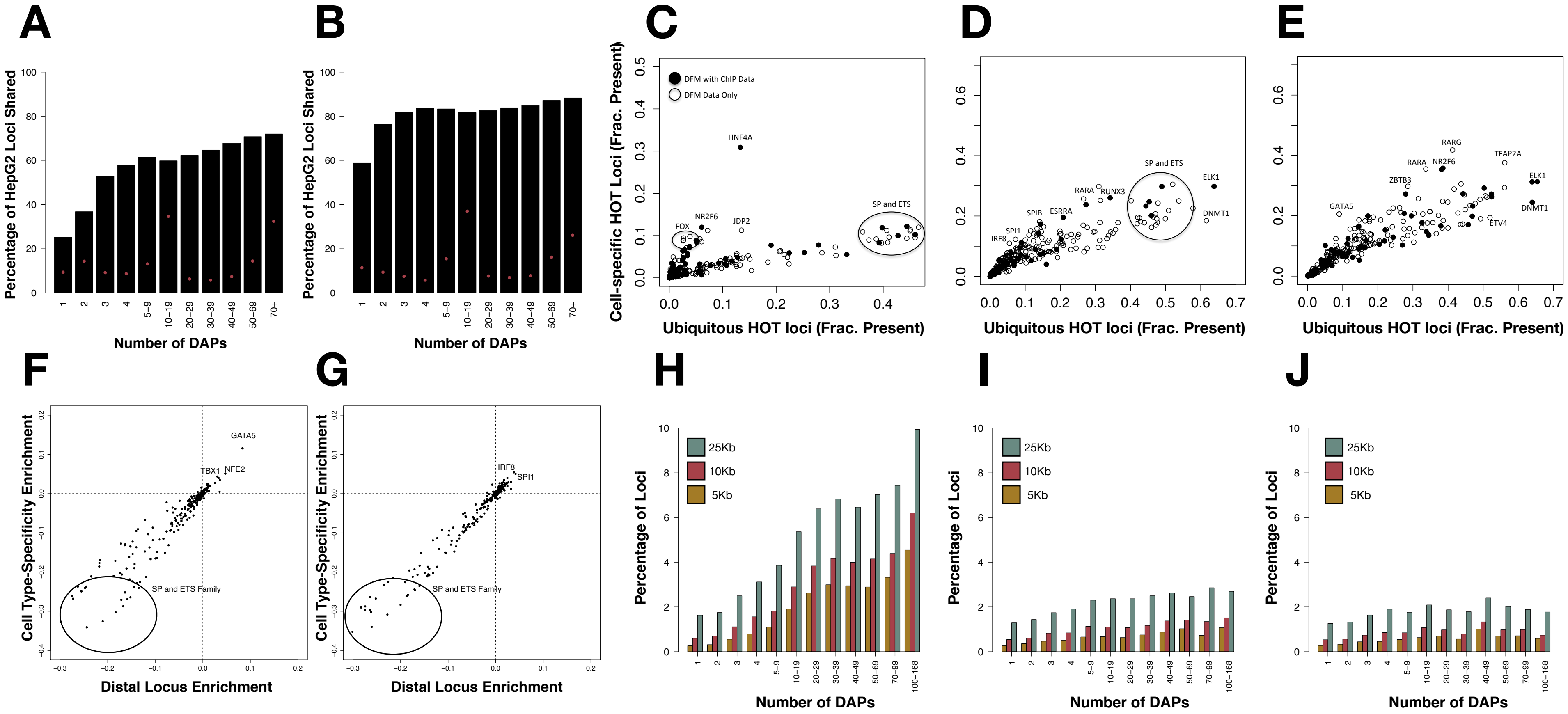
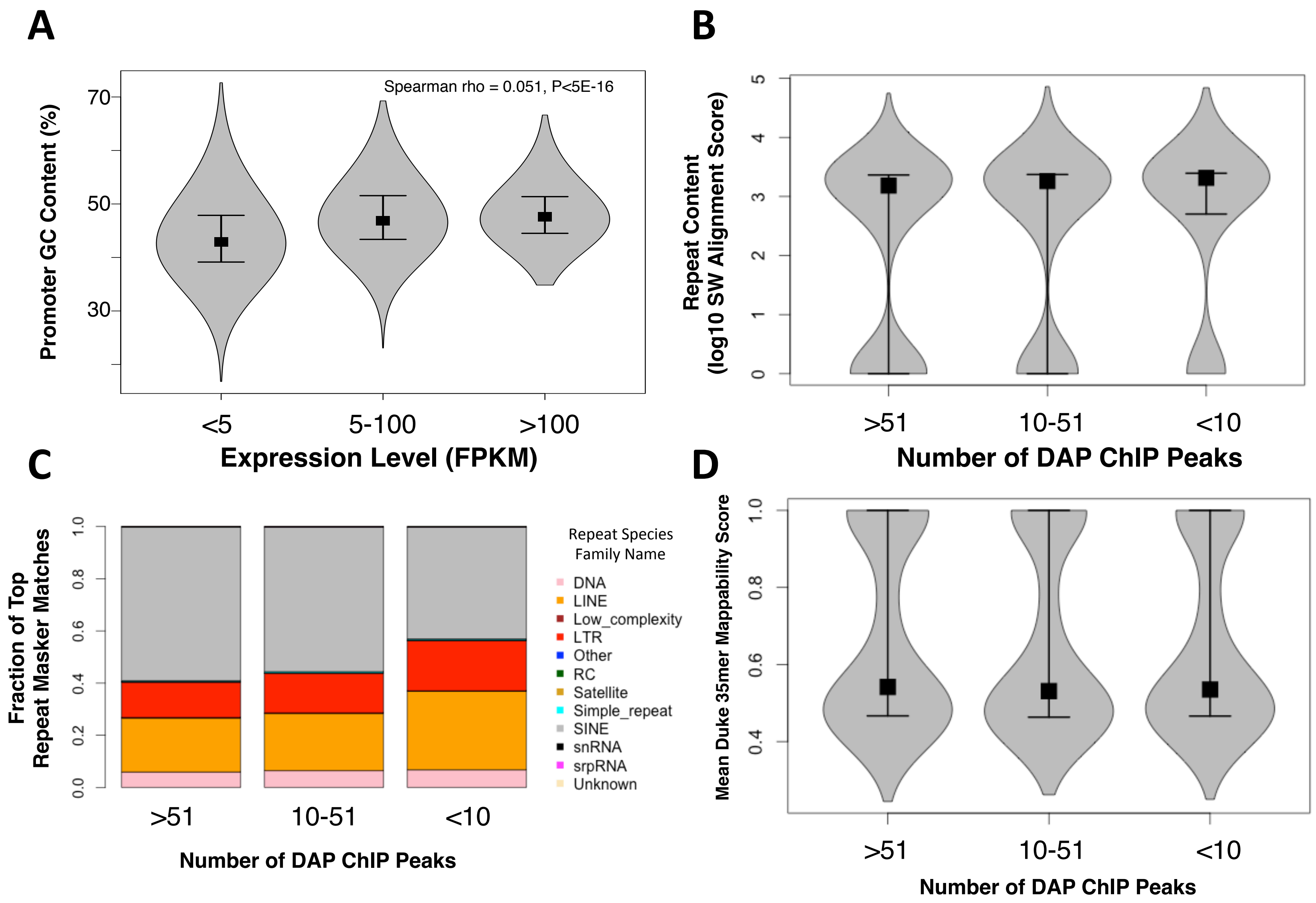


Figure S9. (A-B) The number of HepG2 loci bound by at least one DAP in GM12878 (A) and K562 (B). The red to indicates the number of loci shared by an equivalent number of DAPs in both cell lines. (C-E) Scatterplots demonstrating the fraction of cell type-specific and cell type ubiquitous HOT sites that contain a DFM for each ssTF in HepG2 (C) K562 (D) and GM12878 (E). Points in the top left of this plot indicate ssTFs enriched for cell type-specific HOT loci and points in the bottom right of this plot indicate ssTFs enriched for ubiquitous HOT loci. Filled black circles indicate ssTFs for which we have DFM and ChIP-seq data. Unfilled black circles indicate ssTs for which we only have DFM data. (F-G) Scatter plot demonstrating the association between cell type-specific, HOT loci enrichment and distal, HOT loci enrichment in K562 (F) and GM12878 (G). (H-J) Barplots indicating the fraction of each loci, stratified by the number of ChIP-defined DAP associations in HepG2, that are present near HepG2 (H), GM12878 (G), or K562 (H) specific gnees. 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 five in the cell line of interest



<u>Figure S10.</u> (A) Violin plot indicating the percent GC content at promoters of genes with specified expression levels. (B) Violin plot displaying the distribution of top RepeatMaster Smith-Waterman alignment scores for loci with differing numbers of DAP associations. (C) Stacked bar plots showing the family of repeat elements commonly found across loci. (D) Violin plots showing the distribution of Duke 35mer mappability uniqueness scores for each loci with differing numbers of DAP associations. Violin plot contains a square at the median value and whiskers at the 25th and 75th percentiles. Each violin contains data for 10000 randomly sampled loci meeting the “number of DAPs associated” criteria for each bin.

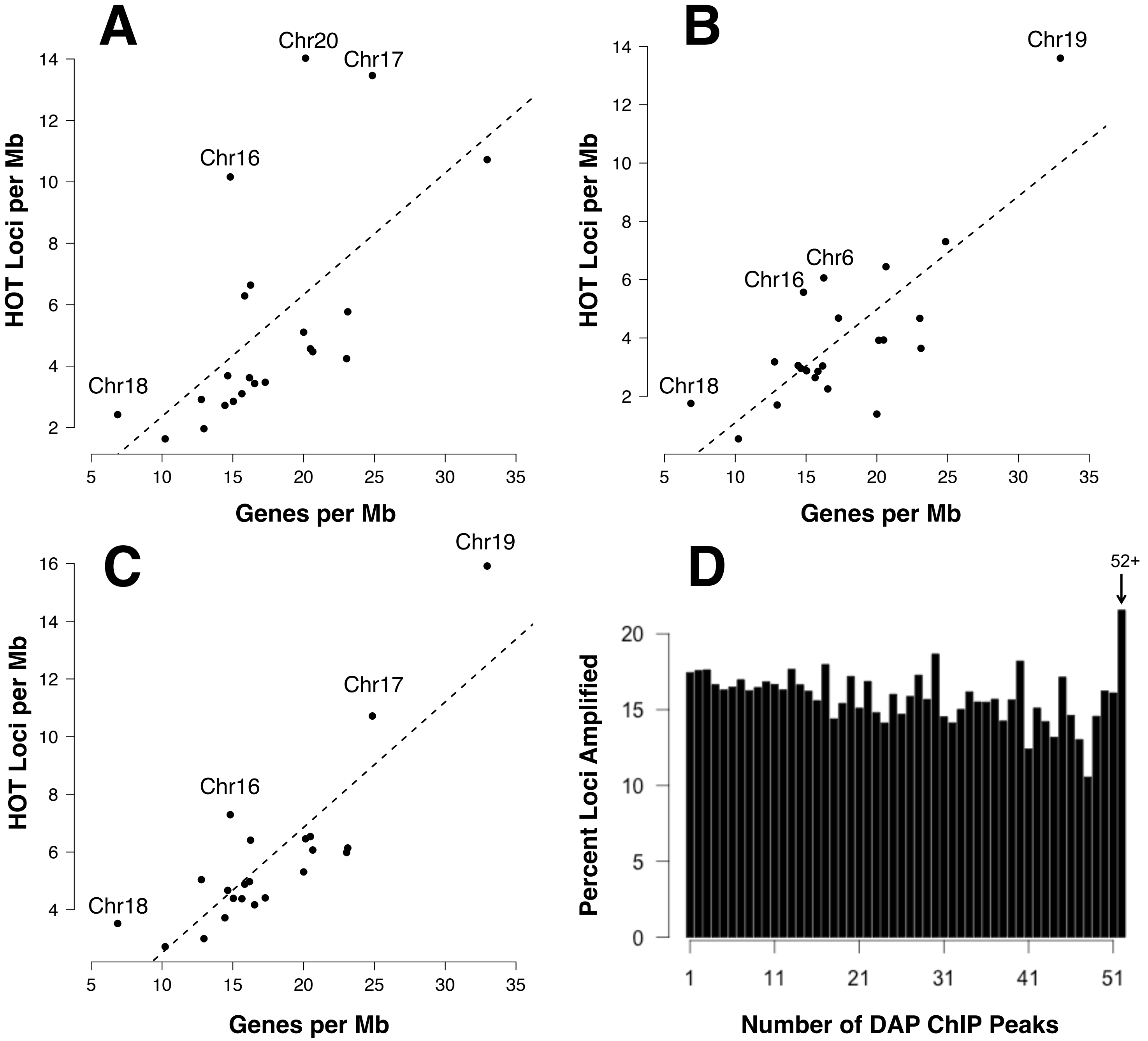


Figure S11. (A-C) Scatter plot indication the observed rate of HOT loci vs. the expected rate based on gene density in each HepG2 (A), K562 (B), or GM12878 (C) chromosome. (D) Bar plots showing the proportion of loci amplified in HepG2 at increasing numbers of ChIP-seq derived DAP associations

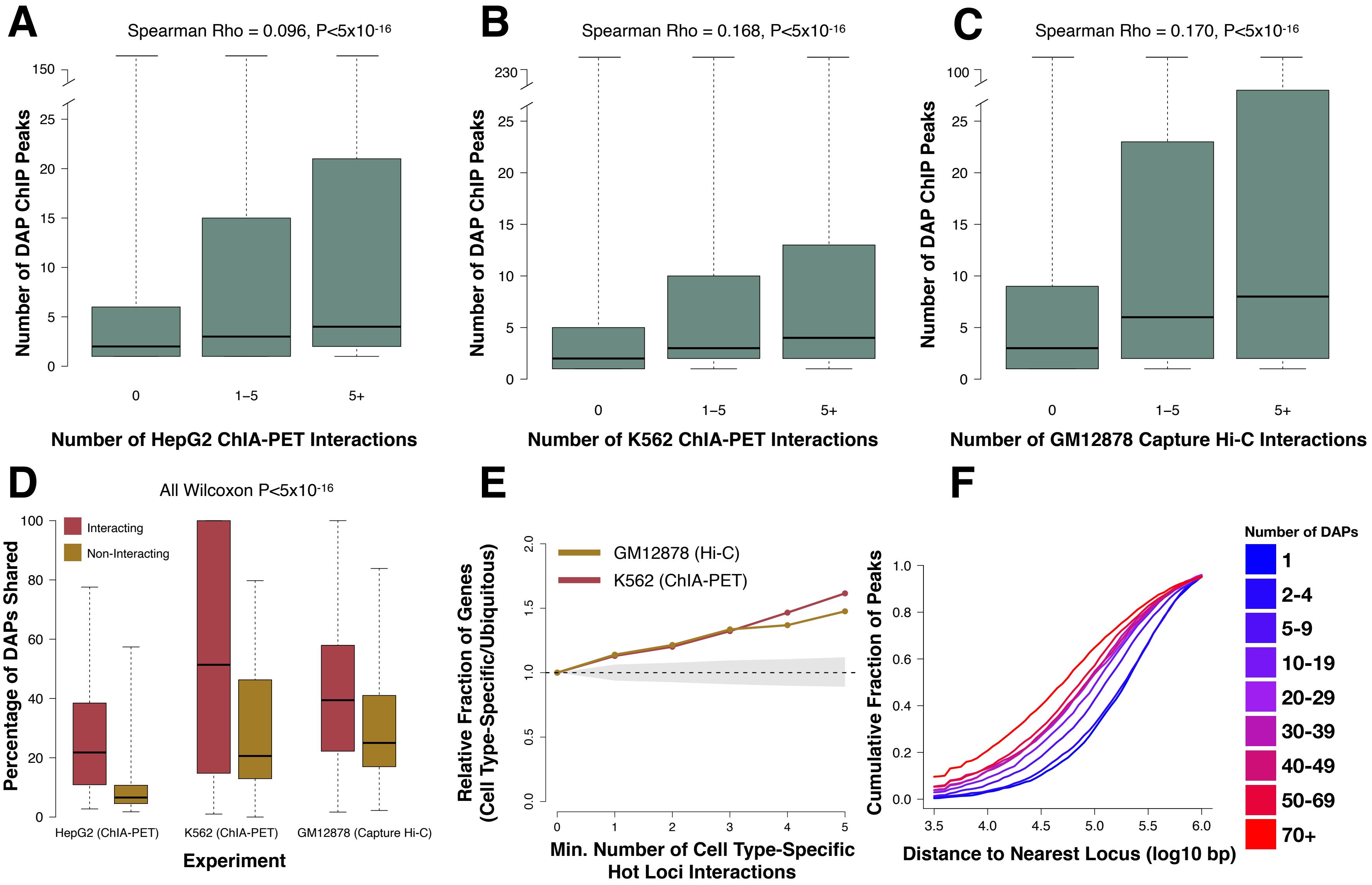


Figure S12. (A-B) Boxplots demonstrating the correlation between the number of DAPs bound (based on ChIP-seq peaks) and the number of ENCODE POLR2A ChIA-PET interactions observed in HepG2 (A) or K562 (B). The middle and high bins are divided at 5 interactions so roughly half of the non-zero loci are in each bin, in both cases. (C) Boxplots 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. P-values reported are derived from Spearman rho correlation of the entire dataset. (D) Boxplots demonstrating the fraction of DAPs in common between interacting loci matched non-interacting loci for HepG2 and K562 ChIA-PET data and GM12878 promoter capture Hi-C data. (E) Line plot indicating the relative fraction (cell type-specific/ubiquitously expressed) of gene promoters with at least the specified number of ChIA-PET interactions with other cell type-specific HOT loci. The gray shaded area represents the 95% confidence, null interval of randomly shuffled loci interactions between cell type-specific and ubiquitously expressed promoters. (F) Cumulative distribution functions indicating the cumulative fraction of loci that contain a neighboring locus with an equivalent number of DAP associations (as specified by the line color) within a given distance in base pairs.