Sequence characteristics and an accurate model of abundant hyperactive loci in the human genome

Abstract

Enhancers and promoters are classically considered to be bound by a small set of TFs in a sequence-specific manner. This assumption has come under increasing skepticism as the datasets of ChIP-seq assays of TFs have expanded. In particular, high-occupancy target (HOT) loci attract hundreds of TFs with seemingly no detectable correlation between ChIP-seq peaks and DNA-binding motif presence. Here, we used a set of 1,003 TF ChIP-seq datasets (HepG2, K562, H1) to analyze the patterns of ChIP-seq peak co-occurrence in combination with functional genomics datasets. We identified 43,891 HOT loci forming at the promoter (53%) and enhancer (47%) regions. HOT promoters regulate housekeeping genes, whereas HOT enhancers are involved in tissue-specific process regulation. HOT loci form the foundation of human super-enhancers and evolve under strong negative selection, with some of these loci being located in ultraconserved regions. Sequence-based classification analysis of HOT loci suggested that their formation is driven by the sequence features, and the density of mapped ChIP-seq peaks across TF-bound loci correlates with sequence features and the expression level of flanking genes. Based on the affinities to bind to promoters and enhancers we detected 5 distinct clusters of TFs that form the core of the HOT loci. We report an abundance of HOT loci in the human genome and a commitment of 51% of all TF ChIP-seq binding events to HOT locus formation thus challenging the classical model of enhancer activity and propose a model of HOT locus formation based on the existence of large transcriptional condensates.

Keywords: Computational genomics, epigenetics, gene regulation, transcriptional regulation, transcription factors

Introduction

Tissue-specificity of gene expression is orchestrated by the combination of transcription factors (TFs) that bind to regulatory regions such as promoters, enhancers, and silencers [1,2]. Classically, an enhancer is thought to be bound by a few TFs that recognize a specific DNA motif at their cognate TF binding site (TFBS) through its DNA-binding domain and recruit other molecules necessary for catalyzing the transcriptional machinery [3–5]. Based on the arrangements of the TFBSs, also called “motif grammar”, the architecture of enhancers is commonly categorized into “enhanceosome” and “billboard” models [6,7]. In the enhanceosome model, a rigid grammar of motifs facilitates the formation of a single structure comprising multiple TFs which then activates

the target gene [8,9]. This model requires the presence of all the participating proteins. Under the billboard model, on the other hand, the TFBSs are independent of each other and function in an additive manner [10]. However, as the catalogs of TF ChIP-seq assays have expanded thanks to the major collaborative projects such as ENCODE [11] and modENCODE [12], this assertion that the TFs interact with DNA through the strictly defined binding motifs has fallen under increasing contradiction with empirically observed patterns of DNA binding regions of TFs. In particular, there have been reported genomic regions that seemingly get bound by a large number of TFs with no apparent DNA sequence specificity. These genomic loci have been dubbed high-occupancy target (HOT) regions and were detected in multiple species [12–16].

Initially, these regions have been partially attributed to technical and statistical artifacts of the ChIP-seq protocol, resulting in a small list of blacklisted regions which are mostly located in unstructured DNA regions such as repetitive elements and low complexity regions [17,18]. These blacklisted regions have been later excluded from the analyses and they represent a small fraction of the mapped ChIP-seq peaks. In addition, various studies have proposed the idea that some DNA elements can serve as permissive TF binding platforms such as GC-rich promoters, CpG islands, R-loops, and G-quadruplexes [17,18]. Other studies have argued that these regions are highly consequential regions enriched in epigenetic signals of active regulatory elements such as histone modification regions and high chromatin accessibility [12,19,20].

Early studies of the subject have been limited in scope due to the small number of available TF ChIP-seq assays. There have been numerous studies in recent years with additional TFs across multiple cell lines. For instance, Partridge et al. [20] studied the HOT loci in the context of 208 chromatin-associated proteins. They observed that the composition of the chromatin-associated proteins differs depending on whether the HOT locus is located in an enhancer or promoter. Wreczycka et al. [18] performed a cross-species analysis of HOT loci in the promoters of highly expressed genes, and established that some of the HOT loci correspond to the “hyper-ChIPable” regions. Remarker et al. [19] conducted a comparative study of HOT regions in multiple cell lines and detected putative driver motifs at the core segments of the HOT loci.

In this study, we used the most up-to-date set of TF ChIP-seq assays available from the ENCODE project and incorporated functional genomics datasets such as 3D chromatin data (Hi-C), eQTLs, GWAS, and clinical disease variants to characterize and analyze the functional implications of the HOT loci. We report that the HOT loci are one of the prevalent modes of regulatory TF-DNA interactions; they represent active regulatory regions with distinct patterns of bound TFs manifested as clusters of promoter-specific, enhancer-specific, and chromatin-associated proteins. They are active during the embryonic stage and are enriched in disease-associated variants. Finally, we propose a model for the HOT regions based on the idea of the existence of large transcriptional condensates.

Results

HOT loci are one of the prevalent modes of TF-DNA interactions

To define and analyze the high-occupancy target (HOT) loci, we used the most up-to-date catalog of ChIP-seq datasets (n=1,003) of TFs obtained from the ENCODE Project assayed in HepG2, K562, and H1-hESC (H1) cells (545, 411, and 47 ChIP-seq assays respectively, see Methods for details). While the TFs are defined as sequence-specific DNA-binding proteins that control the transcription of genes, the currently available ChIP-seq datasets include the assays of many other types of transcription-related proteins such as cofactors, coactivators, histone acetyltransferases as well as RNA Polymerase 2 variants. Therefore we collectively call all of these proteins DNA-associated proteins (DAPs). Using the datasets of DAPs, we overlaid all of the ChIP-seq peaks and obtained the densities of DAP binding sites across the human genome using a non-overlapping sliding window of length 400 bp and considered a binding site to be present in a given window if 8 bp centered at the summit of a ChIP-seq peak as overlapping. Given that the analyzed three cell lines contain varying numbers of assayed DAPs, we binned the loci according to the number of overlapping DAPs in a logarithmic scale with 10 intervals and defined HOT loci as those that fall to the highest 4 bins, which translates to those which contain on average >18% of available DAPs for a given cell line (see Methods for a detailed description and justifications). This resulted in 25,928, 15,231, and 2,732 HOT loci in HepG2, K562, and H1 cells respectively. We applied our definition to the Roadmap Epigenomic ChIP-seq datasets and observed that the number of available ChIP-seq datasets significantly affects the resulting HOT loci. However, the HOT loci defined using the Roadmap Epigenomic datasets were almost entirely composed of subsets of the ENCODE-based HOT loci, comprising 50%, 62%, and 15% in HepG2, K562, and H1, respectively (Table S5). Importantly, we note that the distribution of the number of loci is not multimodal, but rather follows a uniform spectrum, and thus, this definition of HOT loci is ad-hoc (Fig 1A**,** Fig S1). Therefore, in addition to the dichotomous classification of HOT and non-HOT loci, we use all of the DAP-bound loci to extract the correlations with studied metrics with the number of bound DAPs when necessary. Throughout the study, we used the loci from the HepG2 cell line as the primary dataset for analyses and used the K562 and H1 datasets when the comparative analysis was necessary.

Although the HOT loci represent only 5% of all the DAP-bound loci in HepG2, they contain 51% of all mapped ChIP-seq peaks. The fraction of the ChIP-seq peaks of each DAP overlapping HOT loci varies from 0% to 91%, with an average of 65% (Fig 1B, y-axis). Among the DAPs that are present in the highest fraction of HOT loci are (Fig 1B, x-axis) SAP130, MAX, ARID4B, ZGPAT, HDAC1, MED1, TFAP4, and SOX6. The abundance of histone deacetylase-related factors mixed with transcriptional activators suggests that the regulatory functions of HOT loci are a complex interplay of activation and repression. RNA Polymerase 2 (POLR2) is present in 42% of HOT loci arguing for active transcription at or in the proximity of HOT loci (including mRNA and eRNA transcription). When the fraction of peaks of individual DAPs overlapping the

HOT loci are considered (Fig 1B, y-axis), DAPs with >90% overlap are GMEB2 (essential for replication of parvoviruses), ZHX3 (zinc finger transcriptional repressor), and YEATS2 (subunit of acetyltransferase complex). Whereas the DAPs that are least associated with HOT loci (<5%) are ZNF282 (transcriptional repressor), MAFK, EZH2 (histone methyltransferase), and TRIM22 (ubiquitin ligase). The fact that HOT loci harbor more than half of the ChIP-seq peaks suggests that the HOT loci are one of the prevalent modes of TF-DNA interactions rather than an exceptional case, as has been initially suggested by earlier studies [17,18].

Around half of the HOT loci (51%) are located in promoter regions (46% in primary promoters and 5% in alternative promoters), 25% in intronic regions, and only 24% are in intergenic regions with 9% being located >50 kbs away from promoters, suggesting that the HOT loci are mainly clustered in vicinities (promoters and introns) of transcription start sites and therefore potentially playing essential roles in the regulation of nearby genes (Fig 1C). When considering the non-promoter HOT loci, we observed that they were universally located in regions of H3K27ac or H3K4me1, indicating that they are active enhancers (Fig S2 A-D). When comparing the definitions of promoters and enhancers based on chromHMM states and ENCODE SCREEN annotations, the composition of HOT loci in relation to promoters and enhancers showed similar fractions (Fig S2E). Both HOT promoters and enhancers are almost entirely located in the chromatin-accessible regions (97% and 93% of the total sequence lengths, respectively, Fig 1D). We compared our definition of the HOT loci to those reported in Remaker et al. 2020 [19] and Boyle et al. 2014 [21]. We observed that because these two studies define HOT loci using 2 kb windows, they cover a larger fraction of the genome. Our set of HOT loci largely consisted of subsets of those defined in these two studies, with overlap percentages of 81%, 93%, and 100% in HepG2, K562, and H1, respectively (Fig S3). Further analysis revealed that our set of HOT loci primarily constitutes the "core" and more conserved (Fig S4) regions of HOT loci defined in the mentioned studies, while their composition in terms of promoter, intronic, and intergenic regions is similar (Fig S5), suggesting that the three definitions point to loci with similar characteristics.

To further dissect the composition of HOT enhancer loci, we compared them to super-enhancers as defined in the study by White et al. 2013 [22] and a set of regular enhancers (see Methods for definitions). Overall, 31% of HOT enhancers and 16% of HOT promoters are located in super-enhancers, while 97% of all HOT loci overlap H3K27ac or H3K4me1 regions (Fig 1E). While HOT enhancers and promoters appear to provide a critical foundation for super-enhancer formation, they represent only a small fraction of super-enhancer sequences overall accounting for 9% of combined super-enhancer length.

A 400 bp HOT locus, on average, harbors 125 DAP peaks in HepG2. However, the peaks of DAPs are not uniformly distributed across HOT loci. There are 68 DAPs with >80% of all of the peaks located in HOT loci (Fig 1B). To analyze the signatures of unique DAPs in HOT loci, we performed a PCA analysis using the overlapping DAP combinations for each HOT locus. This analysis showed that the principal component 1 (PC1) is correlated with the total number of distinct DAPs located at a given HOT locus (Fig S6A). PC2 separates the HOT promoters and HOT enhancers (Fig 2A, FigS6B), and the PC1-PC2 combination also separates the p300-bound HOT

loci (Fig S6C). This indicates that the HOT promoters and HOT enhancers must have distinct signatures of DAPs. To test if such signatures exist, we clustered the DAPs according to the fractions of HOT promoter and HOT enhancer loci that they overlap with. This analysis showed that there is a large cluster of DAPs (n=458) which on average overlap only 17% of HOT loci which are likely secondary to the HOT locus formation (Fig S7). We focused on the other, HOT-enriched, cluster of DAPs (n=87) which are present in 53% of HOT loci on average (Fig S7) and consist of four major clusters of DAPs (Fig 2D). *Cluster I* comprises 4 DAPs ZNF687, ARID4B, MAX, and SAP130 which are present in 75% of HOT loci on average. The three latter of these DAPs form a PPI interaction network (PPI enrichment p-value=0.001) (Fig S8A). We called this cluster of DAPs essential regulators given their widespread presence in both HOT enhancers and HOT promoters. *Cluster II* comprises 29 DAPs which are present in 47% of the HOT loci and are 1.7x more likely to overlap HOT promoters than HOT enhancers. Among these DAPs are POLR2 subunits, PHF8, GABP1, GATAD1, TAF1 etc. The strongest associated GO molecular function term with the DAPs of this cluster is *RNA Polymerase transcription factor initiation activity* suggestive of their direct role in transcriptional activity (FigS8B). *Cluster III* comprises 16 DAPs which are 1.9x more likely to be present in HOT enhancers than in HOT promoters. These are a wide variety of transcriptional regulators among which are those with high expression levels in liver NFIL3, NR2F6, and pioneer factors HNF4A, CEBPA, FOXA1, and FOXA2. The majority (13/16) of DAPs of this cluster form a PPI network (PPI enrichment p-value $< 10^{-16}$, Fig S8C). Among the strongest associated GO terms of biological processes are those related to cell differentiation (*white fat cell differentiation*, *endocrine pancreas development*, *dopaminergic neuron differentiation*, etc.) suggesting that *cluster III* HOT enhancers underlie cellular development. *Cluster IV* comprises 12 DAPs which are equally abundant in both HOT enhancers and HOT promoters (64% and 63% respectively), which form a PPI network (PPI enrichment p-value $< 6x10^{-16}$, Fig S8D) with HDAC1 (Histone deacetylase 1) being the node with the highest degree, suggesting that the DAPs of the cluster may be involved in chromatin-based transcriptional repression. Lastly, *Cluster V* comprises 26 DAPs of a wide range of transcriptional regulators, with a 1.3x skew towards the HOT enhancers. While this cluster contains prominent TFs such as TCF7L2, FOXA3, SOX6, FOSL2, etc., the variety of the pathways and interactions they partake in makes it difficult to ascertain the functional patterns from the constituent of DAPs alone. Although this clustering analysis reveals subsets of DAPs that are specific to either HOT enhancers or HOT promoters (clusters II and III), it still does not explain what sorts of interplays take place between these recipes of HOT promoters and HOT enhancers, as well as with the other clusters of DAPs with equal abundance in both the HOT promoters and HOT enhancers. To test the significance of the PPI networks described above, we ran 100 trials for each cluster by randomly selecting an equal number of DAPs reported in PPI networks and calculated the significance of the PPI enrichment p-values. All of the reported PPI enrichment p-values were significantly higher than the randomized trials (p-value < 0.01, one-sample t-test).

Notably, PC4 separates HOT loci associated with CTCF (Fig 2C) and Cohesin (Fig S6D). This clear separation of CTCF- and Cohesin-bound HOTs is surprising, given that only relatively

small fractions of their peaks (21% and 38% respectively) reside in HOT loci, and present in 36% of the HOT loci, compared to some other DAPs with much higher presence described above, that do not get separated clearly by the PCA analysis. Furthermore, CTCF- and Cohesin-bound HOT enhancer loci are located significantly closer (p-value<10^{-100}; Mann-Whitney U Test) to the nearest genes (Fig S9B), making it more likely that those loci are proximal enhancers. And the total number of overlapping DAPs is significantly higher (p-value< 10^{-100}; Mann-Whitney U Test) in CTCF- and Cohesin-bound loci compared to the rest of the HOT loci (Fig S9C), suggesting that at least a portion of the number of DAPs in HOT loci can be explained by 3D chromatin contacts between the genomic regions mediated by CTCF-Cohesin complex.

To comprehensively quantify the 3D chromatin interactions involving the HOT loci, we used Hi-C data with 5 kbs resolution [23] (see Methods). First, we obtained statistically significant chromatin interactions using FitHiChIP tool [24] (see Methods) and observed that HOT loci are enriched in chromatin interactions and 1.66x more likely to engage in chromatin interactions than the regular enhancers (p-value< 10^{-20}, Chi-square test). When all of the DAP-bound loci are considered, the number of chromatin interactions positively correlates with the number of bound DAPs (rho=0.3, p-value<10^{-100}, Spearman correlation). Next, we overlayed the chromatin interactions with the loci binned by the number of bound DAPs. We observed that the loci with high numbers of bound DAPs are more likely to engage in chromatin interactions with other loci harboring large numbers of DAPs, i.e. the HOT loci have the propensity to connect through long-range chromatin interactions with other HOT loci (Fig 3A). To further validate this observation, we obtained frequently interacting regions (FIREs) [25], and observed that the FIREs are 2.89x (p-value<10^{-230}, Chi-square test) enriched HOT loci compared to the regular enhancers (see Methods). Moreover, 66% of HOT loci are located in TAD regions and 21% are located in chromatin loops. In particular, the HOT loci are 2.97x (p-value< 10^{-230}, Mann-Whitney U test) enriched in the chromatin loop anchor regions (11% of the HOT loci) compared to regular enhancers. To investigate further, we analyzed the loop anchor regions harboring HOT loci and observed that the number of multi-way contacts on loop anchors correlates with the number of bound DAPs (rho=0.84 p-value< 10^{-4}; Pearson correlation). The number of multi-way interactions in loop anchor regions varies between 1 and 6, with only one locus, in an extreme case, serving as an anchor for 6 overlapping loops on chromosome 2 (Fig 3B). That locus contains one HOT enhancer harboring 101 DAPs, located 23 kbs away from the *LINC02583* gene, which is linked to liver-specific GWAS traits *hematocrit*, *left ventricular diastolic function,* and *high-density lipoprotein cholesterol*, highlighting the important role of the HOT locus. Of the loop anchor regions with >3 overlapping loops, 51% contained at least one HOT locus, suggesting an interplay between chromatin loops and HOT loci (Fig 3B). Overall, 94% of HOT loci are located in regions with at least one chromatin interaction. This observation is consistent with previous reports that much of the long-range 3D chromatin contacts form through the interactions of large protein complexes [26]. While there is a correlation between the HOT loci and chromatin interactions, the causal relation between these two properties of genomic loci is not clear.

A set of DAPs stabilizes the interactions of DAPs at HOT loci

Next, using the ChIP-seq signal intensity as a proxy for the DAP binding affinity, we sought to analyze the patterns of binding affinities of DAP-DNA of interactions in HOT loci. We observed that the overall binding affinities of DAPs correlate with the total number of colocalizing DAPs (Fig 4A, rho=0.97, p-value<10^{-10}; Spearman correlation). Moreover, even when calculated DAP-wise, the average of the overall signal strength of every DAP correlates with the fraction of HOT loci that the given DAP overlaps with (rho=0.6, p-value<10^{-29}; Spearman correlation. Figure 4B), meaning that the overall average value of the signal intensity of a given DAP is largely driven by the ChIP-seq peaks which are located in HOT loci.

While the overall average of the ChIP-seq signal intensity in HOT loci is greater when compared to the rest of the DAP-bound loci, individual DAPs demonstrate different levels of involvement in HOT loci. When sorted by the ratio of the signal intensities in HOT vs. non-HOT loci, among those with the highest HOT-affinities are GATAD1, MAX, NONO as well as POLR2G and Mediator subunit MED1 (Fig 4B,C). Whereas those with the opposite affinity (i.e. those that have the strongest binding sites in non-HOT loci) are REST, RFX5, TP53, etc (Fig 4B,C). By analyzing the signal strengths of DAPs jointly, we observed that a host of DAPs likely has a stabilizing effect on the binding of DAPs in that, when present, the signal strengths of the majority of DAPs are on average 1.7x greater. These DAPs are CREB1, RFX1, ZNF687, RAD51, ZBTB40 and GPBP1L1.

So far, we have treated the DAPs under a single category and did not make a distinction based on their known DNA-binding properties. Previous studies have discussed the idea that sequence-specific DAPs (ssDAPs) can serve as anchors, similar to the pioneer TFs, which could facilitate the formation of HOT loci [19,20,27]. We asked if ssDAPs yield greater signal strength values than non-sequence-specific DAPs (nssDAPs). To test this hypothesis, we classified the DAPs into those two categories using the definitions provided in the study (Lambert et al. 2018) [28] and categorized the ChIP-seq signal values into these two groups. While statistically significant (p-value< 0.001, Mann-Whitney U test), the differences in the average signals of ssDAPs and nssDAPs in both HOT enhancers and HOT promoters are small (Fig 4D). Moreover, while the average signal values of ssDAPs in HOT enhancers are greater than that of the nssDAPs, in HOT promoters this relation is reversed. At the same time, the average signal strength of the DAPs is 3x greater than the average signal strength of H3K27ac peaks in HOT loci. Based on this, we concluded that the ChIP-seq signal intensities do not seem to be a function of the DNA-binding properties of the DAPs.

Sequence features that drive the accumulation of DAPs

We next analyzed the sequence features of the HOT loci. For this purpose, we first addressed the evolutionary conservation of the HOT loci using phastCons scores generated using an alignment of 46 vertebrate species [29]. The average conservation scores of the DAP-bound loci are in strong correlation with the number of bound DAPs (rho=0.98, p-value<10^{-130}; Spearman correlation) indicating that the negative selection exerted on HOT loci are proportional to the

number of bound DAPs (Fig 5A). With 120 DAPs per locus on average, these HOT regions are 1.7x more conserved than the regular enhancers in HepG2 (Fig 5B). We observed a similar trend of conservation levels when the phastCons scores generated from primates and placental mammals and primates were considered, the HOT loci being 1.45x and x1.1 more conserved than the regular enhancers, respectively (Fig S10). In addition, we observed that the HOT loci of all three cell lines (HepG2, K562, and H1) overlap with 22 ultraconserved regions, among which are the promoter regions of 11 genes including SP5, SOX5, AUTS2, PBX1, ZFPM2, ARID1A, OLA1 and the enhancer regions of (within <50kbs of their TSS) 5S rRNA, MIR563, SOX21, etc. (full list in Table S4). Among them are those which have been linked to diseases and other phenotypes. For example, DNAJC1 [30] and OLA1 (which interacts with BRCA1) have been linked to breast cancer in cancer GWAS studies [31]. Whereas AUTS2 [32] and SOX5 [33] have been linked to predisposition to neurological conditions such as Autism spectrum disorder, intellectual disability, and neurodevelopmental disorder. Of these genes, ARID1A, AUTS2, DNAJC1, OLA1, SOX5, and ZFPM2 have been reported to have strong activities in the Allen Mouse Brain Atlas [34].

CpG islands have been postulated to serve as permissive TF binding platforms [35,36] and this has been listed as one of the possible reasons for the existence of HOT loci in a previous study [18]. To test this hypothesis, we extracted the overlap rates of all DAP-bound loci with CpG islands (see Methods). While the overall fraction of loci that overlap CpG islands correlates strongly with the number of bound DAPs (rho=0.7, p-value=0.001; Pearson correlation), only 12% of HOT enhancers overlapped CpG island whereas, for the HOT promoters, this fraction was 83%, suggesting that CpG islands alone do not explain HOT enhancer loci despite accounting for the majority of HOT promoters loci (Fig S11A). Similarly, the average GC content is strongly correlated with the number of bound DAPs (rho=0.89, p-value<10^{-4}; Pearson correlation, Fig S11B), with the average GC content of 64% and 51% in HOT promoters and HOT enhancers respectively (p-value<10^{-100}, Mann-Whitney U test), in both HepG2 and K562.

In addition, we observed that the average content of repeat elements in the loci strongly and negatively correlates with the number of bound DAPs across the cell lines (rho=-0.9, p-value=<10^{-5}; Pearson, Fig S11C), which is likely the result of the fact that the HOTs are under elevated negative selection and reject insertion of repetitive DNA.

Other genomic sequence features that have been considered in the context of HOT loci in previous studies include and are not limited to G-quadruplex, R-loops, methylation patterns, etc., which have concluded that each of them can partially explain the phenomenon of the HOT loci [13,17,18]. Still, one of the central questions remains whether the HOT loci are driven by sequence features or they are the result of cellular biology not strictly related to the sequences, such as the proximal accumulation of DAPs in foci due to the biochemical properties of accumulated molecules, or other epigenetic mechanisms.

To address this question with a broader approach, we asked whether the HOT loci can be accurately predicted based on their DNA sequences alone, and sequence features, including GC, CpG, GpC contents, and CpG island coverage. For sequence-based classification, we trained a Convolutional Neural Network (CNN) model using one-hot encoded sequences and an SVM

classifier trained on gapped k-mers (gkmSVM) [37] (Supplemental Methods 1.6.2, Figure S12). For feature-based classification, we trained logistic regression (LogReg) classifiers and SVM classifiers with linear, polynomial, radial basis function, and sigmoid kernels (See Supplemental Methods 1.6.2.2 for detailed analysis). We carried out the classification experiments using the following control sets: a) randomly selected loci from merged DNaseI Hypersensitivity Sites (DHS) of cell lines in the Roadmap Epigenomics Project, b) promoter regions, and c) regular enhancers (See Supplemental Methods 6.1.1 for definitions).

Using the sequence features, we trained separate models using each of the features in addition to one with all of the features combined. We observed that, when averaged across all the methods, GC content value possesses the highest amount of discrimination power (auROC: 0.73), followed by the combination of all features (auROC: 0.70) (Figure S13A,B). When compared across the classification methods, LogReg and SVM with linear kernel outperformed the other non-linear kernels by 20%, suggesting that the features possess linearly combined or largely overlapping effects in encoding the information in HOT loci (Figure S13A).

When classified using the sequences directly, CNN yielded the highest performance with auROC of 0.91, while for the gkmSVM it was 0.86 (both averaged over cell lines and control sets), suggesting that CNNs capture the motif grammar of the HOT loci better than gapped k-mers (Figure S13C). When the two classification schemes (sequence- and feature-based) are compared, CNNs outperformed the LogReg and linear SVMs by a factor of 1.3x (or 17%), suggesting that there is additional information that is highly relevant to the DNA-DAP interaction density encoded in the DNA sequences, in addition to the GC, CpG, GpC (Fig 5C). This is in line with the observation mentioned above, that 88% of the HOT enhancers do not overlap with annotated CpG islands. This analysis concluded that the mechanisms of HOT locus formation are likely encoded in their DNA sequences.

Of the control regions tested, we observed that CNNs can discriminate with the auROC values of 0.91, 0.89, and 0.87 for DHS, promoters, and regular enhancers respectively (Fig 5C). This observation reflects the fact that HOT loci are themselves located in enhancers and promoters, albeit representing fairly separable and distinct subsets of them.

Extending the input regions from 400 bp to 1 kbs for sequence-based classification did not lead to a significant increase in performance, suggesting that the core 400 bp regions contain most of the information associated with DAP density (Fig S14).

Highly expressed housekeeping genes are commonly regulated by HOT promoters

After characterizing the HOT loci in terms of the DAP composition and sequence features, we sought to analyze the cellular processes they partake in. HOT loci were previously linked to highly expressed genes [18]. In both inspected differentiated cell lines (HepG2 and K562), the number of DAPs positively correlates with the expression level of their target gene (enhancers were assigned to their nearest genes for this analysis; rho=0.56, p-value$<10^{-10}$; Spearman correlation; Fig 15A). In HepG2, the average expression level of the target genes of promoters

with at least one DAP bound is 1.7x higher than that of the target genes of enhancers with at least one DAP bound, whereas when only HOT loci are considered this fold-increase becomes 4.7x. This suggests that the number of bound DAPs of the HOT locus has a direct impact on the level of the target gene expression. Moreover, highly expressed genes (RPKM>50) were 4x more likely to have multiple HOT loci within the 50 kbs of their TSSs than the genes with RPKM<5 (p-value<10^{-12}, chi-square test). In addition, the average distance between HOT enhancer loci and the nearest gene is 4.5x smaller than with the regular enhancers (p-value<10^{-30}, Mann-Whitney U test). Generally, we observed that the distances between the HOT enhancers and the nearest genes are negatively correlated with the number of bound DAPs (rho=-0.9; p-value<10^{-6}; Pearson correlation. Fig S15B), suggesting that the increasing number of bound DAPs makes the regulatory region more likely to be the TSS-proximal regulatory region.

To further analyze the distinction in involved biological functions between the HOT promoters and enhancers, we compared the fraction of housekeeping (HK) genes that they regulate, using the list of HK genes reported by (Hounkpe et al. 2021) [38]. According to this definition, 64% of HK genes are regulated by a HOT promoter and only 30% are regulated by regular promoters (Fig 6A). The HOT enhancers, on the other hand, flank 21% of the HK genes, which is less than the percentage of HK genes flanked by regular enhancers (38%). For comparison, 22% of the flanking genes of super-enhancers constitute HK genes. The involvement of HOT promoters in the regulation of HK genes is also confirmed in terms of the fraction of loci flanking the HK genes, namely, 21% of the HOT promoters regulate 64% of the HK genes. This fraction is much smaller (<9% on average) for the rest of the mentioned categories of loci (HOT and regular enhancers, regular promoters, and super-enhancers, Fig 6A).

We then asked whether the tissue-specificities of the expression levels of target genes of the HOT loci reflect their involvement in the regulation of HK genes. For this purpose, we used the *tau* metric as reported by (Palmer et al. 2021) [39], where a high *tau* score (between 0 and 1) indicates a tissue-specific expression of a gene, whereas a low *tau* score means that the transcript is expressed stably across tissues. We observed that the average *tau* scores of target genes of HOT enhancers are significantly but by a small margin greater than the regular enhancers (0.66 and 0.63, respectively. p-value<10^{-18}, Mann-Whitney U test), with super-enhancers being equal to regular enhancers (0.63). The difference in the average *tau* scores of the HOT and regular promoters is stark (0.57 and 0.74 respectively, p-value<10^{-100}, Mann-Whitney U test), representing a 23% increase (Fig 6B). Combined with the involvement in the regulation of HK genes, average *tau* scores suggest that the HOT promoters are more ubiquitous than the regular promoters whereas HOT enhancers are more tissue-specific than the regular and super-enhancers. Further supporting this, the GO enrichment analysis showed that the GO terms associated with the set of genes regulated by HOT promoters are basic HK cellular functions (such as *RNA processing*, *RNA metabolism*, *ribosome biogenesis,* etc.), whereas HOT enhancers are enriched in GO terms of cellular response to the environment and liver-specific processes (such as *response to insulin, oxidative stress, epidermal growth factors,* etc.) (Fig 6C).

A core set of HOT loci is active during development which expands after differentiation

Having observed that the HOT loci are active regions in many other human cell types, we asked if the observations made on the HOT loci of differentiated cell lines also hold true in the embryonic stage. To that end, we analyzed the HOT loci in H1 cells. It is important to note that the number of available DAPs in H1 cells is significantly smaller (n=47) than in HepG2 and K562, due to a much smaller size of the ChIP-seq dataset generated in H1. Therefore, the criterion of having >17% of available DAPs yields n>15 DAPs for the H1, as opposed to 77 and 55 for HepG2 and K562, respectively. However, many of the features of the loci that we've analyzed so far demonstrated similar patterns (GC contents, target gene expressions, ChIP-seq signal values etc.) when compared to the DAP-bound loci in HepG2 and K562, suggesting that albeit limited, the distribution of the DAPs in H1 likely reflects the true distribution of HOT loci. To alleviate the difference in available DAPs, in addition to comparing the HOT loci defined using the complete set of DAPs, we also (a) applied the HOT classification routing using a set of DAPs (n=30) available in all three cell lines (b) randomly subselected DAPs in HepG2 and K562 to match the number of DAPs in H1.

We observed that, when the complete set of DAPs is used, 85% of the HOT loci of H1 are also HOT loci in either of the other two differentiated cell lines (Fig 7A). However, only <10% of the HOT loci of the two differentiated cell lines overlapped with H1 HOT loci, suggesting that the majority of the HOT loci are acquired after the differentiation. A similar overlap ratio was observed based on DAPs common to all three cell lines (Fig 7B), where 68% of H1 HOT loci overlapped with that of the differentiated cell lines. These overlap levels were much higher than the randomly selected DAPs matching the H1 set (30%, Fig 7C).

Average evolutionary conservation scores (phastCons) of the developmental HOT loci are 1.3x higher than K562 and HepG2 HOT loci (p-value<10^{-10}, Mann-Whitney U test, Fig 7D). It is conceivable to hypothesize that the embryonic HOT loci are located mainly in regions with higher conservation regions, and more regulatory regions emerge as HOT loci after the differentiation. Some of these tissue-specific HOT loci could be those that are acquired more recently (compared to the H1 HOT loci), as it is known that the enhancers are often subject to higher rates of evolutionary turnover than the promoters [40].

GO enrichment analysis showed that H1 HOT promoters, similarly to the other cell lines, regulate the basic housekeeping processes (Fig S16) while the HOT enhancers regulate responses to environmental stimuli and processes active during the embryonic stage such as *TORC1 signaling* and *beta-catenin-TCF assembly*. This suggests that the main processes that the HOT promoters are involved in during the development remain relatively unchanged after the differentiation (in terms of associated GO terms, and due to being the same loci as the HOT promoters in differentiated cell lines), whereas the scope of the cellular activities regulated by HOT enhancers gets expanded after differentiation to be more exclusively tissue-specific.

HOT loci are enriched in causal variants

After establishing the expression and tissue-specificities of the HOT loci, we next analyzed the polymorphic variability in HOT loci and whether these loci are enriched in phenotypically causal variants. First, we analyzed the density of common variants extracted from the gnomAD database [41] (filtered with MAF>5%). We observed that HOT enhancers and HOT promoters are depleted in INDELs (4.7 and 4.1 variants per 1 kbs, respectively), compared to the regular enhancers and regular promoters (5.5 and 6.2 variants per 1 kbs, p-value$<10^{-4}$ and $<10^{-100}$, respectively, Mann-Whitney U test; Fig 8A). Contradicting the pattern of conservation scores described above, the distribution of common SNPs is elevated in HOT enhancers and HOT promoters compared to regular enhancers and regular promoters (1.14x and 1.07x fold-enrichment, p-values $<10^{-20}$ and $<10^{-100}$, respectively, Mann-Whitney U test; Fig 8B). This elevation of common variants in HOT loci, despite being located in conserved loci has been reported in a previous study in which the binding motifs of TFs were observed to colocalize in regions where the density of common variants was higher than average [42].

The eQTLs, on the other hand, are 2.0x enriched in HOT promoters compared to the regular promoters (p-value$<10^{-21}$, Mann-Whitney U test), while HOT enhancers are only moderately enriched in eQTLs compared to the regular enhancers (1.15x, p-value>0.05, Mann-Whitney U test; Fig 8C). eQTL enrichment in HOT promoters and regular promoters (compared to HOT and regular enhancers respectively) is in line with the known characteristics of the eQTL dataset, that the eQTLs most commonly reflect TSS-proximal gene-variant relationships, and therefore are enriched in promoter regions since the TSS-distal eQTLs are hard to detect due to the burden of multiple tests [43].

Unlike the eQTL analysis, we observed that the chromatin accessibility QTLs (caQTLs) are dramatically enriched in the overall enhancer regions (HOT and regular) compared to the promoters (HOT and regular) (4.1x, p-value$<10^{-100}$; Mann-Whitney U test, Fig 8D). This observation confirms the findings of the study which reported the caQTL dataset in HepG2 cells [44], which reported that the likely causal caQTLs are predominantly the variants disrupting the binding motifs of liver-expressed TFs enriched in liver enhancers. However, within the promoters regions, the HOT promoters are 3.0x enriched in caQTLs compared to the regular promoters (p-value=0.001; Mann-Whitney U test), whereas the fold enrichment in HOT enhancers is insignificant (1.2x, p-value=0.22, Mann-Whitney U test).

A similar enrichment pattern displays the reporter array QTLs (raQTLs [45]), with respect to the overall (HOT and regular) promoter and enhancer regions, with 3.3x enrichment in enhancers (p-value$<10^{-10}$, Mann-Whitney U test, Fig 8E). But, within-promoters and within-enhancers enrichments show that the enrichment in HOT promoters is more pronounced than the HOT enhancers (3.6x and 1.8x, p-values<0.01 and $<10^{-11}$, respectively, Mann-Whitney U test). The enrichment of the raQTLs in enhancers over the promoters likely reflects the fact that the SNP-containing loci are first filtered for raQTL detection according to their capacities to function as enhancers in the reporter array [45].

Combined, all three QTL datasets show a pronounced enrichment in HOT promoters compared to the regular promoters, whereas only the raQTLs show significant enrichment in HOT enhancers. This suggests that the individual DAP ChIP-seq peaks in HOT promoters are more likely to have consequential effects on promoter activity if altered, while HOT enhancers are less susceptible to mutations. Additionally, it is noteworthy that only the raQTLs are the causal variants, whereas e/caQTLs are correlative quantities subject to the effects of LD.

Finally, we used the GWAS SNPs combined with the LD SNPs (r2>0.8) and observed that the HOT promoters are significantly enriched in GWAS variants (1.8x, p-value<10^{-100}) whereas the HOT enhancers show no significant enrichment over regular enhancers (p-value>0.1, Mann-Whitney U test) (Fig 8F). We then calculated the fold-enrichment levels GWAS traits SNPs using the combined DHS regions of Roadmap Epigenome cell lines as a background (see Methods). Filtering the traits with significant enrichment in HOT loci (p-value<10^{-3}, Binomial test, Bonferroni corrected, see Methods) left 7 traits, of which all are definitively related to the liver functions (Fig 8G). Of the seven traits, only one (*Blood protein level*) was significantly enriched in regular promoters. While the regular enhancers are enriched in most of the (6 of 7) traits, the overall enrichment values in HOT enhancers are 1.3x greater compared to the regular enhancers. The fold-increase is even greater (1.5x) between the HOT and DHS regions. When the enrichment significance levels are selected using unadjusted p-values, we obtained 24 GWAS traits, of which 22 are related to liver functions (Fig S17). This analysis demonstrated that the HOT loci are important for phenotypic homeostasis.

Discussion

HOT loci have been noticed and studied in different species since the early years of the advent of the ChIP-seq datasets [12–16,27]. Up until recently, most of the studies have extensively studied the reasons through which the ChIP-seq peaks appeared to be binding to HOT loci with no apparent sequence specificity and characterized certain sequence features of the HOT loci which could enable elevated read mapping rates [13,17,18]. As the number of assayed DAPs in multiple human cell types and model organisms has increased, however, the assumption of the HOT loci being exceptional cases and results of false positives in ChIP-seq protocols have given way to the acceptance that the HOT loci, with exorbitant numbers of mapped TFBSs, are indeed hyperactive loci with distinct features characteristic of active regulatory regions [19,20].

In this study, we studied the HOT loci in multiple complementary aspects to the previous works and expanded the scope of characterization extensively using the functional genomics datasets. We used the two most extensively characterized differentiated cell lines of the ENCODE Project; HepG2 and K562. We also included the H1-hESC human stem cells to study the activities of HOT loci during the embryonic stage. The number of assayed DAPs in these cell lines is far from complete [28], therefore it is important to note that as the sizes of the assayed DAP ChIP-seq datasets increase, our understanding of the mechanisms of HOT loci will certainly improve. However, the core principles can already be inferred using the currently available datasets.

Previous studies have used different metrics to define the HOT loci. For example, Wreczycka et al. 2019 [18] used the 99th percentile of the density of TFBSs for a 500 bp sliding window, Remaker et al. 2020 [19] used the window length of 2 kb and required >25% of TFs to be mapped, Partridge et al. 2020 [20] used loci with >70 chromatin-associated proteins in 2 kb window. These heterogeneous definitions, however, fail to appreciate that the histogram of loci binned by the number of harbored TFBSs represents an exponential distribution (Fig 1A, Fig S1). We, therefore, applied our analyses both to the binarily defined HOT and non-HOT loci, as well as to the overall spectrum of loci in the context of TFBS density. This approach allowed us to better understand the correlations of characteristics of loci with the TF activity. Noticeably, this approach showed us that the HOT loci have their propensities to engage in long-range chromatin contacts with other equally or more DAP-bound loci than less active ones, making it more clear that the HOT loci are located in 3D hubs and FIREs (Fig 3A).

Using the datasets generated in H1 we established that only <10% of the HOT loci in two differentiated cell lines overlap with the HOT loci of stem cells. This points to the high tissue-specificity of the HOT loci. Previous studies have also concluded that the HOT loci are not constitutive by nature, and are established in a dynamic manner after the differentiation [21]. Of note, we used the datasets related to H1 cells in order to study the developmental aspects of the HOT phenomenon, and due to the much smaller sizes of the available datasets, we did not include the H1 in other parts of the analyses.

We conducted our analyses using a more comprehensive set of datasets of functional genomics including Hi-C data, eQTLs, raQTLs, etc. Our approach of splitting the HOT loci into enhancer and promoter regions allowed us to detect distinct patterns characteristic of these two categories. While the HOT promoters and enhancers share some sequence features, they are bound by a distinct set of DAPs and possess different biases in enrichments of different types of QTLs.

We have analyzed the patterns of DAPs in HOT loci using PCA in a similar way described in Partridge et al. 2020 [20], which was conducted only on chromatin-associated proteins in HepG2 since we asked if the findings of the study conducted only on chromatin-associated proteins hold true for the HOT loci defined using an unbiased set of DAPs, and we observed that the chromatin-associated DAPs can be distinctly separated from the other transcription-related DAPs.

Previous studies have carried out extensive mapping of the binding motifs of TFs to the HOT loci and identified a small set of "anchor" binding motifs of a few key tissue-specific TFs [13,19], and proposed that perhaps these driver TFs initiated the formation of HOT loci, similar to how the pioneer factors function. More importantly, more studies have come to the conclusion that the overwhelming majority of the peaks do not contain the corresponding motifs and that most of the mapped peaks represent indirect binding through TF-TF interactions [19,20,42,46]. We relied on the conclusions of these studies in making the assumption that the inexplicably high density of DAPs could not be explained by the direct binding events and did not carry out the analyses based on DNA-binding motifs. Interestingly, the high prediction accuracy of our deep learning model is in agreement with the notion of the existence of shared motifs among the HOT loci but also implies that the indirectly bound loci also carry shared sequence features, perhaps other than the binding

motifs or weak motifs which are not detected using the traditional PWM-based tools of motif detection. More studies are needed to further categorize the HOT loci along with the binding affinities of TFs.

Another model that has been increasingly attributed to the formation and maintenance of long-range 3D chromatin interactions involves phase-separated condensates [47–50]. Some enhancers (dubbed MegaTrans enhancers) were shown to drive the formation of large chromosomal assemblies involving a high concentration of TFs [47]. In general, it has been increasingly appreciated that condensates ubiquitously attract and activate enhancers [51–53]. The property of the condensates, which is of special interest to this study, is the capacity to serve as a "storage" of factors and co-factors inside the phase-separated droplets. For instance, the condensates can store hundreds of p300 molecules at active enhancers such that their catalytic histone acetyltransferase activity is decreased while in the phase-separated state, essentially kept in dormant mode until released [54]. The detection of condensates relies on low-throughput live cell imaging methods such as FISH, which often involves only a few tagged molecules. Therefore, currently, there are no datasets of condensate formation with large numbers of molecules simultaneously that we could use to make statistical inferences. However, there is already an increasing body of research reporting that many transcriptional activities are driven by the formation of condensates, where each of them studies individual proteins in their contexts. Based on all this, we postulate that the HOT loci might be the loci where transcriptional condensates form. Once the condensates of sufficient size form, the kinetic trap that it creates can facilitate the accumulation of a soup of DAPs, which then can undergo high-intensity protein-protein and protein-DNA interactions, many constituents of which then get mapped to the involved DNA regions upon ChIP-seq experiments.

Condensates formed at different foci in the nucleus have been shown to acquire physiochemical properties depending on their functions [49]. For instance, the sizes of the transcriptional condensates have been shown to be regulated by the concentration of RNA molecules contained in them [55]. Initially, RNA molecules serve as scaffolds to form the condensates, however, once the concentration of nascent RNA starts to increase due to transcription the condensates dissolve, providing a regulatory feedback loop for the condensates, thus explaining the phenomenon of transcriptional bursts [56]. Another aspect that this RNA-based condensate regulation explains is the enrichment of transcribed RNAs in the active enhancers. Indeed, we observed extreme enrichment of eRNAs in HOT enhancers (Fig S18), further supporting the condensate hypothesis of the HOT loci.

With the condensates assumed, the HOT loci become all the more explainable since ChIP-seq extracts the reads from populations of millions of cells, amounting to an average of many underlying protein-protein and protein-DNA interactions. With the advent of more precise protocols such as CUT&RUN, micro-C, and single-cell versions of ChIP-seq, ATAC-seq combined with bigger databases of experimentally verified condensate studies, we will have a better understanding of how the HOT loci form and gain insights into the causal relations between the high concentrations of DAPs and the transcriptional condensates.

Methods

Datasets

Transcription factor (DAP), histone modification, DNase-I hypersensitivity sites ChIP-seq and ATAC-seq datasets for HepG2, K562, H1-hESC cell lines were batch downloaded from the ENCODE Project [57]. For each DAP of each cell line, if there were multiple datasets, the one with the latest date was selected, prioritizing the ones with the least among of audit errors and warnings (Table S1). The GRCh37/hg19 assembly was used as a reference genome throughout the study. In those cases when ChIP-seq dataset was reported on GRCh38/hg38, the coordinates were converted to hg19 using liftOver. The phastCons evolutionary conservation scores generated from 46 vertebrate species, placental mammals and primates, CpG islands, repeat elements and GENCODE TSS annotations were all obtained from the UCSC genome browser database [11]. Transcribed enhancer regions (eRNAs) were obtained from the FANTOM database [58]. Super-enhancer regions were obtained from (Hnisz et al. 2013) [59].

Hi-C datasets were obtained from ENCODE Project. Please refer to Supplemental Methods 1.3 for detailed description of Hi-C data analysis.

GC contents were calculated using the "nuc" functionality of the bedtools program [60]. Gene expression data was obtained from the Roadmap Epigenomics project. For analyzing the expression levels of target genes, the gene of the overlapping TSS was used for promoters, whereas for enhancers, the nearest genes were selected using the *bedtools closest* function. Tissue-specificity metric *tau* scores for genes were downloaded from (Palmer et al. 2021) [39] which were calculated using the data mined from Gene Expression Omnibus [61].

Definitions

The loci were divided into bins according to a two-part scale. The first part is on a linear scale from 1 to 5, the second part is on a logarithmic scale from 5 to the maximum number of DAPs bound to a single locus in that cell line (Table 1). These nominal numbers are used in cases when the distributions are displayed for individual cell lines (such as Fig1A and Fig). When the figures display the distributions for two cell lines in a joint manner (such as Fig3A,B), the edges are converted to the average percentages of the overall scale lengths for each cell line.

Regular enhancers were defined as central 400bp regions of DNase-I hypersensitivity sites (DHS) which overlap H3K27ac histone modification regions with promoter and exons removed from them.

Promoters were defined as 1.5kbs upstream and 500 downstream regions of the canonical and alternative TSS coordinates were extracted from the knownGenes.txt table obtained from UCSC Genome Browser.

All the genomic arithmetic operations were done using the *bedtools* program [60]. Figures were generated using Matplotlib [62] and Seaborn [63] packages. Statistical and numerical analyses were done using the pandas, *NumPy*, *SciPy* and *sklearn* packages [64] in *Python* programming language. Genomic repeat regions were extracted from *RepeatMasker* table obtained from http://www.repeatmasker.org/. CpG islands were extracted from *cpgIslandExt* table obtained from the UCSC Genome Browser. Protein-protein interaction network information was obtained using the https://string-db.org web interface [65].

Statistical analyses

All the statistical significance analyses were done using the *SciPy* package. Statistical significance of genomic region overlaps was calculated using the "*bedtools fisher*" command. The p-values too small to be represented by the command line output were represented as <1E-100.
Correlation values with the number of bound TFs were calculated using the average of the value for the bins, and the midpoint numbers of the edges of each bin.

GWAS analysis
NHGRI-EBI GWAS database variants were grouped according to their traits (dataset e0_r2022-11-29). For each GWAS SNP, LD SNPs with r2>0.8 were added using the *plink v1.9* [66] program using the parameters *--ld-window-r2 0.8 --ld-window-kb 100 --ld-window 1000000*. Enrichments of GWAS-trait SNPs were calculated as the ratios of densities of SNPs in each class of regions (eg. HOT enhancers, HOT promoters) to either that of the regular enhancers or the DHS regions. Statistical significance of enrichment was calculated using the binomial test. FDR values were calculated using the Bonferroni correction.

Sequence classification analysis
Classification tasks were constructed in a binary classification setup. The control regions were used from: a) Randomly selected (10x the size of the HOT loci) merge DHS regions from all the available datasets from Roadmap Epigenomic Project b) using all of the promoter regions as defined above c) regular enhancers as defined above, with the HOT loci subtracted (see Supplemental Methods 1.6.1 for details).
Sequence-based classification (CNN): sequences were converted to one-hot encoding and a Convolutional Neural Network was trained using each of the control regions as negative set. The model was built using *tensorflow v2.3.1* [67] and trained on NVIDIA k80 GPUs (see Supplemental Methods 1.6.2.1 for details).
Sequence-based classification (SVM): SVM models were trained using the LS-GKM package [37] (see Supplemental Methods 1.6.2.2 for details).
Feature-based classification: sequences were represented in terms of GC, CpG, GpC contents and overlap percentages with annotated CpG islands. Logistic regression and SVM classifiers were trained using these sequence features (see Supplemental Methods 1.6.3 for details).

Variant analysis
Common SNPs and INDELs were extracted from the *gnomAD r2.1.1* dataset [41]. Variants with PASS filter value and MAF>5% were selected using the "view -f PASS -i 'MAF[0]>0.05'" options of *bcftools* program [68]. Loss-of-function variants were downloaded from the *gnomAD* website under the option "all homozygous LoF curation" section of v2.1.1 database. raQTLs were downloaded from https://sure.nki.nl [45]. Liver and blood eQTLs were extracted from the GTEx v8 dataset (https://www.gtexportal.org/home/datasets). Liver caQTLs were obtained from the supplementary material of [44].

Software and Data Availability Statement

Data and codebase used for generating the results presented in this manuscript is available at: https://github.com/okurman/HOT

Acknowledgements

This work utilized the computational resources of the NIH HPC Biowulf cluster. (http://hpc.nih.gov). This research was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health.

Figures

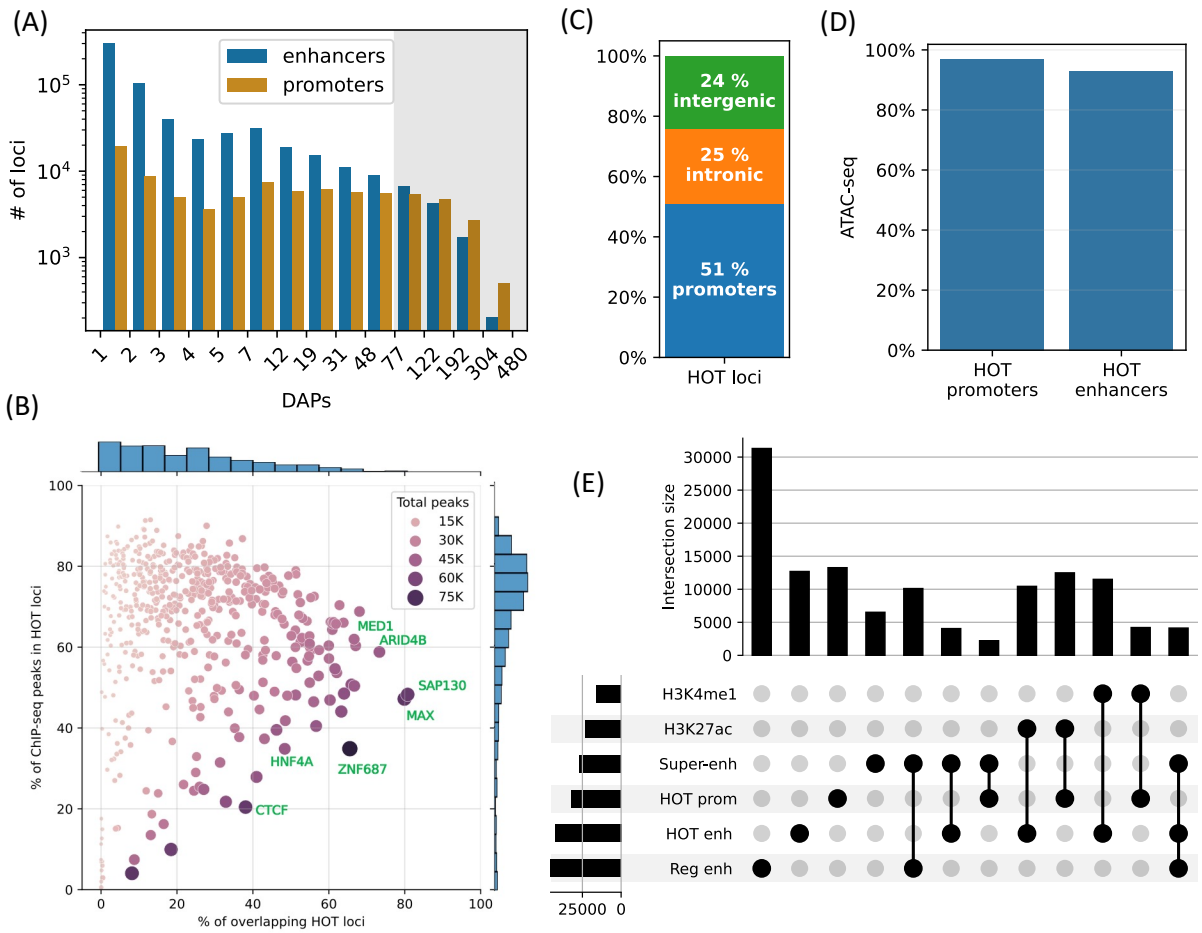


Figure 1. HOT loci are prevalent in the genome. **A**) Distribution of the number of loci by the number of overlapping peaks 400bp loci. Loci are binned on a logarithmic scale (Table S1. Methods). The shaded region represents the HOT loci. **B**) Prevalence of DAPs in HOT loci. Each dot is a DAP. X-axis: percentage

of HOT loci in which DAP is present. Y-axis: percentage of total peaks of DAPs that are located in HOT loci. Dot color and size are proportional to the total number of ChIP-seq peaks of DAP. **C**) Breakdown of HepG2 HOT loci to the promoter, intronic and intergenic regions. **D**) Fractions of HOT enhancer and promoter loci located in ATAC-seq. **E**) Overlaps between the HOT enhancer, HOT promoter, super-enhancer, regular enhancer, H3K27ac, and H4K4me1 regions. All of the visualized data is generated from the HepG2 cell line.

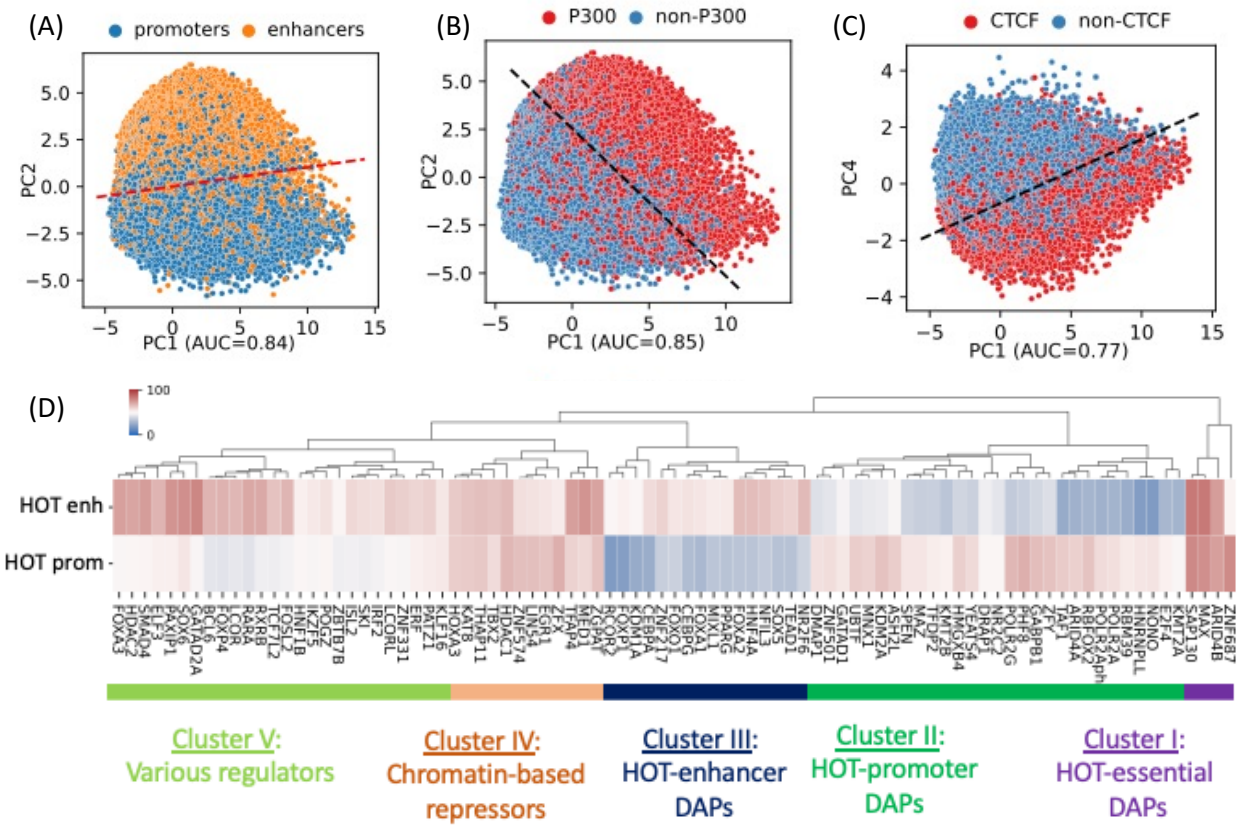


Figure 2. PCA plots of HOT loci based on the DAP presence vectors. Each dot represents a HOT locus: **A**) PC1 and PC2, marked promoters and enhancers. **B**) PC1 and PC2, marked p300 bound HOT loci. **C**) PC1 and PC4, marked CTCF bound HOT loci. The dashed lines in A,B,C are logistic regression lines. auROC values are indicated on x-axes. **D**) DAPs hierarchically clustered by their involvement in HOT promoters and HOT enhancers. Heatmap colors indicate the % of HOT enhancers or promoters that a given DAP overlaps with. All of the visualized data is generated from the HepG2 cell line.

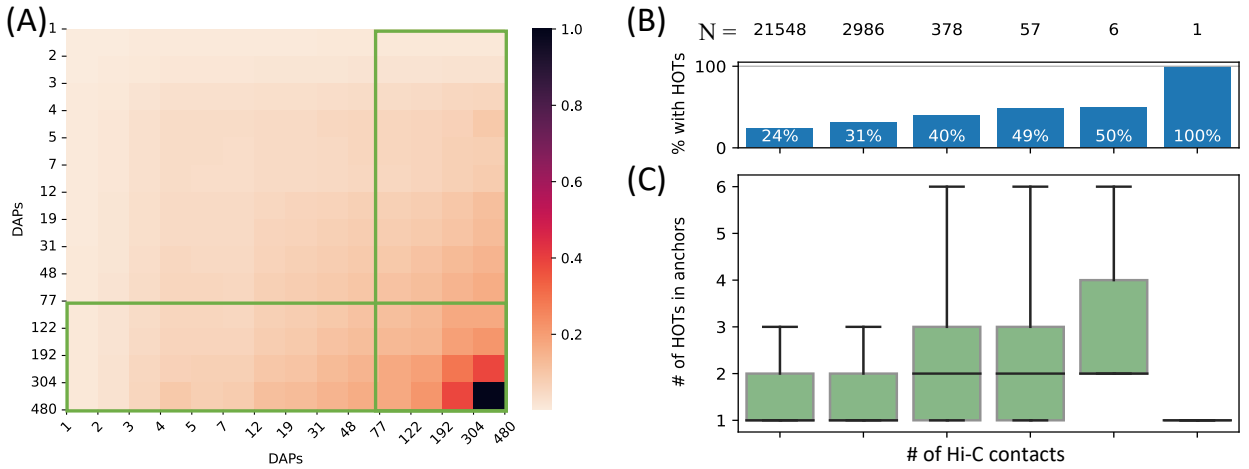


Figure 3. **A**) Densities of long-range Hi-C chromatin contacts between the DAP-bound loci. Each horizontal and vertical bin represents the loci with the number of bound DAPs between the edge values. The density values of each cell are normalized by the maximum value across all pairwise bins. Green boxes represent HOT loci. **B**) Distribution of HOT loci in Hi-C contact regions. X-axis is the number of Hi-C contacts. Numbers in the top row indicate the total number of genomic loci engaging in the given number of Hi-C contacts. Bars indicate the % of loci that contain at least one HOT locus. **C**) Distribution of the number of HOT loci in regions with a given number of Hi-C contacts. X-axis is the same as B. All of the visualized data is generated from the HepG2 cell line.

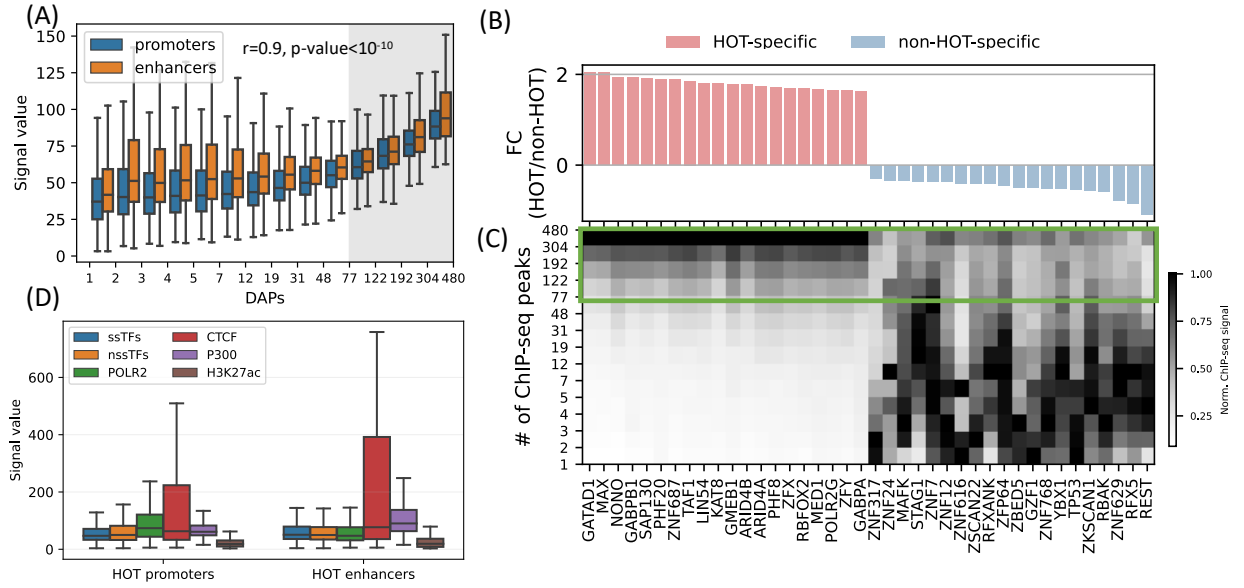


Figure 4. HOT regions induce strong ChIP-seq signals. **A**) Distribution of the signal values of the ChIP-seq peaks by the number of bound DAPs. The shaded region represents the HOT loci. **B,C**) DAPs sorted by the ratio of ChIP-seq signal strength of the peaks located in HOT loci and non-HOT loci. 20 most HOT-specific (red bars) and 20 most non-HOT-specific (blue bars) DAPs are depicted. **B**) Fold change (log2) of the HOT and non-HOT loci ChIP-seq signals. **C**) Distribution of the average ChIP-seq signal in the loci binned by the number of bound DAPs. Rows represent the loci with the bound DAPs indicated by the values of the edges (y-axis). Green box regions demarcate the HOT regions. **D**) Signal values of ssDAPs, nssDAPs

(see the text for description), H3K27ac, CTCF, P300 peaks in HOT promoters and enhancers. All of the visualized data is generated from the HepG2 cell line.

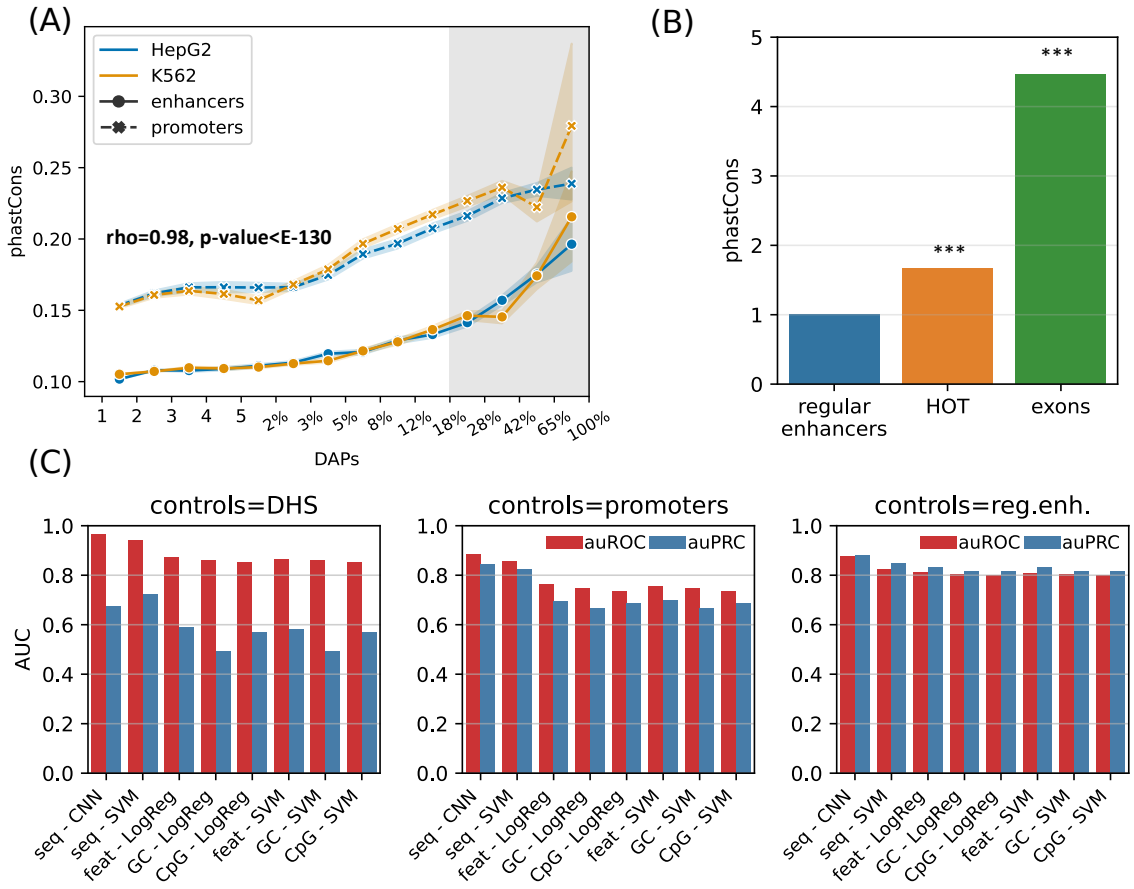
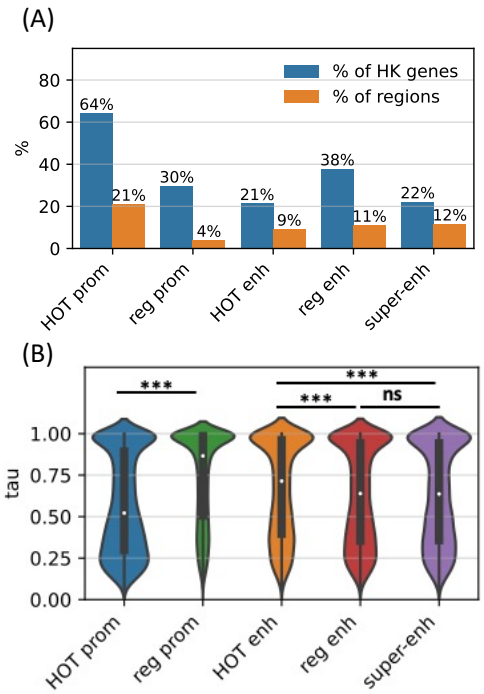


Figure 5. Sequence features of HOT loci. **A)** Distribution of conservation score in loci bound by DAPs in HepG2 and K562. The logarithmic part of the bins is expressed in terms of the percentages of loci that each bin covers, averaged over two cell lines. The correlation value is Pearson. The shaded region represents HOT loci. **B**) phastCons conservation scores of regular enhancer, HOT loci, and exon regions. The values are normalized by the average scores of regular enhancers. **C**) Classification performances (auROC and auPRC values) of HOT loci against the backgrounds of DHS, promoter, and regular enhancer regions. The X-axis values are the methods used for classifications. Methods starting with "seq -" are based on sequences (CNNs and gkmSVM, refer to Methods and main text). Starting with "feat -" are methods where all sequence features are used (GC, CpG, GpC, CpG island). Depicted values for feature-based SVMs are run using linear kernels.



(C)

GO terms	HOT proms	HOT enhs
ncRNA metabolic process	62	0
RNA processing	60	0
ncRNA processing	51	0
ribonucleoprotein complex biogenesis	48	0
rRNA processing	41	0
rRNA metabolic process	41	0
ribosome biogenesis	41	0
mitochondrion organization	41	0
mRNA metabolic process	40	0
translation	40	0
positive regulation of nuclear-transcribed mRNA	0	58
cellular response to oxygen-containing compound	0	57
cellular response to insulin stimulus	0	52
regulation of cytokine production	0	50
cellular response to epidermal growth factor stimulus	0	50
positive regulation of mRNA catabolic process	0	50
response to epidermal growth factor	0	47
neg. regulation of transcription from POLR2 promoter	0	47
regulation of hemopoiesis	0	45
response to insulin	0	45

-log10(*P*)

Figure 6. HOT promoters are ubiquitous and HOT enhancers are tissue-specific. **A)** Fractions of housekeeping genes regulated by the given category of loci (blue). Fractions of the loci which regulate the housekeeping genes (orange) **B**) Tissue-specificity (*tau*) scores of the target genes of different types of regulatory regions **C)** GO enriched terms of HOT promoters and enhancers of HepG2. 0 values in the p-values columns indicate that the GO term was not present in the top 50 enriched terms as reported by the GREAT tool. All of the visualized data is generated from the HepG2 cell line.

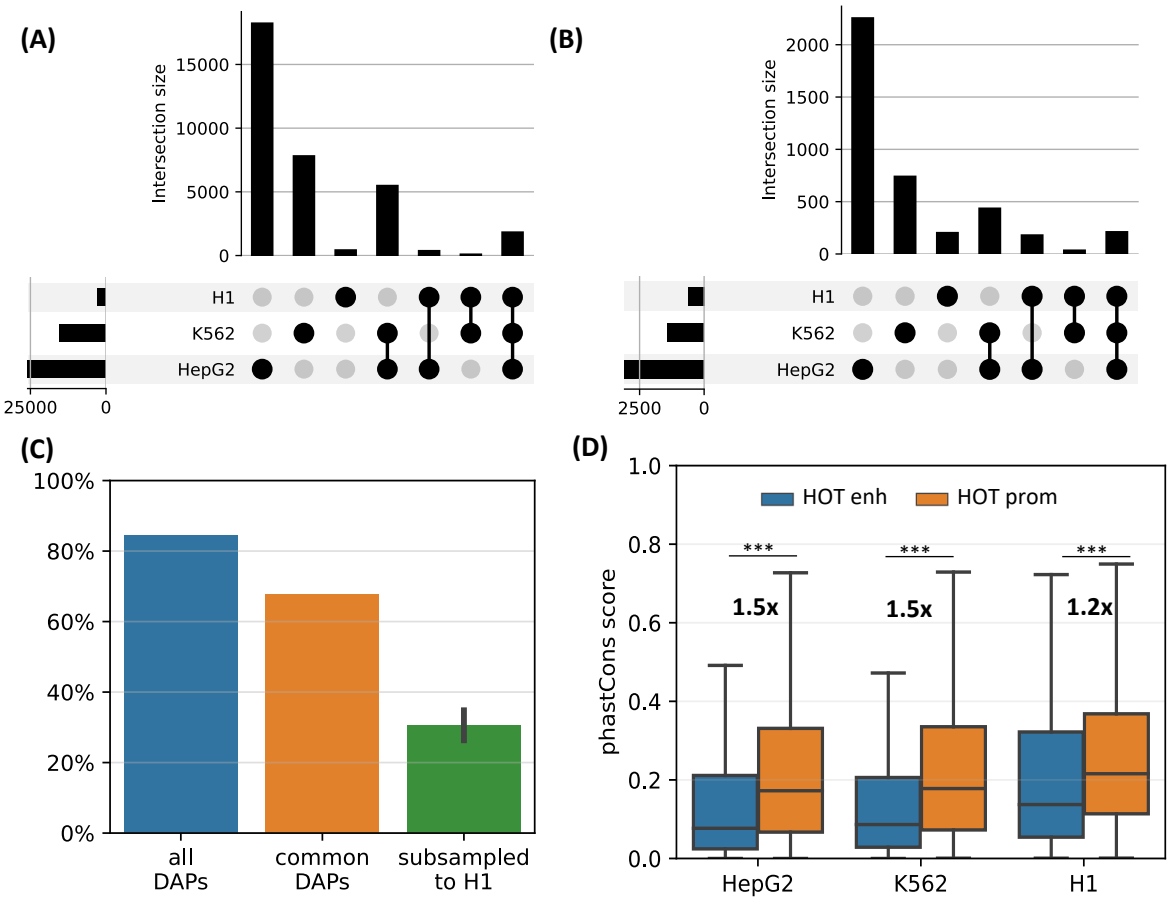


Figure 7. H1-hESC HOT loci **A**) Overlaps between the HOT loci of three cell lines. **B**) Overlaps between the HOT loci of cell lines defined using the set of DAPs available in all three cell lines. **C**) Fractions of H1 HOT loci overlapping that of the HepG2 and K562 using the complete set of DAPs, common DAPs, and DAPs randomly subsampled in HepG2/K562 to match the size of H1 DAPs set **D**) phastCons scores of HOT loci in HepG2, K562, and H1. The ratio of average conservation scores of HOT promoters with that of the HOT enhancers is at the top of every cell line's group.

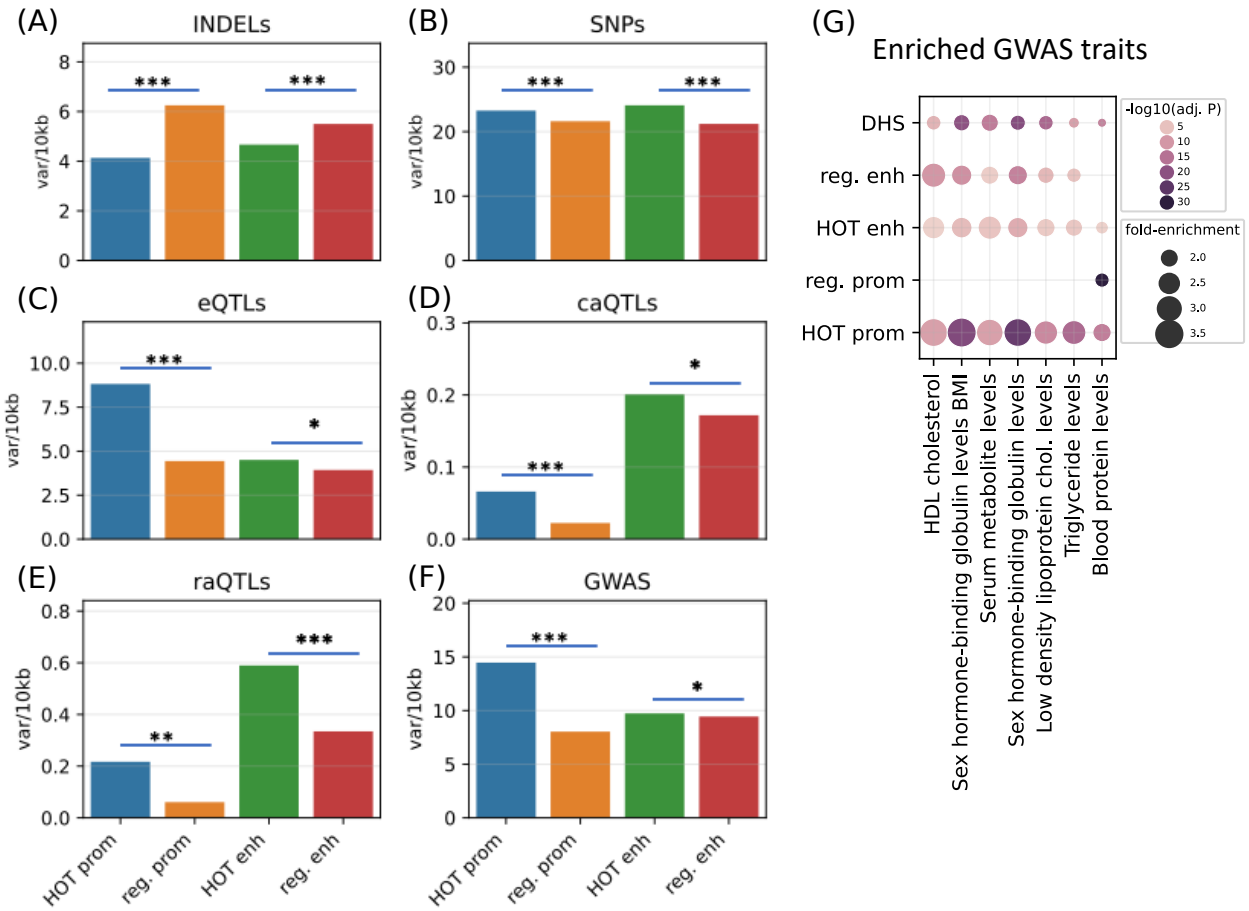


Figure 8. Densities of variants **A**) common INDELs (MAF>5%), **B**) common SNPs (MAF>5%), **C**) eQTLs, **D**) caQTLs **E**) raQTLs, and **F**) GWAS and LD (r2>0.8) variants in HOT loci and regular promoters and enhancers. **G**) Enriched GWAS traits in HOT enhancers and promoters. All of the visualized data is generated from the HepG2 cell line.

Bibliography

1. ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).

2. Gorkin, D. U. *et al.* An atlas of dynamic chromatin landscapes in mouse fetal development. *Nature* **583**, 744–751 (2020).

3. Forsberg, M. & Westin, G. Enhancer activation by a single type of transcription factor shows cell type dependence. *EMBO J.* **10**, 2543–2551 (1991).

4. Serfling, E., Jasin, M. & Schaffner, W. Enhancers and eukaryotic gene transcription. *Trends Genet.* **1**, 224–230 (1985).

5. Sethi, A. *et al.* Supervised enhancer prediction with epigenetic pattern recognition and targeted validation. *Nat. Methods* **17**, 807–814 (2020).

6. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012).

7. Long, H. K., Prescott, S. L. & Wysocka, J. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* **167**, 1170–1187 (2016).

8. Thanos, D. & Maniatis, T. Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. *Cell* **83**, 1091–1100 (1995).

9. Merika, M. & Thanos, D. Enhanceosomes. *Curr. Opin. Genet. Dev.* **11**, 205–208 (2001).

10. Arnosti, D. N. & Kulkarni, M. M. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J. Cell. Biochem.* **94**, 890–898 (2005).

11. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).

12. modENCODE Consortium *et al.* Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330**, 1787–1797 (2010).

13. Moorman, C. *et al.* Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster. *Proc Natl Acad Sci USA* **103**, 12027–12032 (2006).

14. Gerstein, M. B. *et al.* Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* **330**, 1775–1787 (2010).

15. Kvon, E. Z., Stampfel, G., Yáñez-Cuna, J. O., Dickson, B. J. & Stark, A. HOT regions function as patterned developmental enhancers and have a distinct cis-regulatory signature. *Genes Dev.* **26**, 908–913 (2012).

16. Yip, K. Y. *et al.* Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol.* **13**, R48 (2012).

17. Teytelman, L., Thurtle, D. M., Rine, J. & van Oudenaarden, A. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proc Natl Acad Sci USA* **110**, 18602–18607 (2013).

18. Wreczycka, K. *et al.* HOT or not: examining the basis of high-occupancy target regions. *Nucleic Acids Res.* **47**, 5735–5745 (2019).

19. Ramaker, R. C. *et al.* Dissecting the regulatory activity and sequence content of loci with exceptional numbers of transcription factor associations. *Genome Res.* **30**, 939–950 (2020).

20. Partridge, E. C. *et al.* Occupancy maps of 208 chromatin-associated proteins in one human cell type. *Nature* **583**, 720–728 (2020).

21. Boyle, A. P. *et al.* Comparative analysis of regulatory information and circuits across distant species. *Nature* **512**, 453–456 (2014).

22. Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).

23. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).

24. Bhattacharyya, S., Chandra, V., Vijayanand, P. & Ay, F. Identification of significant chromatin contacts from HiChIP data by FitHiChIP. *Nat. Commun.* **10**, 4221 (2019).

25. Schmitt, A. D. *et al.* A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Rep.* **17**, 2042–2059 (2016).

26. Quinodoz, S. A. *et al.* Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell* **174**, 744-757.e24 (2018).

27. Xie, D. *et al.* Dynamic trans-acting factor colocalization in human cells. *Cell* **155**, 713–724 (2013).

28. Lambert, S. A. *et al.* The human transcription factors. *Cell* **172**, 650–665 (2018).

29. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).

30. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–

94 (2017).

31. Liu, J. *et al.* Obg-Like ATPase 1 Enhances Chemoresistance of Breast Cancer via Activation of TGF-β/Smad Axis Cascades. *Front. Pharmacol.* **11**, 666 (2020).
32. Biel, A. *et al.* AUTS2 syndrome: molecular mechanisms and model systems. *Front. Mol. Neurosci.* **15**, 858582 (2022).
33. Schanze, I. *et al.* Haploinsufficiency of SOX5, a member of the SOX (SRY-related HMG-box) family of transcription factors is a cause of intellectual disability. *Eur. J. Med. Genet.* **56**, 108–113 (2013).
34. Daigle, T. L. *et al.* A Suite of Transgenic Driver and Reporter Mouse Lines with Enhanced Brain-Cell-Type Targeting and Functionality. *Cell* **174**, 465-480.e22 (2018).
35. Pachano, T. *et al.* Orphan CpG islands amplify poised enhancer regulatory activity and determine target gene responsiveness. *Nat. Genet.* **53**, 1036–1049 (2021).
36. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25**, 1010–1022 (2011).
37. Lee, D. LS-GKM: a new gkm-SVM for large-scale datasets. *Bioinformatics* **32**, 2196–2198 (2016).
38. Hounkpe, B. W., Chenou, F., de Lima, F. & De Paula, E. V. HRT Atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets. *Nucleic Acids Res.* **49**, D947–D955 (2021).
39. Palmer, D., Fabris, F., Doherty, A., Freitas, A. A. & de Magalhães, J. P. Ageing transcriptome meta-analysis reveals similarities and differences between key mammalian tissues. *Aging (Albany NY)* **13**, 3313–3341 (2021).
40. Domené, S. *et al.* Enhancer turnover and conserved regulatory function in vertebrate evolution. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **368**, 20130027 (2013).
41. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
42. Vierstra, J. *et al.* Global reference mapping of human transcription factor footprints. *Nature* **583**,

729–736 (2020).

43. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).

44. Currin, K. W. *et al.* Genetic effects on liver chromatin accessibility identify disease regulatory variants. *Am. J. Hum. Genet.* **108**, 1169–1189 (2021).

45. van Arensbergen, J. *et al.* High-throughput identification of human SNPs affecting regulatory element activity. *Nat. Genet.* **51**, 1160–1169 (2019).

46. White, S. M., Snyder, M. P. & Yi, C. Master lineage transcription factors anchor trans mega transcriptional complexes at highly accessible enhancer sites to promote long-range chromatin clustering and transcription of distal target genes. *Nucleic Acids Res.* **49**, 12196–12210 (2021).

47. Nair, S. J. *et al.* Phase separation of ligand-activated enhancers licenses cooperative chromosomal enhancer assembly. *Nat. Struct. Mol. Biol.* **26**, 193–203 (2019).

48. Lee, R. *et al.* CTCF-mediated chromatin looping provides a topological framework for the formation of phase-separated transcriptional condensates. *Nucleic Acids Res.* **50**, 207–226 (2022).

49. Feric, M. & Misteli, T. Function moves biomolecular condensates in phase space. *Bioessays* **44**, e2200001 (2022).

50. Ahn, J. H. *et al.* Phase separation drives aberrant chromatin looping and cancer development. *Nature* **595**, 591–595 (2021).

51. Shrinivas, K. *et al.* Enhancer Features that Drive Formation of Transcriptional Condensates. *Mol. Cell* **75**, 549-561.e7 (2019).

52. Wei, M.-T. *et al.* Nucleated transcriptional condensates amplify gene expression. *Nat. Cell Biol.* **22**, 1187–1196 (2020).

53. Boija, A. *et al.* Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. *Cell* **175**, 1842-1855.e16 (2018).

54. Zhang, Y. *et al.* Nuclear condensates of p300 formed though the structured catalytic core can act as a storage pool of p300 with reduced HAT activity. *Nat. Commun.* **12**, 4618 (2021).

55. Sharp, P. A., Chakraborty, A. K., Henninger, J. E. & Young, R. A. RNA in formation and regulation of transcriptional condensates. *RNA* **28**, 52–57 (2022).

56. Henninger, J. E. *et al.* RNA-Mediated Feedback Control of Transcriptional Condensates. *Cell* **184**, 207-225.e24 (2021).

57. Wang, J. *et al.* Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Res.* **41**, D171-6 (2013).

58. Lizio, M. *et al.* Update of the FANTOM web resource: expansion to provide additional transcriptome atlases. *Nucleic Acids Res.* **47**, D752–D758 (2019).

59. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).

60. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

61. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res.* **41**, D991-5 (2013).

62. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

63. Waskom, M. seaborn: statistical data visualization. *JOSS* **6**, 3021 (2021).

64. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

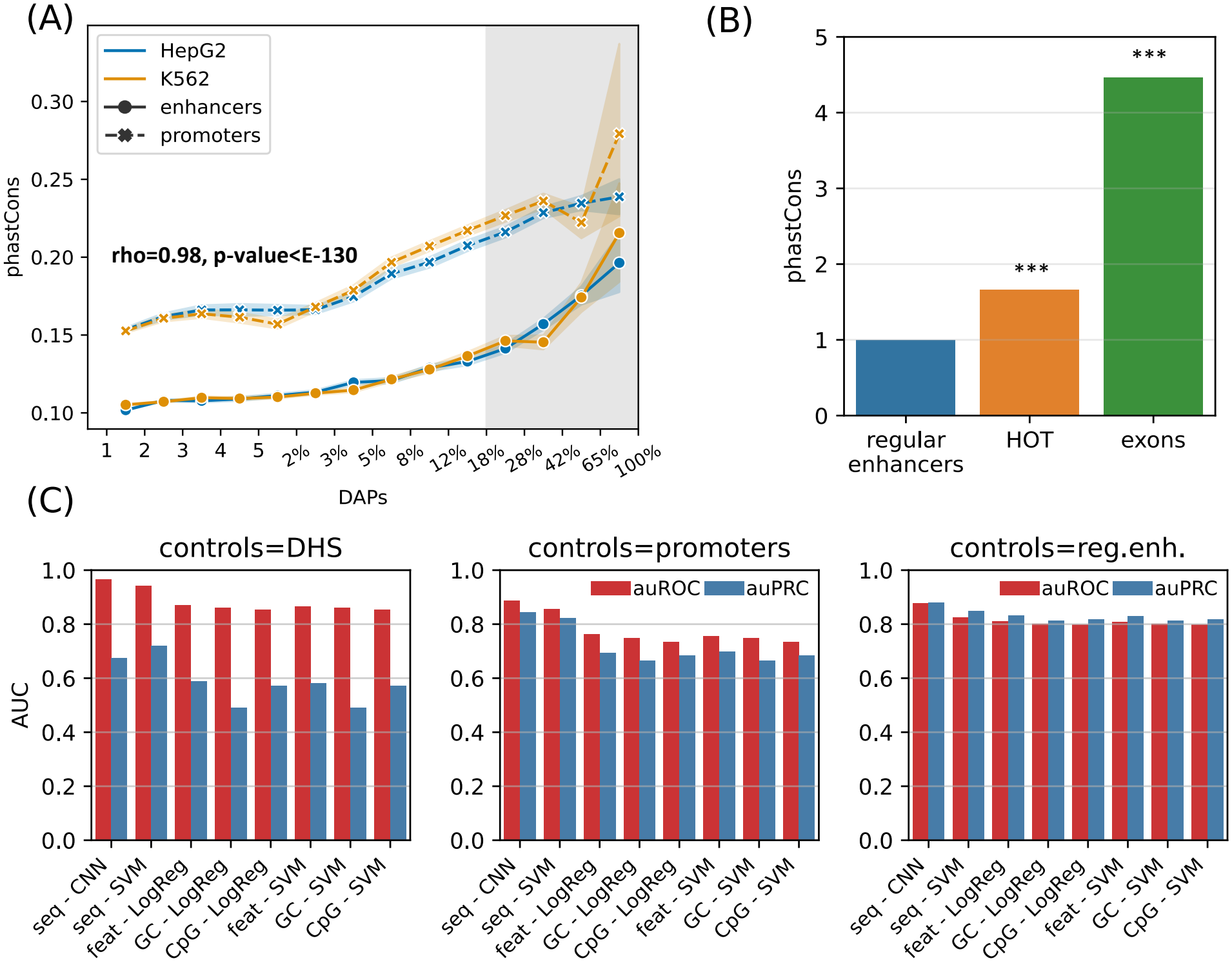
65. Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).

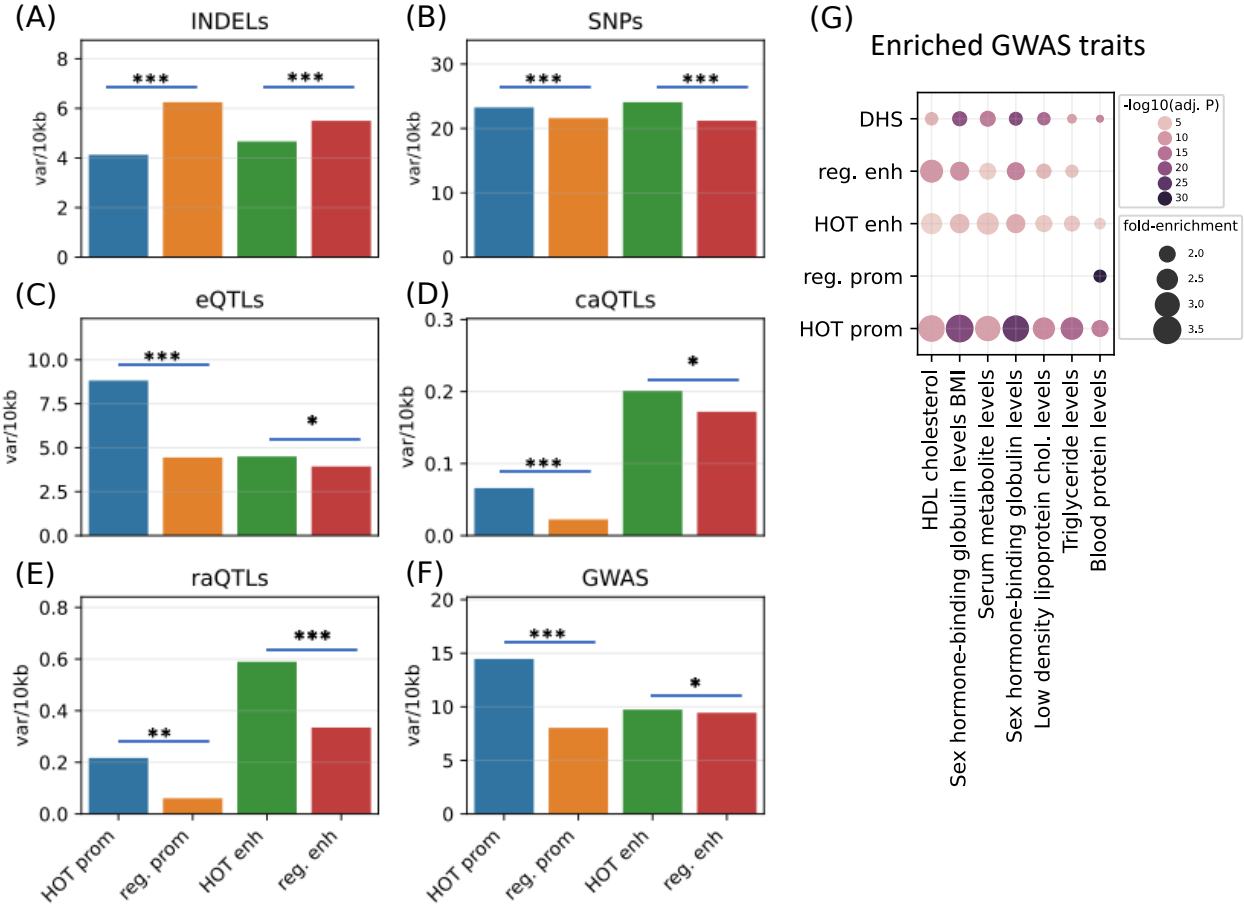
66. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

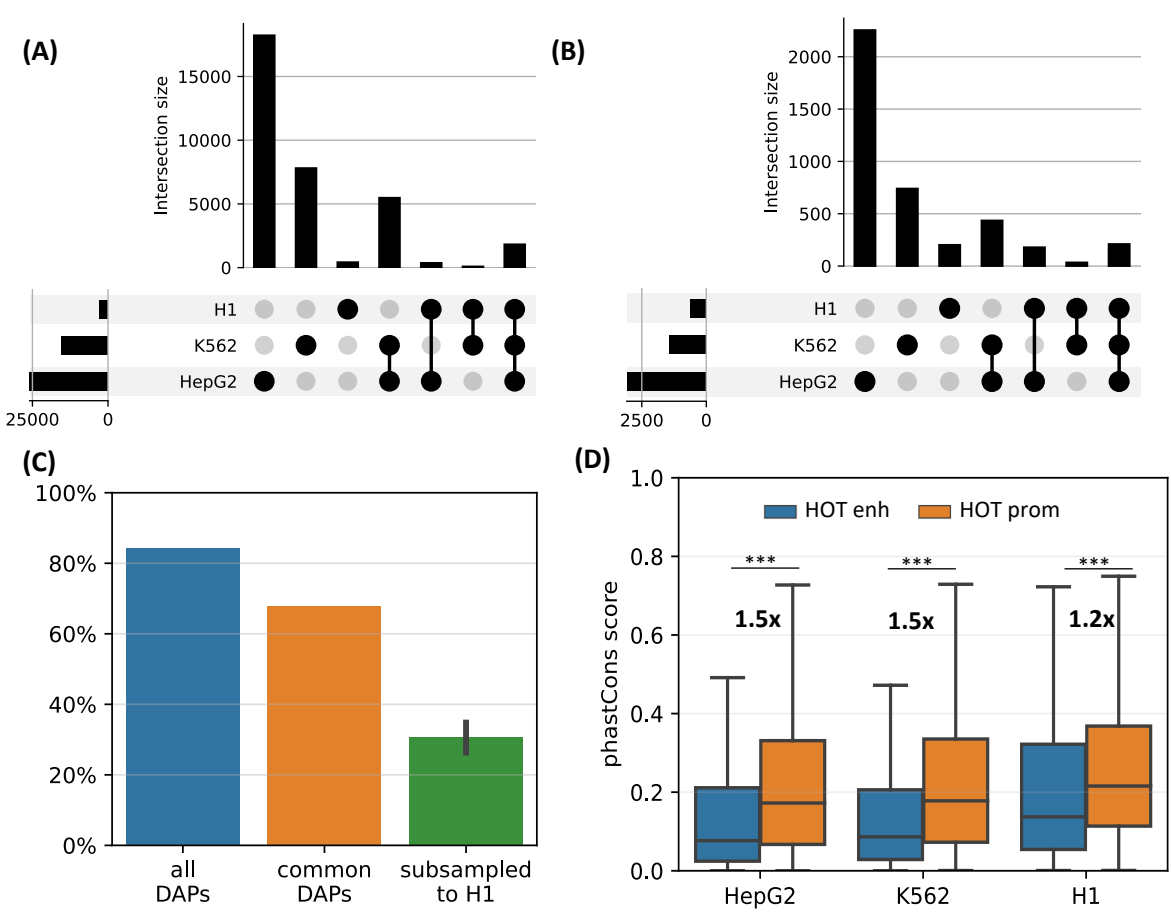
67. Abadi, M. *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv* (2016) doi:10.48550/arxiv.1603.04467.

68. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993

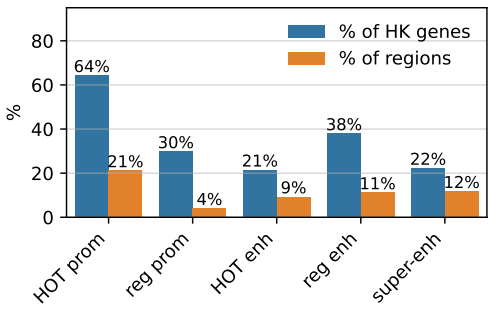
(2011).


(A)
HepG2
K562
enhancers
promoters
rho=0.98, p-value<E-130
phastCons
0.30
0.25
0.20
0.15
0.10
1
2
3
4
5
2%
3%
5%
8%
12%
18%
28%
42%
65%
100%
DAPs
(B)
phastCons
5
4
3
2
1
0
***
***
regular enhancers
HOT
exons
(C)
controls=DHS
controls=promoters
controls=reg.enh.
AUC
1.0
0.8
0.6
0.4
0.2
0.0
auROC
auPRC
seq - CNN
seq - SVM
feat - LogReg
GC - LogReg
CpG - LogReg
feat - SVM
GC - SVM
CpG - SVM

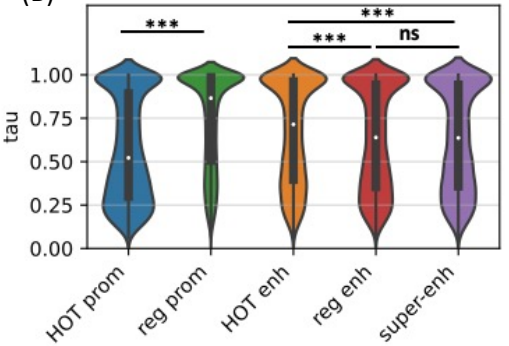

(A)
INDELs
var/10kb
***
***
(B)
SNPs
var/10kb
***
***
(C)
eQTLs
var/10kb
***
*
(D)
caQTLs
var/10kb
***
*
(E)
raQTLs
var/10kb
**
***
(F)
GWAS
var/10kb
***
*
HOT prom
reg. prom
HOT enh
reg. enh
(G)
Enriched GWAS traits
DHS
reg. enh
HOT enh
reg. prom
HOT prom
HDL cholesterol
Sex hormone-binding globulin levels BMI
Serum metabolite levels
Sex hormone-binding globulin levels
Low density lipoprotein chol. levels
Triglyceride levels
Blood protein levels
-log10(adj. P)
5
10
15
20
25
30
fold-enrichment
2.0
2.5
3.0
3.5


(A)
Intersection size
15000
10000
5000
0
H1
K562
HepG2
25000
0
(B)
Intersection size
2000
1500
1000
500
0
H1
K562
HepG2
2500
0
(C)
100%
80%
60%
40%
20%
0%
all DAPs
common DAPs
subsampled to H1
(D)
HOT enh
HOT prom
phastCons score
1.0
0.8
0.6
0.4
0.2
0.0
***
1.5x
1.5x
1.2x
HepG2
K562
H1

(A)

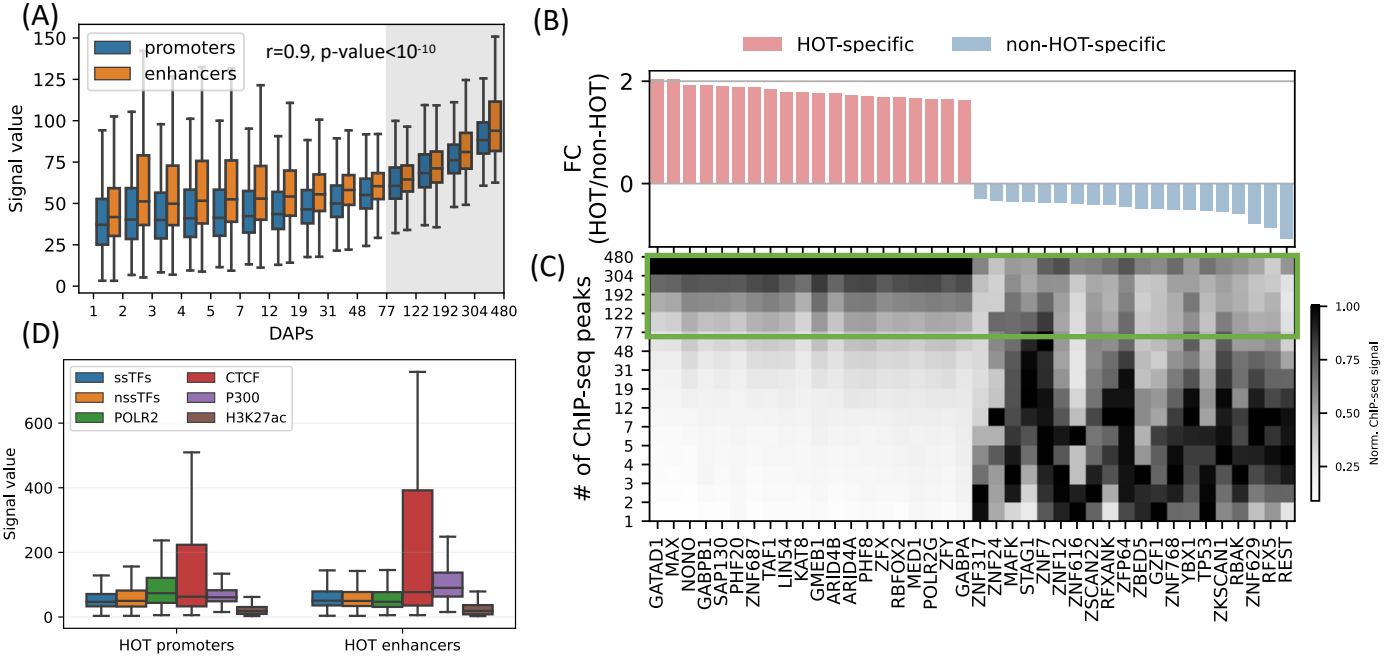


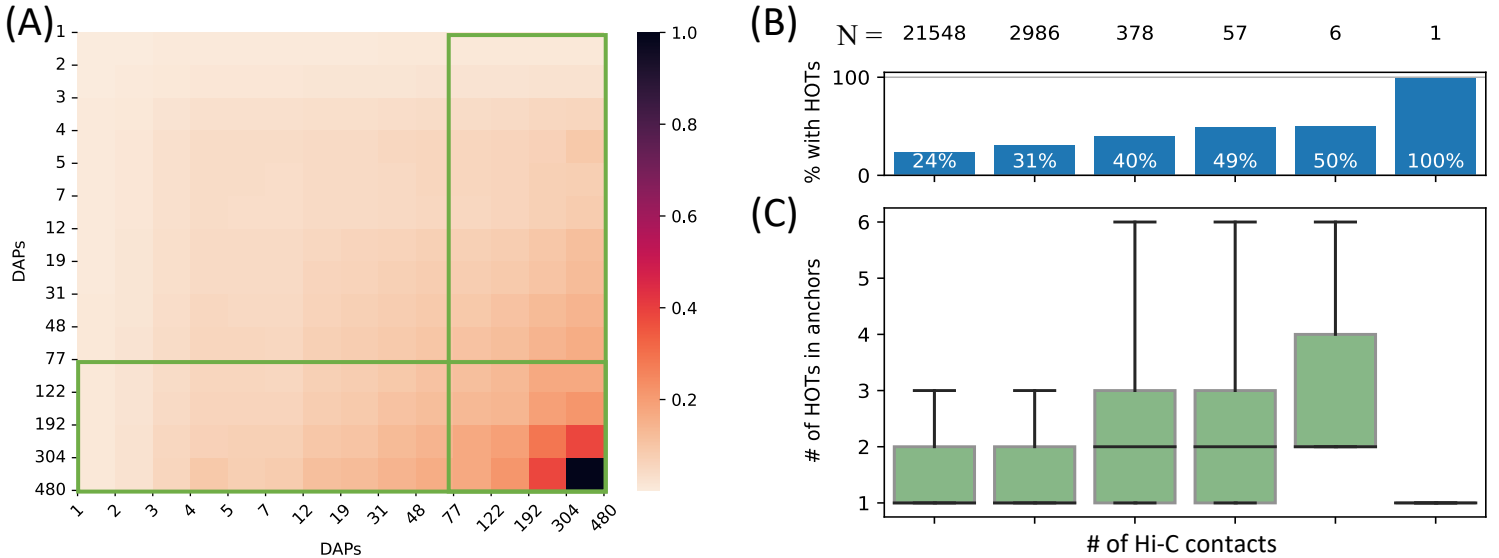
(B)

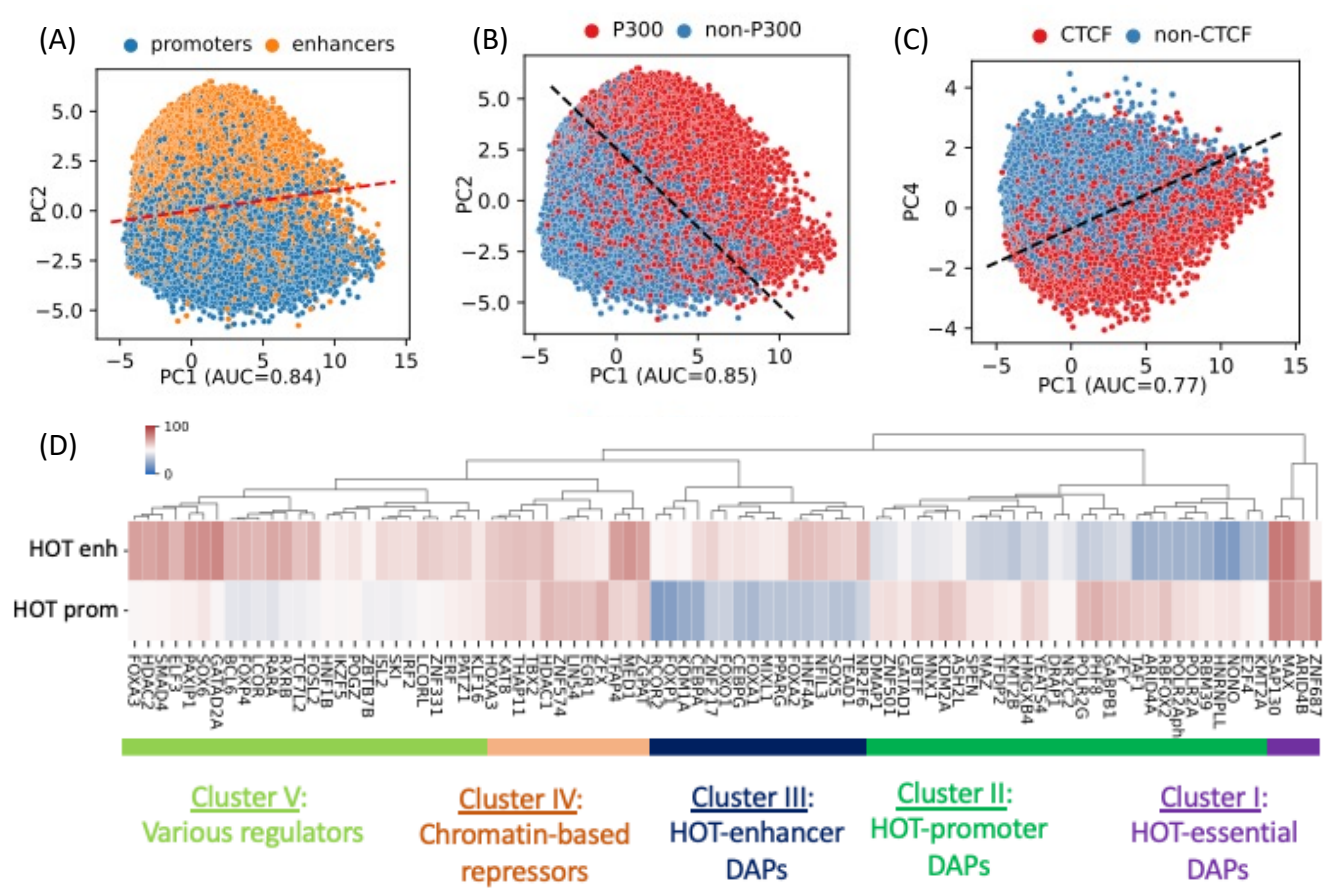


(C)

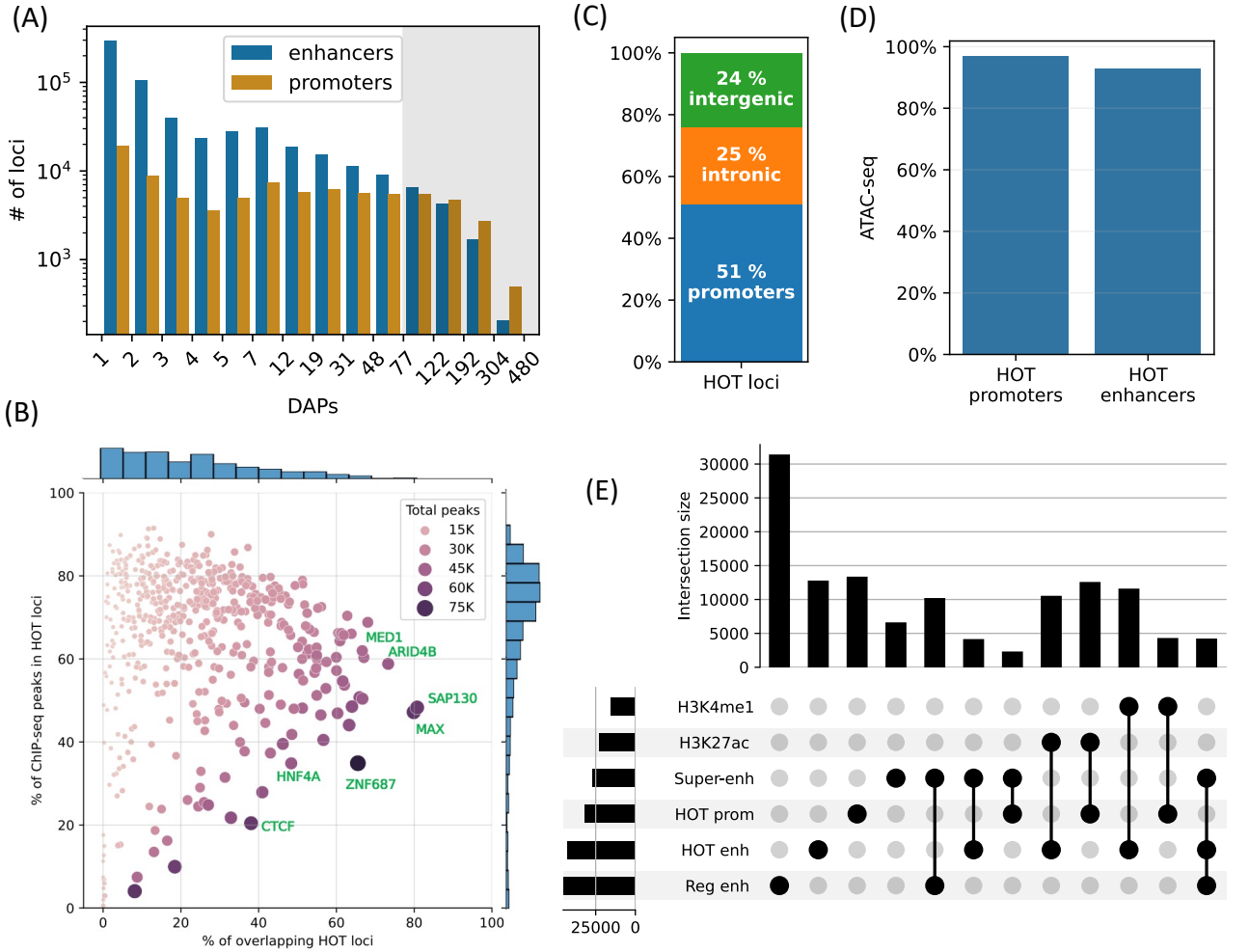
GO terms	-log10(*P*)	
ncRNA metabolic process	62	0
RNA processing	60	0
ncRNA processing	51	0
ribonucleoprotein complex biogenesis	48	0
rRNA processing	41	0
rRNA metabolic process	41	0
ribosome biogenesis	41	0
mitochondrion organization	41	0
mRNA metabolic process	40	0
translation	40	0
positive regulation of nuclear-transcribed mRNA	0	58
cellular response to oxygen-containing compound	0	57
cellular response to insulin stimulus	0	52
regulation of cytokine production	0	50
cellular response to epidermal growth factor stimulus	0	50
positive regulation of mRNA catabolic process	0	50
response to epidermal growth factor	0	47
neg. regulation of transcription from POLR2 promoter	0	47
regulation of hemopoiesis	0	45
response to insulin	0	45
	HOT proms	HOT enhs


(A)
promoters
enhancers
r=0.9, p-value<10-10
Signal value
DAPs
(B)
HOT-specific
non-HOT-specific
FC (HOT/non-HOT)
(C)
# of ChIP-seq peaks
Norm. ChIP-seq signal
GATAD1 MAX NONO GABPB1 SAP130 PHF20 ZNF687 TAF1 LIN54 KAT8 GMEB1 ARID4B ARID4A PHF8 ZFX RBFOX2 MED1 POLR2G ZFY GABPA ZNF317 ZNF24 MAFK STAG1 ZNF7 ZNF12 ZNF616 ZSCAN22 RFXANK ZFP64 ZBED5 GZF1 ZNF768 YBX1 TP53 ZKSCAN1 RBAK ZNF629 RFX5 REST
(D)
ssTFs
nssTFs
POLR2
CTCF
P300
H3K27ac
Signal value
HOT promoters
HOT enhancers


(A)
DAPs
1
2
3
4
5
7
12
19
31
48
77
122
192
304
480
1.0
0.8
0.6
0.4
0.2
(B)
N =
21548
2986
378
57
6
1
% with HOTs
100
0
24%
31%
40%
49%
50%
100%
(C)
# of HOTs in anchors
6
5
4
3
2
1
# of Hi-C contacts



(A)
promoters
enhancers
PC2
PC1 (AUC=0.84)
(B)
P300
non-P300
PC2
PC1 (AUC=0.85)
(C)
CTCF
non-CTCF
PC4
PC1 (AUC=0.77)
(D)
100
0
HOT enh
HOT prom
Cluster V: Various regulators
Cluster IV: Chromatin-based repressors
Cluster III: HOT-enhancer DAPs
Cluster II: HOT-promoter DAPs
Cluster I: HOT-essential DAPs


(A)
enhancers
promoters
# of loci
DAPs
1 2 3 4 5 7 12 19 31 48 77 122 192 304 480
(B)
Total peaks
15K
30K
45K
60K
75K
MED1
ARID4B
SAP130
MAX
HNF4A
ZNF687
CTCF
% of ChIP-seq peaks in HOT loci
% of overlapping HOT loci
(C)
24 % intergenic
25 % intronic
51 % promoters
HOT loci
(D)
ATAC-seq
HOT promoters
HOT enhancers
(E)
Intersection size
H3K4me1
H3K27ac
Super-enh
HOT prom
HOT enh
Reg enh