Sequence characteristics and an accurate model of high-occupancy target loci in the human genome.

Sanjarbek Hudaiberdiev and Ivan Ovcharenko*
Computational Biology Branch, National Library of Medicine, National Institutes of Health, Bethesda, MD, 20892, USA.
*to whom correspondence should be addressed: ovcharen@nih.gov, hudaiber@nih.gov

Abstract

Enhancers and promoters are 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 have expanded. Particularly, 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 1,003 TF ChIP-seq datasets in HepG2, K562, and H1 cells to analyze the patterns of ChIP-seq peak co-occurrence combined with functional genomics datasets. We identified 43,891 HOT loci forming at the promoter (53%) and enhancer (47%) regions and determined that HOT promoters regulate housekeeping genes, whereas the HOT enhancers are involved in extremely tissue-specific processes. HOT loci form the foundation of human super-enhancers and evolve under strong negative selection, with some of them being ultraconserved regions. Sequence-based classification of HOT loci using deep learning suggests that their formation is driven by sequence features, and the density of ChIP-seq peaks correlates with sequence features. Based on their affinities to bind to promoters and enhancers, we detected five distinct clusters of TFs that form the core of the HOT loci. We also observed that HOT loci are enriched in 3D chromatin hubs and disease-causal variants. In a challenge to the classical model of enhancer activity, we report an abundance of HOT loci in human genome and a commitment of 51% of all ChIP-seq binding events to HOT locus formation and propose a model of HOT locus formation based on the existence of large transcriptional condensates.

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 (ENCODE Project Consortium et al., 2020; Gorkin et al., 2020). Classically, an enhancer is thought to be bound by a few TFs that recognize a specific DNA motif at their cognate TF binding sites (TFBSs) through its DNA-binding domains and recruit other molecules necessary for catalyzing the transcriptional machinery (Forsberg and Westin, 1991; Serfling et al., 1985; Sethi et al., 2020). Based on the arrangements of the TFBSs, also called "motif grammar" the architecture of enhancers is commonly categorized into "enhanceosome" and "billboard" models (Spitz and Furlong 2012; Long et al. 2016). 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 (Thanos and Maniatis 1995; Merika and Thanos 2001). This model requires the presence of all the participating proteins. On the other hand, under the billboard model, the TFBSs are independent of each other and function in an additive manner (Arnosti and Kulkarni 2005). However, as the catalogs of TF ChIP-seq assays have expanded due to the major collaborative projects, such as ENCODE (Davis et al. 2018) and modENCODE (modENCODE Consortium et al. 2010), 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. Particularly, genomic regions that seemingly get bound by a large number of TFs with no apparent DNA sequence specificity have been reported. These genomic loci have been dubbed high-occupancy target (HOT) regions and were detected in multiple species (Gerstein et al., 2010; Kvon et al., 2012; modENCODE Consortium et al., 2010; Moorman et al., 2006; Yip et al., 2012).

Initially, HOT regions have been partially attributed to technical and statistical artifacts of the ChIP-seq protocol, and a small list of blacklisted regions, which are mostly located in unstructured DNA regions such as repetitive elements and low complexity regions has been produced (Teytelman et al., 2013; Wreczycka et al., 2019). These blacklisted regions, which represent a small fraction of the mapped ChIP-seq peaks, were later excluded from the analyses. In addition, various studies have proposed that some DNA elements, such as GC-rich promoters, CpG islands, R-loops, and G-quadruplexes, can serve as permissive TF binding platforms (Teytelman et al., 2013; Wreczycka et al., 2019). Yet, 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 (modENCODE Consortium et al., 2010; Partridge et al., 2020; Ramaker et al., 2020).

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., 2020) 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., 2019) 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. (Ramaker et al., 2020) 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 and 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. Based on our findings, 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 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 cells (545, 411, and 47 ChIP-seq assays, respectively, see Supp. 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, and histone acetyltransferases as well as RNA Polymerase 2 variants. We collectively call all these proteins DNA-associated proteins (DAPs). Using the datasets of DAPs, we overlaid all 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. We considered a binding site to be present in each window if 8 bp centered at the summit of a ChIP-seq peak was overlapping. Given that the three cell lines analyzed 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 which fall to the highest four bins, which translates to those that contain on average >18% of available DAPs for a given cell line (see Supp. Methods for a detailed description and justifications). This resulted in 25,928 and 15,231 HOT loci in HepG2 and K562, respectively. However, 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 used all 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 main dataset for analyses and used the K562 and H1 datasets when a 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 which are present in the highest fraction of HOT loci are SAP130, MAX, ARID4B, MED1, (Fig 1B, x-axis) and ZGPAT, HDAC1, 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. POLR2 is present in 42% of HOT loci arguing for active transcription at or in the proximity of HOT loci, including messenger RNA (mRNA) and enhancer RNA (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, ZHX3, and YEATS2, whereas the DAPs least associated with HOT loci (<5%) are ZNF282, MAFK, EZH2, and TRIM22. The fact that HOT loci harbor more than half of the ChIP-seq peaks suggests that HOT loci are one of the prevalent modes of DAP-DNA interactions rather than an exceptional case, as has been initially suggested by earlier studies (Teytelman et al., 2013; Wreczycka et al., 2019).

Approximately 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 (Fig 1C) with 9% being located >50 kbs away from promoters, suggesting that the HOT loci are mainly clustered in vicinities, i.e., in promoters and introns, of transcriptional start sites (TSSs) and, therefore, potentially play important roles in the regulation of nearby genes. Both HOT promoters and enhancers are almost entirely located in chromatin-accessible regions (97% and 93% of the total sequence lengths, respectively, Fig 1D).

To further dissect the composition of HOT enhancer loci, we compared them to super-enhancers as defined in (Whyte et al., 2013) and a set of regular enhancers (see Supp. 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 the peaks located in HOT loci (Fig 1B). To analyze the signatures of unique DAPs in HOT loci, we performed a principal component analysis (PCA) using overlapping DAP combinations for each HOT locus. This analysis showed that principal component 1 (PC1) is correlated with the total number of distinct DAPs located at a given HOT locus (Fig S2A). Principal component 2 (PC2) separates the HOT promoters and HOT enhancers (Fig 2A, Fig S2B), and the PC1-PC2 combination also separates the p300-bound HOT loci (Fig 2B, Fig S2C). 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 with which they overlap. This analysis showed that there is a large cluster of DAPs (n=458) that, on average, overlap only 17% of HOT loci and that are likely secondary to the HOT locus formation (Fig S3). We focused on the other HOT-enriched cluster of DAPs (n=87), which is present in 53% of HOT loci on average (Fig S3) and which consists of four major clusters of DAPs (Fig 2D). *Cluster I* comprises four DAPs – ZNF687, ARID4B, MAX, and SAP130, which are present in 75% of HOT loci on average. The last three of these DAPs form a protein-protein interaction (PPI) network (STRING p-value<0.001) (Fig S4A). We call 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 1.7 times more likely to overlap HOT promoters than HOT enhancers. Among these DAPs are POLR2 subunits, PHF8, GABP1, GATAD1, TAF1, etc. The Gene Ontology (GO) molecular function term most strongly associated with the DAPs of this cluster is *RNA Polymerase transcription factor initiation activity* suggestive of its direct role in transcriptional activity (Fig S4B). *Cluster III* comprises 16 DAPs, which are 1.9 times 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 (STRING p-value<10^{-16}, Fig S4C). Among the GO terms of biological processes most strongly associated with the DAPs of this cluster are those related to cell differentiation, such as *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) and which form a PPI network (STRING p-value<0.001, Fig S4D) with HDAC1 being the node with the highest degree, suggesting that the DAPs of this cluster may be involved in chromatin-based transcriptional repression. Lastly, *Cluster V* comprises 26 DAPs of a wide range of transcriptional regulators, with a 1.3-fold skew toward the HOT enhancers. While this cluster contains prominent TFs, such as TCF7L2, FOXA3, SOX6, FOSL2, etc., the variety of the pathways and interactions in which they participate 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 promoters or HOT enhancers (clusters II and III, respectively), it still does not explain what sorts of interactions take place among 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.

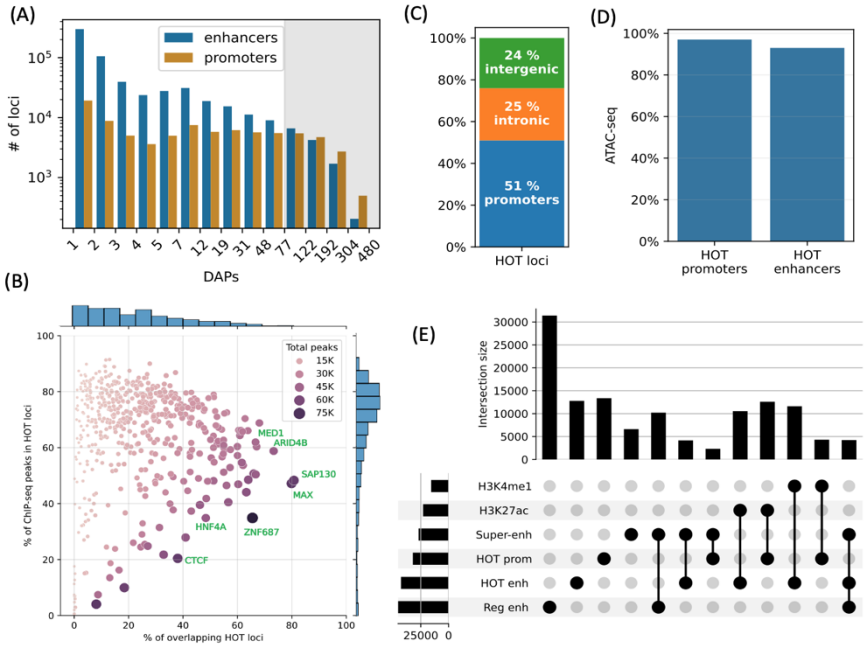

Figure 1. HOT loci are prevalent in the genome. **A**) Distribution of the loci by the number of overlapping peaks 400bp loci in HepG2. Loci are binned on a logarithmic scale (see Supp. Methods). Shaded region represents the HOT loci. **B**) Prevalence of DAPs in HOT loci. Each dot is a DAP. X-axis: 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 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.

Notably, PC4 separates HOT loci characteristic of CTCF (Fig 2C) and Cohesin (Fig S2D). 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 are present in 36% of the HOT loci, compared with some other DAPs with much higher presence described above, that do not get separated clearly by the PCA analysis (Fig 2E). 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 2F), making it more likely that those loci are proximal enhancers. Additionally, 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 2G), suggesting that at least a portion of the number of DAPs in HOT loci can be explained by long-distance chromatin contacts between the genomic regions mediated by CTCF-Cohesin complexes. To quantify the involvement of 3D chromatin interactions involving the HOT loci, we first used Hi-C data with 5 kbs resolution from the ENCODE Project and calculated the density of chromatin contacts between the DAP-bound loci categorized by the total number of bound DAPs (see Supp. Methods). This revealed that the HOT loci are enriched in Hi-C contacts and 1.4 times more likely to engage in long-range chromatin interactions than the regular enhancers (p-value<10^{-6}, binomial test). The density of chromatin contacts in DAP-bound loci correlates with the number of bound DAPs with rho=0.84 (p-value<10^{-4}; Pearson). Moreover, by analyzing the densities of chromatin contacts between the DAP-bound loci, we observed that the loci with high numbers of bound DAPs are more likely to engage in long-range chromatin contacts 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), suggesting that HOT loci are enriched in chromatin 3D hubs. To test this hypothesis, we extracted all the genomic loci that are involved in long-range chromatin interactions using the Hi-C data sorted by the number of interactions and overlapped them with the HOT loci (constructed as a network where each node is a Hi-C locus and the edges are Hi-C contacts between them). The number of long-range interactions in chromatin-interacting loci varies between 1 and 6, with only one locus, in an extreme case, engaging in six different long-range chromatin contacts (Fig 3B). That locus contains one HOT enhancer harboring 101 DAPs, located 23 kbs away from the *LINC02583* gene, which is linked to liver-

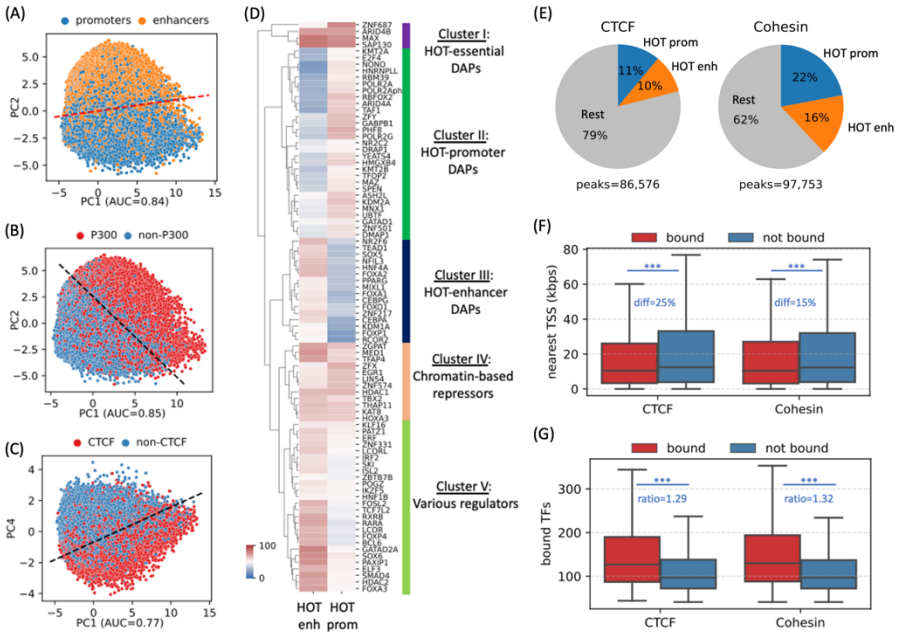

Figure 2. PCA plots of HOT loci in HepG2 based on the DAP presence vectors. Each dot represents a HOT locus: A) PC1 and PC2, marked promoters and enhancers. B) PC1 and PC2, marked p300 bound HOT loci. C) PC1 and PC4, marked CTCF bound HOT loci. D) DAPs hierarchically clustered by their involvements in HOT promoters and HOT enhancers. E) Fractions of the ChIP-seq peaks of CTCF and Cohesin in HOT promoters and enhancers. F) Distribution of distances between the nearest genes and the HOT loci bound by CTCF and Cohesin. G) Distribution of the number of bound DAPs in HOT loci with and without CTCF and Cohesin.

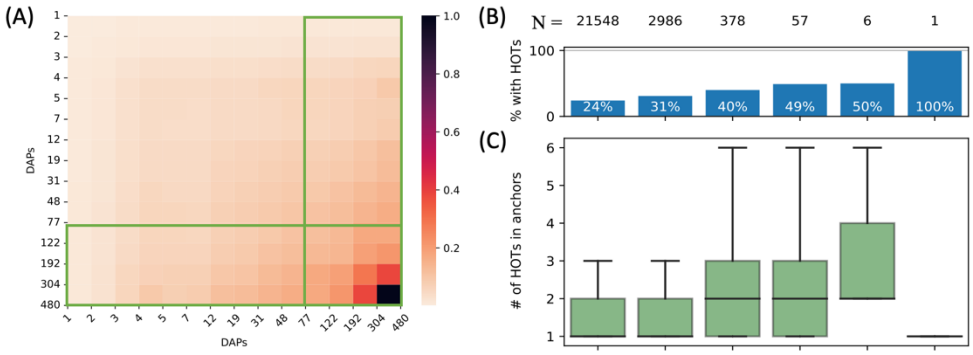

Figure 3. A) Densities of long-range Hi-C chromatin contacts between the DAP-bound loci. Each horizontal and vertical bin represents the loci with the number of bound DAPs between the edge values. Density values of each cell is normalized by the maximum value across all pairwise bins. Green boxes represent HOT loci. **B**) Fractions of genomic regions with Hi-C contacts that overlap HOT loci. X-axis is the number of Hi-C contacts. Top row gives the total number of genomic regions with a given number of Hi-C contacts. **C**) Number of HOT loci in the Hi-C contact regions.

specific GWAS traits *hematocrit*, *left ventricular diastolic function,* and *high-density lipoprotein cholesterol*, highlighting the important role of the HOT locus. Further, defining the loci with three or more chromatin contacts as 3D hubs, we observed that >50% of the 3D hubs contain at least one HOT locus (Fig 3B). The number of chromatin contacts in loci correlates with the number of overlapping HOT loci (rho=0.8, p-value=0.056; Pearson) (Fig 3C) and, by extension, with the average number of bound DAPs (rho=0.95, p-value=0.01; Pearson) (Fig S5). This observation is consistent with previous reports that much of the long-range 3D chromatin contacts form through the interactions of large protein complexes (Quinodoz et al., 2018). However, this explanation alone does not cover all of the HOT loci, since only 11% of the HOT loci overlap with at least one Hi-C contact region.

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

Next, using ChIP-seq signal intensity as a proxy for DAP binding affinity, we sought to analyze the patterns of binding affinities of DAP-DNA 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). Moreover, even when calculated DAP-wise, the average of the overall signal strength of every DAP correlates with the fraction of HOT loci with which the given DAP overlaps (rho=0.6, p-value<10^{-29}; Spearman) (Fig 4B), meaning that the overall average value of the signal intensity of a given DAP is largely driven by the ChIP-seq peaks that are located in HOT loci. While the overall average of the ChIP-seq signal intensity in HOT loci is greater when compared with 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 MED1 (Fig 4D,E,F). Whereas those with the opposite affinity (i.e., those having the strongest binding sites in non-HOT loci) are REST, RFX5, TP53, etc. (Fig 4D) 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, 1.7 times greater. These DAPs are CREB1, RFX1, ZNF687, RAD51, ZBTB40, and GPBP1L1

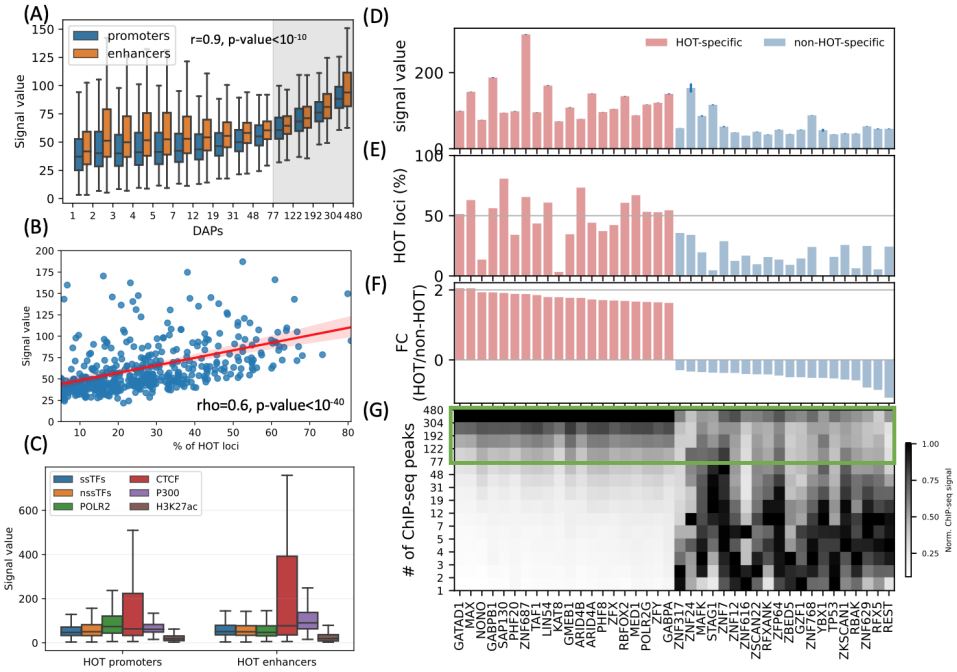

Figure 4. HOT regions induce strong ChIP-seq signals. **A**) Distribution of the signal values of the ChIP-seq peaks by the number of bound DAPs. Shaded region represents the HOT loci. **B**) Correlation of overall ChIP-seq signal strength and HOT overlaps. (x-axis) percentage of the HOT loci that a given DAP is bound to vs. (y-axis) the average overall signal values of the peaks of a given DAP. Each dot represents a DAP. **C**) Signal values of ssDAPs, nssDAPs (see the text for description), H3K27ac, CTCF, P300 peaks in HOT promoters and enhancers. **D,E,F,G**) 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. **D**) Overall ChIP-seq signal values of DAPs. **E**) Fractions of HOT loci that a given DAP is bound to. **F**) Fold change (log2) of the HOT and non-HOT loci ChIP-seq signals. **E**) Distribution of the average ChIP-seq signal in the loci binned by the number of bound DAPs. Horizontal bars represent the loci with the bound DAPs indicated by the values of the edges (y-axis).

So far, we 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 (Partridge et al., 2020; Ramaker et al., 2020; Xie et al., 2013). We asked if ssDAPs yield greater signal strength values than non-sequence-specific DAPs (nssDAPs). To test this hypothesis, we classified the DAPs into these two categories using the definitions provided by (Lambert et al., 2018) and categorized the ChIP-seq signal values into these two groups. While statistically significant (p-value<10^{-3}, Mann-Whitney U test), the differences in the average signals of ssDAPs and nssDAPs in both HOT enhancers and HOT promoters are small (Fig 4C). 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

three times 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 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 evolutionary conservation of the HOT loci using phastCons scores generated using an alignment of 46 vertebrate species (Siepel et al., 2005). 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) 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 1.7 times more conserved than the regular enhancers in HepG2 (Fig 5B). 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 S3). Among the genes are those which have been linked to diseases and other phenotypes. For example, DNAJC1 (Michailidou et al. 2017) and OLA1 (which interacts with BRCA1) have been linked to breast cancer in cancer GWAS studies (Liu et al. 2020) and AUTS2 (Biel et al. 2022) and SOX5 (Schanze et al. 2013) have been linked to predisposition to neurological conditions, such as autism spectrum disorder (ASD), 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 (Daigle et al. 2018).

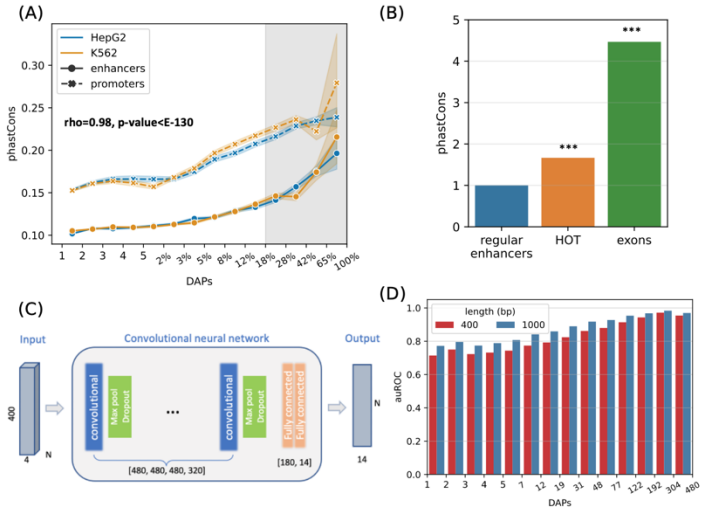

Figure 5. Sequence features of HOT loci. **A)** Distribution of conservation scores in loci. Logarithmic part of the bins are expressed in terms of the percentages of loci that each bin covers, averaged over two cell lines. Correlation value is Pearson. Shaded region represent HOT loci. **B**) phastCons conservation scores of regular enhancer, HOT loci and exon regions. The values are normalized by the average scores of regular enhancers. **C**) The architecture of the deep learning model trained to recognize the HOT loci. Numbers on the input and output layers indicate the dimensions of the input and output data. Numbers at the curly bracket indicate the number of filters in convolutional layers. The architecture of the model is detailed in Supp. Methods, TableS4. **D)** Prediction performance of the DL model on the loci of different bins of bound DAPs in HepG2.

CpG islands have been postulated to serve as permissive TF binding platforms (Deaton and Bird, 2011; Pachano et al., 2021) and this has been listed as one of the possible reasons for the existence of HOT loci in a previous study (Wreczycka et al., 2019). To test this hypothesis, we extracted the overlap rates of all DAP-bound loci with CpG islands (see Supp. Methods for details). 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), only 12% of HOT enhancers contained a CpG island whereas for the HOT promoters this fraction was 83%, suggesting that CpG islands alone do not explain HOT enhancer loci while accounting for most HOT promoter loci (Fig S6A). Similarly, the average GC content is strongly correlated with the number of bound DAPs (rho=0.89, p-value<10^{-4}; Pearson, Fig S6B), 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 S6C), 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, but are not limited to G-quadruplex, R-loops, methylation patterns, etc., suggesting that each of them can partially explain the phenomenon of the HOT loci (Moorman et al., 2006; Teytelman et al., 2013; Wreczycka et al., 2019). Still, one of the central questions remains as to 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 built a deep convolutional neural network model comprising six convolutional layers and asked if the loci binned according to the number of bound DAPs can be predicted using the DNA sequences alone (Fig 5C). We did this analysis using 400 bp loci as data points, as well as including the flanking regions and

extending them to 1 kb. The model predicted the HOT loci with exceptionally high accuracy (average auROC of 0.96) with the auROC values correlating with the number of bound DAPs (rho=0.8, p-value<0.001) (Fig 5D, Fig S6D). The classification accuracy gains in extending the HOT loci to 1 kb were insignificant (p-value>0.5, Mann-Whitney U test), suggesting that the core 400 bp locus encodes most of the identifying information for the HOT loci. This analysis confidently concludes that the mechanisms of HOT locus formation are encoded in their DNA sequences.

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 in which they participate. HOT loci were previously linked to highly expressed genes (Wreczycka et al. 2019). 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; Fig 6A). In HepG2, the average expression level of the target genes of promoters with at least one DAP bound is 1.7 times 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 4 times 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 4.5 times 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. Fig S11), 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). According to this definition, we found that 64% of HK genes are regulated by a HOT promoter and only 30% are regulated by regular promoters (Fig 6B). 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 6B).

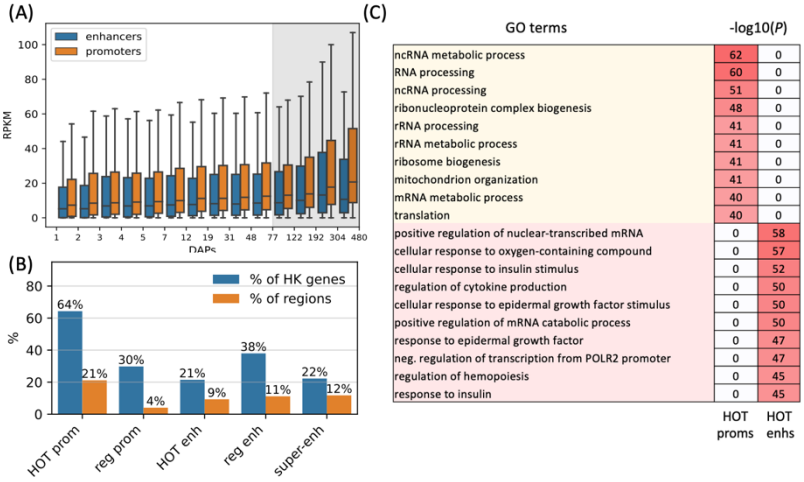

Figure 6. HOT promoters are ubiquitous and HOT enhancers are tissue-specific. **A)** Boxplot distribution of the expression levels of the target genes of HepG2 loci by the number of bound DAPs **B)** Fractions of housekeeping genes regulated by the given category of loci (blue). Fractions of the loci which regulate the housekeeping genes (orange) **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 GREAT tool.

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), 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, Fig S7). 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. 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 shows 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 are 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 have analyzed so far demonstrated similar patterns of GC contents, target gene expressions, ChIP-seq signal values etc.; when compared with 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. 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 to alleviate the difference in available DAPs.

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 those of the differentiated cell lines. These overlap levels were much higher than the randomly selected DAPs matching the H1 set (30%, Fig 7C).

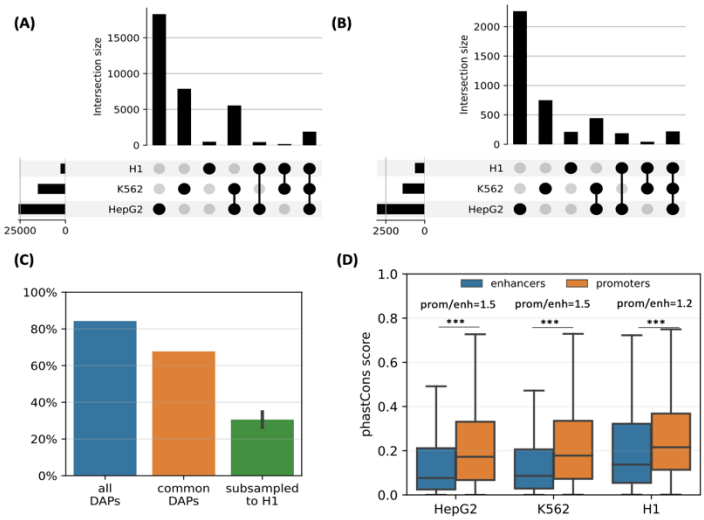

Figure 7. H1 HOT loci **A**) Overlaps between the HOT loci of three cell lines. **B**) Overlaps between the HOT loci of cell lines defined using the set of DAPs available in all three cell lines. **C**) Fractions of H1 HOT loci overlapping that of the HepG2 and K562 using complete set of DAPs, common 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 are at the top of every cell line's group.

Average phastCons scores of the developmental HOT loci are 1.3 times higher than HepG2 and K562 HOT loci (p-value<10^{-10}, Mann-Whitney U test, Fig 7D). We 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. Additionally, 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 (Domené et al. 2013).

GO enrichment analysis showed that H1 HOT promoters, similarly to the other cell lines, regulate the basic housekeeping processes (Fig S8) 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 looked to determine whether these loci are enriched in phenotypically causal variants. First, we analyzed the density of common variants extracted from the gnomAD database (Karczewski et al., 2020) (filtered with MAF>5%). We observed that HOT

enhancers and HOT promoters are depleted in INDELs (4.7 and 4.1 variants per 1 kb, respectively), compared with the regular enhancers and regular promoters (5.5 and 6.2 variants per 1 kb, 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 (1.14-fold and 1.07-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 (Vierstra et al., 2020).

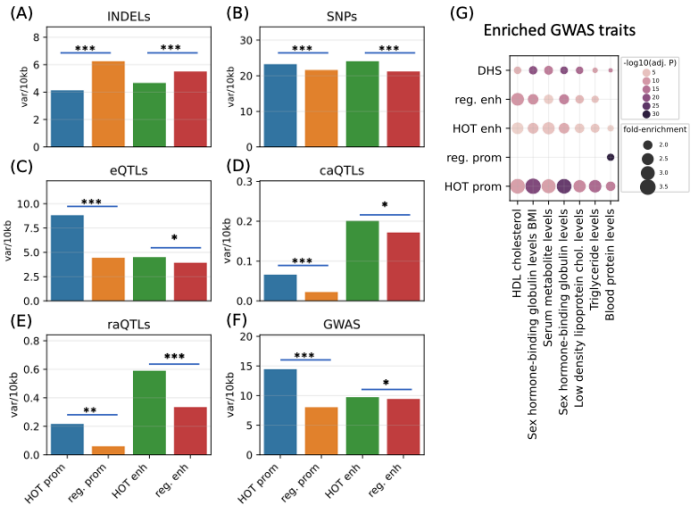

Figure 8. Densities of variants (**A**) common INDELs (MAF>5%), (**B**) common SNPs (MAF>5%), (**C**) eQTLs, (**D**) caQTLs (**E**) raQTLs, and (**F**) GWAS (with r2>0.8 LD variants) variants in HOT loci and regular promoters and enhancers. **G**) Enriched GWAS traits in HOT enhancers and promoters.

The eQTLs, on the other hand, are 2.0-fold enriched in HOT promoters compared to the regular promoters (p-value<10E-21, Mann-Whitney U test), while HOT enhancers are only moderately enriched in eQTLs compared with the regular enhancers (1.15x, p-value>0.05, Mann-Whitney U test; Fig 8C). eQTL enrichment in HOT promoters and regular promoters (compared with 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 (GTEx Consortium, 2015).

Unlike the eQTL analysis, we observed that the chromatin accessibility QTLs (caQTLs) are dramatically enriched in the overall enhancer regions (HOT and regular) compared with the promoters (HOT and regular) (4.1 times, 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 (Currin et al., 2021), 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 promoter, the HOT promoters are 3.0-fold 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 (1.2-times, p-value=0.22, Mann-Whitney U test).

A similar enrichment pattern is seen for the reporter array QTLs (raQTLs van Arensbergen et al., 2019), with respect to the overall (HOT and regular) promoter and enhancer regions, with 3.3-fold enrichment in enhancers (p-value<10^{-10}, Mann-Whitney U test, Fig 8E). In contrast, within-promoters and within-enhancers enrichments show that the enrichment in HOT promoters is more pronounced than the HOT enhancers (3.6-fold and 1.8-fold, 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 (van Arensbergen et al. 2019).

Combined, all three QTL datasets show a pronounced enrichment in HOT promoters compared with 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 (1.8-fold, 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 of GWAS traits SNPs using the combined DNaseI Hypersensitivity Sites (DHSs) of Roadmap Epigenome cell lines as a background (see Supp. Methods). Filtering the traits with significant enrichment in HOT loci (p-value<0.001, Binomial test, Bonferroni corrected, see Supp. Methods) left seven traits, all of which 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 (6/7) traits, the overall enrichment values in HOT enhancers are 1.3 times greater compared to the regular enhancers. The fold-increase is even greater (1.5 times) 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 S9). This analysis demonstrates that the HOT loci are important for phenotypic homeostasis.

Discussion

HOT loci have been noticed and studied in different species since the first compilation of the ChIP-seq datasets (Gerstein et al., 2010; Kvon et al., 2012; modENCODE Consortium et al., 2010; Moorman et al., 2006; Xie et al., 2013; Yip et al., 2012). Up until recently, most of the studies have extensively studied the reasons by 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 that could enable elevated read mapping rates (Moorman et al., 2006; Teytelman et al., 2013; Wreczycka et al., 2019). 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 has 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 (Partridge et al., 2020; Ramaker et al., 2020).

Previous studies have used different metrics to define the HOT loci. For example, (Wreczycka et al., 2019) used the 99th percentile of the density of TFBSs for a 400 bp sliding window, (Ramaker et al., 2020) used the window length of 2 kbs and required >25% of TFs to be mapped, and (Partridge et al., 2020) used loci with >70 chromatin-associated proteins in 2 kbs window. These heterogeneous definitions, however, fail to appreciate that the histogram of loci binned by the number of harbored TFBSs represent 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 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 clearer that the HOT loci are located in 3D hubs (Fig 3A).

Previous studies have extensively mapped 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 (Moorman et al., 2006; Ramaker et al., 2020), and proposed that perhaps these driver TFs initiated the formation of HOT loci, similar to how the pioneer factors function. More importantly, additional studies have concluded 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 (Partridge et al., 2020; Ramaker et al., 2020; Vierstra et al., 2020; White et al., 2021). 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 and further implies that the indirectly bound loci also carry shared sequence features, perhaps other than the binding motifs or weak motifs that are not detected using the traditional PWM-based tools of motif detection.

Another model that has been increasingly attributed to the concept of the formation and maintenance of long-range 3D chromatin interactions involves phase-separated condensates (Ahn et al., 2021; Feric and Misteli, 2022; Lee et al., 2022; Nair et al., 2019). Some enhancers (dubbed MegaTrans enhancers) were shown to drive the formation of large chromosomal assemblies involving a high concentration of TFs (Nair et al., 2019). In general, it has been increasingly appreciated that condensates ubiquitously attract and activate enhancers (Boija et al., 2018; Shrinivas et al., 2019; Wei et al., 2020). The property of the condensates, which is of special interest to our 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 (Zhang et al., 2021). We speculate 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 sequencing through ChIP-seq experiments.

Condensates formed at different foci in the nucleus have been shown to acquire physiochemical properties depending on their functions (Feric and Misteli 2022). For instance, the sizes of the transcriptional condensates have been shown to be regulated by the concentration of RNA molecules contained in them (Sharp et al. 2022). 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 (Henninger et al. 2021). 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 S10), further supporting the condensate hypothesis of the HOT loci.

With the assumption that condensates form, 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 evolution of more precise protocols such as CUT&RUN, micro-C, and single-cell versions of ChIP-seq, ATAC-seq combined with larger databases of experimentally verified condensate studies, we will be able to develop 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.

Acknowledgements

This work utilized the computational resources of the HPC Biowulf cluster (http://hpc.nih.gov) at National Institutes of Health (NIH). This research was supported by the Intramural Research Program (IRP) of the National Library of Medicine (NLM), NIH. We thank Dorothy L. Buchhagen for her help with critical reading.

References

Ahn, J.H., Davis, E.S., Daugird, T.A., Zhao, S., Quiroga, I.Y., Uryu, H., Li, J., Storey, A.J., Tsai, Y.-H., Keeley, D.P., et al. (2021). Phase separation drives aberrant chromatin looping and cancer development. Nature *595*, 591–595.

van Arensbergen, J., Pagie, L., FitzPatrick, V.D., de Haas, M., Baltissen, M.P., Comoglio, F., van der Weide, R.H., Teunissen, H., Võsa, U., Franke, L., et al. (2019). High-throughput identification of human SNPs affecting regulatory element activity. Nat. Genet. *51*, 1160–1169.

Boija, A., Klein, I.A., Sabari, B.R., Dall'Agnese, A., Coffey, E.L., Zamudio, A.V., Li, C.H., Shrinivas, K., Manteiga, J.C., Hannett, N.M., et al. (2018). Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. Cell *175*, 1842-1855.e16.

Currin, K.W., Erdos, M.R., Narisu, N., Rai, V., Vadlamudi, S., Perrin, H.J., Idol, J.R., Yan, T., Albanus, R.D., Broadaway, K.A., et al. (2021). Genetic effects on liver chromatin accessibility identify disease regulatory variants. Am. J. Hum. Genet. *108*, 1169–1189.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., et al. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. *46*, D794–D801.

Deaton, A.M., and Bird, A. (2011). CpG islands and the regulation of transcription. Genes Dev. *25*, 1010–1022.

ENCODE Project Consortium, Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature *583*, 699–710.

Feric, M., and Misteli, T. (2021). Phase separation in genome organization across evolution. Trends Cell Biol. *31*, 671–685.

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

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

Gerstein, M.B., Lu, Z.J., Van Nostrand, E.L., Cheng, C., Arshinoff, B.I., Liu, T., Yip, K.Y., Robilotto, R., Rechtsteiner, A., Ikegami, K., et al. (2010). Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science *330*, 1775–1787.

Gorkin, D.U., Barozzi, I., Zhao, Y., Zhang, Y., Huang, H., Lee, A.Y., Li, B., Chiou, J., Wildberg, A., Ding, B., et al. (2020). An atlas of dynamic chromatin landscapes in mouse fetal development. Nature *583*, 744–751.

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

Henninger, J.E., Oksuz, O., Shrinivas, K., Sagi, I., LeRoy, G., Zheng, M.M., Andrews, J.O., Zamudio, A.V., Lazaris, C., Hannett, N.M., et al. (2021). RNA-Mediated Feedback Control of Transcriptional Condensates. Cell *184*, 207-225.e24.

Hounkpe, B.W., Chenou, F., de Lima, F., and De Paula, E.V. (2021). 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.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.

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

Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., and Weirauch, M.T. (2018). The human transcription factors. Cell *172*, 650–665.

Lee, R., Kang, M.-K., Kim, Y.-J., Yang, B., Shim, H., Kim, S., Kim, K., Yang, C.M., Min, B.-G., Jung, W.-J., et al. (2022). CTCF-mediated chromatin looping provides a topological framework for the formation of phase-separated transcriptional condensates. Nucleic Acids Res. *50*, 207–226.

modENCODE Consortium, Roy, S., Ernst, J., Kharchenko, P.V., Kheradpour, P., Negre, N., Eaton, M.L., Landolin, J.M., Bristow, C.A., Ma, L., et al. (2010). Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science *330*, 1787–1797.

Moorman, C., Sun, L.V., Wang, J., de Wit, E., Talhout, W., Ward, L.D., Greil, F., Lu, X.-J., White, K.P., Bussemaker, H.J., et al. (2006). Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster. Proc Natl Acad Sci USA *103*, 12027–12032.

Nair, S.J., Yang, L., Meluzzi, D., Oh, S., Yang, F., Friedman, M.J., Wang, S., Suter, T., Alshareedah, I., Gamliel, A., et al. (2019). Phase separation of ligand-activated enhancers licenses cooperative chromosomal enhancer assembly. Nat. Struct. Mol. Biol. *26*, 193–203.

Pachano, T., Sánchez-Gaya, V., Ealo, T., Mariner-Faulí, M., Bleckwehl, T., Asenjo, H.G., Respuela, P., Cruz-Molina, S., Muñoz-San Martín, M., Haro, E., et al. (2021). Orphan CpG islands amplify poised enhancer regulatory activity and determine target gene responsiveness. Nat. Genet. *53*, 1036–1049.

Palmer, D., Fabris, F., Doherty, A., Freitas, A.A., and de Magalhães, J.P. (2021). Ageing transcriptome meta-analysis reveals similarities and differences between key mammalian tissues. Aging (Albany NY) *13*, 3313–3341.

Partridge, E.C., Chhetri, S.B., Prokop, J.W., Ramaker, R.C., Jansen, C.S., Goh, S.-T., Mackiewicz, M., Newberry, K.M., Brandsmeier, L.A., Meadows, S.K., et al. (2020). Occupancy maps of 208 chromatin-associated proteins in one human cell type. Nature *583*, 720–728.

Quinodoz, S.A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J.M., Detmar, E., Lai, M.M., Shishkin, A.A., Bhat, P., Takei, Y., et al. (2018). Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. Cell *174*, 744-757.e24.
Ramaker, R.C., Hardigan, A.A., Goh, S.-T., Partridge, E.C., Wold, B., Cooper, S.J., and Myers, R.M. (2020). Dissecting the regulatory activity and sequence content of loci with exceptional numbers of transcription factor associations. Genome Res. *30*, 939–950.
Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.
Serfling, E., Jasin, M., and Schaffner, W. (1985). Enhancers and eukaryotic gene transcription. Trends Genet. *1*, 224–230.
Sethi, A., Gu, M., Gumusgoz, E., Chan, L., Yan, K.-K., Rozowsky, J., Barozzi, I., Afzal, V., Akiyama, J.A., Plajzer-Frick, I., et al. (2020). Supervised enhancer prediction with epigenetic pattern recognition and targeted validation. Nat. Methods *17*, 807–814.
Sharp, P.A., Chakraborty, A.K., Henninger, J.E., and Young, R.A. (2022). RNA in formation and regulation of transcriptional condensates. RNA *28*, 52–57.
Shrinivas, K., Sabari, B.R., Coffey, E.L., Klein, I.A., Boija, A., Zamudio, A.V., Schuijers, J., Hannett, N.M., Sharp, P.A., Young, R.A., et al. (2019). Enhancer Features that Drive Formation of Transcriptional Condensates. Mol. Cell *75*, 549-561.e7.
Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034–1050.
Teytelman, L., Thurtle, D.M., Rine, J., and van Oudenaarden, A. (2013). Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. Proc Natl Acad Sci USA *110*, 18602–18607.
Vierstra, J., Lazar, J., Sandstrom, R., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Haugen, E., et al. (2020). Global reference mapping of human transcription factor footprints. Nature *583*, 729–736.
Wei, M.-T., Chang, Y.-C., Shimobayashi, S.F., Shin, Y., Strom, A.R., and Brangwynne, C.P. (2020). Nucleated transcriptional condensates amplify gene expression. Nat. Cell Biol. *22*, 1187–1196.
White, S.M., Snyder, M.P., and Yi, C. (2021). 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.
Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307–319.
Wreczycka, K., Franke, V., Uyar, B., Wurmus, R., Bulut, S., Tursun, B., and Akalin, A. (2019). HOT or not: examining the basis of high-occupancy target regions. Nucleic Acids Res. *47*, 5735–5745.
Xie, D., Boyle, A.P., Wu, L., Zhai, J., Kawli, T., and Snyder, M. (2013). Dynamic trans-acting factor colocalization in human cells. Cell *155*, 713–724.
Yip, K.Y., Cheng, C., Bhardwaj, N., Brown, J.B., Leng, J., Kundaje, A., Rozowsky, J., Birney, E., Bickel, P., Snyder, M., et al. (2012). Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. Genome Biol. *13*, R48.
Zamudio, A.V., Dall'Agnese, A., Henninger, J.E., Manteiga, J.C., Afeyan, L.K., Hannett, N.M., Coffey, E.L., Li, C.H., Oksuz, O., Sabari, B.R., et al. (2019). Mediator condensates localize signaling factors to key cell identity genes. Mol. Cell *76*, 753-766.e6.
Zhang, Y., Brown, K., Yu, Y., Ibrahim, Z., Zandian, M., Xuan, H., Ingersoll, S., Lee, T., Ebmeier, C.C., Liu, J., et al. (2021). 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.