

Functional characteristics and computational model of abundant hyperactive loci in the human genome

Reviewed Preprint

v2 • October 8, 2024

Revised by authors

Reviewed Preprint

v1 • May 9, 2024

Sanjarbek Hudaiberdiev ✉, Ivan Ovcharenko ✉

National Institute for Biotechnology and Information, National Library of Medicine, National Institutes of Health.
Bethesda, United States

 https://en.wikipedia.org/wiki/Open_access

 Copyright information

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 often 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.

eLife Assessment

This **valuable** study explores the sequence characteristics and conservation of high-occupancy target loci, regions in the human genome such as promoters and enhancers that are bound by a multitude of transcription factors. The computational analyses presented in this study are **solid**. This study would be a helpful resource for researchers performing ChIP-seq based analyses of transcription factor binding.

<https://doi.org/10.7554/eLife.95170.2.sa4>

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 in terms of detectible binding motifs of corresponding motifs. 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 that 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 concluded that these regions are highly functionally 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 proteins including TFs, cofactors, and chromatin regulators which they called 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

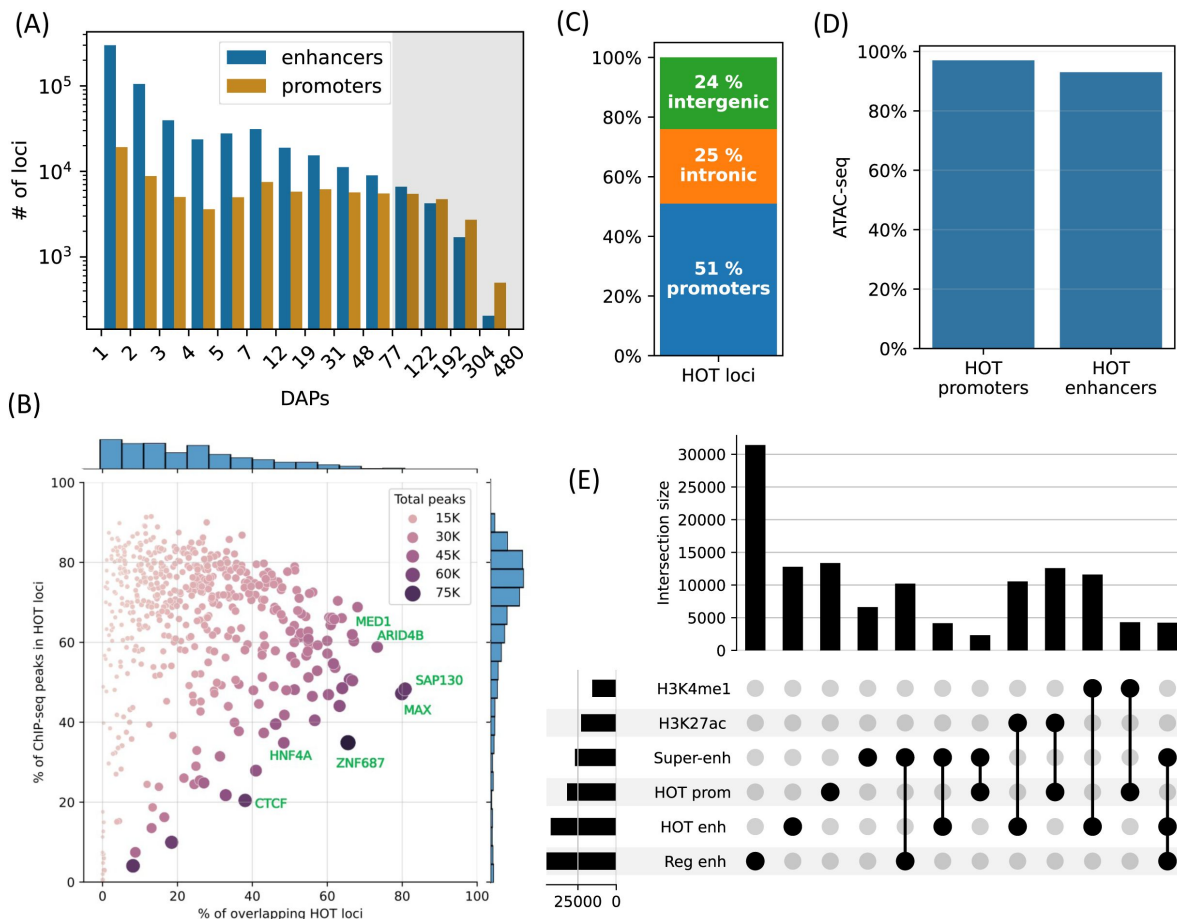


Figure 1.

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 1. Methods). The shaded region represents the HOT loci. **B)** Prevalence of DAPs in HOT loci. Each dot represents a DAP. X-axis: percentage of HOT loci in which DAP is present (e.g. MAX is present in 80% of HOT loci). Y-axis: percentage of total peaks of DAPs that are located in HOT loci (e.g. 45% of all the ChIP-seq peaks of MAX is located in the 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. Horizontal bars on bottom left represent the total number of loci of the corresponding class of loci. All of the visualized data is generated from the HepG2 cell line.

	Bin edges (n=15)														
HepG2	1	2	3	4	5	7	12	19	31	48	77	122	192	304	480
K562	1	2	3	4	5	7	11	16	24	37	55	82	123	184	275
H1	1	2	3	4	5	6	7	8	10	12	15	18	22	26	32
	linear growth (n=4)					logarithmic growth (n=10)									

Table 1.

Schema of classifying loci according to the number of bound DAPs.

The initial 4 bins are loci bound by DAPs increasing linearly from 1 to 5 (gray fields). The remaining 10 bins are defined by edge values increasing on a logarithmic scale from 5 to the maximum number of available DAPs in each cell line (orange and red fields) using the Numpy formula `np.logspace(np.log10(5), np.log10(max_tfs), 11, dtype=int)`. HOT loci correspond to the last 5 bin edges (red fields).

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.¹⁹ and Boyle et al.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. 2013²² and a set of regular enhancers (Methods). 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 where each HOT locus is represented by a binary (presence/absence) vector of length equal to the total number of DAPs analyzed. 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**, Fig S6B), and the PC1-PC2 combination also separates the p300-bound HOT loci (**Fig 2B**, 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 (Fig S8B). *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 $p\text{-value} < 10^{-16}$, 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.

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 ($p\text{-value} < 10^{-100}$; 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 ($p\text{-value} < 10^{-100}$; 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 ($p\text{-value} < 10^{-20}$, 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 ($\rho = 0.3$, $p\text{-value} < 10^{-100}$, Spearman correlation). Next, we overlaid 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 ($p\text{-value} < 10^{-230}$, 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 ($p\text{-value} < 10^{-230}$, 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 (i.e. loci that serve as anchors to multiple loops) correlates with the number of bound DAPs ($\rho = 0.84$ $p\text{-value} < 10^{-4}$; 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**). Of the loop anchor regions with >3 overlapping loops, more than half 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.

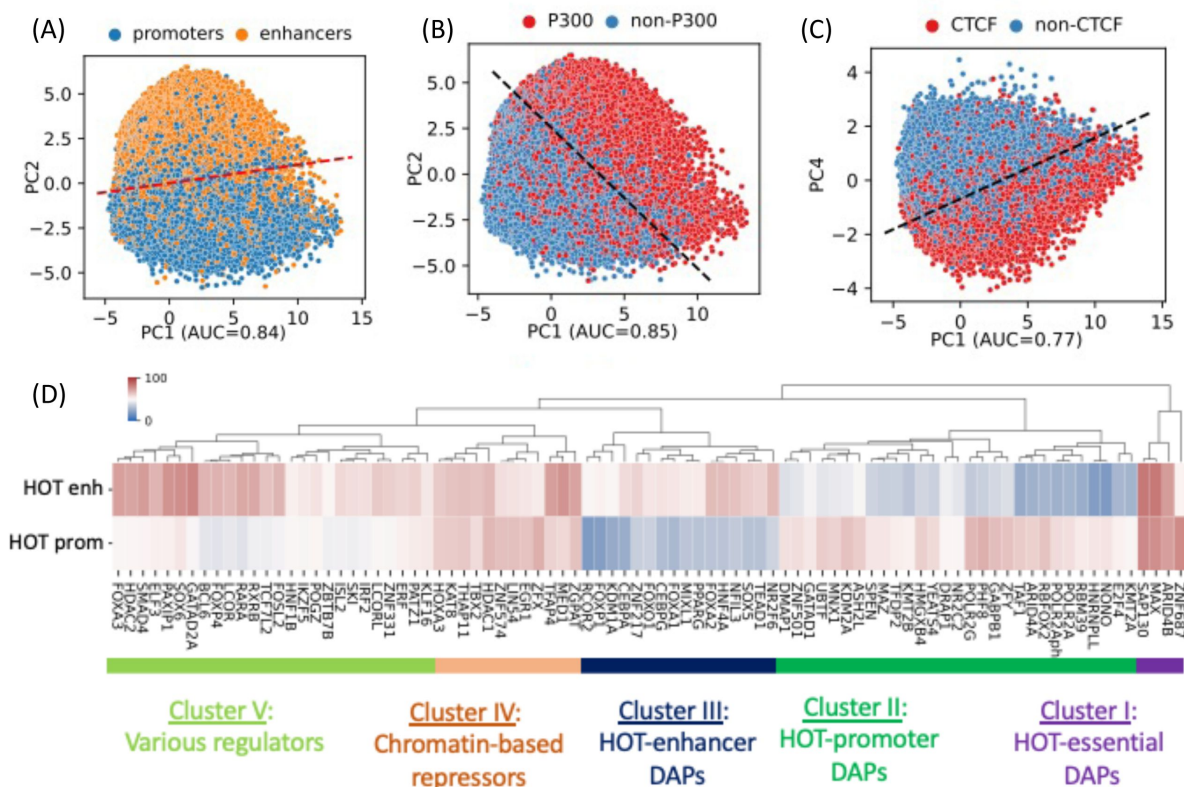


Figure 2.

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 results of logistic regression. **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.

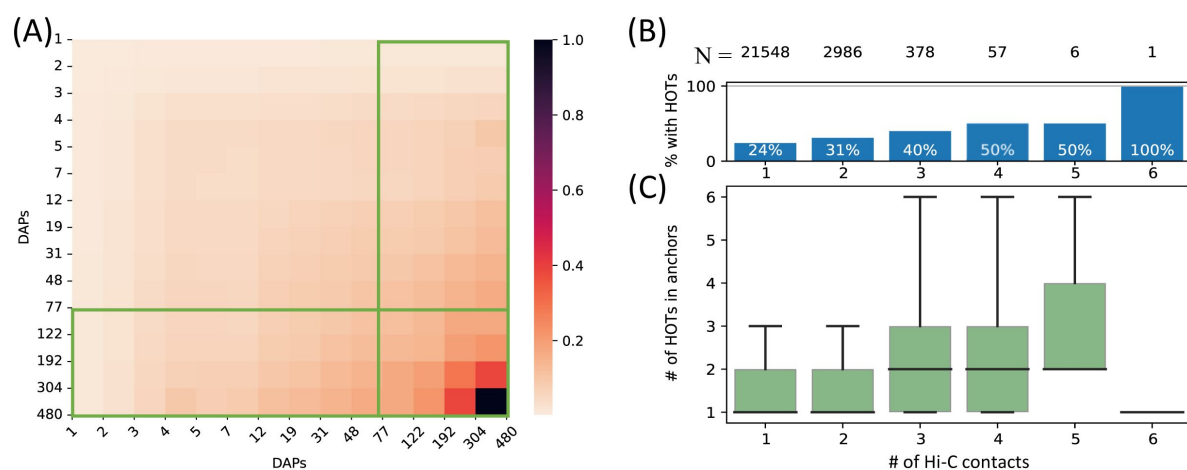


Figure 3.

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 Hi-C 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.

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

Next, we sought to analyze the patterns of ChIP-seq signal values at HOT loci, as a metric for overall DAP occupancy at genomic loci. We observed that the overall signals of DAPs correlate with the total number of colocalizing DAPs (**Fig 4A**^{19,20,27}, $\rho=0.97$, $p\text{-value}<10^{-10}$; 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 ($\rho=0.6$, $p\text{-value}<10^{-29}$; Spearman correlation. **Figure 4B**²⁸), 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**²⁸). 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**²⁸). 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 1.9x greater ($p\text{-value}<10^{-100}$, Mann-Whitney U test). These DAPs are CREB1, RFX1, ZNF687, RAD51, ZBTB40 and GPBP1L1 (Supplemental Methods 1.3, Fig S18,S19).

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, where the TFs are classified by curation through extensive literature review and supported by annotations such as the presence of DNA-binding domains and validated binding motifs. Based on this classification, we categorized the ChIP-seq signal values into these two groups. While statistically significant ($p\text{-value}<0.001$, 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**²⁸). 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 species²⁹. The average conservation scores of the DAP-bound loci are in strong correlation with the number of bound DAPs ($\rho=0.98$, $p\text{-value}<10^{-130}$; Spearman correlation) indicating that the negative selection exerted on HOT loci are proportional to the number of bound DAPs (**Fig 5A**²⁹). With 120 DAPs per locus on average, these HOT regions are 1.7x more conserved than the regular enhancers in HepG2 (**Fig 5B**²⁹). 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

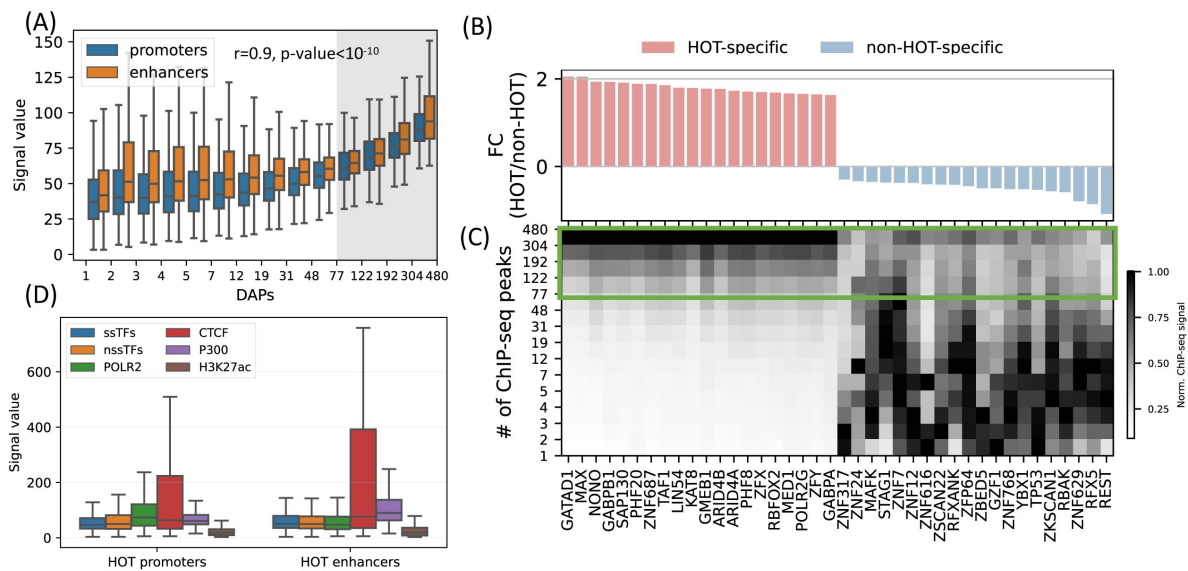


Figure 4.

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 (\log_2) 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.

them are those which have been linked to diseases and other phenotypes. For example, DNJC1³⁰ and OLA1 (which interacts with BRCA1) have been linked to breast cancer in cancer GWAS studies³¹. Whereas AUTS2³² and SOX5³³ have been linked to predisposition to neurological conditions such as Autism spectrum disorder, intellectual disability, and neurodevelopmental disorder. Of these genes, ARID1A, AUTS2, DNJC1, OLA1, SOX5, and ZFPM2 have been reported to have strong activities in the Allen Mouse Brain Atlas³⁴.

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 study¹⁸. To test this hypothesis, we extracted the overlap rates of all DAP-bound loci with CpG islands (Methods). While the overall fraction of loci that overlap CpG islands correlates strongly with the number of bound DAPs ($\rho=0.7$, $p\text{-value}=0.001$; 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 ($\rho=0.89$, $p\text{-value}<10^{-4}$; Pearson correlation, Fig S11B), with the average GC content of 64% and 51% in HOT promoters and HOT enhancers respectively ($p\text{-value}<10^{-100}$, 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 ($\rho=-0.9$, $p\text{-value}<10^{-5}$; 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 (seq-SVM)³⁷. Using the sequence features we trained SVM models with linear kernel function (feature-SVM). We carried out the classification experiments using the following control (i.e. negative) 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 (Supplemental Methods 6.1.1). When averaged over cell lines and control sets, CNN, seq-SVM and feature-SVM models yielded auROC values of 0.91, 0.86, and 0.78 respectively, suggesting that CNNs capture the motif grammar of the HOT loci better than the compared models (Fig 5C). The superiority of sequence-based models over feature-based classification by a factor of 1.3x (or 17%), suggests 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 contents. 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.

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).

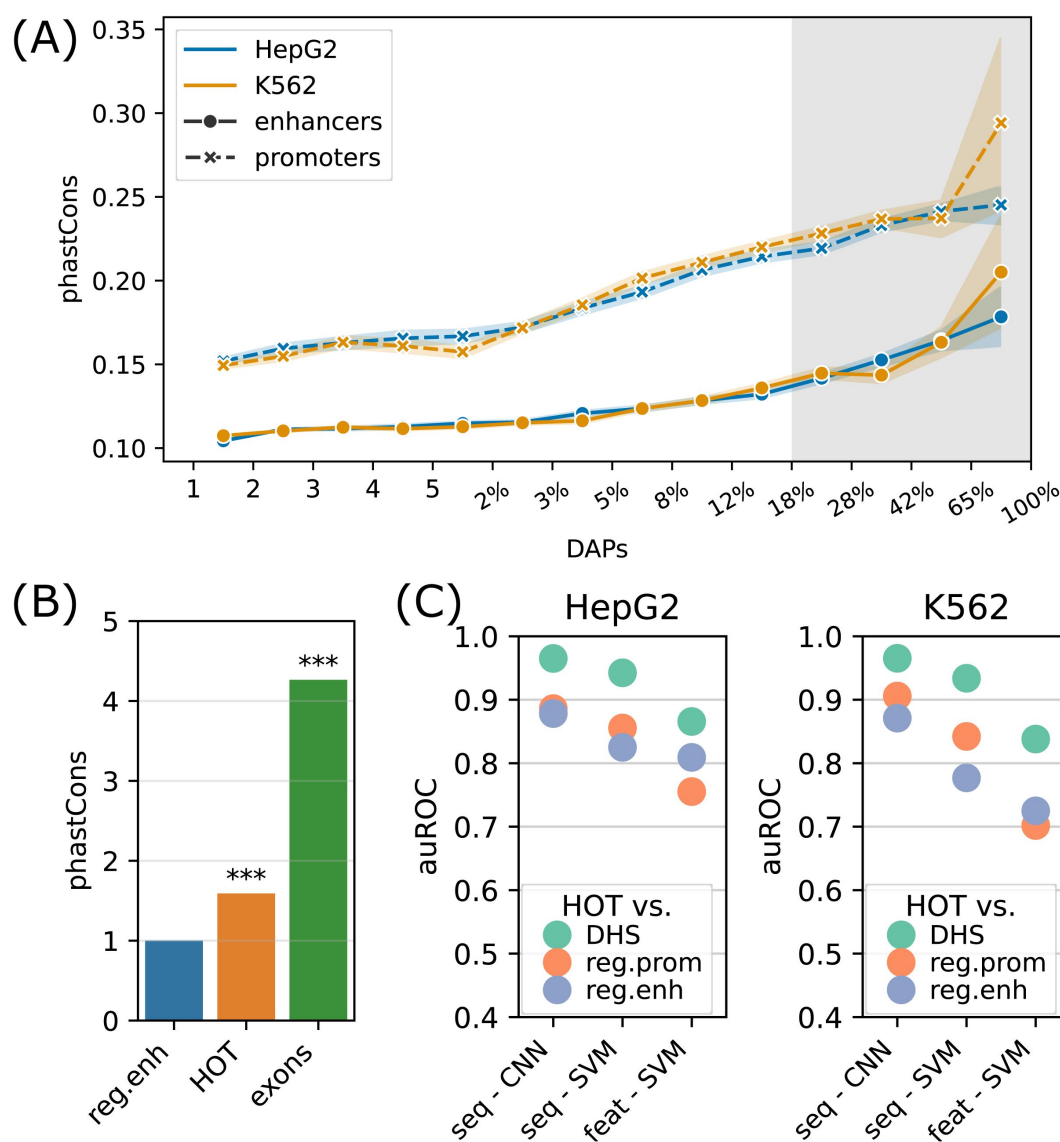


Figure 5.

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 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) 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). Starting with “feat-” are methods where all sequence features are used (GC, CpG, GpC, CpG island).

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; $\rho=0.56$, $p\text{-value}<10^{-10}$; Spearman correlation; Fig S15A). 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 ($p\text{-value}<10^{-12}$, 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 ($p\text{-value}<10^{-30}$, 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 ($\rho=-0.9$; $p\text{-value}<10^{-6}$; 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. $p\text{-value}<10^{-18}$, 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, $p\text{-value}<10^{-100}$, 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).

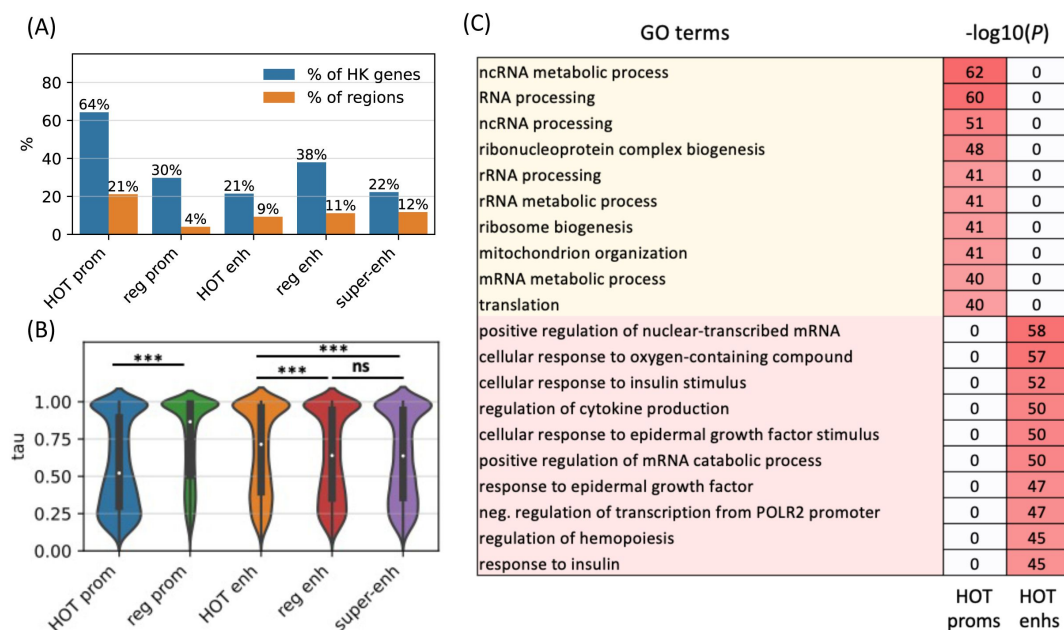


Figure 6.

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.

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 ($p\text{-value}<10^{-10}$, 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 promoters⁴⁰.

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 database⁴¹ (filtered with $\text{MAF}>5\%$). 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, $p\text{-value}<10^{-4}$ 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, $p\text{-values}<10^{-20}$ and $<10^{-100}$, respectively, Mann-Whitney U test; **Fig 8B**). This elevation of common variants in HOT

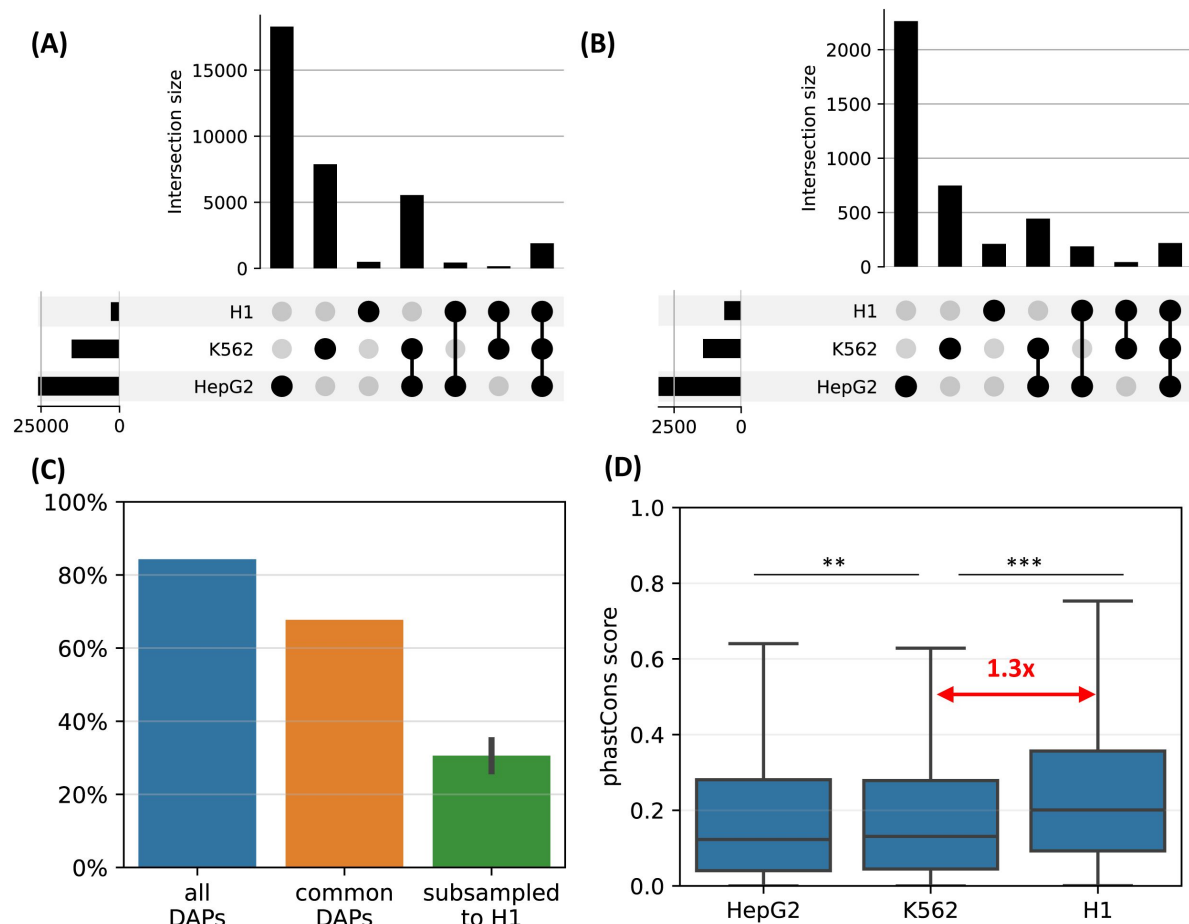


Figure 7.

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.

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<0.001, 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.

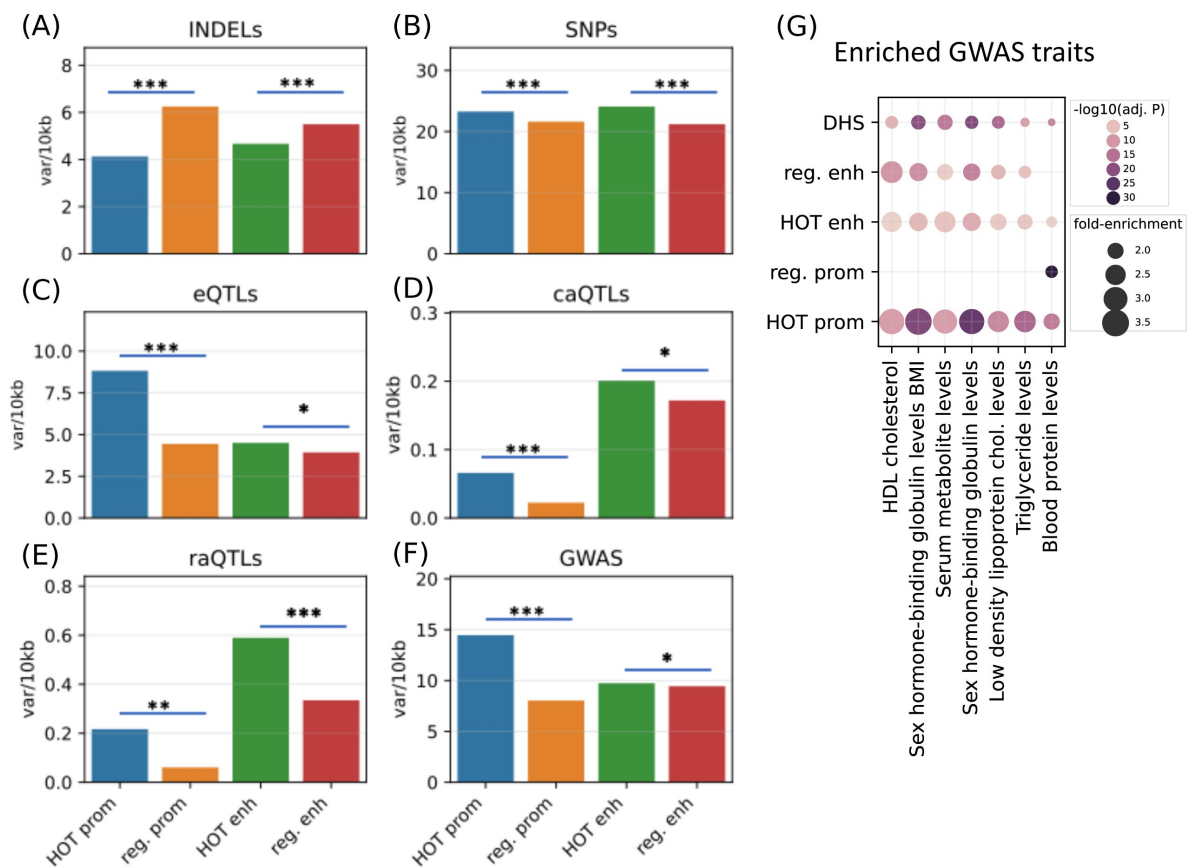


Figure 8.

Densities of variants A) common INDELs (MAF>5%)

B) common SNPs (MAF>5%) **C)** eQTLs, **D)** caQTLs **E)** raQTLs, and **F)** GWAS and LD ($r^2>0.8$) 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.

Transcriptional condensates as a model for explaining the HOT regions

Recent studies on phase-separated condensates have established that condensates are ubiquitous in cells and play crucial roles in gene regulation through transcriptional condensates ^{46–49}. We postulated that the HOT loci could be explainable if it can be shown that the HOT loci demonstrate a high propensity for the formation of transcriptional condensates. The hallmarks of transcriptional condensates include (not limited to) scaffolding proteins that undergo liquid-to-liquid phase separation (LLPS), DNA and RNA molecules, and intrinsically disordered (IDR) proteins. We sought to analyze whether these properties can be attributed to the HOT loci.

First, using CD-CODE database ⁵⁰ we annotated 24% of the DAPs used in the analysis as LLPS-inducing proteins (**Fig 9A**). We observed that LLPS proteins are uniformly distributed in HOT loci (**Fig 9B**). We calculated a null distribution by randomly shuffling the ChIP-seq peaks in HOT loci 10 times, which resulted in a near-zero fraction of LLPS proteins located in >45% of the HOT loci, where the actual observed fraction is 23% (average of the last two bins in **Fig 9B**), strongly suggesting an overrepresentation. Moreover, LLPS proteins yield significantly stronger ChIP-seq signals compared to the rest of the DAPs (**Fig 9C**, p-value=0.002, t-test), and contain a higher percentage of predicted IDR regions (**Fig 9D**, 30% vs. 26%, p-value=0.01, t-test).

Next, we sought to quantify the RNA-related interactions in HOT loci. First, we used ENCODE's set of ChIP-seq datasets extracted using RNA-binding proteins (RBP) and observed that RBPs are more enriched in HOT loci compared to the rest of the DAPs in terms of fold-increase using ATAC-seq regions as background (**Fig 9E**, 1.5 vs. 1.3 in log₂(FC), p-value=0.04, t-test). Second, we quantified the level of transcription using FANTOM, PINTS ⁵¹ (a modern tool for annotating eRNAs combining multiple types of RNA sequencing assays), and CAGE-seq peaks. We observed that all three types of annotations demonstrate high overrepresentation in HOT loci compared to regular promoters and enhancers by a factor of 2.7x on average (**Fig 9F**). Lastly, we used eCLIP datasets of 103 RBSs from the ENCODE Project and calculated the levels of RBP-RNA interactions. We observed that the difference in the levels of eCLIP signals in HOT loci and coding sequences are insignificant (1.31 vs. 1.4 in log₂(FC), p-value=0.4, t-test), while in regular promoter and enhancer regions, the eCLIP signals are depleted compared to the ATAC-seq regions with the log₂(FC) values of -0.1 and -0.05 respectively (p-value<10⁻³⁰, t-test), suggesting a strong RNA-related component in the composition of 3D medium surrounding the HOT loci.

All this data suggests a strong likelihood of involvement of transcriptional condensates in the mechanisms leading to the phenomena of HOT loci.

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 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}.

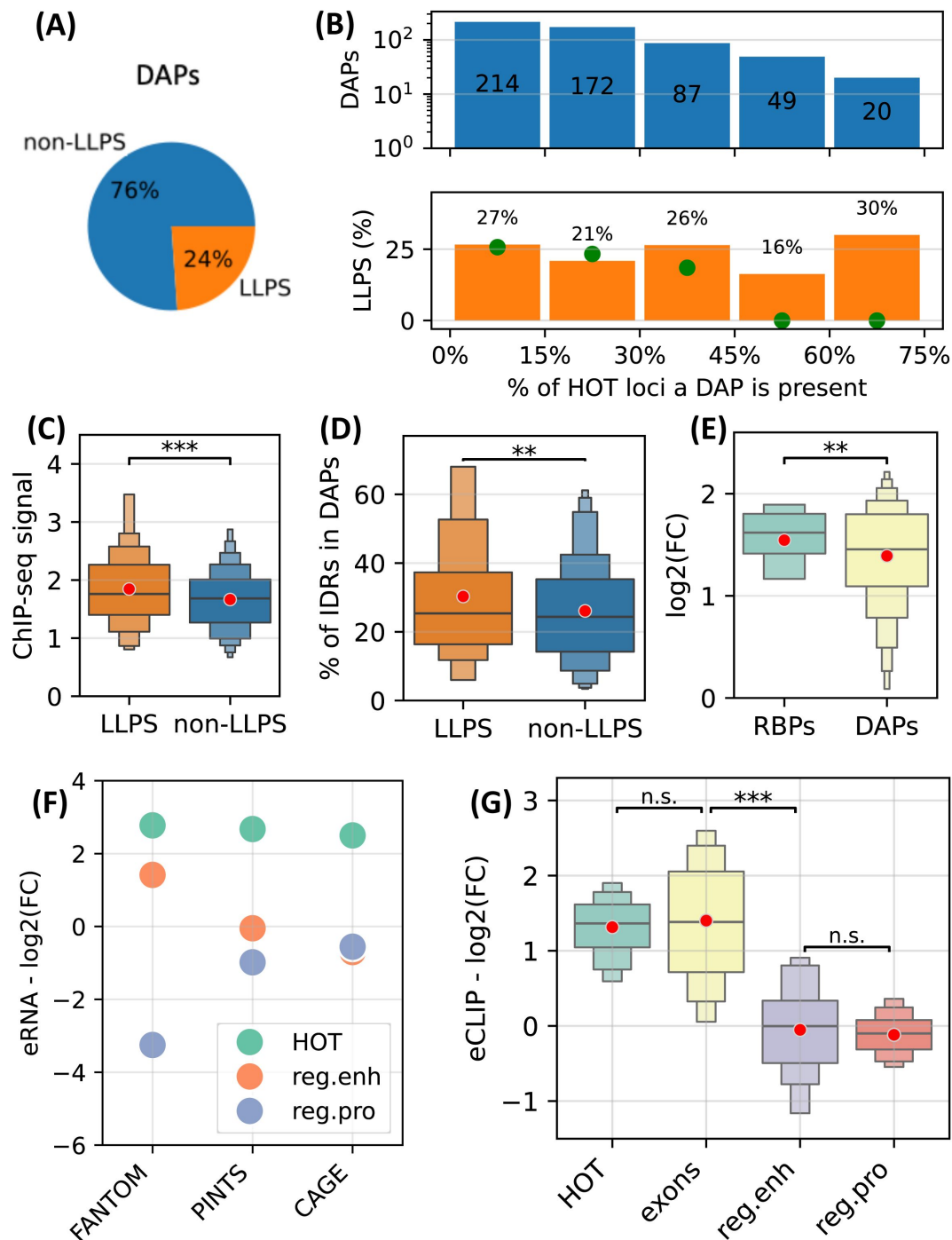


Figure 9.

HOT loci as transcriptional condensates.

A) fraction of DAPs annotated as LLPS proteins in CD-CODE database. **B)** (upper) Distribution of DAPs in HOT loci binned by the % of HOT loci they overlap with. (lower) % of DAPs in the bins annotated as LLPS. Green points are the expected percentage values obtained by randomly shuffling the peaks in HOT loci 10 times. **C)** Z-scores of ChIP-seq signal values of LLPS proteins and the rest of the DAPs in HOT loci. **D)** % of the protein lengths predicted as IDRs (MobiDB) in LLPS proteins and the rest of the DAPs. **E)** Enrichment of ChIP-seq peaks of RNA-binding proteins and the rest of the DAPs. **F)** Enrichment of FANTOM, PINTS, and CAGE regions in HOT, regular enhancers, and regular promoters. **G)** Enrichment of eCLIP RBP-RNA interactions in HOT, exons, regular enhancers, and regular promoters. **E,F,G)** Enrichment values are quantified as log₂(fold-change) with ATAC-seq regions as a background. **C,D,E,G)** red dots represent the mean values of the boxplots.

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 complete²⁸, 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.¹⁸ used the 99th percentile of the density of TFBSs for a 500 bp sliding window, Remaker et al.¹⁹ used the window length of 2 kb and required >25% of TFs to be mapped, Partridge et al.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 differentiation²¹.

Previous studies have carried out extensive mapping of the known 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. Other studies have concluded that the vast majority of the peaks do not contain the corresponding motifs and that most of the mapped peaks represent indirect binding through TF-TF interactions^{19,20,42,52}. We relied on these studies and focused on aspects of the HOT loci other than the quantification of known binding motifs of DAPs in HOT loci. 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.

Another model that has been increasingly attributed to the formation and maintenance of long-range 3D chromatin interactions involves phase-separated condensates^{46–49}. Some enhancers were shown to drive the formation of large chromosomal assemblies involving a high concentration of TFs⁴⁶. In general, it has been increasingly appreciated that condensates ubiquitously attract and activate enhancers^{53–55}. 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, to the best of our knowledge, there are no datasets of condensate formation with large numbers of molecules simultaneously that we could use to draw statistical conclusions. However, there is already an increasing body of research reporting on the characteristic hallmarks that the transcriptional condensates share^{56–60}. We used those hallmarks as telltale signs and made a case for the likelihood of the HOT loci being sites with a high propensity of forming condensates. A condensate can start forming with only one bound TF and a cofactor (e.g. OCT4 and Mediator⁵³), which requires the presence of a strong binding motif of the condensate-initiating TF. 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 and protein-RNA interactions, many constituents of

which then get mapped to the involved DNA regions upon ChIP-seq experiments. This model can incorporate the seemingly contradictory conclusions of a) the vast majority of DAPs lacking the binding motifs in HOT loci and b) a high accuracy of sequence-based classification of HOT loci using the CNN models. It is important to note here that our proposed condensate model is a speculative hypothesis. Further experimental studies in the field are needed to confirm or reject it.

One of the main limitations of our study is the lack of higher-resolution TF-DNA interaction datasets such as CUT&RUN, ChIP-exo, or single-cell versions of the assets used in this study. Furthermore, one of the hallmarks of condensates is the overrepresentation of certain structural motifs in LLPS proteins, which we did not pursue due to size limitations. Further studies addressing these topics hold promise to shed more light on the subject of HOT loci.

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 ⁶¹[61](#). 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. For comparing, averaged values of phastCons scores over the 400 bp loci were used. CpG islands, repeat elements and GENCODE TSS annotations were all obtained from the UCSC genome browser database ¹¹[11](#). Transcribed enhancer regions (eRNAs) were obtained from the FANTOM database ⁶²[62](#). Super-enhancer regions were obtained from (Hnisz et al. 2013) ⁶³[63](#).

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 ⁶⁴[64](#). 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[39](#).

LLPS protein annotations were obtained from CD-CODE website <https://cd-code.org> ⁶⁵[65](#). Predicted intrinsically disordered region annotations of proteins were obtained from MobiDB website <https://mobidb.org> ⁶⁶[66](#). RNA-binding protein ChIP-seq datasets used in the study are in Table S6. eCLIP datasets used in the study are in Table S7. PINTS eRNA dataset was obtained from <https://pints.yulab.org> ⁶⁷[67](#). CAGE datasets were downloaded from ENCODE (ENCFF184VBV, ENCFF246WDH, ENCFF933JJT) and merged.

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 (4 bins), the second part is on a natural logarithmic scale from 5 to the maximum number of DAPs bound to a single locus in that cell line (10 bins) (Table 1 ⁶⁸[68](#)).

We considered an average TF binding site to be 8pb long ⁶⁶[66](#),⁶⁷[67](#). Given that we analyzed the loci in 400bp, we reasoned that, theoretically, there can be at most 50 simultaneous binding events in the locus (8×50=400). Therefore, we considered the bins containing >50 DAPs in K562 as HOT loci,

which meant the last 4 bins in the **Table 1**. The reason we chose K562 for setting the threshold was the fact that K562 is the lesser of the two most TF ChIP-seq abundant cell lines. So, the corresponding threshold number for HepG2 is >77 TFs.

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* program⁶⁴. Figures were generated using Matplotlib⁶⁸ and Seaborn⁶⁹ packages. Statistical and numerical analyses were done using the pandas, NumPy, SciPy and sklearn packages⁷⁰ 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 interface⁷¹.

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 $<10^{-100}$.

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.

For calculating the statistical significance, we used the non-parametric Mann-Whitney U-test when the compared data points are non-linearly correlated and multi-modal. When the data distributions are bell-curve shaped, the Student’s t-test was used.

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 $r^2 > 0.8$ were added using the *plink v1.9*⁷² 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* v2.3.1⁷³ 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 package³⁷ (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. 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 dataset⁴¹. Variants with PASS filter value and MAF>5% were selected using the “view -f PASS -i ‘MAF[0]>0.05’” options of *bcftools* program⁷⁴. 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 of⁴⁴.

Software and Data availability statement

The codebase used for generating the results presented in this manuscript is available at <https://github.com/okurman/HOT>. Supplemental and source datasets used in the study are available at <https://zenodo.org/records/10267278>.

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.

References

1. ENCODE Project Consortium *et al.* (2020) **Expanded encyclopaedias of DNA elements in the human and mouse genomes** *Nature* **583**:699–710
2. Gorkin D.U. *et al.* (2020) **An atlas of dynamic chromatin landscapes in mouse fetal development** *Nature* **583**:744–751
3. Forsberg M., Westin G (1991) **Enhancer activation by a single type of transcription factor shows cell type dependence** *EMBO J* **10**:2543–2551
4. Serfling E., Jasin M., Schaffner W (1985) **Enhancers and eukaryotic gene transcription** *Trends Genet* **1**:224–230
5. Sethi A. *et al.* (2020) **Supervised enhancer prediction with epigenetic pattern recognition and targeted validation** *Nat. Methods* **17**:807–814
6. Spitz F., Furlong E.E.M (2012) **Transcription factors: from enhancer binding to developmental control** *Nat. Rev. Genet* **13**:613–626
7. Long H.K., Prescott S.L., Wysocka J (2016) **Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution** *Cell* **167**:1170–1187
8. Thanos D., Maniatis T (1995) **Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome** *Cell* **83**:1091–1100
9. Merika M., Thanos D (2001) **Enhanceosomes** *Curr. Opin. Genet. Dev* **11**:205–208
10. Arnosti D.N., Kulkarni M.M (2005) **Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?** *J. Cell. Biochem* **94**:890–898
11. Davis C.A. *et al.* (2018) **The Encyclopedia of DNA elements (ENCODE): data portal update** *Nucleic Acids Res* **46**:D794–D801
12. modENCODE Consortium, *et al.* (2010) **Identification of functional elements and regulatory circuits by Drosophila modENCODE** *Science* **330**:1787–1797
13. Moorman C. *et al.* (2006) **Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster** *Proc Natl Acad Sci USA* **103**:12027–12032
14. Gerstein M.B. *et al.* (2010) **Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project** *Science* **330**:1775–1787
15. Kvon E.Z., Stampfel G., Yáñez-Cuna J.O., Dickson B.J., Stark A (2012) **HOT regions function as patterned developmental enhancers and have a distinct cis-regulatory signature** *Genes Dev* **26**:908–913
16. Yip K.Y. *et al.* (2012) **Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors** *Genome Biol* **13**

17. Teytelman L., Thurtle D.M., Rine J., 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
18. Wreczycka K., Franke V., Uyar B., Wurmus R., Bulut S., Tursun B., Akalin A (2019) **HOT or not: examining the basis of high-occupancy target regions** *Nucleic Acids Res* **47**:5735–5745
19. Ramaker R.C., Hardigan A.A., Goh S.-T., Partridge E.C., Wold B., Cooper S.J., 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
20. Partridge E.C. *et al.* (2020) **Occupancy maps of 208 chromatin-associated proteins in one human cell type** *Nature* **583**:720–728
21. Boyle A.P. *et al.* (2014) **Comparative analysis of regulatory information and circuits across distant species** *Nature* **512**:453–456
22. Whyte W.A., Orlando D.A., Hnisz D., Abraham B.J., Lin C.Y., Kagey M.H., Rahl P.B., Lee T.I., Young R.A (2013) **Master transcription factors and mediator establish super-enhancers at key cell identity genes** *Cell* **153**:307–319
23. Lieberman-Aiden E. *et al.* (2009) **Comprehensive mapping of long-range interactions reveals folding principles of the human genome** *Science* **326**:289–293
24. Bhattacharyya S., Chandra V., Vijayanand P., Ay F (2019) **Identification of significant chromatin contacts from HiChIP data by FitHiChIP** *Nat. Commun* **10**
25. Schmitt A.D. *et al.* (2016) **A compendium of chromatin contact maps reveals spatially active regions in the human genome** *Cell Rep* **17**:2042–2059
26. Quinodoz S.A. *et al.* (2018) **Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus** *Cell* **174**:744–757
27. Xie D., Boyle A.P., Wu L., Zhai J., Kawli T., Snyder M (2013) **Dynamic trans-acting factor colocalization in human cells** *Cell* **155**:713–724
28. Lambert S.A., Jolma A., Campitelli L.F., Das P.K., Yin Y., Albu M., Chen X., Taipale J., Hughes T.R., Weirauch M.T (2018) **The human transcription factors** *Cell* **172**:650–665
29. Siepel A. *et al.* (2005) **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes** *Genome Res* **15**:1034–1050
30. Michailidou K. *et al.* (2017) **Association analysis identifies 65 new breast cancer risk loci** *Nature* **551**:92–94
31. Liu J. *et al.* (2020) **Obg-Like ATPase 1 Enhances Chemoresistance of Breast Cancer via Activation of TGF- β /Smad Axis Cascades** *Front. Pharmacol* **11**
32. Biel A., Castanza A.S., Rutherford R., Fair S.R., Chifamba L., Wester J.C., Hester M.E., Hevner R.F (2022) **AUTS2 syndrome: molecular mechanisms and model systems** *Front. Mol. Neurosci* **15**
33. Schanze I., Schanze D., Bacino C.A., Douzgou S., Kerr B., Zenker M (2013) **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

34. Daigle T.L. *et al.* (2018) **A Suite of Transgenic Driver and Reporter Mouse Lines with Enhanced Brain-Cell-Type Targeting and Functionality** *Cell* **174**:465–480
35. Pachano T. *et al.* (2021) **Orphan CpG islands amplify poised enhancer regulatory activity and determine target gene responsiveness** *Nat. Genet* **53**:1036–1049
36. Deaton A.M., Bird A (2011) **CpG islands and the regulation of transcription** *Genes Dev* **25**:1010–1022
37. Lee D (2016) **LS-GKM: a new gkm-SVM for large-scale datasets** *Bioinformatics* **32**:2196–2198
38. Hounkpe B.W., Chenou F., de Lima F., 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
39. Palmer D., Fabris F., Doherty A., Freitas A.A., de Magalhães J.P. (2021) **Ageing transcriptome meta-analysis reveals similarities and differences between key mammalian tissues** *Aging (Albany NY)* **13**:3313–3341
40. Domené S., Bumashny V.F., de Souza F.S.J., Franchini L.F., Nasif S., Low M.J., Rubinstein M. (2013) **Enhancer turnover and conserved regulatory function in vertebrate evolution** *Philos. Trans. R. Soc. Lond. B Biol. Sci* **368**
41. Karczewski K.J. *et al.* (2020) **The mutational constraint spectrum quantified from variation in 141,456 humans** *Nature* **581**:434–443
42. Vierstra J. *et al.* (2020) **Global reference mapping of human transcription factor footprints** *Nature* **583**:729–736
43. GTEx Consortium (2015) **Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans** *Science* **348**:648–660
44. Currin K.W. *et al.* (2021) **Genetic effects on liver chromatin accessibility identify disease regulatory variants** *Am. J. Hum. Genet* **108**:1169–1189
45. van Arensbergen J. *et al.* (2019) **High-throughput identification of human SNPs affecting regulatory element activity** *Nat. Genet* **51**:1160–1169
46. Nair S.J. *et al.* (2019) **Phase separation of ligand-activated enhancers licenses cooperative chromosomal enhancer assembly** *Nat. Struct. Mol. Biol* **26**:193–203
47. Lee R. *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
48. Feric M., Misteli T (2022) **Function moves biomolecular condensates in phase space** *Bioessays* **44**
49. Ahn J.H. *et al.* (2021) **Phase separation drives aberrant chromatin looping and cancer development** *Nature* **595**:591–595
50. Rostam N. *et al.* (2023) **CD-CODE: crowdsourcing condensate database and encyclopedia** *Nat. Methods* **20**:673–676

51. Yao L., Liang J., Ozer A., Leung A.K.-Y., Lis J.T., Yu H (2022) **A comparison of experimental assays and analytical methods for genome-wide identification of active enhancers** *Nat. Biotechnol* **40**:1056–1065
52. White S.M., Snyder M.P., 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
53. Shrinivas K. *et al.* (2019) **Enhancer Features that Drive Formation of Transcriptional Condensates** *Mol. Cell* **75**:549–561
54. Wei M.-T., Chang Y.-C., Shimobayashi S.F., Shin Y., Strom A.R., Brangwynne C.P (2020) **Nucleated transcriptional condensates amplify gene expression** *Nat. Cell Biol* **22**:1187–1196
55. Boija A. *et al.* (2018) **Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains** *Cell* **175**:1842–1855
56. Palacio M., Taatjes D.J (2022) **Merging Established Mechanisms with New Insights: Condensates, Hubs, and the Regulation of RNA Polymerase II Transcription** *J. Mol. Biol* **434**
57. Mitrea D.M., Mittasch M., Gomes B.F., Klein I.A., Murcko M.A (2022) **Modulating biomolecular condensates: a novel approach to drug discovery** *Nat. Rev. Drug Discov* **21**:841–862
58. Gelder K.L., Carruthers N.A., Ball S., Dunning M., Craggs T.D., Twelvetrees A.E., Bose D.A. (2024) **Cooperation between Intrinsically Disordered Regions regulates CBP condensate behaviour** *BioRxiv*
59. Bhat P., Honson D., Guttman M (2021) **Nuclear compartmentalization as a mechanism of quantitative control of gene expression** *Nat. Rev. Mol. Cell Biol* **22**:653–670
60. Rippe K., Papantonis A (2021) **RNA polymerase II transcription compartments: from multivalent chromatin binding to liquid droplet formation?** *Nat. Rev. Mol. Cell Biol* **22**:645–646
61. Wang J. *et al.* (2013) **Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium** *Nucleic Acids Res* **41**:D171–6
62. Lizio M. *et al.* (2019) **Update of the FANTOM web resource: expansion to provide additional transcriptome atlases** *Nucleic Acids Res* **47**:D752–D758
63. Hnisz D., Abraham B.J., Lee T.I., Lau A., Saint-André V., Sigova A.A., Hoke H.A., Young R.A (2013) **Super-enhancers in the control of cell identity and disease** *Cell* **155**:934–947
64. Quinlan A.R., Hall I.M (2010) **BEDTools: a flexible suite of utilities for comparing genomic features** *Bioinformatics* **26**:841–842
65. Barrett T. *et al.* (2013) **NCBI GEO: archive for functional genomics data sets--update** *Nucleic Acids Res* **41**:D991–5
66. Vinson C., Chatterjee R., Fitzgerald P (2011) **Transcription factor binding sites and other features in human and Drosophila proximal promoters** *Subcell Biochem* **52**:205–222

67. Wunderlich Z., Mirny L.A (2009) **Different gene regulation strategies revealed by analysis of binding motifs** *Trends Genet* **25**:434–440
68. Hunter J.D (2007) **Matplotlib: A 2D Graphics Environment** *Comput. Sci. Eng* **9**:90–95
69. Waskom M (2021) **seaborn: statistical data visualization** *JOSS* **6**
70. Virtanen P. *et al.* (2020) **SciPy 1.0: fundamental algorithms for scientific computing in Python** *Nat. Methods* **17**:261–272
71. Szklarczyk D. *et al.* (2019) **STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets** *Nucleic Acids Res* **47**:D607–D613
72. Chang C.C., Chow C.C., Tellier L.C., Vattikuti S., Purcell S.M., Lee J.J (2015) **Second-generation PLINK: rising to the challenge of larger and richer datasets** *Gigascience* **4**
73. Abadi M. *et al.* (2016) **TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems** *ArXiv*
74. Li H (2011) **A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data** *Bioinformatics* **27**:2987–2993

Editors

Reviewing Editor

Nicolas Altemose

Stanford University, United States of America

Senior Editor

Sofia Araújo

University of Barcelona, Barcelona, Spain

Reviewer #1 (Public review):

Summary:

This study explores the sequence characteristics and features of high-occupancy target (HOT) loci across the human genome. The computational analyses presented in this paper provide information into the correlation of TF binding and regulatory networks at HOT loci that were regarded as lacking sequence specificity.

By leveraging hundreds of ChIP-seq datasets from the ENCODE Project to delineate HOT loci in HepG2, K562, and H1-hESC cells, the investigators identified the regulatory significance and participation in 3D chromatin interactions of HOT loci. Subsequent exploration focused on the interaction of DNA-associated proteins (DAPs) with HOT loci using computational models. The models established that the potential formation of HOT loci is likely embedded in their DNA sequences and is significantly influenced by GC contents. Further inquiry exposed contrasting roles of HOT loci in housekeeping and tissue-specific functions spanning various cell types, with distinctions between embryonic and differentiated states, including instances of polymorphic variability. The authors conclude with a speculative model that HOT loci serve as anchors where phase-separated transcriptional condensates form. The findings

presented here open avenues for future research, encouraging more exploration of the functional implications of HOT loci.

Strengths:

The concept of using computational models to define characteristics of HOT loci is refreshing and allows researchers to take a different approach in identifying potential targets. The major strengths of the study lie in the very large number of datasets analyzed, with hundreds of ChIP-seq data sets for both HepG2 and K562 cells as part of the ENCODE project. Such quantitative power allowed the authors to delve deeply into HOT loci, which were previously thought to be artifacts.

Weaknesses:

While this study contributes to our knowledge of HOT loci, there are critical weaknesses that need to be addressed. There are questions on the validity of the assumptions made for certain analyses. The speculative nature of the proposed model involving transcriptional condensates needs either further validation or be toned down. Furthermore, some apparent contradictions exist among the main conclusions, and these either need to be better explained or corrected. Lastly, several figure panels could be better explained or described in the figure legends.

Update After Revisions:

The authors have addressed the above comments and concerns appropriately. The addition of the new Figure 9 is particularly compelling and strengthens the authors' conclusions. This reviewer has no further concerns.

<https://doi.org/10.7554/eLife.95170.2.sa3>

Reviewer #2 (Public review):

Summary:

The paper by Hydaiberdiev and Ovcharenko offers comprehensive analyses and insights about the 'high-occupancy target' (HOT) loci in the human genome. These are considered genomic regions that overlap with transcription factor binding sites. The authors provided very comprehensive analyses of the TF composition characteristics of these HOT loci. They showed that these HOT loci tend to overlap with annotated promoters and enhancers, GC-rich regions, open chromatin signals, and highly conserved regions and that these loci are also enriched with potentially causal variants with different traits.

Strengths:

Overall, the HOT loci' definition is clear and the data of HOT regions across the genome can be a useful dataset for studies that use HepG2 or K562 as a model. I appreciate the authors' efforts in presenting many analyses and plots backing up each statement.

Comments on revised version:

In the second round of review, I think the authors have sufficiently addressed all of my previous comments. The study itself is very comprehensive, tackling all aspects of the HOT loci, though I still find the paper to be unnecessarily long and long-winded. That said, being consistent with the long and detailed paper, the provided Github repository and Zenodo archive is well-documented. I appreciate that the authors include detailed readme about the different datafiles available for readers. The list of HOT loci is probably the most useful asset

in this manuscript and the authors did a good job documenting data availability in both Github and Zenodo.

<https://doi.org/10.7554/eLife.95170.2.sa2>

Reviewer #3 (Public review):

Summary:

Hudaiberdiev and Ovcharenko investigate regions within the genome where a high abundance of DNA associated proteins are located and identify DNA sequence feature enriched in these regions, their conservation in evolution, and variation in disease. Using ChIP-seq binding profiles of over 1,000 proteins in three human cell lines (HepG2, K562, and H1) as a data source they're able to identify nearly 44,000 high-occupancy target loci (HOT) that form at promoter and enhancer regions, thus suggesting these HOT loci regulate housekeeping and cell identity genes. Their primary investigative tool is HepG2 cells, but they employ K562 and H1 cells as tools to validate these assertions in other human cell types. Their analyses use RNA pol II signal, super enhancer, regular enhancer and epigenetic marks to support the identification of these regions. The work is notable, in that it identifies a set of proteins that are invariantly associated with high-occupancy enhancers and promoters and argues for the integration of these molecules at different genomic loci. These observations are leveraged by the authors to argue HOT loci as potential sites of transcriptional condensates, a claim that they provide information in support of. Transcriptional condensates are an important "family" of condensates, regulating different types of genes and this work supports the hypothesis that they possess similar protein partner molecules as those thought to define such bodies.

<https://doi.org/10.7554/eLife.95170.2.sa1>

Author response:

The following is the authors' response to the original reviews.

Reviewer #1 (Public Review):

Summary:

This study explores the sequence characteristics and features of high-occupancy target (HOT) loci across the human genome. The computational analyses presented in this paper provide information into the correlation of TF binding and regulatory networks at HOT loci that were regarded as lacking sequence specificity.

By leveraging hundreds of ChIP-seq datasets from the ENCODE Project to delineate HOT loci in HepG2, K562, and H1-hESC cells, the investigators identified the regulatory significance and participation in 3D chromatin interactions of HOT loci. Subsequent exploration focused on the interaction of DNA-associated proteins (DAPs) with HOT loci using computational models. The models established that the potential formation of HOT loci is likely embedded in their DNA sequences and is significantly influenced by GC contents. Further inquiry exposed contrasting roles of HOT loci in housekeeping and tissue-specific functions spanning various cell types, with distinctions between embryonic and differentiated states, including instances of polymorphic variability. The authors conclude with a speculative model that HOT loci serve as anchors where phase-separated transcriptional condensates form. The findings presented here open avenues for future research, encouraging more exploration of the functional implications of HOT loci.

Strengths:

The concept of using computational models to define characteristics of HOT loci is refreshing and allows researchers to take a different approach to identifying potential targets. The major strengths of the study lies in the very large number of datasets analyzed, with hundreds of ChIP-seq data sets for both HepG2 and K562 cells as part of the ENCODE project. Such quantitative power allowed the authors to delve deeply into HOT loci, which were previously thought to be artifacts.

Weaknesses:

While this study contributes to our knowledge of HOT loci, there are critical weaknesses that need to be addressed. There are questions on the validity of the assumptions made for certain analyses. The speculative nature of the proposed model involving transcriptional condensates needs either further validation or be toned down. Furthermore, some apparent contradictions exist among the main conclusions, and these either need to be better explained or corrected. Lastly, several figure panels could be better explained or described in the figure legends.

We thank the reviewer for their valuable comments.

- We have extended the study and included a new chapter focusing on the condensate hypothesis, added more supporting evidence (including the ones suggested by the reviewer), and made explicit statements on the speculative nature of this model.

- We have restructured the text to remove the sentences which might be construed as contradictory.

Reviewer #2 (Public Review):

Summary:

The paper 'Sequence characteristic and an accurate model of abundant hyperactive loci in human genome' by Hydaiberdiev and Ovcharenko offers comprehensive analyses and insights about the 'high-occupancy target' (HOT) loci in the human genome. These are considered genomic regions that overlap with transcription factor binding sites. The authors provided very comprehensive analyses of the TF composition characteristics of these HOT loci. They showed that these HOT loci tend to overlap with annotated promoters and enhancers, GC-rich regions, open chromatin signals, and highly conserved regions, and that these loci are also enriched with potentially causal variants with different traits.

Strengths:

Overall, the HOT loci' definition is clear and the data of HOT regions across the genome can be a useful dataset for studies that use HepG2 or K562 as a model. I appreciate the authors' efforts in presenting many analyses and plots backing up each statement.

Weaknesses:

It is noteworthy that the HOT concept and their signature characteristics as being highly functional regions of the genome are not presented for the first time here. Additionally, I find the main manuscript, though very comprehensive, long-winded and can be put in a shorter, more digestible format without sacrificing scientific content.

The introduction's mention of the blacklisted region can be rather misleading because when I read it, I was anticipating that we are uncovering new regulatory regions within

the blacklisted region. However, the paper does not seem to address the question of whether the HOT regions overlap, if any, with the ENCODE blacklisted regions afterward. This plays into the central assessment that this manuscript is long-winded.

The introduction also mentioned that HOT regions correspond to 'genomic regions that seemingly get bound by a large number of TFs with no apparent DNA sequence specificity' (this point of 'no sequence specificity' is reiterated in the discussion lines 485-486). However, later on in the paper, the authors also presented models such as convolutional neural networks that take in one-hot-encoded DNA sequence to predict HOT performed really well. It means that the sequence contexts with potential motifs can still play a role in forming the HOT loci. At the same time, lines 59-60 also cited studies that "detected putative drive motifs at the core segments of the HOT loci". The authors should edit the manuscript to clarify (or eradicate) contradictory statements.

We thank the reviewer for their valuable comments. Below are our responses to each paragraph in the given order:

We added a statement in the commenting and summarizing other publications that studied the functional aspects of HOT loci with the following sentence in the introduction part:

“Other studies have concluded that these regions are highly functionally consequential regions enriched in epigenetic signals of active regulatory elements such as histone modification regions and high chromatin accessibility”.

We significantly shortened the manuscript by a) moving the detailed analyses of the computational model to the supplemental materials, and b) shortening the discussions by around half, focusing on core analyses that would be most beneficial to the field.

Given that the ENCODE blacklisted regions are the regions that are recommended by the ENCODE guidelines to be avoided in mapping the ChIP-seq (and other NGS), we excluded them from our analyzed regions before mapping to the genome. Instead, we relied on the conclusions of other publications on HOT loci that the initial assessments of a fraction of HOT loci were the result of factoring in these loci which later were included in blacklisted regions.

We addressed the potential confusion by using the expression of “no sequence specificity” by a) changing the sentence in the introduction by adding a clarification as “... with no apparent DNA sequence specificity in terms of detectable binding motifs of corresponding motifs” and b) removing that part from the sentence in the discussions.

Reviewer #3 (Public Review):

Summary:

Hudaiberdiev and Ovcharenko investigate regions within the genome where a high abundance of DNA-associated proteins are located and identify DNA sequence features enriched in these regions, their conservation in evolution, and variation in disease. Using ChIP-seq binding profiles of over 1,000 proteins in three human cell lines (HepG2, K562, and H1) as a data source they're able to identify nearly 44,000 high-occupancy target loci (HOT) that form at promoter and enhancer regions, thus suggesting these HOT loci regulate housekeeping and cell identity genes. Their primary investigative tool is HepG2 cells, but they employ K562 and H1 cells as tools to validate these assertions in other human cell types. Their analyses use RNA pol II signal, super-enhancer, regular-enhancer, and epigenetic marks to support the identification of these regions. The work is notable, in that it identifies a set of proteins that are invariantly associated with high-occupancy enhancers and promoters and argues for the integration of these molecules at different genomic loci. These observations are leveraged by the authors to argue HOT loci as potential sites of transcriptional condensates, a claim that they are well poised to

provide information in support of. This work would benefit from refinement and some additional work to support the claims.

Comments:

(1) Condensates are thought to be scaffolded by one or more proteins or RNA molecules that are associated together to induce phase separation. The authors can readily provide from their analysis a check of whether HOT loci exist within different condensate compartments (or a marker for them). Generally, ChIPSeq signal from MED1 and Ronin (THAP11) would be anticipated to correspond with transcriptional condensates of different flavors, other coactivator proteins (e.g., BRD4), would be useful to include as well. Similarly, condensate scaffolding proteins of facultative and constitutive heterochromatin (HP1a and EZH2/1) would augment the authors' model by providing further evidence that HOT Loci occur at transcriptional condensates and not heterochromatin condensates. Sites of splicing might be informative as well, splicing condensates (or nuclear speckles) are scaffolded by SRRM/SON, which is probably not in their data set, but members of the serine arginine-rich splicing factor family of proteins can serve as a proxy-SRSF2 is the best studied of this set. This would provide a significant improvement to their proposed model and be expected since the authors note that these proteins occur at the enhancers and promoter regions of highly expressed genes.

(2) It is curious that MAX is found to be highly enriched without its binding partner Myc, is Myc's signal simply lower in abundance, or is it absent from HOT loci? How could it be possible that a pair of proteins, which bind DNA as a heterodimer are found in HOT loci without invoking a condensate model to interpret the results?

(3) Numerous studies have linked the physical properties of transcription factor proteins to their role in the genome. The authors here provide a limited analysis of the proteins found at different HOT-loci by employing go terms. Is there evidence for specific types of structural motifs, disordered motifs, or related properties of these proteins present in specific loci?

(4) Condensates themselves possess different emergent properties, but it is a product of the proteins and RNAs that concentrate in them and not a result of any one specific function (condensates can have multiple functions!)

(5) Transcriptional condensates serve as functional bodies. The notion the authors present in their discussion is not held by practitioners of condensate science, in that condensates exist to perform biochemical functions and are dissolved in response to satisfying that need, not that they serve simply as reservoirs of active molecules. For example, transcriptional condensates form at enhancers or promoters that concentrate factors involved in the activation and expression of that gene and are subsequently dissolved in response to a regulatory signal (in transcription this can be the nascently synthesized RNA itself or other factors). The association reactions driving the formation of active biochemical machinery within condensates are materially changed, as are the kinetics of assembly. It is unnecessary and inaccurate to qualify transcriptional condensates as depots for transcriptional machinery.

1. This work has the potential to advance the field forward by providing a detailed perspective on what proteins are located in what regions of the genome. Publication of this information alongside the manuscript would advance the field materially.

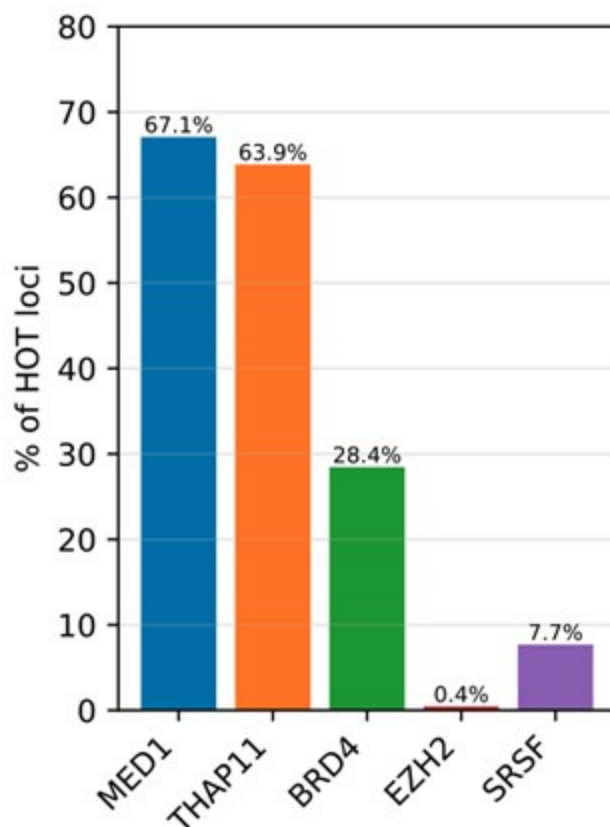
We thank the reviewer for constructive comments and suggestions. Below are our point-by-point responses:

(1) We added a new short section “Transcriptional condensates as a model for explaining the HOT regions” with additional support for the condensate hypothesis, wherein some of the points raised here were addressed. Specifically, we used a curated LLPS proteins (CD-CODE) database and provided statistics of those annotation condensate-related DAPs.

Regarding the DAPs mentioned in this question, we observed that the distributions corresponding ChIP-seq peaks confirm the patterns expected by the reviewer (Author response image 1). Namely:

- MED1 and Ronin (THAP11) are abundant in the HOT loci, being present 67% and 64% of HOT loci respectively.
- While the BRD4 is present in 28% of the HOT loci, we observed that the DAPs with annotated LLPS activity ranged from 3% to 73%, providing further support for the condensate hypothesis.
- ENCODE database does not contain ChIP-seq dataset for HP1A. EZH2 peaks were absent in the HOT loci (0.4% overlap), suggesting the lack of heterochromatin condensate involvement.
- Serine-rich splicing factor family proteins were present only in 7.7% of the HOT loci, suggesting the absence or limited overlap with splicing condensates or nuclear speckles.

Author response image 1.

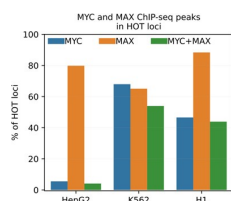


(2) In this study we selected the TF ChIP-seq datasets with stringent quality metrics, excluding those which had attached audit warning and errors. As a result, the set of DAPs analyzed in HepG2 did not include MYC, since the corresponding ChIP-seq dataset had the audit warning tags of "borderline replicate concordance, insufficient read length, insufficient read depth,

extremely low read depth". Analyses in K562 and H1 did include MYC (alongside MAX) ChIP-seq dataset.

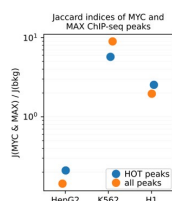
To address this question, we added the mentioned ChIP-seq dataset (ENCODE ID: ENCF800JFG) and analyzed the colocalization patterns of MYC and MAX. We observed that the MYC ChIP-seq peaks in HepG2 display spurious results, overlapping with only 5% of HOT loci. Meanwhile in K562 and H1, MYC and MAX are jointly present in 54% and 44% of the HOT loci, respectively (Author response image 2).

Author response image 2.



These observations were also supported by Jaccard indices between the MYC and MAX ChIP-seq peaks. To do this analysis, we calculated the pairwise Jaccard indices between MYC and MAX and divided them by the average Jaccard indices of 2000 randomly selected DAP pairs. In K562 and H1, the Jaccard indices between MYC and MAX are 5.72x and 2.53x greater than the random background, respectively. For HepG2, the ratio was 0.21x, clearly indicating that HepG2 MYC ChIP-seq dataset is likely erroneous.

Author response image 3.



(3) Despite numerous publications focusing on different structural domains in transcription factors, we could not find an extensive database or a survey study focusing on annotations of structural motifs in human TFs. Therefore, surveying such a scale would be outside of this study's scope. We added only the analysis of intrinsically disordered regions, as it pertains to the condensate hypothesis. To emphasize this shortcoming, we added the following sentence to the end of the discussions section.

"Further, one of the hallmarks of LLPS proteins that have been associated with their abilities to phase-separate is the overrepresentation of certain structural motifs, which we did not pursue due to size limitations."

(4, 5) We agree with these statements and thank the reviewer for pointing out this faulty statement. We modified the sections in the discussions related to the condensates and removed the part where we implied that the condensate model could be because of mostly a single function of TF reservoir.

(6) We added a table to the supplemental materials (Zenodo repository) with detailed annotation of HOT and non-HOT DAP-bound loci in the genome.

Recommendations for the authors:**Reviewing Editor (Recommendations For The Authors):**

The clause with "inadequate" would be dropped if the authors sufficiently address reviewer concerns about clarity of writing, including:

- (1) Editing the title to better reflect the findings of the paper.*
- (2) Making clear that the condensate model is speculative and not explicitly tested in this study (and may be better described as a hypothesis).*
- (3) Resolving apparent contradictions regarding DNA sequence specificity and the interpretation of ChIP-seq signal intensity.*
- (4) Better specifying and justifying model parameters, thresholds, and assumptions.*
- (5) Shortening the manuscript to emphasize the main, well-supported claims and to enhance readability (especially the discussion section).*

We thank the Editor for their work. We followed their advice and implemented changes and additions to address all 5 points.

Reviewer #1 (Recommendations For The Authors):

(1) The title "Sequence characteristics and an accurate model of abundant hyperactive loci in the human genome" does not accurately reflect the findings of the paper. We are unclear as to what the 'accurate model' refers to. Is it the proposed model 'based on the existence of large transcriptional condensates' (abstract)? If so, there are concerns below regarding this statement (see comment 2). If the authors are referring to the computational modeling presented in Figure 5, it is unclear that any one of them performed that much better than the others and the best single model was not identified. Furthermore, the models being developed in the study constitute only a portion of the paper and lacked validation through additional datasets. Additionally, sequence characteristics were not a primary focus of the study. Only figure 5 talks about the model and sequence characteristics, the rest of the figures are left out of the equation.

We agree with and thank the reviewer for this idea of clarifying the intended meaning.

- (1) We changed the title and clarified that the computational model is meant:

“Functional characteristics and a computational model of abundant hyperactive loci in the human genome”.

- (2) Shortened the part of the manuscript discussing the computational models and pointed out the CNNs as “the best single model”.

(2) The abstract and discussion (and perhaps the title) propose a model of transcriptional condensates in relation to HOT loci. However, there is no data provided in the manuscript that relates to condensates. Therefore, anything relating to condensates is primarily speculative. This distinction needs to be properly made, especially in the abstract (and cannot be included in the title). Otherwise, these statements are misleading. Although the field of transcriptional condensates is relatively new, there have been several factors studied. The authors could include in Figure 2d which factors have been shown to form transcriptional condensates. This might provide some support for the model, though it would still largely remain speculative unless further testing is done.

We added a new short chapter “Transcriptional condensates as a model for explaining the HOT regions”, with additional analyses testing the condensates hypothesis. We provided supportive evidence by analyzing the metrics used as hallmarks of condensates including the distributions of annotated condensate-related proteins, nascent transcription, and protein-RNA interaction levels in HOT loci. Still, we acknowledge that this is a speculative hypothesis and we clarified that with the following statement in the discussions:

“It is important to note here that our proposed condensate model is a speculative hypothesis. Further experimental studies in the field are needed to confirm or reject it.”

(3) Several apparent contradictions exist throughout the manuscript. For example, "HOT locus formation are likely encoded in their DNA sequences" (lines 329-330) vs the proposed model of formation through condensates (abstract). These two statements do not seem compatible, or at the very least, the authors can explain how they are consistent with each other. Another example: "ChIP-seq signal intensity as a proxy for... binding affinity" (line 229) vs. "ChIP-seq signal intensities do not seem to be a function of the DNA-binding properties of the DAPs" (lines 259-260). The first statement is the assumption for subsequent analyses, which has its own concerns (see comment 4). But the conclusion from that analysis seems to contradict the assumption, at least as it is stated.

In this study, we argue that the two statements may not necessarily contradict each other. We aimed to a) demonstrate that the observed intensity of DAP-DNA interactions as measured by ChIP-seq experiments at HOT loci cannot be explained with direct DNA-binding events of the DAPs alone and b) propose a hypothesis that this observation can be at least partially explained if the HOT loci have the propensity to either facilitate or take part in the formation of transcriptional condensates.

One of the conditions for condensates to form at enhancers was shown to be the presence of strong binding sites of key TFs (Shrinivas et al. 2019 “Enhancer features that drive the formation of transcriptional condensates”), where the study was conducted using only one TF (OCT4) and one coactivator (MED1). To the best of our knowledge, no such study has been conducted involving many TFs and cofactors simultaneously. We also know that the factors that lead to liquid-to-liquid phase separation include weak multivalent IDR-IDR, IDR-DNA, and IDR-RNA interactions. As a result, the observed total sum of ChIP-seq peaks in HOT loci is the direct DNA-binding events combined with the indirect DAP-DNA interactions, some of which may be facilitated by condensates. And, the fact that CNNs can recognize the HOT loci with high accuracy suggests that there must be an underlying motif grammar specific to HOT loci.

We emphasized this conclusion in the discussions.

The comment on using the ChIP-seq signal as a proxy for DNA-binding affinity is addressed under comment 4.

(4) In lines 229-230, the authors used "the ChIP-seq signal intensity as a proxy for the DAP binding affinity." What is the basis for this assumption? If there is a study that can be referenced, it should be added. However, ChIP-seq signal intensity is generally regarded as a combination of abundance, frequency, or percentage of cells with binding. RNA Pol2 is a good example of this as it has no specific binding affinity but the peak heights indicate level of expression. Therefore, the analyses and conclusions in Figure 4, particularly panel A, are problematic. In addition, clarification from lines 258-260 is needed as it contradicts the earlier premise of the section (see comment 3).

We thank the reviewer for pointing out this error. The main conclusion of the paragraph is that the average ChIP-seq signal values at HOT loci do not correlate well with the sequence-specificity of TFs. We rewrote the paragraph stating that we are analyzing the patterns of ChIP-seq signals across the HOT loci, removing the part that we use them as a proxy for sequence-specific binding affinity.

(5) In Figure 1A, the authors show 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" (lines 92-95). The threshold to determine how a locus is considered to be HOT is unclear. How did the authors decide to use the current threshold given the uniform spectrum observed? How does this method of calling HOT loci compare to previous studies? How much overlap is there in the HOT loci in this study versus previous ones?

We moved the corresponding explanation from the supplemental methods to the main methods section of the manuscript.

Briefly, our reasoning was as follows: assuming that an average TFBS is 8bp long and given that we analyze the loci of length 400bp, we can set the theoretical maximum number of simultaneous binding events to be 50. Hence, if there are >50 TF ChIP-seq peaks in a given 400bp locus, it is highly unlikely that the majority of ChIP-seq peaks can be explained by direct TF-DNA interactions. The condition of >50 TFs corresponded to the last four bins of our binning scale, which was used as an operational definition for HOT loci.

We have compared our definition of HOT loci to those reported in previous studies by Remaker et al. and Boyle et al. The results of our analyses are in lines 147-154.

(6) In Figure 3B, the authors state that 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." However, it is unclear how "51%" is calculated from the figure. Similarly, in the following sentence, "94% of HOT loci are located in regions with at least one chromatin interaction". It is unclear as to how the number was obtained based on the referenced figure.

Initially, the x-axis on the Figure 3B was missing, making it hard to understand what we meant. We added the x-axis numbers and changed the "51%" to "more than half". We intend to say that, of the loci with 4 and 5 overlapping loops, exactly 50% contain at least one HOT locus. However, since for x=6 the percentage is 100% (since there's only one such locus), the percentage is technically "more than half".

The percentage of HOT loci engaging in chromatin interaction regions (91%) was calculated by simply overlapping the HOT regions with Hi-C long-range contact anchors. The details of extracting these regions using FitHiChip are described in Supplemental Methods 1.3.

(7) While we have a limited basis to evaluate computational models, we would like to see a clearer explanation of the model set-up in terms of the number of trained vs. test datasets. In addition, it would be interesting to see if the models can be applied to data from different cell lines.

We added the table with the sizes of the datasets used for classification in Supplemental Methods 1.6.1.

Evaluating the models trained on the HOT loci of HepG2 and K562 on other cell lines would pose challenges since the number of available ENCODE TF ChIP-seq datasets is significantly less compared to the mentioned cell lines. Therefore, we conducted the proposed analysis between the studied cell lines. Specifically, we used the CNN models trained on HOT and

regular enhancers of HepG2 and K562. Then, we evaluated each model on the test sets of each classification experiment (Author response image 4). We observed that the classification results of the HOT loci demonstrated a higher level of tissue-specificity compared to the same classification results of the regular enhancers.

Author response image 4.

Strain		HepG2		K562	
		HOT	reg.enh	HOT	reg.enh
HepG2	HOT	0.905	0.883	0.88	0.736
	reg.enh	0.913	0.932	0.824	0.772
K562	HOT	0.882	0.711	0.965	0.838
	reg.enh	0.834	0.753	0.909	0.9

(8) Lines 349-351. The significance of highly expressed genes being more prone to having multiple HOT loci, and vice versa, appears conventional and remains unclear. Intuitively, it makes sense for higher expressed genes to have more of the transcriptional machinery bound, and would bias the analysis. One way to circumvent this is to only analyze sequence-specific TFs and remove ones that are directly related to transcription machinery.

We thank the reviewer for this suggestion. Our attempt to re-annotate the HOT loci with only sequence-specific TFs led to a significantly different set of loci, which would not be strictly comparable to the HOT loci defined by this study. Analyzing these new sets of loci would create a noticeable departure from the flow of the manuscript and further extend the already long scope of the study.

Moreover, numerous studies have shown that super-enhancers recruit large numbers of TFs via transcriptional condensates (Boija et al., 2018; Cho et al., 2018; Sabari et al., 2018). We hope that our results can serve as data-driven supportive evidence for those studies.

(9) Lines 393-396. We would like to see a reference to the models shown in the figures, if these models have been published previously.

We could not understand the question. The lines 393-396 contains the following sentence:

“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.”

In case the question was about the models that we trained to classify the HOT loci, we included the models and codebase to Zenodo and GitHub repository.

(10) Values in Figure 7D are not reflected in the text. Specifically, the text states "Average ... phastCons of the developmental HOT loci are 1.3x higher than K562 and HepG2 HOT loci (Figure 7D)" (lines 408-409). Figure 7D shows conservation scores between HOT enhancers vs promoters for each cell line, and does not seem to reflect the text.

We modified the figure to reflect the statement appropriately.

(11) Methodology should include a justification for the use of the Mann-Whitney U-test (non-parametric) over other statistical tests.

We added the following description to the methods section:

“For calculating the statistical significance, we used the non-parametric Mann-Whitney U-test when the compared data points are non-linearly correlated and multi-modal. When the data distributions are bell-curve shaped, the Student’s t-test was used.”

Minor:

(1) Figure 2b was never mentioned in the paper. This can be added alongside Figure S6C, line 148.

Indeed, Figure 2B was supposed to be listed together with Figure S6C, which was omitted by mistake. It was corrected.

(2) Supplementary Figure 8 has two Cs. Needs to be corrected to D.

Fixed.

(3) Figure 3B is missing labels on the x-axis.

Fixed.

(4) The horizontal bar graph on the bottom left of Figure 1E needs to be described in the figure legend.

Description added to the figure caption.

(5) Line 345, Fig 15A should be Fig S15A.

Corrected.

Reviewer #2 (Recommendations For The Authors):

I listed all my concerns about the paper in the public comments. I think the manuscript is very comprehensive and it is valuable, but it should be cut short and presented in a more digestible way.

We thank the reviewer for their valuable comments and suggestions. We addressed all the concerns listed in the public comments. We shortened the manuscript by reducing the paragraph that focuses on computational classification models and reduced the discussions by about half in length.

Line 55: What are chromatin-associated proteins, i.e. are they histone modifications?

To clarify the definition used from the citation we changed the sentence to the following:

“For instance, Partridge et al. studied the HOT loci in the context of 208 proteins including TFs, cofactors, and chromatin regulators which they called chromatin-associated proteins.”

Though most of the paper can be cut short to avoid analysis paralysis for readers, there are details that still need filling in. For example, how did the authors perform PCA analysis, i.e. what are the features of each data point in the PCA analysis? Lines 214-215: How do we calculate the number of multi-way contacts in Hi-C data?

We added clarifying descriptions and changed the mentioned sentences to the following:

PCA:

“To analyze the signatures of unique DAPs in HOT loci, we performed a PCA analysis where each HOT locus is represented by a binary (presence/absence) vector of length equal to the total number of DAPs analyzed.”

Multi-way contacts on loop anchors:

“To investigate further, we analyzed the loop anchor regions harboring HOT loci and observed that the number of multi-way contacts on loop anchors (i.e. loci which serve as anchors to multiple loops) correlates with the number of bound DAPs ($\rho=0.84$ p-value<10E-4; Pearson correlation). “

- Lines 251-252: How did the referenced study categorize DAPs? It is important for any manuscript to be self-contained.

We added the explanation and changed the sentence to the following:

“To test this hypothesis, we classified the DAPs into those two categories using the definitions provided in the study (Lambert et al. 2018) 28, where the TFs are classified by manual curation through extensive literature review and supported by annotations such as the presence of DNA-binding domains and validated binding motifs. Based on this classification, we categorized the ChIP-seq signal values into these two groups.”

- Lines 181-185, sentences starting with 'To test' can be moved to the methods, leaving only brief mentions of the statistic tests if needed.

We removed the mentioned sentence and moved to the supplemental methods (1.4).

- Lines 217-220: I find this sentence extremely redundant unless it can offer more specific insights about a particular set of DAPs or if the DAPs are closer/or a proven distal enhancer to a confirmed causal gene.

We removed the mentioned sentence from the text.

- Lines 243-246: How did the authors determine the set DAPs that have stabilizing effects, and how exactly are the 'stabilizing effects' observed/measured?

We added explanations to Supplemental Methods 3.1 and Fig S18, S19.

While addressing this comment we realized that the reported value of the ratio is 1.91x, not 1.7x. We corrected that value in the main text and added the p-value.

- When discussing the phastCons scores analyses, such as in lines 268-271, how did the authors calculate the relationship between phastCons scores and HOT loci, i.e. was the score averaged across the 400-bp locus to obtain a locus-specific conservation score?

Yes, per-locus conservation scores were averaged over the bps of loci. We added this clarification to the methods.

- Line 311: What is the role of the 'control sets' in the analyses of the sequence's relationship with HOT?

In this specific case, the control sets are used as background or negative sets to set up the classification tasks. In other words, we are asking, whether the HOT loci can be distinguished when compared to random chromatin-accessible regions, promoters, or regular enhancers. We clarified this in the text.

- I also find the discussion about different machine learning methods that classify HOT loci based on sequence contexts quite redundant UNLESS the authors decide to go further into the features' importance (such as motifs) in the models that predict/ are associated with HOT loci, which in itself can constitute another study.

We agree with the reviewer, and shortened the part with the discussions of models by limiting it to only 3 main models and moved the rest to the supplemental materials.

- Can the authors clarify where they obtain data on super-enhancers?

We obtained the super-enhancer definitions from the original study (Hnisz et al. 2013, PMID: 24119843) where the super-enhancers were defined for multiple cell lines. We clarified this in the methods.

- Figure 1B, the x and y axis should be clarified.

We clarified it by using MAX as an example case in the figure caption as follows:

“Prevalence of DAPs in HOT loci. Each dot represents a DAP. X-axis: percentage of HOT loci in which DAP is present (e.g. MAX is present in 80% of HOT loci). Y-axis: percentage of total peaks of DAPs that are located in HOT loci (e.g. 45% of all the ChIP-seq peaks of MAX is located in the HOT loci). Dot color and size are proportional to the total number of ChIP-seq peaks of DAP.”

Reviewer #3 (Recommendations For The Authors):

The list of proteins associated with different types of genomic loci at a meta level (enhancers, promoters, and gene body etc.), and an annotation of the genome at the specific loci level.

The authors use a wide range of acronyms throughout the text and figure legends, they do a reasonably good job, but the main text section "HOT-loci are enriched in causal variants" and Figure 8 would be materially improved if they held it to the same standard.

Size is a physical property and not a physicochemical property.

We thank the reviewer for their comments and suggestions. We added a table to supplemental files with detailed annotations of analyzed loci.

We reviewed the section “HOT loci are enriched in causal variants” and corrected a few mismatches in the acronyms.

<https://doi.org/10.7554/eLife.95170.2.sa0>