MOCHI enables discovery of heterogeneous interactome modules in 3D nucleome

(Supplemental Material)

A Supplemental Methods

A.1 Properties of the identified HIMs

Here we describe the properties of HIMs that are not defined in the main text. The features are the topological structural features related to connection patterns of the genes and TFs in a given HIM. The features include motif density, chromatin interaction edge density (or Hi-C edge density), and GRN edge density, which quantify the connection strength between genes/TFs in a HIM in terms of different connection patterns. All of these properties range from 0 to 1.

- 4-node motif M density. It is the ratio between the number of occurrences of the motif M and the total number of possible occurrences of the motif M in a HIM. The maximum motif density is 1, which is achieved when every pair of genes in the cluster have Hi-C interactions and every gene is regulated by each TF in the cluster. The triangle motif density used later in Supplemental Results B.3 is defined similarly.
- Hi-C edge density. It is the density of the sub-Hi-C interaction network induced by the genes in the HIM. The Hi-C edge density at 1 means that every pair of genes are connected by a chromatin interaction with O/E$>$1, where 0 means that no pair of genes are connected. A higher density reflects that the genes in HIM as a unit are more densely packed.
- GRN edge density. It is the density of the sub-GRN induced by the genes and TFs in the HIM. The maximum 1 is achieved when every gene is regulated by every TF in the HIM. The minimum 0 is achieved when TF-gene interaction does not exist in a HIM.

A.2 Data collection and processing

In this work, we use data from five human cell types: GM12878, HeLa, HUVEC, K562, and NHEK. Hi-C data used to construct chromatin interactome were downloaded from GEO with the accession number GSE63525 (Rao et al. 2014). Nodes in chromatin interactome are 10kb genomic loci that contain transcription start sites (TSS) of expressed genes. Undirected edges between two nodes are defined as follows. For two genomic loci on the same chromosome, an edge is defined if (1) their O/E based on KR normalized contact frequency is greater than 1; and (2) they are less than 10Mb away from each other. The distance constraint is used to control the 1D distance between genes farthest apart in HIMs. The cutoff 10Mb is chosen based on the size of A/B compartments (the 99-th percentile of the size of compartments is around 10Mb in 4 out of the 5 cell types). For two genomic loci on different chromosomes, an edge is defined if their Hi-C contact frequency is among top 1% highest inter-chromosomal contact frequencies. We also calculated the A/B compartments for each chromosome using the first principal component of the O/E contact matrix (Lieberman-Aiden et al. 2009). Processed replication timing data with the GEO accession number GSE34399 (Hansen et al. 2010; Thurman et al. 2007) were downloaded from the UCSC Genome Browser (Rosenbloom et al. 2012).

GRN data were downloaded from Marbach et al. (2016), where TF-gene interactions were inferred by simultaneously considering TF binding and gene expression. Briefly, an interaction between a TF and a gene is defined if (1) the TF has enriched binding motifs in the enhancer or promoter regions of the gene; and (2) the co-expression level between the TF and the target gene is high.

Protein-protein interactions (PPIs) were downloaded from BioPlex2 (Huttlin et al. 2017), BioGrid (Chatr-Aryamontri et al. 2012), CORUM (Ruepp et al. 2009), and STRING (Franceschini et al.

2012). We first extracted the PPIs of the 591 TFs in the GRNs from these public sources. We then combined them into a PPI network after merging duplicated PPIs. Note that the GRNs have the same set of TFs. Thus the PPI network is suitable for all 5 different cell types. The density of the PPI network is 0.158, which is also the expected density of a sub-PPI network of a set of randomly sampled TFs.

Essential genes were downloaded from Wang et al. (2015). However, among the cell lines in the original study, only K562 matches the cell type used in this work. The essential genes identified in K562 were only used to analyze the K562 HIMs. Because the majority of the essential genes are shared among cell lines (Wang et al. 2015), we used the union of them (2741 genes) as the essential gene list in GM12878, HeLa, HUVEC, and NHEK cell types.

RNA-seq data with GEO accession GSE33480 were downloaded from the ENCODE portal (Davis et al. 2018) at: https://www.encodeproject.org. The gene expression level quantified in FPKM values across the 5 cell types were normalized by quantile normalization followed by log-transformation by the function $\log_{10}(1+x)$. From the expression data, we constructed a list of cell type-specific genes for each cell type by the following two criteria: Given a cell type, (1) the gene expression value in the given cell type is higher than 0.1; and (2) the ratio between the gene expression value in the given cell type and the median gene expression value in the other 4 cell types is higher than 2.

Note that the analysis in this work integrates chromosome structural properties and transcriptional regulation features using multiple datasets with hg19 (GRCh37) as the reference human genome. The hg38 (GRCh38) version of the reference human genome mainly improves the centromeric and repetitive regions as compared to hg19, which were not part of our analysis. Therefore, using the hg38 reference human genome would not significantly affect our conclusions in this work.

A.3 Clusters are near optimal when $\alpha = 4/3$

Here we briefly prove that the two clusters from Steps (1) and (2) described in the algorithm in the main text are near optimal when $\alpha = 4/3$ (in Eq. (3) in the main text). Without loss of generality, we prove that the two clusters S and $\bar{S}$ of the original heterogeneous network G are near optimal. By definition, $\varphi_M(S) = \varphi_M(\bar{S})$, we therefore only show that S is near optimal. Our proof follows the same strategy of the proofs in Benson et al. (2016). We first formally state the near optimal claim and prove it in the subsequent paragraphs. Here S being near optimal refers to the motif Cheeger inequality:

$$\varphi_M(S) \leq 4\sqrt{\varphi_M^\star} \leq 1, \tag{7}$$

where $\varphi_M^\star$ is the minimum of subgraph conductance over all possible sets of nodes in G. In other words, near optimal means that S is at most a quadratic factor away from the optimal cluster that achieves $\varphi_M^\star$.

We recall and define some notations first. Let N be the total number of nodes in G. Let M be the subgraph with four nodes and five interactions, where the four nodes are 2 TFs and 2 genes. Four of the five interactions are interactions between TFs and genes. The fifth interaction is between the two genes. Let V_M be the set of the four nodes of M. Let $|V_M|$ denote the cardinality of V_M. Here $|V_M| = 4$. Let $\mathbb{M}$ be the set of the occurrences of M in G. Let W_M denote the subgraph adjacency matrix where $[W_M]_{ij} = \sum_{M \in \mathbb{M}} \mathbb{1}(i \in V_M, j \in V_M)$. The undirected, weighted network induced by W_M is denoted as G_M. The subgraph conductance $\varphi_M(S)$ for G is defined in Eq. (8) and the conductance $\varphi_{G_M}(S)$ for G_M

is defined in Eq. (9):

$$\varphi_M(S) = \frac{\text{cut}_M(S,\bar{S})}{\min[\text{Vol}_M(S), \text{Vol}_M(\bar{S})]} \tag{8}$$

$$\varphi_{G_M}(S) = \frac{\text{cut}_{G_M}(S,\bar{S})}{\min[\text{Vol}_{G_M}(S), \text{Vol}_{G_M}(\bar{S})]} \tag{9}$$

First, we prove that $\text{cut}_M(S,\bar{S}) = \frac{1}{3}\text{cut}_{G_M}(S,\bar{S})$. This proof follows the proof in Benson et al. (2016) and we refer readers to Benson et al. (2016) for further details. Let $X = (x_1, x_2, \ldots, x_N)$ be the vector denoting which nodes belong to S. If node i belongs to S, then $x_i = -1$; otherwise, $x_i = 1$. Let v_1, v_2, v_3, v_4 be the four nodes in an occurrence of the subgraph $M \in \mathbb{M}$. We first transform $|V_M \cap S|$ into sum of $x_{v_i}x_{v_j}$, $1 \leq i < j \leq 4$. $|V_M \cap S| = 3$ means that only one node in M is not assigned to S. Without loss of generality, we assume that node v_1 is not assigned to S and nodes v_2, v_3, v_4 are assigned to S. Thus $x_{v_1} = 1$ and $x_{v_2} = x_{v_3} = x_{v_4} = -1$. $x_{v_1}x_{v_2}$, $x_{v_1}x_{v_3}$, and $x_{v_1}x_{v_4}$ all equal to -1. Also, $x_{v_2}x_{v_3}$, $x_{v_2}x_{v_4}$, and $x_{v_3}x_{v_4}$ all equal to 1. We have:

$$|V_M \cap S| = 3 \text{ is equivalent to } -x_{v_1}x_{v_2} - x_{v_1}x_{v_3} - x_{v_1}x_{v_4} - x_{v_2}x_{v_3} - x_{v_2}x_{v_4} - x_{v_3}x_{v_4} = 0.$$

Note that this also holds when $|V_M \cap S| = 1$. Next, $|V_M \cap S| = 2$ means that two nodes in M are not assigned to S. Again, without loss of generality, we assume node v_1, v_2 are not assigned to S and nodes v_3, v_4 are assigned to S. Thus $x_{v_1} = 1$, $x_{v_2} = 1, x_{v_3} = -1, x_{v_4} = -1$. $x_{v_1}x_{v_2}$ and $x_{v_3}x_{v_4}$ equal to 1. Also, $x_{v_1}x_{v_3}$, $x_{v_1}x_{v_4}$, $x_{v_2}x_{v_3}$, and $x_{v_2}x_{v_4}$ all equal to -1. We have:

$$|V_M \cap S| = 2 \text{ is equivalent to } -x_{v_1}x_{v_2} - x_{v_1}x_{v_3} - x_{v_1}x_{v_4} - x_{v_2}x_{v_3} - x_{v_2}x_{v_4} - x_{v_3}x_{v_4} = 2.$$

Therefore,

$$\begin{aligned}
\text{cut}_M(S,\bar{S}) &= \sum_{M\in\mathbb{M}} \mathbb{1}\left(|V_M \cap S| \in \{1,3\}\right) + \frac{4}{3}\sum_{M\in\mathbb{M}} \mathbb{1}\left(|V_M \cap S| = 2\right) \\
&= \sum_{M\in\mathbb{M}} \frac{6\,\mathbb{1}\left(|V_M \cap S| \in \{1,3\}\right) + 8\,\mathbb{1}\left(|V_M \cap S| = 2\right)}{6} \\
&= \sum_{M\in\mathbb{M}} \frac{6 - x_{v_1}x_{v_2} - x_{v_1}x_{v_3} - x_{v_1}x_{v_4} - x_{v_2}x_{v_3} - x_{v_2}x_{v_4} - x_{v_3}x_{v_4}}{6} \\
&= \sum_{M\in\mathbb{M}} \frac{\frac{3}{2}(x_{v_1}^2 + x_{v_2}^2 + x_{v_3}^2 + x_{v_4}^2) - (x_{v_1}x_{v_2} + x_{v_1}x_{v_3} + x_{v_1}x_{v_4} + x_{v_2}x_{v_3} + x_{v_2}x_{v_4} + x_{v_3}x_{v_4})}{6}
\end{aligned}$$

Following the proof in Benson et al. (2016), we have:

$$\text{cut}_M(S,\bar{S}) = \frac{\frac{1}{2}x^T D_M x - \frac{1}{2}x^T W_M x}{6} = \frac{2 \times \text{cut}_{G_M}(S,\bar{S})}{6} = \frac{1}{3}\text{cut}_{G_M}(S,\bar{S}).$$

Next, we show that $\text{Vol}_M(S) = \frac{1}{3}\text{Vol}_{G_M}(S)$. One key step is to connect whether $i \in V_M$ to whether $|\{i,j\} \cap V_M| = 2$, $j \neq i$. Note that the motif M has 4 nodes, i.e., $|V_M| = 4$. If node i is a node of a given M ($i \in V_M$), there are 3 other nodes in M. For any one of the other 3 nodes, denoted by j, i being a node of M ($i \in V_M$) is equivalent to both i and j being part of M ($|\{i,j\} \cap V_M| = 2$). Therefore, both $\mathbb{1}(i \in V_M)$ and $1/3 \sum_{j \in V_M, j\neq i} \mathbb{1}(|\{i,j\} \cap V_M| = 2)$ equal to 1. On the other hand, if i is not a node of M

($i \notin S$), then i and j cannot both be the nodes of M ($|\{i,j\} \cap V_M| \neq 2$) for any node j in M. Therefore, both $\mathbb{1}(i \in V_M)$ and $1/3 \sum_{j \in V_M, j \neq i} \mathbb{1}(|\{i,j\} \cap V_M| = 2)$ equal to 0. Taken together, regardless whether i is a node of M, $\mathbb{1}(i \in V_M) = 1/3 \sum_{j \in V_M, j \neq i} \mathbb{1}(|\{i,j\} \cap V_M| = 2)$. Therefore,

$$\begin{aligned}
\mathrm{Vol}_M(S) &= \sum_{i \in S} \sum_{M \in \mathbb{M}} \mathbb{1}(i \in V_M) = \sum_{i \in S} \sum_{M \in \mathbb{M}} \frac{1}{3} \sum_{j \in V_M, j \neq i} \mathbb{1}(|\{i,j\} \cap V_M| = 2) \\
&= \sum_{i \in S} \sum_{M \in \mathbb{M}} \frac{1}{3} \sum_{j=1}^{N} \mathbb{1}(|\{i,j\} \cap V_M| = 2) = \frac{1}{3} \sum_{i \in S} \sum_{j=1}^{N} \sum_{M \in \mathbb{M}} \mathbb{1}(|\{i,j\} \cap V_M| = 2) \\
&= \frac{1}{3} \sum_{i \in S} \sum_{j=1}^{N} [W_M]_{ij} = \frac{1}{3} \mathrm{Vol}_{G_M}(S).
\end{aligned}$$

Given that $\mathrm{cut}_M(S, \bar{S}) = \frac{1}{3}\mathrm{cut}_{G_M}(S)$ and $\mathrm{Vol}_M(S) = \frac{1}{3}\mathrm{Vol}_{G_M}(S)$, we have $\varphi_M(S) = \varphi_{G_M}(S)$. For the undirected, weighted graph G_M, Chung (2007) showed that S satisfies the Cheeger inequality, i.e.,

$$\varphi_{G_M}(S) \leq 4\sqrt{\varphi^{\star}_{G_M}} \leq 1, \tag{10}$$

where $\varphi^{\star}_{G_M}$ is the minimum of conductance over all possible sets of nodes G_M. Replacing $\varphi_{G_M}(S)$ by $\varphi_M(S)$ and $\varphi^{\star}_{G_M}$ by $\varphi^{\star}_M$ in Inequality (10) yields that S satisfies the Cheeger inequality for G, i.e., we have proved the Inequality (7) and that S is near optimal.

Comparison with the proofs in Benson et al. (2016)

In Step (2) of the algorithm described in the main text, we apply a spectral clustering method to find two sets S and $\bar{S}$ in the undirected, weighted network G_M that is induced by W_M. The spectral clustering method is the same as the method in Benson et al. (2016) where W_M is computed based on a homogeneous motif and homogeneous network. Benson et al. (2016) proved that S and $\bar{S}$ are near optimal for their case. However, the results in Benson et al. (2016) are not directly applicable to our situation, because our input network and motif are heterogeneous, and converting the heterogeneous network and motif to homogeneous network and motif will mis-count the occurrences of the heterogeneous motif M.

A.4 Pseudocode for the MOCHI algorithm

Algorithm 1 MOCHI

```
Require: Original graph $G_0$, motif $M$, threshold $t_1$.
Ensure: Network motif based clusters
function ITERATIVE SPECTRAL CLUSTERING($G_0, M, t_1$)
    $W_M(G_0) \leftarrow$ Motif adjacency matrix for $G_0$ based on motif $M$
    $S_0, \bar{S}_0, Score_0 =$ SPECTRAL CLUSTERING($W_M(G_0), N$)
    $L \leftarrow \{G_0\}$
    while $\exists G_i \in L$ such that $Score_i < t_1$, do
        $G_k \leftarrow \mathrm{argmin}_i Score_i$ the graph with the lowest corresponding score
        $G_{S_k}, G_{\bar{S}_k} \leftarrow$ Graph for node set $S_k, \bar{S}_k$, respectively
        $W_M(G_{S_k}), W_M(G_{\bar{S}_k}) \leftarrow$ Motif adjacency matrix for $G_{S_k}, G_{\bar{S}_k}$
```

```
            Drop all-zero rows and columns in $W_M(G_{S_k}), W_M(G_{\bar{S}_k})$ and the corresponding nodes in $G_{S_k}, G_{\bar{S}_k}$
            $N_{S_k}, N_{\bar{S}_k} \leftarrow$ Node size of $G_{S_k}, G_{\bar{S}_k}$, respectively
            $S_{S_k}, \bar{S}_{S_k}, Score_{S_k} = \textsc{Spectral Clustering}(W_M(G_{S_k}), N_{S_k})$
            $S_{\bar{S}_k}, \bar{S}_{\bar{S}_k}, Score_{\bar{S}_k} = \textsc{Spectral Clustering}(W_M(G_{\bar{S}_k}), N_{\bar{S}_k})$
            $L \leftarrow \{..., G_{k-1}, G_{k+1}, ..., G_{S_k}, G_{\bar{S}_k}\}$
        end while
    end function

    function Spectral Clustering($W_M, N$)
        $D \leftarrow$ Diagonal Matrix$_{(1:N \times 1:N)}$ given by $D_{ii} = \sum_{j=1}^{N} [W_M]_{ij}$
        $L \leftarrow D^{-\frac{1}{2}}(D - W_M)D^{-\frac{1}{2}}$
        $v = \{v_1, v_2, ..., v_N\} \leftarrow$ Eigenvector of $L$
        $v_k \in v \leftarrow$ Eigenvector of the second smallest eigenvalue
        $O \leftarrow D^{-\frac{1}{2}} v_k$
        $\alpha_i \leftarrow$ Index of the $i$-th smallest value in $O$
        $S \leftarrow \text{argmin}_{S_k} \varphi_{G_M}(S_k)$, where $S_k = \{\alpha_1, ..., \alpha_k\}$
        $Score \leftarrow \varphi_{G_M}(S)$
        return $S, \bar{S}, Score$
    end function
```

A.5 Runtime analysis

Here we analyze the computational complexity of MOCHI. In practice, the most time-consuming step would be the construction of the motif adjacency matrix W_M and the calculation of the eigenvector for the normalized Laplacian matrix. Although in general for eigenvalue decomposition of a matrix size of $N \times N$ the runtime would be $O(N^3)$, using fast symmetric diagonally dominant solvers for Laplacian matrix, we can reach near linear time for this process (Kelner et al. 2013). Therefore, in the rest of this section, we will only discuss the runtime of the matrix construction part.

Intuitively, for a 4-node motif, we can calculate W_M by checking every combination of 4 nodes in the graph, and has the runtime of $O(N^4)$, where N is the number of nodes in the graph. However, since we only deal with a special 4-node motif that consists of 2 different types of nodes, which can be treated as a combination of two specific 3-node motifs (one TF regulates two genes), if we use T for the TF nodes in the graph, and C for the chromatin loci nodes, and t, c for the size of these nodes, respectively, we can derive the runtime as follows. For $[W_M]_{ij}$ in the motif adjacency matrix, where $i \in C, j \in C$, it is equivalent to finding the number $n_{3_{ij}}$ of specific 3-node motif that i and j share, then using the combination number to get the number of 4-node motif $[W_M]_{ij} = \binom{n_{3_{ij}}}{2}$. For the search of triangles, given i and its neighbor j, we consider all the TF nodes that they share, which gives us the runtime of $O(tc^2)$. For $[W_M]_{ij}$ where $i \in T, j \in C$, since we already know how many 3-node motifs would form between j and any another locus k (as calculated above), we can calculate the number of 4-node motifs involving i, j by counting the 3-node motifs using the similar method. The runtime of this part would also be $O(tc^2)$. Finally, for the $[W_M]_{ij}$ where $i \in T, j \in T$, we cannot count the 3-node triangle anymore. To count the number of 4-node motifs involving i, j, we find out the common loci they share, and then calculate the total number of edges between these common loci. Together, the runtime for the whole algorithm is dominated by the construction of W_M, especially for the part between TF and TF. The worst case runtime would be $O(t^2c^2)$, where we have to go over all the combination of two TFs and two gene loci. Note that, however, TFs only make up a small proportion of the nodes (about 4.5%). Also, since the network is sparse, we usually do not need to go over all the combination of nodes, which would accelerate the computation further. In addition, in the actual implementation, we use parallel computation to speed up the process, which makes the entire algorithm efficient in practice.

B Supplemental Results

B.1 HIMs are robust to the parameters used to construct the heterogeneous networks

We show that the identified HIMs are robust to the parameters used to define the heterogeneous networks. In all the analyses presented in the main text, we use a cutoff at 1 for "observed over expected" (O/E) quantity to filter intra-chromosomal Hi-C contacts when defining chromatin interaction networks. To test the robustness of HIMs, we constructed multiple sets of chromatin interaction networks by varying the cutoff between 1 and 2 (for O/E). The number of intra-chromosomal chromatin interactions with the cutoff at 2 is 64.6-81.9$\%$ of the number of intra-chromosomal interactions with the cutoff at 1 across five cell types. We denote the chromatin interaction networks derived from cutoffs higher than 1 as sub-Hi-C. Regarding GRNs, we also constructed sub-GRNs for each cell type by only keeping the top $k\%$ interactions with the highest scores, $k = 60, 70, 80, 90$. We then constructed different heterogeneous networks for each cell type by enumerating different combinations of O/E cutoffs and different GRNs.

We then applied MOCHI with the 4-node motif M to each of these heterogeneous networks of each cell type. We use adjusted Rand index to quantify the similarities on gene memberships between the HIMs from two different heterogeneous networks. For example, if the assignments of genes to HIMs are identical between two sets of HIMs, then adjusted Rand index would be 1. We use hierarchical clustering to group the sets of HIMs with similar adjusted Rand index. As shown in Fig. S3, we found that the sets of HIMs from the heterogeneous networks of the same cell type are much more similar to each other than the sets of HIMs from the other cell types. Hierarchical clustering produces five major clusters (Fig. S3). Each cluster contains different heterogeneous networks in the same cell type. This analysis suggests that HIMs are robust to the parameters used to construct the chromatin interactome and GRNs.

B.2 HIMs share similar connections with 3D genome features across cell types

We found that the HIMs in 5 cell types in this study share similar connections with 3D genome organization features, such as A/B compartments, TADs, and loops. We looked at the genomic regions of each HIM that is the smallest genomic block containing the TSS of the genes in the HIM. The median size of the genomic regions of the HIMs ranges from 4.9Mb in NHEK to 8Mb in GM12878 (Table S2), comparable to the size of A/B compartments (median size is 5Mb). The median numbers of TADs in the genomic regions of the HIMs are 3-4 in different cell types (Table S2). The genomic regions of the HIMs have, on average, 7 chromatin loops in GM12878 and 2-4 loops in the other cell types (Table S2). The HIMs in GM12878 involve a higher number of loops, which is perhaps because GM12878 has at least 60$\%$ more detected loops than the other cell types possibly due to higher Hi-C coverage.

B.3 Justification of the 4-node motif *M*

To justify the choice of the motif M, we compared it with two different types of motifs. One is the triangle motif with 3 nodes (Fig. S1A), where two of them are genes with a chromatin interaction and the third node is a TF that regulates both genes. The triangle motif does not explicitly encode co-regulation between TFs. Another motif is the bifan motif with 4 nodes (Fig. S1A), where two nodes are TFs that co-regulate two genes but there is no chromatin interaction between the two genes. The bifan motif does not explicitly encode spatial proximal relationship between genes. For bifan motif, we applied MOCHI with the bifan on the GRN and then split the identified HIMs by chromosome number.

We found that our 4-node motif M and the triangle motif are better than the bifan motif in terms of identifying clusters (Fig. S1B). Compared to the HIMs identified by the bifan motif, the HIMs by the motif M and triangle motif have higher Hi-C edge density, higher triangle density, higher motif M density. Moreover, the genes in a HIM are closer to each other in the 1D sequence space, although the HIMs from the bifan motif have a smaller number of genes as compared to the HIMs based on the 4-node motif M and the triangle motif. This result highlights that the chromatin interaction between the two target genes in a motif is important to capture spatial proximity between the genes in HIMs.

In addition, we found that our 4-node motif M is better than the triangle motif. We comprehensively compared the identified HIMs by the motif M and the triangle motif. The HIMs identified by the two motifs have similar numbers of genes as the median numbers of genes are equivalent in 4 out of 5 cell types. A similar pattern is observed for the Hi-C edge density. Specifically, the density is only significantly different in two cell types: GM12878 and NHEK ($p \leq 0.04$), although the difference is small (the median density is 0.015 in GM12878 and 0.028 in NHEK) (Fig. S1B). However, the identified HIMs by the two motifs are very different in other features. Compared to the HIMs from the triangle motif, the HIMs from the 4-node motif M have much higher numbers of TFs ($p \leq 2.45 \times 10^{-16}$). The difference in the median number of TFs ranges from 4 to 8 (Fig. S2). Even though the HIMs from the 4-node motif M have higher number TFs, they have comparable numbers of genes as compared to the HIMs from the triangle motif. They also have higher triangle density and higher 4-node motif M density ($p \leq 0.027$ and $p \leq 1.43 \times 10^{-3}$, respectively; Fig. S1B). The HIMs from M also have a higher proportion of genes in A compartment in 4 cell types ($p \leq$0.022), replicate much earlier ($p \leq 6.74 \times 10^{-3}$), and have smaller replication timing coefficient of variation ($p \leq 4.38 \times 10^{-5}$) (Fig. S2).

Next, we compared the features after adjusting the number of TFs and the number of genes of the identified HIMs. The HIMs from the 4-node motif have much higher numbers of TFs than the HIMs from the triangle motif. The number of genes is slightly different in some cell types. Since the features could be biased to the number of TFs and the number of genes, we compared the features by adjusting the number of genes and the number of TFs in HIMs based on a linear regression model:

$$Y = \beta_0 + \beta_1 \times \text{\# TFs} + \beta_2 \times \text{\# genes} + \beta_3 \times \mathbb{1}_{\text{motif}}, \tag{11}$$

where Y is a given feature, $\mathbb{1}_{\text{motif}} = 1$ if the HIM is identified with the 4-node motif M, $\mathbb{1}_{\text{motif}} = 0$ if the HIM is identified with the triangle motif, and $\hat{\beta}_3$ indicates the averaged difference in the feature Y between the HIMs identified by the two motifs after adjusting the number of TFs and the number of genes in the HIMs. Specifically, a positive $\hat{\beta}_3$ means that HIMs from the 4-node motif have higher feature Y than the HIMs from the triangle motif. On the other hand, a negative $\hat{\beta}_3$ means lower Y in HIMs from the 4-node motif M. The detailed $\hat{\beta}_3$ values for the features are in Table S3. Overall, the differences are still significant after adjusting the number of TFs and the number of genes. Take together, our 4-node motif M is better than the triangle motif in identifying HIMs.

B.4 Comparison between HIMs and conventional GRN clusters

One key difference between HIMs and conventional GRN clusters is that HIMs have spatial constraints such that genes in HIMs are in spatial contact more frequently than expected in the nucleus. To characterize the potential advantages of HIMs over conventional GRN clusters, we constructed gene clusters based on the GRN data used in this study by modifying the MOCHI framework. Note that the conventional GRN clusters were constructed using GRN data only. For a fair comparison, the MOCHI framework was modified such that the motif is a single GRN edge and the parameters are tuned such that identified clusters have the same median number of genes as in HIMs. We call the resulting clusters GRN clusters.

In the main text, we reported the comparison between HIMs and GRN clusters in terms of gene expression, replication timing, and the enrichment over genes affected by eQTLs and GWAS SNPs. Additionally, we found that genes in HIMs are more spatially proximal to each other than those in GRN clusters (Hi-C edge density, Fig. S5), again confirming the utility of adding spatial constraint to HIMs. Regarding cluster enrichment in KEGG pathways and Gene Ontology (GO) terms, we found that HIMs and GRN clusters have 25.86-41.65% of clusters enriched in at least one functional terms in each category (% of clusters enriched in KEGG pathways, GO biological process levels 5 and 6 terms ranges 25.96-29.09%, 36.42-41.65% and 25.86-32.43%, respectively). However, the differences in the proportions between HIMs and GRN clusters are not statistically significant (percentage increase in the proportion of enriched clusters in KEGG pathways and GO biological process ranges from -11.61% to 21.47%, $p \geq 0.057$), possibly because KEGG pathways and GO terms have weak association with genome spatial organization.

B.5 HIMs harbor long-range enhancers

We determine the existence of a long-range enhancer in a HIM if the enhancer connects to genes in the HIM through long-range chromatin loops. We downloaded enhancers inferred by ChromHMM and Segway from the ENCODE portal (https://www.encodeproject.org) (Davis et al. 2018) and obtained the chromatin loops from Rao et al. (2014). We call an long-range enhancer-gene interaction for genes in a HIM if (1) the enhancer and gene are within 20kb to the loop anchor loci, (2) the enhancer and gene are 100kb to 1Mb away from each other, and (3) the enhancer is located within the genomic region covered by the HIM. An example in K562 is highlighted by two crossed gray boxes in Fig. S4H, where enhancer Chr10:6443523-6443803 is connected to upstream essential gene *RBM17* (Wang et al. 2015) and *PFKFB3* by a chromatin loop that is 270kb in length. Enhancer-gene interactions satisfy conditions (1) and (3) but <100kb away to gene in condition (2) are called short-range enhancer-gene interactions. We found that a large proportion of (80.3%, 522) GM12878 HIMs have at least one long-range enhancer-gene interactions. Among those 522 GM12878 HIMs, on average, 8 long-range enhancers are connected to 2 (25%) genes. Only a small proportion (7.08%, 46) of GM12878 HIMs have their genes connected to short-range but not long-range enhancers. Similar patterns are observed in other cell types (Fig. S7A). Next, we focused on HIMs having multiple enhancers, including both long-range and short-range enhancers, connected to the majority (≥50%) of their genes. Three such HIMs are shown in detail in Fig. S4F-H. Overall, we found that 118 (18.15%) such HIMs in GM12878 cell line. Among these HIMs, the maximum 1D distance between the connected enhancer and the genes in each HIM ranges from ∼50kb to ∼933kb with median at 416kb. The proportion of such HIMs is significantly higher than that of the conventional GRN clusters (Fig. S7B). HIMs have consistently significant improvement over conventional GRN clusters in other cell types, and with a smaller (500kb) window in condition (2) to define potential long-range enhancer-gene interaction in a HIM (Fig. S7B). Together, we found that long-range enhancers involve in HIMs, further demonstrating the benefit of having spatial constraints to HIMs and their biological significance.

B.6 Comparisons between conserved and cell type-specific HIMs

Overall, 40.69-47.38% of the identified HIMs in each cell type are cell type-specific HIMs. We found that the conserved and cell type-specific HIMs have distinct properties of interactomes across cell types. Compared to cell type-specific HIMs, conserved HIMs exhibit stronger clustering features with higher Hi-C edge density, higher GRN edge density, and higher motif M density ($p \leq 8.15 \times 10^{-4}$; Fig. S16). Also, conserved HIMs tend to be closer to the nuclear interior with a higher proportion of their genes in A compartment, and their genes replicate earlier and more synchronously (Fig. S16). Moreover, we found that conserved HIMs and cell type-specific HIMs tend to have large differences in gene expression level and cell type-specific genes. The conserved HIMs have higher mean gene expression level than the cell type-specific HIMs in 3 cell types except for NHEK and HUVEC (p <0.05; Fig. S17). On the other hand, cell type-specific HIMs have a higher proportion of cell type-specific genes (p ≤0.02) than conserved HIMs across 5 cell types (Fig. S17). Note that the genes in both types of HIMs are expressed significantly higher than the genes (in the networks) that are not assigned to HIMs ($p < 2.22 \times 10^{-16}$). Taken together, our results demonstrate that conserved and cell type-specific HIMs in general have distinct network properties, spatial location preference, and functional characteristics.

References

A. R. Benson, D. F. Gleich, and J. Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, 2016.

A. Chatr-Aryamontri, B.-J. Breitkreutz, S. Heinicke, L. Boucher, A. Winter, C. Stark, J. Nixon, L. Ramage, N. Kolas, L. ODonnell, et al. The BioGRID interaction database: 2013 update. *Nucleic Acids Research*, 41(D1):D816–D823, 2012.

F. Chung. Four Cheeger-type inequalities for graph partitioning algorithms. *Proceedings of ICCM, II*, pages 751–772, 2007.

C. A. Davis, B. C. Hitz, C. A. Sloan, E. T. Chan, J. M. Davidson, I. Gabdank, J. A. Hilton, K. Jain, U. K. Baymuradov, A. K. Narayanan, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Research*, 46(D1):D794–D801, 2018.

N. C. Durand, J. T. Robinson, M. S. Shamim, I. Machol, J. P. Mesirov, E. S. Lander, and E. L. Aiden. Juicebox provides a visualization system for hi-c contact maps with unlimited zoom. *Cell Systems*, 3 (1):99–101, 2016.

A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. Von Mering, et al. STRING v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, 41(D1):D808–D815, 2012.

R. S. Hansen, S. Thomas, R. Sandstrom, T. K. Canfield, R. E. Thurman, M. Weaver, M. O. Dorschner, S. M. Gartler, and J. A. Stamatoyannopoulos. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences*, 107(1): 139–144, 2010.

E. L. Huttlin, R. J. Bruckner, J. A. Paulo, J. R. Cannon, L. Ting, K. Baltier, G. Colby, F. Gebreab, M. P. Gygi, H. Parzen, et al. Architecture of the human interactome defines protein communities and disease networks. *Nature*, 545(7655):505, 2017.

J. A. Kelner, L. Orecchia, A. Sidford, and Z. A. Zhu. A simple, combinatorial algorithm for solving sdd systems in nearly-linear time. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 911–920. ACM, 2013.

E. Lieberman-Aiden, N. L. Van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.

D. Marbach, D. Lamparter, G. Quon, M. Kellis, Z. Kutalik, and S. Bergmann. Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nature Methods*, 13(4):366, 2016.

J. Paulsen, T. M. L. Ali, and P. Collas. Computational 3D genome modeling using chrom3d. *Nature Protocols*, 13(5):1137, 2018.

S. S. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014.

K. R. Rosenbloom, C. A. Sloan, V. S. Malladi, T. R. Dreszer, K. Learned, V. M. Kirkup, M. C. Wong, M. Maddren, R. Fang, S. G. Heitner, et al. ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Research*, 41(D1):D56–D63, 2012.

A. Ruepp, B. Waegele, M. Lechner, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, C. Montrone, and H.-W. Mewes. CORUM: the comprehensive resource of mammalian protein complexes-2009. *Nucleic Acids Research*, 38(suppl_1):D497–D501, 2009.

R. E. Thurman, N. Day, W. S. Noble, and J. A. Stamatoyannopoulos. Identification of higher-order functional domains in the human ENCODE regions. *Genome Research*, 17(6):917–927, 2007.

T. Wang, K. Birsoy, N. W. Hughes, K. M. Krupczak, Y. Post, J. J. Wei, E. S. Lander, and D. M. Sabatini. Identification and characterization of essential genes in the human genome. *Science*, 350(6264):1096–1101, 2015.

C Supplemental Tables

Table S1: Summary of the input heterogeneous networks and the identified HIMs across five cell types. 'Overlapping HIMs (%)' is the proportion of the identified HIMs that share TFs with other HIMs. 'Genes in HIMs (%)' represents the proportion of genes in a heterogeneous network that are assigned to HIMs.

		GM12878	HeLa	HUVEC	K562	NHEK
Input	TFs	591	591	591	591	591
	Genes	11,627	12,036	11,927	12,391	12,161
	TF→gene	1,078,893	998,174	828,303	1,119,395	814,017
	Gene−gene	337,036	164,007	184,866	253,218	139,385
Output	HIMs	650	806	773	802	664
	Overlapping HIMs (%)	72.8	74.7	71.9	74.7	79.4
	Genes in HIMs (%)	69.1	77.2	75.3	76.5	62.1

Table S2: Statistics of the identified HIMs across five cell types.

	GM12878	HeLa	HUVEC	K562	NHEK
Median TF number	9	17	17	15	14
Median gene number	9	9	9	9	9
Median loop number	7	2	2	4	2
Median TAD number	4	3	3	4	3
Median 1D distance between The farthest apart genes (Mb)	8	5.4	6.1	7.2	4.9
# of HIMs inherited TFs	451	599	546	596	497
Median proportion of inherited TFs	28.6	27.8	24.6	25.0	28.0

Table S3: Comparison between the identified HIMs from the 4-node motif M and the triangle motif while adjusting the numbers of TFs and genes in the HIMs by the linear regression model $Y = \beta_0 + \beta_1 \times \text{\# TFs} + \beta_2 \times \text{\# genes} + \beta_3 \times \mathbb{1}_{\text{motif}}$, where Y is a continuous feature, $\mathbb{1}_{\text{motif}} = 1$ if a HIM is identified by the 4-node motif M and 0 otherwise, $\hat{\beta_3} \geq 0$ means that the HIMs identified by the 4-node motif M have higher Y than the HIMs identified by the triangle motif after adjusting the numbers of TFs and genes. *P*-value is computed for the hypothesis that $\beta_3 \neq 0$. The features with *P*-value < 0.05 across 5 cell types are highlighted with bold font.

Y	GM12878		HeLa		HUVEC		K562		NHEK	
	$\hat{\beta_3}$	*P* value	$\hat{\beta_3}$	*P* value	$\hat{\beta_3}$	*P* value	$\hat{\beta_3}$	*P* value	$\hat{\beta_3}$	*P* value
Hi-C edge density	0.005	0.586	0.012	0.155	0.02	0.015	-0.001	0.903	0.026	3.02×10^{-3}
Triangle density	0.11	9.9×10^{-17}	0.071	3.19×10^{-10}	0.074	6.04×10^{-11}	0.056	4.45×10^{-7}	0.077	1.43×10^{-11}
4-node motif M density	0.15	1.93×10^{-22}	0.082	1.83×10^{-11}	0.089	6.25×10^{-13}	0.074	1.35×10^{-9}	0.09	6.91×10^{-13}
% of genes in A compartment	0.038	3.61×10^{-4}	0.036	8.4×10^{-3}	0.022	0.147	0.033	7.81×10^{-4}	0.015	0.289
Mean replication timing	2.76	8.87×10^{-7}	2.269	1.36×10^{-8}	2.485	2.59×10^{-6}	2.527	2.13×10^{-7}	2.673	1×10^{-10}
Replication timing CV	-0.04	3.27×10^{-13}	-0.029	6.88×10^{-9}	-0.032	8.82×10^{-10}	-0.029	2.03×10^{-9}	-0.029	9.36×10^{-12}

Table S4: Blood-related disorders and SNPs that are enriched in HIMs but not in conventional GRN clusters. On the other hand, there is no blood-related disorders and SNPs that are enriched in conventional GRN clusters but not in HIMs. 'Genes shared' refers to the shared genes between HIM and the genes affected by SNPs associated with blood-related disorders.

Cell line	HIM	Chr.	Genes shared	SNP ID	Disease	*P* value
GM12878	484	Chr17	*ZFP3*;*ZNF232*	rs73331351	PEG-asparaginase hypersensitivity without enzyme activity in childhood acute lymphoblastic leukaemia	0.001
GM12878	291	Chr7	*EIF4H*;*LIMK1*	rs193107685	Systemic seropositive rheumatic diseases (Systemic sclerosis or systemic lupus erythematosus or rheumatoid arthritis or idiopathic inflammatory myopathies)	0.001
K562	607	Chr6	*ANKS1A*;*TAF11*;*UHRF1BP1*	rs4646949	Fasting blood insulin	2.3×10^{-4}
K562	332	Chr17	*ZFP3*;*ZNF232*	rs73331351	PEG-asparaginase hypersensitivity without enzyme activity in childhood acute lymphoblastic leukaemia	0.002
K562	39	Chr2	*IL18R1*;*IL1RL1*	rs1420101	White blood cell count (eosinophil)	4.1×10^{-4}

Table S5: The top GO terms or pathways that are enriched in the genes assigned to HIMs consistently or in a cell type-specific manner. The number of genes in each category is shown in Fig. 4A.

	GO term/Pathway	**Count**	**Fold Enrichment**	***P* value**
Constitutive genes	chromosome organization	371	1.50	5.7×10^{-19}
	macromolecular complex subunit organization	653	1.30	1.2×10^{-12}
	regulation of gene expression, epigenetic	105	1.90	3.4×10^{-12}
	RNA processing	287	1.40	4.1×10^{-12}
	nucleosome organization	74	2.10	4.2×10^{-12}
	DNA conformation change	104	1.80	2.2×10^{-11}
	mRNA processing	163	1.60	9.3×10^{-11}
	protein-DNA complex subunit organization	98	1.80	2.0×10^{-10}
	DNA packaging	77	1.90	5.2×10^{-10}
	mRNA metabolic process	213	1.40	2.5×10^{-9}
	RNA splicing	139	1.60	3.4×10^{-9}
	protein localization to organelle	268	1.40	4.2×10^{-9}
	intracellular transport	436	1.30	6.5×10^{-9}
GM12878 specific genes	regulation of lymphocyte activation	28	3.40	6.5×10^{-8}
	regulation of T cell activation	24	3.80	7.4×10^{-8}
	regulation of leukocyte cell-cell adhesion	24	3.60	1.8×10^{-7}
	T cell activation	28	3.00	5.5×10^{-7}
HeLa specific genes	cell development	74	1.50	4.8×10^{-4}
	cell-cell signaling	57	1.60	5.6×10^{-4}
K562 specific genes	phospholipase C-activating G-protein coupled receptor signaling pathway	11	4.80	8.5×10^{-5}
	reproduction	55	1.70	1.3×10^{-4}
	G-protein coupled receptor signaling pathway	33	2.00	1.7×10^{-4}
NHEK specific genes	keratinocyte differentiation	23	10.30	2.5×10^{-16}
	skin development	30	6.60	1.2×10^{-15}
	keratinization	16	18.30	2.0×10^{-15}
	peptide cross-linking	16	17.00	7.0×10^{-15}
	epidermis development	32	5.60	2.6×10^{-14}

Table S6: The dynamics of chromatin interaction networks and GRNs across 5 cell types. Here we show the numbers/proportions of interactions that exist in the corresponding number of cell types in each column. For example, column '1' corresponds to the interactions that only exist in one cell type. Column '5' corresponds to the interactions that exist in all 5 different cell types. Overall, a large proportion of the edges in GRNs and chromatin interaction networks only exist in one cell type.

	Type	1	2	3	4	5
Hi-C networks	# interactions	755,574	242,833	86,790	36,945	23,442
	% of interactions	66.00	21.20	7.60	3.20	2.00
GRNs	# interactions	637,950	457,289	309,733	234,033	389,722
	% of interactions	31.40	22.50	15.30	11.50	19.20

D Supplemental Figures

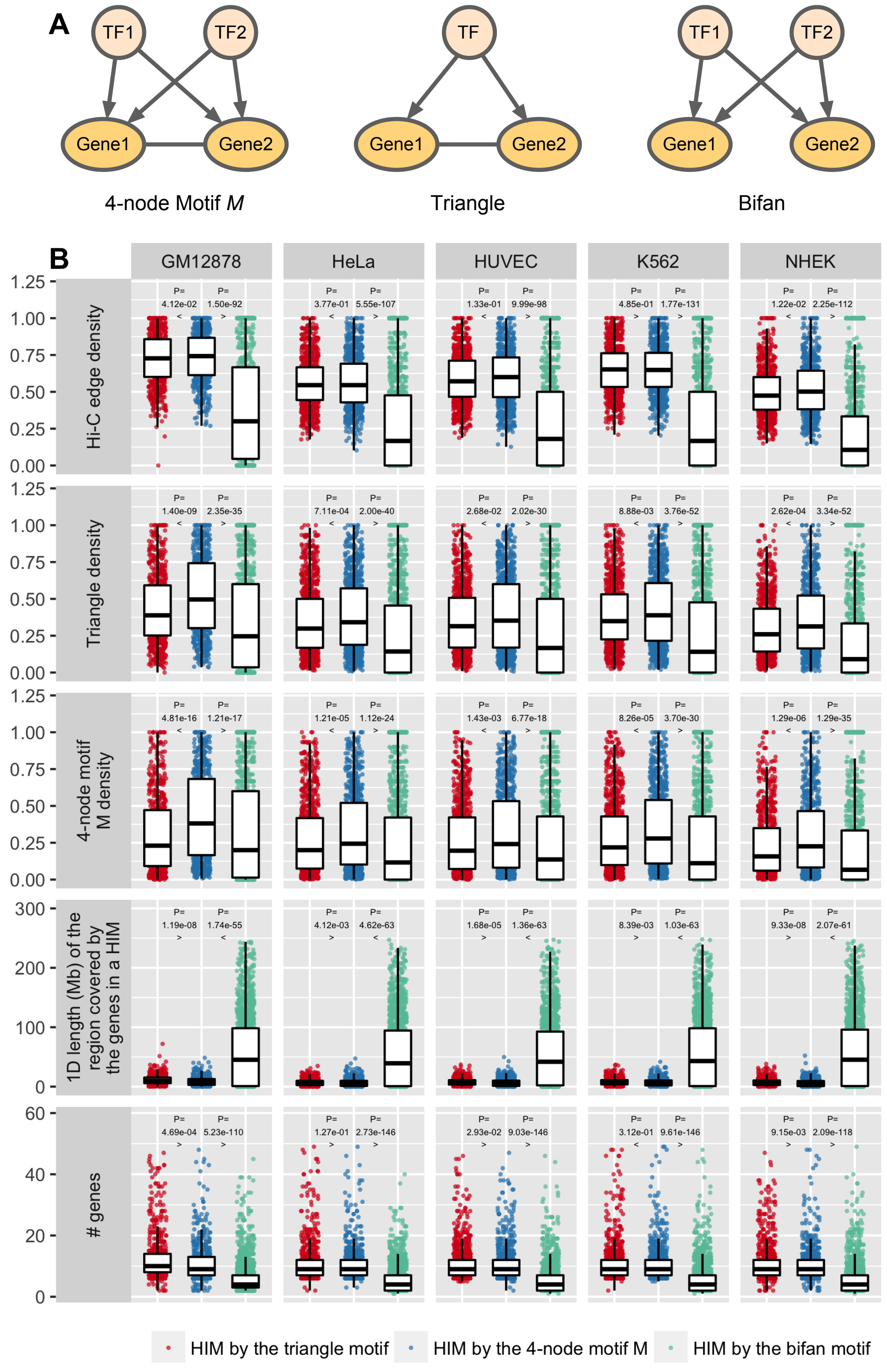

Figure S1: Comparison of the identified HIMs based on the 4-node motif M, the triangle motif, the bifan motif across 5 cell types. **(A)** Illustration of the 3 different motifs. **(B)** Comparison of the HIMs identified by the 3 different motifs. Rows correspond to the features. Columns correspond to the cell types. Each dot represents a HIM.

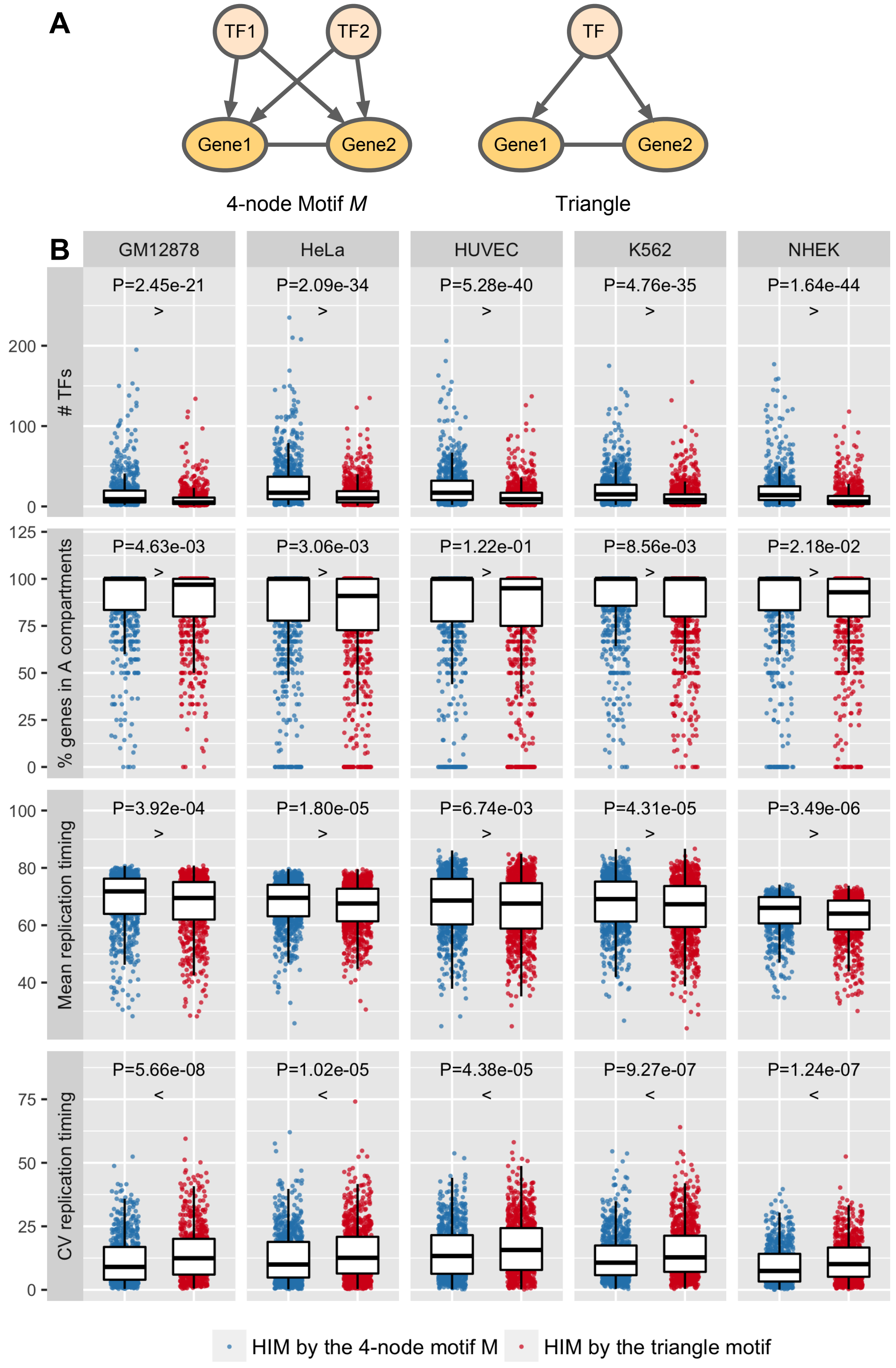

Figure S2: Comparison of the identified HIMs based on the 4-node motif M and the triangle motif across the 5 cell types. **(A)** Illustration of the 4-node motif M and the triangle motif. **(B)** Rows correspond to the features. Columns correspond to the cell types. Each dot represents a HIM identified either by the 4-node motif M or the triangle motif. The patterns are consistently observed after adjusting the numbers of TFs and genes in HIMs by a linear regression model (Table S3).

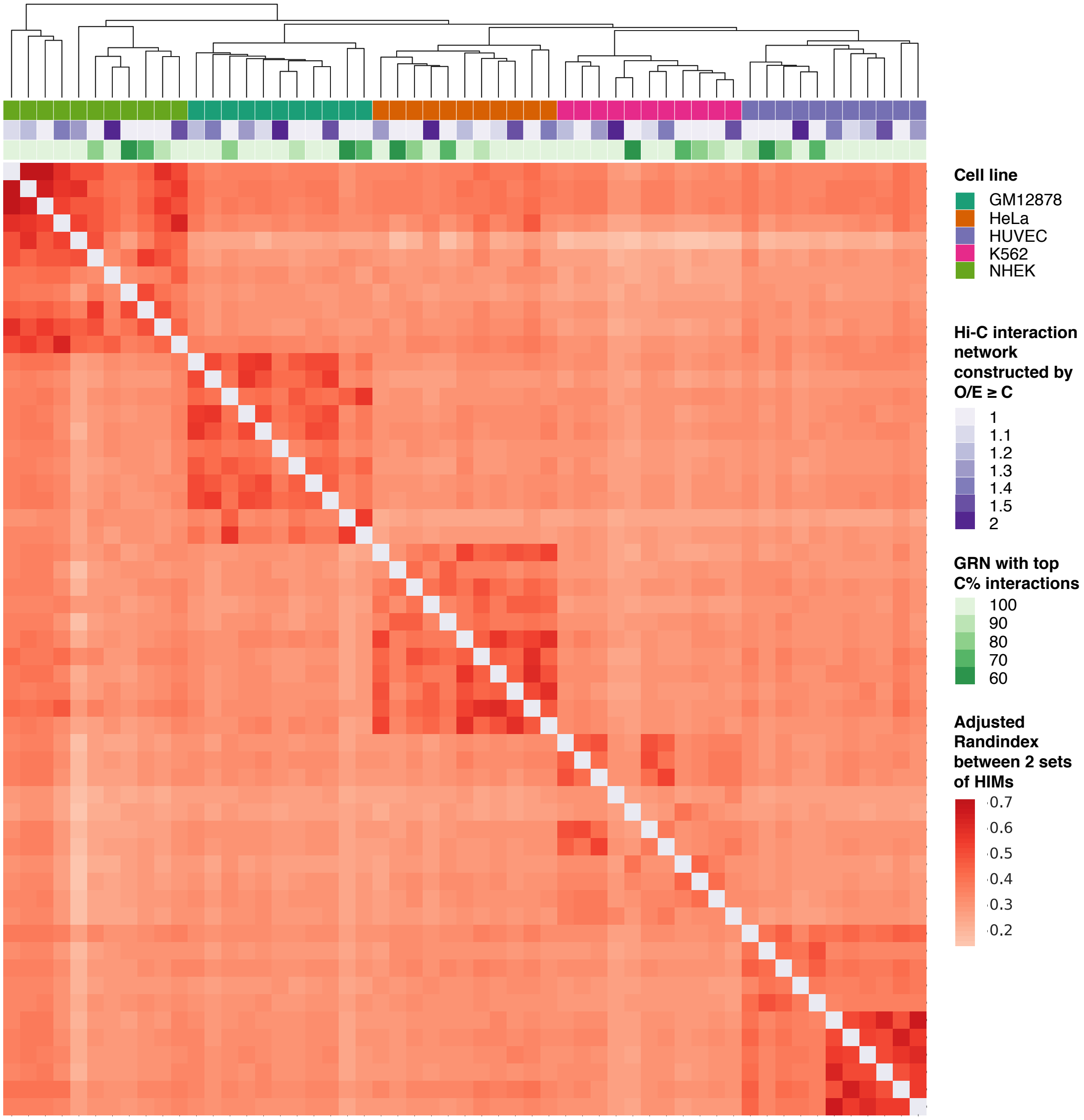

Figure S3: HIMs are robust to the input heterogeneous networks. For each cell type, there are multiple heterogeneous networks derived from different combinations of the Hi-C interaction networks (based on different O/E cutoff ranging from 1 to 2) and the GRN networks. Here we constructed multiple GRNs by keeping only the top k% interactions with highest scores, $60 \leq k \leq 100$, in the whole GRN. In the main text, we used the heterogeneous networks that combines the Hi-C interaction network with O/E≥ 1 and the whole GRN. Adjusted Rand index is used to quantify the similarities on gene memberships between two different sets of HIMs derived from two different heterogeneous networks. The hierarchical clustering is applied to the adjusted Rand index. Note that the sets of HIMs from the same cell type are all in the same cluster.

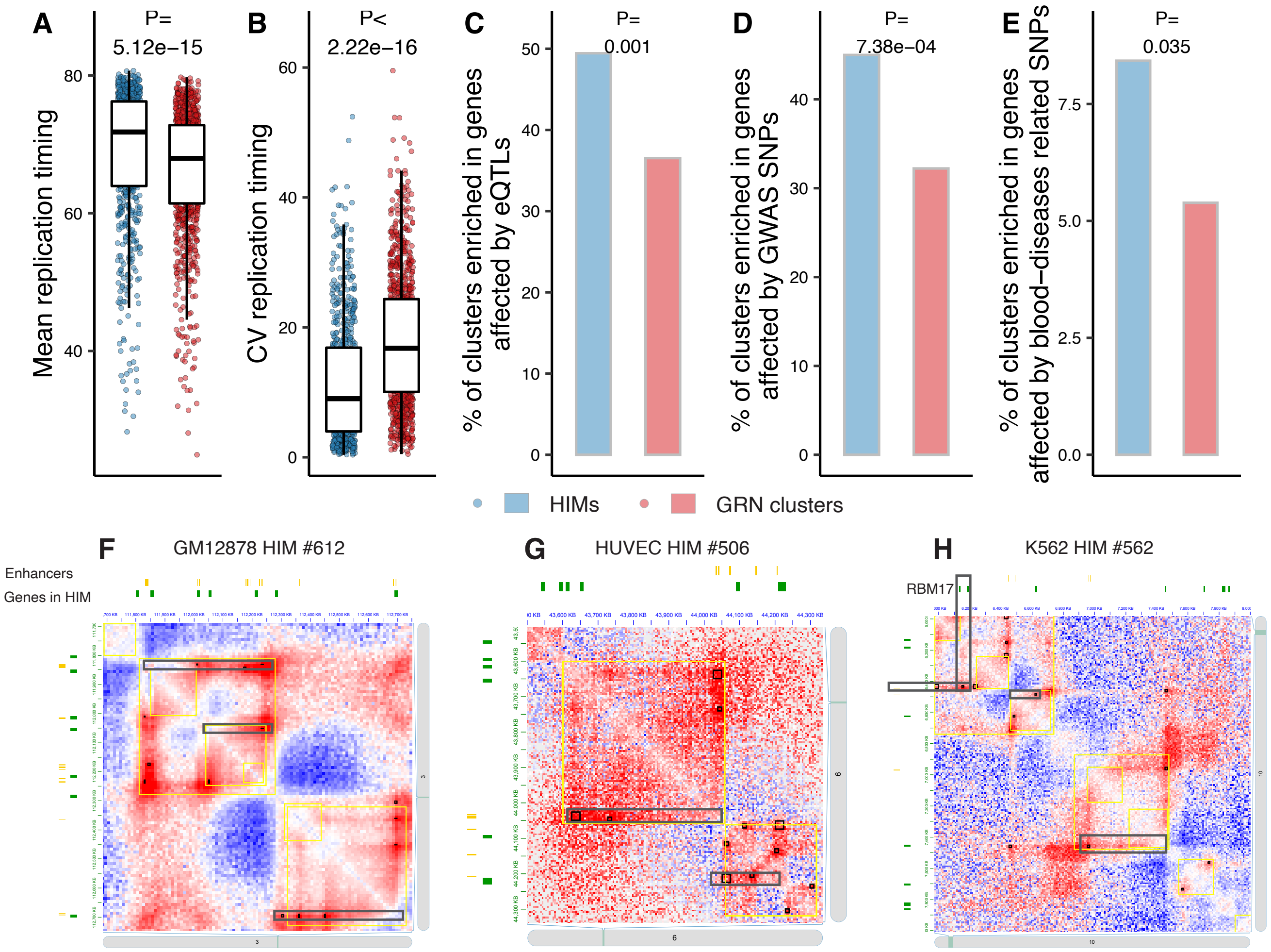

Figure S4: Summary figure of the comparison between HIMs and GRN clusters in GM12878 (**A-E**) and three examples of enhancer-gene interactions in HIMs (**F-H**). **(A)** Boxplot shows the mean replication timing of genes in each cluster. **(B)** Boxplot shows the coefficient of variation (CV) of replication timing of genes in each cluster. **(C)** Barplot shows the proportion of clusters that are enriched in genes affected by eQTLs. **(D)** Barplot shows the proportion of clusters that are enriched in genes affected by GWAS SNPs. **(E)** Barplot shows the proportion of clusters that are enriched in genes affected by SNPs associated with blood-related disorders. **(F-H)** Visualization of 3 HIMs where the majority ($\geq 50\%$) of the genes have enhancer-gene interactions supported by chromatin loops (small black rectangles). Enhancer track only shows enhancers involved in enhancer-gene interactions in HIM. Long-range enhancer-gene interactions are highlighted by gray boxes. Visualization is done using Juicebox (Durand et al. 2016). Heatmap shows Hi-C contacts within a region (red: O/E$>$1; blue: O/E$<$1). Yellow boxes represent TADs. In **(H)**, essential gene *RBM17* in K562 and its long-range enhancer are highlighted by two crossed gray boxes. Comparisons in other cell types and comparison between HIMs and another set of GRN clusters based on the bifan motif are shown in Fig. S5, Fig. S6, and Fig. S7.

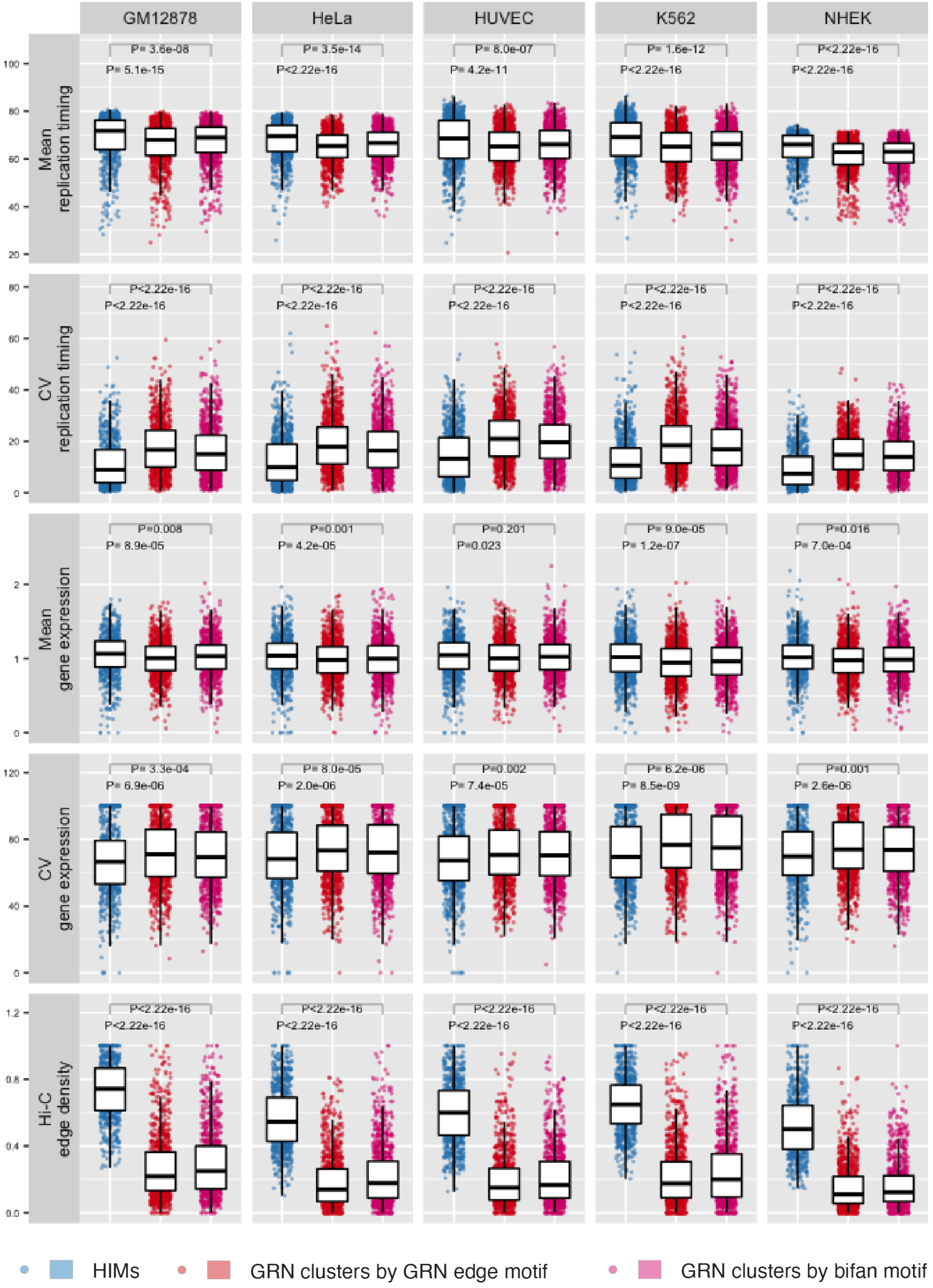

Figure S5: Comparison between HIMs and GRN clusters across cell types in terms of replication timing, gene expression, and Hi-C edge density. Each row represents one feature summarizing genes in each cluster. Each column represents one cell type. Each dot represents a feature value in a cluster. First row is the mean replication timing. Second row is the coefficient of variation (CV) of replication timing. Smaller CV means that genes in a cluster have more similar replication timing. Third and fourth rows are the mean and CV of gene expression levels. Fifth row is the Hi-C edge density where 1 indicates that genes in a cluster are fully connected by chromatin interactions.

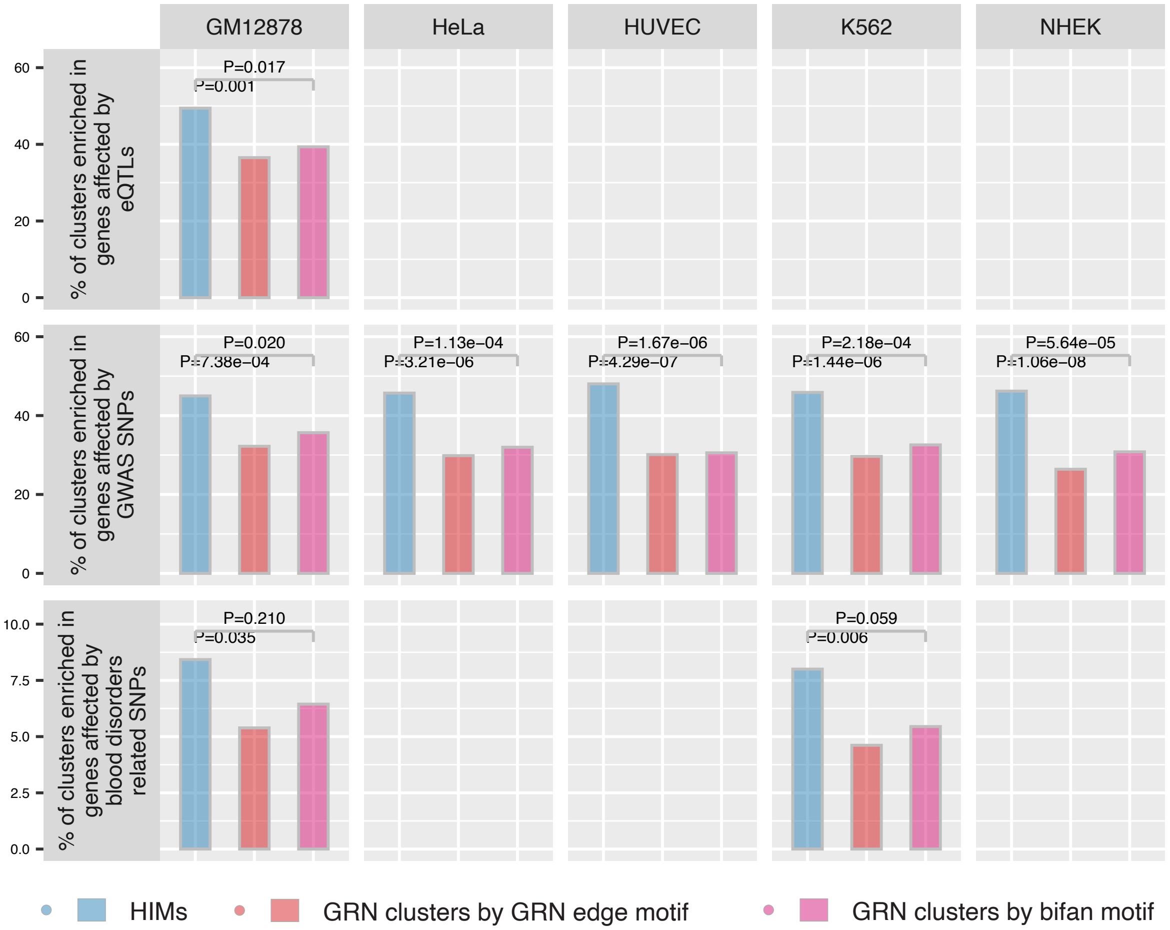

Figure S6: Comparison between HIMs and GRN clusters across cell types in terms of eQTLs (first row), GWAS SNPs (second row), and GWAS SNPs associated with blood-related disorders (third row). Enrichment is defined in Supplemental Methods. Note that eQTLs data are only available in GM12878. GWAS SNPs associated with blood disorders are only applied to blood-related cell lines GM12878 and K562.

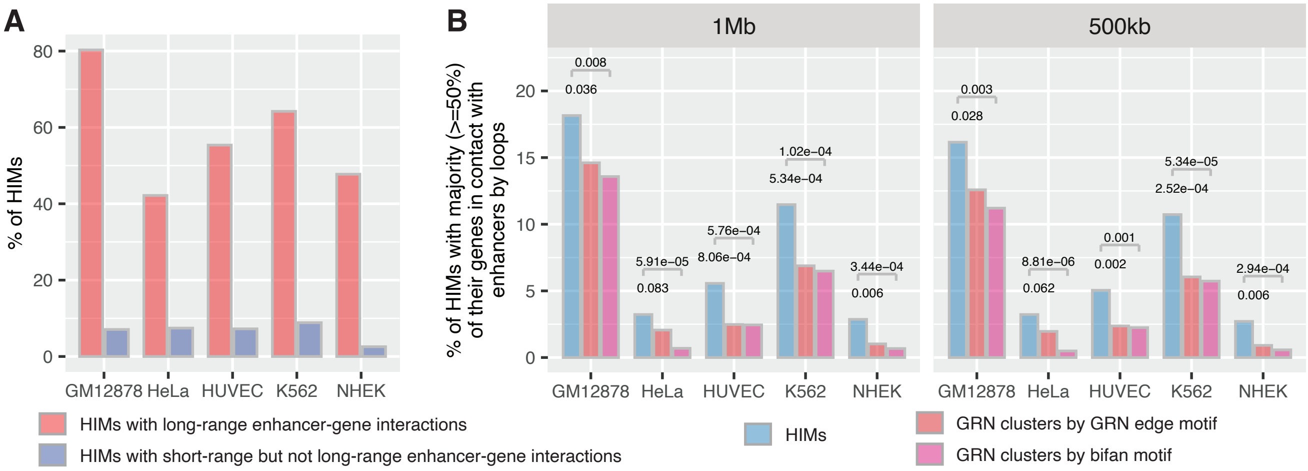

Figure S7: HIMs harbor long-range enhancers. **(A)** Barplots show the proportion of HIMs that have at least one long-range enhancer-gene interactions, and the proportion of HIMs that have short-range (but not long-range) enhancer-gene interactions. Note that both long-range and short-range enhancer-gene interactions in HIMs are defined based on chromatin loops. Some HIMs may not have long-range and short-range enhancer-gene interactions supported by long-range chromatin loops. The sum of the proportions in a cell line therefore does not necessarily equal to 100%. **(B)** Barplots show the proportion of clusters with the majority ($\geq 50\%$) of their genes connected to enhancers by enhancer-gene interactions. HIMs have statistically significant improvement over two sets of conventional GRN clusters. The maximum 1D distance between the potential long-range enhancer and its target genes is set at 1Mb and 500kb, respectively. Note that the proportion in **(A)** and **(B)** varies considerably across cell types because the number of chromatin loops used to define enhancer-gene interactions is different across cell types (minimum number of loops is 3094 in HeLa, maximum is 9448 in GM12878).

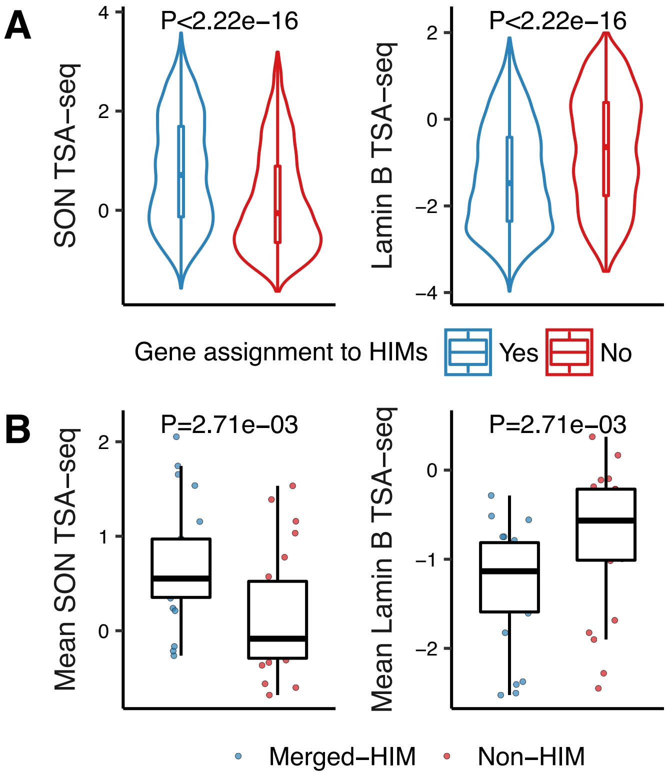

Figure S8: Genes assigned to HIMs are closer to nuclear speckles as compared to the genes that are not assigned to HIMs in K562. **(A)** Violin plots show the distributions of TSA-seq scores of the two sets of genes. **(B)** Boxplots show the distributions of mean TSA-seq scores of the merged-HIM clusters and non-HIM clusters. Here we merge the genes assigned to HIMs on the same chromosome into one cluster and call it a merged-HIM cluster. Similarly, we merge the genes not assigned to HIMs on the same chromosome into one cluster and call it a non-HIM cluster.

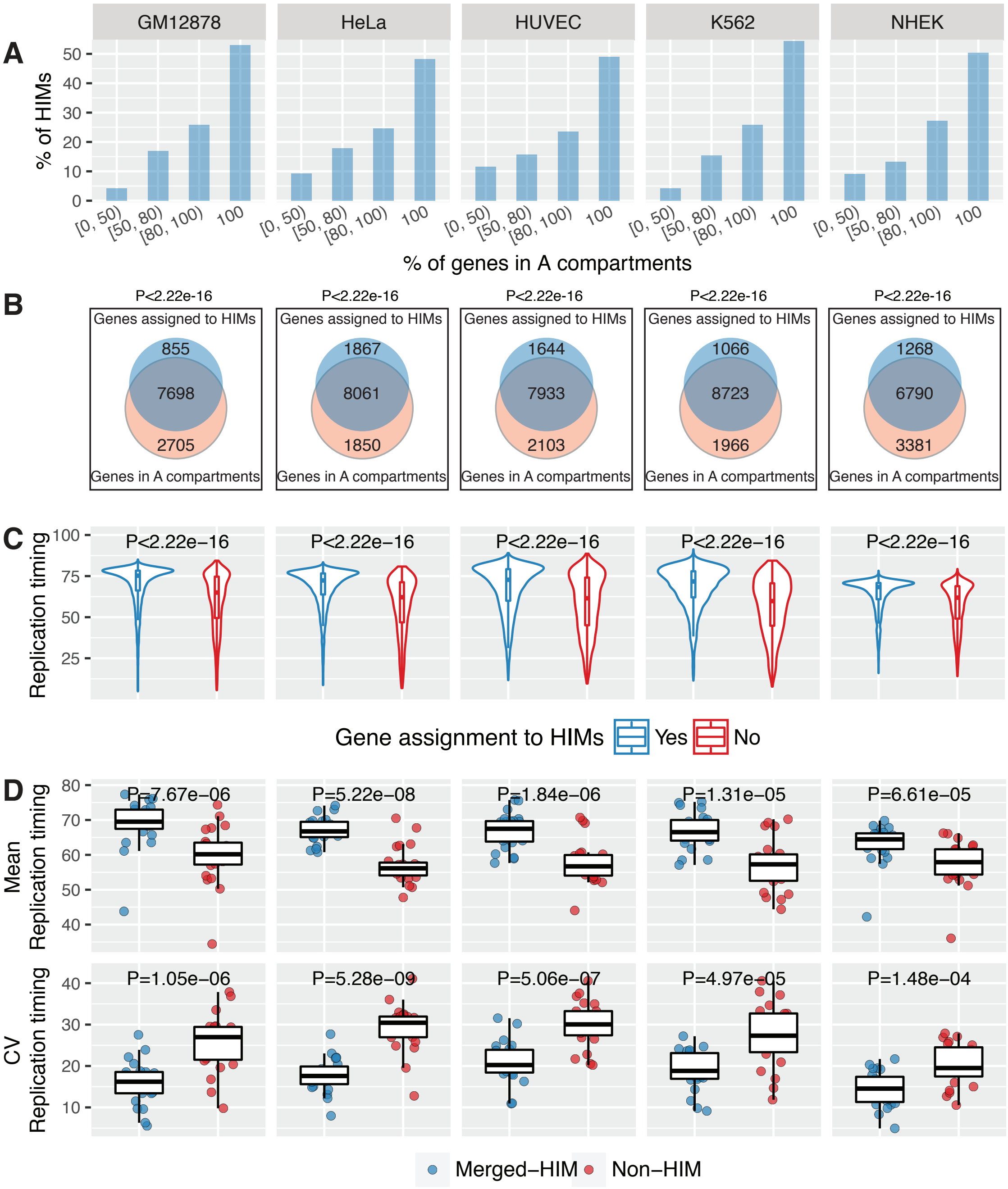

Figure S9: HIMs consistently have spatial localization preferences as compared to non-HIMs in five cell types. Rows correspond to the spatial localization features. Columns correspond to the cell types. **(A)** Barplot shows the distribution of HIMs with a varied proportion of genes that are in A compartment. **(B)** Venn diagram shows that the genes assigned to HIMs are enriched in A compartment. **(C)** Boxplots compare the replication timing of the genes that are assigned to HIMs against the genes that are not assigned to HIMs. **(D)** Boxplots show the mean and coefficient of variation (CV) of replication timing of the genes in merged-HIMs or a non-HIMs. Each dot represents a merged-HIM or non-HIM. A lower CV means that the genes in a cluster have a lower variability in replication timing thus they are more likely to replicate together.

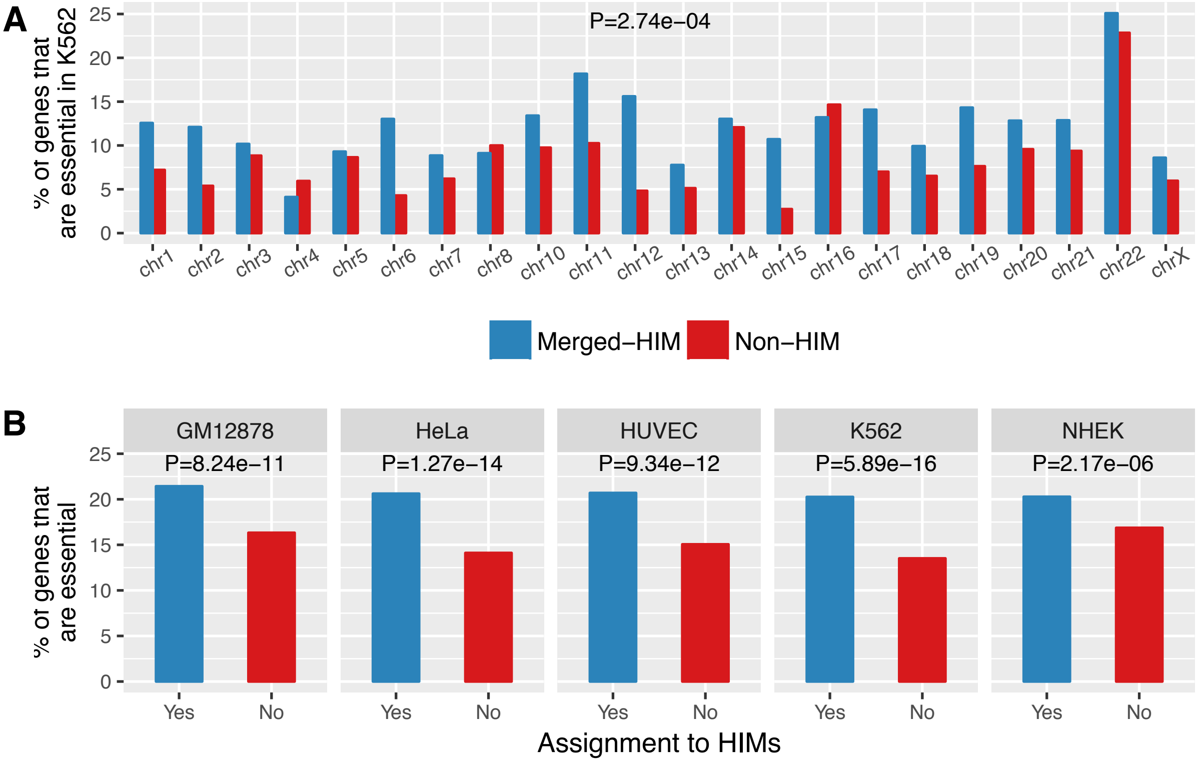

Figure S10: HIMs have more essential genes across cell types. **(A)** Barplots show the proportions of genes that are K562 essential genes among the genes assigned to HIMs (merged-HIM) and the genes not assigned to HIMs (non-HIM) in K562 across different chromosomes. The *P*-value is computed by the paired two-sample Wilcoxon rank-sum test. **(B)** Barplots show the proportions of essential genes in the genes assigned to HIMs and the genes not assigned to HIMs. *P*-value is computed by the Chi-squared test of independence.

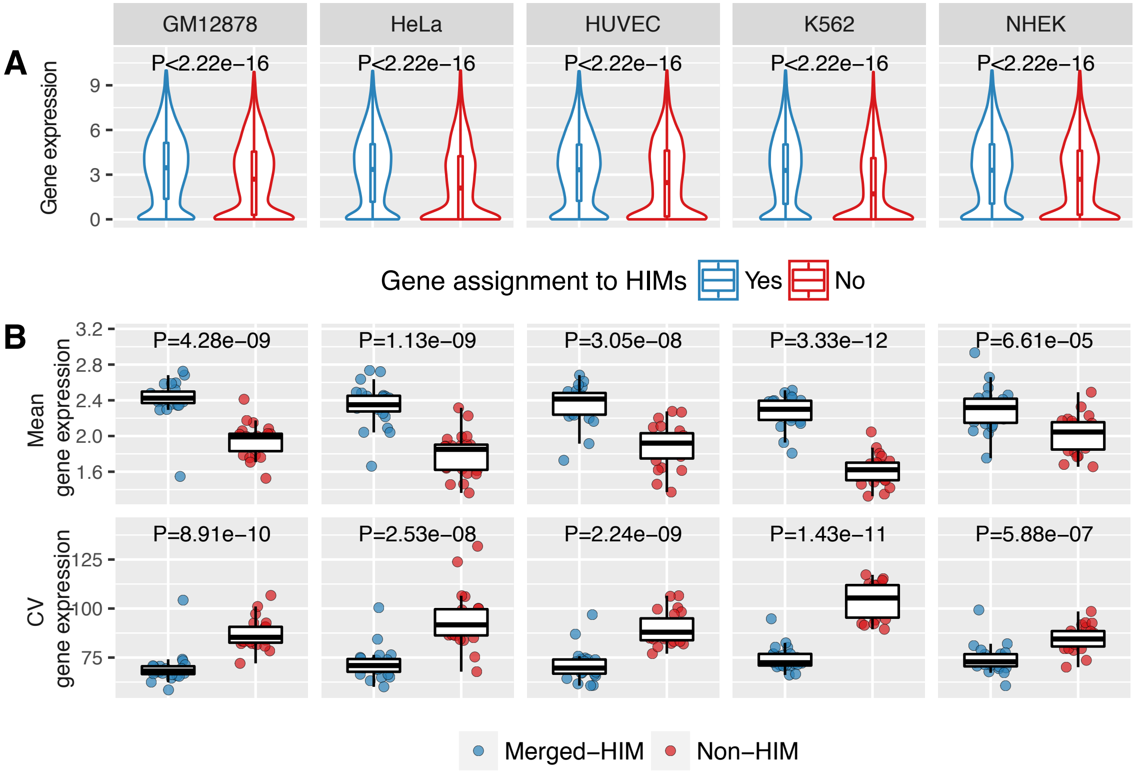

Figure S11: The genes assigned to HIMs have higher expression levels and are expressed at more similar levels. **(A)** Violin plots show that the genes assigned to HIMs have higher expression than the genes not assigned to HIMs. **(B)** Boxplots show that the merged-HIMs have higher mean and lower CV of gene expression than the non-HIMs.

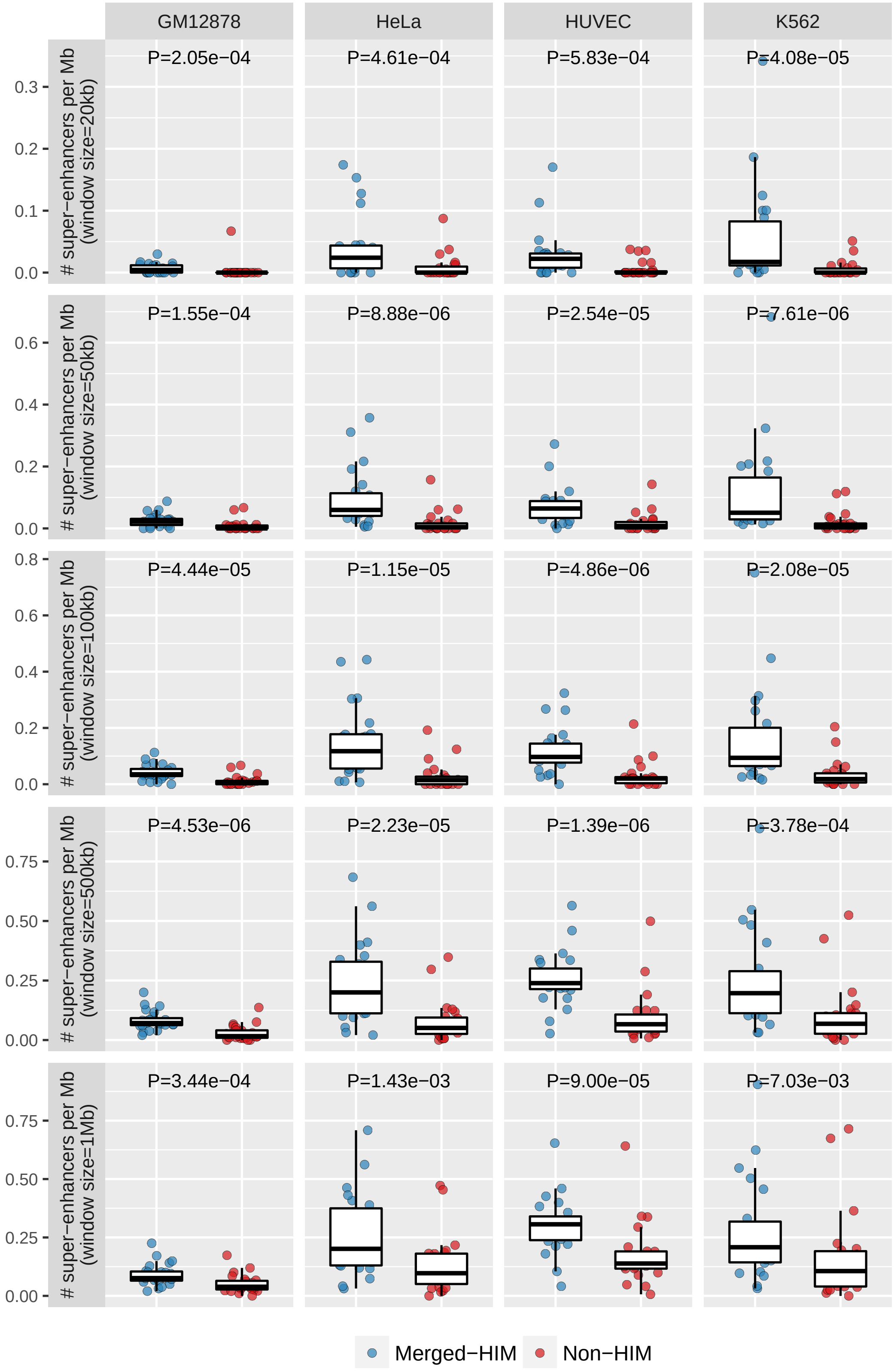

Figure S12: HIMs are enriched with super-enhancers and this observation is robust to the window size. The window size is used to define the genes that are close to a given super-enhancer. The window size ranges from 20kb to 1Mb. The distribution of super-enhancers in NHEK is missing due to lack of super-enhancer data in NHEK.

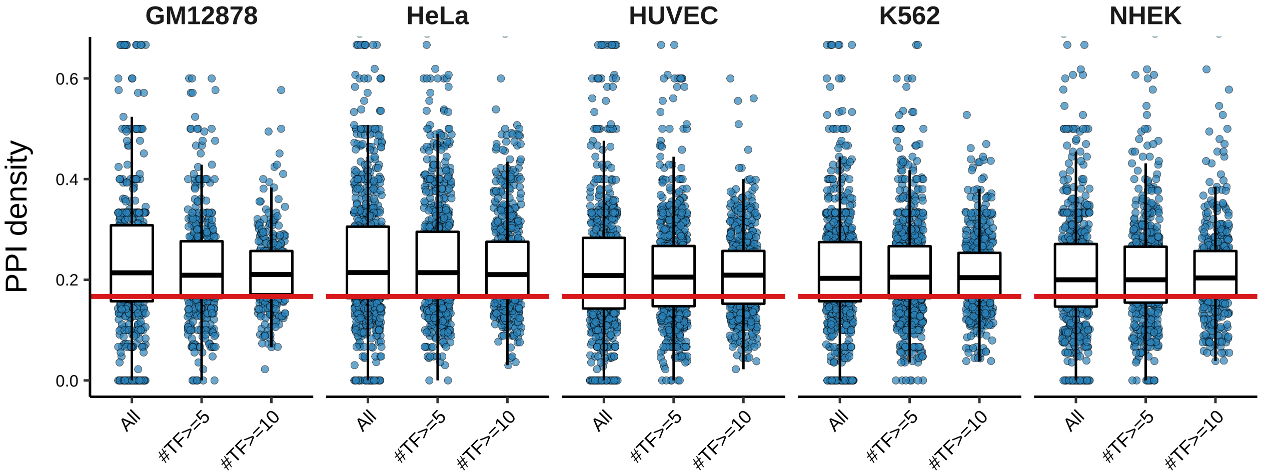

Figure S13: TFs in HIMs are enriched with protein-protein interactions (PPIs). Boxplots show the distribution of the sub-PPI network density across HIMs and the subsets of HIMs with at least n TFs, $n = 5, 10$. Here for each HIM, we compute the density of the sub-PPI network induced by the TFs in the HIM from the PPI network based on 591 TFs used in this study. The medians are all higher than the expected density (0.158, red line) of the sub-PPI networks induced by randomly sampled TFs (p <2.22e-16).

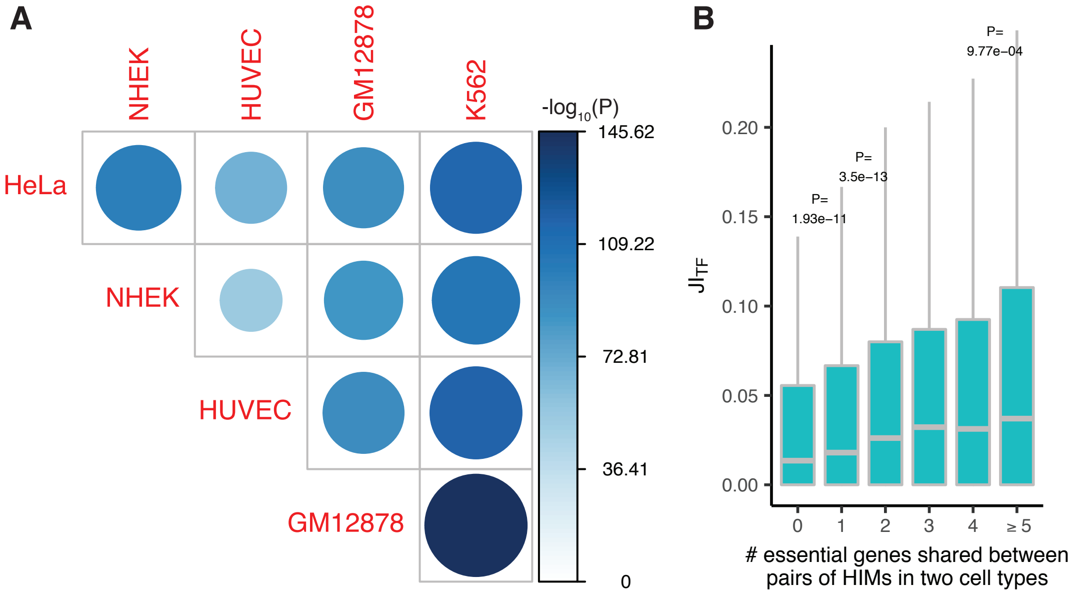

Figure S14: HIM comparisons regarding genes and TFs across cell types. **(A)** Heatmap shows the level of significance of the overlaps between the genes assigned to HIMs in two different cell types. GM12878 and K562 have the highest overlap. Statistical significance is evaluated by the hypergeometric test. **(B)** Boxplots show the distribution of Jaccard index on the TFs of paired HIMs with different numbers of shared essential genes.

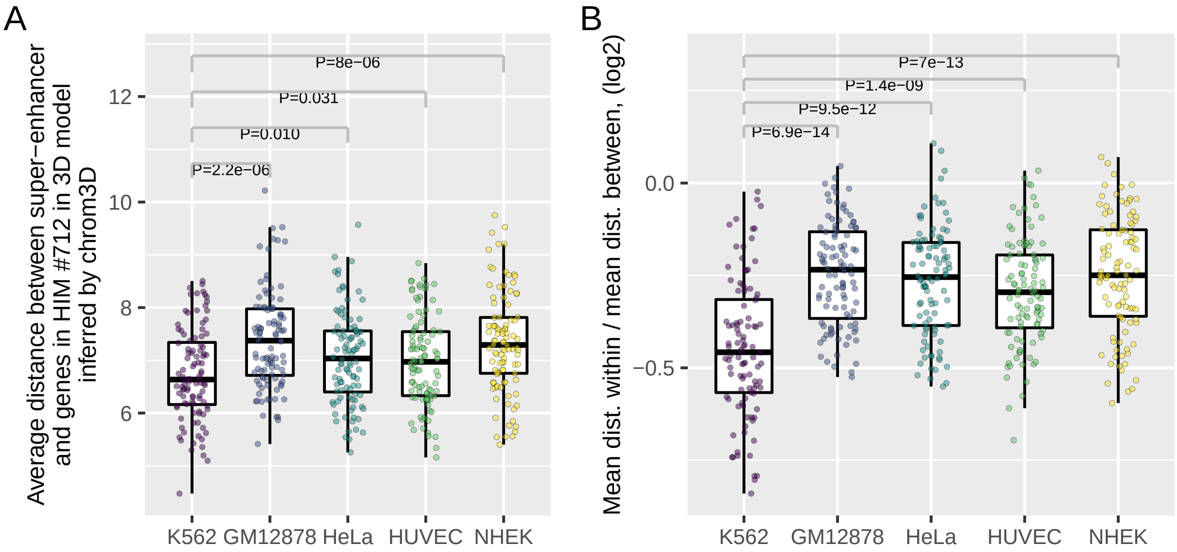

Figure S15: Comparison on 100 possible 3D structural representations centering on HIM #712 across cell lines. Chrom3D (Paulsen et al. 2018) was run 100 times to generate 100 possible 3D structures in each cell type. One representation in each cell line is shown in Supplemental Movie S1. **(A)** Boxplot shows the average distance between the upstream super-enhancer and the genes in HIM #712. The super-enhancer is spatially closest to the genes in HIM #712 in K562. **(B)** Boxplot shows the log2 transformed ratio of two average distances. The first is the average distance between the 10kb chromosomal bins within the genomic region that covers the super-enhancer and genes in HIM #712. The second is the average distance between the bins in the genomic region and the bins in the flanking regions (+/-500kb) of the genomic region. The boxplot shows that in K562 the genomic region covering the super-enhancer and the genes in HIM #712 preferentially contact itself rather than its flanking regions (+/-500kb) in 3D space. Here each dot represents a value of one 3D structure. Chrom3D was run with 10kb resolution Hi-C data (the number of iterations is 100,000 with other default parameters).

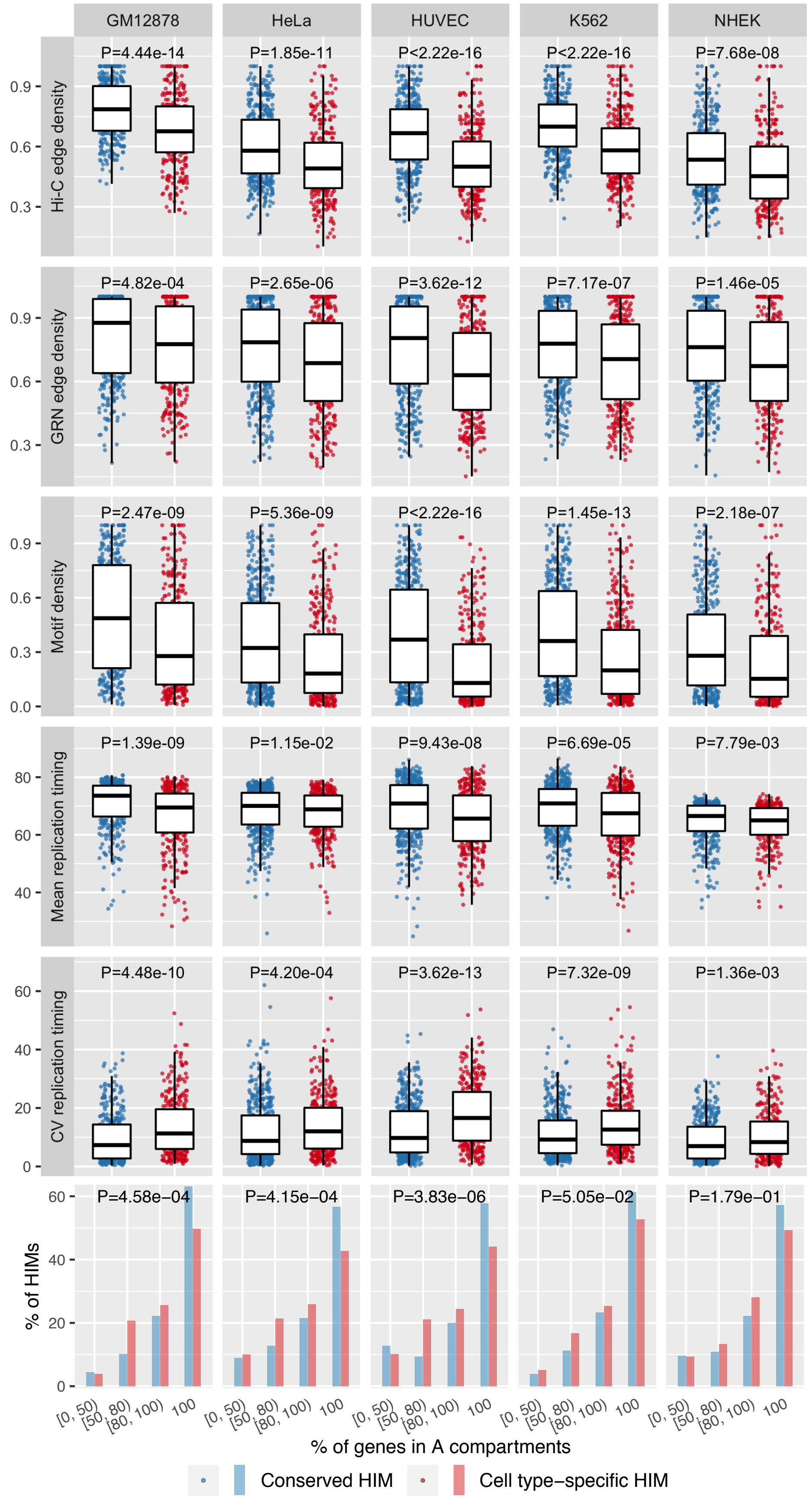

Figure S16: Conserved and cell type-specific HIMs have distinct cluster features (density), replication timing, and connections to A compartment.

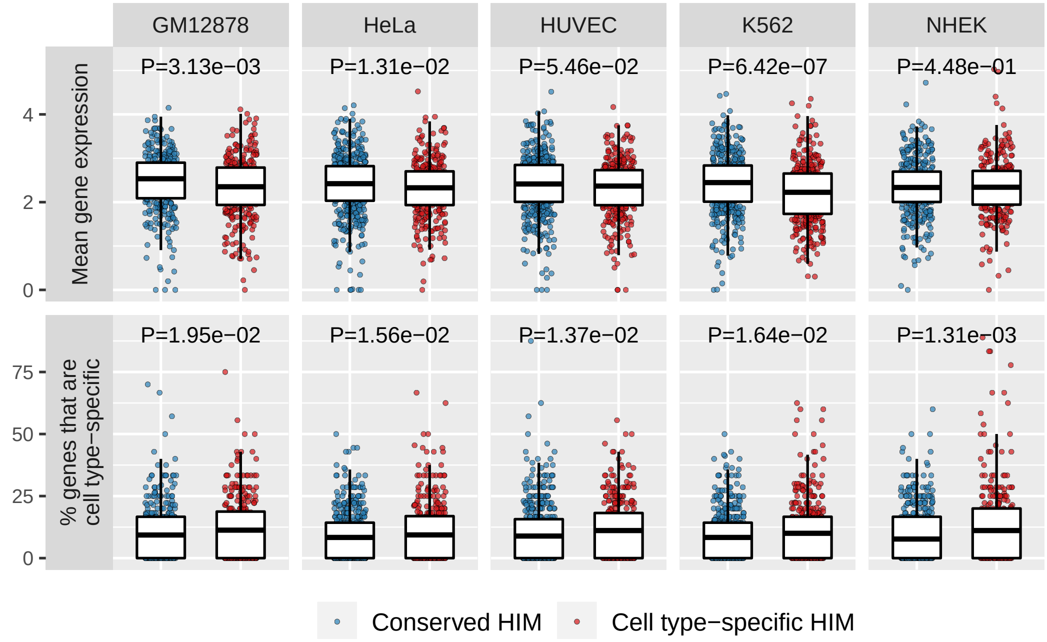

Figure S17: Functional differences and similarities between conserved and cell type-specific HIMs. Conserved HIMs have higher average gene expression in 3 cell types (first row). On the other hand, cell type-specific HIMs tend to have a higher proportion of cell type-specific genes in all 5 cell types (second row).