MOCHI enables discovery of heterogeneous interactome modules in cell nucleus

Dechao Tian[1,#], Ruochi Zhang[1,#], Yang Zhang[1], and Jian Ma[1,*]

[1]Computational Biology Department, School of Computer Science,
Carnegie Mellon University, Pittsburgh, PA 15213, USA
[#]These two authors contributed equally
[*]Corresponding author: jianma@cs.cmu.edu

Abstract

The composition of the cell nucleus is highly heterogeneous with different constituents forming complex interactomes. However, the global patterns of these intertwined different interactomes in the nucleus remain poorly understood. In this paper, we focus on two different interactomes in the nucleus, chromatin interaction network and gene regulatory network as a proof-of-principle, to identify heterogeneous interactome modules (HIMs). Each HIM represents a cluster of gene loci that are in spatial contact more frequently than expected and that are regulated by the same group of transcription factor proteins. We developed a new algorithm, called MOCHI (MOtif Clustering in Heterogeneous Interactomes), to facilitate the discovery of HIMs based on network motif clustering in heterogeneous interactomes. By applying MOCHI to five different cell types, we found that HIMs have strong spatial preference within the nucleus and exhibit distinct functional roles. In addition, more conserved HIMs have different spatial and functional properties as compared to cell type-specific HIMs. This work demonstrates the utility of MOCHI to identify HIMs as a type of nuclear genome organization unit, which may provide new perspective of nucleome structure and function.

Introduction

The cell nucleus is an organelle that contains heterogeneous components such as chromatins, proteins, RNAs, and subnuclear compartments. These different constituents of the nucleus form complex interactions that are spatially and temporally dynamic [1–4]. In human and other higher eukaryotes, chromosomes are folded and organized in three-dimensional space by compartmentalizing the cell nucleus [5, 6], and different chromosomal loci also interact with each other [3, 4, 7]. Recent development of new high-throughput whole-genome mapping approaches for chromatin interactome such as Hi-C [7] has allowed us to identify genome-wide chromatin organization and interactions comprehensively, revealing important chromatin interactome features such as loops [8, 9], topologically associating domains (TADs) [10], and A/B compartments [7, 8].

Gene regulation can be influenced by changes in spatial position and interactions of genomic loci [11, 12]. In particular, correlations between higher order chromatin structure, chromosome positioning relative to subnuclear compartments, and transcriptional activity of the genes have been demonstrated [13–15]. It is also known that gene regulatory networks (GRNs) are dynamic among various cellular conditions [16, 17]. GRNs model the phenomena of selective binding of transcription factor (TF) proteins to *cis*-regulatory elements in the genome to regulate target genes [18–20]. At the systems level, we can view the cell nucleus as being organized by a collection of complex interacting networks among different constituents, such as chromatin interactome, GRN, RNA interactome, RNA-chromatin interactome, protein interactome, that are also intertwined with each other. However, how these heterogeneous interactomes are organized to form functionally relevant, global patterns remain poorly understood.

In this work, as a proof-of-principle, we specifically consider two different types of interactomes in the nucleus: (1) chromatin interactome – a network of chromosomal interactions between different genomic loci – and (2) a GRN where TF proteins bind to the genomic loci to regulate target genes' transcription. Many studies in the past have separately analyzed the structure and dynamics of chromatin interactomes and GRNs in

different contexts [8, 17, 21–25]. However, the global patterns between chromatin interactome and GRN have not been fully revealed [26–30]. In particular, algorithms that can simultaneously analyze these heterogeneous networks in the nucleus to discover important network structures have not been developed.

Here we aim to discover significant heterogeneous network-level patterns by integrating chromatin interactome and GRN. Specifically, we want to identify network structures where nodes of TFs (from GRN) and gene loci (from both chromatin interactome and GRN) cooperatively form distinct types of network clusters We developed a new algorithm MOCHI (MOtif Clustering in Heterogeneous Interactomes) that can effectively discover such network clusters, which we call heterogeneous interactome modules (HIMs), based on network motifs. Each identified HIM represents a collection of gene loci and TFs that (1) the gene loci have higher than expected chromatin interactions between themselves, and (2) the gene loci are regulated by the same group of TFs. We applied MOCHI to five different human cell types to discover patterns of HIMs. We found that HIMs have strong spatial preference within the nucleus and may harbor important functional roles with more enriched essential genes and super-enhancers. Additionally, we found that more conserved HIMs have different structural and functional properties as compared to cell type-specific HIMs. These findings demonstrate the utility of MOCHI to identify HIMs based on complex heterogeneous interactomes in cell nucleus. We believe that HIMs as a nuclear genome organizational unit could provide potential new insights into nucleome structure and function. The source code of our MOCHI method can be accessed at: https://github.com/ma-compbio/MOCHI.

Results

Overview of the MOCHI algorithm

Fig. 1 shows the overview of our method; detailed algorithms are described in the Methods section. Our goal is to reveal higher-order network clusters in a heterogeneous network such that certain higher-order network structures (e.g., the network motif M in Fig. 1A) frequently contained within the same cluster. The input heterogeneous network in this work considers two types of interactomes in the nucleus: a GRN (directed) between TF proteins and target genes; and chromatin interaction network (undirected) between gene loci on the genome. For chromatin interactome, for each pair of gene loci within 10Mb, we use the "observed over expected" (O/E) quantity in the Hi-C data (we use O/E$>$1 as the cutoff in this work but we found that our main results are largely consistent with different cutoffs; see Supplementary Results) to define the edges in the chromatin interaction network [7, 8]. For GRN, we use the transcriptional regulatory networks from [31], which were constructed by combining enrichment of TF binding sites in enhancer and promoter regions and co-expressions between TFs and genes. If a TF protein regulates a gene, we add a directed edge from the TF to the gene. We then merge the chromatin interaction network and the GRN from the same cell type to form a heterogeneous network G with nodes that are either TF proteins or gene loci together with the directed/undirected edges defined above (Fig. 1B).

In this paper, we specifically consider the type of network motif M with four nodes, i.e., two gene loci and two TFs in the heterogeneous network with two genes whose genomic loci are spatially more proximal to each other (than expected) within the nucleus and that are also co-regulated by the two TFs (Fig. 1A). Our goal is to reveal higher-order network clusters based on this network motif. In other words, we want to partition the nodes in the network such that this 4-node network motif occurs mostly within the same cluster. Based on the motif, our MOCHI algorithm constructs an undirected, weighted network G_M (Fig 1D) based on subgraph adjacency matrix W_M (Fig. 1C). We then apply recursive bipartitioning in G_M to find multiple clusters (Fig 1E). We call such clusters as heterogeneous interactome modules (HIMs), which, in this work, represent network structures containing the same group of TFs that regulate many target genes whose spatial contact frequencies are greater than expected. Since TFs can regulate multiple sets of genes that may belong to different clusters, different HIMs may overlap by sharing TFs.

MOCHI identifies HIMs in multiple cell types

We applied MOCHI to five different human cell types: GM12878, HeLa, HUVEC, K562, and NHEK. The input heterogeneous network of each cell type has 591 TFs, $\sim$12,000 genes, and $\sim$1 million regulatory interactions

(Table S1). A few example HIMs identified in GM12878 are shown in Fig. 1F-H, including overlapping HIMs in Fig. 1H. We found that the identified HIMs in five cell types have similar basic characteristics. The number of identified HIMs ranges from 650 to 806 in different cell types, with at least 71.9% of the HIMs share TFs with other HIMs in each cell type. Notably, HIMs cover majority (62.1-77.2%) of the genes in the heterogeneous networks (Table S1). For example, in GM12878, there are 591 TFs co-regulating 7,617 (69.1%) genes in 650 HIMs. The HIMs have, on average, 9 TFs regulating 9 genes in GM12878 and 14-17 TFs regulating 9 genes in the other cell types (Table S2).

In addition, MOCHI is robust to the parameters used to construct the heterogeneous networks (Supplementary Section B.2). Importantly, we also justified the choice of the 4-node motif M by showing its advantages over a triangle motif and a bifan motif (Supplementary Section B.3). The triangle and bifan motifs do not explicitly encode the co-regulations between TFs and spatial proximities between genes, respectively. These results demonstrate that MOCHI is able to reliably identify HIMs across the 5 different cell types.

HIMs have strong preference in spatial locations in the nucleus

Next, we specifically looked at the nuclear localization of HIMs. Recently published SON TSA-seq and lamin TSA-seq quantify cytological distance of chromosome regions to nuclear speckles and lamina, respectively [15]. In K562, which is currently the only cell type that has TSA-seq data [15], 60.7% of the HIMs have mean SON TSA-seq score higher than 0.284 (80-th percentile of the SON TSA-seq score), suggesting that the genes in these HIMs, on average, are within $0.518\,\mu\mathrm{m}$ (estimated by [15]) of nuclear speckles (Fig. 2A). Compared to the genes not assigned to HIMs, the genes in HIMs have significantly higher SON TSA-seq and lower Lamin B TSA-seq (P<2.22e-16; Fig. S1). We specifically looked at HIMs that are away from the nuclear interior. In Fig. 2B, we show one HIM (#541) that is close to nuclear lamina (mean Lamin B TSA-seq 0.593, mean SON TSA-seq -0.642). This HIM has 6 genes spanning 6.78 Mb on chr3 and are co-regulated by 9 TFs. The Hi-C edge density among these genes is 0.667. The SON TSA-seq scores of the 6 genes are low but tend to be the local maxima (i.e., small peaks within valleys), while the Lamin B TSA-seq scores are high but tend to be the local minima (i.e., small valleys within peaks). The gene RPL15 in this HIM is a K562 essential gene [32]. The TFs proteins CDX1, HOXA9, and HOXA10 are involved in leukemia and hematopoietic lineage commitment, This suggests that even though HIM 541 is one of the HIMs that are away from nuclear speckle, it may play relevant functional roles in K562.

Recently, Quinodoz et al. [33] reported that chromosomal interactions across large distance are clustered into two distinct nuclear bodies as hubs, including nuclear speckles and nucleolus. By comparing with the genomic regions organized around nucleolus based on SPRITE data in GM12878 [33], we found that a large proportion (85.4%) of the GM12878 HIMs do not have genes close to nucleolus. There are 30 (4.62%) GM12878 HIMs with all their genes near nucleolus. 16 out of these 30 HIMs have at least one TF protein located close to nucleoli according to protein subcellular locations from the human protein atlas [34]. For example, HIM #267 has 4 TF regulators: ETS1, ETV6, PPARG, and PTEN, where ETV6 is known to be mainly localized to the nucleoli in cell nucleus. Earlier work estimated that only 4% of the human genome are within nucleolus-associate domains [35], it is therefore expected that only a small number of HIMs would be close to nucleoli.

Earlier work from Hi-C analysis showed that human chromosomes are segregated into A and B compartments that are largely active and inactive in transcription, respectively [3, 8, 36]. Chromosome regions in A compartments tend to be toward the interior of the nucleus, while B compartments tend to be near the nuclear periphery/lamina [28]. In addition, it is known that replication timing correlates well with A/B compartments, where early replicating regions tend to be in A compartments and late replicating regions are mostly B compartments [36, 37]. We found that the genes in the HIMs are preferentially in A compartments and early replicated across the cell types. For example, 57.4% of the HIMs have genes that are all in A compartments in K562. For the other HIMs, there are genes that are in A or B compartments. Only a small proportion (4.49%) of HIMs have over 50% of genes in B compartments (Fig. 2C). We found that the genes in HIMs as a whole are significantly more enriched in A compartments with 89.1% of them are in A compartments (P<2.22e-16, hypergeometric test; Fig. 2D) Additionally, we found that the genes assigned to the HIMs have much earlier (P<2.22e-16) replication timing than the other genes (Fig. 2E). The genes on the same chromosomes that are

also in HIMs tend to have more similar replication timing as compared to the genes (on the same chromosomes) that are not in HIMs (Fig. S2). These patterns are also observed in other cell types (Fig. S2). Our results suggest that HIMs tend to be near the interior of the nucleus where there is more transcription.

HIMs are enriched with essential genes, super-enhancers, and PPIs

Next, we explored the functional properties of HIMs. We merged the genes assigned to HIMs into one set and the genes in the heterogeneous network but are not assigned to HIMs into another set. For a fair comparison, we also stratify the gene sets by chromosome number. We call these clusters merged-HIM clusters and non-HIM clusters accordingly. First, we investigated the gene essentiality [32] (see Supplementary Methods A.5). We found that genes assigned to HIMs are enriched with essential genes across all 5 cell types. For example, 12.7% of the genes assigned to HIMs in K562 are K562 essential genes, which is significantly (P=1.13e-12) higher than the proportion (7.79%) of the genes not assigned to HIMs (Fig. 3A). The merged-HIMs have significantly (P=2.74e-4) higher proportion of K562 essential genes than the non-HIMs across the chromosomes (Fig. S3A). Across the cell types, the genes assigned to HIMs consistently have significantly (P$\leq$2.17e-6) higher proportions of the essential genes than the genes not assigned to HIMs (Fig. S3B). Regarding gene expression level, we found that the genes assigned to HIMs are more highly expressed and expressed at similar levels (Fig. 3B, Fig. S4).

Super-enhancers are known to be associated with many cell type-specific functions in both normal and cancer cell types [38]. To study the connections between HIMs and super-enhancers, we computed 1D distance normalized number of super-enhancers [38] that (1) have Hi-C contacts with; and (2) are close to (window size=50kb) at least one gene in each cluster. We found that HIMs are enriched with spatial contacts with super-enhancers. Specifically, the merged-HIMs have at least 6-fold higher normalized number of super-enhancer than the non-HIMs across the cell types (Fig. 3C, Fig. S5). The significant pattern is consistent with a varied window size from 20kb to 1Mb (Fig. S5).

Protein-protein physical interactions (PPIs) can further stabilize TF-DNA binding of the interacting TFs [20]. We asked whether TFs in the same HIM tend to have more PPIs with each other, given that they co-regulate a set of spatially proximal genes in HIMs. We computed the density of the sub-PPI network induced by the TFs in a HIM from a whole PPI network, where the whole PPI network is between TF proteins and constructed from public databases (see Supplementary Methods A.5). We found that TFs within HIMs are enriched with PPIs among themselves as compared to random cases where the same number of TFs are randomly sampled from the whole PPI network. For example, in GM12878, master TFs NR3C1 and TFEB [38] co-regulate 8 genes with the other 7 TFs in a HIM (Fig. 3D). The sub-PPI network induced by the 9 TFs from the whole PPI network has 14 PPIs. The density of the sub-PPI network is 0.389 and is 2.46 times higher than the average density (0.158) of the random cases, which is also the density of the whole PPI network. Moreover, the median density of the sub-PPI networks induced by TFs in the identified HIMs in GM12878 is 0.214, which is significantly (P<2.22e-16) higher than the average of the random cases (Fig. 3E). In the other cell types, the average densities are no less than 0.2 and are significantly (P< 2.22e-16) higher than the average of the random cases (Fig. S6). Besides, the significance is not affected by a varied number of TFs across HIMs (Fig. S6). For example, the subset HIMs with at least $n, n = 5, 10$, TFs in GM12878 still have significantly (P<2.22e-16) higher density than the random cases (Fig. 3E).

HIMs undergo changes across cell types

To study how HIM changes across different cell types, we first focused on assignment of genes to HIMs in the cell lines. There are 3,025 genes constitutively assigned to HIMs, accounting for 30.91% to 40.06% genes that are in the HIMs in each cell type (Fig. 4A). In contrast, only a small fraction ($\leq$5.93%) of genes are uniquely assigned to the HIMs in each cell type. For example, only 344 (4.28%) genes out of the 8,034 genes in the GM12878 HIMs are only assigned to HIMs in GM12878 (Fig. 4A). The genes constitutively and uniquely assigned to HIMs are enriched with distinct functional terms (DAVID [39, 40], Fig. 4B, Table S5). The genes constitutively assigned to HIMs are strongly enriched with biological processes related to chromosome organization including nucleosome organization, DNA conformation, and DNA packaging. They are also enriched

with transcription and splicing related biological processes, suggesting that they are responsible for essential cellular machineries maintaining cellular activity. On the other hand, the genes uniquely assigned to HIMs in a particular cell type are enriched with cell type-specific functions. For example, such genes in NHEK are enriched for keratinocyte related GO terms including keratinization, keratinocyte differentiation, and epidermis development. Such genes in GM12878 are enriched in regulation of lymphocyte activation. Additionally, we found a few HIMs where majority of their genes are uniquely assigned to HIMs in one cell type. An example is NHEK HIM #107 (Fig. 4C). Among its 6 genes, DSC1, DSC3, DSG1 are not assigned to HIMs in the other cell types. The 6 genes are involved in keratinization pathway [41].

We next looked at HIM dynamic changes in spatial proximity to subnuclear compartments. We found that a considerable proportion of the 30 GM12878 HIMs close to nucleoli in GM12878 [33] are spatially proximal to speckles in K562 cell type. For example, 15 out of the 30 GM12878 HIMs have mean SON TSA-seq score $\geq$ 0.284 in K562 (Fig. 4D). One standout example is HIM #267 which has the highest (2.41) mean SON TSA-seq score. Interestingly, the genes and TFs in HIM #267 change in K562. The 10 genes in HIM #267 plus other 8 genes form a new HIM (#628) in K562. The GM12878 HIM #267 has 4 TFs: ETS1, ETV6, PPARG, and PTEN. On the other hand, K562 HIM #628 has 4 totally different TFs: KLF4, NFKB1, STAT3, and WT1, where KLF4, STAT3, and WT1 are involved in multiple functions in leukemia [42–44].

We further zoomed into the membership dynamics of HIMs across the cell types. We computed Jaccard indices, denoted by JI_{TF} and JI_{gene}, on the TF members and gene members between HIMs from two different cell types. We found that the gene members undergo moderate change from one cell type to another, and the TF members change at a much higher rate. The median of JI_{gene} and JI_{TF} are 0.096 and 0.017, respectively. JI_{TF} is $\sim$10 times smaller than JI_{gene} (Fig. 4E). There are at least two possible explanations for the observations. First, the Hi-C interaction networks and GRNs are highly cell type-specific (Table S3). Second, given a HIM identified in a cell type, the motif M density of the HIM has significantly (P<2.22e-16) higher fold change than the Hi-C edge density of the HIM in another cell type (Fig. 4F). In other words, the co-regulation relationships of the TFs on the genes in HIMs change more often in another cell type than the spatial proximity relationships between the gene loci.

Conserved and cell type-specific HIMs have distinct features

Motivated by the gene membership dynamics of HIMs across the cell types, we further classified HIMs into more conserved and cell type-specific HIMs. For the HIMs in a given cell type, we called a HIM a more conserved HIM if it shares high proportions of genes ($\text{JI}_{gene} \geq 1/3$) with at least one HIM in the other cell types. The others are called cell type-specific HIMs. As a result, 34-38.8% of the identified HIMs in each cell type are cell type-specific HIMs. We first showcased a cell type-specific HIM, HIM #712, in K562 and its changes in other cell types (Fig. 5). The HIM covers 9 genes in a region on chr11. These genes spatially contact to each other at higher frequencies than expected (Fig. 5A) and are co-regulated by TF protein BCL6B and CPEB1 in K562 (Fig. 5B). In the other cell types, at most 4 genes are assigned to HIMs (Fig. 5C). We found that the HIM has K562 specific chromosomal structures and functional annotations. The genomic region covering the genes in the HIM is in A compartment in K562 but switches to B compartments in other cell types (Fig. 5D). One nearby upstream region is annotated as a super-enhancer only in K562 [38] (Fig. 5E). Many sites are annotated as active states such as enhancers, promoters, or transcribed states in K562 but not in other cell types by ChromHMM [45, 46] (Fig. 5F). The genes MRPL16, OSBP, and PATL1 are essential genes in K562 [32]. This example demonstrates that the K562 specific HIM has specific chromatin organization and biological functions.

We found that the more conserved and cell type-specific HIMs have distinct spatial location features across the 5 cell types. Compared to the cell type-specific HIMs, the more conserved HIMs have better cluster quality. Because the more conserved HIMs have significantly (P$\leq$8.15e-4) higher Hi-C edge density, higher GRN edge density, and higher motif M density (Fig. S7). Also, the more conserved HIMs tend to be closer to the nuclear interior as a higher proportion of them have all their genes in A compartments, their genes are earlier replicated and replicated at more similar phases (Fig. S7). Moreover, we found that the more conserved HIMs and cell type-specific HIMs tend to have significant differences in gene expression level and cell type-

specific genes. The more conserved HIMs have significantly (P<0.05) higher averaged gene expression level than the cell type-specific HIMs in 3 cell types except for NHEK and HUVEC (Fig. S8). On the other hand, the cell type-specific HIMs have significantly (P≤0.014) higher proportion of the cell type-specific genes (see Supplementary Methods A.5) than the more conserved HIMs across the 5 cell types (Fig. S8), indicating their cell type-specific functions. Our results demonstrate that the more conserved and cell type-specific HIMs in general have distinct spatial location and functional properties.

Methods

Brief introduction on homogeneous network clustering by motif conductance

We first review a recently published higher-order network clustering method that can identify a cluster of nodes S based on motif conductance (defined below). We then introduce our algorithm MOCHI in the next subsection. Let G be an undirected graph with N nodes and A be the adjacency matrix of G. $[A]_{ij} \in \{0, 1\}$ represents the connection between nodes i and j. The *conductance* of a cut$(S, \bar{S})$, where S is a subset of the nodes is defined as:

$$\varphi_G(S) = \frac{\text{cut}_G(S, \bar{S})}{\min[\text{Vol}_G(S), \text{Vol}_G(\bar{S})]}, \tag{1}$$

where $\text{cut}_G(S, \bar{S}) = \sum_{i \in S, j \in \bar{S}} [A]_{ij}$ is the number of edges connecting nodes in S and $\bar{S}$. $\text{Vol}_G(S) = \sum_{i \in S} \sum_{j=1}^{N} [A]_{ij}$ is the sum of the node degree in S. Moreover, the conductance of the graph G, φ^G, is defined as $\min_S \varphi_G(S)$. The S that minimizes the function is the optimal solution. Finding the optimal S is NP-hard but spectral methods such as Fiedler partitions can obtain clusters effectively [47]. Recently, the conductance metric is generalized to motif conductance [48, 49], where a motif refers to an induced subgraph. The motif conductance computes $\text{cut}_G(S, \bar{S})$ and $\text{Vol}_G(S)$ based on a chosen n-node motif. When $n = 2$, the motif is an interaction that reduces the motif conductance to conductance in Eq. (1). When $n \geq 3$, the motif conductance may reveal new higher-order organization patterns of the network [48]. A more recent network clustering method that incorporates network higher-order structures is developed in the setting of hypergraph clustering [50], which includes the motif conductance as a special case. However, one key limitation of the aforementioned methods is that they cannot identify overlapping clusters, which is a crucial feature of the heterogeneous networks used in this paper.

Higher-order network clustering to identify HIMs in a heterogeneous network

We developed a higher-order network clustering method based on network motif to identify overlapping HIMs in a heterogeneous network inspired by the approach in [48]. We call our method MOCHI (MOtif Clustering in Heterogeneous Interactomes). We illustrate the workflow of MOCHI in Fig. 1. First, we select a specific heterogeneous 4-node network motif M (Fig. 1A). In M, two nodes are TFs and the other two nodes are genes. Both TFs regulate the two genes and the two genes are spatially more proximal to each other than expected. The motivation for choosing the subgraph M is that it is the building block of HIMs given that our goal is to discover a group of genes that contact with each other more frequently than expected and are regulated by the same set of TFs. As compared to simpler motifs (e.g., 3-node motif where one node is TF), our 4-node motif defined here has the advantage of simultaneously considering a pair of genomic loci that interact with each other higher than expected and that are co-regulated by the same pair of TFs. Conceptually, our method searches for a HIM with two goals. The TFs and genes in the same HIM should be involved in many occurrences of M. Additionally, HIM should avoid cutting occurrences of M, where a cut of occurrences of M means that only a subset of TFs and genes in the occurrences of M are in the HIM node set. More formally, our method aims to find HIMs with the nodes set S that minimizes the motif conductance

$$\varphi_M(S) = \frac{\text{cut}_M(S, \bar{S})}{\min[\text{Vol}_M(S), \text{Vol}_M(\bar{S})]}. \tag{2}$$

We first introduce some notations before we explain $\varphi_M(S)$. Let G be the given heterogeneous network (e.g., Fig. 1B). Let $\mathbb{M}$ be the set of occurrences of the motif M in G. For simplicity and without confusion, we

also denote an occurrence of the motif M as M. Let V_M be the vertex set of the 2 TFs and 2 genes in $M \in \mathbb{M}$. In Eq. (2), $\text{cut}_M(S, \bar{S})$ is the number of occurrences of the subgraph M that are cut by S. Formally,

$$\text{cut}_M(S, \bar{S}) = \sum_{M \in \mathbb{M}} \mathbb{1}(|V_M \cap S| \in \{1, 3\}) + \alpha \sum_{M \in \mathbb{M}} \mathbb{1}(|V_M \cap S| = 2), \quad \alpha > 1, \tag{3}$$

where $\mathbb{1}$ is a indicator function. Here, $\text{cut}_M(S, \bar{S})$ distinguishes the number of nodes of the 4-node motif M being assigned to S and $\bar{S}$. Specifically, it adds a higher penalty for the cut to the cases where two nodes in M are assigned to S and two nodes are assigned to $\bar{S}$, as compared to the case where one node or three nodes are assigned to S, by letting $\alpha > 1$ in Eq. (3). This is because the 1-vs-3 split could still keep interaction information from both GRN and chromatin interaction network, and the 2-vs-2 split will lose either of the information. We will discuss the choice of α later in this section. $\text{Vol}_M(S)$ is the number of nodes in the occurrences of M that are in S, which is defined as:

$$\text{Vol}_M(S) = \sum_{i \in S} \sum_{M \in \mathbb{M}} \mathbb{1}(i \in V_M). \tag{4}$$

Similarly, we define the subgraph conductance of the graph G based on motif M, φ_M^G as $\text{argmin}_S \varphi_M(S)$. In the following procedures of the algorithm, we show that the motif conductance is equivalent to the normal conductance in a projection of the graph by calculating the subgraph adjacency matrix. Thus, finding the set S that achieves the minimum subgraph conductance is also NP-hard, following that it is NP-hard to find the minimal $\varphi_G(S)$. We describe our algorithm MOCHI to find HIMs that approximate the solution.

1 – Calculate subgraph adjacency matrix $W_M(G)$

We first calculate the subgraph adjacency matrix $W_M(G)$ by

$$[W_M(G)]_{ij} = \sum_{M \in \mathbb{M}} \mathbb{1}(i \in V_M, j \in V_M), \tag{5}$$

where $[W_M(G)]_{ij}$ is the number of occurrences of the subgraph M in G that cover both i and j (see example in Fig. 1C). For example, if both i and j are TFs, $[W_M]_{ij}$ reflects the number of paired gene loci that are spatially proximal to each other than expected and that are also co-regulated by TFs i and j. If both i and j are genes, $[W_M]_{ij} = 0$ if i and j are not spatially proximal to each other than expected. Otherwise, $[W_M]_{ij}$ is the number of paired TFs that co-regulate i and j. Generally, $W_M(G)$ is symmetric and $[W_M(G)]_{ij} \geq 0$. Thus $W_M(G)$ can be viewed as the adjacency matrix of an undirected weighted network. Let G_M denote the network with $W_M(G)$ as the adjacency matrix (see Fig. 1D for example). It is important to note that there are genes or TFs that may not be in any occurrence of M, which would lead to zero vectors in the corresponding rows and columns in $W_M(G)$. These singleton nodes in G_M would be removed before the next step.

2 – Apply Fiedler partitions to find a cluster in G_M

We utilize Fiedler partitions similar to the algorithm in [48] to find a cluster S in graph G_M, where $\varphi_{G_M}(S)$ is close to the global optimal conductance of the graph: $\varphi(G_M)$. Recall that $\varphi(G_M)$ is the minimum of $\varphi_{G_M}(S_1)$ over all possible sets S_1. We show that S is a near optimal cluster in G in Supplementary Methods A.3. The method is described as follows:

- Calculate the normalized Laplacian matrix of $W_M(G)$:

$$\boldsymbol{\mathcal{L}} = \boldsymbol{\mathcal{I}} - D_{G_M}^{-1/2} W_M(G) D_{G_M}^{-1/2}, \tag{6}$$

  where $\boldsymbol{\mathcal{I}}$ is a identity matrix, D_{G_M} with $[D_{G_M}]_{ii} = \sum_{j=1}^{N} (W_M(G))_{ij}$ is the diagonal degree matrix of G_M.
- Calculate the eigenvector v of the second smallest eigenvalue of $\boldsymbol{\mathcal{L}}$
- Find the index vector $(\alpha_1, \ldots, \alpha_N)$, where α_k is the k-th smallest value of $D_{G_M}^{-1/2} v$.
- $S = \underset{S_k, 1 \leq k \leq N}{\text{argmin}} \varphi_{G_M}(S_k)$, where $S_k = \{\alpha_1, \ldots, \alpha_k\}$.

The sets S and $\bar{S}$ are two disjoint clusters for the heterogeneous network G.

3 – Apply recursive bipartitioning to find multiple HIMs

We then utilize recursive bipartitioning to find multiple HIMs. We use a very different strategy than the one in [48] to select which cluster to split at each iteration, in order to specifically allow overlapping motif clusters (HIMs) with shared TFs. At each iteration, we split one HIM into 2 child HIMs. After iteration $\ell - 1$, there are ℓ HIMs: $S_1, S_2, \ldots, S_\ell$.

At next iteration ℓ, one HIM S_k is selected if the graph it forms, G_k, has the lowest subgraph conductance value $\varphi_M^{G_k}$ among $\varphi_M^{G_j}$, $1 \leq j \leq \ell$. We set a threshold t_1 for $\varphi_M^{G_k}$. If $\varphi_M^{G_k} \leq t_1$, S_k will be split into two child HIMs $S_k(c)$ and $\overline{S_k(c)}$ by treating the induced heterogeneous subnetwork as a new network G_k and repeating Steps (1) and (2) for graph G_k. However, if the partition of graph G_k would lead to zero motif occurrence in either of its child graphs, we would stop partitioning this graph, add penalty to its conductance value (to make sure it would not be selected to partition again), and move on to the next iteration. Otherwise, when $\varphi_M^{G_k} > t_1$, the recursive bipartitioning process will stop as all the HIM's subgraph conductance value passes the threshold.

4 – Find overlapping HIMs

Finally, we reconcile the HIMs from the clustering history tree to find overlapping HIMs. This step is added because the HIMs after Step (3) share no TFs. To reconcile the results, we first trace back the ancestor HIMs up to certain generations for each HIM based on the conductance value of its ancestor $\varphi_M^{G_{ancestor_i}}$, where $i = \{1, 2, 3...\}$ denotes for the 'parent', 'grandparent' of the HIM. We trace along the tree until $\varphi_M^{G_{ancestor_i}} \leq t_2$, where t_2 denotes another threshold. Clearly, t_2 has to be smaller than t_1 to make this process practical. Next, we pool together the TFs from the HIM and from its ancestor HIMs. We sequentially remove pooled TFs from the HIM based on their contribution to the number of occurrences of the subgraph M based on the heterogeneous graph this HIM represents, and stop this process when removing certain TF would cause a large decrease in the number of subgraphs.

Pseudocode and complexity of our algorithm

The pseudocode of our MOCHI algorithm is presented in the Supplementary Methods A.1. The runtime of the algorithm is $O(t^2c^2)$, where t and c are the number of TFs and the number of gene loci in the graph, respectively (detailed analysis in Supplementary Methods A.2).

Summary of our algorithm

Given a heterogeneous network from chromatin interactome network and GRN, we developed MOCHI to identify multiple and overlapping HIMs, which represent clusters of genes and TFs where the genes are interacting more frequently than expected and are also co-regulated by the same set of TFs. Our algorithm has a few key differences as compared to the subgraph conductance method in [48]. First, the input of our algorithm is a heterogeneous network with different types of nodes (TFs and gene loci), which are treated differently, while the input network for the method [48] is homogeneous. Second, the algorithm in [48] will not explicitly find multiple overlapping clusters. In MOCHI, we further developed a recursive bipartitioning method to find multiple HIMs that may overlap. Specifically, we selected a HIM to split if it has the smallest motif conductance among the HIMs at each interaction. In other words, we split the HIM that has the clearest pattern of multiple clusters. HIMs with overlapping TFs will be split in the late stage of iterations, and the overlapping information is encoded in the clustering history tree.

The recent method on hypergraph clustering [50] can be applied to identify non-overlapping HIMs where a hyperedge is defined as the motif M. However, similar to the method in [48], it was not designed to identify overlapping clusters, i.e., the method would not be able to find multiple overlapping HIMs. Our method also has clear differences as compared to previous works on multilayer network clustering (see reviews in [51–53]). First, the inputs are different. A multilayer network has only one type of nodes and different types of interactions connecting nodes within the same layer and between layers. The heterogeneous network in this paper has different types of nodes (TF proteins and gene loci) and also edges. Previous multilayer network clustering methods are not directly applicable to identify HIMs. Second, the outputs are different. The majority of multilayer network clustering methods aim to find clusters that are either consistently observed across multiple layers or observed only in a specific layer, which are conceptually different from HIMs.

Conclusion and Discussion

To better understand the heterogeneous nature of different molecules in cell nucleus, new computational models are needed to consider different types of molecular interacting networks. In this paper, we developed MOCHI to specifically consider two types of different interactomes in cell nucleus: (1) a network of chromosomal interactions between different gene loci, and (2) a GRN where TF proteins bind to the genomic loci to regulate target genes' transcription. MOCHI is able to identify network organizations where nodes of TFs (from GRN) and genomic bins (from both chromatin interactome and GRN) cooperatively form distinct types of network clusters, which we call HIMs, utilizing a motif clustering framework for heterogeneous networks. To the best of our knowledge, this is the first algorithm that can simultaneously analyze these heterogeneous networks within the nucleus to discover important network structures. By applying MOCHI to five different human cell types, we made new observations to demonstrate the biological relevance of HIMs.

Our method has multiple methodological contributions. We further developed the motif conductance clustering method [48] to find overlapping HIMs in heterogeneous networks. We successfully applied the modified method to identify HIMs. Moreover, our method can be further modified to identify other types of potentially interesting HIMs in heterogeneous biological networks by replacing the 4-node motif M with relevant motifs. Taken together, our study shows the utility of MOCHI to identify HIMs based on complex heterogeneous molecular interactomes.

HIMs as a type of unit have a few advantages over spatially colocalized TF pairs. First, the pairwise spatial colocalization does not necessarily mean simultaneously spatial colocalization between 3 or more TFs that are pairwise colocalized, because different pairs could colocalize around different chromosome regions that are distal to each other in the nucleus. For example, a recent study investigated spatial co-localization between paired TFs from their binding sites' spatial proximity that is revealed by Hi-C data [30]. Among the 14 significantly colocalized TF pairs in A1 subcompartments in GM12878 [30], 10 pairs are in at least one GM12878 HIMs. However, among the 8 triplets of TFs that are pairwise colocalized, only 3 triplets are in HIMs. Second, in comparison to colocalized TFs, HIMs provide an extra layer of information: the set of genes that are in spatial contact more frequently than expected and that are regulated by the same group of TFs. Thus HIMs represent a finer scale unit to study the connection between nucleome structure and function.

How to explain the formation of HIMs? In Fig. 6, we illustrate a possible model of HIMs within cell nucleus. HIMs (light pink domains) are more toward interior with a group of interacting TFs and chromatin loci. Such enrichment may contribute to interpreting the set of TFs in a HIM that cooperatively regulate target genes, which also have higher contact frequency than expected. This is conceptually consistent with recently reported co-localization of TF pairs [30]. Some of these TF clusters may come from nuclear compartments, such as nuclear speckle, that are enriched by transcription factories and promote integrated regulation of gene expression [54]. Indeed, we found that the majority of the identified HIMs in K562 are close to nuclear speckles. The definitions of HIMs may also have some intrinsic connections with the recently formulated mechanism of nuclear subcompartment formation, where TFs and their potential regulated genes/chromatin are trapped by localized liquid-like chambers through phase separation process [55, 56]. Evidence has been shown that phase separation can well explain the formation of super-enhancer mediated gene regulation [56], transcription regulation complex in nucleoli [57], and the formation of heterochromatin [58, 59]. From our analysis, we found that the genes assigned to HIMs are enriched with contacts with super-enhancers. The genes more consistently in HIMs are enriched with biological processes related to chromosomal organizations and transcription. However, there could be other possible formation mechanisms for HIMs. More experimental data are needed to further evaluate the formation mechanisms of HIMs and their functional significance. Nevertheless, HIMs may become a useful type of nuclear genome unit in integrating heterogeneous nucleome mapping data, which has the potential to provide new insights into the interplay among different constituents in nucleus and its role in nucleome structure and function.

References

[1] Wendy A Bickmore and Bas van Steensel. Genome architecture: domain organization of interphase chromosomes. *Cell*, 152(6):1270–1284, 2013.

[2] Christian Lanctôt, Thierry Cheutin, Marion Cremer, Giacomo Cavalli, and Thomas Cremer. Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nature Reviews Genetics*, 8(2):104–115, 2007.

[3] Boyan Bonev and Giacomo Cavalli. Organization and function of the 3d genome. *Nature Reviews Genetics*, 17(11):661–678, 2016.

[4] M Jordan Rowley and Victor G Corces. Organizational principles of 3d genome architecture. *Nature Reviews Genetics*, page 1, 2018.

[5] Andrew S Belmont. Large-scale chromatin organization: the good, the surprising, and the still perplexing. *Current Opinion in Cell Biology*, 26:69–78, 2014.

[6] Bas van Steensel and Andrew S Belmont. Lamina-associated domains: Links with chromosome architecture, heterochromatin, and gene repression. *Cell*, 169(5):780–791, 2017.

[7] Erez Lieberman-Aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.

[8] Suhas SP Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric 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.

[9] Zhonghui Tang, Oscar Junhong Luo, Xingwang Li, Meizhen Zheng, Jacqueline Jufen Zhu, Przemyslaw Szalaj, Pawel Trzaskoma, Adriana Magalska, Jakub Wlodarczyk, Blazej Ruszczycki, et al. CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell*, 163(7): 1611–1627, 2015.

[10] Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 2012.

[11] Tom Misteli. Beyond the sequence: cellular organization of genome function. *Cell*, 128(4):787–800, 2007.

[12] David U Gorkin, Danny Leung, and Bing Ren. The 3d genome in transcriptional regulation and pluripotency. *Cell Stem Cell*, 14(6):762–775, 2014.

[13] Michael H Kagey, Jamie J Newman, Steve Bilodeau, Ye Zhan, David A Orlando, Nynke L van Berkum, Christopher C Ebmeier, Jesse Goossens, Peter B Rahl, Stuart S Levine, et al. Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467(7314):430–435, 2010.

[14] Jop Kind, Ludo Pagie, Sandra S de Vries, Leila Nahidiazar, Siddharth S Dey, Magda Bienko, Ye Zhan, Bryan Lajoie, Carolyn A de Graaf, Mario Amendola, et al. Genome-wide maps of nuclear lamina interactions in single human cells. *Cell*, 163(1):134–147, 2015.

[15] Yu Chen, Yang Zhang, Yuchuan Wang, Liguo Zhang, Eva K Brinkman, Stephen A Adam, Robert Goldman, Bas van Steensel, Jian Ma, and Andrew S Belmont. Mapping 3d genome organization relative to nuclear compartments using tsa-seq as a cytological ruler. *J Cell Biol*, pages jcb–201807108, 2018.

[16] Mark B Gerstein, Anshul Kundaje, Manoj Hariharan, Stephen G Landt, Koon-Kiu Yan, Chao Cheng, Xinmeng Jasmine Mu, Ekta Khurana, Joel Rozowsky, Roger Alexander, et al. Architecture of the human regulatory network derived from encode data. *Nature*, 489(7414):91–100, 2012.

[17] Shane Neph, Andrew B Stergachis, Alex Reynolds, Richard Sandstrom, Elhanan Borenstein, and John A Stamatoyannopoulos. Circuitry and dynamics of human transcription factor regulatory networks. *Cell*, 150(6):1274–1286, 2012.

[18] Eric H Davidson. *The regulatory genome: gene regulatory networks in development and evolution*. Aca-

demic Press, San Diego, 2006.

[19] Juan M Vaquerizas, Sarah K Kummerfeld, Sarah A Teichmann, and Nicholas M Luscombe. A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics*, 10(4):252–263, 2009.

[20] Samuel A Lambert, Arttu Jolma, Laura F Campitelli, Pratyush K Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R Hughes, and Matthew T Weirauch. The human transcription factors. *Cell*, 172(4):650–665, 2018.

[21] Fulai Jin, Yan Li, Jesse R Dixon, Siddarth Selvaraj, Zhen Ye, Ah Young Lee, Chia-An Yen, Anthony D Schmitt, Celso A Espinoza, and Bing Ren. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, 503(7475):290–294, 2013.

[22] Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(1):S7, 2006.

[23] Nir Yosef, Alex K Shalek, Jellert T Gaublomme, Hulin Jin, Youjin Lee, Amit Awasthi, Chuan Wu, Katarzyna Karwacz, Sheng Xiao, Marsela Jorgolli, et al. Dynamic regulatory network controlling th17 cell differentiation. *Nature*, 496(7446):461–468, 2013.

[24] Dechao Tian, Quanquan Gu, and Jian Ma. Identifying gene regulatory network rewiring using latent differential graphical models. *Nucleic Acids Research*, 44(17):e140–e140, 2016.

[25] Casey S Greene, Arjun Krishnan, Aaron K Wong, Emanuela Ricciotti, Rene A Zelaya, Daniel S Himmelstein, Ran Zhang, Boris M Hartmann, Elena Zaslavsky, Stuart C Sealfon, et al. Understanding multicellular function and disease with human tissue-specific networks. *Nature Genetics*, 47(6):569–576, 2015.

[26] Indika Rajapakse, David Scalzo, Stephen J Tapscott, Steven T Kosak, and Mark Groudine. Networking the nucleus. *Molecular Systems Biology*, 6(1):395, 2010.

[27] Indika Rajapakse and Mark Groudine. On emerging nuclear order. *The Journal of Cell Biology*, 192(5): 711–721, 2011.

[28] Ralph Stadhouders, Enrique Vidal, François Serra, Bruno Di Stefano, François Le Dily, Javier Quilez, Antonio Gomez, Samuel Collombet, Clara Berenguer, Yasmina Cuartero, et al. Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. *Nature Genetics*, page 1, 2018.

[29] Ruggero Cortini and Guillaume Filion. Principles of transcription factor traffic on folded chromatin. *Nature Communications*, pages doi:10.1038/s41467–018–04130–x, 2018.

[30] Xiaoyan Ma, Daphne Ezer, Boris Adryan, and Tim J Stevens. Canonical and single-cell Hi-C reveal distinct chromatin interaction sub-networks of mammalian transcription factors. *Genome Biology*, 19 (174), 2018.

[31] Daniel Marbach, David Lamparter, Gerald Quon, Manolis Kellis, Zoltán Kutalik, and Sven Bergmann. Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nature Methods*, 2016.

[32] Tim Wang, Kıvanç Birsoy, Nicholas W Hughes, Kevin M Krupczak, Yorick Post, Jenny J Wei, Eric S Lander, and David M Sabatini. Identification and characterization of essential genes in the human genome. *Science*, 350(6264):1096–1101, 2015.

[33] Sofia A Quinodoz, Noah Ollikainen, Barbara Tabak, Ali Palla, Jan Marten Schmidt, Elizabeth Detmar, Mason M Lai, Alexander A Shishkin, Prashant Bhat, Yodai Takei, et al. Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell*, 2018.

[34] Peter J Thul, Lovisa Åkesson, Mikaela Wiking, Diana Mahdessian, Aikaterini Geladaki, Hammou Ait Blal, Tove Alm, Anna Asplund, Lars Björk, Lisa M Breckels, et al. A subcellular map of the human proteome. *Science*, 356(6340):eaal3321, 2017.

[35] Attila Németh, Ana Conesa, Javier Santoyo-Lopez, Ignacio Medina, David Montaner, Bálint Péterfia, Irina Solovei, Thomas Cremer, Joaquin Dopazo, and Gernot Längst. Initial genomics of the human nu-

cleolus. *PLoS genetics*, 6(3):e1000889, 2010.

[36] Annette Denker and Wouter de Laat. The second decade of 3C technologies: detailed insights into nuclear organization. *Genes & Development*, 30(12):1357–1382, 2016.

[37] Tyrone Ryba, Ichiro Hiratani, Junjie Lu, Mari Itoh, Michael Kulik, Jinfeng Zhang, Stephen Dalton, and David M Gilbert. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome research*, 2010.

[38] Denes Hnisz, Brian J Abraham, Tong Ihn Lee, Ashley Lau, Violaine Saint-André, Alla A Sigova, Heather A Hoke, and Richard A Young. Super-enhancers in the control of cell identity and disease. *Cell*, 155(4):934–947, 2013.

[39] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57, 2009.

[40] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13, 2008.

[41] Marilyn Safran, Irina Dalah, Justin Alexander, Naomi Rosen, Tsippi Iny Stein, Michael Shmoish, Noam Nativ, Iris Bahir, Tirza Doniger, Hagit Krug, et al. GeneCards Version 3: the human gene integrator. *Database*, 2010, 2010.

[42] C Rosenfeld, MA Cheever, and A Gaiger. WT1 in acute leukemia, chronic myelogenous leukemia and myelodysplastic syndrome: therapeutic potential of WT1 targeted therapies, 2003.

[43] KA Dorritie, JA McCubrey, and DE Johnson. STAT transcription factors in hematopoiesis and leukemogenesis: opportunities for therapeutic intervention. *Leukemia*, 28(2):248, 2014.

[44] Y Huang, J Chen, C Lu, J Han, G Wang, C Song, S Zhu, C Wang, G Li, J Kang, et al. HDAC1 and KLF4 interplay critically regulates human myeloid leukemia cell proliferation. *Cell Death & Disease*, 5(10): e1491, 2014.

[45] Jason Ernst and Manolis Kellis. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, 28(8):817, 2010.

[46] Jason Ernst, Pouya Kheradpour, Tarjei S Mikkelsen, Noam Shoresh, Lucas D Ward, Charles B Epstein, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael Coyne, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43, 2011.

[47] Fan Chung. Four cheeger-type inequalities for graph partitioning algorithms. *Proceedings of ICCM, II*, pages 751–772, 2007.

[48] Austin R Benson, David F Gleich, and Jure Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, 2016.

[49] Charalampos E Tsourakakis, Jakub Pachocki, and Michael Mitzenmacher. Scalable motif-aware graph clustering. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1451–1460. International World Wide Web Conferences Steering Committee, 2017.

[50] Pan Li and Olgica Milenkovic. Inhomogeneous hypergraph clustering with applications. In *Advances in Neural Information Processing Systems*, pages 2308–2318, 2017.

[51] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. *Journal of complex networks*, 2(3):203–271, 2014.

[52] Stefano Boccaletti, Ginestra Bianconi, Regino Criado, Charo I Del Genio, Jesús Gómez-Gardenes, Miguel Romance, Irene Sendina-Nadal, Zhen Wang, and Massimiliano Zanin. The structure and dynamics of multilayer networks. *Physics Reports*, 544(1):1–122, 2014.

[53] Brandon Oselio, Sijia Liu, and Alfred Hero. Multilayer social networks. In *Cooperative and Graph Signal Processing*, pages 679–697. Elsevier, 2018.

[54] Lukasz Galganski, Martyna O Urbanek, and Wlodzimierz J Krzyzosiak. Nuclear speckles: molecular organization, biological function and role in disease. *Nucleic Acids Research*, 45(18):10350–10368, 2017.

[55] Yongdae Shin and Clifford P Brangwynne. Liquid phase condensation in cell physiology and disease. *Science*, 357(6357), September 2017.

[56] Denes Hnisz, Krishna Shrinivas, Richard A Young, Arup K Chakraborty, and Phillip A Sharp. A phase separation model for transcriptional control. *Cell*, 169(1):13–23, March 2017.

[57] Marina Feric, Nilesh Vaidya, Tyler S Harmon, Diana M Mitrea, Lian Zhu, Tiffany M Richardson, Richard W Kriwacki, Rohit V Pappu, and Clifford P Brangwynne. Coexisting liquid phases underlie nucleolar subcompartments. *Cell*, 165(7):1686–1697, June 2016.

[58] Amy R Strom, Alexander V Emelyanov, Mustafa Mir, Dmitry V Fyodorov, Xavier Darzacq, and Gary H Karpen. Phase separation drives heterochromatin domain formation. *Nature*, 547(7662):241–245, July 2017.

[59] Adam G Larson, Daniel Elnatan, Madeline M Keenen, Michael J Trnka, Jonathan B Johnston, Alma L Burlingame, David A Agard, Sy Redding, and Geeta J Narlikar. Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. *Nature*, 547(7662):236–240, July 2017.

[60] Jonathan A. Kelner, Lorenzo Orecchia, Aaron Sidford, and Zeyuan Allen Zhu. A simple, combinatorial algorithm for solving SDD systems in nearly-linear time. *CoRR*, abs/1301.6628, 2013. URL http://arxiv.org/abs/1301.6628.

[61] R Scott Hansen, Sean Thomas, Richard Sandstrom, Theresa K Canfield, Robert E Thurman, Molly Weaver, Michael O Dorschner, Stanley M Gartler, and John 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.

[62] Robert E Thurman, Nathan Day, William S Noble, and John A Stamatoyannopoulos. Identification of higher-order functional domains in the human ENCODE regions. *Genome Research*, 17(6):917–927, 2007.

[63] Kate R Rosenbloom, Cricket A Sloan, Venkat S Malladi, Timothy R Dreszer, Katrina Learned, Vanessa M Kirkup, Matthew C Wong, Morgan Maddren, Ruihua Fang, Steven G Heitner, et al. ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Research*, 41(D1):D56–D63, 2012.

[64] Edward L Huttlin, Raphael J Bruckner, Joao A Paulo, Joe R Cannon, Lily Ting, Kurt Baltier, Greg Colby, Fana Gebreab, Melanie P Gygi, Hannah Parzen, et al. Architecture of the human interactome defines protein communities and disease networks. *Nature*, 2017.

[65] Andrew Chatr-Aryamontri, Bobby-Joe Breitkreutz, Sven Heinicke, Lorrie Boucher, Andrew Winter, Chris Stark, Julie Nixon, Lindsay Ramage, Nadine Kolas, Lara ODonnell, et al. The BioGRID interaction database: 2013 update. *Nucleic Acids Research*, 41(D1):D816–D823, 2012.

[66] Andreas Ruepp, Brigitte Waegele, Martin Lechner, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Gisela Fobo, Goar Frishman, Corinna Montrone, and H-Werner Mewes. CORUM: the comprehensive resource of mammalian protein complexes2009. *Nucleic Acids Research*, 38(suppl_1):D497–D501, 2009.

[67] Andrea Franceschini, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguez, Peer Bork, Christian Von Mering, et al. STRING v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, 41(D1):D808–D815, 2012.

[68] ENCODE Project Consortium et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57, 2012.

[69] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL https://www.R-project.org/.

[70] Job Dekker, Marc A Marti-Renom, and Leonid A Mirny. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics*, 14(6):390, 2013.

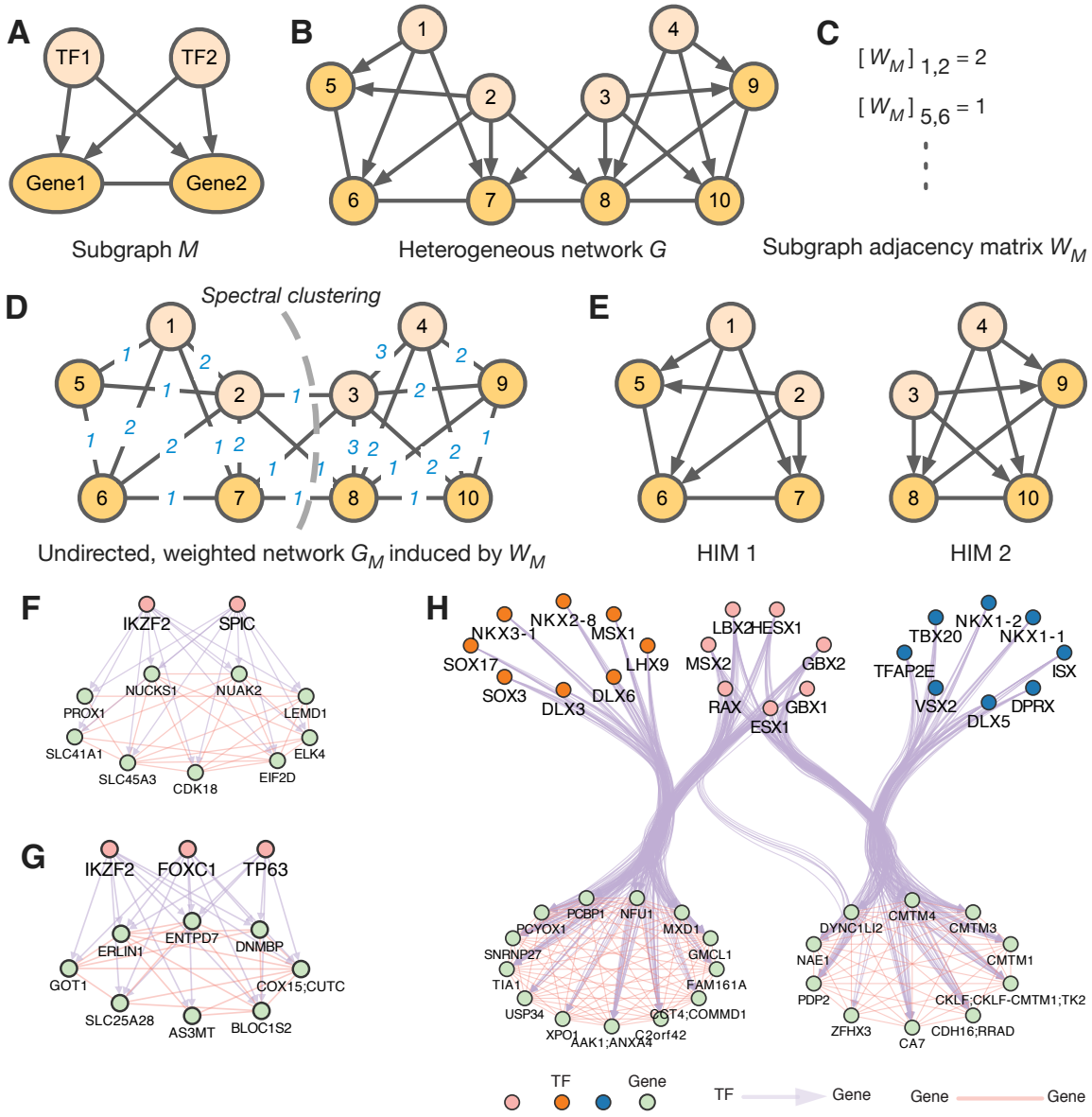

Figure 1: Workflow of our MOCHI algorithm and output example of HIMs. The network has both gene-gene spatial proximity and TF-gene regulation relationships. **(A)** A 4-node motif M represents the smallest HIM. Here directed interaction represents a TF-gene regulation relationship, an undirected interaction represents that the two genes are spatially more proximal to each other than expected. **(B)** Given a heterogeneous network G, we find HIMs by minimizing the motif conductance (see Eq. 2). **(C)** We compute the subgraph adjacency matrix W_M with $[W_M]_{ij}$ being the number of occurrences of M that have both nodes i and j. **(D)** The weighted network G_M is defined from adjacency matrix W_M. **(E)** Spectral clustering will find clusters in G_M. We recursively apply the method to find multiple HIMs and overlapping HIMs. **(F-G)** 2 HIMs as example in GM12878. **(H)** Example of two overlapping HIMs in GM12878 sharing 7 TFs (the group with pink nodes). TFs in orange and pink nodes form one HIM with their target genes (bottom left). TFs in pink and blue nodes form another HIM with their target genes (bottom right). Note that the directed interactions from TFs to their target genes are bundled in **(H)**.

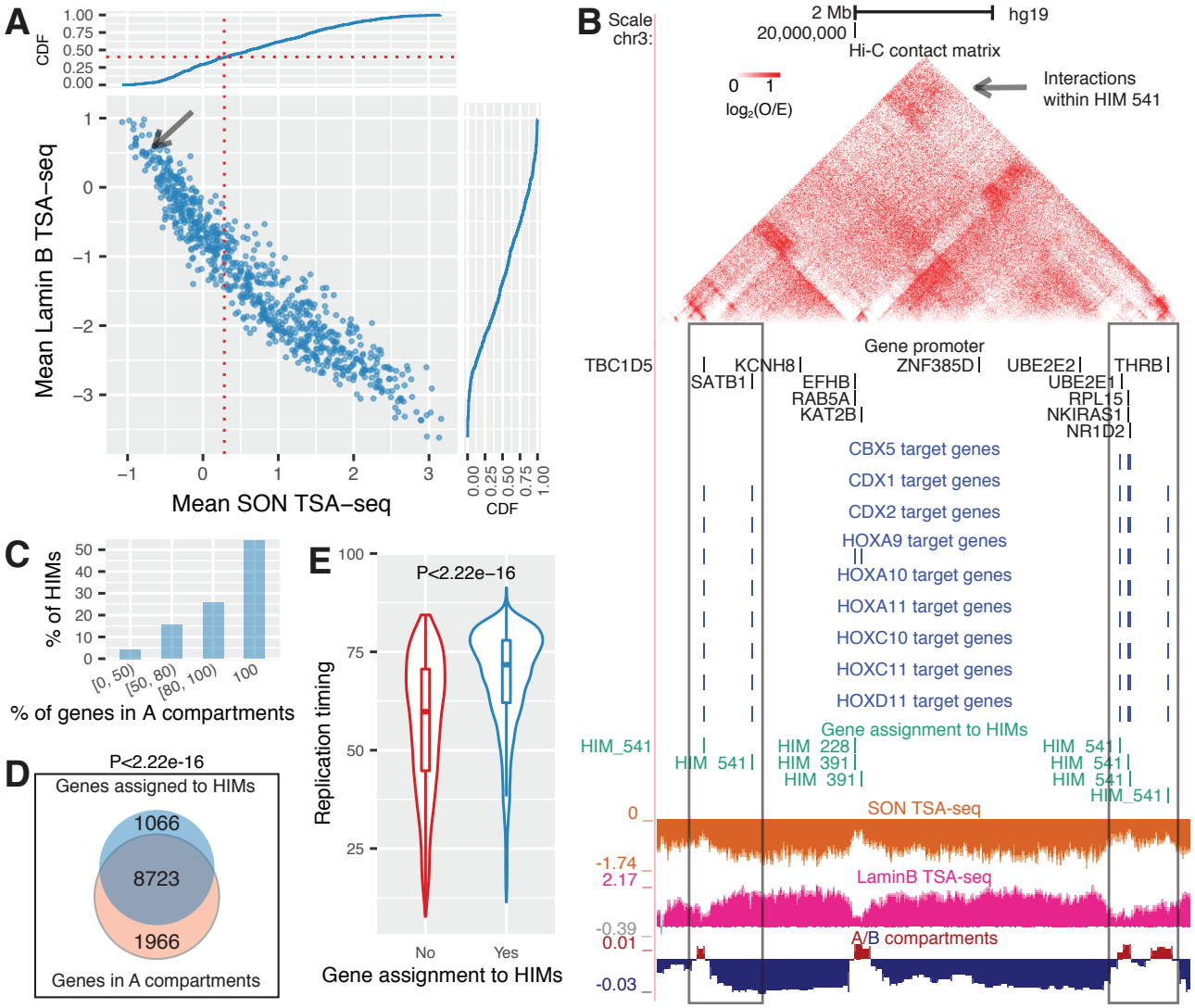

Figure 2: HIMs tend to be close to nuclear interior, in particular, speckles. **(A)** Scatter plot shows the mean SON TSA-seq score and mean Lamin B TSA-seq score of the genes in each HIM. Each dot represents a HIM. The curves on top and right are cumulative density functions (CDF). The red vertical dotted line represents the mean SON TSA-seq at 0.284 (approx. within $0.518\,\mu\mathrm{m}$ of nuclear speckles [15]). **(B)** HIM #541 with low mean SON TSA-seq (pointed by the arrow). The heatmap shows the upper-triangle part of the Hi-C contact matrix (O/E) of the 10kb-sized bins in the chromosome region that covers the genes in the HIM. **(C)** Barplot shows the proportion of HIMs with a varied proportion of genes in A compartments. **(D)** Venn diagram shows that the genes assigned to HIMs are enriched in A compartments. **(E)** Violin and boxplot compare the replication timing of the genes assigned to HIMs and the other genes in the heterogeneous network of K562. Here the HIMs are identified in K562 cell line. The spatial location features of HIMs in other cell types are in Fig. S2.

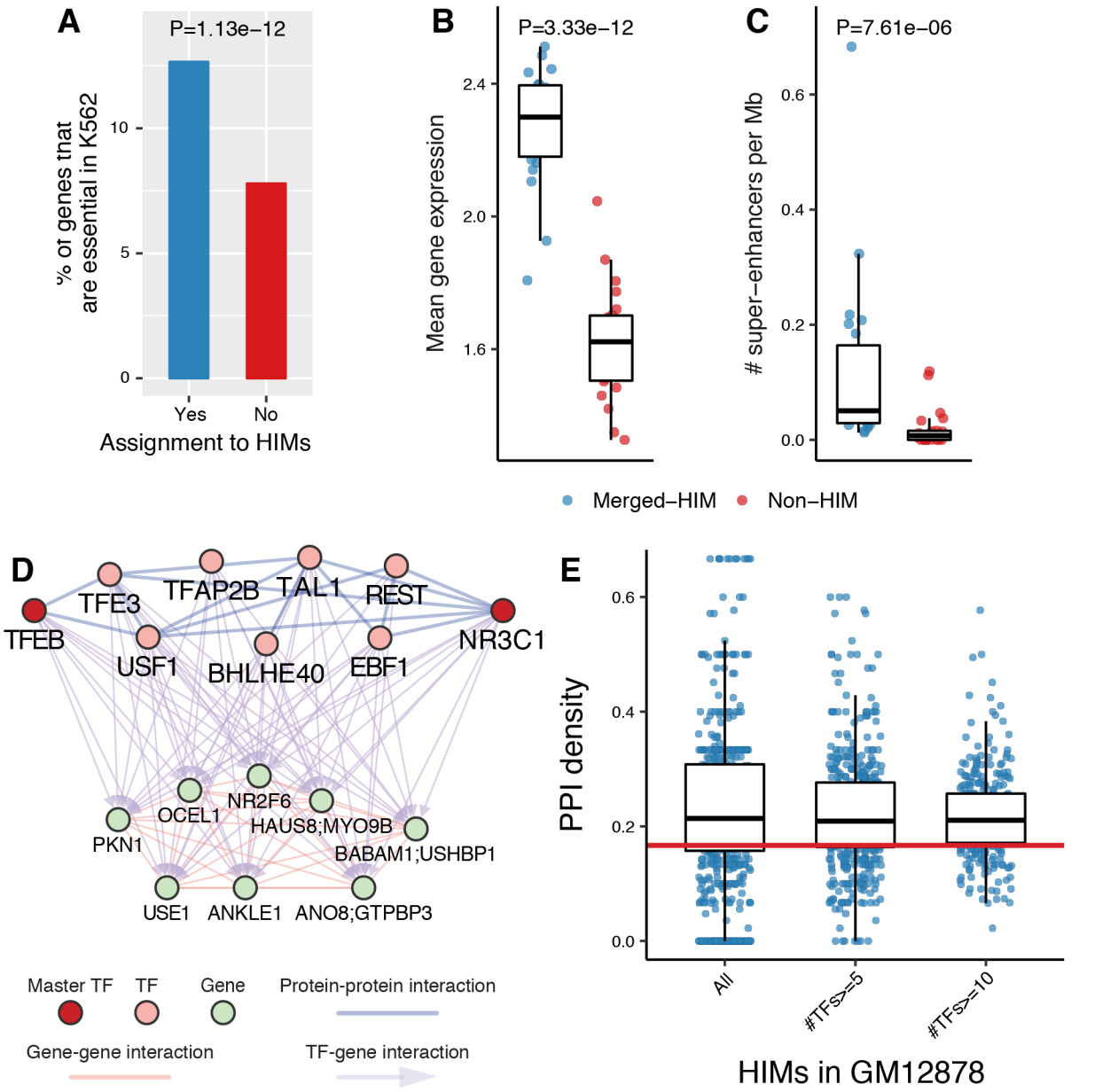

Figure 3: HIMs are enriched with essential genes, super-enhancers, and protein-protein interactions. **(A)** Barplots show the proportions of genes that are K562 essential genes in the genes assigned to HIMs and not to HIMs. **(B-C)** Functional properties of the genes in the identified HIMs in K562 cell line. To make fair comparison, we stratify the genes assigned to HIMs by chromsome number and called resulted clusters as merged-HIM clusters. Similarly, we derived non-HIM clusters from the genes in the heterogeneous networks but not assigned to HIMs. P values are computed by the paired two-sample Wilcoxon rank-sum test. **(B)** Boxplot shows the average gene expression level of the genes in a cluster. **(C)** Boxplot shows the normalized number of super-enhancers related to a cluster. **(D-E)** TFs in HIMs are enriched with protein-protein interactions (PPIs) between themselves. **(D)** One example HIM from GM12878 cell line shows that the 9 TFs in the HIM are connected by 14 PPI interactions. The sub-PPI network has a density at 0.389. The TFs NR3C1 and TFEB are master TFs in GM12878. **(E)** Boxplots show the distribution of the sub-PPI network density of the HIMs and the subsets of HIMs with at least n TFs, $n = 5, 10$. The red line is the density (0.158) of the whole PPI network.

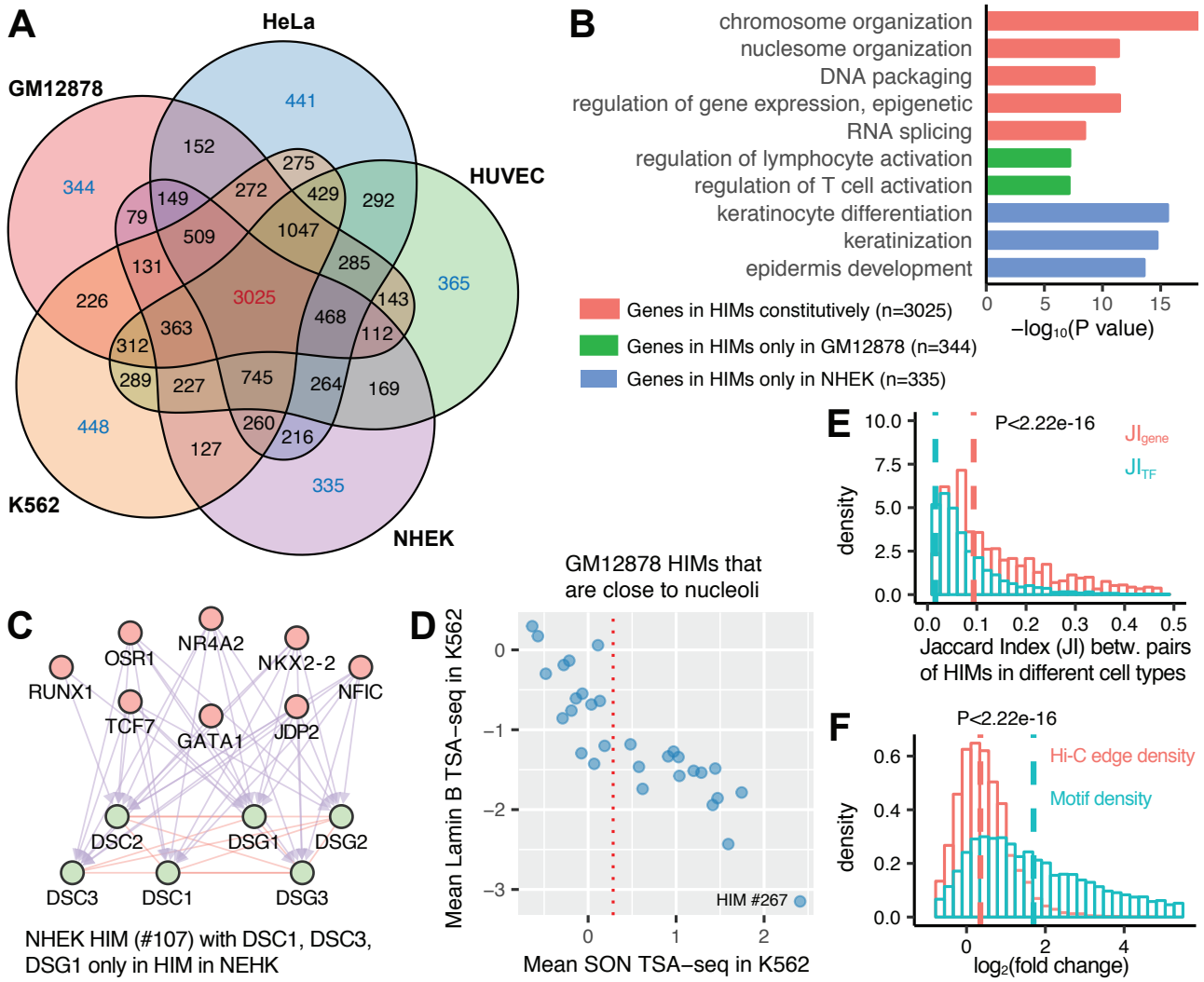

Figure 4: HIM comparisons in terms of genes and TFs across the cell types. **(A)** Venn diagram shows the assignment of genes in HIMs across five cell type. Numbers in each facet represents the gene number in each possible logic intersection relationship across five cell types. **(B)** The barplot shows the top GO terms or pathways enriched with the genes that are in HIMs constitutively or only in a particular cell type. **(C)** A NHEK HIM with 3 genes only assigned to HIMs in NHEK. All of its genes involve in keratinization pathway. Here the top and bottom nodes are the TFs and genes in the HIM, respectively. **(D)** Scatter plot shows the mean SON and Lamin B TSA-seq scores of the 30 GM12878 HIMs that are inferred as close to nuclear nucleoli in GM12878. The red vertical dotted line represents the mean SON TSA-seq score at 0.284. **(E)** Jaccard index on the genes/TFs between paired HIMs from different cell types. **(F)** Fold changes of motif M density and Hi-C edge density of each HIM between the cell type it is identified and another cell type. Here a vertical dashed line reprsents the median of a variable.

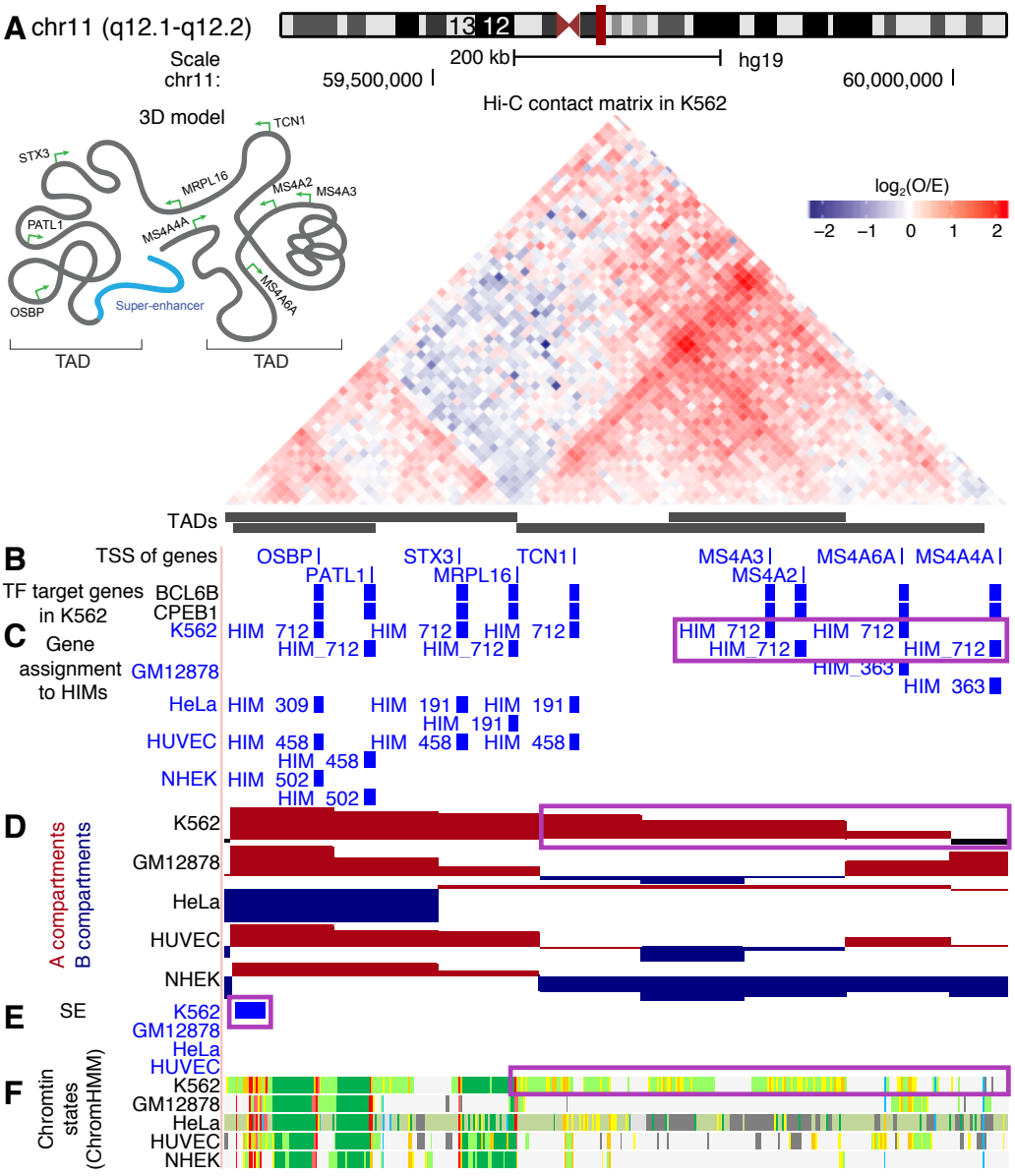

Figure 5: A K562 specific HIM with K562 specific chromatin interactome and functional annotations. **(A)** The 45 degree rotated upper triangle part of the contact matrix between the 10kb-sized bins in a chromosome region in K562. The region is segregated into 4 nested TADs. **(B)** Thin bars represent the transcriptional start sites (TSSs) of the genes that are in the heterogeneous networks. Thick bars represent the genes that are regulated by BCL6B or CPEB1 in K562. **(C)** The assignment of the genes to HIMs in K562 and the other cell types. **(D)** The assignment of the bins to A/B compartments. **(E)** The regions that are annotated as super-enhancers (SE). **(F)** The chromatin states inferred by ChromHMM based on multiple histone modification marks, where red and purple colors represent promoter, orange and yellow stand for enhancer, green represents transcribed regions, gray represents other functional regions such as repressed regions.

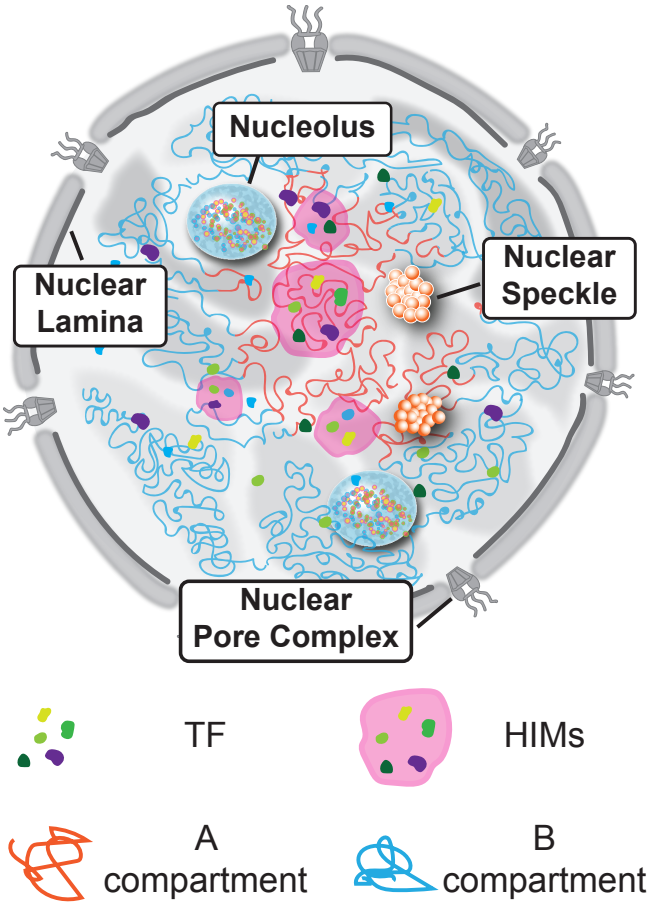

Figure 6: A possible model of HIMs within the cell nucleus.

Supplementary Materials

A Supplementary Methods

A.1 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**
    $G_k \leftarrow \text{argmin}_i Score_i$ 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 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
    $S_{S_k}, \bar{S}_{S_k}, Score_{S_k} =$ SPECTRAL CLUSTERING($W_M(G_{S_k}), N_{S_k}$)
    $S_{\bar{S}_k}, \bar{S}_{\bar{S}_k}, Score_{\bar{S}_k} =$ 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 ith smallest value in O
  $S \leftarrow \text{argmin}_k \varphi_{G_M}(S_k)$, where $S_k = \{\alpha_1, ..., \alpha_k\}$
  $Score \leftarrow \varphi_{G_M}(S)$
  return $S, \bar{S}, Score$
end function

A.2 Computational complexity

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 nearly linear time for this process [60]. Therefore, in the rest of this section, we will only discuss the computational complexity 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 complexity of $O(N^4)$, where N is the number of nodes in the graph. However, here 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 actually 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 $n_{4_{ij}} = \binom{n_{3_{ij}}}{2}$. For the search of triangles, given i and its neighbor j, we sum up all the TF nodes that they share, which gives us the complexity of $O(tc^2)$. For $W_{M_{ij}}$, where $i \in T, j \in C$, as we already know how many 3-node motifs would form between locus j and any another locus k (as calculated above), we can also calculate the number of 4-node motifs involving i, j by counting the 3-node motifs using the similar method. The complexity 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. To summarize, 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 TF and two gene loci. Note that, however, TFs only make up a small part of the nodes (about 4.5%). Also, as the network is somehow sparse, usually, we 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 further speed up the process, which makes the entire algorithm quite efficient in practice.

A.3 Clusters are near optimal

Here we 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). 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, $\varphi_M(S) = \varphi_M(\bar{S})$, thus we only need to show that S is near optimal.

We first formally state the near optimal claim and prove it in the subsequent paragraphs. Let $\varphi_M^\star$ be the minimum of subgraph conductance over all possible sets of nodes in G. Then S satisfies motif Cheeger inequality [47], i.e.,

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

which means that S is at most a quadratic factor away from the optimal cluster that achieves $\varphi_M^\star$.

We recall and define some mathematical notations. 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 by G_M. The subgraph conductance $\varphi_M(S)$ for G is defined in Eq. (2) and the conductance $\varphi_{G_M}(S)$ is defined in Eq. (3):

$$\varphi_M(S) = \frac{\text{cut}_M(S, \bar{S})}{\min[\text{Vol}_M(S), \text{Vol}_M(\bar{S})]} \tag{2}$$

$$\varphi_{G_M}(S) = \frac{\text{cut}_{G_M}(S, \bar{S})}{\min[\text{Vol}_{G_M}(S), \text{Vol}_{G_M}(S)]} \tag{3}$$

First, we prove that $\text{cut}_M(S, \bar{S}) = \frac{1}{3}\text{cut}_{G_M}(S, \bar{S})$. 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 have:

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

Next, we show that $\text{Vol}_M(S) = \frac{1}{3}\text{Vol}_{G_M}(S)$. Note that $|V_M| = 4$.

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

We have $\varphi_M(S) = \varphi_{G_M}(S)$ by definitions in Eq. (2), Eq. (3), and that $\text{cut}_M(S, \bar{S}) = \text{cut}_{G_M}(S)$ and $\text{Vol}_M(S) = \text{Vol}_{G_M}(S)$.

Finally, let $\varphi^\star_{G_M}$ be the minimum of conductance over all possible sets of nodes G_M. Then S satisfies the Cheeger inequality, i.e.,

$$\varphi_{G_M}(S) \leq 4\sqrt{\varphi^\star_{G_M}} \leq 1.$$

Comparison with the proofs in [48]

In Step (2), 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 [48] where W_M is computed based on a homogeneous motif and homogeneous network. The authors of [48] proved that S and $\bar{S}$ are near optimal for their case. However, the results in [48] are not 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. However, our proofs follow the same strategy as the proofs in [48].

A.4 Some features of the identified HIMs

Here we describe the HIM featurse that are not defined in the main text. The featuers are the topological structural features related to connection patterns of the genes and TFs in a given HIM within a heterogeneous network. They are the motif density, 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 range from 0 to 1.

- 4-node motif M density. It is the ratio of the number of occurrences of the motif M to the total number of possible occurrences of the motif M in a HIM. The maximal of the motif density is 1, which is achieved when every pair of genes in the cluster are connected with Hi-C interactions and every gene is regulated by each TF in the cluster. The triangle motif density used latter in Supplementary section 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 is connected by a Hi-C interaction with O/E$>$1. Where 0 means that no pair of genes is connected. Thus a higher density means that the HIM genes as a unit are more densely packed in the nucleus.
- GRN edge density. It is the density of the sub-GRN induced by the genes and TFs in the HIM. The maximal 1 is achieved when every gene is regulated by every TF in the HIM. The minimal 0 is achieved when no gene is regulated by no TFs in the HIM.

A.5 Collection and processing of data used in this study

In this work, we use data for five human cell types: GM12878, HeLa, HUVEC, K562, and NHEK. For Hi-C related data, including KR normalized contact frequency matrices by in-situ Hi-C and O/E contact frequency matrices were from [8]. We downloaded the data from GEO with the accession number GSE63525. We calculated the A/B compartments for each chromosome using the first principal component of the O/E contact frequency matrix as the same in [7]. For intra-chromosomal contacts, we first filtered the genome-wide KR normalized contact matrix by only keeping intra-chromosomal contacts higher than expected values, aiming to reduce intra-chromosomal contacts due to random chromatin collisions. In addition, we focused on analyzing HIMs with the 1D distance between farthest apart genes in a HIM at most comparable to the size of compartments. The 99-th percentile of the size of compartments is around 10Mb in 4 out of the 5 cell types. Thus we chosen the 1D distance cutoff universally as 10Mb and only kept the remaining intra-chromosome contacts that connect bins within 10Mb across the cell types. Then only the top 1% inter-chromosomal contacts were kept by choosing the cutoff as the 99-th percentile of the genome-wide inter-chromosomal contacts. The remaining inter-chromosomal contacts have at least 2.17 KR normalized Hi-C contacts. Processed replication timing data with the GEO accession number GSE34399 [61, 62] were downloaded from the UCSC Genome Browser [63].

GRN data were downloaded from [31], where directed TF-gene interactions were inferred by simultaneously considering the TF binding motifs and gene expression level. Briefly, an interaction between a TF and a gene is called if (1) the TF has binding motifs on the enhancer or promoter regions of the gene; and (2) the co-expression level between the TF and the enhancer or promoter of the gene is high.

Protein-protein interactions (PPIs) were downloaded from BioPlex2 [64], BioGrid [65], CORUM [66], and STRING [67]. We first extracted the PPIs between the 591 TFs in the GRNs from these public data. We then combined them into a whole PPI network after merging duplicated PPIs. Note that the GRNs have the same set of TF protein. Thus the whole PPI network is suitable for all the 5 different cell types.

Essential genes that are important for proliferation and survival in 4 cancer cell lines were downloaded from [32]. Among the 4 cancer cell lines, only K562 cell line matches the cell lines used in this study. The essential genes identified in K562 were only used to analyze identified HIMs in K562. Because majority of the essential genes are shared between the 4 cancer cell lines [32], we used the union of them as the essential gene list in GM12878, HeLa, HUVEC, and NHEK cell types. The union has 2741 essential genes.

Poly-A RNA-seq data with the GEO accession number GSE33480 were downloaded from ENCODE project [68]. The gene expression level quantified by FPKM value across the 5 cell types are normalized by quantile normalization then logarithm transformed by the function $\log(1+x)$. 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 ratio of the gene expression value in the given cell type to the median gene expression value in the other 4 cell types is higher than 2. (2) the gene expression value in the given cell type is higher than 0.1.

A.6 Statistical test

All statistical tests are done by the R software [69]. Wilcoxon rank-sum test is performed by the *wilcox.test* function to test that two continuous vectors having equal means or the mean of one vector is equal to a given constant such as 0 or 1. The argument *paired* is set to *TRUE* in case that the two vectors are paired. For example, one vector is the average gene expression of the genes assigned to HIMs across the 23 chromosomes. The other vector is the average gene expression of the genes that are in the heterogeneous networks but are not assigned to HIMs across the 23 chromosomes. Hypergeometric test for enrichment analysis is done by the *phyper* function. Chi-squared test of independence is performed by the *chisq.test* function to test that two categorical variables are independent. P values smaller than 2.22e-16 are reported as $P<2.22\text{e-}16$.

B Supplementary Results

B.1 HIMs share similar connections with organization units in 3D chromatin interactome

We found that the HIMs in the 5 cell types share similar connections with well-known chromatin interactome features, such as A/B compartments, TADs, and loops. To do this, we looked at the genomic region of each HIM which is the smallest constitutive genomic block that contains the transcription start sites 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 (Supplementary Table S2), which are comparable to the size of A/B compartments whose median size is 5Mb [8, 70]. The median numbers of sub-TADs in the genomic regions of the HIMs are 3-4 in different cell types (Supplementary Table S2). The genomic regions of the HIMs have, on average, 7 loops in GM12878 and 2-4 loops in the other cell types (Supplementary Table S2). The HIMs in GM12878 involve a higher number of loops because GM12878 has at least 60% more called loops than the other cell types possibly due to higher sequencing depth [8].

B.2 HIMs are robust to the parameters used to construct the heterogeneous networks

In this section, we will show that the identified HIMs in the main text are robust to the parameters used to define the heterogeneous networks. In all the analysis presented in the main text, we use a cutoff at 1 for "observed over expected" (O/E) quantity to filter out intra-chromosomal Hi-C contacts when defining Hi-C interaction networks. To test the robustness of HIMs, we construct another set of Hi-C interaction networks by the cutoff at 2. The number of intra-chromosomal Hi-C interactions with cutoff at 2 is 64.6%-81.9% of the number of intra-chromosomal Hi-C interactions with cutoff at 1 across the five cell types. We denote the sub-Hi-C interaction networks resulted from cutoff at 2 as sub-Hi-C. Regarding GRNs, we also construct a sub-GRN for each cell type by only keeping the top 90% interactions with highest scores. Then we construct 3 different heterogeneous networks for each cell type as the following.

- The heterogeneous network combines the Hi-C interaction network with cutoff at 1 and the whole GRN. This is the heterogeneous network used in the main text. The heterogeneous network is referred as Hi-C + GRN.
- The heterogeneous network combines the Hi-C interaction network with cutoff at 2 and the whole GRN. The heterogeneous network is referred as sub-Hi-C + GRN.
- The heterogeneous network combines the Hi-C interaction network with cutoff at 1 and the sub-GRN. The heterogeneous network is referred to as Hi-C + sub-GRN.

We apply MOCHI with the 4-node motif M to each of the 3 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 assignment of genes to HIMs are identical between two sets of HIMs, then adjusted Rand index is 1. We then use hierarchical clustering to group the sets of HIMs with similar adjusted Rand index. We found that the sets of HIMs from the heterogeneous networks of the same cell type are more similar to each other than to the sets of HIMs from the other cell types. Hierarchical clustering produces five major clusters (Supplementary Fig. S9). Each cluster contains exact 3 nodes corresponding to the 3 different heterogeneous networks in the same cell type. Overall, the result suggests that the HIMs are not sensitive to the parameters for constructing the chromatin interactome and GRNs thus the input heterogeneous network.

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

To justify the choice of the motif M, we compared it with two different motifs. One motif is a triangle motif (Supplement Fig. S10A). The triangle motif has 3 nodes, two of them are genes with a Hi-C interaction. The third node is a TF and it regulates both the genes. Thus the triangle motif does not explicitly encode co-regulations between TF proteins. Another motif is a bifan motif with 4 nodes (Supplement Fig. S10A). Two nodes are TF proteins that co-regulate two genes but there is no Hi-C interactions between the two genes. Thus 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. Each identified HIM is split

based on chromosome number such that only the genes on the same chromosome are in the same HIM.

We found that the motif M and triangle motif are better than the bifan motif (Supplementary Fig. S10B). In detail, compared to the HIMs identified by the bifan motif, the HIMs by the motif M and triangle have higher Hi-C edge density, higher triangle density, higher motif M density. Moreover, the genes in a HIM are more close to each other in the 1D sequence space, although the HIMs by the bifan motif have smaller number of genes as compared to the HIMs by the motifs M and triangle. Thus the results highlight that the Hi-C interaction between the two target genes in a motif is vital to capture spatial proximities between the genes in a HIM.

We found that the motif M is better than the triangle motif. To do this, we comprehensively compared the identified HIMs by the motif M and the triangle motif. The HIMs identified by the two motifs have similar number of genes as the median numbers of genes are equal in the 4 out of 5 cell types. Similar pattern is observed on the Hi-C edge density. Specifically, the density is only significantly (P$\leq$ 0.04) different in two cell types: GM12878 and NHEK. Even though statistically significant, the difference is small because the difference in the median density is 0.015 in GM12878 and 0.028 in NHEK (Supplementary Fig. S10B). The identified HIMs by the two motifs are significantly different in other features. In detail, compared to the HIMs identified by the triangle motif, the HIMs identified by the 4-node motif M have significantly (P$\leq$ 2.45e-21) higher number of TFs. The difference in the median number of TFs ranges from 4-8 (Supplementary Fig. S11). Even though the HIMs by the 4-node motif M have higher number TFs and comparable number of genes as compared to the HIMs by the triangle. They have significantly (P$\leq$2.68e-02) higher triangle density and significantly (P$\leq$1.43e-03) higher 4-node motif M density (Supplementary Fig. S10B). They have significantly (P$\leq$2.18e-02) higher proportion of genes in the A compartments in 4 cell types, are significantly (P$\leq$6.74e-03) earlier replicated, and have significantly (P$\leq$4.38e-05) smaller replication timing CV (Supplementary Fig.s S11).

Next, we compared the features after adjusting the number of TFs and the number of genes of the identified HIMs. Because the HIMs by the 4-node motif have significantly higher number of TFs than the HIMs by the triangle motif. The number of genes is slightly different in some cell types. The features could be biased to the number of TFs and the number of genes. Thus, to make a more fair comparison, we compared the features by adjusting the number of genes and the number of TFs in HIMs by a linear regression model (Eq. (4)).

$$Y = \beta_0 + \beta_1 \times \text{\# TFs} + \beta_2 \times \text{\# genes} + \beta_3 \times \mathbb{1}_{\text{motif}}. \tag{4}$$

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. $\hat{\beta}_3$ 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 with the 4-node motif is higher in the feature Y than the HIMs with the triangle motif. On the other hand, a negative $\hat{\beta}_3$ means lower Y in HIMs with the 4-node motif M. The detailed $\hat{\beta}_3$ for the features are reported in the Supplementary Table S4. Overall, the differences are still significant after adjusting the number of TFs and the number of genes. Which suggests that the advantages of the 4-node motif M to the triangle motif are not artifacts. Take together, the 4-node motif M is better than the triangle motif in identifying HIMs.

C Supplementary Tables

Table S1: Summary of the input heterogeneous networks and the identified HIMs across five cell types. Overlapping HIMs (%) is the proportion of identified HIMs that share TFs with other HIMs. Genes in HIMs (%) is 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$\rightarrow$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 furthest apart genes (Mb)	8	5.4	6.1	7.2	4.9
Number 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: The dynamics of Hi-C interaction networks and GRNs across the 5 cell types. Cells in the table are 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 the 5 different cell types. Overall, a large proportion of the edges in the GRNs and Hi-C 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

Table S4: Comparing the identified HIMs by 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 $Y = \beta_0 + \beta_1 \times$ # TFs $+ \beta_2 \times$ # 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 the 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	5.86e-01	0.012	1.55e-01	0.02	1.51e-02	-0.001	9.03e-01	0.026	3.02e-03
Triangle density	0.11	9.9e-17	0.071	3.19e-10	0.074	6.04e-11	0.056	4.45e-07	0.077	1.43e-11
4-node motif M density	0.15	1.93e-22	0.082	1.83e-11	0.089	6.25e-13	0.074	1.35e-09	0.09	6.91e-13
% of genes in A compartments	0.038	3.61e-04	0.036	8.4e-03	0.022	1.47e-01	0.033	7.81e-04	0.015	2.89e-01
Mean replication timing	2.76	8.87e-07	2.269	1.36e-08	2.485	2.59e-06	2.527	2.13e-07	2.673	1e-10
Replication timing cv	-0.04	3.27e-13	-0.029	6.88e-09	-0.032	8.82e-10	-0.029	2.03e-09	-0.029	9.36e-12

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

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

D Supplementary Figures

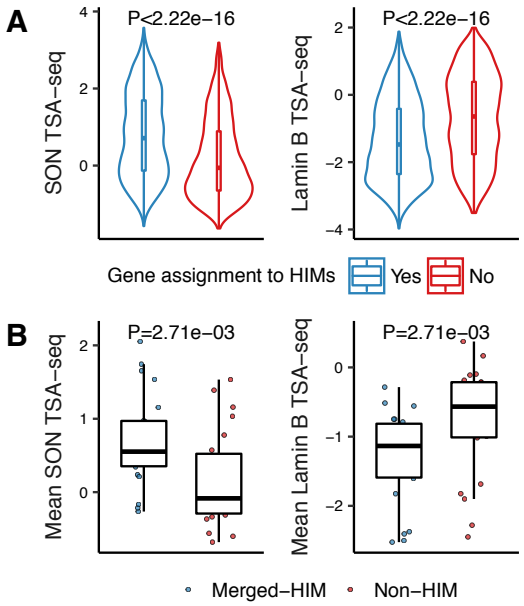

Figure S1: Genes assigned to HIMs are closer to nuclear speckles as compared to the genes in the heterogeneous network but not assigned to HIMs in K562 cell line. **(A)** Violin plots show the distributions of TSA-seq scores of the two sets of genes. **(B)** Boxplots show the distributions of mean TSA-seq scores of the merged-HIM clusters and non-HIM clusters. Here we merged the genes assigned to HIMs on the same chromosome into one cluster and called it a merged-HIM cluster. Similarly, we merged the genes not assigned to HIMs on the same chromosome into one cluster and called it a non-HIM cluster. As a result, there are one merged-HIM and one non-HIM on each chromosome.

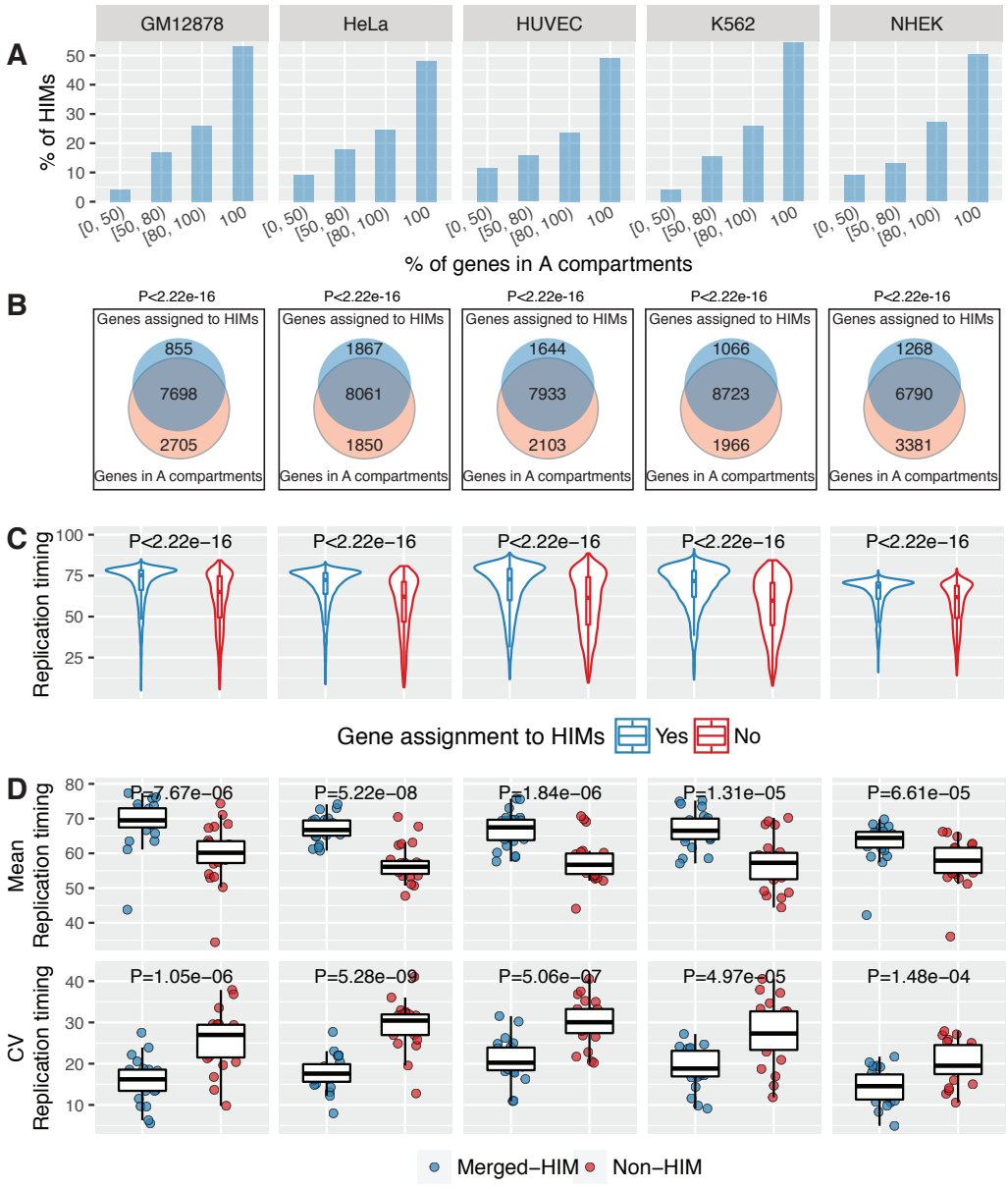

Figure S2: HIMs consistently have spatial location preferences as compared to non-HIMs in the 5 different cell types. Rows correspond to the spatial location features. Columns correspond to the cell types. **(A)** Barplot shows the distribution of HIMs with a varied proportion of genes that are in A compartments. **(B)** Venn diagram shows that the genes assigned to HIMs, as a whole, are enriched in the genes in A compartments. **(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 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 be replicated in the same phase.

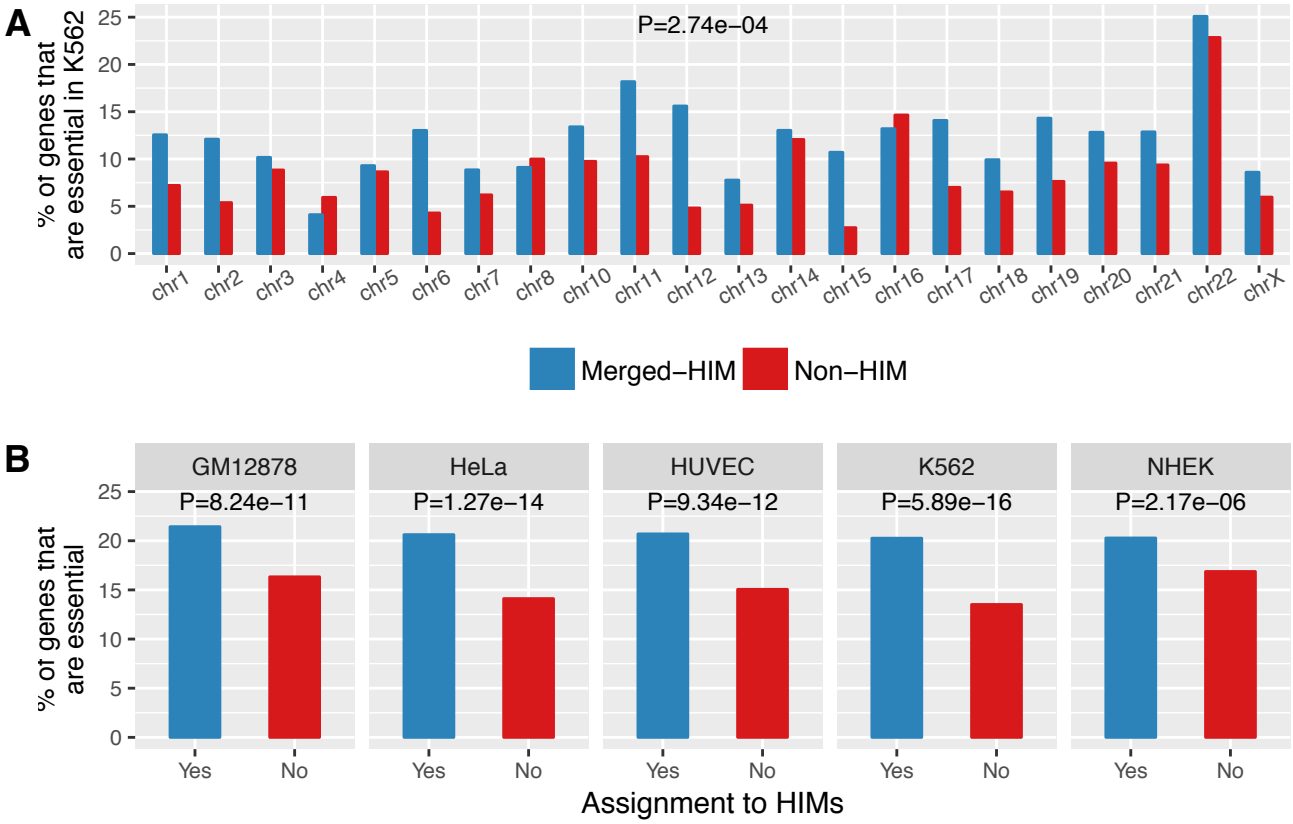

Figure S3: HIMs enrich with essential genes across the cell types. **(A)** Barplots show the proportions of genes that are K562 essential genes in the genes assigned to HIMs and not to HIMs. **(B)** Barplots show the proportions of K562 essential genes in merged-HIMs and non-HIMs across the chromosomes. The P value is computed by the paired two-sample Wilcoxon rank-sum test. **(C)** Barplots show the proportions of essential genes in the genes assigned to HIMs and not to HIMs. The essential gene list is created by merging the identified essential genes in the 4 cancer cell lines in Wang et al. [32]. P value is computed by the Chi-squared test of independence using the R *chisq.test* function. The proportions of the essential genes in merged-HIMs and non-HIMs are similar to **(A)** thus are not shown here.

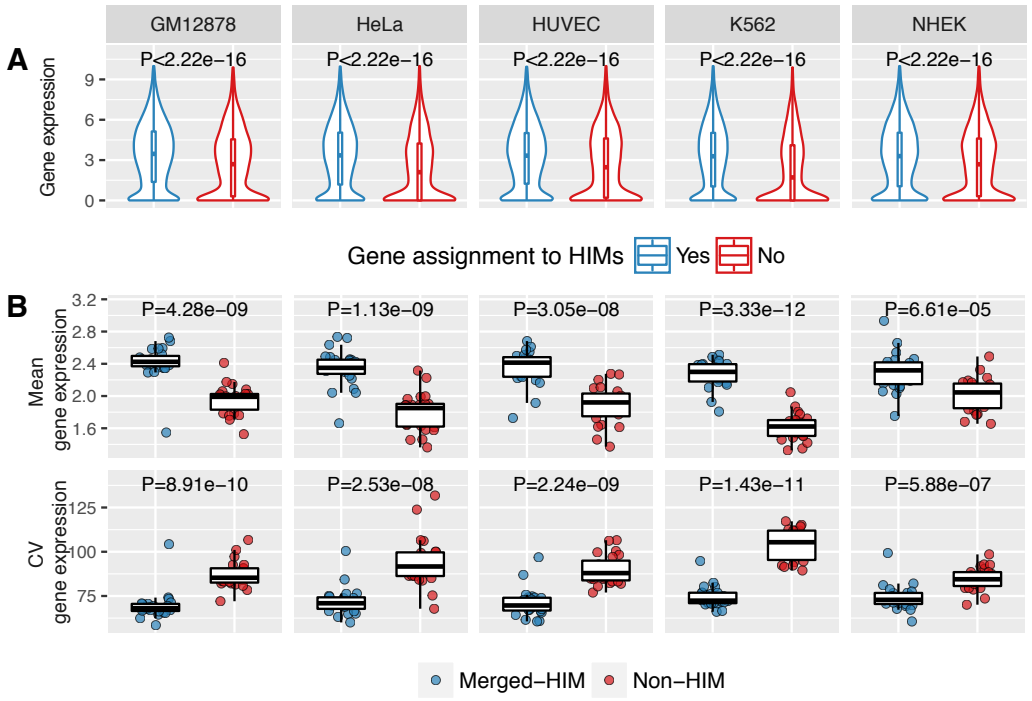

Figure S4: The genes assigned to HIMs express at higher levels than the genes not assigned to HIMs across the cell types. **(A)** Violin plots show that the genes assigned to HIMs, as a whole, are significantly higher expressed than the genes not assigned to HIMs. **(B)** Boxplots show that the merged-HIMs have significantly higher mean and lower CV of expression level than the non-HIMs.

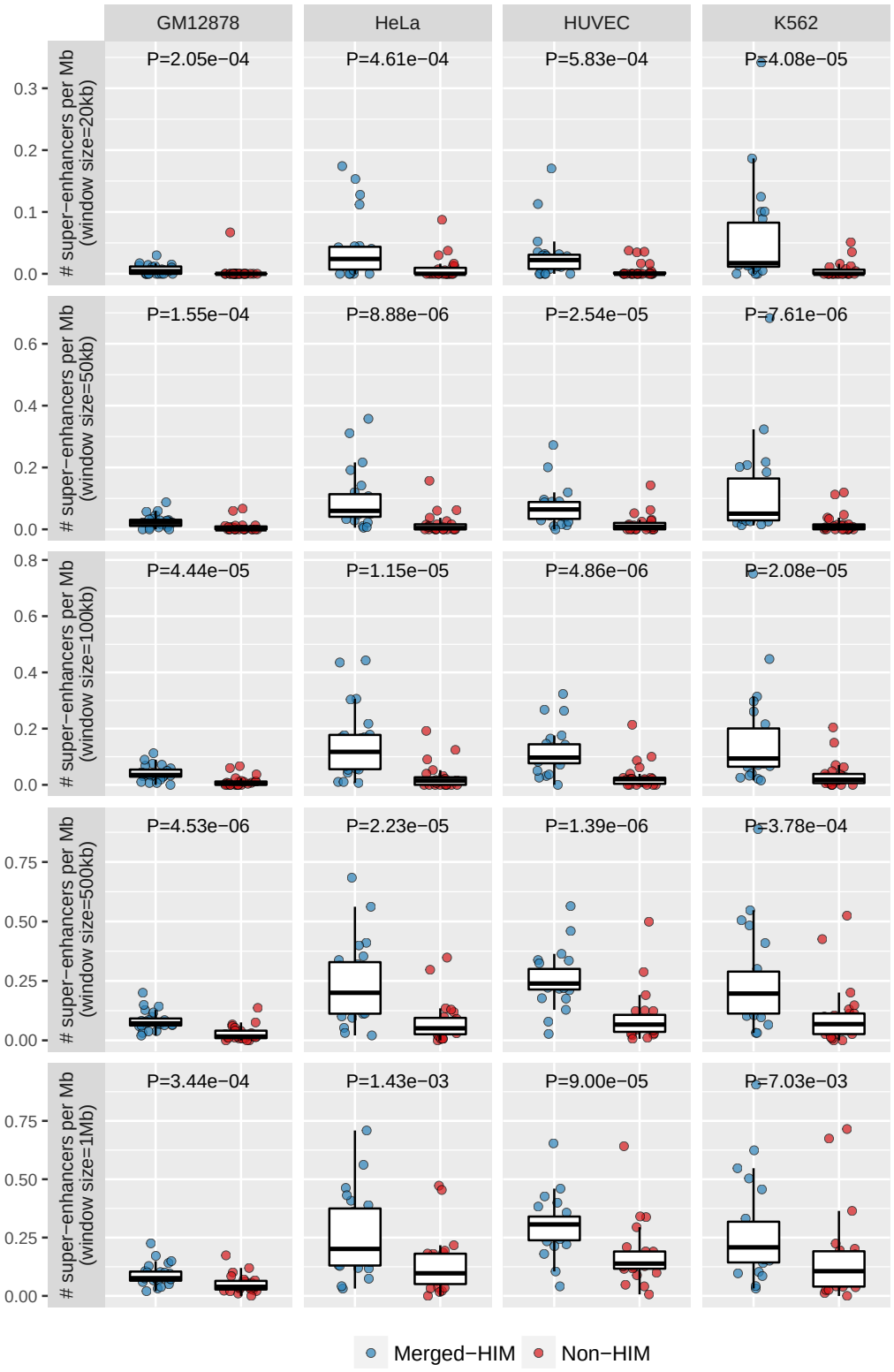

Figure S5: HIMs are enriched with super-enhancers, and the enrichment is robust to the window size. The window size is used to define the genes that are close to a given super-enhancer. The window size ranges from 20kb to 1Mb. The distribution of super-enhancers in NHEK is missing due to lack of super-enhancer data in the cell type.

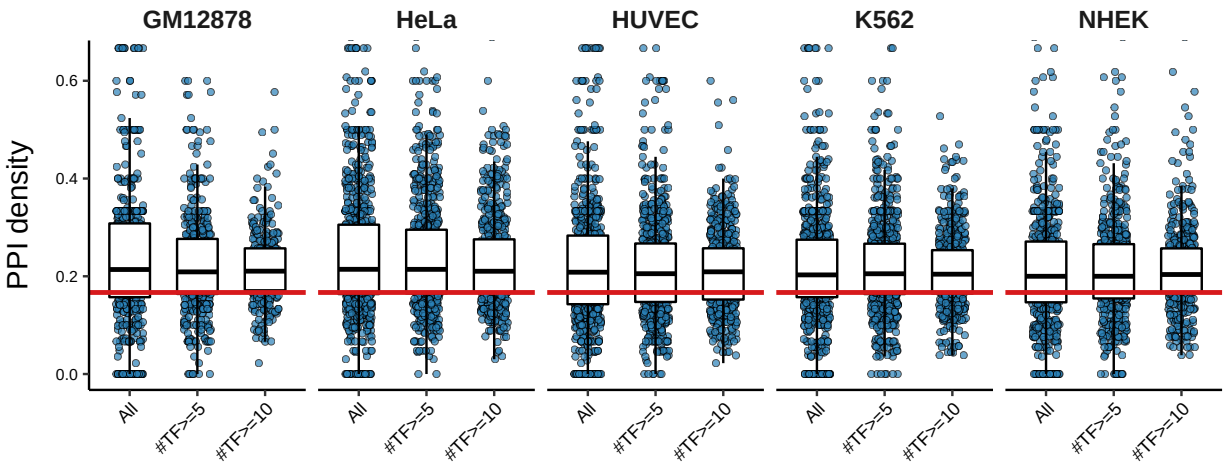

Figure S6: TFs in HIMs are enriched with protein-protein interactions (PPIs). Boxplots show the distribution of the sub-PPI network density across HIMs and subsets of HIMs with at least n TFs, $n = 5, 10$. Here for each HIM, we computed the density of the sub-PPI network induced by the TFs in the HIM from the whole PPI network between TFs. The red line is the density (0.158) of the whole PPI network. The medians of the boxplots are all significantly (P$\leq$4.56e-29) higher than the density of the global PPI network.

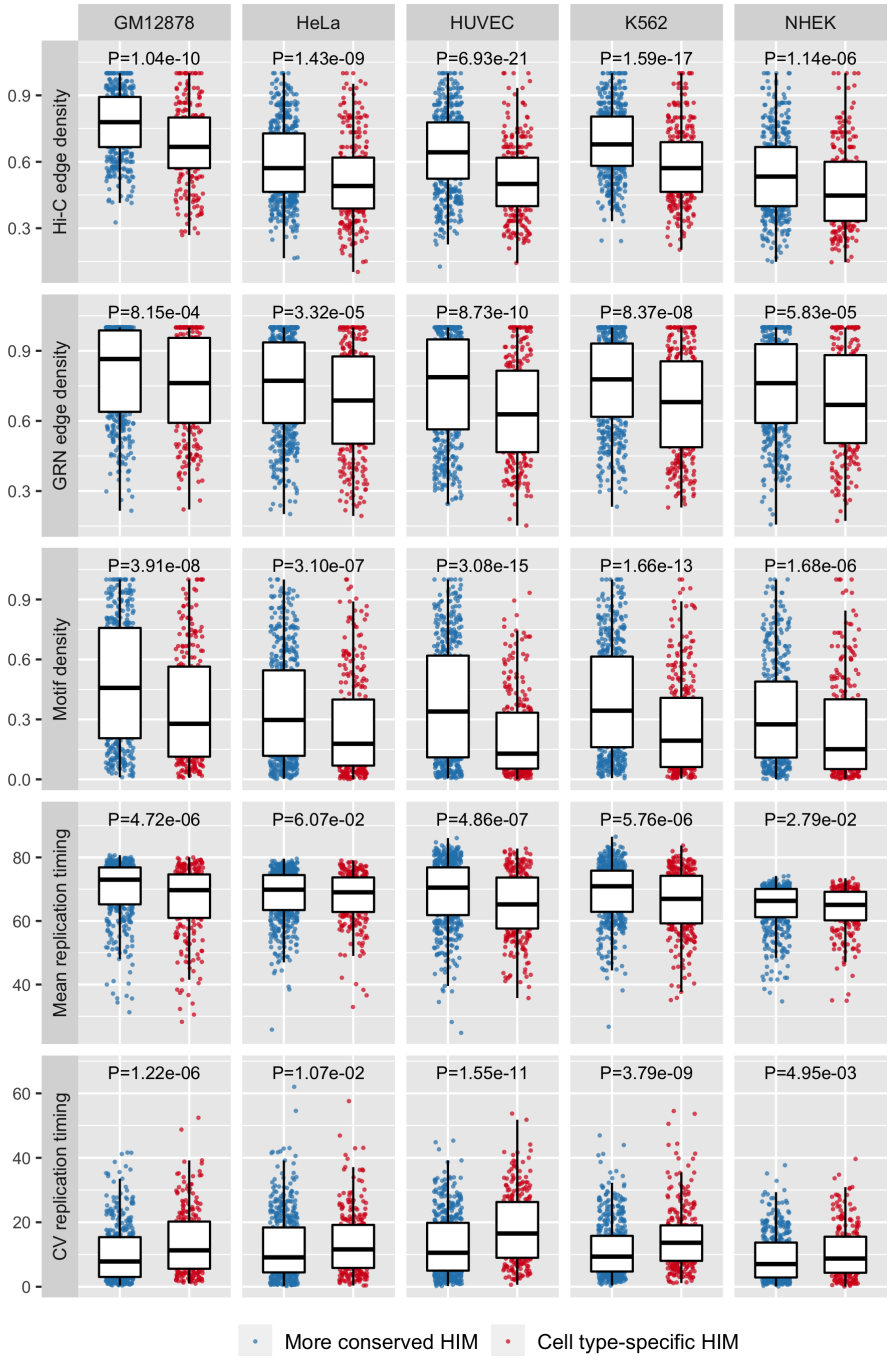

Figure S7: More conserved and cell type-specific HIMs have distinct spatial location features.

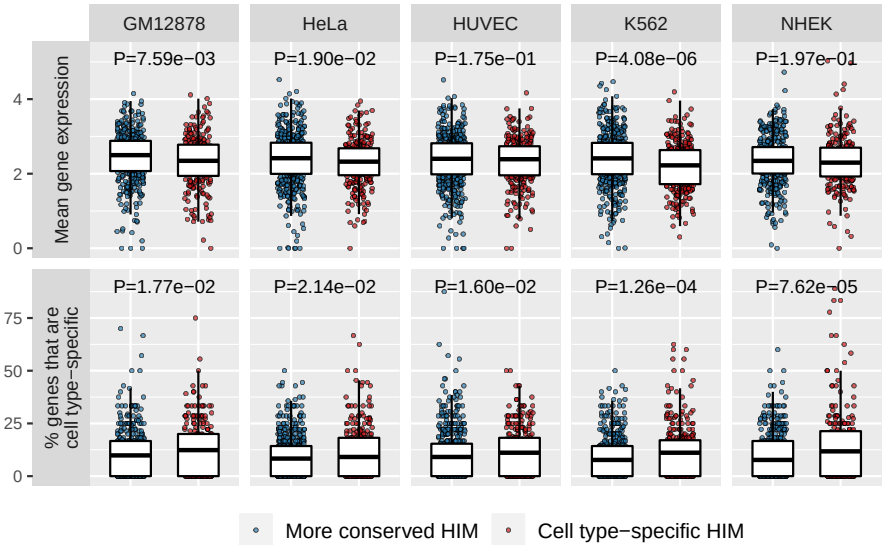

Figure S8: Functional differences and similarities between more conserved and cell type-specific HIMs. More conserved HIMs have significantly higher average gene expression in 3 cell types (first row). On the other hand, cell type-specific HIMs tend to have a higher proportion of cell type-specific genes in all the 5 cell types (second row).

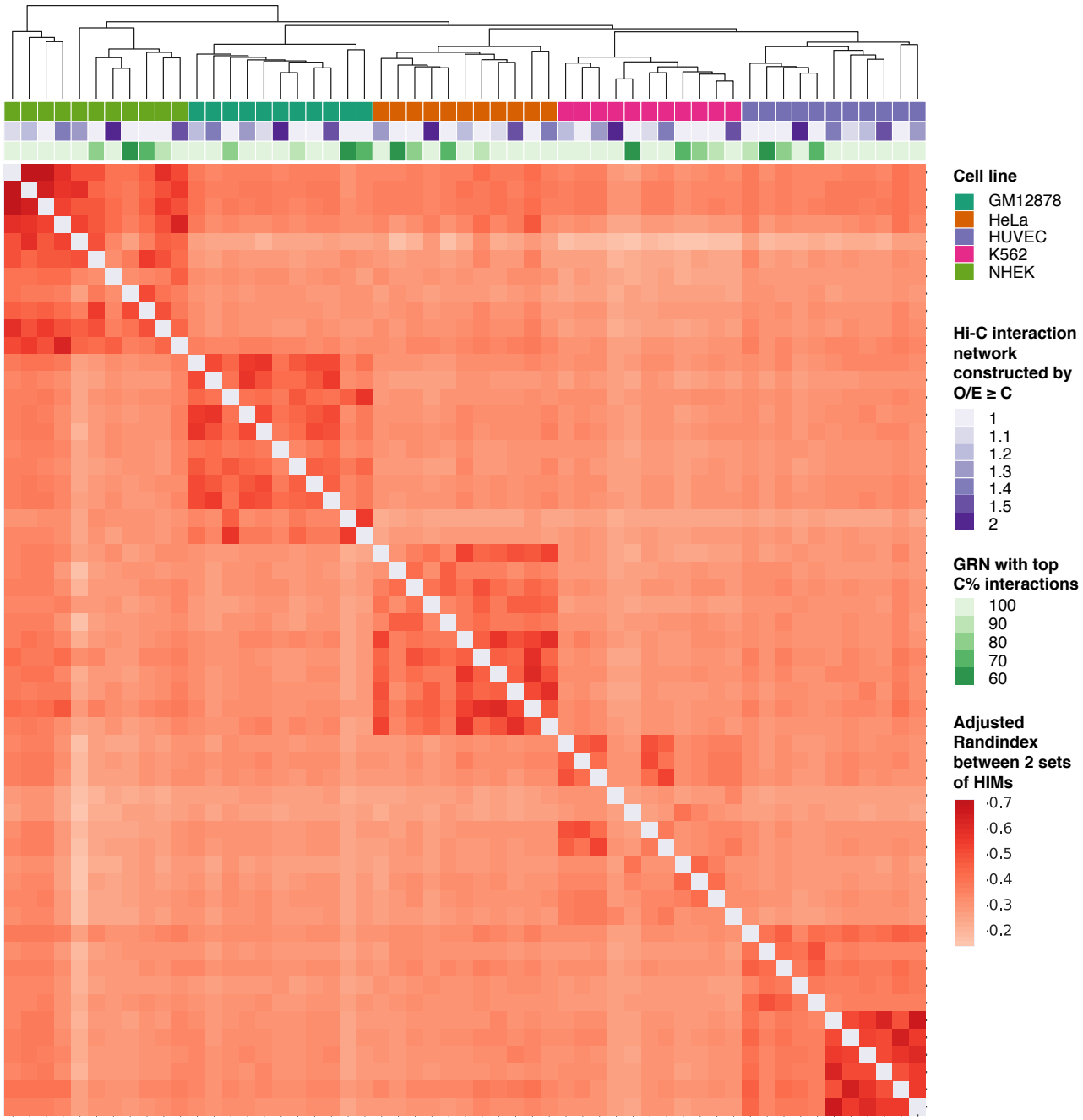

Figure S9: HIMs are robust to the input heterogeneous networks. For each cell type, there are 3 heterogeneous networks: one used in the main text and two sub-heterogeneous networks derived from different combinations of Hi-C interaction networks and GRNs. A sub-Hi-C represents a sub-Hi-C interaction network with interactions satisfying O/E≥ 2. The sub-GRN represents a sub-GRN with interactions having top 90$\%$ scores. The heterogeneous network (Hi-C + GRN) is used in the main text. The other two (sub-Hi-C + GRN, Hi-C + sub-GRN) are sub-heterogeneous networks. Adjusted Rand index is used to quantify similarities on gene memberships between two different sets of HIMs that are resulted from two different heterogeneous networks. Then hierarchical clustering on adjusted Rand index is used to cluster the sets of HIMs. Overall, the result shows that the sets of HIMs from the same cell type are in the same cluster thus are more similar as compared to the sets of HIMs from different cell types.

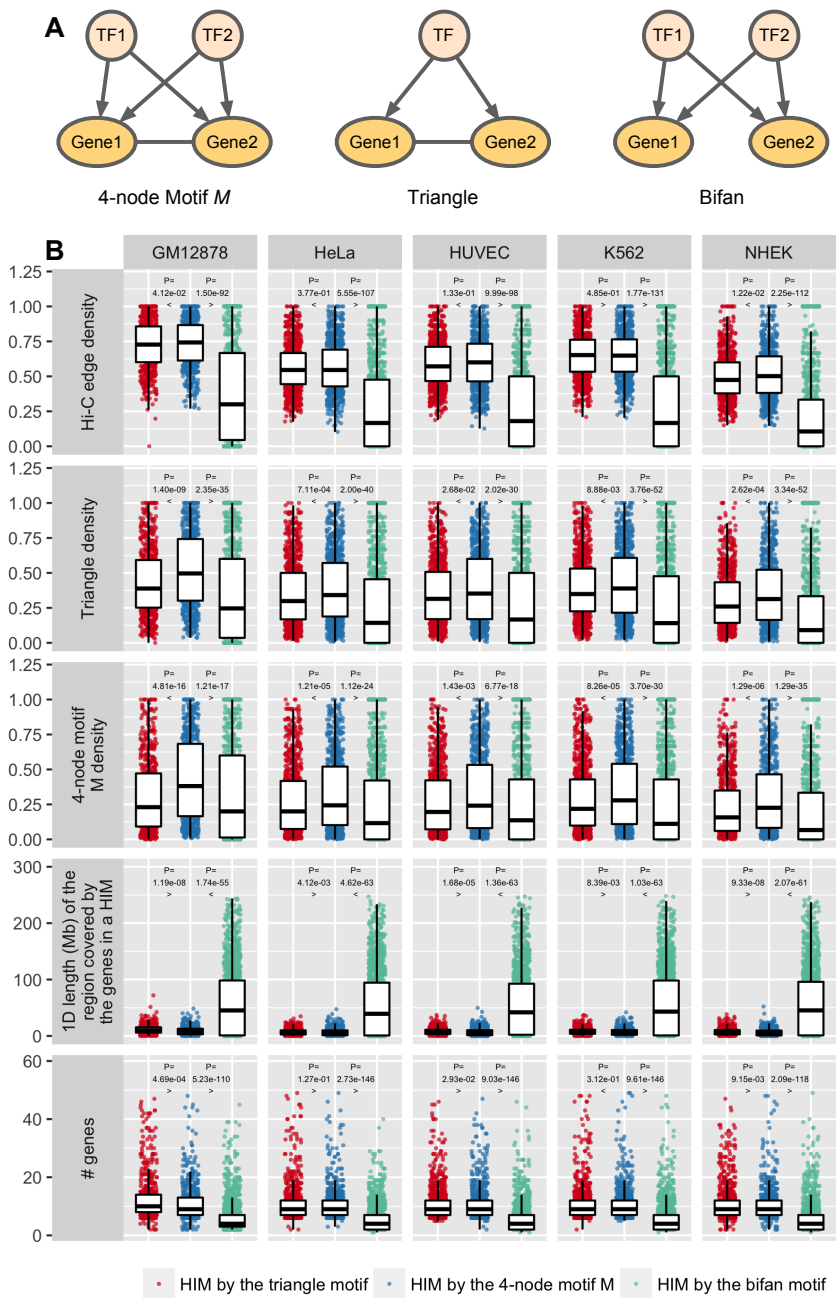

Figure S10: Comparison of the identified HIMs between the 4-node motif M, a triangle motif, bifan motif across the 5 cell types. **(A)** Demonstration of the 3 different motifs. **(B)** Comparison between the HIMs identified by the 3 different motifs. Rows correspond to the features. Columns correspond to the cell types. A dot represents a HIM. Overall, the 4-node motif M and triangle are better than the bifan motif.

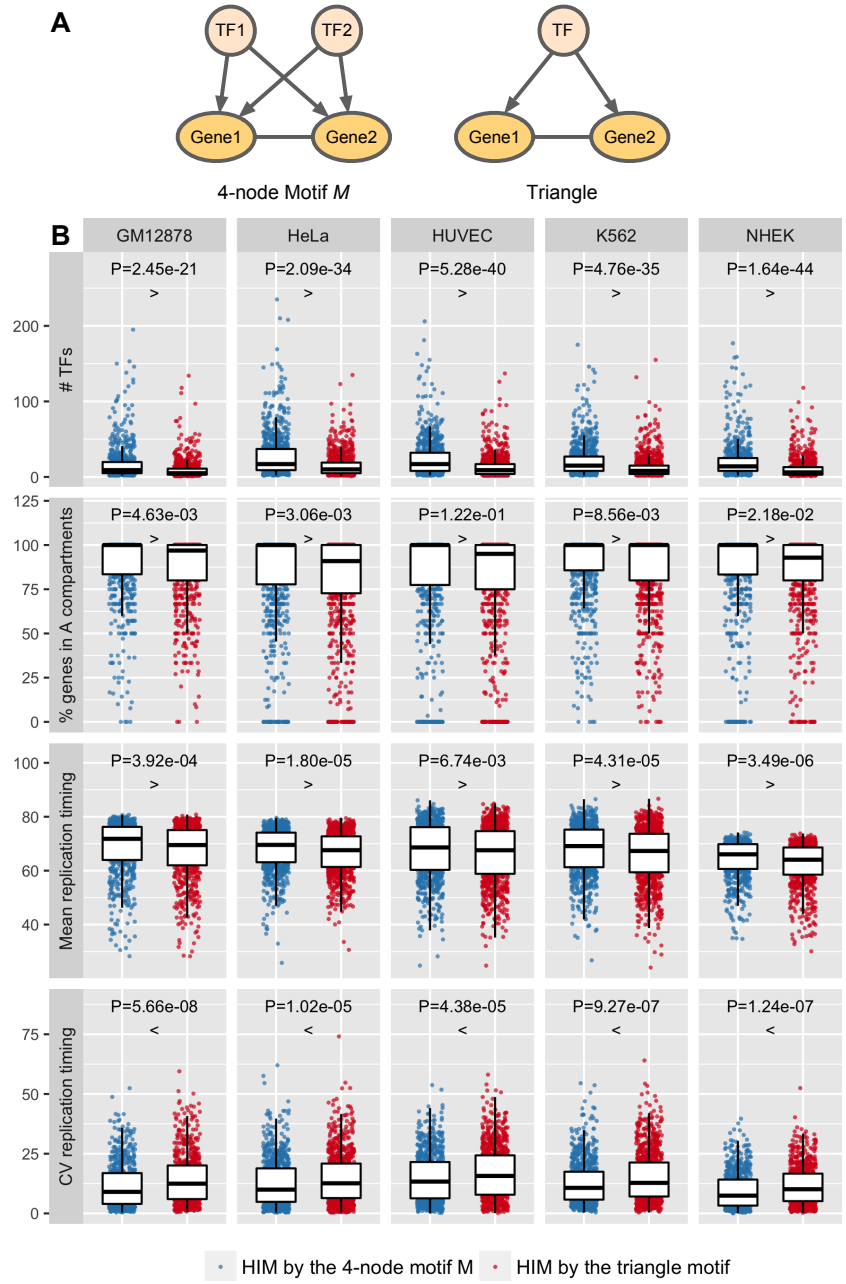

Figure S11: Comparison of identified HIMs between the 4-node motif M and the triangle motif across the 5 cell types. Where the triangle motif has 3 nodes **(A)**. Two of them are genes connected by a Hi-C interaction and co-regulated by a TF. **(B)** Rows correspond to the features. Columns correspond to the cell types. A dot represents a HIM identified either by the 4-node motif M or the triangle motif. Overall, the 4-node motif M is better than the triangle motif in the majority features. The patterns are consistently observed after adjusting the numbers of TFs and genes in HIMs by a linear regression model (Supplementary Table S4).