# MOCHI enables discovery of heterogeneous interactome modules in 3D nucleome

Dechao Tian[1,#], Ruochi Zhang[1,#], Yang Zhang[1], Xiaopeng Zhu[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
[*]Correspondence: jianma@cs.cmu.edu

## Contact

To whom correspondence should be addressed:

Jian Ma
School of Computer Science
Carnegie Mellon University
7705 Gates-Hillman Complex
5000 Forbes Avenue
Pittsburgh, PA 15213

Phone: +1 (412) 268-2776

Email: 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 interwoven heterogeneous interactomes remain poorly understood. Here we focus on two different interactomes, chromatin interaction network and gene regulatory network, as a proof-of-principle, to identify heterogeneous interactome modules (HIMs) in the nucleus. 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 develop a new algorithm MOCHI 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 properties. Through integrative analysis, this work demonstrates the utility of MOCHI to identify HIMs, which may provide new perspectives on 3D genome organization and function.


## Introduction

The cell nucleus is an organelle that contains heterogeneous components such as chromosomes, proteins, RNAs, and subnuclear compartments. These different constituents form complex organizations that are spatially and temporally dynamic [1, 2]. Interphase chromosomes are folded and organized in three-dimensional (3D) space by compartmentalizing the cell nucleus [3, 4], and different chromosomal loci also interact with each other [2]. The development in whole-genome mapping approaches such as Hi-C [5] to probing chromatin interactome has enabled comprehensive identification of genome-wide chromatin interactions, revealing important nuclear genome features such as loops [6, 7], topologically associating domains (TADs) [8, 9], and A/B compartments [5]. Nuclear genome organization has intricate connections with gene regulation [3, 10]. In particular, correlations between higher-order genome organization (including chromatin interactions and chromosome compartmentalization) and transcriptional activity have been demonstrated [6, 11, 12].

Systems level transcriptional machinery can often be represented by gene regulatory networks (GRNs), which are dynamic among various cellular conditions [13, 14]. GRN models the phenomena of selective binding of transcription factor (TF) proteins to *cis*-regulatory elements in the genome to regulate target genes [15, 16]. Transcription of co-regulated genes in GRN can be facilitated by long-range chromosomal interactions [17] and chromatin interactome has been shown to exhibit strong correlations with GRN [18, 19]. Indeed, network-based representation of both chromatin interactome and GRN has been suggested to consider different subnuclear components holistically [20, 21]. The paradigm of viewing the nucleus as a collection of interacting networks among various constituents can potentially be extended to account for other types of related interactomes in the nucleus. However, whether these interactomes, in particular chromatin interactome and GRN, are organized to form functionally relevant, global patterns remains unknown.

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 analyzed the structure and dynamics of chromatin interactomes and

GRNs as well as the coordinated binding of transcription factors on folded chromatin [6, 7, 14, 22, 23]. However, the global network level patterns between chromatin interactome and GRN are still unclear, and algorithms that can simultaneously analyze these heterogeneous networks in the nucleus to discover important network structures have not been developed.

Here we aim to identify mesoscale network structures where nodes of TFs (from GRN) and gene loci (from both chromatin interactome and GRN) cooperatively form distinct types of modules (i.e., clusters). We develop a new algorithm, MOCHI (MOtif Clustering in Heterogeneous Interactomes), that can effectively uncover such network modules, which we call heterogeneous interactome modules (HIMs), based on network motif clustering using a 4-node motif specifically designed to reveal HIMs. Each identified HIM represents a collection of gene loci and TFs for which (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. To demonstrate the utility of MOCHI to identify HIMs based on complex heterogeneous interactomes in the nucleus, we apply MOCHI to five different human cell types, identifying patterns of HIMs and their functional properties through integrative analysis. HIMs have the potential to provide new insights into the nucleome structure and function, in particular, the interwoven interactome patterns from different components of the nucleus. The source code of our MOCHI method can be accessed at: https://github.com/ma-compbio/MOCHI.

# Results

## Overview of the MOCHI algorithm

The overview of our method is illustrated in Fig. 1, with detailed algorithms described in the Methods section. Our goal is to reveal network clusters in a heterogeneous network such that certain higher-order network structures (e.g., the network motif $M$ in Fig. 1A) are frequently contained within the same cluster. The input heterogeneous network in this work considers two types of interactomes: 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 Supplemental Information B.1) to define the edges in the chromatin interaction network. For GRN, we use the transcriptional regulatory networks from [14], which were constructed by combining enrichment of TF binding sites in enhancer and promoter regions and co-expression 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 network $G$ with nodes that are either TF proteins or gene loci together with the directed and undirected edges defined above (Fig. 1B).

We specifically consider the network motif $M$ which has 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) (see the Methods section and Supplemental Information for the justification of this motif). Our goal is to reveal higher-order network clusters based on this particular 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, which extends the original algorithm in [24],

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 HIMs, which, in this work, represent network structures containing the same group of TFs that regulate many target genes whose spatial contact frequencies are higher than expected. Since TFs can regulate multiple sets of genes that may belong to different clusters, different HIMs may overlap by sharing TFs. The algorithm details of MOCHI are in the Methods section.

## 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 expressed genes, and $\sim$1 million regulatory interactions (Table S1). A few examples of 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 share several basic characteristics. The number of identified HIMs ranges from 650 to 806 in different cell types, with at least 71.9% of the HIMs sharing TFs with other HIMs in each cell type. Notably, HIMs cover a 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, 9-17 TFs regulating 9 genes in different cell types (Table S2). In addition, we found that the identified HIMs in different cell types share similar connections to 3D genome features (Supplemental Information B.3, Table S2).

To further assess that the genes in a HIM are indeed co-regulated by the same TF, we used the available ChIP-seq data of 26 TFs in GM12878 and K562 cells from the ENCODE project [25]. We found that for all the HIMs in GM12878 or K562 with these 26 TFs, more than half (55.85%) of them have $\geq 50\%$ of their genes with corresponding TF ChIP-seq peaks within 10kb of the transcription start site, further suggesting that the genes in HIMs identified by MOCHI share regulatory TFs. In addition, MOCHI can reliably identify HIMs with different parameters in various cell types (Supplemental Information B.1). Importantly, we justified the choice of the 4-node motif $M$ by showing its advantages over a triangle motif and a bifan motif (Supplemental Information B.2). The triangle and bifan motifs do not explicitly encode the co-regulation between TFs and the spatial proximity between genes. These results demonstrate that MOCHI can reliably identify HIMs across multiple cell types.

## HIMs show strong preference in spatial location relative to subnuclear structures

Next, we specifically analyzed the spatial localization of HIM in the nucleus. Recently published SON TSA-seq and Lamin B TSA-seq datasets quantify cytological distance of chromosome regions to nuclear speckles and nuclear lamina, respectively [12]. In K562, which is currently the only cell type with TSA-seq data, 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$\mu$m (estimated in [12]) of nuclear speckles (Fig. 2A). Compared to the genes in the K562 heterogeneous network but not assigned to HIMs, the genes in HIMs have higher SON TSA-seq score and lower Lamin B TSA-seq score ($p$<2.22e-16; Fig. S1).

We specifically looked at the HIMs that are away from the nuclear interior. Fig. 2B shows one HIM (#541) that is close to nuclear lamina (mean Lamin B TSA-seq score 0.593, mean SON TSA-seq score -0.642). This HIM has 9 TFs co-regulating 6 genes that span 6.78 Mb on chromosome 3. The Hi-C edge density (see Supplemental Information A.4) among these genes is 0.667, suggesting that these 6 genes

as a group are spatially closer to each other than expected (i.e., connected with chromatin interaction). 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), suggesting that these gene loci are localized more towards the nuclear interior than their surrounding chromatin. Five out of the 6 genes are expressed with FPKM>=3.4. The gene RPL15 in this HIM is a K562 essential gene [26]. The TF proteins CDX1, HOXA9, and HOXA10 are involved in leukemia and hematopoietic lineage commitment provided by Genecards [27]. This suggests that even though HIM #541 is a HIM away from nuclear speckle, it may play relevant functional roles in K562.

Recently, Quinodoz et al. [28] reported that inter-chromosomal interactions are clustered around two distinct nuclear bodies as hubs, including nuclear speckles and nucleoli. By comparing with the genomic regions organized around nucleolus based on data from the SPRITE method in GM12878 [28], we found that vast majority (85.4%) of the GM12878 HIMs do not have genes close to the nucleolus. Earlier work estimated that only 4% of the human genome is within nucleolus-associated domains [29]. It is therefore expected that only a small number of HIMs would be close to the nucleolus. Indeed, we found that there are only 30 (4.62%) GM12878 HIMs with all their genes near the nucleoli. Notably, 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 [30]. For example, HIM #267 has 4 TF regulators: ETS1, ETV6, PPARG, and PTEN, where ETV6 is known to localize to the nucleoli.

Earlier work from Hi-C data showed that at megabase resolution the interphase chromosomes are segregated into A and B compartments that are largely active and inactive in transcription, respectively [5]. Chromosome regions in B and A compartments have nearly identical agreements with lamina associated domains (LADs) and inter-LADs (i.e., more towards interior) [4]. Compartment A regions also replicate earlier than compartment B regions [31]. We found that the genes in HIMs are preferentially in A compartments and replicated earlier across cell types. Specifically, 57.4% of HIMs have genes that are all in A compartments in K562. 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 more enriched in A compartments, with 89.1% of them in A compartments ($p$<2.22e-16, hypergeometric test; Fig. 2D). Compartment A can be further subdivided into A1 and A2 subcompartments in GM12878 [6] at a finer scale. Among the 369 GM12878 HIMs with genes all in A compartments, 198 (53.66%) HIMs have $\geq$80% of their genes in A1 subcompartments, 60 (16.26%) HIMs are in A2 subcompartments, and the rest 111 HIMs span both A1 and A2 compartments. Additionally, we found that the genes assigned to HIMs have much earlier replication timing than the other genes ($p$<2.22e-16; Fig. 2E). We also observed that the genes on the same chromosome that are in HIMs tend to have more similar replication timing as compared to the genes (on the same chromosome) that are not in HIMs (Fig. S2). These patterns can also be observed in other cell types (Fig. S2).

Taken together, these results suggest that HIMs have strong preference to localize towards the nuclear interior in active compartments with the majority of them being in proximity of the nuclear speckles and replicating earlier.

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

Next, we explored the functional properties of HIMs. We again grouped 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. We first compared with the information of gene essentiality [26] (see Supplemental Information A.5). We found that genes assigned to HIMs are enriched with essential genes across all five cell types. For example, 12.7% of the genes assigned to HIMs in K562 are K562 essential genes, which is significantly higher than the proportion (7.79%) of the genes not assigned to HIMs ($p$=1.13e-12; Fig. 3A). This observation is also present across chromosomes (Fig. S3A). Across the cell types, genes assigned to HIMs consistently have higher proportions of essential genes than genes not assigned to HIMs ($p \leq$2.17e-6; Fig. S3B). Regarding gene expression level, we found that 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 [32]. To study the connections between HIMs and super-enhancers, we computed the cluster-size normalized number of super-enhancers annotated in [32] 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-enhancers than the non-HIMs across cell types (Fig. 3C, Fig. S5). The significant pattern is consistent with a varied window size from 20kb to 1Mb (Fig. S5).

Protein-protein interactions (PPIs) can further stabilize TF-DNA binding of the interacting TFs [16]. We ask whether TFs in the same HIM tend to have more PPIs with each other. We computed the density of the sub-PPI network induced by the TFs in a HIM, where the PPI network is based on 591 TF proteins used in this study (see Supplemental Information A.5). We found that TFs within HIMs are enriched with PPIs among themselves as compared to random cases selected from the 591 TFs. For example, in GM12878, TFs NR3C1 and TFEB, which are master regulators [32], co-regulate 8 genes with the other 7 TFs proteins in a HIM (Fig. 3D). This particular sub-PPI network of the 9 TFs has 14 interactions. The density of this sub-PPI network is 0.389 which is 2.46 times higher than the average density (0.158) of the random cases. Overall, the median density of the sub-PPI networks induced by TFs in the identified HIMs in GM12878 is 0.214, much higher than the random cases ($p$<2.22e-16; Fig. 3E). This observation is also consistent in other cell types in this study (Fig. S6). We also found that the significance is not affected by a varied number of TFs across HIMs (Fig. S6).

These results suggest that the genes and TFs involved in HIMs likely perform critical roles, which are manifested by the level of gene essentiality of target genes, engagement of super-enhancers, and enrichment of PPI among TFs.

## Genes in HIMs exhibit stability and variability across cell types

To study how HIMs change across different cell types, we first focused on the assignment of genes to HIMs in different cell types. Through pairwise comparison, we found that the genes assigned to HIMs have the highest degree of overlap between GM12878 and K562 as compared to the other cell types, which is consistent with the fact that both GM12878 and K562 are from human hematopoietic cells (Fig. S7A). Comparisons among all five cell types showed that 3,025 genes are consistently assigned to HIMs, accounting for 30.91% to 40.06% of 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, out of the 8,034 genes in the GM12878 HIMs, only 344 (4.28%) genes are not assigned to HIMs in other cell types (Fig. 4A).

The genes consistently and uniquely assigned to HIMs are enriched with distinct functional terms

using DAVID [33] (Table S5). The genes consistently assigned to HIMs are strongly enriched with functions related to essential cellular machinery, whereas the genes uniquely assigned to HIMs in a particular cell type are enriched with more cell type-specific functions. An example is NHEK HIM #107 (Fig. 4B). Among the 6 genes in this HIM, DSC1, DSC3, DSG1 are not assigned to HIMs in the other cell types. These 6 genes are involved in the keratinization pathway based on GeneCards [27]. We further assessed the assignment of housekeeping genes [34] and essential genes to HIMs. We found that for both sets of genes, the majority ($\geq$84%) of them are assigned to HIMs consistently or in at least 3 out of the 5 cell types (Fig. 4C), suggesting that the genes with crucial functions tend to form spatial clusters across multiple cell types.

We next analyzed the variability of HIMs in terms of spatial proximity to subnuclear compartments. We found that 15 out of the 30 HIMs close to nucleoli in GM12878 (based on the data from [28]) have mean SON TSA-seq score $\geq$ 0.284 in K562 (based on the data from [12]) (Fig. 4D), in other words, these HIMs are involved in a change of spatial position from nucleoli to speckle between GM12878 and K562. One notable example is HIM #267 in GM12878 which has the highest mean SON TSA-seq score (2.41) in K562. Interestingly, the 10 genes (in HIM #267 in GM12878) together with another 8 genes form a new HIM (#628) in K562. This GM12878 HIM #267 has four TFs: ETS1, ETV6, PPARG, and PTEN. On the other hand, the K562 HIM #628 has four different TFs: KLF4, NFKB1, STAT3, and WT1, where KLF4, STAT3, and WT1 are known to be involved in the progression of leukemia.

To compare the detailed membership changes of HIMs across cell types, we computed Jaccard indices, denoted by $\text{JI}_{TF}$ and $\text{JI}_{gene}$, of the TF members and gene members between HIMs from two different cell types, respectively. We found that the gene members undergo a moderate change from one cell type to another, whereas the TF members change at a much higher rate. $\text{JI}_{gene}$ has a median of 0.096 and it is higher than expected $\text{JI}_{gene}$ between random gene sets while controlling the set size and chromosome number (median ratio=14.12, Fig. 4E). On the other hand, $\text{JI}_{TF}$ has a median of 0.017 and it is close to expected $\text{JI}_{TF}$ between randomly selected control TF sets (median ratio=0.878, Fig. 4E). There are at least two phenomena jointly contributing to these observations. First, the Hi-C interaction networks and GRNs are highly cell type-specific, as 66% Hi-C interactions and 31.4% GRN interactions only exist in one cell type (Table S3). Second, given a HIM identified in a cell type, the motif $M$ density of the HIM (see Supplemental Information A.4) has higher fold change than the Hi-C edge density of the HIM in another cell type ($p$<2.22e-16; Fig. 4F). In other words, the co-regulation relationships of the TFs on the genes in HIMs change more often across cell types than the spatial proximity relationships between the gene loci. However, we observed that if HIMs from two different cell types share a higher number of housekeeping genes, they tend to have a higher $\text{JI}_{TF}$ (Fig. 4G). We found a similar pattern for essential genes (Fig S7B).

## Conserved and cell type-specific HIMs have distinct properties

Motivated by the gene membership dynamics of HIMs across cell types, we further classified HIMs into conserved and cell type-specific HIMs. For HIMs in a given cell type, we call a HIM conserved if it shares a significantly high proportion of genes ($\text{JI}_{gene} \geq$ 1/3, $p\leq$0.001, Bonferroni adjusted hypergeometric test) with at least one HIM in other cell types (i.e., the HIM is recurrent). Note that $\text{JI}_{gene} \geq$1/3 represents that two equal-sized gene sets share than half of their genes. The rest are called cell type-specific HIMs. As a result, 40.69-47.38% of the identified HIMs in each cell type are cell type-specific HIMs. Fig. 5 shows a cell type-specific HIM, HIM #712, in K562 and its changes in other cell types.

This HIM covers 9 genes on chromosome 11. These genes spatially contact each other at higher frequencies than expected (Fig. 5A) and are co-regulated by TF protein BCL6B and CPEB1 in K562 (Fig. 5B). In other cell types, at most 4 out of the 9 genes are assigned to HIMs (Fig. 5C). We found that this 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 compartment in other cell types (Fig. 5D). One nearby upstream region is annotated as a super-enhancer only in K562 [32] (Fig. 5E). Many sites are annotated as transcriptionally active states, such as enhancers, promoters, or transcribed states in K562, but not in other cell types based the results from ChromHMM [35] (Fig. 5F). The genes MRPL16, OSBP, and PATL1 are essential genes in K562. This example demonstrates that the K562-specific HIM has specific chromatin organization and biological functions.

Overall, we found that the conserved and cell type-specific HIMs have distinct properties of interactomes across cell types. Compared to cell type-specific HIMs, conserved HIMs exhibit stronger clustering features with higher Hi-C edge density, higher GRN edge density, and higher motif $M$ density ($p \leq$8.15e-4; Fig. S8). Also, conserved HIMs tend to be closer to the nuclear interior with a higher proportion of their genes in A compartment, and their genes replicate earlier and synchronously (Fig. S8). Moreover, we found that conserved HIMs and cell type-specific HIMs tend to have large differences in gene expression level and cell type-specific genes. The conserved HIMs have higher mean gene expression level than the cell type-specific HIMs in 3 cell types except for NHEK and HUVEC ($p<0.05$; Fig. S9). On the other hand, cell type-specific HIMs have a higher proportion of cell type-specific genes ($p \leq 0.02$; see Supplemental Information A.5) than conserved HIMs across five cell types (Fig. S9). Note that genes in both types of HIMs are significantly ($p<$2.22e-16) higher expressed than the genes in the heterogeneous networks but not assigned to HIMs. Taken together, our results demonstrate that conserved and cell type-specific HIMs, in general, have distinct network properties, spatial location preference, and functional characteristics.

## Discussion

To better understand the heterogeneous nature of different components in the nucleus, new computational models are needed to consider different types of molecular interacting networks. In this work, we developed MOCHI to specifically consider two types of different interactomes in the 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. MOCHI is able to identify network structures where nodes of TFs (from GRN) and gene loci (from both chromatin interactome and GRN) cooperatively form distinct network clusters, which we call HIMs, by utilizing a new 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 and properties. By applying MOCHI to five different human cell types, we made new observations to demonstrate the biological relevance of HIMs in 3D nucleome.

Our method has multiple methodological contributions. We further extended the motif conductance clustering method [24] to find overlapping HIMs in heterogeneous networks. Our study shows the utility of our new algorithm to identify HIMs based on complex heterogeneous molecular interactomes. In addition, our method can be further modified to identify other types of potentially interesting HIMs in heterogeneous networks by replacing the 4-node motif $M$ with relevant motifs, especially when addi-

tional types of interactomes are included. For example, in addition to considering chromatin interactions and protein-DNA interactions as we did in this work, it would be of interest to incorporate other types of relevant interactomes in the nucleus, such as the RNA-chromatin interactome [36].

How can we explain the formation of HIMs? In Fig. 6, we illustrate a possible model of HIMs within the nucleus. HIMs (light pink domains) are toward the interior with a group of interacting TFs and chromatin loci. The set of TFs in a HIM cooperatively regulate target genes, which also have higher contact frequency than expected. Note that this is conceptually consistent with recently reported co-localization of TF pairs [22]. Some of these TF clusters may be related to the localization preferences of TF proteins in nuclear compartments, such as nuclear speckles that are enriched with various transcriptional activities [12, 37]. Indeed, we found that the majority of the identified HIMs are close to nuclear speckles. The definitions of HIMs may also have intrinsic connections with the emerging findings on the mechanism of nuclear subcompartment formation, where TFs and their potential regulating genes/chromatin are trapped by localized liquid-like chambers through the phase separation [38, 39]. Evidence has been shown that phase separation can help explain the formation of super-enhancer mediated gene regulation [39, 40]. From our analysis, we found that genes assigned to HIMs are enriched with contacts with super-enhancers. The genes consistently assigned to HIMs are enriched with essential biological processes related to chromosomal organization and transcription. However, the detailed formation mechanisms for HIMs, which may involve both *cis* elements and *trans* factors, remain to be investigated. It would also be important to delineate the different roles of both different TFs and different genes in forming the HIMs, as some of them may be necessary and others may be redundant for the stability of HIMs. In addition, more experimental data are needed to further evaluate the functional significance of HIMs. For example, although we observed connections between HIMs and 3D genome organization features, the intricate functional relevance among these different higher-order nucleome units that jointly contribute to gene regulation in different cellular conditions is yet to be revealed. 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 the nucleus and their roles in 3D nucleome structure and function.

# Methods

## Brief introduction to homogeneous network clustering by motif conductance

We first review 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 $G$, $\varphi^G$, 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 [41]. Recently, the conductance metric is generalized to motif conductance [24, 42], 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 [24]. A more recent network clustering method that incorporates network higher-order structures has been developed in the setting of hypergraph clustering [43], 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 that we want to achieve in this work.

## MOCHI – 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 by extending the approach in [24]. 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 HIMs 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 node 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 an 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 show that when $\alpha = 4/3$ in Eq. 3 the clustering results would be near optimal (Supplemental Information A.3). Thus, $\alpha$ is set to $4/3$ in this paper. $\text{Vol}_M(S)$ 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 $M$, $\varphi_M^G$ as $\min_S \varphi_M(S)$. 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 more 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 more 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 [24] 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 $S_1$. 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 $\alpha_k$ is the $k$-th smallest value of $D_{G_M}^{-1/2} v$.
- $S = \underset{1 \le k \le N}{\operatorname{argmin}} \varphi_{G_M}(S_k)$, 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 [24] 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 $\ell$ HIMs: $S_1, S_2, \ldots, S_\ell$.

At next iteration $\ell$, one HIM $S_k$ is selected if the graph it forms, $G_k$, has the lowest subgraph conductance value $\varphi_M^{G_k}$ among $\varphi_M^{G_j}$ , $1 \le j \le \ell$. We set a threshold $t_1$ for $\varphi_M^{G_k}$. If $\varphi_M^{G_k} \le t_1$, $S_k$ 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 a large enough penalty value 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 ancestral HIMs up to certain generations for each HIM based on the conductance value of its ancestor $\varphi_M^{G_{anc_i}}$, where $i = \{1, 2, 3...\}$ denotes for the 'parent', 'grandparent' of the HIM. We trace along the tree until $\varphi_M^{G_{anc_i}} \le t_2$, 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 according to their contribution to the number of occurrences of the subgraph $M$ based on the 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 Supplemental Information A.1. The runtime of the algorithm is $O(t^2 c^2)$, where $t$ and $c$ ($t \ll c$) are the number of TFs and the number of gene loci in the input heterogeneous network, respectively (detailed analysis in Supplemental Information A.2).

## Summary of the algorithm

Given a heterogeneous network from chromatin interactome network and GRN, our algorithm MOCHI will 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. MOCHI has a few key differences as compared to the subgraph conductance method in [24]. 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 in [24] is rather homogeneous. Second, the algorithm in [24] will not explicitly identify 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 [43] can be applied to identify non-overlapping HIMs where a hyperedge is defined as the motif $M$. However, similar to the method in [24], it was not designed to identify overlapping clusters, i.e., the method would not be able to find multiple overlapping HIMs that we would need. Our method also has clear differences as compared to previous works on multi-layer network clustering (see review in [44]). First, the inputs are different. A multi-layer network typically 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 work has different types of nodes (TF proteins and gene loci) and also edges. Previous multi-layer network clustering methods are not directly applicable to identify HIMs. Second, the outputs are different. The majority of multi-layer 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.

## Acknowledgement

This work was supported in part by the National Institutes of Health Common Fund 4D Nucleome Program grant U54DK107965 (J.M.), National Institutes of Health grant R01HG007352 (J.M.), and National Science Foundation grant 1717205 (J.M.). The authors would like to thank Bas van Steensel and members of Jian Ma's laboratory (Ben Chidester, Tianming Zhou, Kyle Xiong, and Yang Yang) for helpful comments to improve the manuscript.

## Author Contributions

Conceptualization, J.M.; Methodology, D.T., R.Z., and J.M.; Software, D.T., R.Z.; Investigation, D.T., R.Z., Y.Z., X.Z., and J.M.; Writing – Original Draft, D.T., R.Z., and J.M.; Writing – Review & Editing, D.T. and J.M.; Funding Acquisition, J.M.

## Declaration of Interests

The authors declare no competing interests.

# References


[1] 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.

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

[3] Thomas Cremer and Christoph Cremer. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Reviews Genetics*, 2(4):292, 2001.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Elphège P Nora, Bryan R Lajoie, Edda G Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, Nynke L van Berkum, Johannes Meisig, John Sedat, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, 485(7398):381, 2012.

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

[11] Lars Guelen, Ludo Pagie, Emilie Brasset, Wouter Meuleman, Marius B Faza, Wendy Talhout, Bert H Eussen, Annelies de Klein, Lodewyk Wessels, Wouter de Laat, et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, 453(7197): 948, 2008.

[12] 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*, 217(11):4025–4048, 2018.

[13] 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.

[14] 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.

[15] Eric H Davidson. *The regulatory genome: gene regulatory networks in development and evolution*. Academic Press, San Diego, 2006.

[16] 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.

[17] Stephanie Fanucchi, Youtaro Shibayama, Shaun Burd, Marc S Weinberg, and Musa M Mhlanga. Chromosomal contact permits transcription between coregulated genes. *Cell*, 155(3):606–620, 2013.

[18] Steven T Kosak, David Scalzo, Sam V Alworth, Fusheng Li, Stephanie Palmer, Tariq Enver, James SJ Lee, and Mark Groudine. Coordinate gene regulation during hematopoiesis is related to genomic organization. *PLoS Biology*, 5(11):e309, 2007.

[19] Daniel S Neems, Arturo G Garza-Gongora, Erica D Smith, and Steven T Kosak. Topologically associated domains enriched for lineage-specific genes reveal expression-dependent nuclear topologies during myogenesis. *Proceedings of the National Academy of Sciences*, page 201521826, 2016.

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

[21] Haiming Chen, Jie Chen, Lindsey A Muir, Scott Ronquist, Walter Meixner, Mats Ljungman, Thomas Ried, Stephen Smale, and Indika Rajapakse. Functional organization of the human 4D nucleome. *Proceedings of the National Academy of Sciences*, 112(26):8002–8007, 2015.

[22] 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.

[23] Ruggero Cortini and Guillaume J Filion. Theoretical principles of transcription factor traffic on folded chromatin. *Nature Communications*, 9, 2018.

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

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

[26] 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.

[27] 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.

[28] 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.

[29] 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 nucleolus. *PLoS Genetics*, 6(3):e1000889, 2010.

[30] 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.

[31] Benjamin D Pope, Tyrone Ryba, Vishnu Dileep, Feng Yue, Weisheng Wu, Olgert Denas, Daniel L Vera, Yanli Wang, R Scott Hansen, Theresa K Canfield, et al. Topologically associating domains are stable units of replication-timing regulation. *Nature*, 515(7527):402, 2014.

[32] 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.

[33] 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.

[34] Eli Eisenberg and Erez Y Levanon. Human housekeeping genes, revisited. *Trends in Genetics*, 29 (10):569–574, 2013.

[35] Jason Ernst and Manolis Kellis. Chromhmm: automating chromatin-state discovery and characterization. *Nature Methods*, 9(3):215, 2012.

[36] Tri C Nguyen, Kathia Zaleta-Rivera, Xuerui Huang, Xiaofeng Dai, and Sheng Zhong. Rna, action through interactions. *Trends in Genetics*, 2018.

[37] David L Spector and Angus I Lamond. Nuclear speckles. *Cold Spring Harbor Perspectives in Biology*, 3(2):a000646, 2011.

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

[39] 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.

[40] Ann Boija, Isaac A Klein, Benjamin R Sabari, Alessandra DallAgnese, Eliot L Coffey, Alicia V Zamudio, Charles H Li, Krishna Shrinivas, John C Manteiga, Nancy M Hannett, et al. Transcription factors activate genes through the phase-separation capacity of their activation domains. *Cell*, 175 (7):1842–1855, 2018.

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

[42] 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.

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

[44] 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.

[45] Jonathan A Kelner, Lorenzo Orecchia, Aaron Sidford, and Zeyuan Allen Zhu. A simple, combinatorial algorithm for solving sdd systems in nearly-linear time. In *Proceedings of the forty-fifth*

*annual ACM symposium on Theory of computing*, pages 911–920. ACM, 2013.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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 complexes-2009. *Nucleic Acids Research*, 38(suppl_1): D497–D501, 2009.

[52] 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.

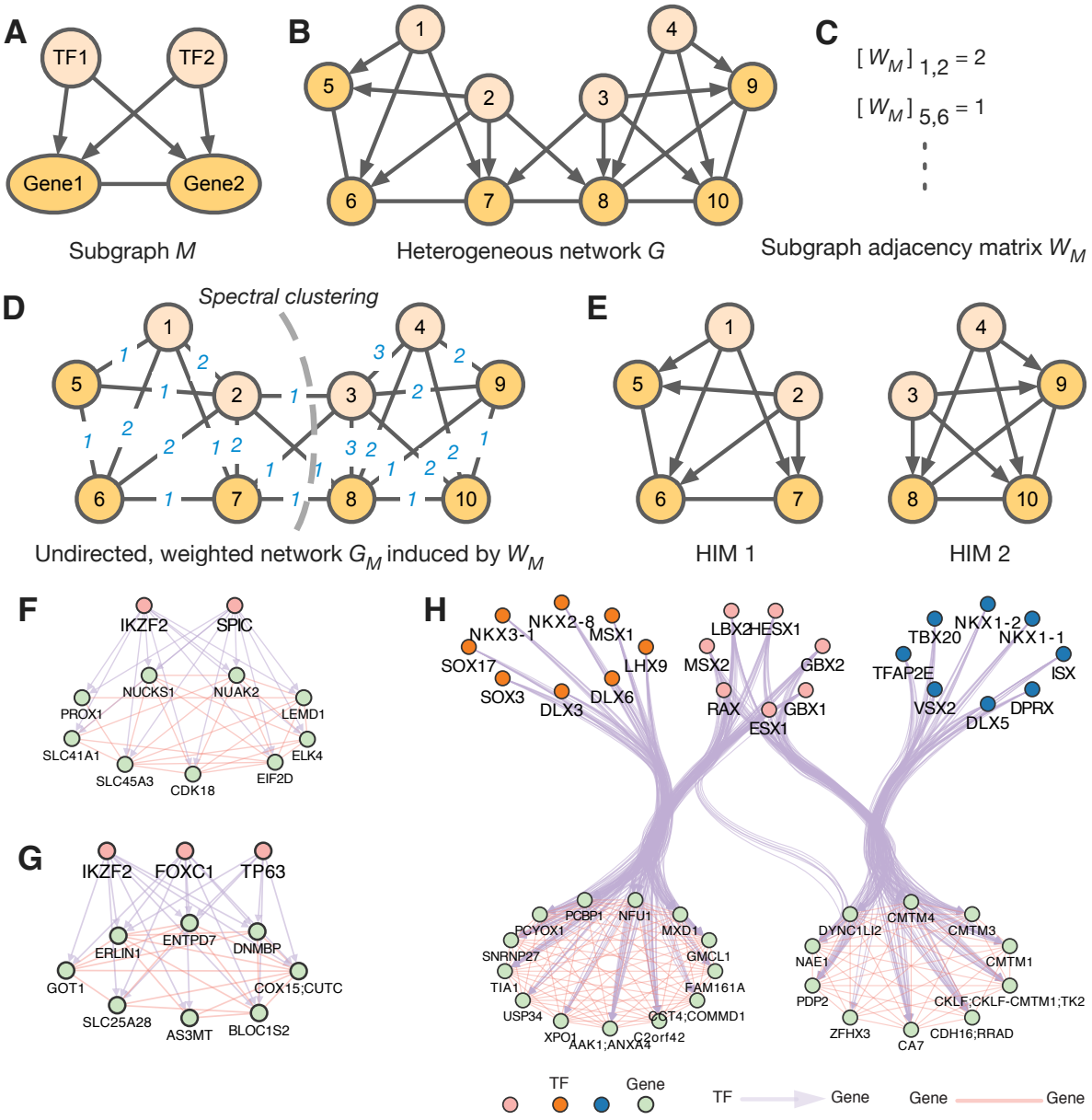


**Figure 1:** Workflow of our MOCHI algorithm and output examples of HIMs. The network has both gene-gene spatial proximity and TF-gene regulation relationships. **(A)** A 4-node motif $M$ represents the smallest HIM. Here directed interaction represents a TF-gene regulation relationship, an undirected 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)** Two HIMs as examples 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.

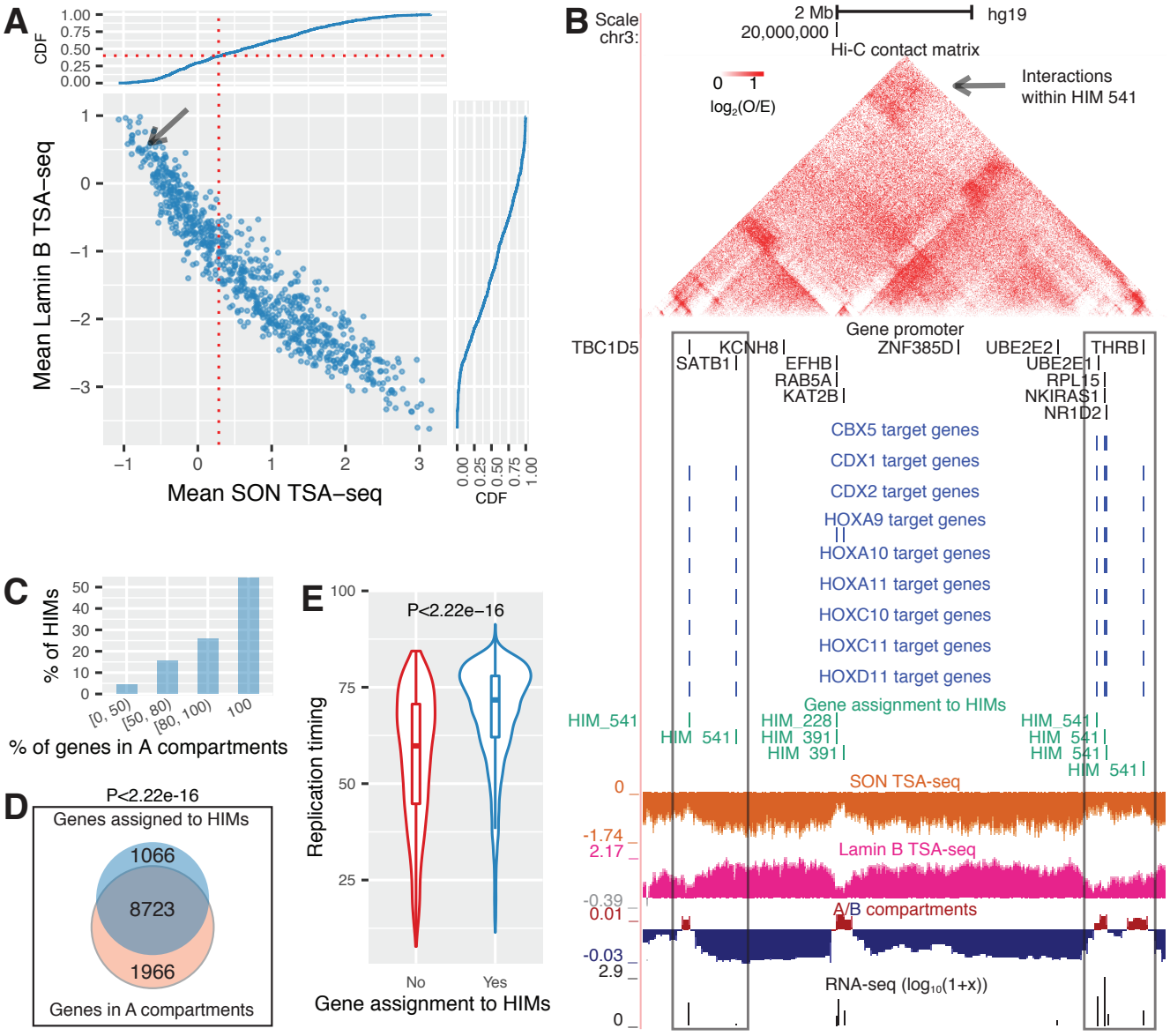


**Figure 2:** HIMs tend to be close to nuclear interior, in particular, speckles. **(A)** Scatter plot shows the mean SON TSA-seq score and mean Lamin B TSA-seq score of the genes in each HIM. Each dot represents a HIM. The curves on top and right are cumulative density functions (CDF). The red vertical dotted line represents the mean SON TSA-seq at 0.284 (approx. within $0.518\,\mu\mathrm{m}$ of nuclear speckles [12]). The black arrow points to HIM #541. **(B)** HIM #541 with low mean SON TSA-seq (pointed by the arrow in **(A)**). 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. Target genes of different TFs, gene members of HIM, SON TSA-seq, LaminB TSA-seq, A/B compartments, and RNA-seq signals are shown in different tracks. **(C)** Barplot shows the proportion of HIMs with a varied proportion of genes in A compartment. **(D)** Venn diagram shows that the genes assigned to HIMs are enriched in A compartment. **(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.

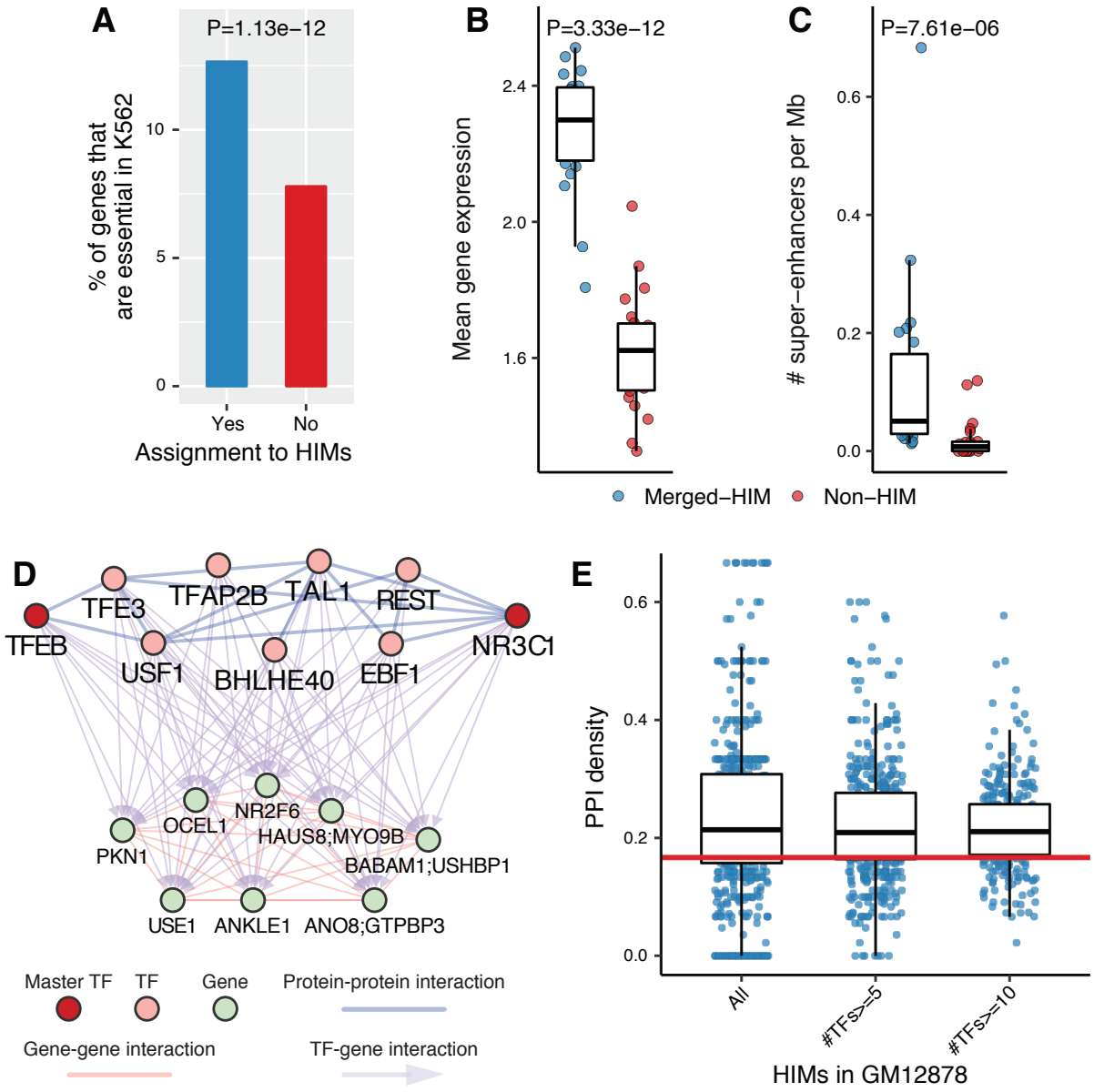


**Figure 3:** HIMs are enriched with essential genes, super-enhancers, and protein-protein interactions. **(A)** Barplots show the proportions of genes that are K562 essential genes among the genes assigned to HIMs and those not assigned to HIMs. **(B-C)** Functional properties of the genes in the identified HIMs in K562. To make a fair comparison, we stratified the genes assigned to HIMs by chromosome 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) among themselves. **(D)** One example of HIM from GM12878 cell line shows that 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 medians are significantly ($p$<2.22e-16) higher than the expected density (0.158, red line) of the sub-PPI networks induced by randomly sampled TFs.

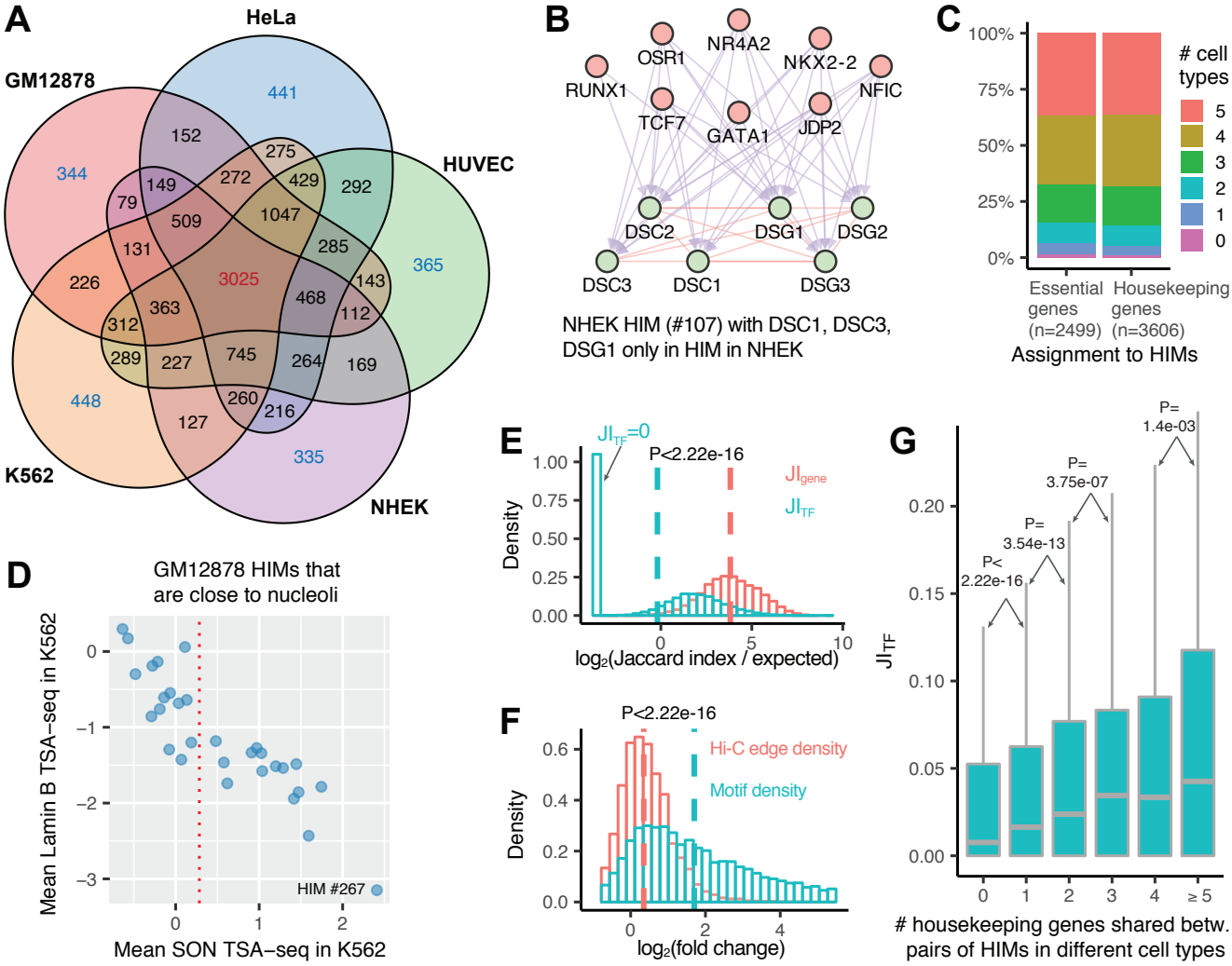


**Figure 4:** HIM comparisons in terms of genes and TFs across the cell types. **(A)** Venn diagram shows the assignment of genes in HIMs across five cell type. Numbers in each facet represent the gene number in each possible logic intersection relationship across five cell types. **(B)** A NHEK HIM with 3 genes only assigned to HIMs in NHEK. All of its genes are involved in keratinization pathway. Here the top and bottom nodes are the TFs and genes in the HIM, respectively. **(C)** Barplot shows the assignment of essential genes and housekeeping genes to HIMs across five cell types. **(D)** Scatter plot shows the mean SON TSA-seq and Lamin B TSA-seq scores (in K562 [12]) of the 30 GM12878 HIMs that are inferred as close to nucleoli in GM12878 [28]. The red vertical dotted line represents the mean SON TSA-seq score at 0.284. **(E)** The log-transformed ratio of Jaccard index on the genes/TFs between paired HIMs from different cell types over the expected Jaccard index between random control sets. **(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 dash line represents the median of a variable. **(G)** Boxplots show the distribution of Jaccard index on the TFs of paired HIMs with different numbers of shared housekeeping genes.

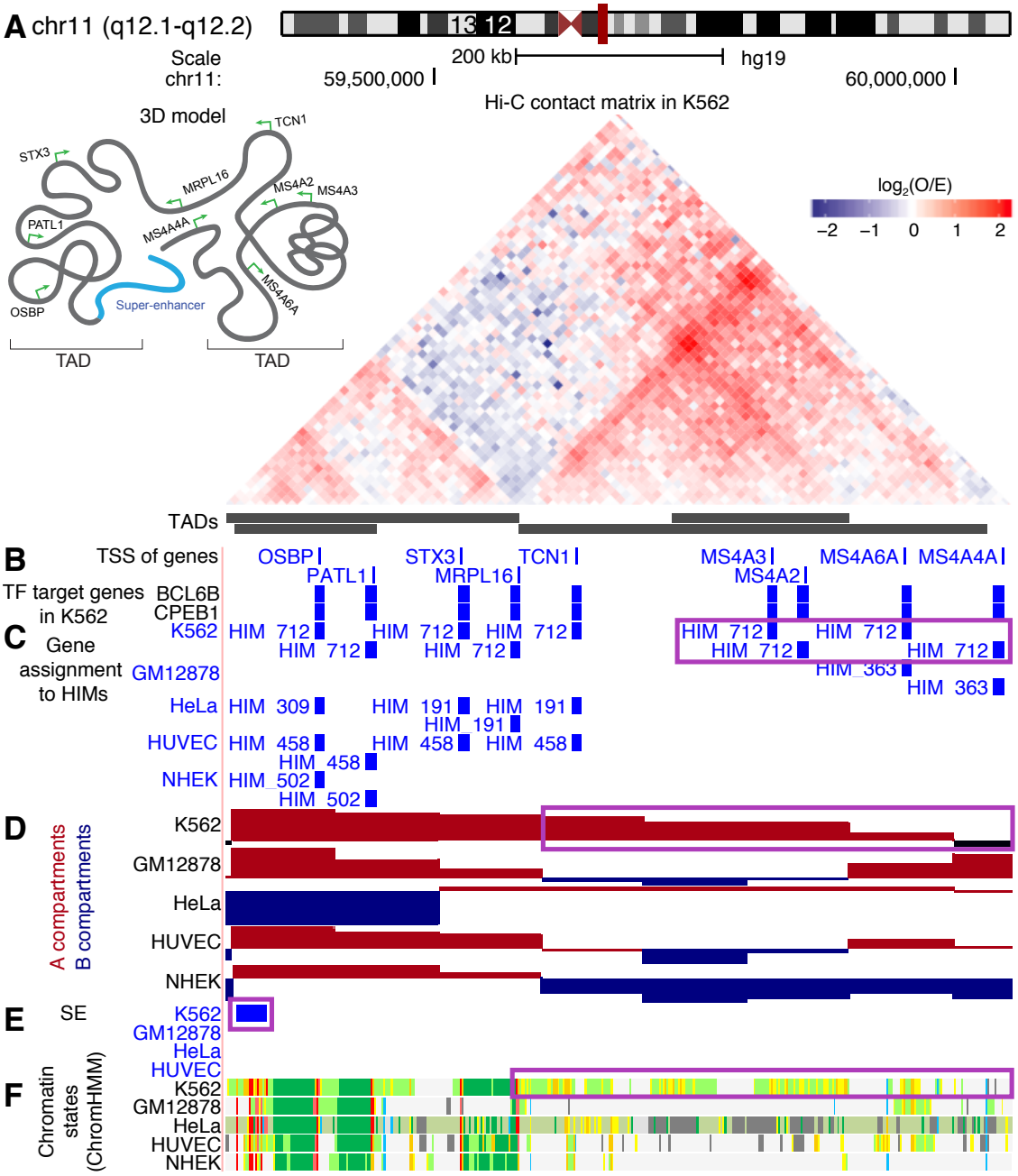


**Figure 5:** A K562 specific HIM with K562 specific chromatin interactome and functional annotations. **(A)** The 45 degree rotated upper triangle part of the contact matrix between the 10kb-sized bins in a chromosome region in K562. The region is segregated into 4 nested TADs. **(B)** Thin bars represent 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 promoters, orange and yellow stand for enhancers, green represents transcribed regions, gray represents other types of regions such as repressed regions.

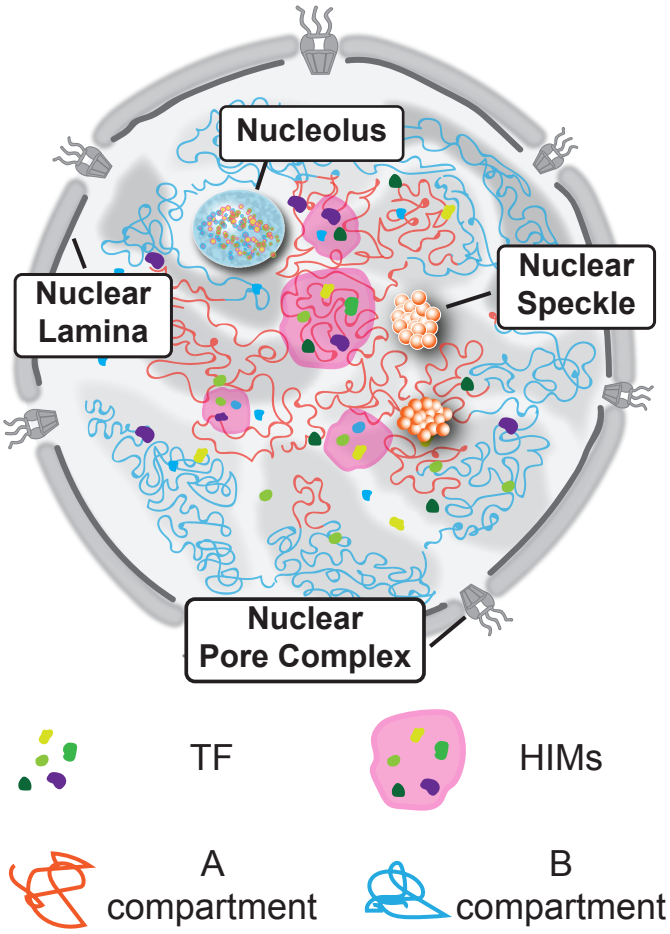


**Figure 6:** A possible model of HIMs within the nucleus.

# Supplemental Information

## 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 $i$-th 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 near linear time for this process [45]. 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 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 [41], i.e.,

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

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. (8) and the conductance $\varphi_{G_M}(S)$ is defined in Eq. (9):

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

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

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. (8), Eq. (9), 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.,

$$\varphi_{G_M}(S) \leq 4\sqrt{\varphi^\star_{G_M}} \leq 1. \tag{10}$$

*Comparison with the proofs in Benson et al. [24]*

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 [24] where $W_M$ is computed based on a homogeneous motif and homogeneous network. Benson et al. [24] proved that $S$ and $\bar{S}$ are near optimal for their case. However, the results in [24] 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 [24].

## A.4 Properties of the identified HIMs

Here we describe the properties of HIMs that are not defined in the main text. The features are the topological structural features related to connection patterns of the genes and TFs in a given HIM within a heterogeneous network, including motif density, chromatin interaction edge density (or Hi-C edge density), and GRN edge density, which quantify the connection strength between genes/TFs in a HIM in terms of different connection patterns. All 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 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 later in Supplemental Information B.2 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 chromatin 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 TF-gene interaction does not exist in a 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 [6]. 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 [5]. 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 used the size of compartments to control the 1D distance between

genes farthest apart in HIMs. The 99-th percentile of the size of compartments is around 10Mb in 4 out of the 5 cell types. Thus we chose 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 [46, 47] were downloaded from the UCSC Genome Browser [48].

GRN data were downloaded from [14], 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 enriched 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 [49], BioGrid [50], CORUM [51], and STRING [52]. We first extracted the PPIs between the 591 TFs in the GRNs from these public sources. We then combined them into a PPI network after merging duplicated PPIs. Note that the GRNs have the same set of TF protein. Thus the PPI network is suitable for all 5 different cell types. The density of the PPI network is 0.158, which is also the expected density of a sub-PPI network of a set of randomly sampled TFs.

Essential genes in four cell lines were downloaded from [26]. However, among the four cell lines, only K562 matches the cell types used in this study. The essential genes identified in K562 were only used to analyze identified HIMs in K562. Because the majority of the essential genes are shared between the 4 cancer cell lines [26], 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.

RNA-seq data with GEO accession # GSE33480 were downloaded from the ENCODE project [25]. The gene expression level quantified in FPKM value across the 5 cell types were normalized by quantile normalization then logarithm transformed by the function $\log_{10}(1 + x)$.From the expression data, we constructed a list of cell type-specific genes for each cell type by the following two criteria. Given a cell type, (1) the gene expression value in the given cell type is higher than 0.1; (2) 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.

# B Supplementary Results

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

We 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 chromatin interaction networks. To test the robustness of HIMs, we construct another set of chromatin interaction networks by the cutoff at 2 (for O/E). The number of intra-chromosomal chromatin interactions with the cutoff at 2 is 64.6%-81.9% of the number of intra-chromosomal interactions with the cutoff at 1 across five cell types. We denote the sub-Hi-C interaction networks resulted from the 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 the highest scores. We then construct 3 different heterogeneous networks for each cell type as follows:

- The heterogeneous network combines the chromatin interaction network with the 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 chromatin interaction network with the cutoff at 2 and the whole GRN. The heterogeneous network is referred as sub-Hi-C + GRN.
- The heterogeneous network combines the chromatin interaction network with the 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 would be 1. We 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 much more similar to each other than to the sets of HIMs from the other cell types. Hierarchical clustering produces five major clusters (Fig. S10). Each cluster contains 3 different heterogeneous networks in the same cell type. Overall, the result suggests that the HIMs are less sensitive to the parameters used in constructing the chromatin interactome and GRNs.

## B.2 Justification of the 4-node motif *M*

To justify the choice of the motif $M$, we compared it with two different types of motifs. One is a triangle motif with 3 nodes (Fig. S11A), where two of them are genes with a chromatin interaction and the third node is a TF that regulates both the genes. The triangle motif does not explicitly encode co-regulation between TF proteins. Another motif is a bifan motif with 4 nodes (Fig. S11A), where two nodes are TF proteins that co-regulate two genes but there is no chromatin interaction between the two genes. The bifan motif does not explicitly encode spatial proximal relationship between genes. For bifan motif, we applied MOCHI with the bifan on the GRN and then split the identified HIMs by chromosome number.

We found that the motif $M$ and triangle motif are better than the bifan motif in terms of identifying the clusters (Fig. S11B). Compared to the HIMs identified by the bifan motif, the HIMs by the motif $M$ and triangle motif have higher Hi-C edge density, higher triangle density, higher motif $M$ density.

Moreover, the genes in a HIM are closer to each other in the 1D sequence space, although the HIMs by the bifan motif have a smaller number of genes as compared to the HIMs by the motifs $M$ and triangle. This result highlights that the chromatin interaction between the two target genes in a motif is important to capture spatial proximity between the genes in HIMs.

In addition, we found that our motif $M$ is better than the triangle motif. We comprehensively compared the identified HIMs by the motif $M$ and the triangle motif. The HIMs identified by the two motifs have similar numbers of genes as the median numbers of genes are equal in the 4 out of 5 cell types. A similar pattern is observed on the Hi-C edge density. Specifically, the density is only significantly different in two cell types: GM12878 and NHEK ($p\leq$ 0.04), although the difference is small (the median density is 0.015 in GM12878 and 0.028 in NHEK) (Fig. S11B). However, the identified HIMs by the two motifs are very different in other features. Compared to the HIMs identified by the triangle motif, the HIMs identified by the 4-node motif $M$ have much higher numbers of TFs ($p\leq$ 2.45e-21). The difference in the median number of TFs ranges from 4 to 8 (Fig. S12). Even though the HIMs by the 4-node motif $M$ have higher number TFs, they have comparable numbers of genes as compared to the HIMs from the triangle. They also have much higher triangle density and much higher 4-node motif $M$ density ($p\leq$2.68e-02 and $p\leq$1.43e-03, respectively; Fig. S11B). The HIMs from $M$ also have a higher proportion of genes in the A compartment in 4 cell types ($p\leq$2.18e-02), are much earlier replicated ($p\leq$6.74e-03), and have smaller replication timing coefficient of variation ($p\leq$4.38e-05) (Fig. S12).

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

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

where $Y$ is a given feature, $\mathbb{1}_{\text{motif}} = 1$ if the HIM is identified with the 4-node motif $M$, $\mathbb{1}_{\text{motif}} = 0$ if the HIM is identified with the triangle motif, and $\hat{\beta}_3$ indicates the averaged difference in the feature $Y$ between the HIMs identified by the two motifs after adjusting the number of TFs and the number of genes in the HIMs. Specifically, a positive $\hat{\beta}_3$ means that HIMs 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 Table S4. Overall, the differences are still significant after adjusting the number of TFs and the number of genes. Take together, the 4-node motif $M$ is better than the triangle motif in identifying HIMs.

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

We found that the HIMs in 5 cell types in this study share similar connections with 3D genome organization features, such as A/B compartments, TADs, and loops. We looked at the genomic regions of each HIM that is the smallest genomic block containing the 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 (Table S2), comparable to the size of A/B compartments (median size is 5Mb). The median numbers of TADs in the genomic regions of the HIMs are 3-4 in different cell types (Table S2). The genomic regions of the HIMs have, on average, 7 chromatin loops in GM12878 and 2-4 loops in the other cell types (Table S2). The HIMs in GM12878 involve a higher number of loops, which perhaps is due to the

fact that GM12878 has at least 60% more detected loops than the other cell types possibly due to higher sequencing depth [6].

**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 (%)' represents the proportion of genes in a heterogeneous network that are assigned to HIMs.

| | | GM12878 | HeLa | HUVEC | K562 | NHEK |
|---|---|---|---|---|---|---|
| Input | TFs | 591 | 591 | 591 | 591 | 591 |
| | Genes | 11,627 | 12,036 | 11,927 | 12,391 | 12,161 |
| | TF$\rightarrow$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 farthest apart genes (Mb) | 8 | 5.4 | 6.1 | 7.2 | 4.9 |
| # of HIMs inherited TFs | 451 | 599 | 546 | 596 | 497 |
| Median proportion of inherited TFs | 28.6 | 27.8 | 24.6 | 25.0 | 28.0 |

**Table S3:** The dynamics of chromatin interaction networks and GRNs across 5 cell types. Cells in the table are numbers/proportions of interactions that exist in the corresponding number of cell type in each column. For example, column '1' corresponds to the interactions that only exist in one cell type. Column '5' corresponds to the interactions that exist in all 5 different cell types. Overall, a large proportion of the edges in the GRNs and chromatin interaction networks only exist in one cell type.

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

**Table S4:** Comparison between 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 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 compartment | 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 consistently or in a cell type-specific manner. The number of genes in each category is shown in Fig. 4A.

| | **GO term/Pathway** | **Count** | **Fold Enrichment** | ***P* value** |
|---|---|---|---|---|
| **Constitutive genes** | chromosome organization | 371 | 1.50 | 5.7e-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 |

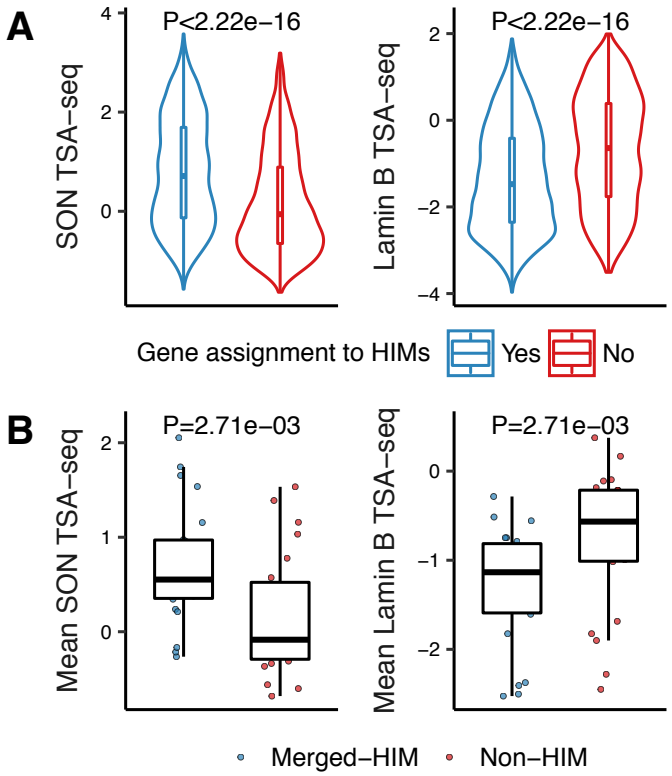


**Figure S1:** Genes assigned to HIMs are closer to nuclear speckles as compared to the genes in the heterogeneous network that are not assigned to HIMs in K562. **(A)** Violin plots show the distributions of TSA-seq scores of the two sets of genes. **(B)** Boxplots show the distributions of mean TSA-seq scores of the merged-HIM clusters and non-HIM clusters. Here we 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.

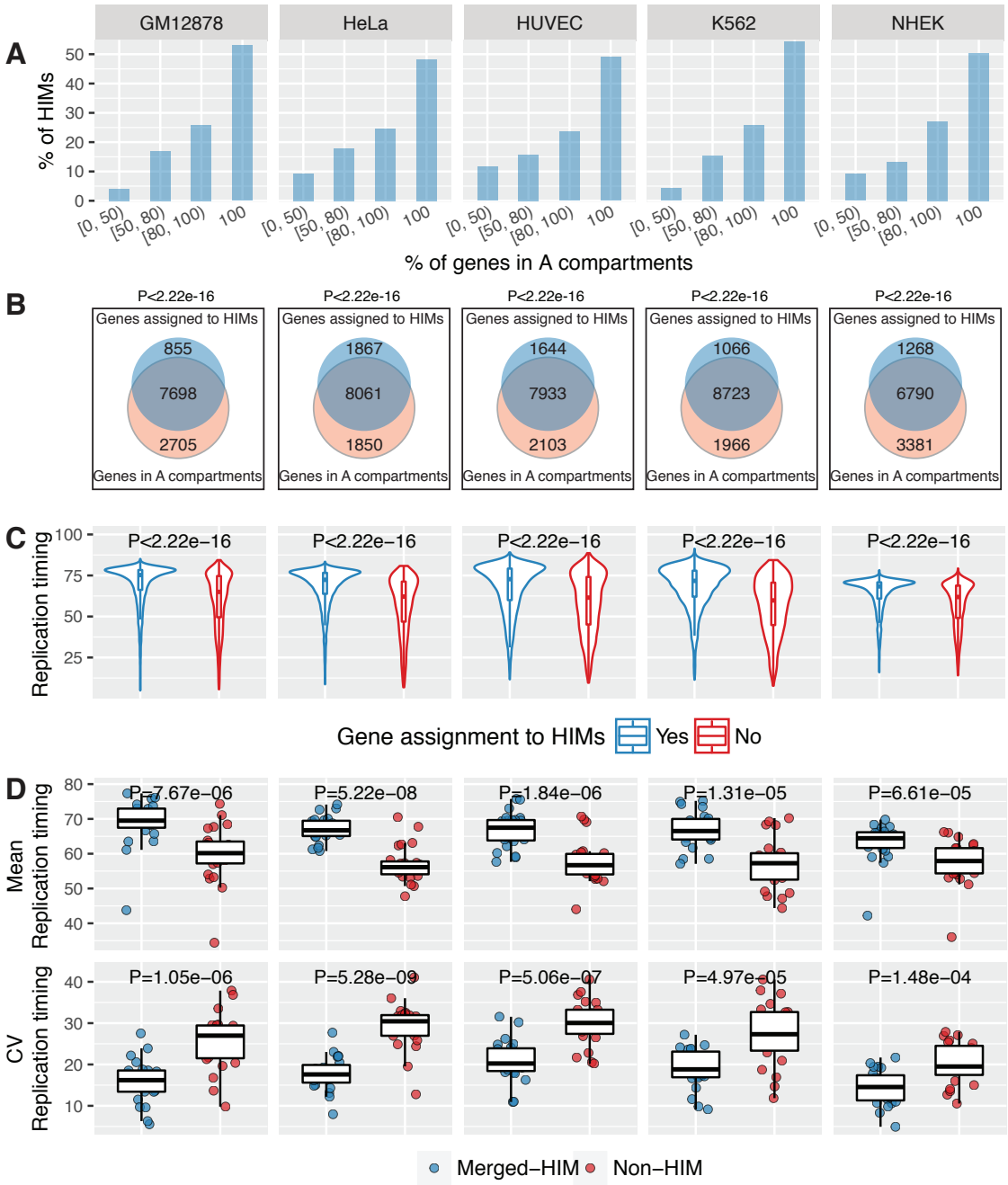


**Figure S2:** HIMs consistently have spatial location preferences as compared to non-HIMs in five cell types. Rows correspond to the spatial location features. Columns correspond to the cell types. **(A)** Barplot shows the distribution of HIMs with a varied proportion of genes that are in A compartment. **(B)** Venn diagram shows that the genes assigned to HIMs, as a whole, are enriched with the genes in A compartment. **(C)** Boxplots compare the replication timing of the genes that are assigned to HIMs against the genes that are not assigned to HIMs. **(D)** Boxplots show the mean and coefficient of variation (CV) of replication timing of the genes in merged-HIMs or 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 synchronously.

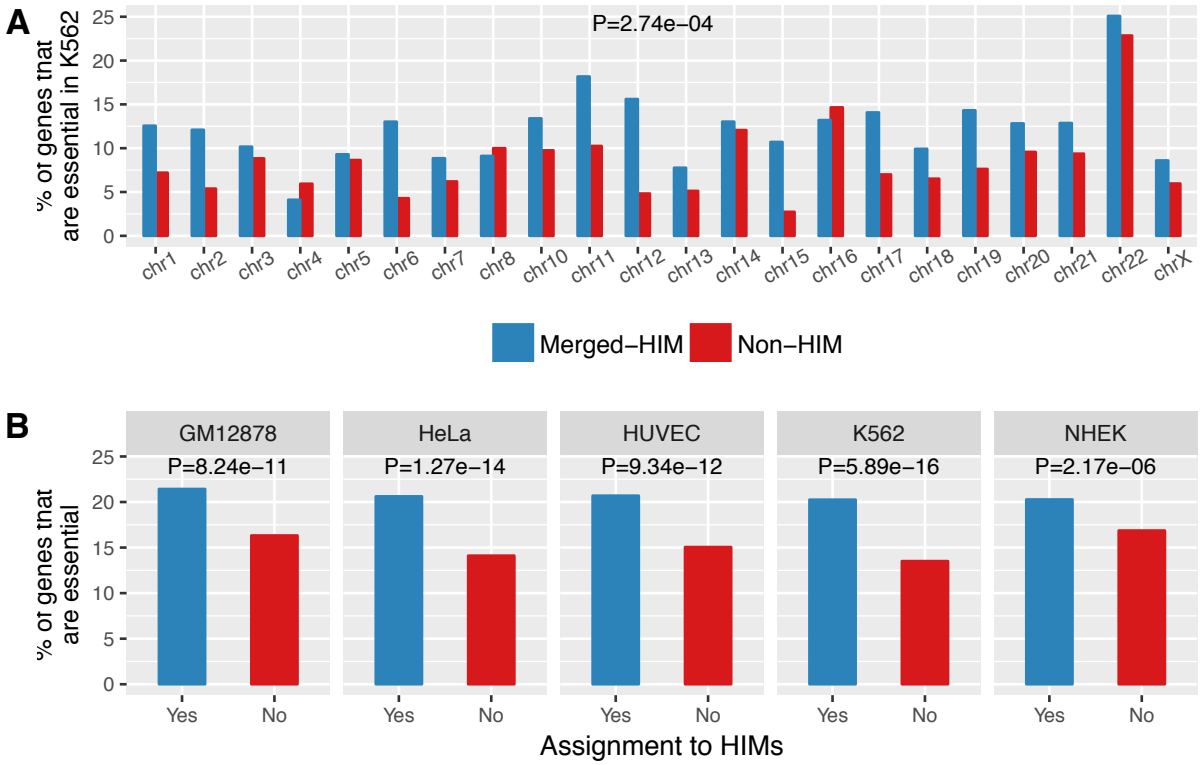


**Figure S3:** HIMs have more essential genes across cell types. **(A)** Barplots show the proportions of genes that are K562 essential genes in the genes assigned to HIMs and the genes not assigned to HIMs. **(B)** Barplots show the proportions of K562 essential genes in merged-HIMs and non-HIMs across 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 the genes not assigned to HIMs. P-value is computed by the Chi-squared test of independence. The proportions of the essential genes in merged-HIMs and non-HIMs are similar to **(A)** thus are not shown here.

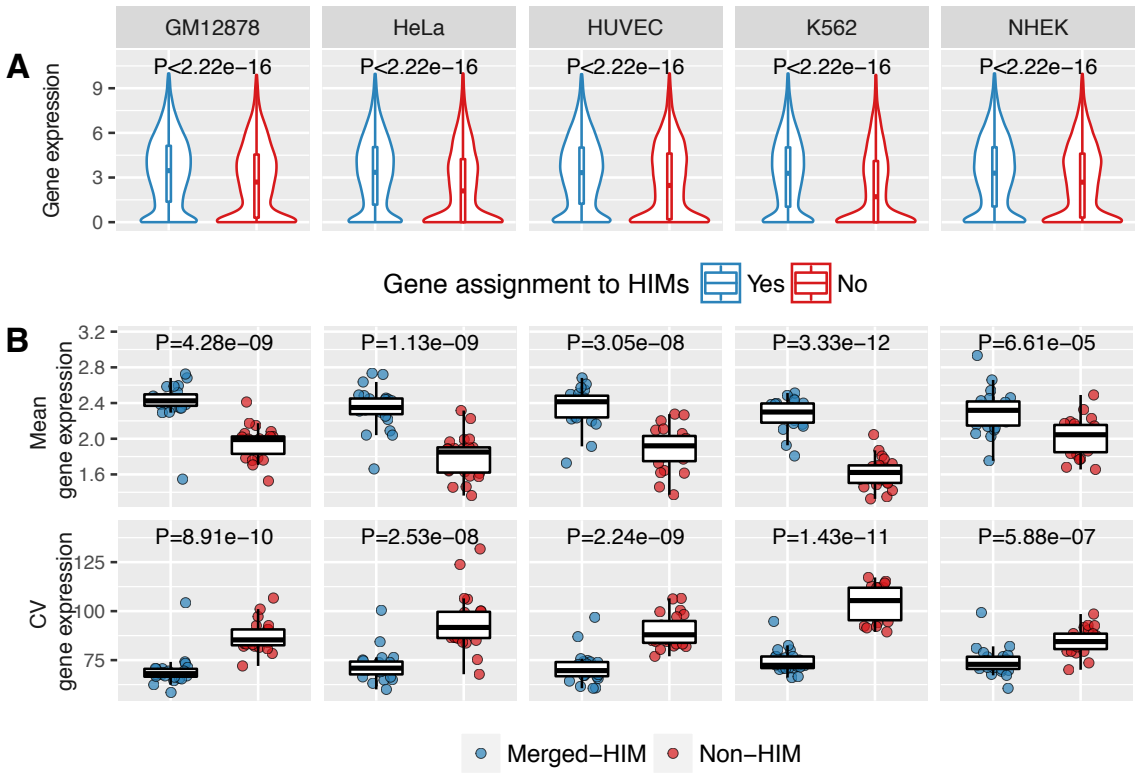


**Figure S4:** The genes assigned to HIMs express at higher levels than the genes not assigned to HIMs across cell types. **(A)** Violin plots show that the genes assigned to HIMs, as a whole, are more expressed than the genes not assigned to HIMs. **(B)** Boxplots show that the merged-HIMs have higher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

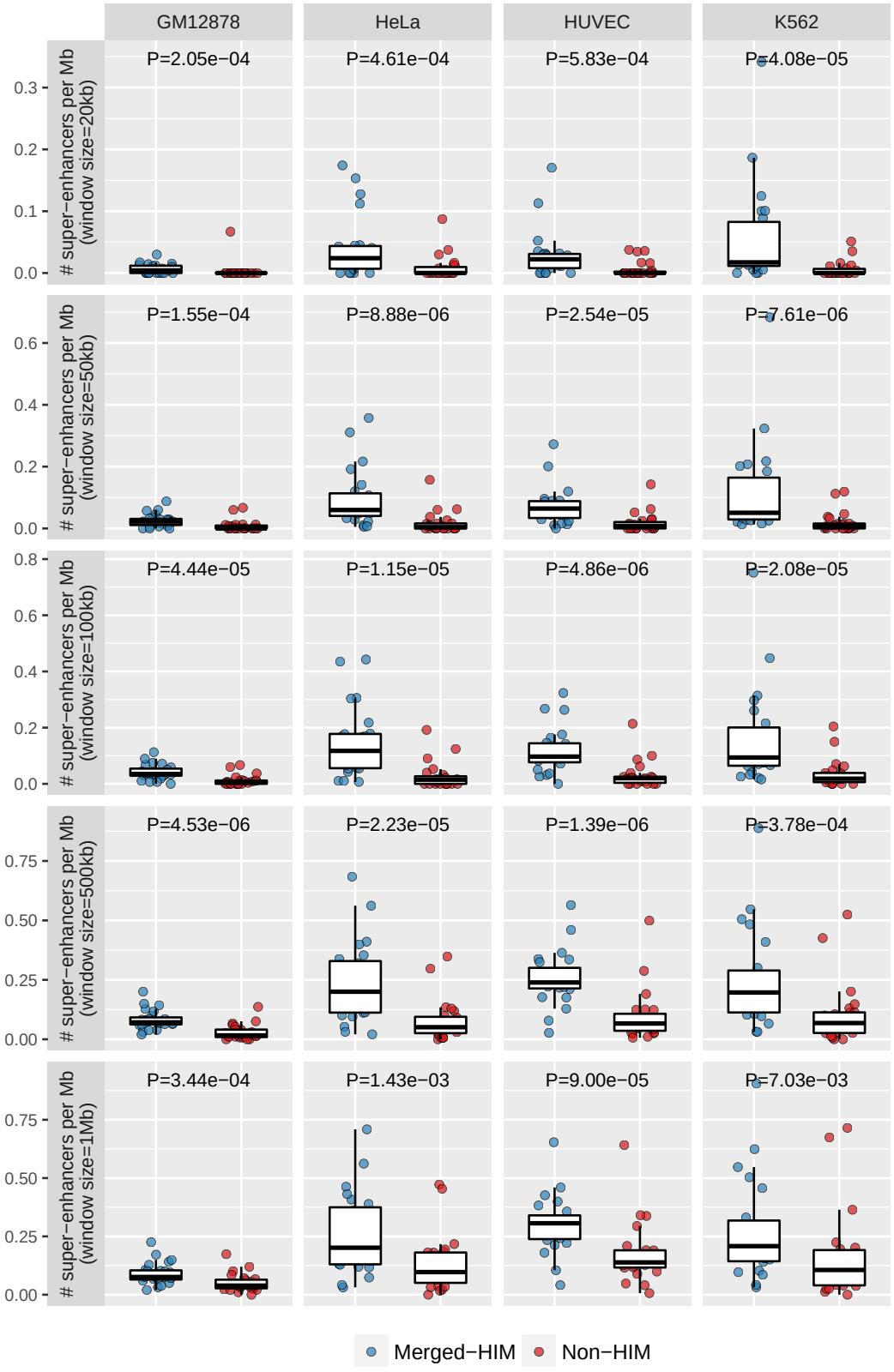


**Figure S5:** HIMs are enriched with super-enhancers and this observation is robust to the window size. The window size is used to define the genes that are close to a given super-enhancer. The window size ranges from 20kb to 1Mb. The distribution of super-enhancers in NHEK is missing due to lack of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

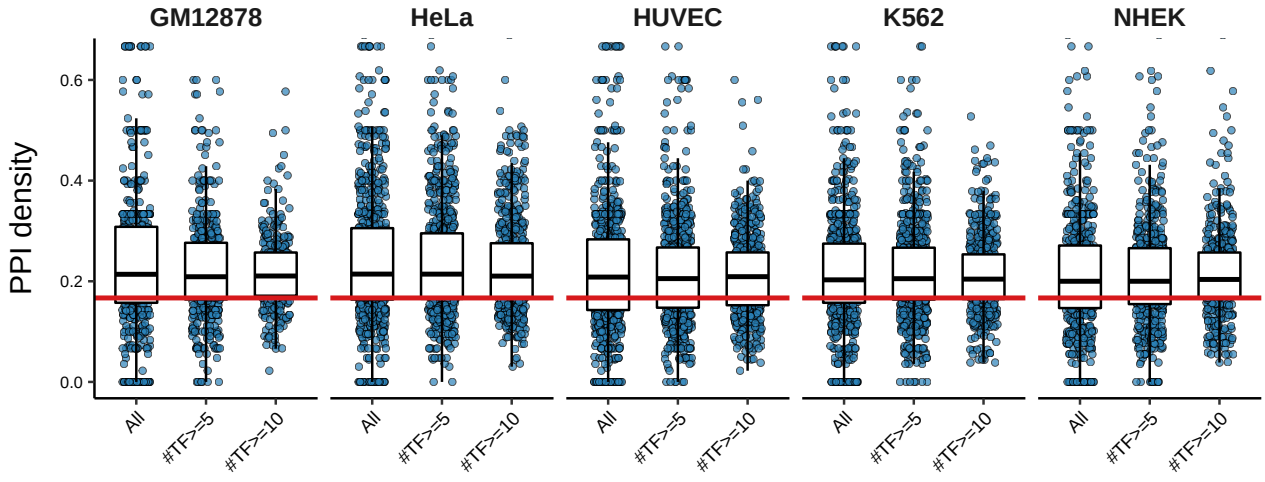


**Figure S6:** TFs in HIMs are enriched with protein-protein interactions (PPIs). Boxplots show the distribution of the sub-PPI network density across HIMs and subsets of HIMs with at least $n$ TFs, $n = 5, 10$. Here for each HIM, we computed the density of the sub-PPI network induced by the TFs in the HIM from the PPI network based on 591 TF proteins used in this study. The medians are all higher than the expected density (0.158, red line) of the sub-PPI networks induced by randomly sampled TFs ($p$<2.22e-16).

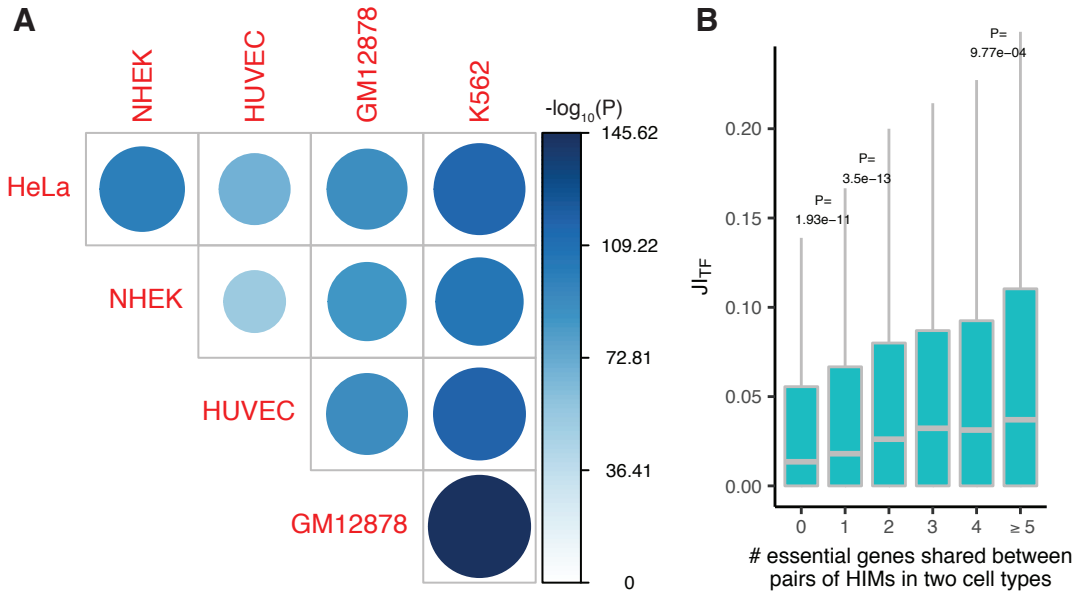


**Figure S7:** HIM comparisons regarding genes and TFs across cell types. **(A)** Heatmap shows the level of significance in overlapping between the genes assigned to HIMs in two different cell types. GM12878 and K562 have the highest overlap. Statistical significance is evaluated by the hypergeometric test. **(B)** Boxplots show the distribution of Jaccard index on the TFs of paired HIMs with different numbers of shared essential genes.

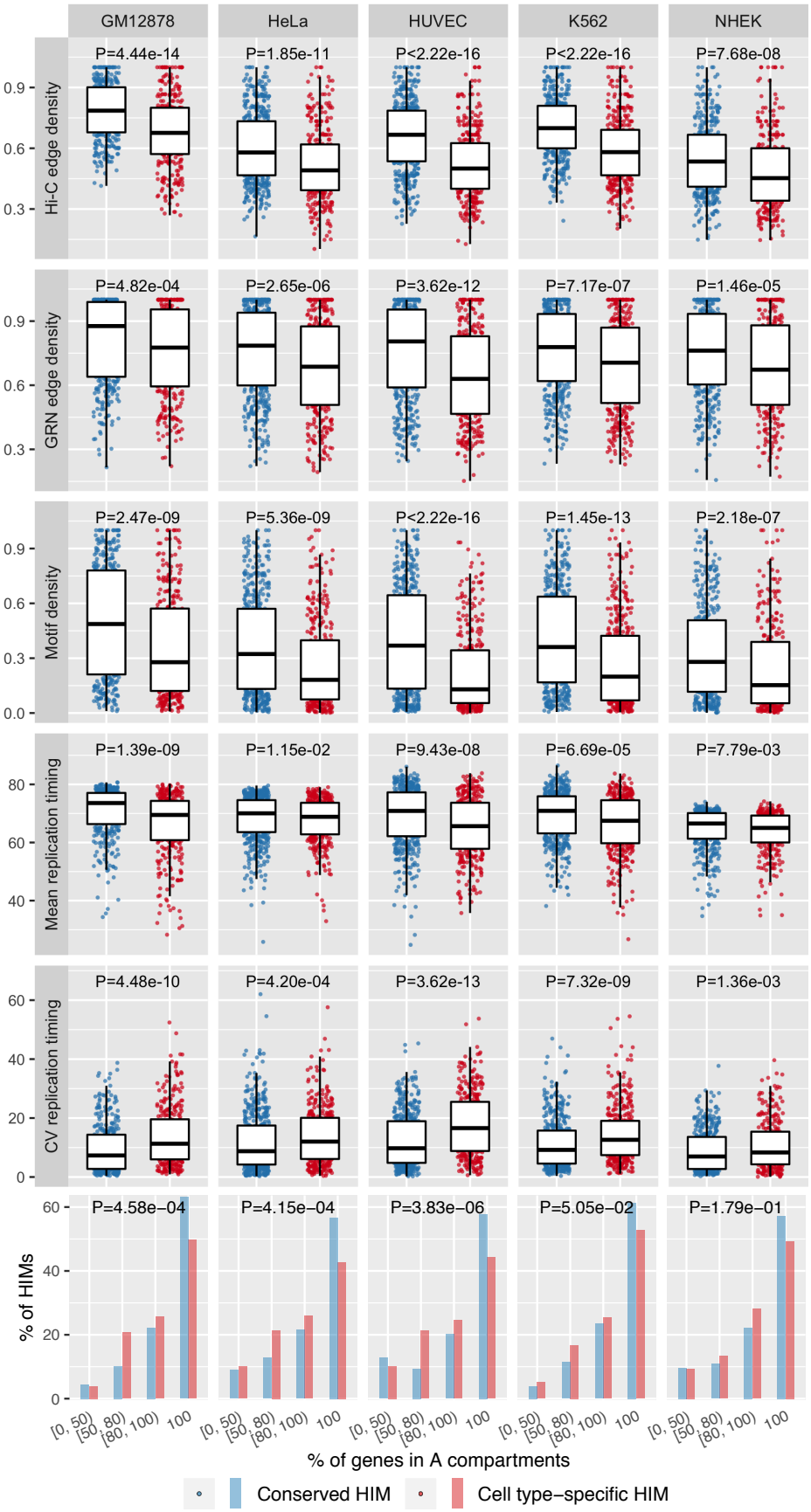


**Figure S8:** Conserved and cell type-specific HIMs have distinct spatial location features.

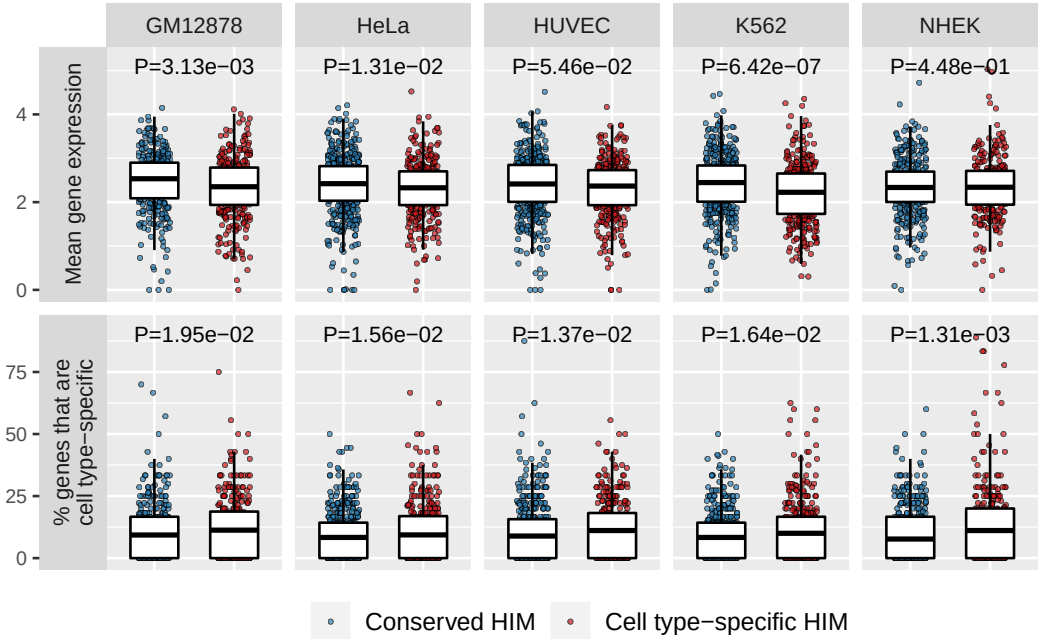


**Figure S9:** Functional differences and similarities between conserved and cell type-specific HIMs. Conserved HIMs have higher average gene expression in 3 cell types (first row). On the other hand, cell type-specific HIMs tend to have a higher proportion of cell type-specific genes in all 5 cell types (second row).

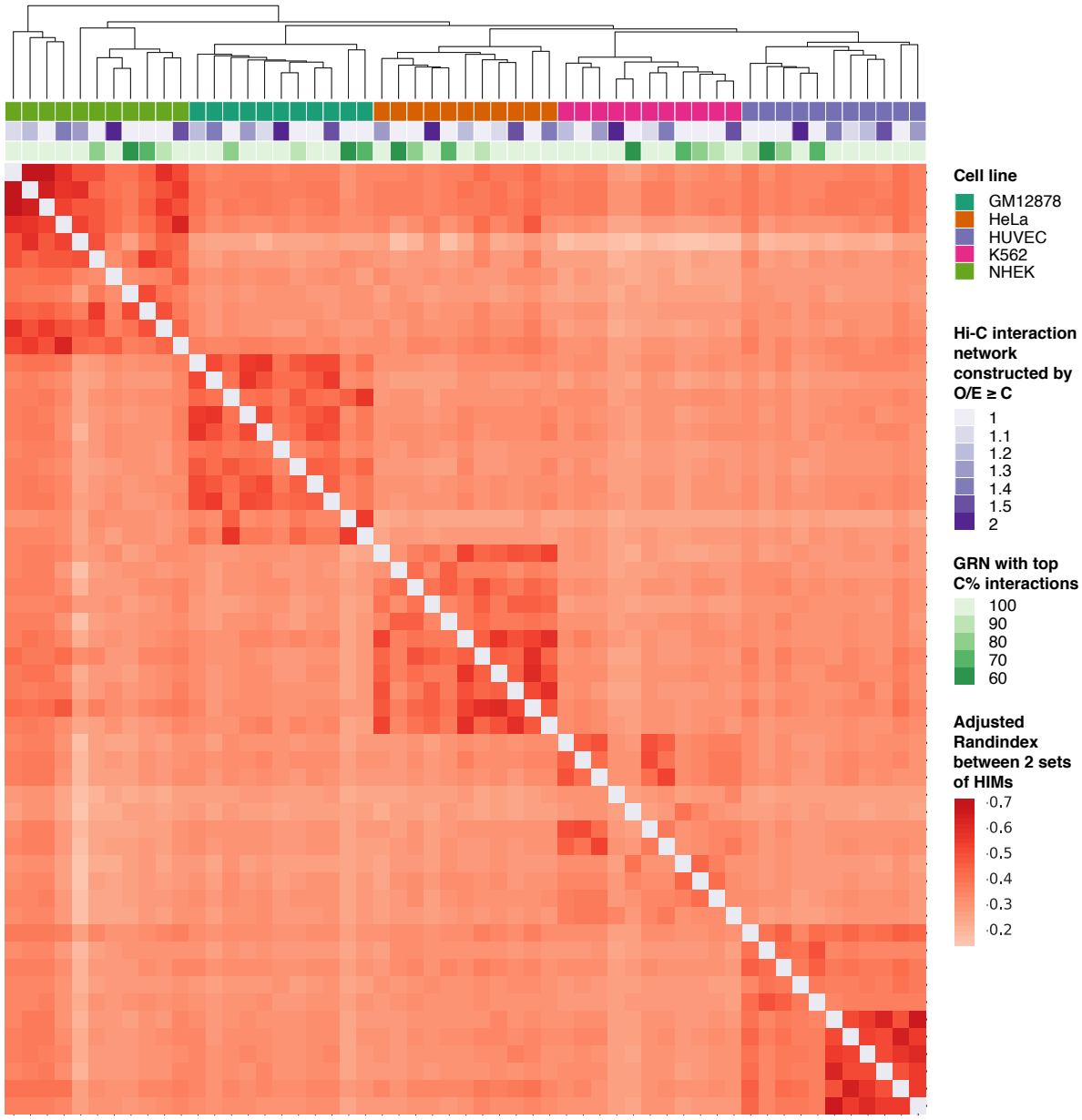


**Figure S10:** HIMs are robust to the input heterogeneous networks. For each cell type, there are 3 heterogeneous networks: one used in the main text and two sub-heterogeneous networks derived from different combinations of Hi-C interaction networks and GRNs. A sub-Hi-C represents a sub-Hi-C interaction network with interactions satisfying O/E$\geq 2$. 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 much more similar as compared to the sets of HIMs from different cell types.

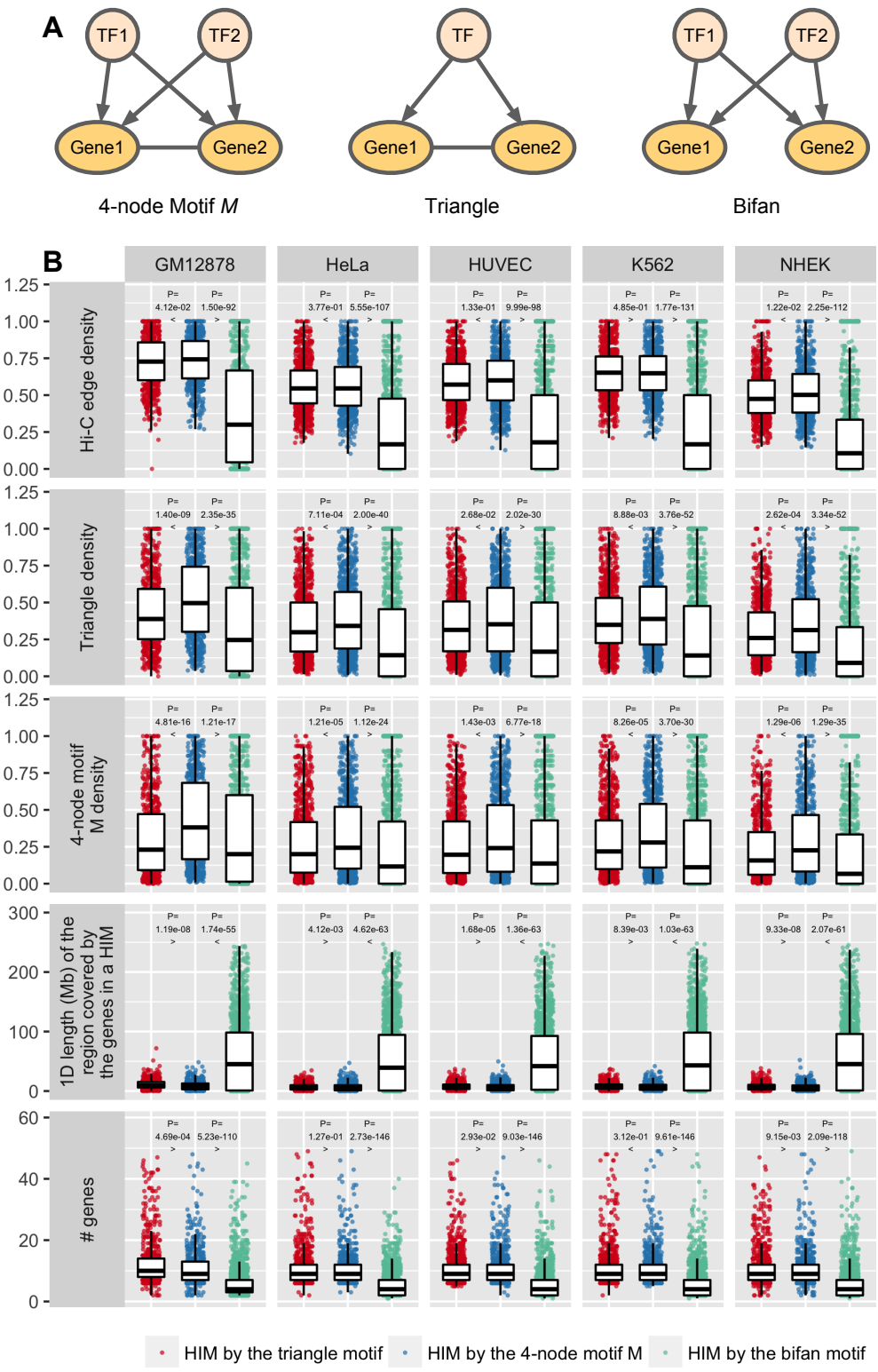


**Figure S11:** Comparison of the identified HIMs between the 4-node motif $M$, a triangle motif, bifan motif across 5 cell types. **(A)** Illustration of the 3 different motifs. **(B)** Comparison between the HIMs identified by the 3 different motifs. Rows correspond to the features. Columns correspond to the cell types. Each dot represents a HIM.

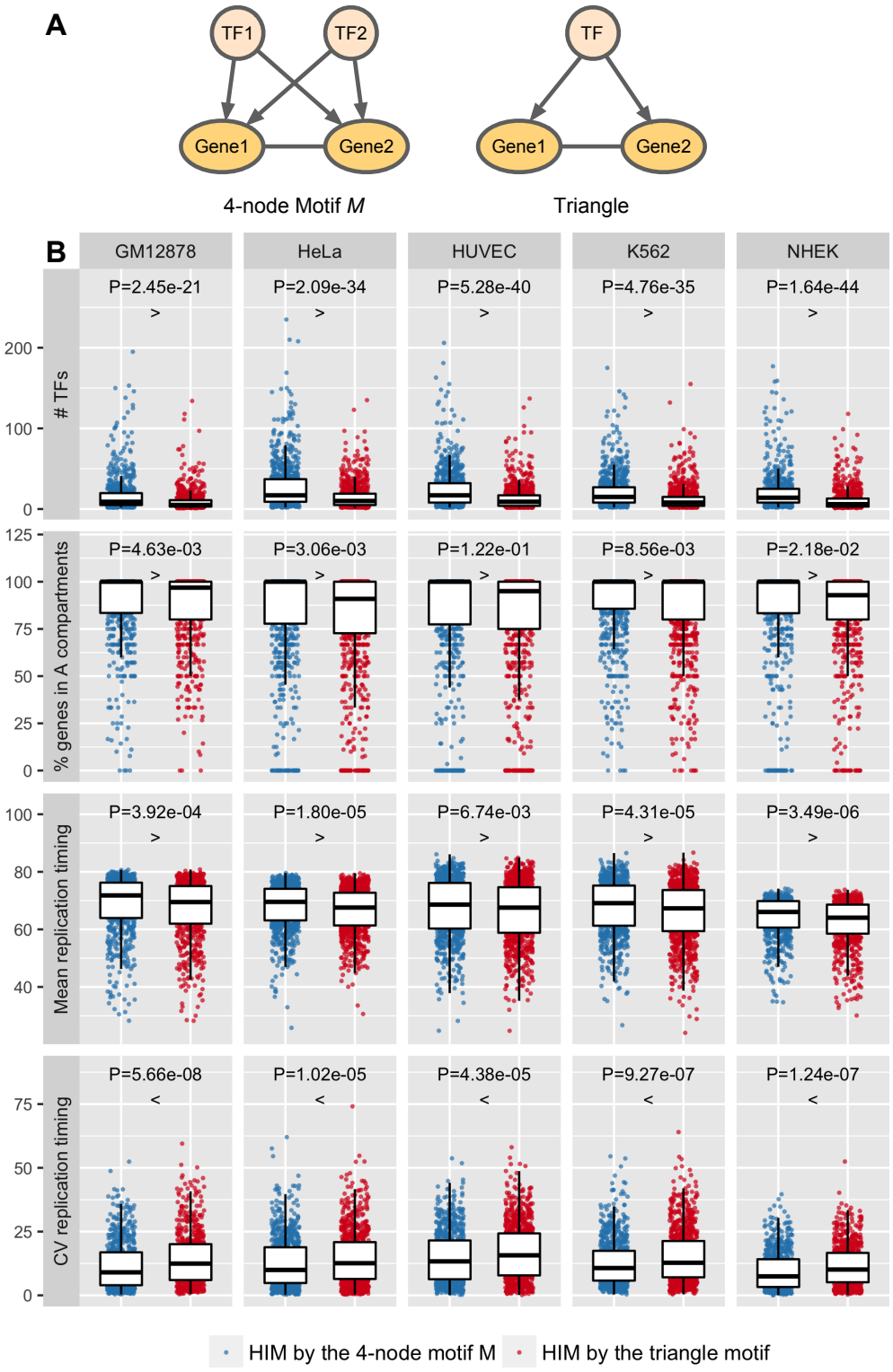


**Figure S12:** Comparison of identified HIMs between the 4-node motif $M$ and the triangle motif across the 5 cell types. Where the triangle motif has 3 nodes **(A)**. Two of them are genes connected by a chromatin interaction and co-regulated by a TF. **(B)** Rows correspond to the features. Columns correspond to the cell types. Each dot represents a HIM identified either by the 4-node motif $M$ or the triangle motif. The patterns are consistently observed after adjusting the numbers of TFs and genes in HIMs by a linear regression model (Table S4).