

# SCIENTIFIC REPORTS



OPEN

## Chromatin interaction networks revealed unique connectivity patterns of broad H3K4me3 domains and super enhancers in 3D chromatin

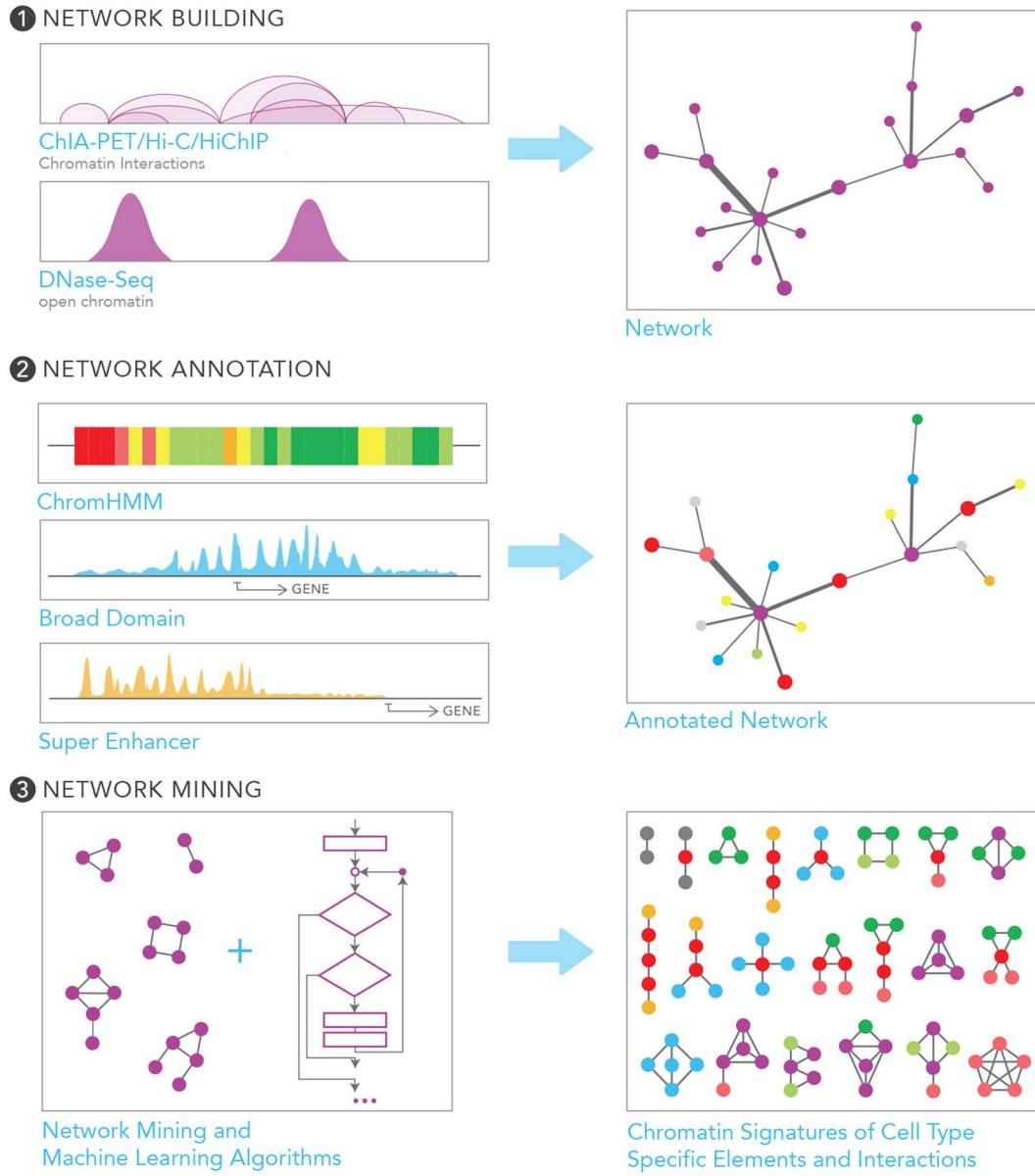
Asa Thibodeau<sup>1</sup>, Eladio J. Márquez<sup>2</sup>, Dong-Guk Shin<sup>1</sup>, Paola Vera-Licona<sup>3,4,5</sup> & Duygu Ucar<sup>2,5</sup>

Broad domain promoters and super enhancers are regulatory elements that govern cell-specific functions and harbor disease-associated sequence variants. These elements are characterized by distinct epigenomic profiles, such as expanded deposition of histone marks H3K27ac for super enhancers and H3K4me3 for broad domains, however little is known about how they interact with each other and the rest of the genome in three-dimensional chromatin space. Using network theory methods, we studied chromatin interactions between broad domains and super enhancers in three ENCODE cell lines (K562, MCF7, GM12878) obtained via ChIA-PET, Hi-C, and Hi-CHIP assays. In these networks, broad domains and super enhancers interact more frequently with each other compared to their typical counterparts. Network measures and graphlets revealed distinct connectivity patterns associated with these regulatory elements that are robust across cell types and alternative assays. Machine learning models showed that these connectivity patterns could effectively discriminate broad domains from typical promoters and super enhancers from typical enhancers. Finally, targets of broad domains in these networks were enriched in disease-causing SNPs of cognate cell types. Taken together these results suggest a robust and unique organization of the chromatin around broad domains and super enhancers: loci critical for pathologies and cell-specific functions.

Cell-type-specific functions of super enhancers and broad domains have been extensively studied and well established across diverse cell types and organisms<sup>1–4</sup>, where their distinct epigenomic profiles were instrumental in their discovery. Super enhancers are demarcated by high levels of enhancer-associated histone modification mark H3 lysine 27 acetylation (H3K27ac) and are catalogued in 86 human cell and tissue types using this mark<sup>5</sup>. Moreover, super enhancers have been shown to harbor Single Nucleotide Polymorphisms (SNPs) associated with the diseases of the cognate cell type, including cancer<sup>2,4</sup>. Pharmacological molecules have been used to effectively and specifically target super enhancer domains at oncogenes<sup>6</sup>, further reinforcing their significance for disease biology. Similarly, cell type-specific promoters (i.e., broad domains) are associated with expanded deposition of histone H3 lysine 4 tri-methylation (H3K4me3) mark - a signature conserved across diverse cell types (>99 in human cells) and organisms<sup>3</sup>. Shortening of broad domains has been observed in cancer cells at tumor suppressor genes, enabling the discovery of novel tumor suppressors<sup>4</sup>. Recently, super enhancers and broad domains overlapping super enhancers were shown to be more associated with chromatin interactions than their typical counterparts<sup>6</sup> suggesting a unique organization of chromatin around cell-specific loci.

Chromatin structure plays a major role in governing cellular functions in a cell type- and condition-specific manner<sup>7</sup>. Advances in genomewide chromatin interaction profiling have shown that many regulatory elements

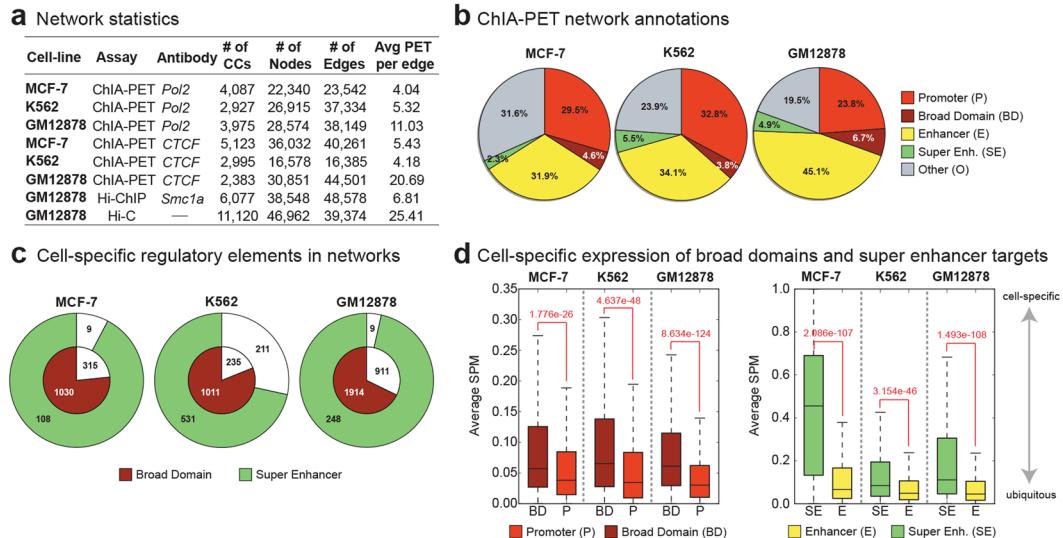
<sup>1</sup>Department of Computer Science & Engineering, University of Connecticut, Storrs, CT, USA. <sup>2</sup>The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA. <sup>3</sup>Center for Quantitative Medicine, University of Connecticut Health Center, Farmington, CT, USA. <sup>4</sup>Department of Cell Biology, University of Connecticut Health Center, Farmington, CT, USA. <sup>5</sup>Institute of Systems Genomics, University of Connecticut Health Center, Farmington, CT, USA. Correspondence and requests for materials should be addressed to P.V.-L. (email: [veralicona@uchc.edu](mailto:veralicona@uchc.edu)) or D.U. (email: [duygu.ucar@jax.org](mailto:duygu.ucar@jax.org))



**Figure 1.** Our data analysis framework. Our three-step data analyses framework is composed of (1) network building; (2) network annotation using ChromHMM states, broad domain, and super enhancer definitions; and (3) network mining using network measures, graphlets, and machine-learning models.

(*i.e.*, enhancers and promoters) that are distal on the linear genome map are actually in close physical proximity with each other as a result of the 3D chromatin structure<sup>8–10</sup>. Among these technologies, the Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) combines chromatin immunoprecipitation with chromatin conformation capture to identify chromatin interactions that are mediated by a protein<sup>8</sup>, such as RNA Polymerase II (Pol2) which mediates interactions between promoters and enhancers<sup>11</sup>. More recently, an alternative method has been developed, HiChIP<sup>12</sup>, to detect protein-centric chromatin interactions<sup>12</sup> using 100-fold less input material, providing an opportunity to generate such maps in primary human cells and tissues. These data-sets, particularly the ones capturing protein-mediated promoter and enhancer interactions enable genomewide study of chromatin interactions between broad domains and super enhancers.

This study utilizes advanced computational methods to uncover how broad domains and super enhancers interact in the 3D chromatin space, in particular, whether they are associated with distinct connectivity patterns, whether these patterns are conserved across cell types and assays, and whether they are predictive of the cell-specific nature of promoters and enhancers. For this, we built chromatin interaction networks using diverse assays (*i.e.*, ChIA-PET, Hi-C, HiChIP) in three ENCODE cell lines: MCF-7 (breast adenocarcinoma), K562 (chronic myeloid leukemia), and GM12878 (lymphoblastoid cell line). These networks were annotated using ChromHMM states<sup>13,14</sup>, super enhancer<sup>2</sup>, and broad domain<sup>3</sup> definitions in the corresponding cell types (Fig. 1). We studied interaction frequencies, network centrality measures and graphlets (*i.e.*, small connected



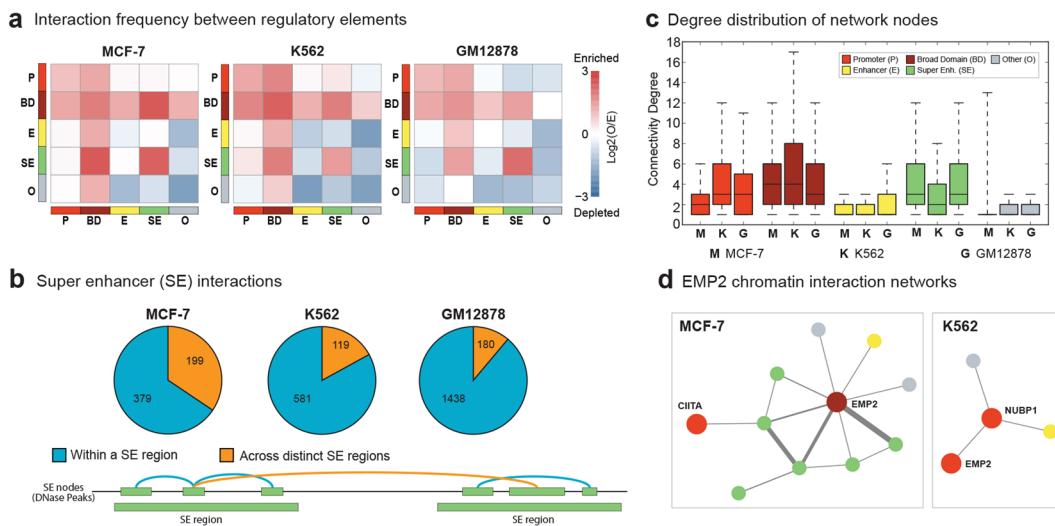
**Figure 2.** Regulatory elements in chromatin interaction networks. **(a)** Network statistics for chromatin interaction networks built from ChIA-PET, HiChIP, and Hi-C interactions in three ENCODE cell lines. **(b)** Regulatory annotations of Pol2 ChIA-PET network nodes. Note the enrichment of promoter and enhancer nodes in these networks. **(c)** Number of broad domains (inner chart) and super enhancers (outer chart) represented in ChIA-PET networks. Note that most broad domains and super enhancers in these cell lines are represented in the networks. **(d)** (Left) Cell-specific expression of genes associated to broad domains (dark red) and regular (red) promoters. Broad domain promoters are more cell-specific than regular promoters. (Right) Cell specific expression of genes associated to super enhancers (green) and regular enhancers (yellow). Super enhancer targets are more cell-specific than targets of regular enhancers.

non-isomorphic induced subnetworks)<sup>15</sup> to uncover distinct connectivity patterns associated with broad domains and super enhancers. Using machine learning models based on support vector machines (SVM)<sup>16,17</sup>, we showed that these chromatin connectivity patterns can effectively discriminate broad domains from regular promoters and super enhancers from regular enhancers. Our results suggest a unique and conserved chromatin organization around critical regulatory elements. Finally, we studied the clinical relevance of these annotated chromatin interaction networks by demonstrating that enhancers targeting broad domains harbor more SNPs associated to diseases of the cognate cell type.

## Results

**Chromatin interaction networks capture interactions among diverse regulatory elements.** We built chromatin interaction networks using Pol2 ChIA-PET data in three ENCODE cell lines: MCF-7 (derived from metastatic mammary gland epithelium), K562 (derived from chronic myelogenous leukemia cells), and GM12878 (lymphoblastoid cell line) and using Hi-C<sup>18</sup> and HiChIP<sup>12</sup> (targeting cohesion subunit *Smc1a*) data in GM12878. These networks consisted of 20–50 thousand network nodes/edges and thousands of connected components (Fig. 2a, Table S1). Next, nodes in these networks were annotated using ChromHMM states<sup>13,14</sup> in conjunction with broad domain<sup>3</sup> and super enhancer<sup>2</sup> definitions in corresponding cell types (Methods). In Pol2 ChIA-PET networks, a majority of nodes (68–80%) overlapped promoters and enhancers, showing the utility of Pol2-mediated ChIA-PET interactions to capture interactions between regulatory elements<sup>7</sup> (Fig. 2b, Supplementary Figure 1a). Majority of super enhancers and broad domains (> ~70%) were represented in these networks (Fig. 2c), in agreement with recent reports on super enhancers being more involved in chromatin interactions<sup>6</sup>. In comparison we also built chromatin interaction networks using CTCF-mediated ChIA-PET interactions. As expected these networks captured far fewer promoters, enhancers, broad domain and super enhancers (Supplementary Figure 2a-b, Table S1) and more insulator regions, suggesting that Pol2 ChIA-PET data is more suitable to study interactions between promoter and enhancer elements. Therefore, the rest of the ChIA-PET data analyses are conducted in Pol2 datasets. We noted that Hi-C networks include less number (~25–39% fewer) of promoters, broad domains and super enhancers compared to Pol2-associated assays (Supplementary Figure 3a-b), since Hi-C captures all DNA-DNA contacts. As previously noted<sup>3</sup> genes associated with broad domains were expressed in a more cell-specific manner than genes that are active yet not associated with broad domains in the same cell type (Fig. 2d, left panel). Similarly, gene targets of super enhancers were expressed in a more cell-specific manner than the gene targets of regular enhancers in the same cell type (Fig. 2d, right panel).

**Increased interaction frequency among broad domains and super enhancers.** We calculated the frequencies of interactions between all pairs of annotations (*i.e.*, broad domains, typical promoters, super enhancers, typical enhancers, and other annotations) and compared against theoretical expectations (Methods). These analyses showed that in Pol2 ChIA-PET networks, broad domains were more connected to all other nodes than theoretically expected (2.9 times more than expected) in all three cell lines (Fig. 3a, Supplementary Figure 1b).



**Figure 3.** Interactions between regulatory elements in Pol2 ChIA-PET Networks. **(a)** Enrichment of interactions between pairs of annotation classes in ChIA-PET networks. Colors represent log<sub>2</sub> ratio of observed over expected number of edges, where red represents enrichment of interactions and blue represents depletion of interactions. **(b)** (Top) Distribution of interactions within (turquoise) and across distinct (orange) super enhancer regions. (Bottom) Illustration of super enhancer nodes as defined by DNase-seq peaks within original super enhancer calls and the two different types of super enhancer interactions. **(c)** Connectivity degree distribution for different annotation classes. M, K, G represents MCF-7, K562, and GM12878, respectively. Broad domains and super enhancers are more connected on average than regular promoters and enhancers. **(d)** Example chromatin interaction networks around oncogene *EMP2*. (Left) Broad domain node associated with *EMP2* in MCF-7 is highly connected with super enhancers. (Right) Regular promoter node associated with *EMP2* is loosely connected in K562, with a single interaction with another promoter.

Furthermore, super enhancers interacted more frequently with broad domains (2.7–5.5 times more than expected) across the three cell types (Fig. 3a, Supplementary Figure 1b). Interestingly, super enhancer nodes also interacted more frequently among themselves (2.7–5 times more than expected), raising the possibility that distinct enhancer elements within a super enhancer region form highly interacting enhancer clusters in the 3D space. Indeed, further investigation of super enhancer-super enhancer interactions revealed that most of these (60–90%) take place within the same super enhancer region (Fig. 3b). We repeated these analyses after accounting for interactions within a single super enhancer region by representing the multiple nodes that belong to the same super enhancer domain as a single node (Methods). After this adjustment enrichment of interactions among super enhancer nodes were mostly lost (Supplementary Figure 4). Our analyses suggest that constituent enhancers within a super enhancer domain are in close proximity in the 3D chromatin space, however, these interactions do not typically span multiple distinct super enhancer domains. Finally, we noted an enrichment of interactions among promoter elements (both cell-specific and non-specific) (1.9–5.0 fold over expected, Fig. 3a). HiChIP, Hi-C, and CTCF ChIA-PET assays revealed similar interaction frequency patterns: i) high interactions between broad domains and super enhancers, ii) high interactions among constituent enhancers of super enhancer regions (Supplementary Figures 2c and 3c). Robustness of our results across assays and across cell types suggests a strong link between 3D configuration of the genome and distinct characteristics of regulatory elements. We summarized the characteristics of networks generated different assays in Table S1.

**Broad domains and super enhancers are hubs in chromatin interaction networks.** Network centrality measures suggest that cell-specific regulatory elements are more connected and exhibit hub-like connectivity in these networks in comparison to their typical counterparts (Fig. 3c, Supplementary Figure 5). On average, promoters were connected to 2.63, 4.21, and 3.56 other nodes in MCF-7, K562, and GM12878 Pol2 ChIA-PET networks respectively, whereas the corresponding values for broad domains were 5.03, 5.83, and 4.49 (one-sided Wilcoxon test p-values < 4.4.e-32 for all three cell lines). The increased connectivity of broad domain promoters taken together with their frequent interactions with super enhancers might be essential in maintaining their robust and increased gene expression patterns. On the other hand, enhancer nodes were connected to an average of 1.81, 1.90, 2.25 other nodes respectively in MCF-7, K562 and GM12878, whereas super enhancer nodes averaged 4.61, 3.27, and 4.22 connections respectively (one-sided Wilcoxon test p-values < 5.1e-114 for all comparisons). Our results also revealed a higher betweenness score for broad domains relative to non-specific promoters (Supplementary Figure 5) suggesting that broad domains act as connectors in the networks. For example, Fig. 3d shows Pol2 ChIA-PET network involving the *EMP2* oncogene that is upregulated in invasive breast cancer patients<sup>19</sup>. In the breast cancer cell line MCF-7, *EMP2* maps to a broad domain node and is connected to multiple super enhancers (Fig. 3d, left panel), which are also connected to each other. In contrast, in K562 where *EMP2* is an active yet non-specific promoter, this locus is connected differently and less densely (Fig. 3d, right panel). On the other hand, this locus was repressed in GM12878, and was not represented in the corresponding networks.

This example illustrates that connectivity of a locus in the 3D chromatin reflects the functional importance of that region in the cognate cell type. Our analyses revealed that cell-specific regulatory elements are connected more frequently in the 3D genome in comparison to their non-specific counterparts.

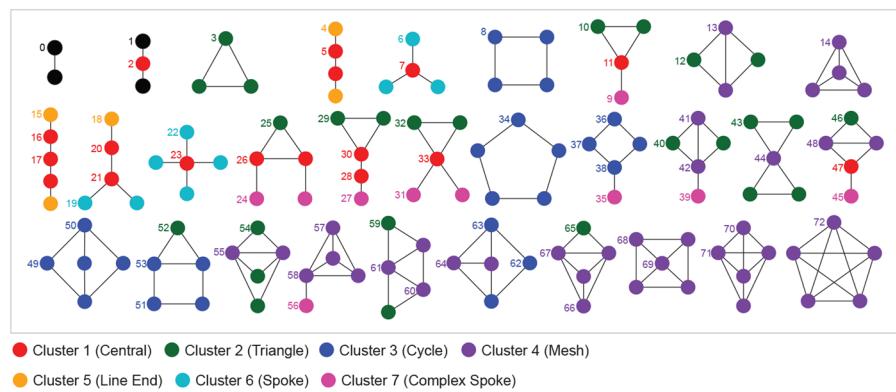
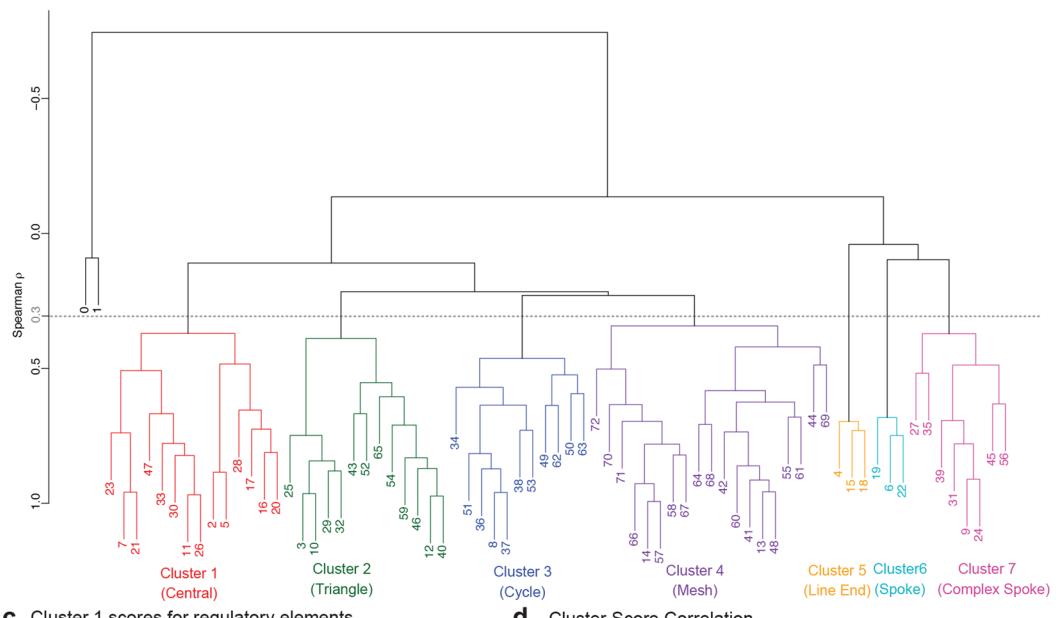
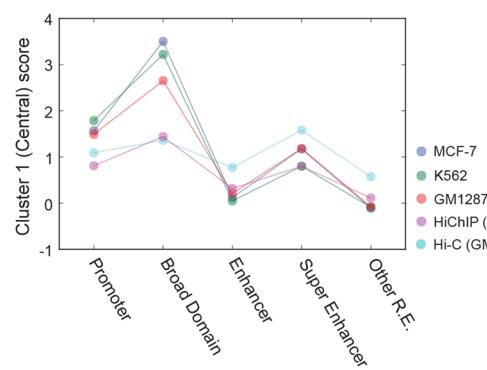
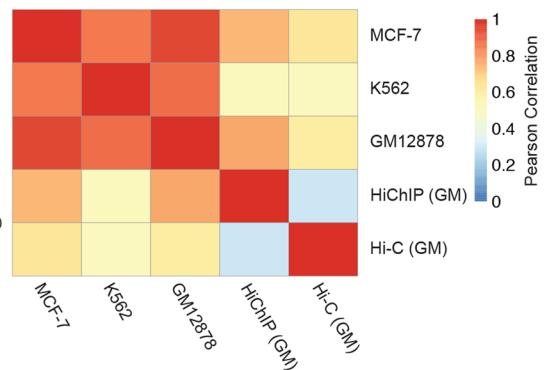
**Broad domains and super enhancers have unique connectivity patterns.** Chromatin interaction networks offer the opportunity to explore higher-level chromatin connectivity patterns (*i.e.*, network motifs) beyond immediate interactions. However, enumerating all possible configurations in a large network is computationally intractable (subgraph isomorphism problem). To effectively and systematically uncover chromatin interaction patterns associated with cell-specific regulatory elements, we utilized graphlets<sup>15</sup>. Graphlets are small, connected, and non-isomorphic (*i.e.*, topologically different) subnetworks within a large network that enables systematically studying and quantifying the local network structure around a node of interest. Topologically distinct nodes within a graphlet are referred to as orbits. We studied the local structure of chromatin interaction networks using all possible graphlets composed of two to five nodes, which encompasses 73 orbits (Fig. 4a). For each node an orbit signature vector is compiled by counting the number of times each node possessed the local structure of the 73 orbits. Orbit cluster scores of different functional elements showed very consistent patterns across cell types and across different assays (Fig. 4d), suggesting that these connectivity patterns are not stochastic and have functional relevance. These results suggest that cell type-specific regulatory elements have unique connectivity patterns and they tend to be central and tend to form tightly connected subnetworks in the 3D chromatin space.

**Network connectivity patterns are predictive of cell-specific activity.** To determine whether chromatin connectivity patterns of regulatory elements can be predictive of their cell-specific activity, we used machine-learning models based on support vector machines (SVM) (Fig. 5a). Each network node was represented using two types of data features: (1) network related and (2) genomic-data related and we built SVM-based<sup>16,17</sup> classification models to discriminate i) broad domains from regular promoters and ii) super enhancers from regular enhancers. The discriminative power of these models was quantified using receiver operating characteristic (ROC) curves and area under these curves (AUC), where a perfect predictor (*i.e.*, one that only identifies a ‘real’ broad domain as a broad domain) has an AUC score of 1.

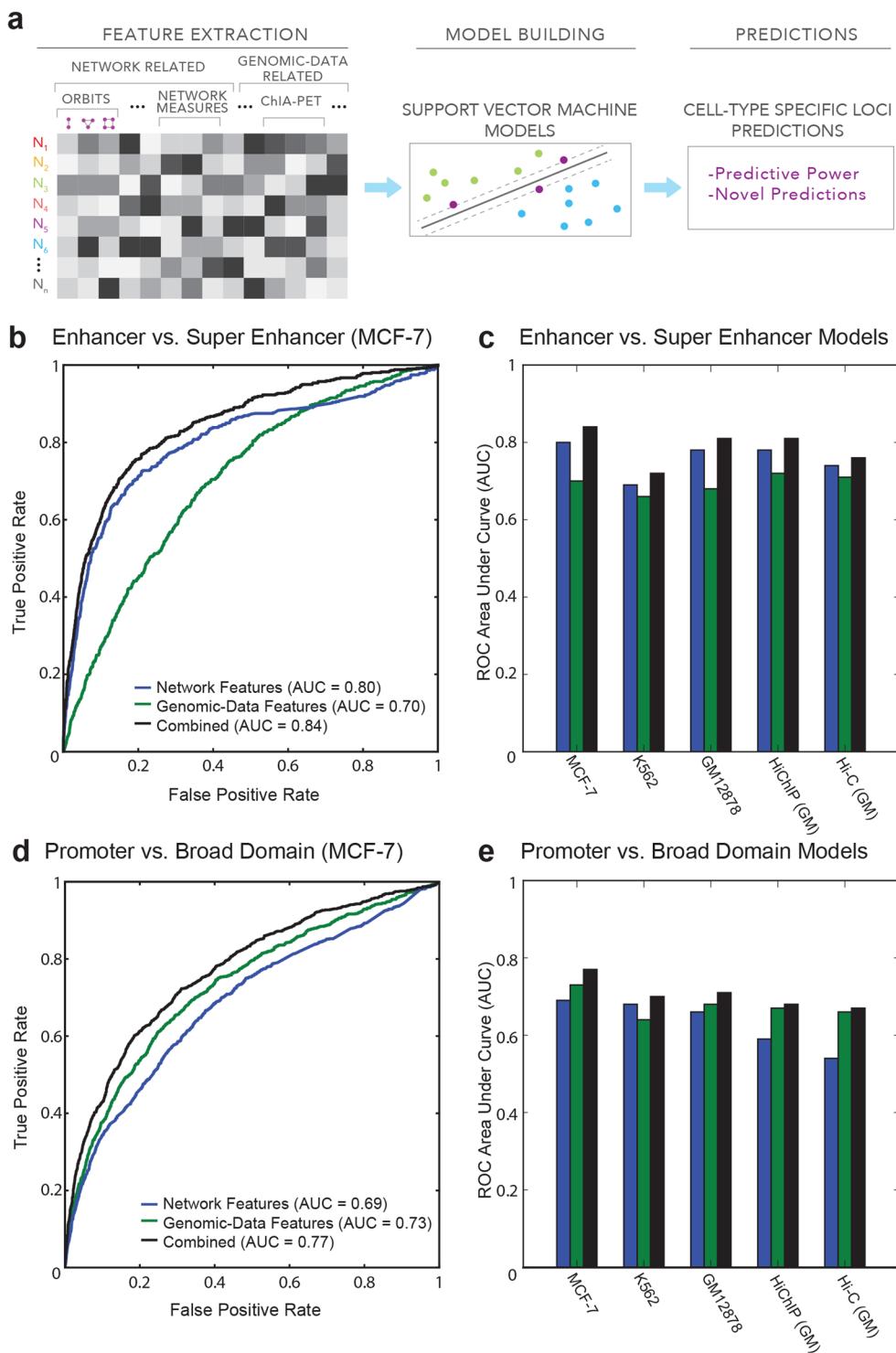
SVM models efficiently discriminated super enhancers from regular enhancers with high accuracy (Accuracy = 0.91, 0.74, 0.84 at 0.2 probability threshold, AUC = 0.84, 0.72, 0.81 in MCF-7, K562, and GM12878 Pol2 ChIA-PET networks respectively) (Fig. 5b-c, Supplementary Figure 7). Similarly, these models were also effective in discriminating broad domain promoters from regular promoters (AUC scores = 0.77, 0.70, 0.71 for MCF-7, K562, and GM12878 respectively) (Fig. 5d-e, Supplementary Figure 8). Furthermore, both analyses revealed that integration of network related features (*e.g.*, orbit cluster scores) with other genomic features improved the predictive ability of these models, suggesting that chromatin interaction networks and connectivity patterns harbor functional and non-redundant information. Prediction of super enhancers in HiChIP and Hi-C was as effective (AUC = 0.81 and 0.76 respectively) (Supplementary Figure 9). Similar results for broad domain classification models were achieved from Hi-C and HiChIP data (Supplementary Figure 10), although the relatively smaller impact of network features in these analyses suggests that Pol2-mediated interactions may be better suited for capturing network patterns associated with promoters. Precision-recall curves (Supplementary Figures 7–10) further emphasized the value of integrating network and genomic data features and the ability of these datasets to predict cell-specific regulatory elements.

The most predictive data features in these models were obtained using forward selection (Supplementary Figures 11 and 12), which uncovered consistent ranking of discriminatory features. For broad domain predictions, the two most predictive features were node size and network features related to their centrality (Supplementary Figure 11). Further investigation revealed that indeed broad domains are associated with expanded chromatin accessibility around their promoters (Supplementary Figure 13), in agreement with the expanded H3K4me3 deposition observed at these loci<sup>3</sup>. On the other hand, the most predictive features for super enhancers were related to their clique-like connectivity (*i.e.*, mesh) and high degree, reinforcing the importance of tight connectivity around super enhancers (Supplementary Figure 12). In summary, results from these analyses showed that network connectivity patterns of a regulatory element is predictive of its importance for regulating critical cellular functions in that cell type. Prediction results from our models can be found in Table S2.

**Enhancers targeting broad domains in interaction networks are enriched in disease-causing SNPs.** Previous studies have established that disruption of enhancer activity is associated with pathologies, particularly at cell-specific enhancers<sup>4,14</sup>. Here we investigate whether chromatin interaction networks can be utilized in further prioritizing loci associated to disease-causing genetic variants and in identifying their gene targets. To

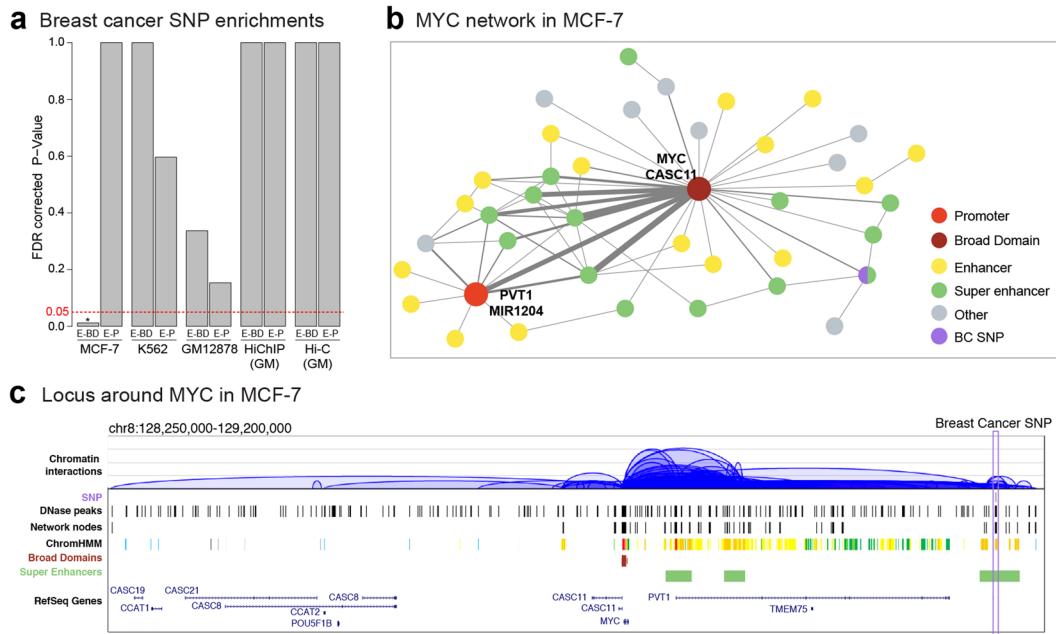
**a** Orbit and graphlets with clustering assignments**b** Hierarchical clustering of orbits**c** Cluster 1 scores for regulatory elements**d** Cluster Score Correlation

**Figure 4.** Cell-specific regulatory elements are associated with distinct network patterns. **(a)** Representation of 73 orbits used to drive orbit cluster signatures. Orbit nodes are color-coded with respect to their cluster assignments from **(b)**. For example red nodes represent central nodes that are in cluster 1. **(b)** Hierarchical clustering of 73 orbits after pooling data from all Pol2 ChIA-PET networks. Seven distinct clusters are identified based on the topology of network graphlets. **(c)** Trimmed mean Cluster-1 score (central nodes) for different annotations in ChIA-PET, Hi-ChIP, and Hi-C networks. Broad domain nodes have the highest C1 score in these networks. **(d)** Pairwise correlations (Pearson's coefficient) of 7 cluster scores computed over each annotation between all network pairs. Note the high correlation across different assays and cell types (except for Hi-C), suggesting that network connectivity patterns are non-stochastic.



**Figure 5.** Network connectivity patterns are predictive of cell-specific activity. **(a)** Summary of our machine-learning framework. **(b)** Receiver Operating Characteristic (ROC) curves for SVM models discriminating enhancers from super enhancers in MCF-7. **(c)** Receiver Operating Characteristic (ROC) area under the curve (AUC) values for enhancer vs. super enhancer predictions over all networks. **(d)** ROC curves for SVM models separating promoters from broad domains in MCF-7. **(e)** Receiver Operating Characteristic (ROC) area under the curve (AUC) values for promoter vs. broad domain prediction over all networks.

test this, we calculated the enrichment of NHGRI GWAS SNPs<sup>20</sup> in enhancers targeting regular promoters and enhancers targeting broad domains. Our analyses revealed that in MCF-7, enhancers (regular and super) targeting broad domains were significantly enriched in breast-cancer associated SNPs (Fig. 6a, Benjamini-Hochberg adjusted p-value = 0.012). This includes 7 out of 1757 enhancer/super enhancer nodes harboring breast cancer



**Figure 6.** Enhancers targeting broad domains are enriched in variants associated to the diseases of the cognate cell type. (a) Enrichment p-values for breast-cancer SNPs at enhancers targeting promoters (E-P) and enhancers targeting broad domain promoters (E-BD) in different cell types. Breast-cancer associated SNPs are significantly enriched in enhancers targeting BDs in MCF-7 breast cancer cell line. (b) Example MCF-7 network representing breast-cancer associated SNP rs1121946 in a super enhancer node targeting the *MYC* broad domain promoter. (c) Genome browser representation of the locus around *MYC*. Note that network representation is effective in detecting connectivity patterns.

SNPs (Table S3), one of which is a super enhancer and targets the promoter of the *MYC* oncogene (Fig. 6b-c) that is constitutively expressed in many cancers, including breast and blood cancers<sup>21</sup>. In breast cancers, *MYC* de-regulation is associated with breast cancer development and poor prognosis<sup>22</sup>. The locus around *MYC* is active and represented in the chromatin interaction networks in all three cell types, however connectivity around this locus was different in these cells. In the MCF-7 network, *MYC* is a broad domain node and connected to multiple enhancer and super enhancer nodes, one of which harbors rs1121946 SNP that is in high linkage disequilibrium (LD) with index SNP rs11780156 ( $r^2 > 0.7$ ), a known breast-cancer associated variant (Fig. 6b-c). Our network representation revealed that although this SNP is within an enhancer domain that is 414 kb away from the *MYC* promoter, it is directly connected to *MYC*'s promoter in MCF-7 networks, emphasizing the importance of studying disease-associated SNPs using 3D chromatin interaction networks. Similar to the enrichment of breast cancer associated SNPs in MCF-7, we observed that in GM12878, enhancers targeting broad domains were enriched in SNPs associated with immune-related disorders, such as Systemic Lupus and Crohn's disease in comparison to enhancers targeting regular promoters (Table S4). This analysis revealed the utility of chromatin interaction networks to uncover and prioritize non-coding loci associated with pathologies as well as their gene targets.

## Discussion

In this study, we showed that broad domains and super enhancers have distinct connectivity patterns in chromatin interaction networks that are conserved across cell types and can be captured using diverse assays. Broad domains tend to be central nodes in these networks and are frequently targeted by multiple constituent enhancers that belong to a super enhancer domain. A recent study employed CRISPR/Cas9 to delete enhancers and super enhancers in mouse embryonic stem cells (mESCs) and suggested that super enhancer clusters act in a redundant manner to fine tune the expression of their target genes<sup>23</sup>. Our observations from chromatin interaction networks provide an explanation for this redundancy. We observed high connectivity among super enhancer elements and high connectivity between broad domains and constituent super enhancers. This tightened connectivity among cell-specific regulatory elements ensures that the activity of super enhancer target genes (likely to be a broad domain) is robust to the disruption of any single enhancer element within the enhancer cluster. Such a tightened connectivity pattern might also be critical in establishing and maintaining robust (i.e., low variation) expression patterns associated with broad domain genes<sup>3</sup>. Our machine learning models showed that connectivity patterns of super enhancers and broad domains are predictive of their cell-specific nature.

Recent studies revealed that regulatory elements with frequent chromatin interactions (i.e., hubs) are enriched in super enhancers and harbor more GWAS SNPs<sup>6,24,25</sup>. Our study is in alignment with these findings where we observed increased connectivity for super enhancer nodes in comparison to regular enhancers using three different assays (ChIA-PET, Hi-C, and HiChIP) and in three different cell types (K562, MCF-7, GM12878). Our study furthered our knowledge on how super enhancers are connected in these networks and revealed that super

enhancers form clique-like structures and typically connect to broad domains and other super enhancer elements within the same domain. Broad domains were associated with more chromatin interactions compared to typical H3K4me3 domains<sup>6</sup>. Our study established that in comparison to regular promoters, broad domains had more interactions overall. Moreover, these interactions were especially connecting them to super enhancers. Increased targeting of broad domains by super enhancers likely ensures robust and increased expression of these important genes in cognate cells. Furthermore, we showed that enhancers targeting broad domains in chromatin interaction networks harbor disease-associated SNPs, further reinforcing the importance of studying these genes and their regulators in the context of human diseases.

A challenge we faced in our analyses was the difficulty of representing super enhancers that span long genomic regions (10–20 kb) in our networks, which would harbor more interactions merely due to their genomic coverage. We overcame this challenge by defining the network nodes using open chromatin regions (i.e., DHS sites), which represent putative active regulatory elements. This methodology also enabled the study of individual enhancer elements within a super enhancer domain. We observed that a super enhancer element typically is not a single and expanded regulatory element but a combination of constituent active enhancers that are in close proximity to each other in both linear and 3D space, and they frequently interact with one another and with their target gene to regulate that gene's expression levels (see examples of such networks in Figs 3d, and 6b). Our analyses also showed that although members of a super enhancer domain heavily interact among themselves, these enhancers typically do not interact with other super enhancer domains.

In conclusion, our study uncovered the unique chromatin interaction patterns around loci that are critical for cellular functions and disease etiology by taking advantage of network theory and machine learning models. By building interaction networks we were able to study complex network patterns (such as graphlets) and to systematically quantify differences in the connectivity of different regulatory elements. Our findings revealed that chromatin connectivity patterns around super enhancers and broad domains are non-stochastic and conserved across cell types and can be captured via different assays. However, we acknowledge the caveat that genome-wide chromatin interaction maps analyzed here are generated from millions of cells. Therefore it is not possible to dissect whether the connectivity patterns observed for super enhancers and broad domains take place in individual cells. Advances in single-cell chromatin interaction profiling techniques will be essential in studying these patterns at the single cell level. As a step towards this direction, Hi-ChIP<sup>12</sup> a recent technique significantly reduced the input material required to profile protein-mediated chromatin interactions: a 100 fold decrease from 100 million cells to 1 million cells. Similarly, recent developments in single-cell Hi-C profiling techniques open the doors to studying cell-to-cell-heterogeneity for DNA-DNA interactions<sup>26</sup>. With these technological advances, we expect more chromatin interaction data to be generated from human cells in healthy and disease states at the single-cell resolution in the near future. Advanced computational methods we present in our study will be critical in furthering our understanding on how chromatin interactions might relate to establishing and maintaining critical cellular functions and how changes in these interactions might be associated with pathologies.

## Methods

**ChIA-PET data analyses and network construction.** Pol2 and CTCF ChIA-PET chromatin interactions called using ChIA-PET Tool<sup>27</sup> were obtained from<sup>28,29</sup>. Accession numbers for these datasets are as follows: MCF-7 (GSE39495), K562 (GSE39495), and GM12878 (GSE72816). ChIA-PET Tool<sup>27</sup> interactions were preferred over the alternative Mango<sup>30</sup>, since calls obtained via Mango were sparse and did not include as many broad domain or super enhancer interactions (Table S1). To minimize the number of false positive interactions, we instead filtered interaction calls using QuIN<sup>31</sup>, by selecting only those having both anchors overlapping DNase I hypersensitive sites (DHS) defined from DNase-seq open chromatin peaks, which reduces false positive calls and likely captures active regulatory loci. These peaks were then used to define the nodes of the interaction networks, which were constructed using QuIN<sup>31</sup>. MCF-7, K562, and GM12878 open chromatin peaks were called using MACS2 software<sup>32</sup> (version 2.1) after pooling replicates (GSE32970 and GSE29692<sup>28</sup>). For network building, we used 250 bp extension and removed inter-chromosome interactions and edges spanning distances greater than 1 Mb. Network nodes were first annotated using ChromHMM states<sup>13,14</sup>. If a node overlapped with multiple ChromHMM annotations, we applied the following priority schema to dissolve ambiguities in annotations due to the genomic segmentation framework in ChromHMM: 1) enhancers/promoters 2) insulators 3) poised promoters 4) repressed elements 5) transcribed elements 6) low signal. If a node was annotated both as a promoter and an enhancer, known transcription start site (TSS) definitions were used to define promoter nodes found within 2 kb of a known TSS. Broad domains and super enhancers annotations were then assigned respectively to promoter and enhancer nodes using the previously defined broad domain<sup>3</sup> and super enhancer<sup>2</sup> regions in corresponding cells. If a broad domain was found to overlap multiple promoter nodes, the promoter node with the largest overlap with the broad domain in terms of base pairs (bps) was assigned as the broad domain node. In the case of ties, both promoters were labeled as broad domain nodes.

**Processing and analyses of Hi-C and HiChIP data.** We called interaction pairs from HiChIP<sup>12</sup> (GSE80820) (targeting cohesion subunit *Smc1a*) and Hi-C<sup>18</sup> (GSE63525) data in GM12878 cell line. Valid interaction pairs from HiChIP biological and technical replicates were pooled and filtered by extending 250 bp in both directions, keeping only the pairs with both ends overlapping a DNase-seq peak. Significant HiChIP interactions between peaks were called based on the hyper-geometric distribution as described in<sup>27</sup> and filtered using the Benjamini-Hochberg procedure ( $FDR < 0.05$ ). Significant Hi-C intra-chromosomal interactions at 1 kb resolution were identified using in-house software implemented based on the HiCCUPS method<sup>18</sup>. For each contact, expected values based on donut, vertical, horizontal and lower left filters were calculated using parameters  $P = 20$  and  $W = 40$  to calculate a P-value based on the Poisson distribution and filtered based on the Benjamini-Hochberg procedure ( $FDR < 0.025$ ). A more stringent FDR cutoff is used for Hi-C data to make it

more comparable with other assays, since Hi-C data were more deeply sequenced than the others. The intersection of these four filters was used to identify the final set of contacts. Network construction was performed similar to ChIA-PET data using 0 bp and 1,250 bp extension parameters for HiChIP and Hi-C networks, respectively. These values were chosen to account for differences in read extension introduced in interaction calling steps. HiChIP edges were further filtered to remove edges with less than 4 supporting valid pairs.

**Processing and analyses of RNA-seq data for cell-type specificity.** RNA-seq data was obtained from the ENCODE project repository for all available cell lines and adult tissues<sup>28,33</sup>. These data were filtered to exclude samples that were generated under an experimental treatment or audited for quality concerns. This resulted in 23 cell and tissues types (including GM12878, K562, and MCF-7) that were used to quantify the degree of cell-specificity of gene expression levels. For this, we calculated the cell-type specificity measure (SPM)<sup>34</sup> for each gene in the three cell types. SPM was used to quantify cell-specificity of genes associated to regular and broad domain promoters, as well as target genes of regular and super enhancers. For super enhancers and enhancers, we considered both nearest TSS definitions (up to 10kb) as well as promoter targets identified from ChIA-PET networks. Network targets were chosen by selecting promoter/broad domain targets of up to 4 edges using breadth first search between the enhancer and promoter node, stopping at the first promoter node found. If a network node was associated with multiple genes, the average SPM measure of all associated genes was used.

**Interaction frequency between functional annotations.** We calculated the enrichment of interactions between every pair of annotations (e.g., promoter versus enhancer nodes) by calculating a theoretical expectation of interaction frequency and by comparing against the observed interaction frequency for this annotation pair. More formally, expected interaction frequency between all nodes annotated as “A” and all nodes annotated as “B” is calculated as:

$$\text{Expected}(A, B) = \sum_{C_i} |E_{C_i}| \frac{\sum_{a_j \in C} \sum_{b_j \in C_i} r(\text{gd}(a_j, b_j))}{\sum_{n_j \in C_i} \sum_{n_k \in C_i, j < k} r(\text{gd}(n_j, n_k))} \quad r(x) = \begin{cases} 1 & \text{if } x = 1000000 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where  $C_i$  represents the set of all nodes in chromosome  $i$ ,  $|E_{C_i}|$  represent the number of observed edges within the node subset  $C_i$ , and  $gd(x, y)$  is the genomic distance from node  $x$  to node  $y$ . The expected frequency obtained corrects for both analyzing only intra-chromosomal interactions as well as the limitation that connected nodes within the network are within 1 Mb. For nodes with the same annotation, the expected value was adjusted accordingly to account for two nodes being selected from the same set.

To study whether super enhancer interactions are restricted within a domain, we generated interactions networks again by merging super enhancer nodes that belong to the same super enhancer domain and representing the whole domain as a single node in the network. Then we counted the number of edges that are connecting these merged nodes with the rest of the nodes in the network and repeated the interaction enrichment analyses.

**Centrality measures.** Network metrics were calculated using QuIN<sup>31</sup>. Connectivity degree refers to the number of edges connected to a node. Closeness centrality was computed as:

$$\text{Closeness}(v) = \frac{1}{\sum_{u \in N_c} \text{nd}(u, v)} \quad (2)$$

where  $\text{nd}$  is the network distance (*i.e.*, number of edges) between nodes  $u$  and  $v$ , and  $N_c$  represents the set of nodes within the connected component containing  $v$ . A connected component is defined as a collection of nodes where every node pair is connected to each other through some path (*i.e.*, a sequence of adjacent edges) such that nodes are not connected to other nodes outside this collection. Harmonic centrality was computed as follows:

$$\text{Harmonic}(v) = \sum_{u \in N_c, u \neq v} \frac{1}{\text{nd}(u, v)} \quad (3)$$

Finally, betweenness centrality was calculated as:

$$\text{Betweenness}(v) = \sum_{u \neq v \neq x \in N_c} \frac{|\text{sp}(u, x, v)|}{|\text{sp}(u, x, x)|} \quad (4)$$

where  $\text{sp}(u, x, v)$  denotes all shortest paths between nodes  $u$  and  $x$  that include node  $v$ . Normalized centrality measures of a node were calculated with respect to the size of the network component  $|N_c|$ .

Comparisons of connectivity degree measures of promoters to broad domains and super enhancer to enhancers were performed using the Wilcoxon rank sum test. The Wilcoxon rank sum test is a non-parametric test that does not assume normally distributed measurements. Corresponding  $n$ -values for each annotation subset and network are reported in Table S1.

**Graphlet and orbit scores.** Graphlets are small, connected, and non-isomorphic (*i.e.*, topologically distinct) subnetworks components that can be used to decompose and describe a large network<sup>15</sup>. Topologically distinct nodes of a graphlet are referred to as orbits. There exist 73 possible orbits for all graphlets of size 2–5, which were counted using the Cytoscape plugin GraphletCounter<sup>35</sup>. For each ChIA-PET network node, a vector of 73 orbit counts was obtained. The proportion of each orbit with respect to all 73 orbits counted was calculated

to normalize each vector. To identify orbits that are topologically similar, we performed hierarchical clustering based on Spearman rank correlation coefficient on orbit scores of ChIA-PET network nodes. Seven orbit clusters were obtained at a Spearman cutoff of 0.3 (Fig. 4b) and visually interpreted and labeled based on their connectivity patterns after excluding orbits 0 and 1, which were present and have similar scores for all nodes. Finally, scores for each orbit clusters is calculated as follows:

$$CS_{ki} = \max_{j \in C_k} \left( \frac{O_{ij} - \mu_j}{\sigma_j} \right) \quad (5)$$

where  $CS_{ki}$  denotes the score of node  $i$  for cluster  $k$ ,  $O_{ij}$  denotes the orbit score for node  $i$  and orbit  $j$ ,  $C_k$  denotes the set of orbits in cluster  $k$ , and  $\mu_j$  and  $\sigma_j$  represent the mean and standard deviation of orbit  $j$ 's score over all nodes.

**Support Vector Machine models.** We used support vector machines (SVMs) due to their efficacy in handling high dimensional and noisy data<sup>36</sup>, which is the case in genomics data. SVMs are supervised learning algorithms that effectively handle classification problems<sup>37</sup> by mapping each example as a point in a higher dimensional data space and identifying a hyperplane that separates classes by the widest margin. To build SVM models, we extracted two types of data features we extracted for each network node: (1) network related and (2) genomic-data related. We extracted sixteen network related features, which included network measures (both raw and normalized by component size) ( $n=8$ ), connected component size, and graphlet cluster scores ( $n=7$ ). For genomic-data related features, we extracted and used seven features: open chromatin peak length in bps (equivalent to node size in our networks), the ratio of a node's size to the average size of all nodes connected to it, the average PET of a node's edges, and gene definition-related features ( $n=4$ ), which include the distance to the closest upstream and downstream TSS and the direction of transcription for these two genes. SVM<sup>16,17</sup> models were trained with a radial basis function (RBF) kernel using scikit-learn Python libraries<sup>38</sup>. Grid search was used to tune hyper-parameters of the SVM models ( $C$  and  $gamma$ ). For feature ranking, we implemented forward feature selection using a greedy algorithm that implemented a stepwise addition of features to the model, keeping the best-performing feature in the model for each round until all features were included. Matthews correlation coefficient (MCC) was used as a performance measure both during hyper-parameter tuning and feature selection analyses, since this measure is effective for unbalanced class sizes (i.e., 500–1400 super enhancer nodes compared to 7000–13000 enhancer nodes).

**GWAS SNP enrichments.** We tested for enrichment of GWAS SNPs in enhancer loci targeting regular and broad domain promoters using GREGOR<sup>39</sup>. Index SNPs from the GWAS catalogue<sup>20</sup> were used for enrichment analyses ( $n=657$  phenotypes) along with the SNPs that are in high linkage disequilibrium (LD) with the index SNPs ( $R^2 >= 0.7$ ). Benjamini-Hochberg FDR method was used for multiple hypothesis correction of enrichment p-values.

**Availability of data and methods.** Datasets and scripts are available on the Ucar Lab github website: <https://github.com/UcarLab/CellSpecificNetworks>.

## References

- Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319, <https://doi.org/10.1016/j.cell.2013.03.035> (2013).
- Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947, <https://doi.org/10.1016/j.cell.2013.09.053> (2013).
- Benayoun, B. A. *et al.* H3K4me3 breadth is linked to cell identity and transcriptional consistency. *Cell* **158**, 673–688, <https://doi.org/10.1016/j.cell.2014.06.027> (2014).
- Chen, K. *et al.* Broad H3K4me3 is associated with increased transcription elongation and enhancer activity at tumor-suppressor genes. *Nature genetics* **47**, 1149–1157, <https://doi.org/10.1038/ng.3385> (2015).
- Loven, J. *et al.* Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**, 320–334, <https://doi.org/10.1016/j.cell.2013.03.036> (2013).
- Cao, F. *et al.* Super-Enhancers and Broad H3K4me3 Domains Form Complex Gene Regulatory Circuits Involving Chromatin Interactions. *Scientific Reports* **7**, <https://doi.org/10.1038/s41598-017-02257-3> (2017).
- Zhang, Y. *et al.* Chromatin connectivity maps reveal dynamic promoter–enhancer long-range associations. *Nature* **504**, 306–310, <https://doi.org/10.1038/nature12716> (2013).
- Fullwood, M. J. *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**, 58–64, <https://doi.org/10.1038/nature08497> (2009).
- Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293, <https://doi.org/10.1126/science.1181369> (2009).
- Nagano, T. *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59–64 (2013).
- Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84–98, <https://doi.org/10.1016/j.cell.2011.12.014> (2012).
- Mumbach, M. R. *et al.* HiChIP: Efficient and sensitive analysis of protein-directed genome architecture. *Nature Methods* (2016).
- Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**, 215–216, <https://doi.org/10.1038/nmeth.1906> (2012).
- Parker, S. C. *et al.* Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 17921–17926, <https://doi.org/10.1073/pnas.1317023110> (2013).
- Pržulj, N. Biological network comparison using graphlet degree distribution. *Bioinformatics* **23**, e177–e183 (2007).
- Boser, B. E., Guyon, I. M. & Vapnik, V. N. In *Proceedings of the fifth annual workshop on Computational learning theory*. 144–152 (ACM).
- Cortes, C. & Vapnik, V. Support-vector networks. *Machine Learning* **20**, 273–297, <https://doi.org/10.1007/bf00994018> (1995).

18. Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680, <https://doi.org/10.1016/j.cell.2014.11.021> (2014).
19. Fu, M. *et al.* Rationale and preclinical efficacy of a novel anti-EMP2 antibody for the treatment of invasive breast cancer. *Molecular cancer therapeutics* **13**, 902–915, <https://doi.org/10.1158/1535-7163.MCT-13-0199> (2014).
20. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research* **42**, D1001–1006, <https://doi.org/10.1093/nar/gkt1229> (2014).
21. Dang, C. V. MYC on the path to cancer. *Cell* **149**, 22–35, <https://doi.org/10.1016/j.cell.2012.03.003> (2012).
22. Xu, J., Chen, Y. & Olopade, O. I. MYC and Breast Cancer. *Genes & cancer* **1**, 629–640, <https://doi.org/10.1177/1947601910378691> (2010).
23. Moorthy, S. D. *et al.* Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. *Genome Research* **27**, 246–258, <https://doi.org/10.1101/gr.210930.116> (2017).
24. Huang, J., Marco, E., Pinello, L. & Yuan, G.-C. Predicting chromatin organization using histone marks. *Genome Biology* **16**, doi:<https://doi.org/10.1186/s13059-015-0740-z> (2015).
25. Schmitt, A. D. *et al.* A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Reports* **17**, 2042–2059, <https://doi.org/10.1016/j.celrep.2016.10.061> (2016).
26. Ramani, V. *et al.* Massively multiplex single-cell Hi-C. *Nature Methods* **14**, 263–266, <https://doi.org/10.1038/nmeth.4155> (2017).
27. Li, G. *et al.* ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome biology* **11**, 1 (2010).
28. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74, <https://doi.org/10.1038/nature11247> (2012).
29. Tang, Z. *et al.* CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell* **163**, 1611–1627, <https://doi.org/10.1016/j.cell.2015.11.024> (2015).
30. Phanstiel, D. H., Boyle, A. P., Heidari, N. & Snyder, M. P. Mango: a bias-correcting ChIA-PET analysis pipeline. *Bioinformatics* **31**, 3092–3098, <https://doi.org/10.1093/bioinformatics/btv336> (2015).
31. Thibodeau, A. *et al.* QuIN: A Web Server for Querying and Visualizing Chromatin Interaction Networks. *PLoS computational biology* **12**, e1004809, <https://doi.org/10.1371/journal.pcbi.1004809> (2016).
32. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**, R137 (2008).
33. Sloan, C. A. *et al.* ENCODE data at the ENCODE portal. *Nucleic acids research* **44**, D726–732, <https://doi.org/10.1093/nar/gkv1160> (2016).
34. Xiao, S. J., Zhang, C., Zou, Q. & Ji, Z. L. TiSGeD: a database for tissue-specific genes. *Bioinformatics* **26**, 1273–1275, <https://doi.org/10.1093/bioinformatics/btq109> (2010).
35. Whelan, C. & Sonmez, K. Computing graphlet signatures of network nodes and motifs in Cytoscape with GraphletCounter. *Bioinformatics* **28**, 290–291, <https://doi.org/10.1093/bioinformatics/btr637> (2012).
36. Noble, W. S. in *Kernel Methods in Computational Biology* (eds B. Schölkopf, K. Tsuda, & JP. Vert) 71–92 (MIT Press, 2004).
37. Byvatov, E. & Schneider, G. Support vector machine applications in bioinformatics. *Applied bioinformatics* **2**, 67–77 (2003).
38. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
39. Schmidt, E. M. *et al.* Gregor: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. *Bioinformatics* **31**, 2601–2606 (2015).

## Acknowledgements

We thank Jane Cha for aid in graphic design for Figures 1 and 5a. We thank members of Ucar and Stitzel labs for critical feedback during the progress of the study. This study was made possible by generous financial support of the National Institute of General Medical Sciences (NIGMS) under award number GM124922 (to DU) and by The Jackson Laboratory startup funds (to DU). Opinions, interpretations, conclusions, and recommendations are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health (NIH).

## Author Contributions

D.U. and P.V.L. conceived the study and data analysis strategies. A.T. performed analyses of all corresponding data and developed the methods. A.T. and D.U. wrote the paper. E.J.M. helped with data analyses. D.G.S. provided advice on data analyses. All authors edited the manuscript and approved it prior to submission.

## Additional Information

**Supplementary information** accompanies this paper at <https://doi.org/10.1038/s41598-017-14389-7>.

**Competing Interests:** The authors declare that they have no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017