NAR Genomics and Bioinformatics, 2024, **6**, lqae136
https://doi.org/10.1093/nargab/lqae136
Advance access publication date: 3 October 2024
Standard Article

Genome wide clustering on integrated chromatin states and Micro-C contacts reveals chromatin interaction signatures

Corinne E. Sexton[1,2], Sylvia Victor Paul[1,2], Dylan Barth[1,2] and Mira V. Han[1,2,*]

[1]School of Life Sciences, University of Nevada, Las Vegas, NV 89154, USA
[2]Nevada Institute of Personalized Medicine, University of Nevada, Las Vegas, NV 89154, USA

[*]To whom correspondence should be addressed. Tel: +1 702 774 1503; Fax: +1 702 895 3956; Email: mira.han@unlv.edu
Present address: Corinne E. Sexton, Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA.

Abstract

We can now analyze 3D physical interactions of chromatin regions with chromatin conformation capture technologies, in addition to the 1D chromatin state annotations, but methods to integrate this information are lacking. We propose a method to integrate the chromatin state of interacting regions into a vector representation through the contact-weighted sum of chromatin states. Unsupervised clustering on integrated chromatin states and Micro-C contacts reveals common patterns of chromatin interaction signatures. This provides an integrated view of the complex dynamics of concurrent change occurring in chromatin state and in chromatin interaction, adding another layer of annotation beyond chromatin state or Hi-C contact separately.

Introduction

Chromatin states are regulatory annotations defined by patterns of chromatin marks in genomic regions (1). These patterns have elucidated regulatory roles of the noncoding genome (2). ChromHMM and Segway softwares use a Hidden Markov model and dynamic Bayesian network respectively to infer 'hidden' chromatin states based on the combination of various epigenomic marks (3–5). This approach has been further extended to integrate additional molecular information in the model such as RNA-Seq, transcription factor binding and even 3D Hi-C data (6–10). Each approach has in common the goal of annotating regions of the genome with interpretable clusters to be used in subsequent studies of gene regulation, genetic association and others.

Of particular interest to this research is the integration of Hi-C data in the state inference framework. Hi-C sequencing captures the conformation of chromatin in the cell and this folding of DNA is an integral element of gene regulation. Though it is still of debate whether chromatin conformation is a cause or consequence of gene expression (11), the canonical model assumes that distal enhancers bound by transcription factors require contact with promoters through DNA folding to influence transcription (12).

A challenge to integrating Hi-C in the previously stated chromatin state framework is the interacting nature of the data: a Hi-C contact involves two genomic regions, whereas chromatin mark ChIP-seq defines peaks in one region alone. To incorporate interaction into a feature space defined for each region, one has to summarize both the strength of the interaction that the region has with its multiple contacts, as well as the characteristics of each region in contact. A few softwares have attempted to integrate Hi-C in hidden state models in the following ways.

Segway-GBR (6) and SPIN (7) both use integrative methods which encourage contacting regions to be clustered within the same state. This is based on the phenomena that large regions with similar chromatin marks are often in contact (13–15). SPIN uses a Hidden Markov random field using a resolution of 25 kb windows. Segway-GBR assigns a pairwise prior to significant Hi-C contacts at 10 kb windows. Shokraneh *et al.* (10) employ graph embedding to learn structural vector features of Hi-C data which were then passed on to an HMM along with chromatin marks to be segmented together. The resulting combinatorial domains recapitulated known Hi-C sub-compartment categories (14) and provided additional granularity to sub-compartments. But, since the segments resulting from combined information tended to be broader than traditional chromatin-based segments, it didn't perform as well as the traditional methods in explaining gene expression (10).

Beyond these few examples, to our knowledge, no method exists to broadly investigate both chromatin interactions and chromatin states across the entire genome. The closest studies were the recent studies that used deep learning approaches that integrate chromatin interactions and chromatin states to predict gene expression (15,16), but since these studies were framed as classification or regression studies, they were not interested in unsupervised discovery of the patterns of interaction between different chromatin states at a genome-wide scale. We propose a new method for integrating chromatin

Received: February 22, 2024. **Revised:** August 21, 2024. **Editorial Decision:** September 19, 2024. **Accepted:** September 20, 2024

interaction data into a traditional framework for chromatin state annotation. Using this approach, we can apply genome-wide clustering to annotate patterns of chromatin interaction. By summarizing all contacts at each individual genomic region of interest, we capture a chromatin interaction signature (CIS) of both the summed strength of contacts as well as the chromatin states of those contacts. These CISs provide an additional layer of genome-wide annotation beyond contact or chromatin state alone. Applying this approach to Micro-C data (at 1 kb window size) (17,18) rather than lower resolution Hi-C, we were able to annotate chromatin interaction signatures for chromatin state segments that span smaller regions, such as enhancers or promoters, rather than looking at broad compartment activity.

Materials and methods

Data sources

Chromatin states generated by ChromHMM were obtained from EpiMap for HFF, endoderm, H1-hESC and HeLa cells for hg38 (19). We used the 18-state Roadmap model, re-generated by EpiMap (19) from H3K27ac, H3K4me1, H3K4me3, H3K36me3, H3K9me3 and H3K27me3 chromatin marks. We refer to this data as ChromHMM18. To investigate the effect of chromatin annotation variation, in addition to the 18-state EpiMap model we used as our main annotation, we also used the 15-state ChromHMM model originally generated by Roadmap epigenomics (20) (ChromHMM15), and the Segway annotations (5). Segway annotations were downloaded from https://noble.gs.washington.edu/proj/encyclopedia/interpreted/ (21). To make the labels comparable across cell types, we translated from the 33 Segway labels down to the 8 Segway labels that were available for endoderm and HFF (Supplementary Table S1).

Micro-C mcool files were downloaded from the 4dnucleome.org and were originally from Oksuz *et al.* (22). The Formaldehyde + DSG protocol was used for each of these Micro-C datasets. The HeLa cell data used was non synchronized as opposed to the cell cycle synchronized HeLa Micro-C data available.

A common issue with integrating chromatin mark data and Hi-C is the issue of window size. In general Hi-C data is interpreted in >5 kb windows whereas ChIP-Seq peaks are much smaller. For example, the smallest ChromHMM segmented regions are 200 bp. To remedy this, we take advantage of available Micro-C data. Micro-C data is a derivative of Hi-C which uses a micrococcal nuclease (MNase) to create much smaller windows sizes (18,23). These high resolution fragments accurately capture fine scale interactions rather than the broader scale of traditional Hi-C enabling us to use 1 kb windows for contact data.

Creation of data matrix

Through a process we named sum of chromatin state by contact (SCC), we create a matrix of Micro-C scores summed by chromatin state. The mathematical definition of the SCC matrix has been described in the results, here we describe the algorithm. Using Cooler (24), all Micro-C normalized weighted contacts were pulled 2 Mb upstream and downstream for each non-overlapping 1 kb window which contained an annotated chromatin state (omitting quiescent regions for feasibility purposes). Next, for each base region with a chromatin state, all regions in contact with the base were assigned a chromatin state based on results from bedtools (25) intersect between ChromHMM segments and Micro-C bedpe files. In 1 kb windows where multiple chromatin states are present, each is included as a separate entry. If a ChromHMM segment overlapped multiple windows, the mean Micro-C contact score across the windows was used.

Finally, the scores are summed for each unique chromatin state, resulting in a matrix where each chromatin state annotated base region (row) has 17 scores (columns), one for each chromatin state except for quiescent regions. Quiescent regions were excluded for this analysis because they signify regions with low chromatin marks. For clarity, the score is the summed amount of KR normalized contact the base region has with a particular chromatin state annotated region. All code for this process can be found at https://doi.org/10.6084/m9.figshare.25270645.

Comparison across annotations.

To compare the effect of different annotations, we generated independent SCC matrices based on ChromHMM18, ChromHMM15 and Segway annotations. The segments that are common between the SCC matrices were identified with bedtools intersect requiring at least 80% of overlap reciprocally between annotations. Mapping between annotation labels were determined based on the frequency of joint labels for common segments. Correlation was calculated for each mapped columns of the SCC matrices.

k-Means clustering

After creating the matrix, *k*-means clustering was performed. To assess an appropriate *k* value of clusters, we employ the elbow plot method (26). In brief, the sum of squared error for each cluster is calculated for several *k* values and the 'elbow' point of the resulting plot is used to determine an appropriate *k* value (Supplementary Figure S1). We chose $k = 18$ based on the elbow method and to maximize understandability.

Differential expression and CIS association

Bulk RNA-Seq count data from Chu *et al.* (27) (GSE75748) was analyzed with DESeq2 to obtain differentially expressed genes between HFF and H1 cells. Up regulation was determined by >2 $\log_2$ fold change and down regulation was determined by ←2 $\log_2$ fold change and significance was detected at adjusted *P*-value <1e-10. TSS sites were obtained from Fulco *et al.* (28), which narrowed the scope by selecting the single TSS for each gene with the largest number of coding isoforms. Bedtools (29) intersect was then used to overlap those TSS regions with CIS and chromatin state annotations.

Specifically, every CIS-defined region for each cell type was intersected with the TSS regions. Then the different cell types were merged based on the genes they overlapped. These regions were then used for analysis of CIS changes (Figure 5). The 18-CIS model was collapsed to seven labels for ease of interpretation. The membership of the seven labels is as follows:

TX Contact: cont_Tx_EnhG1, cont_Tx_EnhG2, cont_TxWk_Enh, cont_TxWk_EnhWk

TSS Contact: cont_Tss_Enh, cont_Tss_EnhG, cont_Tss_noEnh, cont_Tss_EnhWk

Enhancer Contact: cont_EnhA, cont_EnhA_EnhWk

Bivalent Contact: cont_Biv

ReprPC Contact: cont_ReprPC_Biv, cont_ReprPC

Het Contact: cont_Het_Rpts, cont_Het_Rpts_strong, cont_Het_Rpts_strongest

Low Contact: cont_Low_TssBiv, cont_Low_EnhWk

Mutual information analysis. To quantify mutual information between differential gene expression (up or down regulation) and differential CIS/chromatin states, we concatenated the states for the HFF and H1 cell types. We used the aricode (30) package to calculate the normalized mutual information using the NMI function for three different variables. (i) NMI between chromatin states and gene expression, (ii) NMI between CIS and gene expression and (iii) NMI between the concatenated vector of chromatin state and CIS and gene expression.

Enrichment analyses

Data sources. To determine functional significance of the clusters, we calculated the enrichment statistics of enhancer and promoter elements for each cluster. FANTOM5 active CAGE-defined enhancers were obtained for each of the four cell types (HFF, HeLa, H1-hESC and definitive endoderm) (24). Super enhancer annotations were taken from SEdb (25). All TSS for the hg38 genome were obtained from refTSS (31) and high versus low expression genes were determined by taking the top 25% and bottom 25% of normalized gene expression counts for the following datasets: H1-GSE102311, Endoderm (32)-Additional File 1:Table S8C, HFF-GSM2448852, Hela-Encode ENCSR000CPR.

Chi-squared tests. To determine significance of enrichment, chi-squared tests were performed for each CIS cluster within each cell type for each of the five sub groupings: active enhancers, super enhancers, high expressed gene TSS, low expressed gene TSS and zero expressed gene TSS. Resulting *P*-values were Bonferroni corrected.

Results

Genomic contacts are widespread across multiple unique chromatin states

Different chromatin states as marked by distinct combinations of histone modification exhibit distinct chromatin folding and spatial organization. (33). Active (A) and inactive (B) chromatin compartments characterized by Hi-C are enriched in corresponding active and inactive chromatin marks respectively (14). Interestingly, a recent paper also proposed an intermediate (I) state between A and B which is enriched for poised-promoter and polycomb-repressed chromatin states, marked mainly by presence of H3K27me3 marks (34).

Though the interplay of several chromatin states within one compartment is well characterized, the degree of an individual genomic region in contact with several different chromatin states has not been studied in detail. To explore the global view of chromatin state interactions across different cell states, we focused on the four cell types (H1-hESC, HeLa, HFF, definitive endoderm) that have been recently assayed with Micro-C (22). Chromatin states used in this analysis correspond to the 18-state ROADMAP model inferred using ChromHMM (20) and were obtained for the four cell types from the EpiMap repository (19).

Important to note is that the distribution of chromatin states varies greatly between the four cell types (Figure 1). In particular, the endoderm cell type has many more quiescent characterized regions (those with low chromatin marks) compared to the other three cell types used here. Though this affects downstream analysis, we choose to include endoderm in our clustering because including a cell type with larger proportion of quiescent regions that have very low signals for all available histone marks shows that the method is robust to different distributions of marks.

To quantify the frequency of different chromatin state interactions for a single region, we counted the number of unique ChromHMM states that each genomic region is in contact with (Supplementary Figure S2). Here, the region is defined by the segmentation of ChromHMM and thus can be of variable length. The median number of unique ChromHMM chromatin states in contact with any single region is 12 for endoderm and HFF cells, 11 for H1 and 13 for HeLa cells suggesting a high degree of connectivity between several different chromatin state types for any single region.

Importantly, this high degree of interaction is not due to the Hi-C window spanning multiple chromatin states, as one usually finds with window size of 5 kb or more. As we are examining micro-C interactions, we can observe the interactions between segments that are as small as 1 kb.

Integrating chromatin interactions through the contact-weighted sum of chromatin states.

Given the complex nature of interaction across different chromatin states, we need an approach that will summarize this high degree of interaction for any single region. We used a straight-forward approach of summing across the different chromatin states, weighted by the contact intensity. For each region segmented with a chromHMM state annotation, all contacts 2 Mb upstream and downstream were summarized by summing all KR-normalized Micro-C scores across corresponding chromatin states, as shown in Figure 2 (see Materials and methods).

Sum of chromatin state by contact is defined as a $N \times M$ matrix, where N is the total number of segments annotated by ChromHMM and M is the number of possible chromatin states defined by ChromHMM. Each row of the matrix, SCC_i, $i = (1, \ldots N)$, is defined for the focal segment i, which we call the base region. SCC_i is a vector of length M, and is defined as the contact-weighted sum of all chromatin states that are interacting with segment i.

$$SCC_i = \sum_{j=1}^{L_i} c_{i,\,j} \cdot z_j \tag{1}$$

$$SCC_{ik} = \sum_{j=1}^{L_i} c_{i,\,j} \cdot z_{jk}, \;\; k = (1, \ldots M) \tag{2}$$

$z_j = (z_{j1}, \ldots z_{jM})$ is a vector that represents the chromatin state of the interacting segment j that is in contact with segment i. The vector consists of M binary indicator variables that are mutually exclusive and exhaustive (i.e. one and only one of the z_{jk}'s is equal to 1, and the others are 0). This indicator vector is multiplied by the contact intensity $c_{i,\,j}$ between the chromatin segment i, and the interacting segment j. When the segment spans many Micro-C windows, the mean contact intensity is used. Then the contact weighted indicator vector is summed across all interacting segments $j = (1, \ldots L_i)$, within the ± 2 Mb window.

This approach assumes that each interaction with a chromatin state segment contributes additively and independently to the base region's annotation. A similar additive approach was proposed for enhancer-target prediction in a model called

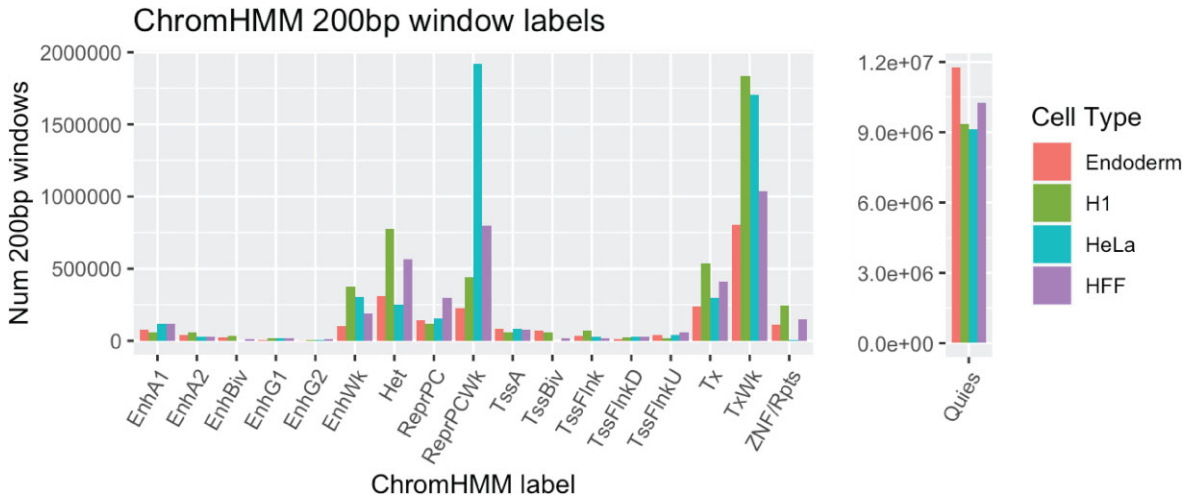

Figure 1. Chromatin state distribution between cell types. For endoderm, H1, HeLa and HFF cells, annotated ChromHMM segments were divided into equidistant 200 bp windows to show the amount of ChromHMM state distribution between cell types.

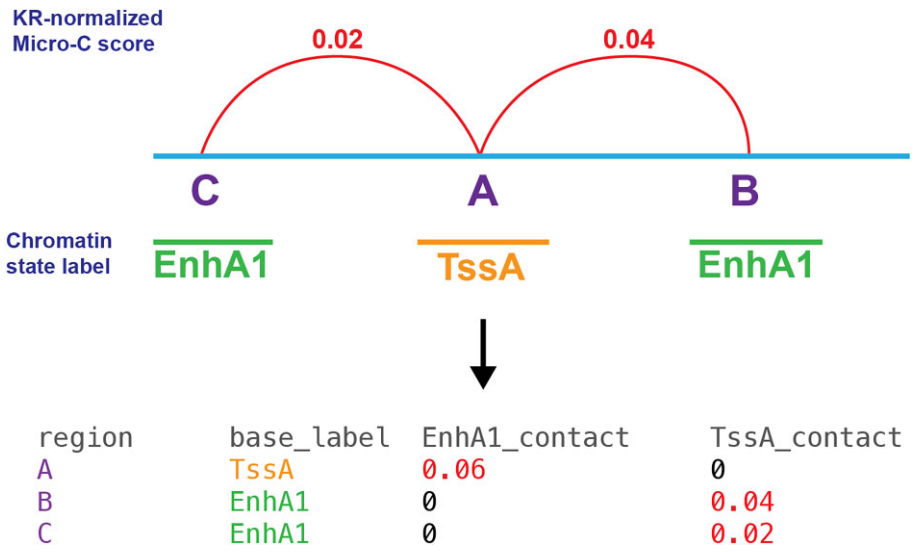

Figure 2. Creation of sum of chromatin state by contact matrix. Region A is in contact with EnhA1 regions B and C and therefore the KR-normalized scores are summed for those regions and added to the matrix. Similarly, region B is in contact with region A and the KR-normalized score is included in the matrix for region B and likewise with region C.

activity by contact in Fulco *et al.* (28). Our approach here is to extend the model to all chromatin states, rather than focusing on enhancers, as our goal is to identify broad patterns rather than specific enhancer-gene pairs.

As mentioned before, these regions were of variable length depending on the chromatin state segmentation. We chose to define base regions by chromatin state segmentation rather than a fixed window size because we are most interested in how annotated regions interact with other annotated regions. Having a fixed window size for the annotated regions can be misleading in this regard, because due to our assumptions above, a large segment that is broken into multiple smaller windows would then be summed to result in a stronger value linear to the size of the segment, and we did not want that effect of segment size influencing our results. Also, further subdivision of chromatin states into smaller windows would not provide additional information to our clustering because the Micro-C resolution used was 1 kb. To describe the relationship between the variable segment size and the fixed size of the Micro-C windows, an example is shown in Supplementary Figure S3 that describes how segments with variable sizes that can span multiple Micro-C windows were incorporated.

We plotted the resulting SCC matrix in Figure 3. We concatenated the matrices of all four cell types to detect patterns found across cell types. We excluded quiescent segments from the rows because patterns of contact with low marked regions are of little interest but being the most widespread segment in the genome (Figure 1), they occupy too many rows in the matrix if included. However, we included contacts to quiescent states in the columns to show the overall distribution of contact. A clear diagonal enrichment is evident which confirms the previously reported pattern that similar chromatin states interact with each other (13,35,36). More interestingly, there are also off-diagonal hotspots suggesting that specific interactions between different chromatin states can happen frequently.

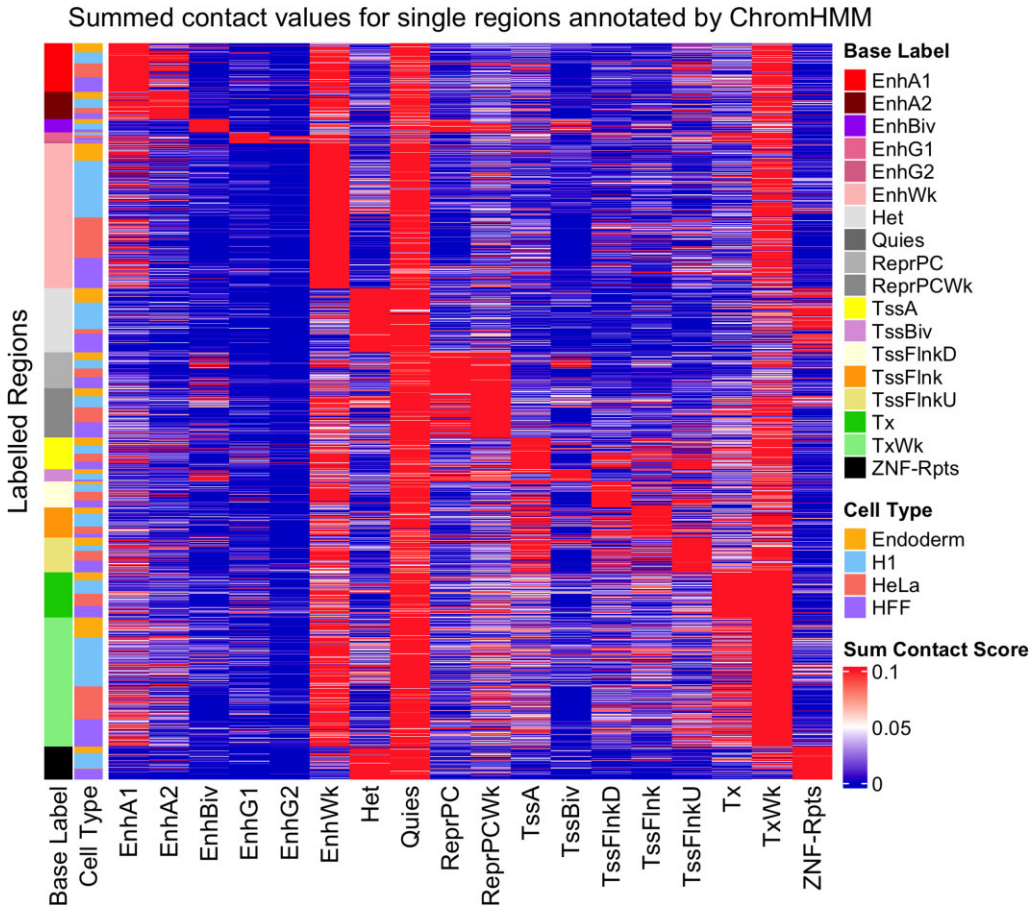

Figure 3. Sum of chromatin state by contact matrix. Each row is a ChromHMM annotated segment. Contact score in each column represents the KR-normalized Micro-C contact scores summed across all the interacting segments annotated with the corresponding chromatin state. The heatmap is ordered by base chromatin state, then by cell type.

Clustering on integrated chromatin states and Micro-C contacts reveals chromatin interaction signatures.

Based on the observation of off-diagonal contacts in the SCC matrix, we employed unsupervised clustering to characterize the patterns of chromatin interactions for each region across the whole genome. The concatenated SCC matrix for all four cell types (Figure 3) was clustered using *k*-means and a *k* of 18 was chosen based on cluster interpretability and the 'elbow' plot method (see Materials and methods and Supplementary Figure S1). While we included contact with quiescent regions in the columns of Figure 3, we excluded contacts with quiescent regions in our clustering analysis because they represent broad regions with low chromatin signals and can overwhelm the clustering.

Results of *k*-means clustering (Figure 4) yielded several clusters with distinct chromatin interaction signatures (CIS). We coin this term to describe a distinct pattern of chromatin interactions for a single genomic region. After clustering, we named each CIS according to its pattern of contact. For example, regions in the cont_Tss_Enh cluster exhibit high contact with both transcription start sites (TSS) and enhancer regions. All 18 CIS names and their ratio of contact scores are shown in Figure 4. The length distribution of CIS annotations showed longer contiguous CIS segments (median 2000 bp, mean 5203 bp) compared to the lengths of contiguous ChromHMM annotation segments (median 800 bp, mean 2485 bp) (Supplementary Figure S4).

The left panel of Figure 4 corresponds to the $\log_2$ ratio of the mean Micro-C score in contact with each chromatin state within a CIS cluster divided by the mean Micro-C scores in contact with each chromatin state across all CIS clusters. This shows the enrichment of each contact type for all 18 CIS clusters. The right panel is for annotation purposes only and was not included in clustering. The percentages correspond to the percent of the chromatin state of the base regions which belong to each CIS cluster, with each column summing to 100%. Lastly, the rightmost cell type annotation shows what percent of the cluster is made of regions from each cell type. For example, the cont_ReprPC cluster is mostly made of HeLa cell type regions.

Several clear patterns emerge. The first notable pattern is that across all regions with various base ChromHMM states, a subset of the regions are grouped into the Low contact clusters (cont_Low_TssBiv and cont_Low_EnhWk), while the rest are distributed across various clusters of high contact, showing clear difference in the contact intensity even among regions with the same chromatin state annotation.

The second pattern is that the enriched interaction between the same chromatin states that showed up as the high scores along the diagonal in the heatmap (Figure 3), are also present as enriched concentrations in Figure 4. For example, a dis-

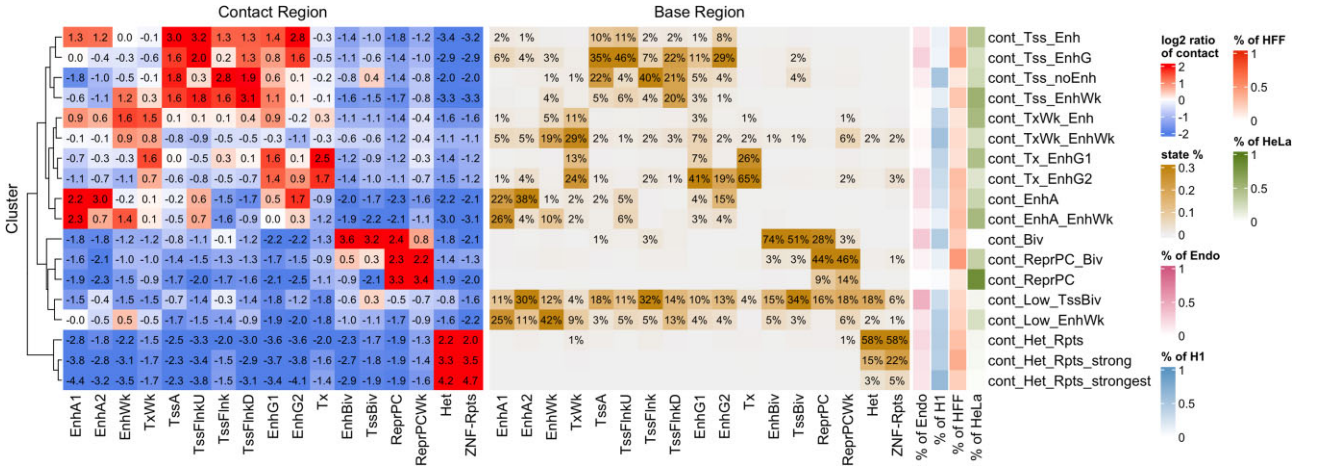

Figure 4. K-means clustering of the SCC matrix. Summed contact data for ChromHMM segmented regions (of variable length) for all four cell types were included in *k*-means clustering with a $k = 18$. The left panel shows the $\log_2$ ratio of mean summed contact scores for each segment divided by the mean summed contact scores across all clusters. In essence, this shows the enrichment of contacts to different chromatin states in each cluster. Included for annotation on the right is the percent of the base chromatin states found as well as the cell type amount present in each cluster.

tinct signature of contact with bivalent states for bivalent enhancers and bivalent TSS, and enriched contact with repressed polycomb states for repressed polycomb regions are clearly notable.

The third pattern we see is that regions with the same base chromatin annotations are now subdivided into different clusters based on the chromatin states they are interacting with. For example, the cont_Tss_Enh cluster is contrasted with the cont_Tss_noEnh cluster, the latter of which is depleted of enhancer contact, but is highly enriched for TSS contact. If one focuses on the base region TssA column (active TSS), we can see that 35% of TssA regions cluster into cont_Tss_EnhG, but a sizable minority also group into cont_Tss_noEnh or cont_Tss_Enh clusters. Likewise for EnhG2, 29% of these regions cluster into cont_Tss_EnhG, but less often it also groups into cont_Tss_Enh cluster. This perhaps shows a subdivision of enhancers with some elements preferring singular contact with TSS and others interacting with enhancers as well.

Each CIS is present in all cell types, but individual regions often change CIS between cell types.

To understand how the genome-wide chromatin interaction changes across cell types, we compared the CIS clusters across cell types. Four cell types were included in this analysis: H1-hESC, HeLa, HFF and definitive endoderm. These cell types were chosen based on the availability of human Micro-C data. Each of the cell types have regions with membership in all 18 CISs, hinting at the common pattern of chromatin interaction even among diverse cell types. This is similar to how A and B chromatin conformation compartments are observed across cell types.

A notable exception is the cont_Biv cluster which has much less HeLa membership than the other cell types (Figure 4). This is most likely due to the lack of TssBiv and EnhBiv chromatin state regions in HeLa cells (Figure 1). Similarly, HeLa cells also have a lower proportion of Het regions and therefore have less membership in the three cont_Het_Rpts clusters.

Though the same CISs are consistently found in all four cell types, individual regions often change CIS between cell types (Figure 5). Of particular interest are regions where chromatin state stays constant, but CIS is different between cell types. Depending on cell type comparison, between 12% and 40% of 200 bp base regions cluster into a different CIS despite retaining the same chromatin state between cell types (Supplementary Figure S5). This suggests that contact itself may indeed provide another layer of annotation beyond chromatin state. We chose to use a fixed window size of 200 bp in this visualization, because it is the smallest unit of the chromatin state segmentation observed in the data, and shows more accurately the distribution of chromatin state and CIS changes between the four cell types which each have varying window sizes due to the nature of cell type specific chromatin state segmentation.

Sum of chromatin state by contact (SCC) is influenced by the variation in annotation.

To understand how variation in chromatin annotation can influence our SCC measure, we generated an independent SCC matrix with a different annotation of ChromHMM15 (Supplementary Figure S6). We identified common segments (rows) between the two SCC matrices, requiring 80% overlap in length in both directions. To find the comparable labels (columns) between two annotations we looked at the frequency of joint occurrence among the common segments (Figure 6A). For eight labels, we found clear one-to-one correspondence between the two annotation versions, e.g. 5_TxWk in ChromHMM15 to TxWk in ChromHMM18. For the rest of the labels, there were more than one mapping between the versions. We report the correlation between one-to-one labels and the many-to-many labels separately (Figure 6B, C, Table 1). The SCC values showed good correlation between annotation versions, on average $\rho = 0.789$ across all the one-to-one labels. This was despite the two annotations being generated by different labs [(21) versus (20)], using different version of software (ChromHMM v1.12 versus ChromHMM v.1.10), and ChromHMM18 incorporating imputed chromatin tracks. The correlation between the labels without one-to-one mappings were less well correlated (average $\rho = 0.564$) (Figure 6C).

In addition to the SCC values, we also generated the downstream CIS clusters using k-means clustering on the SCC matrix based on ChromHMM15. The CIS clusters and the

Figure 5. CIS and chromatin state changes between cell types. Sankey diagrams show the change of CIS (**A**) and chromatin state (**B**) for 200 bp regions between H1, endoderm, HFF and HeLa cell types.

change in CIS across cell type based on ChromHMM15 are reported in Supplementary Figures S7 and S8. As seen with the SCC matrix, the overall patterns we saw in CIS clusters based on ChromHMM18 were replicated consistently in results based on ChromHMM15.

Although we found that the results were largely replicated with ChromHMM15, this was not true with the Segway annotations. Overall, the lengths of the segments and the total number of segments were significantly different between ChromHMM18 and Segway, as such there were fewer segments that were common between the two annotations. Among those common segments, there were little straightforward correspondence between the labels from ChromHMM18 and Segway (Supplementary Figure S9). In addition, there were certain labels entirely missing in cell types endoderm and HFF for Segway annotations. Since we could not find comparable columns to compare between the two SCC matrices, we instead went ahead and generated the downstream results, to compare qualitatively. We did not see that the overall patterns in the SCC matrix, CIS clusters nor CIS change replicated when using Segway annotations. (Supplementary Figures S10–S12).

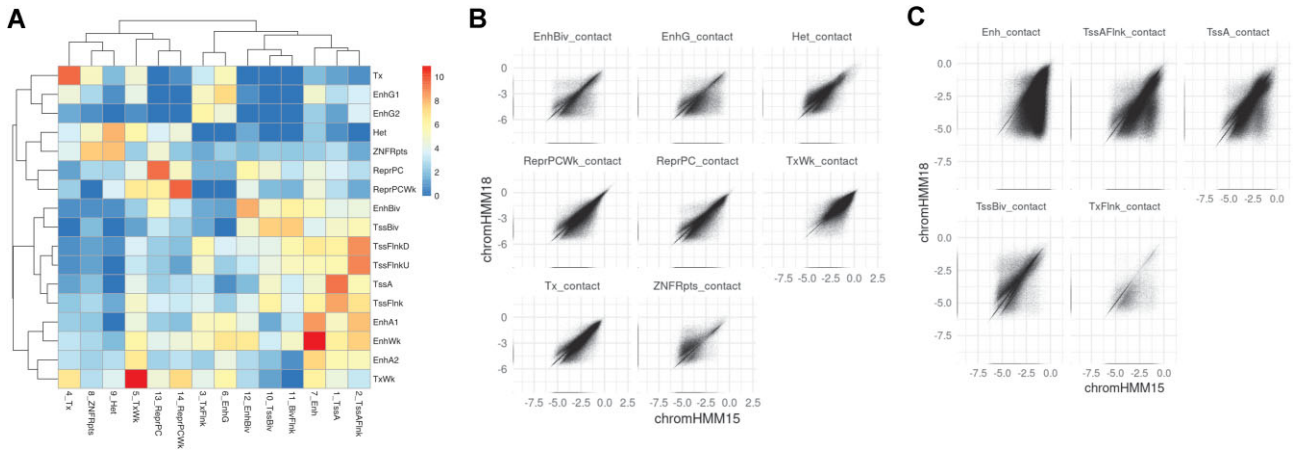

Figure 6. Correlation in SCC values between ChromHMM annotations. (ChromHMM18 vs chromHMM15). (**A**) Log-transformed frequency of joint occurrences of ChromHMM18 labels and ChromHMM15 labels, for segments that are common in both annotations. (**B**, **C**) SCC values summarizing contact with each state for segments that are common between ChromHMM18 and ChromHMM15. (B) eight states that have 1:1 mapping between ChromHMM18 and ChromHMM15. (C) other states that have uncertain mapping between ChromHMM18 and ChromHMM15.

Table 1. Correlation in SCC values between ChromHMM annotations. (ChromHMM18 versus ChromHMM15)

1:1 states		*n*:*n* states	
Label	***r***	**Label**	***r***
EnhBiv_contact	0.757	Enh_contact	0.451
EnhG_contact	0.597	TssAFlnk_contact	0.624
Het_contact	0.813	TssA_contact	0.633
ReprPCWk_contact	0.905	TssBiv_contact	0.657
ReprPC_contact	0.890	TxFlnk_contact	0.456
TxWk_contact	0.845		
Tx_contact	0.886		
ZNFRpts_contact	0.619		

Pearson correlation between the columns of the SCC matrix generated based on two different annotations. The columns represent summarized contact with the chromatin state. Left side shows contact with chromatin states that have 1:1 correspondence between annotations. Right side shows contact with chromatin states that have uncertain correspondence between annotations.

Changes in CIS in the transcription start sites are associated with change in gene expression.

To assess functional significance of CIS, we investigated CIS clusters at the transcription start sites (TSS) of upregulated and downregulated genes (Figure 7). We looked at genes with >2 $\log_2$-fold gene expression change in HFF versus H1 cell types (Figure 7). In general, TSS chromatin states are TssA, TssFlnkU, TssFlnkD or TssFlnk for upregulated genes and TssBiv and ReprPC for downregulated genes as expected. However, CIS are more variable and therefore provide additional information beyond chromatin marks alone. The TSS of genes upregulated in HFF (Figure 7A) show increased proportions of cont_TSS, cont_EnhA, cont_Tx and the TSS of genes downregulated in HFF (Figure 7B) show increased proportions of cont_Biv, cont_Het, cont_ReprPC and cont_Low. A similar pattern can be seen in H1, when the regions are ordered by H1 CIS as well, with specific CIS being more pronounced (Supplementary Figure S13). To understand the information gain provided by CIS, we quantified the normalized mutual information (NMI) between the chromatin state and CIS change in the TSS and the gene expression change, comparing HFF and H1. The NMI between gene expression (up or downregulation) and chromatin state change was 0.19, and the NMI between gene expression and CIS change was 0.15. While CIS shows less amount of information dependency with gene expression, it is notable that without knowledge of the chromatin state of the TSS itself, one can gain about three quarters of information just by observing the chromatin states of interacting partners, as one would by observing the chromatin state of the TSS directly. When we concatenate the CIS and ChromHMM states, the NMI is 0.21, showing we gain information by observing CIS together with chromatin state, compared to just observing the chromatin alone. For reference, CIS clusters across TSS of all genes regardless of differential expression are shown in (Supplementary Figure S14).

We looked at the well-characterized H1 pluripotency gene NANOG (37), to understand the utility of CIS clustering compared to chromatin state alone. We show a detailed example of changes between H1 and endoderm cells for the NANOG TSS region (Figure 8).

The chromatin state is largely identical in both endoderm and H1 cells, except for a change from TSSFlnkU to TSSBiv for a small region near the promoter, consistent with the pattern observed for TSS regions in Figure 7. However, the CIS shows more extensive change between the two cell types, with endoderm cells clustered as Low contact and H1 cells clustered as Enhancer contacting. In this case, the CIS adds extra information which correlates with the increased NANOG expression found in H1 cells.

Understanding the reasons behind the CIS change are crucial as clustering is performed on a summed contact matrix. What is the main driver for the difference between endoderm and H1 CIS? Because of the way that we define SCC as a summation vector for each region (Figure 2), $SCC_i = \sum_{j=1}^{L} c_{i,j} \cdot z_j$, there are only two ways that can result in change in this vector. Either the strengths of the contacts can change ($c_{i,j}$), or the chromatin states of the interacting regions can change (z_j). We decided to look at the TSS of NANOG in detail, to understand the components driving the change in CIS. The details of the chromatin interactions and the chromatin state of the interacting regions are shown for the NANOG TSS in Figure 9.

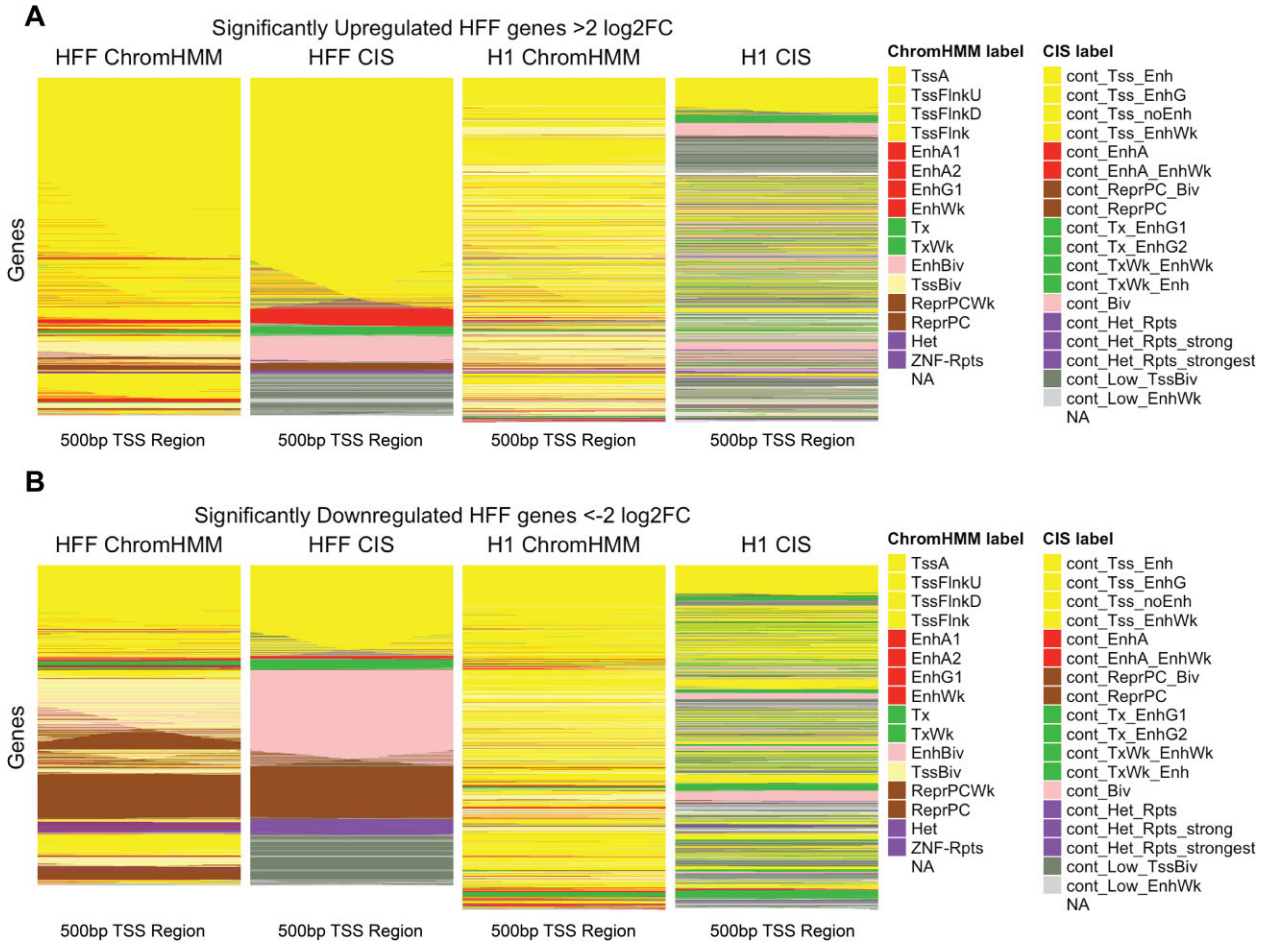

Figure 7. CIS and chromatin states for differentially expressed genes. Genes with adjusted *P*-value less than 1e-10 and with an HFF versus H1 $\log_2$ fold change >2 for upregulated genes (**A**) and <−2 for downregulated genes (**B**) were selected between H1 and HFF cell types. CIS and chromatin states are shown for each gene's 500bp TSS region. The genes (heatmap rows) are ordered by proportion of clusters in the order of HFF CIS, HFF Chromatin state, H1 CIS, then H1 Chromatin state. Alternative ordering can be found in Supplemental Figure S5.

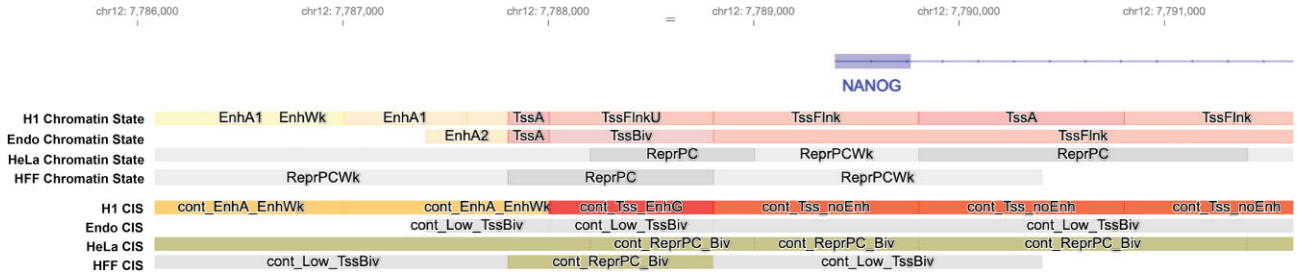

Figure 8. Chromatin state and CIS annotations at the NANOG TSS. Though the endoderm and H1 chromatin state annotations are largely similar, the CIS shows extensive differences. This corresponds with decreasing NANOG expression as H1 develops into endoderm cells.

In this example, we observe a dynamic change in both contact strength and chromatin state for most regions in contact with the TSS. This is to be expected as chromatin interactions are known to occur between similar chromatin states and therefore a change in histone marks can affect contact strength. There are however a few regions in this example which maintain chromatin state such as the TssFlnk regions found around the NANOG TSS in both cell types and the EnhA2 region downstream of the NANOG TSS. This example shows how our method is able to integrate the complex dynamics of concurrent change in chromatin state and in chromatin contact and summarize it down to lower interpretable dimensions. The CIS annotations that represent the combination of chromatin state and Micro-C data make it easier to identify regions of interest that are undergoing change in both chromatin interaction and in the chromatin state of its interacting regions. This is a unique strength of our approach, since such identification is not possible based on chromatin state or Micro-C data alone.

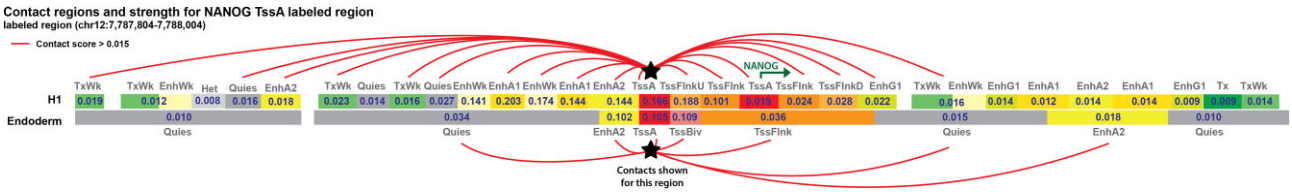

Figure 9. Bars are representative of genomics regions in contact with NANOG. They are in genomic order, but are not to scale. Regions are labeled by chromatin state and weighted Micro-C strength. Red connections signify Micro-C contact with a KR-normalized score above 0.015.

Active enhancers and super enhancers are enriched in CIS clusters

To determine if CIS clusters are enriched in cell-type specific regulatory elements, we conducted enrichment analyses looking at relevant enhancer and promoter regions in each cell type. For each cell type, we conducted chi-squared tests within each CIS for super enhancer annotations (25) and FANTOM5 CAGE-defined active enhancers (24) as well as TSS regions for highly expressed genes and lowly expressed genes (see Materials and methods).

As shown in Figure 10, several CIS show significant and strong enrichment across cell types for unique annotations. Of particular interest are the CIS cont_Tss_Enh and cont_Tss_EnhG which are enriched in active enhancers in both H1 and HeLa cells. These CIS represent regions that are in strong contact with both TSS chromatin states and enhancer chromatin states.

Regions clustered into cont_Tss_noEnh are enriched in the TSS of highly expressed genes in each cell type and less enriched for active enhancer regions. CIS cont_Tss_noEnh represent regions that are mainly in contact with TSS chromatin states but not in contact with enhancer chromatin states.

Another interesting pattern is the super enhancer enrichment in enhancer-contacting CIS across all cell types. CIS cont_EnhA and cont_EnhA_EnhWk represent regions that are in strong contact with enhancer chromatin states and in weak contact with TSS chromatin states, which is to be expected as super enhancers are generally defined as clusters of active enhancers (38).

A last observation is the enrichment of cont_Biv CIS in lowly and zero expressed genes, in line with our observation based on differential gene expression in Figure 7B. CIS cont_Biv represent regions in strong contact with bivalent TSS and bivalent enhancer chromatin states. Bivalent chromatin states are regions of the genome marked simultaneously by both active and inactive chromatin marks purported to be repressed but 'poised' for rapid activation upon a cell differentiation signal (39). The pattern we see here is consistent with such hypothesis, but it emphasizes the role of contact with bivalent chromatin in addition to the bivalent state of the TSS itself (40).

Discussion

Methods for integrating 3D Hi-C data and traditional 1D data such as ChIP-Seq are a crucial step towards understanding the shifting dynamics of cell regulation. We show that clustering on an integrated matrix of Micro-C scores and chromatin states adds an additional layer of annotation to the genome.

Though chromatin states are known to interact with similar chromatin states in the genome (13,35,36), we show that there are regions in contact with different chromatin states. These associations present an opportunity for pattern detection which further aid with interpretation.

As shown in our enrichment analysis, cont_EnhA and cont_EnhA_EnhWk clusters are enriched for super enhancers, cont_Tss_Enh and cont_Tss_EnhG are enriched for active enhancers. Therefore, by clustering across the entire genome, we can effectively narrow the list of candidate regions with regulatory potential based on both contact and chromatin state evidence.

Most importantly, we present a framework for integrating chromatin conformation data into a more traditional vector representation across the genome. This creates possibilities for future machine learning and clustering approaches. Additionally, this summative approach could be extended to include ATAC-seq, transcription factor binding, or DNA methylation data to further enhance regulatory predictions.

It is important to caution against the uncertainty and instability of the chromatin annotations that our method relies on. There is growing evidence in the literature that chromatin state annotations are not completely reproducible across different experimental replicates or even different runs of annotation (41,42). The effect of this variation was seen in our comparison between two different ChromHMM annotations. There were good correlations in the SCC values, and the overall patterns were replicated, but there was considerable noise in the values for individual segments. In the case of the comparison with Segway, even qualitative patterns were not replicated. This suggests that although unsupervised learning based on chromatin states and contact can provide insight for global patterns, relying on these annotations for inference on individual loci may be fraught with uncertainty.

Another limitation of our approach is the loss of information on specific pairwise interactions as a result of summing across all interactions with the same chromatin states. Future work to remedy this could include clustering based on pairwise relationships rather than on a summarized vector of contacts for one each base region. An alternative approach would be to utilize the graph structure of the contact data in order to preserve the individual contacts, and use graph embedding techniques to transform relational data into a vector form (10,15,16). This would allow one to learn on segments (nodes) as well as pairwise contacts (edges), and has been shown to be a powerful approach in prediction of gene expression (15,16).

In conclusion, we show a simple and straight-forward methodical approach to integrate contact and chromatin mark data across the genome, allowing researchers to distill complex chromatin interaction information into a vector representation, and then to an interpretable annotation, by further clustering on the vector. Using this method, we present the first set of chromatin interaction signatures for the human genome that summarizes the genome-wide pattern of contact

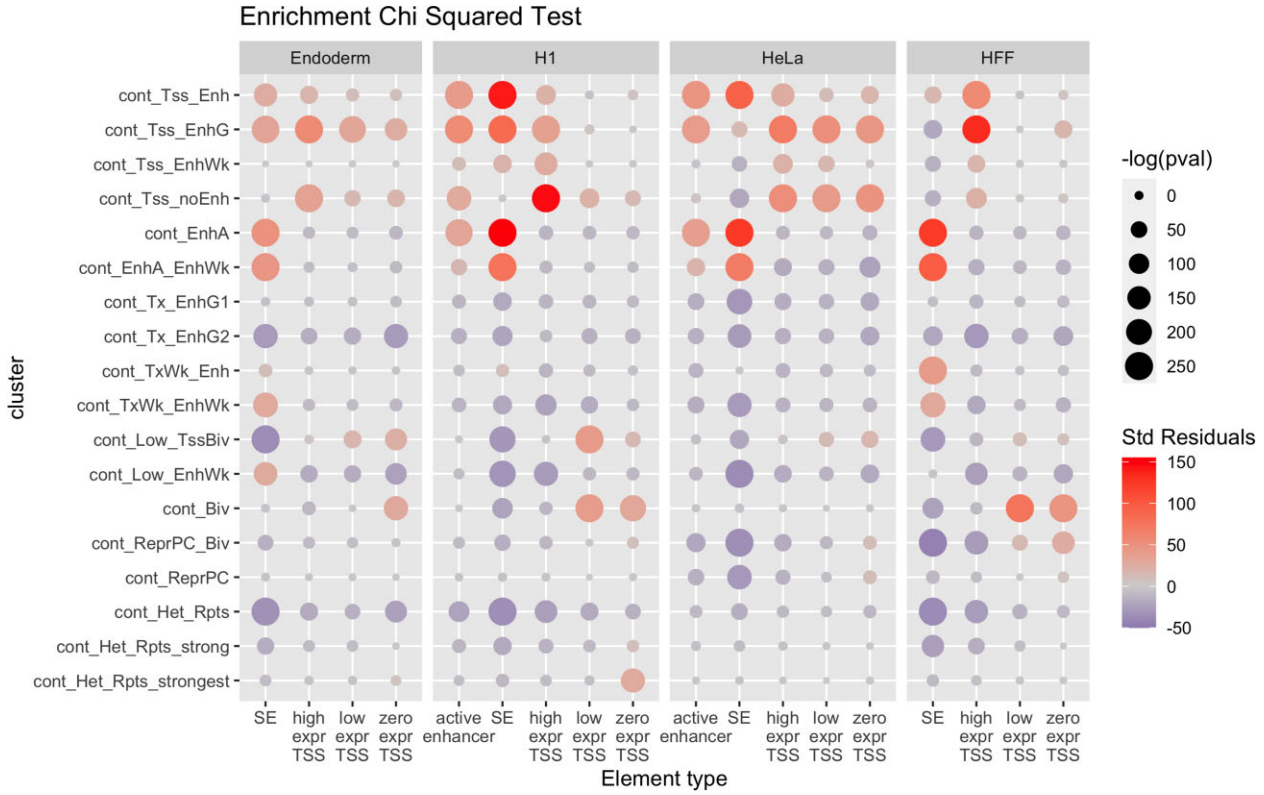

Figure 10. Enrichment results for CIS of all four cell types. Based on data availability, we performed chi-squared tests to test enrichment for super enhancers, high, low, and zero gene expression TSSs and active enhancers in each CIS cluster.

between chromatin states at the Micro-C resolution. Our results show that most chromatin interaction signatures are frequently and repeatedly found across the genome and in all four cell types investigated here. Between 12–40% of the regions change chromatin interaction signatures between the cell types despite maintaining chromatin state, hinting at the dynamic nature of chromatin conformation. Although regions with similar chromatin states are often in contact as expected, subcategories of enhancers and transcription start sites have distinct chromatin interaction signatures that are associated with gene expression. Thus, these chromatin interaction signatures allow a genome-wide view of chromatin interaction, and provide more information about gene regulation than either chromatin state or Hi-C contacts alone.

Data availability

The code underlying this article are available at https://doi.org/10.6084/m9.figshare.25270645. The data underlying this article are available in zenodo at https://doi.org/10.5281/zenodo.10694854.

Supplementary data

Supplementary Data are available at NARGAB Online.

Funding

National Science Foundation [1750532, 1946082]. Funding for open access charge: National Science Foundation [1750532].

Conflict of interest statement

None declared.

References

1. Day,N., Hemmaplardh,A., Thurman,R.E., Stamatoyannopoulos,J.A. and Noble,W.S. (2007) Unsupervised segmentation of continuous genomic data. *Bioinformatics*, **23**, 1424–1426.
2. Hoffman,M.M., Ernst,J., Wilder,S.P., Kundaje,A., Harris,R.S., Libbrecht,M., Giardine,B., Ellenbogen,P.M., Bilmes,J.A., Birney,E., *et al.* (2013) Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.*, **41**, 827–841.
3. Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.
4. Ernst,J. and Kellis,M. (2017) Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.*, **12**, 2478–2492.
5. Chan,R.C.W., Libbrecht,M.W., Roberts,E.G., Bilmes,J.A., Noble,W.S. and Hoffman,M.M. (2018) Segway 2.0: gaussian mixture models and minibatch training. *Bioinformatics*, **34**, 669–671.
6. Libbrecht,M.W., Ay,F., Hoffman,M.M., Gilbert,D.M., Bilmes,J.A. and Noble,W.S. (2015) Joint annotation of chromatin state and chromatin conformation reveals relationships among domain types and identifies domains of cell-type-specific expression. *Genome Res.*, **25**, 544–557.
7. Wang,Y., Zhang,Y., Zhang,R., van Schaik,T., Zhang,L., Sasaki,T., Peric-Hupkes,D., Chen,Y., Gilbert,D.M., van Steensel,B., *et al.* (2021) SPIN reveals genome-wide landscape of nuclear compartmentalization. *Genome Biol.*, **22**, 36.
8. Liu,Q., Bonneville,R., Li,T. and Jin,V.X. (2017) Transcription factor-associated combinatorial epigenetic pattern reveals higher

transcriptional activity of TCF7L2-regulated intragenic enhancers. *Bmc Genomics [Electronic Resource]*, **18**, 375.
9. Mendez,M., FANTOM Consortium Main Contributors,Scott,M.S. and Hoffman,M.M. (2020) Unsupervised analysis of multi-experiment transcriptomic patterns with SegRNA identifies unannotated transcripts. bioRxiv doi: https://doi.org/10.1101/2020.07.28.225193, 29 July 2020, preprint: not peer reviewed.
10. Shokraneh,N., Arab,M. and Libbrecht,M. (2023) Integrative chromatin domain annotation through graph embedding of Hi-C data. *Bioinformatics*, **39**, btac813.
11. Oudelaar,A.M. and Higgs,D.R. (2021) The relationship between genome structure and function. *Nat. Rev. Genet.*, **22**, 154–168.
12. Furlong,E.E.M. and Levine,M. (2018) Developmental enhancers and chromosome topology. *Science*, **361**, 1341–1345.
13. Esposito,A., Annunziatella,C., Bianco,S., Chiariello,A.M., Fiorillo,L. and Nicodemi,M. (2019) Models of polymer physics for the architecture of the cell nucleus. *WIREs Syst. Biol. Med.*, **11**, e1444.
14. Rao,S.S.P., Huntley,M.H., Durand,N.C., Stamenova,E.K., Bochkov,I.D., Robinson,J.T., Sanborn,A.L., Machol,I., Omer,A.D., Lander,E.S., *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.
15. Bigness,J., Loinaz,X., Patel,S., Larschan,E. and Singh,R. (2022) Integrating long-range regulatory interactions to predict gene expression using graph convolutional networks. *J. Comput. Biol.*, **29**, 409–424.
16. Karbalayghareh,A., Sahin,M. and Leslie,C.S. (2022) Chromatin interaction–aware gene regulatory modeling with graph attention networks. *Genome Res.*, **32**, 930–944.
17. Krietenstein,N., Abraham,S., Venev,S.V., Abdennur,N., Gibcus,J., Hsieh,T.-H.S., Parsi,K.M., Yang,L., Maehr,R., Mirny,L.A., *et al.* (2020) Ultrastructural details of mammalian chromosome architecture. *Mol. Cell*, **78**, 554–565.
18. Hsieh,T.-H.S., Fudenberg,G., Goloborodko,A. and Rando,O.J. (2016) Micro-C XL: assaying chromosome conformation from the nucleosome to the entire genome. *Nat. Methods*, **13**, 1009–1011.
19. Boix,C.A., James,B.T., Park,Y.P., Meuleman,W. and Kellis,M. (2021) Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature*, **590**, 300–307.
20. Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J., Ziller,M.J., *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
21. Libbrecht,M.W., Rodriguez,O.L., Weng,Z., Bilmes,J.A., Hoffman,M.M. and Noble,W.S. (2019) A unified encyclopedia of human functional DNA elements through fully automated annotation of 164 human cell types. *Genome Biol.*, **20**, 180.
22. Akgol Oksuz,B., Yang,L., Abraham,S., Venev,S.V., Krietenstein,N., Parsi,K.M., Ozadam,H., Oomen,M.E., Nand,A., Mao,H., *et al.* (2021) Systematic evaluation of chromosome conformation capture assays. *Nat. Methods*, **18**, 1046–1055.
23. Hsieh,T.-H.S., Weiner,A., Lajoie,B., Dekker,J., Friedman,N. and Rando,O.J. (2015) Mapping nucleosome resolution chromosome folding in yeast by Micro-C. *Cell*, **162**, 108–119.
24. Andersson,R., Gebhard,C., Miguel-Escalada,I., Hoof,I., Bornholdt,J., Boyd,M., Chen,Y., Zhao,X., Schmidl,C., Suzuki,T., *et al.* (2014) An atlas of active enhancers across human cell types and tissues. *Nature*, **507**, 455–461.
25. Jiang,Y., Qian,F., Bai,X., Liu,Y., Wang,Q., Ai,B., Han,X., Shi,S., Zhang,J., Li,X., *et al.* (2019) SEdb: a comprehensive human super-enhancer database. *Nucleic Acids Res.*, **47**, D235–D243.
26. Marutho,D., Hendra Handaka,S., Wijaya,E. and Muljono (2018) The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. In: *2018 International Seminar on Application for Technology of Information and Communication*. pp. 533–538.
27. Chu,L.-F., Leng,N., Zhang,J., Hou,Z., Mamott,D., Vereide,D.T., Choi,J., Kendziorski,C., Stewart,R. and Thomson,J.A. (2016) Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol.*, **17**, 173.
28. Fulco,C.P., Nasser,J., Jones,T.R., Munson,G., Bergman,D.T., Subramanian,V., Grossman,S.R., Anyoha,R., Doughty,B.R., Patwardhan,T.A., *et al.* (2019) Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.*, **51**, 1664–1669.
29. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
30. Chiquet,J., Rigaill,G., Sundqvist,M. and Dervieux,V. (2020) Package 'aricode'. R Package Version.
31. Abugessaisa,I., Noguchi,S., Hasegawa,A., Kondo,A., Kawaji,H., Carninci,P. and Kasukawa,T. (2019) refTSS: a reference data set for human and mouse transcription start sites. *J. Mol. Biol.*, **431**, 2407–2422.
32. Blake,L.E., Thomas,S.M., Blischak,J.D., Hsiao,C.J., Chavarria,C., Myrthil,M., Gilad,Y. and Pavlovic,B.J. (2018) A comparative study of endoderm differentiation in humans and chimpanzees. *Genome Biol.*, **19**, 162.
33. Boettiger,A.N., Bintu,B., Moffitt,J.R., Wang,S., Beliveau,B.J., Fudenberg,G., Imakaev,M., Mirny,L.A., Wu,C. and Zhuang,X. (2016) Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. *Nature*, **529**, 418–422.
34. Vilarrasa-Blasi,R., Soler-Vila,P., Verdaguer-Dot,N., Russiñol,N., Di Stefano,M., Chapaprieta,V., Clot,G., Farabella,I., Cuscó,P., Kulis,M., *et al.* (2021) Dynamics of genome architecture and chromatin function during human B cell differentiation and neoplastic transformation. *Nat. Commun.*, **12**, 651.
35. Hildebrand,E.M. and Dekker,J. (2020) Mechanisms and functions of chromosome compartmentalization. *Trends Biochem. Sci.*, **45**, 385–396.
36. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O., *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
37. Boyer,L.A., Lee,T.I., Cole,M.F., Johnstone,S.E., Levine,S.S., Zucker,J.P., Guenther,M.G., Kumar,R.M., Murray,H.L., Jenner,R.G., *et al.* (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, **122**, 947–956.
38. Hnisz,D., Abraham,B.J., Lee,T.I., Lau,A., Saint-André,V., Sigova,A.A., Hoke,H.A. and Young,R.A. (2013) Super-enhancers in the control of cell identity and disease. *Cell*, **155**, 934–947.
39. Bernstein,B.E., Mikkelsen,T.S., Xie,X., Kamal,M., Huebert,D.J., Cuff,J., Fry,B., Meissner,A., Wernig,M., Plath,K., *et al.* (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, **125**, 315–326.
40. Jia,J., Zheng,X., Hu,G., Cui,K., Zhang,J., Zhang,A., Jiang,H., Lu,B., Yates,J., Liu,C., *et al.* (2012) Regulation of pluripotency and self-renewal of ESCs through epigenetic- threshold modulation and mRNA pruning. *Cell*, **151**, 576–589.
41. Zhang,Y. and Hardison,R.C. (2017) Accurate and reproducible functional maps in 127 human cell types via 2D genome segmentation. *Nucleic Acids Res.*, **45**, 9823–9836.
42. Foroozandeh Shahraki,M., Farahbod,M. and Libbrecht,M.W. (2024) Robust chromatin state annotation. *Genome Res.*, **34**, 469–483.

Received: February 22, 2024. **Revised:** August 21, 2024. **Editorial Decision:** September 19, 2024. **Accepted:** September 20, 2024