A graph-based comparative analysis of three-dimensional organization of chromosomes in yeast and mammals

Robert Atlas¹, Sushmita Roy¹,²,³
¹Dept. of Computer Science, ²Dept of Biostatistics and Medical Informatics, ³Wisconsin Institute for Discovery
University of Wisconsin, Madison, WI 53717, USA

Abstract

Genome-wide maps of chromosomal interactions are becoming increasingly common. Computational tools to analyze such maps, and more importantly, comparing such maps across multiple contexts, and organisms are scarce. We here develop a novel graph-based clustering approach to detect sets of interacting genomic loci and compare them across multiple cellular and organismal contexts. We used a hybrid approach to detect an interaction between two loci to enable our approach to be applicable to 3C data from both simple eukaryotes as well as higher eukaryotes such as mouse and humans. To determine the number of clusters we used a penalized cluster quality criteria and developed numerous statistics to systematically examine properties of chromosomal organization. Application of our approach identified several principles: (a) the proportion of inter-chromosomal interactions is much higher in yeast, compared to mammalian species, (b) in addition to replication forks, distally interacting regions in yeast also exhibited a tendency to be co-regulated (based on gene expression, targets of knockouts and targets of ChIP), (c) most of the chromosomal organization is the same between two mammalian tissues, but there are some regions that exhibit a tissue-specific interaction pattern, (d) comparison of interaction maps between human and mouse identified significant conservation of clusters of regions, but, we did not find a similar conservation between yeasts, and yeast and mammals. Our graph-based clustering approach enabled us to perform a systematic comparison of multiple chromosomal region interaction maps, corroborated known findings of such interaction maps, and also identified novel aspects of such maps.

1 Introduction

Three-dimensional organization of chromatin is emerging as an important layer of regulation of gene expression [10, 14]. Recent advances in chromosome conformation capture (3C) technology coupled with next-generation sequencing allows us to assay the 3D organization of a genome at an unprecedented precision and detail [3, 4]. These comprehensive maps have enabled us to examine the folding, organizational principles of the chromosome and identify long-range interactions among genomic loci. Studies in yeast have shown that such long range interactions are enriched for loci involving tRNA genes, centromeres, and origins of replication [6] and transcription factories for regulation of gene expression [15]. In mammalian systems, such long range interactions occur pervasively across different genomic loci, and are often enriched for regulatory elements and associated with major developmental changes [10, 14, 5, 8].

The recent availability of these genome-wide 3C datasets, namely HiC and 4C, for multiple species and tissues gives us the unique opportunity to compare chromatin organization across multiple tissues, multiple organisms and to identify evolutionary principles of this organization. However, computational tools to
analyze these datasets are still in their infancy. While statistical techniques have been developed to identify significant interacting genomic loci [19, 18], methods that systematically compare these maps across multiple tissues or multiple organisms are scarce [5]. In particular, systematic comparison of such distal interaction patterns and the general 3D organization of the genome across multiple conditions, cell-types, and organisms, is still largely unexplored and is a significant computational challenge.

One approach is to first represent a set of interacting loci as a graph of interactions, where each node corresponds to a genomic region, and an edge represents a distal interaction between the two loci (as determined by the statistical significance of the number of observed reads mapped to both regions). Then to compare two 3D organization maps entails comparing the edge-set interaction between these two maps. However such a “pairwise” approach assumes that a pair of regions interact independently, an assumption, which is not necessarily true for most of these data. In particular, two chromosomal regions that are close together in linear space are likely to share many of their interaction patterns in 3D space. Further more, comparing entire graphs might not reveal properties that are true for some subgraphs in the network.

We develop a novel approach based on the identification of dense sub-graphs for systematically comparing 3C datasets from multiple cell types and organisms. Our approach projects interacting loci onto a graph and applies a clustering algorithm on the graph to identify densely connected subgraphs. Such subgraphs bring together regions close together in linear space, producing tight clusters of closely entwined regions. We use the Markov Cluster algorithm (MCL, [16]) which we adapted to automatically define the optimal clustering to detect densely-connected components of these sparse graphs. Then to compare such sets of clusters across two cell types we develop statistical tests to detect enrichment between different clusters.

Our contributions are: (1) a hybrid approach of detecting interactions between genomic loci controlling both for inter and intra-chromosomal biases; (2) new graph-based clustering algorithm for analyzing 3C data; (3) a framework that enables comparison of multiple genome-wide 3C datasets from multiple cell types and organism, which takes into account complications arising from genomes of different size, structure, gene density; (4) application of our approach to seven 3C datasets from yeast, humans and mouse, making this one of the most comprehensive studies of 3C datasets.

Our results on the yeast genome-wide 3C data recapitulates the known enrichment of interacting loci in centromeres, tRNA, origins of replication (ORCs), and double-stranded breaks (DSBs) [18]. We also find several clusters enriched for targets of TFs including SWI3, a chromatin remodeler. Comparison of interaction maps across pluripotent and differentiated cells in mouse shows that there is significant overlap between the clusters from two cell types, and this holds for both mouse and human. We observe significant conservation of clusters obtained between human and mouse ES cells, but no conservation between fission and budding yeasts. Furthermore, the proportion of inter-chromosomal interactions in budding yeast is more than three times greater than the interactions in mammalian species, suggesting different organizational principles of the chromatin at play between simple eukaryotes and mammalian species.

2 Results

2.1 Markov Cluster on Chromosome Capture (MCL-CC): A graph-based framework to detect chromosomal interaction modules.

To enable comparison across different genome-wide chromosome capture maps, we develop a graph-based clustering approach, Markov Cluster for Chromosome Capture (MCL-CC, Fig. 1). We use datasets from both 4C as well as HiC but most of our datasets are from HiC, and thus we will use HiC to refer to all our datasets. Our approach uses the Markov Cluster Algorithm (MCL) within a penalized framework to
Figure 1: Overview of approach. We begin with mapped reads to the genome and estimate the interactions among loci while controlling for inter and intra-chromosomal biases. This interaction graph is analyzed using a pair-wise approach (Witten et al., [18]) in yeast. The interaction map is then clustered using our penalized Markov Cluster algorithm to automatically detect the number of clusters. Such clusters are produced for multiple cell types and conditions. Finally, these clusters are compared across conditions.

partition the three-dimensional chromosomal organization of each genome into tightly co-located clusters [?]. The MCL algorithm is flow-based algorithm that automatically detects densely connected components in the graph by taking increasing powers of the adjacency matrix. The graph itself was defined by a set of interactions among genomic regions. We represented the three-dimensional organization of a genome as a graph where the nodes represent regions, and the edges represent a statistically significant interaction. To define an interaction between two genomic loci we computed a q-value statistical significance of the number of paired-end reads that map between the two loci (See Methods). Two loci were said to interact, if their associated q-value < 0.01. We tried various thresholds of q-values and our results are not sensitive to the particular value of q. Our method of detecting significance explicitly controls for interactions among inter and intra-chromosomal loci, thus is applicable to both mammalian and yeast species. In particular, controlling for intra-chromosomal loci is especially important for mammalian species because of their larger genome size [10]. To determine the number of clusters, we need to select the I parameter in the MCL algorithm (Methods). We select I using a penalized cluster quality approach approach that optimizes the cluster quality while preventing too many small clusters. We estimate the cluster quality using a modified version of Davies-Bouldin index that uses within-cluster and between-cluster edge-density to measure cluster quality. This approach is suited for “graph” based clusters and enables us to automatically determine the number of clusters.
2.2 Higher-order organization of yeast genome revealed by MCL-CC

We first applied our approach to chromosome conformation capture data for the unicellular yeast, *S. cerevisiae* from Duan et al. [6], as this is a very well characterized model organism, and serves as a good test for our approach. At an FDR \(< 0.01\) we recovered 252,659 interactions (71,096 intra and 181,563 inter-chromosomal) among 3,596 regions, which is comparable to 306,312 reported in Duan et al., with an overlap of 43.6% co-locations common in both analysis methods, p-value\(< 1E-100\). We next used clustered these interactions into 241 clusters (Fig. 2A). These clusters were non-uniform in size, with \(~90\%\) of the regions present in three clusters (Fig. 2B). We discarded clusters with \(< 4\) regions, resulting in 121 clusters over 3,377 which we used for downstream analysis.

We examined whether the interaction map we inferred using our q-value calculation was enriched for specific biological properties, including those that have already been identified [6, 18]. To assess statistical significance of enrichment of a functional category in our clusters, we used two level of stringency: first we estimated a “null” distribution of p-values using enrichment in random clusters, second, we calculated a p-value from this distribution, which we corrected using the Benjamini-Hochberg FDR correction of the p-value (Methods). A category is considered enriched in particular cluster if it has a q-value \(< 0.05\). We used both a pairwise approach of Witten et al., but with our q-value calculation, and also our MCL-CC approach. Regions associated with centromeres, tRNA, and two kinds of breakpoints were significantly enriched in two of our clusters, Cluster 0 and 1 (Fig 2C). A similar enrichment was observed for the pairwise approach.

We next asked if genes associated with interacting regions are enriched for targets of particular transcription factors or for certain biological processes. We projected regions onto genes; with \(~2\) genes mapped to a region on average, primarily concentrated in the three largest clusters. Using Reimand et al.’s TF knockout we found several TFs whose targets were enriched in different clusters (TUP1, UME6, SNF6, SWI3, SIN4, SNF2). Interestingly, SNF6, SNF3 and SWI3 are subunits of the replication fork complex. TUP1 is involved in repressive chromatin and interacts with H3 and H4 histones. SIN4 is RNA pol II subunit and is a general regulator, and finally UME6 is known to interact with Sin3-RPd3 complex to regulate meiosis related genes. All of these factors are general chromatin modelers, or are associated with DNA replication.

In addition, we found that 37 of 193 factors for which ChIP-chip data were recently generated [17] to be enriched in clusters including SWI3 and RPB3. We compared our clusters also with data from MacIsaac et al., but did not find any significant enrichment. This is consistent with previous findings [11]. Notably, we found SWI3, which is a core component of the SWI/SNF chromatin remodeling complex, subunit was found in both knockout and ChIP data, which strongly suggests that the corresponding cluster is enriched for functional distal interactions. While previous work has shown that distal interactions are enriched for ORCs and replication forks, and SWI is known regulator of replication, we here provide the first evidence of the role of distal interactions in the regulation of genes regulated by SWI3.

Several of our clusters were associated with GO processes associated with cell-fate determination and RNA processing, (using the pairwise approach we did not observe any enrichment). We also observed a modest, but significant tendency of genes close to interacting loci to be co-expressed than random pairs of genes (KS test p-value \(< 1E-10\), in an expression compendium measuring expression of the entire yeast genome under multiple stresses [7] (Fig 2C).

Overall, of the 121 clusters, 12 were associated with ORCs, tRNAs, targets of various factors based on ChIP-chip, TF knockouts, gene co-expression or GO processes. These clusters included 82\% of the regions (87.2\% of the genes), thus comprising the majority of the regions. Some of these properties were observed using the pairwise approach as well. Thus both these approaches are complementary, with both agreeing on the core properties associated with distal interactions, but each also finding new properties that are missed by the other. The advantage of using our cluster based-approach is that it enables comparison of
two co-location maps more easily as signal for “sets” of regions is much greater than a pair of regions.

2.3 MCL-CC clusters in differentiated and pluripotent cells of mouse and human identifies tissue specific as well as common domains

Having established the utility of our MCL-CC approach in yeast, we next applied our approach to the recently generated HiC data that measures the three-dimensional chromosomal organization in two mouse and two human cell types [5]. We found 294,102 (277,977 intra-chromosomal, 16,125 inter-chromosomal) interactions in mouse embryonic stem (mES) cells and 200,668 (187,964 intra-chromosomal, 12,704 inter) interactions in mouse cortex (mCortex) cells. These interactions clustered into 106 clusters in mESC and 104 clusters in mCortex. Consistent with the identification of long topological domains in mammalian chromosomes [5], most clusters spanned a single chromosome. However, 2 mESC and 4 mCortex clusters included regions from a wide variety of chromosomes, demonstrating the ability of our method to detect very long range interactions (Fig. 3). We examined these multi-chromosomal clusters between tissues and found that one pair of clusters (cluster 13 from mES and cluster 20 from mCortex) that had significant overlap with each other, suggesting that such inter-chromosomal interactions are common between tissues. Genes that corresponded to the common regions between these two clusters were similarly expressed between the two tissues suggesting common regulatory mechanisms acting on these genes.

We next asked to what extent regions that were in one cluster in mES cells were also in a conserved cluster in mCortex cells. We found that the cluster membership for the vast majority of the clusters agreed between the two tissues. In particular 6939 of regions in 45 clusters in mES cells were also in the same cluster in mCortex cells, and 6844 regions in 39 clusters in mouse cortex cells were in the same cluster in mESC cells (Fig 4). We asked whether genes associated with regions that are in conserved clusters exhibit tissue specific expression. Of the 45 clusters that overlapped with 39 clusters in mCortex, 7 were enriched in a GO function including sensory perception and 1 for cell differentiation. We considered the genes that contribute to this enrichment, and found that the 69 genes involved in epidermal cell differentiation have a higher expression in mES cells compared to mCortex cells (ranksum test p-value < 0.06, Fig 3F). However, the 371 genes involved in sensory perception were not as differently expressed between the two tissues.

To examine the generality of these properties in other mammals, we applied MCL-CC to HiC data from four human cell lines, human ES cells and IMR90 fibroblasts from Dixon et al. 2012, and GM90 and K562 cells from leiberman2009 et al. [10]. In both human ES and IMR90 cells, the clusters primarily contained regions from the same chromosomes (Fig 5A,B), with a handful of clusters harboring regions from multiple chromosomes. As in mouse cells, we observed substantial agreement in cluster membership of regions between hES and IMR90 cells comprising with 89 (of 424) human ES cell clusters, and 126 (of 326) IMR90 fibroblast clusters contributing to overlap of ~75% of regions common between 7812 regions in ESC and 7718 regions in IMR90. Several of these clusters were enriched in processes in differentiation, sensory perception of smell and response to stimulus. To confirm that the results for different tissues in the same species are not just an artifact of an experimental methodology particular to one study, we compared the human GM cells with the human K562 cells from the Lieberman-Aiden study, and found a relatively lower, but significant extent of overlap between regions: 55 of 319 clusters in GM, overlapping with 50 of 590 clusters in K562 including 6679 regions in the intersection of 7769 GM regions and 6689 K562 regions. Thus this similarity pattern between these tissues is likely a general organizational principle of chromatin in mammalian cells.

Lastly, we compared interaction maps across species, we used a hyper-geometric p-value to quantify the overlap between two clusters. We considered matching cell types between human and mouse and considered orthologous genes in the other species, and calculated a hyper-geometric p-value with FDR correction for
Figure 2: Analysis of 3C data in yeast using MCL-CC. A. Heatmap showing the pattern of significant interactions between regions within the yeast MCL clusters. Interactions (red) are ordered by cluster (dotted lines) and chromosome, with interactions between cluster 0 and itself in the upper left. The bottom row and rightmost column contains regions outside the three largest clusters. The proportion of regions and genes contained in the three largest yeast clusters. Circos diagrams displaying the pattern of interactions within and between different chromosomes in cluster 0 (blue), cluster 1 (green), and cluster 2 (red). The distribution of regions according to chromosome for the ten largest yeast clusters. As shown, the three large clusters contain numerous regions from a variety of chromosomes. Functional categories and GO terms for which the yeast 3C interaction network is enriched. Top row shows results from Witten et al’s pairwise approach, and the remaining rows represent MCL-CC clusters. Intensity of blue corresponds to our log(qvalue). Only enrichments at FDR<0.05 are displayed.
Figure 3: Distal interactions of genomic loci in mouse stem cells. A. The distribution of regions according to chromosome for the 26 largest mouse ESC clusters. Most clusters draw from a single chromosome. B. Heatmap showing the pattern of significant interactions between regions within the mouse ESC clusters. Interactions (red) are ordered by cluster (dotted lines) and chromosome, with interactions between cluster 0 and itself in the upper left. The close-up shows interactions between cluster 2 and 7, which both contain a majority of regions from chromosome 1. As shown, the interaction pattern between each cluster and itself (upper-left and lower-right) is much stronger than the interaction pattern between the two. C. In this circos diagram of chromosome 1, blue lines represent interactions between pairs of regions in cluster 2, while red lines are interactions between pairs in cluster 7. As suggested in B, each cluster tends to represent distinct portions of the chromosome with little interaction between chromosomes.
Figure 4: Comparison of tissue-specific organization of chromatin in mouse
A. The distribution of regions per chromosome for the 26 largest mouse Cortex clusters. B. Overlap of clusters between cortex and ES cells in mouse. C. The proportion of regions and genes in each tissue contributing to the conservation in interactions. Since the majority of large clusters in one tissue overlap clusters in the other, most regions contribute to this conservation. D. Shown are boxplots for tissue-specific expression of sensory perception and epidermal cell differentiation genes. The red line indicates the median log expression value and diamonds represent the mean. E. A sampling of the GO-enrichments at FDR<0.05 for clusters that are conserved between mEsc and mCortex.
Figure 5: Identification of interaction subnetworks in human stem cells A. The distribution of regions per chromosome for the 26 largest ESC clusters, and B. 26 largest IMR90 clusters. C. GO enrichment of clusters overlapping between IMR90 and hES tissues. Only clusters that overlap and have an enrichment in GO process are shown. D. GO enrichment of clusters conserved between tissues in human and mouse ES cells. Only conserved clusters with any GO enrichment are shown.
each directional comparison. We applied our approach to mouse ES cells and human ES cells to make sure that we were comparing similar tissues across organisms. Of the 106 mESC clusters, 7 were conserved among the 424 hES cell clusters. Interestingly, 6 of these 7 clusters were similar between human tissues and 4 of these clusters were among clusters overlapping between mouse tissues. Thus those regions that preserve their interaction partners among organisms, also tend to preserve their interactions between tissues.

3 Discussion

In this paper we have introduced a new approach to analyze the chromosome conformation capture (3C) data and established a framework to compare such comprehensive maps across multiple contexts, where a context can be different organisms or different tissues. Computational approaches for analyzing such datasets, and more importantly comparing such maps across multiple tissues are still in their infancy. The underlying philosophy of our approach is to compare sets of regions between contexts, instead of pairs of regions. Our approach first defines an interaction network using a statistical significance of two genomic loci to interact, followed by a graph clustering approach to find groups of interacting loci for each context. We first determined that our approach works at least as well in yeast, and identifies biological properties that agree with those identified by an edge-based approach, and finds additional features. However, we do miss some aspects that the edge-based approach captures, suggesting that the two approaches are complementary and should be used in concert.

Our approach for determining whether two regions interact is a hybrid approach that controls for both intra and inter-chromosomal biases [6, 10]. We revised our approach of determining statistical significance for a pair of regions to interact. We did this because we wanted to have the same approach to be universally applicable to yeast and mammalian species. In particular we control for inter and intra-chromosomal interactions. This is not as much an issue in yeast, but is important in mammalian genomes which are many times more complex and have inherent biases of high reads between regions that are close together. Our results in yeast agree with existing approaches, demonstrating the validity of our approach. By having the same method applicable for human, mouse and yeasts, we found that the percentage of inter-chromosomal interactions in yeast is much more higher (72% inter-chromosomal interactions versus 21.4% in human ES cells and 5% in mouse ES cells).

We performed a comparative analysis of human, mouse ES cells and we found significant conservation of genes that are interacting distally between the two organisms. Interestingly, we did not find any significant conservation of clusters between *S. cerevisiae* and *S. pombe* (results not shown), which are both respiro-fermentative yeasts. Multiple reasons could be responsible for this. We suspect that this is due to their evolutionary distance, ~300 mya. It is also possible that the chromosomal organizational is much more constrained in higher eukaryotes than in simple unicellular organisms.

As such maps become available for multiple cell types in multiple organisms [1, 12], methods such as ours will be increasingly useful to perform systematic comparisons of such maps to identify properties that common and specific to different cell types, and organisms, revealing principles governing organization of chromosomes from simple to complex eukaryotic species.
4 Methods

4.1 Identification of interactions among chromosomal regions

To define interactions among chromosomal regions, we segmented the chromosome into equal sized, non-overlapping bins. We acquired paired-end reads measuring three-dimensional organization of the genome for yeast [6], mouse [5] and humans [5, 10]. For the yeast data, reads were already mapped to 5KB bins. For the other organisms we tested with different numbers of bins, to try to generate a coverage similar to the yeast data. In particular, we estimated the average number of reads in the 5KB bins, and selected the bin size to have similar average numbers of reads per bin in the other datasets. We chose 350KB because it was a good balance between Lieberman-Aiden et al. and Dixon et al. data. For yeast we directly used the mapped reads to the genome. For the mouse and human data we used Maq [9] to map them to their respective genomes as necessary. We call each bin on a chromosome as a region.

We developed a uniform way to compute the random probability of two regions interact that applies to both yeast and mammalian systems. Specifically, because Hi-C experiments produce a disproportionately large number of reads for regions that are close on the same chromosome, potentially disguising distal interactions. That is, two regions that are \( n \) base pairs apart have more reads assigned to them than two regions that \( m > n \) base pairs apart where. Similar biases exist for regions within and between chromosomes. Therefore, p-values and subsequent q-values need to take into account this inherent bias into account. We here use a hybrid approach that uses Duan et al.’s approach to calculate within chromosomal interaction, but controlling for regions that are at different linear distances, and an expected count approach that controls for the relatively high number of reads for regions within the same chromosome versus regions in two different chromosomes.

Let \( r_{ic} \) denote the \( i^{th} \) region on chromosome \( c \). To define the interaction between two regions \( r_{ic} \) and \( r_{jd} \) on two different chromosomes \( c \) and \( d \) we calculate the expected number of reads for two regions as follows:

\[
m(r_{ic}, r_{jd}) = \frac{n_{ic} + n_{jd}}{\sum_{c=1}^{C} \sum_{l=1}^{R_c} n_{lc}},
\]

where \( C \) is the total number of chromosomes, \( R_c \) is the number of regions in chromosome \( c \), and \( n_{ic} \) is the number of reads between \( r_{ic} \) and every region on another chromosome. Thus the probability that the two regions interact, given that there are \( k \) reads between two regions, and there are \( n \) total number of reads that involve the two chromosomes, is given from the binomial distribution

\[
p(r_{ic}, r_{jc} | d) = \frac{n}{\sum a=p_{r_{ic}, r_{jd}}} \left( \binom{n}{a} p_{r_{ic}, r_{jd}} (1 - p_{r_{ic}, r_{jd}})^{n-a},
\right)
\]

where \( p_{r_{ic}, r_{jd}} = \left( \frac{m(r_{ic}, r_{jd})}{\sum_{a=c} m(r_{ac}, r_{ad})} \right) \).

To define the \( p \)-value of interaction of regions in the same chromosome, we used the approach of Duan et al. that accounts for the bias that nearby regions on the same chromosome tend to exhibit a large number of paired-end reads. Thus we consider pairs of regions at different distances, where the distance is defined by the number of intermediate regions between two regions. Thus two contiguous regions, \( r_{ic} \) and \( r_{(i+1)c} \) are at a distance of \( d = 1 \), where as two regions \( r_{ic} \) and \( r_{(i+2)c} \) are at distance \( d = 2 \) etc. To calculate the probability of two regions at distance \( d \) to interact we use the binomial distribution:

\[
p(r_{ic}, r_{jc} | d) = \frac{n_d}{m} \left( \binom{n_d}{m} p_d (1 - p_d)^{n_d-m},
\right)
\]

where \( p_d = \frac{1}{\sum_{c=1}^{C} R_c - d} \), \( k \) is the number of reads assigned to both regions and \( n_d \) is the total number of reads assigned to all regions that are at distance \( d \). Note that \( n_d \) is generally less than \( n_{d-1} \), so the \( p \)-value
for \( d = 1 \) will be larger than the \( p \)-value for \( d = 2 \) for the same \( k \). After calculating the \( p \)-values, we convert them to \( q \)-values and consider a \( q \)-value \(< 0.01 \) to determine whether two regions interact.

### 4.2 Markov Cluster algorithm to identify sets of co-located regions

We view the set of interacting regions as a graph where the nodes are regions and edges represent interactions between the regions. To identify densely connected subgraphs in this interaction network we used the Markov Cluster algorithm on the graph [16]. The Markov Cluster algorithm uses a parameter \( I \) for inflation that controls how many clusters one gets. To select \( I \) we first define how good a Cluster is using a modified form of the Davies-Bouldin (DB) index [2] to select the optimal \( I \). Given \( n \) clusters, the standard DB index approach is to minimize the ratio of the within cluster distance and the across cluster distance:

\[
\text{DBI} = \frac{1}{n} \sum_{i=1}^{n} \max_{j: i \neq j} \left( \frac{s_i + s_j}{m_{ij}} \right)
\]

(1)

where \( s_i \) is the “scatter” within a cluster \( i \). In other words, it measures the distance of all regions in cluster \( i \) from its centroid. \( m_{ij} \) is defined as the distance between the two clusters \( i \) and \( j \).

As our clusters are based on edges between two regions, we did not calculate a centroid to estimate \( s_i \) but rather an edge-based interaction density to define the scatter: \( s_i = 1 - \frac{2|e_i|}{v_i(v_i-1)} \), where \( v_i \) is the number regions in cluster \( i \) and \( e_i \) is the number of regions that interact in cluster \( i \). The between cluster scatter \( m_{ij} \) is defined as \( 1 - \frac{e_{ij}}{v_i v_j} \) where \( e_{ij} \) is the number of edges between regions in clusters \( i \) and \( j \). This metric allows us to define the cluster heterogeneity or scatter using the density of edges between two clusters. We add a pseudocount of 0.01 to the possible edges to avoid dividing by zero when two clusters are completely collocated.

Next we use a penalty term that on the one hand penalizes too many clusters, that is overly fragmented clusters, but also at the same time tries to avoid too few clusters. First, we only consider clusters with at least 3 regions, but we track the total number of clusters before this filtering. \( n_b \) is the total number clusters and \( n_s \) is the total number of clusters with three or less regions We select the \( I \) value that minimizes a penalized DBI, \( \lambda = \frac{n_b+n_s}{n_b \log(n_b-n_s)} \), \( \lambda \) defines the penalty term. The term in the numerator penalizes overly fragmented Clusters, while the \( \log(n_b-n_s) \) term in the denominator penalizes Clusters resulting in few large clusters.

### 4.3 Calculating enrichment of biological categories clusters from a single interaction map.

We evaluate MCL clusters for enrichment with sets of biological categories by calculating a \( p \)-value based on a random set background distribution. For the purpose of this description, we use region and gene interchangeably. We first create five randomized versions of the MCL clusters while maintaining chromosome distribution, as in the Witten and Noble approach [18]. After removing regions unique to either the MCL cluster or the query set of interest, we produce five randomizations of the functional regions using weighted selection without replacement from the full set of regions while maintaining chromosome distribution. The likelihood for selecting a particular gene or region is proportional to the number of times that region appears in the functional set (relevant when, for instance, two TFs regulate the same gene). We calculate hypergeometric \( p \)-values for the overlap between a cluster and functional sets for both the random and non-random data using the python rpy2 package. We use the random \( p \)-values to create a distribution and assign a final \( p \)-value corresponding to the percent of random \( p \)-values that are less than the non-random value.
4.4 Calculating conserved co-locations between different tissues and species

Our method for calculating enrichment between clusters from the HiC networks of different tissues resembles the approach to calculating enrichment between MCL clusters and various biological categories. We always exclude clusters that do not contain at least four regions from our analysis. We compare the number of overlapping regions between each cluster from the first and second set exactly as described above. When comparing clusters from different species, we first map regions to genes that are closest to the region. Multiple genes can be mapped to the same region. We found on average 2 genes per region in yeast, 5 genes per region in mouse and 8 genes per region in mouse. Using known orthology between two species, we produce a many-to-many mapping of genes from one species to the other [13]. We calculate \( p \)-values for the overlap of orthologous genes from clusters in the first species to the second using hyper-geometric distributions. If two genes from the same cluster in one species are orthologous to the same gene in the second species, we consider it only once. We apply the Benjamini-Hotchberg correction for the estimated \( p \)-values.

5 Acknowledgement

This work was supported by startup funds from UW Madison and Wisconsin Institutes of Discovery to S.R. and Morgridge research fellowship to R.A. We thank Rupa Sridharan for helpful discussions.
References


