Bioinformatics, 32, 2016, i111–i120
doi: 10.1093/bioinformatics/btw256
ISMB 2016

OXFORD

A novel method for discovering local spatial clusters of genomic regions with functional relationships from DNA contact maps

Xihao Hu[1,†], Christina Huan Shi[1] and Kevin Y. Yip[1,2,3,4,*]

[1]Department of Computer Science and Engineering, [2]Hong Kong Bioinformatics Centre, [3]CUHK-BGI Innovation Institute of Trans-omics, [4]Hong Kong Institute of Diabetes and Obesity, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong

[†]Present address: Key Laboratory of RNA Biology, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China

*To whom correspondence should be addressed.

Abstract

Motivation: The three-dimensional structure of genomes makes it possible for genomic regions not adjacent in the primary sequence to be spatially proximal. These DNA contacts have been found to be related to various molecular activities. Previous methods for analyzing DNA contact maps obtained from Hi-C experiments have largely focused on studying individual interactions, forming spatial clusters composed of contiguous blocks of genomic locations, or classifying these clusters into general categories based on some global properties of the contact maps.

Results: Here, we describe a novel computational method that can flexibly identify small clusters of spatially proximal genomic regions based on their local contact patterns. Using simulated data that highly resemble Hi-C data obtained from real genome structures, we demonstrate that our method identifies spatial clusters that are more compact than methods previously used for clustering genomic regions based on DNA contact maps. The clusters identified by our method enable us to confirm functionally related genomic regions previously reported to be spatially proximal in different species. We further show that each genomic region can be assigned a numeric affinity value that indicates its degree of participation in each local cluster, and these affinity values correlate quantitatively with DNase I hypersensitivity, gene expression, super enhancer activities and replication timing in a cell type specific manner. We also show that these cluster affinity values can precisely define boundaries of reported topologically associating domains, and further define local sub-domains within each domain.

Availability and implementation: The source code of BNMF and tutorials on how to use the software to extract local clusters from contact maps are available at http://yiplab.cse.cuhk.edu.hk/bnmf/.

Contact: kevinyip@cse.cuhk.edu.hk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Although genomes are commonly depicted as a linear sequence of base pairs, they actually contain complex three-dimensional (3D) structures. There has long been an interest in identifying biologically meaningful territories from these structures (Cremer and Cremer, 2010). Fluorescence in situ hybridization (FISH) has provided insights into genome organization by visualizing individual pairs of spatially interacting genomic regions (Cremer and Cremer, 2001), but it has limited resolution and cannot be scaled up to study a large number of interacting regions at the same time. These limitations have been overcome by high-throughput experimental methods that can probe interacting regions genome-wide by means of cross-linking and deep sequencing, such as ChIA-PET, Hi-C and TCC (Fullwood *et al.*, 2009; Kalhor *et al.*, 2012; Lieberman-Aiden *et al.*, 2009). Each sequencing read produced provides information about two interacting regions. The whole set of reads is summarized by a contact map in the form of a square matrix, where the whole genome is binned into contiguous genomic locations with each bin

corresponding to a row and a column of the matrix, and each matrix element represents the number of supporting reads linking up regions from the two respective bins, which we refer to as the 'contact counts' (Supplementary Fig. S1). Two bins that are closer to each other in the 3D genome structures in a cell population receive a larger contact count in general, subject to various types of bias to be discussed below.

The appropriate size of each bin depends on the amount of data produced. If a small bin size is used but there are not enough sequencing reads, many values in the matrix would be small or even zero, and would thus be heavily affected by background noise and random sampling effects. Therefore, in order to produce contact maps at high resolution, the amount of data produced in each experiment has been increasing rapidly (Ay *et al.*, 2014; Dixon *et al.*, 2012; Jin *et al.*, 2013; Kalhor *et al.*, 2012; Lieberman-Aiden *et al.*, 2009). A recent *in situ* Hi-C dataset contains billions of pairwise interactions obtained from a human cell line, which allows for a high resolution of 1 kb bin size while the average contact count is still sufficiently large for proper analyses (Rao *et al.*, 2014). These massive datasets have created computational challenges in extracting useful information about the underlying 3D genome structures.

Various methods have been applied to analyze DNA contact maps with different goals. Statistical methods have been used to test whether a set of genomic locations of interest are co-localized in the 3D genome structure (Duan *et al.*, 2010). Their reliability depend on the normalization procedure (Cournac *et al.*, 2012) and the suitability of the null model (Paulsen *et al.*, 2013). Conversely, methods have been proposed for discovering groups of interacting genomic regions without a predefined set of candidates. One way is to create a global model for the whole 3D genome structure using the contact map to define model constraints. Existing methods differ from each other by their assumptions about global genome structures, their optimization procedures, and whether a single structure or a population of structures is used to explain the observed contact map (Duan *et al.*, 2010; Kalhor *et al.*, 2012; Lieberman-Aiden *et al.*, 2009; Nagano *et al.*, 2013; Sexton *et al.*, 2012). Another way to identify groups of interacting regions is to cluster genomic regions according to their contact counts with other regions. One previous study used principle component analysis (PCA) to identify a low-dimensional representation of the contact map based on which genomic regions were clustered. The clustering results revealed general open and close compartments of the genome at 1 Mb resolution (Lieberman-Aiden *et al.*, 2009). More refined topologically associating domains (TADs) (Dixon *et al.*, 2012; Nora *et al.*, 2012) at 100 kb resolution were later identified by using hidden Markov models. Recently, chromatin loops at even higher resolutions were identified using dynamic programming. Clustering of the inter-chromosomal interactions revealed six general groups using a Gaussian hidden Markov model clustering algorithm (Rao *et al.*, 2014).

In general, the accuracy of these methods depends on how the contact map is processed to remove biases. Some genomic regions tend to have more contact counts than others, caused by factors such as GC content and uniqueness of sequence (Yaffe and Tanay, 2011). These factors lead to a non-uniform distribution of contact read counts, which could confuse the analysis and should be properly corrected. In an early correction method, the observed contact counts are normalized by the expected counts based on a background model (Lieberman-Aiden *et al.*, 2009). An iterative correction approach was later proposed to enforce equal sums for all rows in the contact map after normalization, assuming that after removing the biases, every bin should have approximately the same number of interactions with other bins. This normalization method was proved to be more effective and have a better convergence property (Imakaev *et al.*, 2012). Using the same framework, another effective normalization method adopts a new updating rule by dividing row sums with their Euclidean norms (Cournac *et al.*, 2012). Yaffe and Tanay's work followed a different direction by explicitly modeling three types of biases observed in Hi-C data (Yaffe and Tanay, 2011). This method was further sped up by relaxing the objective function (Hu *et al.*, 2012). In the recent high resolution study, a faster iterative approach based on matrix balancing was used to achieve the assumption of equal row sums (Knight and Ruiz, 2013; Rao *et al.*, 2014).

Together, these data correction and analysis methods have led to many interesting findings about genome structures. On the other hand, so far most analyses have focused on either individual interactions, domains composed of a contiguous segment on the primary sequence, or grouping of these domains into general domain categories based on some global properties of the contact map. Given the complexity of genome structures, it would be useful to have a way to flexibly identify local clusters of genomic regions that are spatially close in the 3D structure but not necessarily adjacent in the primary sequence, such as multiple non-adjacent TADs that form a higher-level spatial cluster. As to be shown in the Results section, many of these local clusters are related to particular molecular activities such as transcription and DNA replication. Compared to the domains identified in previous studies, these local clusters could provide information about local organizations of the genome structure involving small subsets of genomic regions that are particularly related to each other within larger domains, thereby supplementing the previous methods.

In this study, our main goal is to identify these local clusters from contact maps obtained from Hi-C experiments using matrix factorization. In general, matrix factorization aims at decomposing a matrix into two or more matrices with a certain objective. One of the most well-known methods is eigenvalue decomposition, which decomposes a diagonalizable square matrix X into $Q\Lambda Q^{-1}$, where Q contains the orthogonal eigenvectors and Λ is a diagonal matrix containing the eigenvalues. A popular application of it is principal component analysis, which uses eigenvalue decomposition to factorize the covariance matrix of a dataset. Each eigenvector (i.e. a principal component, PC) represents the direction with the largest data variance after subtracting out the projections on the previous PCs with larger eigenvalues. PCA has been used to identify key contributing factors of Hi-C contact maps (Lieberman-Aiden *et al.*, 2009). Since each bin could have a negative coordinate along a PC, the biological meanings of the PCs are sometimes difficult to interpret. In one of the studies, permutation tests have also indicated that only the first few PCs were statistically significant (Imakaev *et al.*, 2012).

Conceptually, a perfect DNA contact map can be considered as the superposition of contact counts from different local clusters and a small number of counts linking different clusters. For example, the contact map submatrix shown in Supplementary Figure S1 can be well approximated by read counts within the first local cluster (consisting of bins a, b and c) and read counts within the second local cluster (consisting of bins d and e). Each bin can have an affinity value indicating its participation in each local cluster, and bins at the interface of multiple clusters (bins c and d) can have partial affinities to these clusters. When only the contact counts between different bins are available but not the 3D coordinates of each bin in the genome structure, these local clusters can be identified by non-negative matrix factorization (NMF), which decomposes the data matrix into matrices containing only non-negative values (Devarajan, 2008). For a DNA contact map, after considering some

global effects due to general open and close domains or chromosome structures, the remaining factors could correspond well to the local clusters. Comparing with analyzing DNA contact maps with PCA, NMF enjoys the advantages of (i) having clear biological interpretations for many factors, (ii) allowing only positive coefficients (i.e. the affinity values) and (iii) focusing more on local sub-structures. In fact, NMF was popularized by Lee and Seung's work on image processing (Lee and Seung, 1999), which clearly demonstrated that each image of human face can be automatically factorized by NMF into distinct facial features, such as the nose, the mouth and so on. NMF has also been shown effective in many biological applications such as analyzing protein-protein interactions (Greene *et al.*, 2008).

As discussed, DNA contact maps contain various types of bias that should be properly corrected before applying NMF. Here, we propose a novel method called balanced non-negative matrix factorization (BNMF), which was specially designed for DNA contact maps to handle data biases and relationships between data bins in the primary sequence when performing NMF. We show that BNMF can identify compact spatial clusters of genomic regions according to simulated data that highly resemble real Hi-C data. The clusters were identified without making any assumptions about the whole genome structure. They are found to contain genomic regions closer to each other than those identified by some other clustering methods previously used to analyze Hi-C data. We further describe how statistical tests can be performed to check whether a group of genomic regions of interest are spatially close to each other in the 3D genome structure based on the clusters identified by BNMF. We use this testing procedure to confirm that various functionally related groups of genomic regions are spatially close to each other. We then show that the BNMF clusters are not only good indicators of particular groups of functionally related genomic regions, but their affinity values are actually quantitatively correlated with various molecular activities. Finally, we show that BNMF cluster affinities can precisely define boundaries of TAD, and provide more fine-grained information about sub-domains within these large domains.

2 Results

2.1 BNMF identifies local spatial clusters from DNA contact maps

BNMF takes a contact map X as input, and finds an approximation Y of it that can be decomposed into the product of several matrices:

$$X \approx Y = BHSH^TB,$$

where B is a diagonal matrix that captures the position-specific contact count biases, S is a diagonal matrix that defines the clusters, and H contains the cluster membership values (Fig. 1). B, S and H all contain only non-negative values. To understand this formula, first we define a new matrix R as follow:

$$R := B^{-1}YB^{-1} = HSH^T$$

R is a balanced contact map required to have equal sums for all rows and equal sums for all columns except for those bins with no contact counts in the original contact map X. BNMF then looks for non-negative matrices H and S such that $R = HSH^T$. This decomposition does not have a unique solution. For instance, if every element in H is multiplied by a positive value and each element in S is divided by the square of that value, HSH^T would stay unchanged. To make the decomposition result easy to interpret, we require each column of H to sum to a given constant, and correspondingly each column of $W = SH^T$ to sum to another constant (Fig. 1). With this constraint, each column of H contains the memberships of different bins to a cluster when each cluster has a fixed membership quota for all the bins, and each column of W contains the affinity of a bin to join the different clusters when each bin has a fixed affinity budget for all the clusters.

To illustrate the decomposition procedure, consider the following raw contact map:

$$X = \begin{pmatrix} 8 & 4 & 0 \\ 4 & 2 & 0 \\ 0 & 0 & 9 \end{pmatrix}$$

In this idealized example, we can find Y that is identical to X. The position-specific biases can be removed by the bias matrix B:

$$\begin{aligned} R = B^{-1}YB^{-1} &= B^{-1}XB^{-1} \\ &= \begin{pmatrix} 4 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}^{-1} \begin{pmatrix} 8 & 4 & 0 \\ 4 & 2 & 0 \\ 0 & 0 & 9 \end{pmatrix} \begin{pmatrix} 4 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}^{-1} \\ &= \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0.5 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{aligned}$$

R is in turn decomposed into HSH^T, where each column in H and W sums to 1:

$$R = HSH^T = \begin{pmatrix} 0.5 & 0 \\ 0.5 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$W = SH^T = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

From matrix S, the two diagonal entries state that there are two clusters, with two members belonging to the first and one member to the second. From matrix H, the first cluster assigns 0.5 membership to bin 1 and 0.5 membership to bin 2, and the second cluster assigns 1 membership to bin 3. From matrix W, bins 1 and 2 both have full affinity to cluster 1 and bin 3 has full affinity to cluster 2.

We developed an algorithm that searches for B, S and H at the same time given an input contact map X. This algorithm also uses the location of the bins in the primary sequence to define a manifold component in the objective function, such that the resulting affinity values are smooth (Section 4, Supplementary Fig. S2).

When applying BNMF to a published yeast Hi-C contact map (Duan *et al.*, 2010), the identified clusters contain adjacent bins in the primary sequence, as evidenced by blocks of large values in each row of W (Fig. 1); On the other hand, there are also clusters with member bins not adjacent in the primary sequence but have high contact counts between them, seen as off-diagonal blocks. Overall, when all clusters are considered, the approximated contact map Y is highly similar to the raw contact map X.

An important parameter of BNMF is the number of clusters to form, which is equal to the number of rows in matrix S. We have developed an automatic procedure to search for an appropriate number of clusters such that each cluster is sufficiently homogeneous but not too fragmented (Section 4).

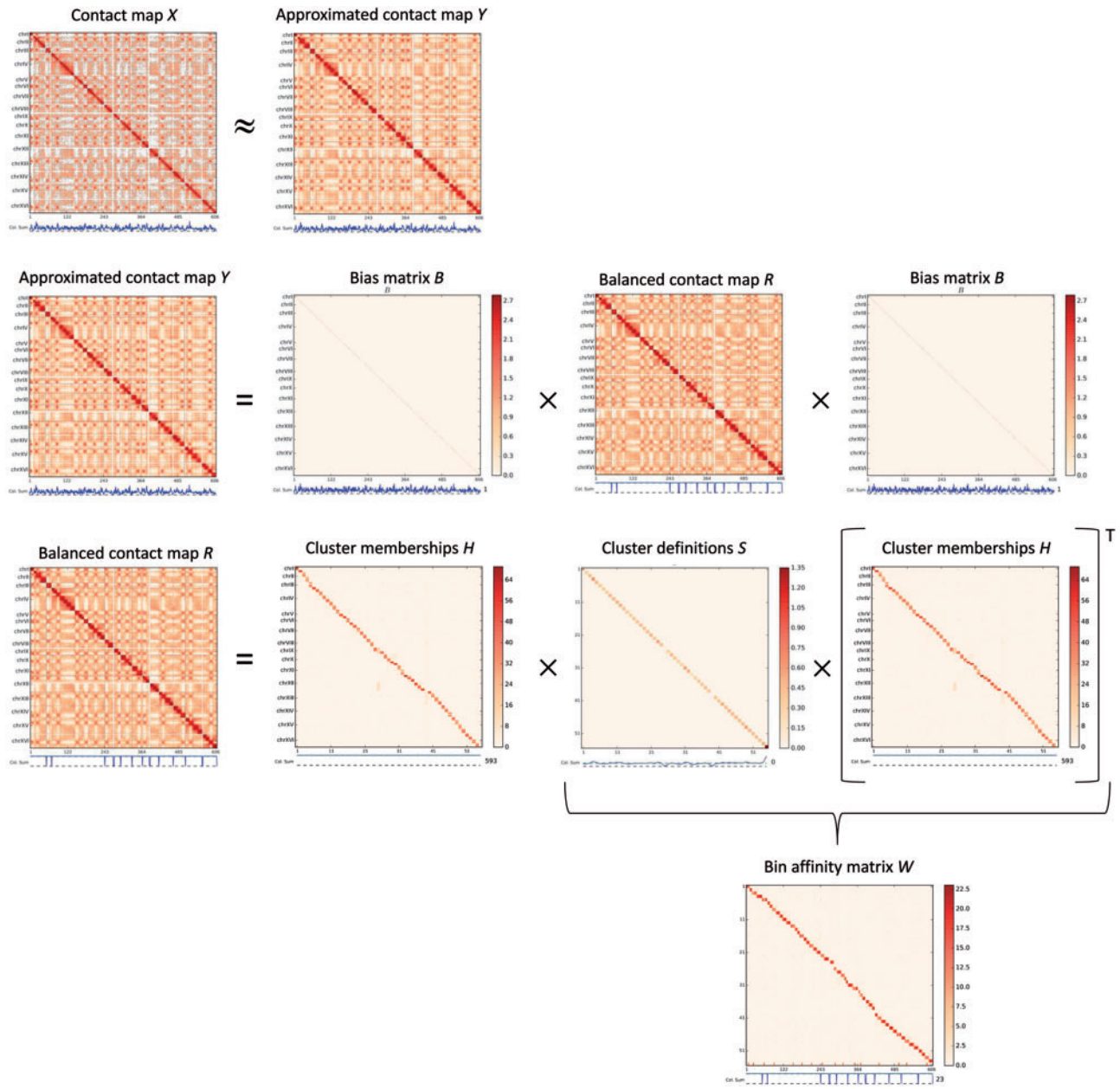

Fig. 1. Summary of the BNMF method. The Hi-C contact map based on HindIII interactions at 1% FDR from Duan *et al.* (2010) is used as the input contact map *X*. The goal of BNMF is to find a matrix *Y* that can be decomposed into the product of several matrices and is close to *X*. *Y* contains similar position-specific biases as *X*, as shown by the column sums below the heat maps. These biases can be captured by a bias matrix *B*. When these biases are removed from *Y*, we get a balanced contact map *R*. *R* is in turn decomposed into HSH^T, where *S* defines the local spatial clusters and *H* contains the membership values that associate the bins to the clusters

2.2 BNMF produces spatial clusters more compact than other clustering methods

We used two methods to test whether BNMF produces clusters that contain spatially proximal genomic regions. In the first method, we used a space-filling Hilbert curve to draw an artificial 2D chromosome in a 16×16 area. Each of the 256 points mimics a bin of consecutive genomic locations. We then generated an idealized synthetic contact map by assigning $\frac{d_{\max}}{(1+d)^2}$ contact counts between any two points at a Euclidean distance of d in the 2D structure not necessarily adjacent in the 1D sequence, where $d_{\max}$ is the maximum Euclidean distance between any two points on the curve. The reason for having contact counts inversely proportional to d^2 in this 2D chromosome is based on previous work of the 3D case in which contact counts are inversely proportional to d^3 due to the volume in 3D space (Varoquaux *et al.*, 2014). We next decomposed the matrix using eigenvalue decomposition (EIG) and BNMF, respectively. For EIG, we formed four clusters of points that have coefficients larger than the corresponding means along the first four eigenvectors. For BNMF, we formed four clusters comprising points with an affinity value larger than the mean.

Figure 2 shows the resulting clusters produced by the two methods. For EIG, the first cluster separates the interior and exterior of the structure. The second and third clusters separate points at different sides of the whole area. In order to be orthogonal to the first

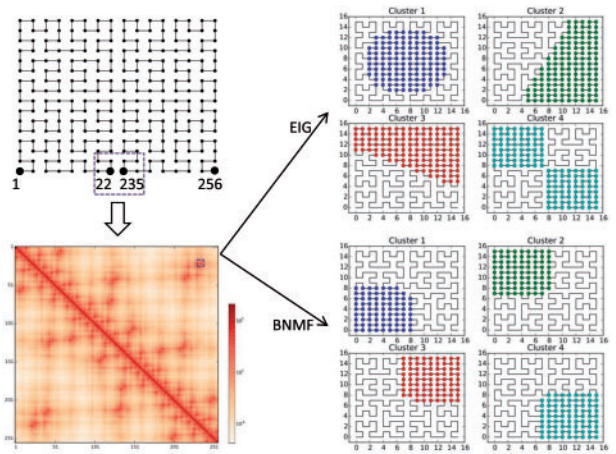

Fig. 2. Comparison between eigenvalue decomposition and BNMF based on a space-filling curve. A Hilbert curve is used to generate an artificial 2D chromosome and a corresponding synthetic contact map. The contact map is decomposed by eigenvalue decomposition (EIG) and BNMF. In both cases, four clusters are obtained by grouping the points with the highest coefficients along the eigenvectors or affinity values to the clusters. A region with long-range contacts, including points 22 and 235, is highlighted in the Hilbert curve, the heat map and cluster 4 produced by BNMF

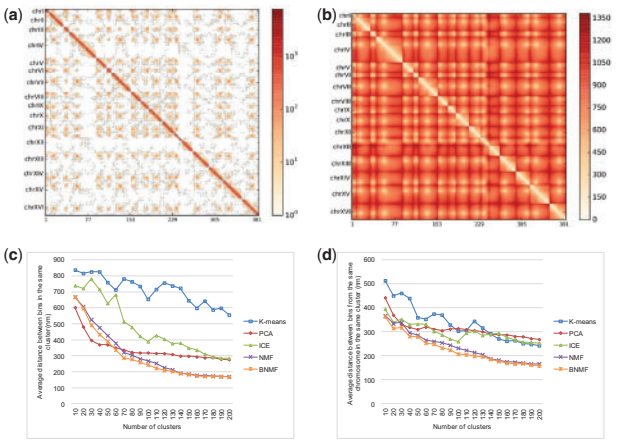

Fig. 3. Comparison between five different clustering methods based on the noisy synthetic yeast Hi-C contact map. (**a**) The contact map obtained from 3000 simulated yeast genome structures with random biases added to each bin. (**b**) The corresponding average Euclidean distances between different bins in the 3000 structures. (**c**) The average distance between bins in the same cluster. (**d**) The average distance between bins in the same cluster, considering only bin pairs from the same chromosome

three eigenvectors, the fourth eigenvector results in a cluster that involves two isolated parts of the structure. The four clusters overlap each other substantially and are all based on some global data properties. In contrast, the clusters formed by BNMF clearly correspond to spatial clusters at the four corners. These clusters include points that are spatially close in the structure but not adjacent in the primary sequence. For example, points 22 and 235 are far away in the primary sequence, but due to their spatial locality, they have a high contact count, and are grouped into the same cluster. Similar results could also be obtained from a 3D Hilbert curve. While BNMF is not intended for identifying global compartments as was done in some previous studies (Lieberman-Aiden *et al.*, 2009), this example clearly illustrates how BNMF identifies spatial clusters.

In the second method, we generated a noisy synthetic yeast Hi-C contact map with position-specific biases, and used it to compare different clustering methods (Section 4). A volume exclusion model was previously proposed for generating a population of simulated 3D structures of the yeast genome, which was shown to produce a contact map similar to one obtained from real data (Tjong *et al.*, 2012). We used this model to generate 3000 simulated yeast genome structures at 3.2 kb resolution. The resulting contact map has data patterns highly similar to a real contact map at 32 kb resolution (Supplementary Fig. S3a,b). The average Pearson correlation between the rows in the two matrices is larger than 0.9 (Supplementary Fig. S3e). When we decomposed these two matrices using BNMF, the cluster affinity values were highly similar (Supplementary Fig. S3c,d), with clear correspondence between most clusters obtained from the two maps (Supplementary Fig. S3f). Having this highly realistic synthetic contact map, we used it to check whether the bins grouped into the same cluster by BNMF are spatially close, based on the known average distance between any two bins in the 3000 simulated structures.

In addition to BNMF, we also applied four other methods to produce clusters from the noisy synthetic contact map. The first method is PCA with a normalization step for removing position biases (Lieberman-Aiden *et al.*, 2009), in which case each PC was used to define a cluster (see below). The second method is Iterative Correction and Eigenvalue Decomposition (ICE) (Imakaev *et al.*, 2012), which uses an iterative procedure to remove biases in the contact map before applying eigenvalue decomposition. The third method is the standard k-means clustering method (MacQueen, 1967), in which case each row vector was used as the features of the corresponding bin. The fourth method is standard NMF without correcting for the biases. We required all methods to produce the same number of clusters, with each bin assigned to exactly one cluster.

Figure 3a shows the synthetic contact map and Figure 3b shows the corresponding pairwise Euclidean distances between the bins based on the actual synthetic structures. The centromeres of the different chromosomes are in general close to each other as previously reported in whole-genome models (Duan *et al.*, 2010). For each cluster, we computed the average distance between the member bins (Fig. 3c). For most cluster numbers, the clusters identified by BNMF were the tightest. We found that some methods tended to put only bins from the same chromosome in the same cluster. If we focused on only bin pairs from the same chromosome, the average distances were again smallest for BNMF (Fig. 3d).

Since BNMF uses an iterative optimization procedure, we checked how the objective score and the average Euclidean distance between bins assigned to the same cluster change across the iterations. Supplementary Figure S4 shows that the objective score decreased progressively across iterations as expected. Correspondingly, the average distance between the bins in the same cluster also decreased, and it had a good correlation with the objective score, suggesting that the objective function is a good indicator of intra-cluster bin distances in DNA contact maps.

Taking the results of both analyses together, the clusters identified by BNMF are shown to represent compact groups of genomic bins coming from spatially proximal regions.

2.3 BNMF confirms spatial locality of functionally related genomic regions

As an application of BNMF, we developed statistical tests for checking whether a given set of genomic regions are spatially proximal using the clusters identified by BNMF (Section 4). The basic idea is that if a set of genomic regions are spatially close, they would have large affinity values to a cluster simultaneously. The affinity values to this cluster can thus well separate bins belonging to this group of

genomic regions from other bins. Our testing procedure performs this analysis for every cluster, and reports a *P*-value corrected for multiple hypothesis testing. One major difference between our testing procedure and the ones used in previous studies is that we use cluster affinity values to check the spatial locality of the bins of interest while most previous methods use contact counts directly. If a set of genomic regions of interest simultaneously have a high affinity to a cluster, other genomic regions with a high affinity to this cluster are also spatially co-localized with the regions of interest, thus providing a simple way to discover new spatial relationships.

To check the effectiveness of our testing procedure, we applied it to check the co-localization of previously reported groups of genomic regions, including early replication sites in yeast (Imakaev *et al.*, 2012), VRSM genes in the parasite Plasmodium falciparum (Ay *et al.*, 2014) and tRNA genes in human, mouse and yeast (Duan *et al.*, 2010). Table 1 shows the resulting *P*-values. All these three previously reported functional groups were found to contain bins that could be well distinguished from other bins based on the cluster affinity values of BNMF. As a control, we also applied the testing procedure on late replication sites, which are not expected to be co-localized. Indeed, the resulting *P*-value was not significant. These results demonstrate BNMF's ability to test the spatial locality of functionally related genomic regions, and its general applicability to data from different species.

Table 1. Spatial locality of various functional groups of genomic regions

Genomic regions	Species	Corrected *P*-value
Early replication sites	Yeast	<0.001
Late replication sites (control)	Yeast	0.3
VRSM genes	*P. falciparum*	<0.001
tRNA genes	Human	<0.001
tRNA genes	Mouse	<0.001
tRNA genes	Yeast	0.03

Several groups of functionally related genomic regions previously reported to be spatially proximal were re-tested for their spatial co-localization using BNMF cluster affinity values based on the CCD method. Late replication sites, which are not expected to be co-localized in the 3D genome structure, were used as a control.

2.4 BNMF cluster affinities quantitatively correlate with genomic features in a cell type specific manner

For functionally related regions, if their activities differ in different cell types, their spatial locality may also change accordingly. To test if such cell type specific structural information is captured by the BNMF clusters, we collected Hi-C data from four human cell lines and used BNMF to identify local spatial clusters from them. We also collected cell type specific experimental data that indicate various molecular activities, including genes with cell type specific expression, locations of super enhancers, DNase I hypersensitivity values and replication time (Section 4). The first two types of data identify genomic regions belonging to these categories, and we used the same statistical procedure described in the last section to test the spatial locality of these bins. The last two types of data provide numeric values for each bin across the whole genome. We used a modified testing procedure to check if these values were quantitatively correlated with cluster affinity values.

Table 2 shows the analysis results. In 61 of these 64 cases, the correlation between the cluster affinity values and the genomic features is statistically significant, except for the three cases involving highly expressed genes in K562 and local clusters from other cell lines, indicating that genomic regions defined by these features are in general close to each other regardless of their activities. On the other hand, in many cases the highest correlations (in terms of AUC or SPC, defined in Section 4) are observed when both the DNA contact map and the genomic feature were obtained from the same cell type. For example, the genes specifically expressed in a cell type appear to be most structurally proximal in the same cell type. The results for DNase I hypersensitivity and replication time further show that the BNMF cluster affinity values not only can help identify the local clusters related to particular genomic features, but also quantitatively reflect the activity level of some features. These results indicate that cell type specific structural changes related to differential molecular activities are captured by the BNMF clusters.

Table 2. Correlation between genomic features and BNMF cluster affinities in human cell lines

	Highly expressed genes (AUC)				Super enhancers (AUC)			
Hi-C	GM12878	h1-hESC	IMR90	K562	GM12878	H1-hESC	IMR90	K562
GM12878	**0.79****	0.76**	0.68**	0.68	**0.82****	0.64**	0.76**	0.82**
H1-hESC	0.75**	**0.80****	0.67**	0.71*	0.72**	0.67**	0.73**	**0.79****
IMR90	0.71**	0.73**	**0.75****	0.70*	0.73**	0.63**	**0.80****	0.76**
K562	0.73**	0.77**	0.64**	**0.81****	0.73**	0.63**	0.71**	**0.79****
	DNase I hypersensitivity (SPC)				Replicating time (SPC)			
Hi-C	GM12878	h1-hESC	IMR90	K562	GM12878	H1-hESC	IMR90	K562
GM12878	**0.91****	0.82**	0.84**	0.84**	**0.82****	0.67**	0.66**	0.75**
H1-hESC	0.79**	**0.83****	0.82**	0.78**	0.64**	0.65**	0.55**	**0.68****
IMR90	0.69**	0.65**	**0.72****	0.65**	0.58**	0.53**	**0.74****	0.57**
K562	0.75**	0.73**	0.75**	**0.77****	0.63**	0.58**	0.56**	**0.70****

The four sub-tables show the correlations between BNMF cluster affinities and four types of genomic features. For the first two features (genes highly expressed in specific cell lines and super enhancers), the correlations are quantified by the area under the receiver operator characteristics curve (AUC). For the last two features (DNase I hypersensitivity and replication time), the correlations are quantified by the Spearman rank correlation (SPC). In each sub-table, the rows correspond to the cell types from which the Hi-C data were obtained, and the columns correspond to the cell types from which the genomic features were obtained. The highest value in each row is highlighted in bold face. Corrected *P*-values, $*P < 0.05$, $**P <$ 1E-5.

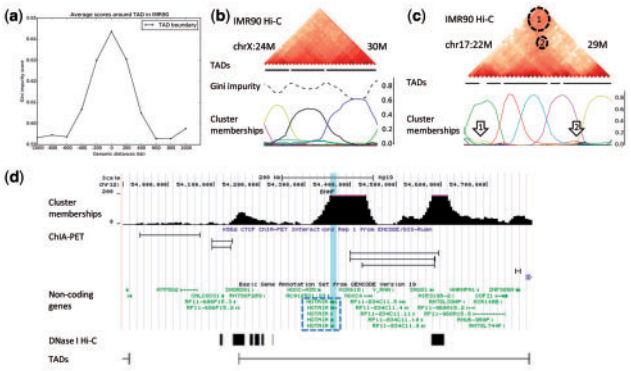

Fig. 4. Relationships between BNMF clusters and TADs. (**a**) Bins around TAD boundaries have increased Gini impurity scores, indicating that they are also at the boundaries of BNMF clusters. (**b**) An example showing a simple case in which each TAD corresponds to exactly one BNMF cluster. In the cluster affinity panel, each color represents a different cluster. (**c**) A more complex example showing that some TADs contain multiple clusters, and some long-range DNA contacts across multiple TADs (circles 1 and 2) appear as small peaks of cluster affinity values (marked by arrows 1 and 2). (c) An example showing a BNMF cluster with almost the same boundaries with a TAD, but the affinity values further show the long-range contacts between the HOTAIR gene and two distal regions that have been independently shown by ChIA-PET or DNase I Hi-C

2.5 Local clusters identified by BNMF are consistent with and supplement topologically associating domains

Previous work has identified topologically associating domains (TADs), consecutive blocks of DNA locations that define structural domains (Dixon *et al.*, 2012; Nora *et al.*, 2012). We investigated how the clusters identified by BNMF are related to TADs.

We first reasoned that if TADs mark structural domains, the boundaries of some TADs should also be boundaries of BNMF clusters. We identified BNMF cluster boundaries by checking the affinity values of the bins. Bins at cluster boundaries have affinities to multiple clusters, and thus their vectors of cluster affinities have a higher Gini impurity index (Section 4). When plotting all these Gini impurity values from a contact map from the human IMR90 cell line, a clear enrichment is seen around TAD boundaries (Fig. 4a). In general, there is substantial overlap between TAD and BNMF cluster boundaries (Supplementary Fig. S5a).

By inspecting individual TADs, we found that in some simple cases each TAD corresponds to exactly one BNMF cluster (Fig. 4b). In some more complex cases, one TAD contains multiple BNMF clusters, and long-range contacts across TADs are captured by the BNMF affinity values and appear as small distal peaks (Fig. 4c, Arrows 1 and 2, Supplementary Fig. S5b). Alternatively, some BNMF clusters span multiple TADs, suggesting the spatial proximity between them as indicated by the contact map (Fig. 4c, green cluster, Supplementary Fig. S5c).

One feature of BNMF is that it can identify clusters at different resolution by decomposing contact maps at different bin sizes and/or setting different number of clusters. We demonstrate this feature by extracting the balanced contact map of a segment (from 53 to 56 Mb) of chromosome 12 from the human K562 cell line at 5 kb resolution (Rao *et al.*, 2014), and using BNMF to produce local spatial clusters. There was a recent study using a novel DNase Hi-C method to study the 3D genome structures around long intergenic noncoding RNAs of human cell lines (Ma *et al.*, 2015). Two distal regions were clearly observed to have interactions with the promoter of the HOTAIR gene, one of which is also supported by ChIA-PET (Fig. 4d). We checked the BNMF clusters around this genomic locus, and found a cluster with almost identical boundaries with a surrounding TAD. The two distal interacting regions have strong enrichment in the affinity values of this cluster as compared to other regions within the TAD, showing that BNMF identified these two distal regions and the HOTAIR gene to be particularly close among all the bins in this TAD.

These results suggest that the BNMF clusters are consistent with TADs, but also provide additional information about more detailed interactions within a single TAD and across multiple TADs.

3 Discussion and conclusion

In this paper, we have described the BNMF method for decomposing a contact map and identifying local spatial clusters from it. We have used synthetic contact maps based on both an artificial space-filling chromosome and simulated yeast genome structures that resemble real yeast genome structures to show that the clusters identified by BNMF correspond to compact clusters of genomic regions proximal in the 3D genome structure. Genomic regions in the same spatial clusters are usually, but not necessarily, close to each other in the primary sequence. Comparisons with other clustering methods showed that the clusters produced by BNMF are more compact. This is partially attributed to the non-negative requirements for the decomposing matrices, which naturally attempts to explain the overall contact map by the summation of contact counts from different local clusters. In contrast, if the decomposing matrices allow negative values, as in eigenvalue decomposition, the resulting clusters may not correspond to local spatial clusters and their meanings could be difficult to interpret.

Since BNMF does not require any prior knowledge about the overall genome structure and does not make any assumptions about it, the identified clusters are based purely on the contact map data and could contain novel groups of genomic regions. We have shown that BNMF can be applied to data obtained from a variety of species.

One could also apply BNMF to the same contact map at different bin sizes, to identify spatial clusters at different resolution. Currently one limitation of BNMF is that when the input contact map is too large, the computation involved in the decomposition could be prohibitive due to expensive matrix multiplications. One way to deal with this problem is to extract a subset of the contact map of interest, such as one chromosome, and apply BNMF on this subset only. When we used this strategy to study the sub-contact map around the HOTAIR gene at 5 kb resolution, we were able to identify the enhancer-promoter interactions around the HOTAIR gene within a larger TAD. Some other technical methods for speeding up the calculations are discussed in the Supplementary Materials.

4 Materials and Methods

4.1 Formulation of balanced non-negative matrix factorization, BNMF

Given a DNA contact map X, the BNMF method looks for positive matrix B and non-negative matrices H and S such that $X \approx BHSH^TB$. This searching process requires an objective function to evaluate how similar the reconstructed matrix $BHSH^TB$ and the input matrix X are, and an algorithm for finding matrices B, H and S that result in a good objective score. Here, we provide the details of these two components of BNMF.

4.1.1 Defining the basic objective function

We begin with the simplified problem of decomposing X into HSH^T, assuming X is bias free. Later we will explain how the bias matrix B can be identified to remove position-specific biases in X.

Let $W := SH^T$ be a non-negative matrix. In order to approximate X by HW, we assume that X is produced by adding Poisson noise to HW. Maximizing the data likelihood is related to minimizing the Kullback–Leibler (KL) divergence between X and HW (Lee and Seung, 1999) defined as follows:

$$d(X||HW) = \sum_{i,j}\left(X_{ij}\ln\frac{X_{ij}}{(HW)_{ij}} - X_{ij} + (HW)_{ij}\right)$$
$$= \sum_{i,j}\left[(HW)_{ij} - X_{ij}\ln(HW)_{ij}\right] + \sum_{i,j}\left[X_{ij}\ln X_{ij} - X_{ij}\right]$$

Since only the first term varies with HW, minimizing the KL divergence is equivalent to minimizing the following objective function:

$$J_{Poi} = \sum_{i,j}\left[(HW)_{ij} - X_{ij}\ln(HW)_{ij}\right]$$

We also add to the objective function a manifold component, such that bins adjacent to each other in the primary sequence are more likely to have similar cluster membership values. The details are given in Section S1.1.1.

4.1.2 Algorithm for solving X = HW

For the optimization problem based on the basic objective function J_{Poi} with non-negative constraints on H and W, locally optimal solutions can be found by the Multiplicative Update Rules (MUR) (Lee and Seung, 1999). These rules are obtained by gradient descent.

The partial derivative of the objective function J_{Poi} with respect to any element H_{ab} in H is

$$\frac{\partial J_{\text{Poi}}}{\partial H_{ab}} = \sum_{ij}\frac{\partial (HW)_{ij}}{\partial H_{ab}} - \sum_{ij}\frac{X_{ij}}{(HW)_{ij}}\frac{\partial (HW)_{ij}}{\partial H_{ab}}$$
$$= \sum_{j} W_{bj}\left(1 - \frac{X_{aj}}{(HW)_{aj}}\right)$$
$$= \sum_{j}\left(1 - \frac{X}{HW}\right)_{aj} W_{bj}$$
$$= \left(\left(1 - \frac{X}{HW}\right)W^T\right)_{ab},$$

where the divisions between two matrices are calculated element-wise, and $\mathbf{1}$ is a matrix of all ones with the same size of X. Similarly, the partial derivative of J_{Poi} with respect to any element W_{ab} in W is

$$\frac{\partial J_{\text{Poi}}}{\partial W_{ab}} = \left(H^T\left(1 - \frac{X}{HW}\right)\right)_{ab}$$

Therefore, the gradients of J_{Poi} are

$$\nabla_H J_{\text{Poi}} = \left(1 - \frac{X}{HW}\right)W^T$$
$$\nabla_W J_{\text{Poi}} = H^T\left(1 - \frac{X}{HW}\right)$$

By setting the step sizes in the gradient decent method to be

$$\eta_H = \frac{H}{\mathbf{1}W^T}$$
$$\eta_W = \frac{W}{H^T\mathbf{1}},$$

we obtain the MUR update rules:

$$H_{ab} \leftarrow H_{ab}\frac{\left(\left(\frac{X}{HW}\right)W^T\right)_{ab}}{(\mathbf{1}W^T)_{ab}}$$
$$W_{ab} \leftarrow W_{ab}\frac{\left(H^T\left(\frac{X}{HW}\right)\right)_{ab}}{(H^T\mathbf{1})_{ab}}$$

The update rules for the extended objective function with manifold information, the way to decompose contact map matrix X into HSH^T and the way to handle sparse contact maps are discussed in the Sections S1.1.2, S1.1.3 and S1.1.4, respectively.

4.1.3 Removing biases in contact maps

When studying Hi-C contact maps, it is desirable to correct the data matrix such that signals coming from different genomic locations are equally visible (Imakaev *et al.*, 2012; Rao *et al.*, 2014). Suppose R is a matrix derived from X with all position-specific biases removed, we require each row and each column of R to sum to the same constant. We relate R and X by $X = BRB$, where B is a diagonal matrix with its diagonal elements indicating the biases at each bin. BNMF applies non-negative factorization on matrix R, and thus $X = BHSH^TB = (BH)S(BH)^T$.

To determine B and R from X, we first initialize B to be the identity matrix and iteratively normalize $W = SH^T$ and H:

$$b \leftarrow \text{Diag}\left(\frac{n\sum_{i=1}^{r}(SH^TB)_i}{\sum_{i=1}^{r}\sum_{j=1}^{n}(SH^TB)_{ij}}\right)$$
$$H \leftarrow b^{-1}BH$$
$$B \leftarrow b$$
$$h \leftarrow \text{Diag}\left(\frac{1}{n}\sum_{i=1}^{n}H_i\right)$$
$$H \leftarrow Hh^{-1}$$
$$S \leftarrow hSh,$$

where $\text{Diag}(v)$ is the diagonal matrix with the diagonal entries taken from vector v. This process will produce suitable B and S that make each column of SH^T to sum to a constant and each column of H to have an average of one, leading to HSH^T having equal column sums and equal row sums. Supplementary Figure S6a shows that our method produces bias vectors highly similar to the ones produced by the ICE method (Imakaev *et al.*, 2012) based on a Hi-C contact map obtained from the human GM06990 cell line (SRR027956). The main difference between the two methods is that our method optimizes a function for $X \approx BHSH^TB$ while the ICE method optimizes $X \approx BTB$, which has much more free parameters. Supplementary Figure S6b shows that after bias removal, the correlation between contact maps obtained from the same cell line increased more than the correlation between contact maps from different cell lines.

4.1.4 The final BNMF algorithm

Our BNMF algorithm combines the techniques discussed in the previous sections. For a non-negative symmetric matrix $X \in \mathbb{R}^{n\times n}$ and a given number of clusters r, our algorithm finds diagonal matrix

$B \in \mathbb{R}^{n\times n}$, cluster membership matrix $H \in \mathbb{R}^{n\times r}$ and cluster size matrix $S \in \mathbb{R}^{r\times r}$ to approximate X by $BHSH^TB$ by solving the following optimization problem:

Minimize

$$\sum_{i=1}^{n}\sum_{j=1}^{n} I_{ij}((BHSH^TB)_{ij} - X_{ij}\ln(BHSH^TB)_{ij}) + \lambda\sum_{k=1}^{r}(H^TLH)_{kk}$$

subject to

$$B_{ij} \geq 0, H_{ij} \geq 0, S_{ij} \geq 0\ \forall i,j, \sum_{i=1}^{n} H_{ij} = c_1\ \forall j, \sum_{i=1}^{r}(SH^T)_{ij} = c_2\ \forall j,$$

where λ is a parameter, and $L = D - E$ is a user-defined nearest-neighbor matrix with a default to take value 1 for entries corresponding to bins adjacent in the primary sequence and 0 for all other entries. Note that since

$$\sum_{i=1}^{n}(HSH^T)_{ij} = \sum_{i=1}^{n}\sum_{k=1}^{r} H_{ik}(SH^T)_{kj} = \sum_{i=1}^{n} H_{ik}\sum_{k=1}^{r}(SH^T)_{kj} = c_1c_2$$

for all j, the balanced contact map $R = HSH^T$ should have constant column sums.

Our algorithm initializes the matrices by the Non-negative Double Singular Value Decomposition (NNDSVD) method (Boutsidis and Gallopoulos, 2008). It then iteratively updates B, H and S using the following rules:

$$G \leftarrow BH$$

$$G \leftarrow G \otimes \frac{\left(I \otimes \frac{X}{GSG^T + \epsilon}\right)GS + 2\lambda EH}{IGS + 2\lambda DH + \epsilon}$$

$$B \leftarrow \mathrm{Diag}\left(\frac{n\sum_{i=1}^{r}(SG^T)_i}{\sum_{i=1}^{r}\sum_{j=1}^{n}(SG^T)_{ij}}\right)$$

$$H \leftarrow B^{-1}G$$

$$h \leftarrow \mathrm{Diag}(\frac{1}{n}\sum_{i=1}^{n} H_i)$$

$$H \leftarrow Hh^{-1}$$

$$S \leftarrow hSh$$

$$S \leftarrow S \otimes \frac{BH^T(I \otimes \frac{X}{BHSH^TB + \epsilon})HB}{BH^TIHB + \epsilon},$$

where $\otimes$ represents element-wise multiplication and a small constant $\epsilon = 1 \times 10^{-30}$ is added to avoid division-by-zero errors. The parameter λ was always set to 1 in the current study. The algorithm terminates when the decrease of objective score is less than 1×10^{-6} of the score in the previous iteration, or a maximum number of iterations is reached. In our analyses of the human and yeast datasets, we observed that the objective score usually became stable after 1500 iterations. We therefore set the maximum number of iterations to 3000, which guaranteed the convergence of the algorithm in most cases.

BNMF contains a procedure for automatically determining the number of clusters. The details are given in Section 9.1.5.

4.2 Statistical testing of co-localization

Using the BNMF clusters, we tested whether some genomic features define bins that are co-localized. We considered two scenarios. In the first, 'binary' scenario, genomic bins containing the feature or have an average feature value exceeding a certain threshold will be given a label of 1 (the positive bins), and all other regions will be given a label of 0 (the negative bins). In other words, the feature provides a binary classification of the genomic bins. We used this setting to test the co-localization of early replication sites, tRNA genes, VRSM genes, genes highly expressed in specific cell lines, and super enhancers (the data sources will be discussed in Section 4.3). In the second, 'continuous' scenario, each bin is labeled by the average feature value across all genomic locations in the bin. We applied this setting to the correlation between cluster affinity and either DNase I hypersensitivity or replication time. For each scenario, we considered two different ways to perform the statistical tests.

For the binary scenario, we checked whether the positive bins were spatially co-localized by using the cluster affinity values. The first testing method was based on the area under the receiver operator characteristics (ROC), which was also used previously (Cournac *et al.*, 2012; Duan *et al.*, 2010). Specifically, for each cluster, we sorted all bins in descending order of their cluster affinity values. We then used this order to make an ROC curve according to the bin labels assigned by the feature. The statistical significance of the AUC was evaluated by the Mann-Whitney U test. The resulting *P*-values from all the clusters were then collected and the most significant one was reported after Bonferroni correction for multiple hypothesis testing.

The second method was inspired by the conserved consecutive distances (CCD) method (Paulsen *et al.*, 2013). The basic idea is to compare a test statistic of the positive bins with a background distribution of the test statistic. In our case, we used the average affinity value to a cluster as the test statistic. The CCD method draws a large number of random bins and computes the average affinity to the cluster to form a corresponding background distribution of average affinity values. In each sample, the bins are required to preserve the same distance distribution between the positive bins in the primary sequence. By having this requirement, the positive bins are considered significantly co-localized at a cluster only if their co-localization is not simply due to their proximity in the primary sequence. The fraction of background bin samples with an average affinity higher than that of the positive bins is defined as the *P*-value. Again, the most significant *P*-value after Bonferroni correction from the different clusters was reported.

For the continuous scenario, we checked whether the cluster affinity values were quantitatively correlated with the feature values of the bins. In the first testing method, we used Pearson correlation coefficient (PCC), and in the second method, we used the Spearman rank correlation coefficient (SPC). In either case, we used the 'Numpy' package in Python to calculate the *P*-values, and then applied Bonferroni correction.

The key difference between our statistical tests and the ones previously used for analyzing DNA contact maps is that our tests involve cluster affinity values instead of contact counts, which enables us to test the relationship between a genomic feature and a specific local cluster.

As a basic check of our statistical testing procedures, we examined the *P*-values computed from 100 randomly shuffled contact maps. We used the yeast Hi-C contact map as a template and shuffled each row and each column simultaneously to produce random contact maps. In each random map, the bin labels were carried over from the real case, but the contact counts were changed by the shuffling. Supplementary Figure S8 shows that all four tests gave a uniform distribution of the *P*-value as expected.

4.3 Data collection and pre-processing

Hi-C and TCC data were collected from multiple sources (Supplementary Table S1). Human Hi-C data were mapped to the *hg19* reference and pre-processed by the *hiclib* pipeline (Imakaev et al., 2012). For the other Hi-C datasets, we downloaded the processed data provided by the authors from public databases.

The synthetic yeast contact map was generated by using the volume exclusion model to produce 3000 simulated genomes and the corresponding combined contact map, using the source codes provided by the author of a published study (Tjong *et al.*, 2012). We then added random position-specific biases to it to produce a noisy contact map.

DNase I hypersensitive sites (DHS) and gene expression levels based on RNA-seq in the four human cell lines GM12878, H1-hESC, IMR90 and K562 were produced by the ENCODE consortium (The Encode Project Consortium, 2012) and downloaded from the UCSC Genome Browser (Kent *et al.*, 2002). Cell-type specific highly expressed genes in the four cell lines were defined as the genes having an expression level of RPKM > 4 in a cell line but RPKM < 1 in the other three cell lines. Replication time of different regions in the human genome was defined in a previous study (Ryba *et al.*, 2010) and we downloaded the data from http://www.replicationdomain.org. Early replication sites in yeast were defined in a previous work (Duan *et al.*, 2010). The lists of super enhancers in the four human cell lines were defined in a published study (Hnisz *et al.*, 2013).

Acknowledgment

We would like to thank Harianto Tjong for providing the source code for generating synthetic yeast genome structures.

Funding

KYY is partially supported by the Hong Kong Research Grants Council General Research Fund CUHK418511. XH is partially supported by the National Natural Science Foundation of China (NSFC: 61402423).

Conflict of Interest: none declared.

References

Ay,F. *et al.* (2014) Three-dimensional modeling of the p. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. *Genome Res.*, **24**, 974–988.

Boutsidis,C. and Gallopoulos,E. (2008) Svd based initialization: a head start for nonnegative matrix factorization. *Pattern Recogn.*, **41**, 1350–1362.

Cai,D. *et al.* (2008). Non-negative matrix factorization on manifold. In: *Eighth IEEE International Conference on Data Mining, 2008. ICDM'08*, IEEE. pp. 63–72.

Cournac,A. *et al.* (2012) Normalization of a chromosomal contact map. *BMC Genomics*, **13**, 436.

Cremer,T. and Cremer,C. (2001) Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat. Rev. Genet.*, **2**, 292–301.

Cremer,T. and Cremer,M. (2010) Chromosome territories. *Cold Spring Harb. Perspect. Biol.*, **2**, a003889.

Devarajan,K. (2008) Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. *PLoS Comput. Biol.*, **4**, e1000029.

Ding,C. *et al.* (2005). On the equivalence of nonnegative matrix factorization and spectral clustering. In: *Proc. SIAM Data Mining Conf*, **4**, pp. 606–610.

Dixon,J.R. *et al.* (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.

Duan,Z. *et al.* (2010) A three-dimensional model of the yeast genome. *Nature*, **465**, 363–367.

Fullwood,M.J. *et al.* (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, **462**, 58–64.

Greene,D. *et al.* (2008) Ensemble non-negative matrix factorization methods for clustering protein-protein interactions. *Bioinformatics*, **24**, 1722–1728.

Hnisz,D. *et al.* (2013) Super-enhancers in the control of cell identity and disease. *Cell*, **155**, 934–947.

Hu,M. *et al.* (2012) Hicnorm: removing biases in hi-c data via poisson regression. *Bioinformatics*, **28**, 3131–3133.

Imakaev,M. *et al.* (2012) Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nat. Methods*, **9**, 999–1003.

Jin,F. *et al.* (2013) A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, **503**, 290–294.

Kalhor,R. *et al.* (2012) Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.*, **30**, 90–98.

Kent,W.J. *et al.* (2002) The human genome browser at ucsc. *Genome Res.*, **12**, 996–1006.

Knight,P.A. and Ruiz,D. (2013) A fast algorithm for matrix balancing. *IMA J. Numer. Anal.*, **33**, 1029–1047.

Lee,D.D. and Seung,H.S. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**, 788–791.

Lieberman-Aiden,E. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.

Ma,W. *et al.* (2015) Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. *Nat. Methods*, **12**, 71–78.

MacQueen,J.B. (1967) Some methods for classification and analysis of multivariate observations. In: *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297.

Nagano,T. *et al.* (2013) Single-cell hi-c reveals cell-to-cell variability in chromosome structure. *Nature*, **502**, 59–64.

Nora,E.P. *et al.* (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, **485**, 381–385.

Paulsen,J. *et al.* (2013) Handling realistic assumptions in hypothesis testing of 3d co-localization of genomic elements. *Nucleic Acids Res.*, **41**, 5164–5174.

Rao,S.S.P. *et al.* (2014) A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.

Ryba,T. *et al.* (2010) Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res.*, **20**, 761–770.

Sexton,T. *et al.* (2012) Three-dimensional folding and functional organization principles of the drosophila genome. *Cell*, **148**, 458–472.

The Encode Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Tjong,H. *et al.* (2012) Physical tethering and volume exclusion determine higher-order genome organization in budding yeast. *Genome Res.*, **22**, 1295–1305.

Varoquaux,N. *et al.* (2014) A statistical approach for inferring the 3d structure of the genome. *Bioinformatics*, **30**, i26–i33.

Yaffe,E. and Tanay,A. (2011) Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.*, **43**, 1059–1065.

Zhang,Y. *et al.* (2012) Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell*, **148**, 908–921.