Novel AI-powered computational method using tensor decomposition can discover the common optimal bin sizes when integrating multiple Hi-C datasets

Y-h Taguchi[1*] and Turki Turki[2]

[1*]Department of Physics, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo, 112-8551, Japan.
[2]Department of Computer Science, King Abdulaziz University, Street, Jeddah, 21589, Saudi Arabia.

*Corresponding author(s). E-mail(s): tag@granular.com;
Contributing authors: tturki@kau.edu.sa;

Abstract

Discovering the optimal bin sizes (or resolutions) for the integration of multiple Hi-C datasets is difficult, since bin sizes must be common over multiple datasets whereas dependence of quality upon bin sizes can vary from dataset to dataset. Moreover, we are not supposed to seek common structures in the smaller bin sizes than optimal bin sizes if we cannot suppress the appearance of phase transition-like phenomena even after increasing the number of mapped short reads per bins, since it might mean that there are no common structures any more in finer resolutions and individual Hi-C datasets might have to be analyzed separately. Thus, quality assessments of individual datasets have the limited ability to determine the best bin size common for all datasets. In this study, we propose, the first-ever to our knowledge, a novel adoption of tensor decomposition (TD) based unsupervised feature extraction (FE) to choose the optimal bin sizes for the integration of multiple Hi-C datasets. The use of TD-based unsupervised FE can exhibit something like phase-transition phenomena, by which we can automatically estimate the possible smallest bin size (or the highest resolution) empirically without manually setting some threshold value by hand for the integration of multiple Hi-C datasets.

Keywords: Hi-C, genome, AI, tensor decomposition, feature extraction, advances in unsupervised learning

1 Introduction

The decision of bin sizes for Hi-C datasets is a critical step, since the bin sizes heavily affect the outcome. For too large bin sizes (i.e., low resolution), we might overlook the biologically important factor, which can be seen only in the finer resolution. On the other hand, too small bin sizes (i.e., high resolution) also might result in the different problem, if not large enough short reads are mapped to the individual bins because of increase of the number of bins; small number of short reads mapped to a bin causes the fluctuation that might prevent us from observing the biologically important structure. Thus there is a trade-off between resolution and fluctuation. While suppressing fluctuation small enough, it is difficult to achieve the high enough resolution to capture the biologically important fine structures. It is important to estimate how small the bin sizes can be without having too large fluctuation.

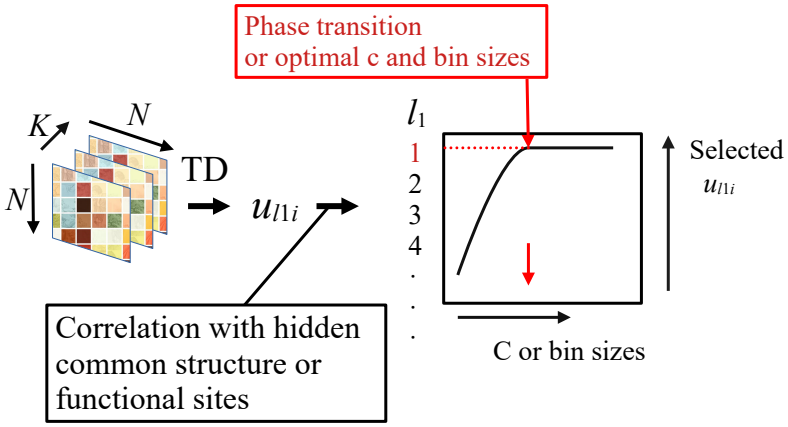

Fig. 1 Starting from K Hi-C datasets or synthetic data represented as $N \times N$ matrix, a tensor $x_{ii'k} \in \mathbb{R}^{N\times N\times K}$ is generated. TD was applied to $x_{ii'k}$ and we get $u_{\ell_1 i}$ attributed to ith bins or features. As long as c, which is a contribution of hidden common structure, or bin sizes is large enough, $\ell_1 = 1$ is most correlated with hidden common structure or functional sites. Optimal c or bin size is decided at the phase transition after which $\ell_1 = 1$ is not most correlated with hidden common structure or functional sites anymore.

This difficulty of the estimation of optimal bin sizes increases even more when integrating multiple Hi-C datasets, since we need to find the optimal bin sizes commonly for all datasets apart from the investigation of the optimal bin sizes toward the individual Hi-C datasets. In actual, it is hardly said that the strategy by which we can decide the smallest bin sizes under the reasonable amount of fluctuation for multiple Hi-C datasets simultaneously when we integrate them is very successful. For example, although Yardımcı et al [1] performed quality assessment of various tools, HiCRep [2], GenomeDISCO [3], HiC-Spector [4], and QuASAR-Rep [5] to process multiple Hi-C datasets, their conclusions were "we do not recommend using reproducibility scores to attempt to select an appropriate resolution." On the other hand, although there was once a major tool, HSA [6], to process multiple Hi-C datasets, it is not available any

more. Although multiHiCcompare [7] apparently deals with multiple Hi-C datasets, since it can perform only pairwise comparisons, it is useless to simultenously decide optimal bin sizes for multiple datasets. Although Dedoc2 [8] can decide optimal bin size using entropy, it can be applicable to single Hi-C dataset (i.e., to individual single cells), not in the integrated mannar. Although these are only a few examples that attempt to find optimal bin sizes for multiple Hi-C datasets, to our knowledge, there are no successful tools to decide optimal bin sizes over multiple Hi-C datasets when integrating them.

In this study, to resolve this problem, i.e., estimating the common optimal bin size for the multiple Hi-C datasets simultaneously when integrating them, we propose a novel use of TD based unsupervised FE [9], which allows us to observe something like phase transition by which we can decide the optimal bin sizes without specifying any threshold values (Fig. 1). As far as we know, this is the first successful trial that decides the optimal bin sizes for multiple Hi-C datasets simultaneously when integrating them.

2 Results

2.1 Synthetic data

At first, we investigated which singular value vector, $u_{\ell_1 i}$, attributed to the ith feature, which corresponds to the ith bin when we consider real Hi-C data later, is the most associated with the hidden common (clustering) structure, y_i, (see Methods) embedded into the synthetic dataset where there are 1,000 features (i), top 10 among which are clustered with each other (for more details, see Methods). Figure 2 shows the dependence of selected ℓ_1 upon c, which represents the contribution of hidden common structure among all datasets (see Methods). Although $\ell_1 = 1$, which has the largest

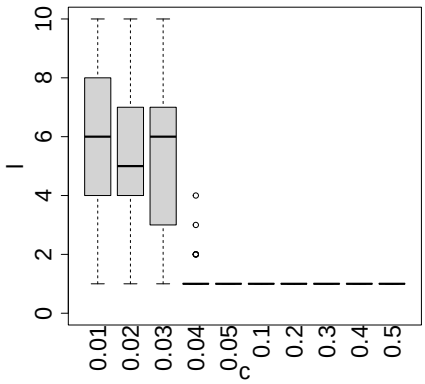

Fig. 2 The boxplot of ℓ_1 selected upon c

contribution, is always selected as those most correlated with common structure for

larger c, the frequency of the selection of ℓ_1 other than 1 increases as c decreases. In addition to this, although ℓ_3, which is most associated with the selected ℓ_1, is always 1, which corresponds to simple averaging over samples (Fig. 3), for large enough c, other ℓ_3s than 1 start to be selected as c decreases (Fig. 4). In actual, although the

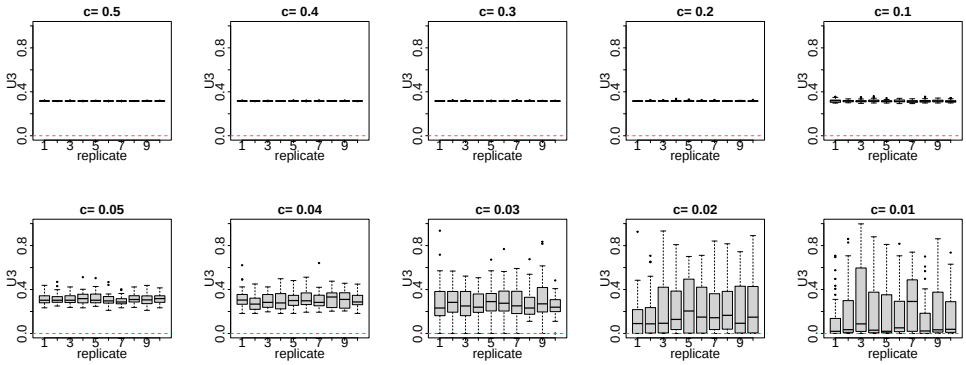

Fig. 3 Boxplot of $u_{\ell_3 k}$ over replicates ($1 \leq k \leq 10$) for various c values. Red horizontal broken lines represent $u_{\ell_3 k} = 0$

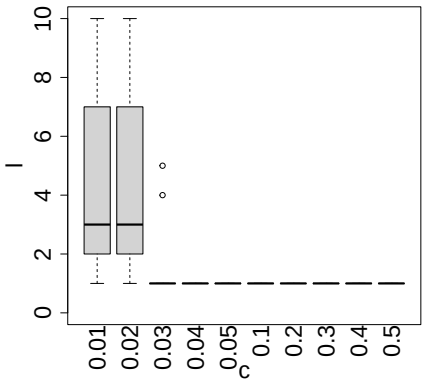

Fig. 4 The boxplot of ℓ_3 selected upon c

correlation between the selected $u_{\ell_1 i}$ and hidden common structure, y_i, keeps 1 for larger c, it starts to drop when c decreases less than some threshold value (Fig. 5) and the correlations are not significant anymore for smaller cs (Fig. 6).

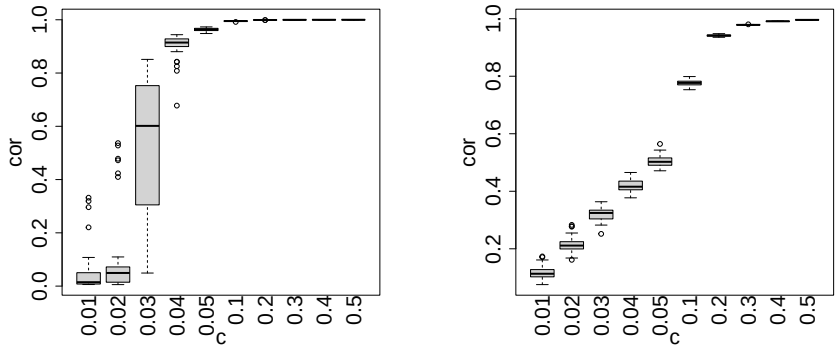

Fig. 5 The boxplot of the Pearson's correlation coefficients between $u_{\ell_1 i}$ selected (left) or simple average $\langle x_{ii'k} \rangle_i$ (right) and hidden common structure, y_i, as a function of c

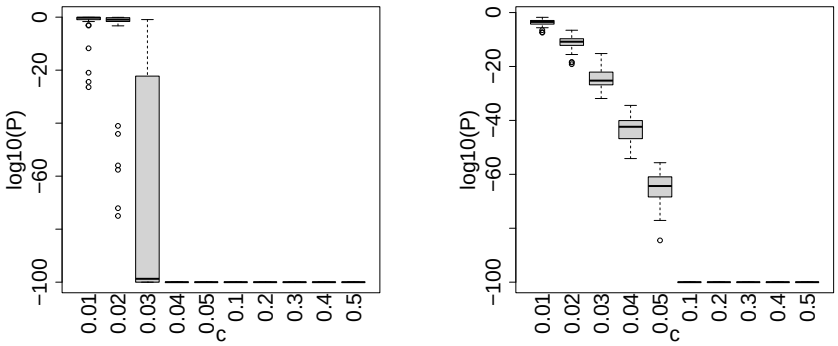

Fig. 6 The boxplot of P-values in the logarithmic scale associated with the Pearson's correlation coefficient between $u_{\ell_1 i}$ selected (left) or simple average $\langle x_{ii'k} \rangle_i$ (right) and hidden common structure, y_i, as a function of c. $P < 10^{-100}$ are truncated to be $P = 10^{-100}$.

Finally, we have checked if TD based unsupervised FE correctly selected 10 features ($i \leq 10$) which are interconnected with each other (Table 1). Although false positives (i.e., $i > 10$ but wrongly associated with adjusted P-values less than 0.01) heavily fluctuated, true positives (i.e, $i \leq 10$ and correctly associated with adjusted P-values less than 0.01) monotonically decrease and false negatives (i.e, $i \leq 10$ but wrongly associated with adjusted P-values larger than 0.01) monotonically increase as c decreases after $c \leq 0.05$. It is coincident with the above observation that TD based unsupervised FE cannot capture the hidden common structure, y_i, correctly if c is too small.

Table 1 Confusion matrix between sites selected (associated with adjusted P-values less than 0.01) by TD based unsupervised FE and sites inter connected each other ($y_i = 1, i \leq 10$). The correspondences to TP (True positive), FP (False positive), TN (True negative) and FN (False negatives) are also shown at the end of table.

		$i > 10$	$i \leq 10$	$i > 10$	$i \leq 10$	$i > 10$	$i \leq 10$	$i > 10$	$i \leq 10$
	c	0.5		0.4		0.3		0.2	
Adjusted	≥ 0.01	990.0	0.0	990.0	0.0	846.0	0.0	989.9	0.0
P-values	< 0.01	0.0	10.0	0.0	10.0	144.0	10.0	0.1	10
	c	0.1		0.05		0.04		0.03	
Adjusted	≥ 0.01	960.0	0.0	990.0	3.8	989.5	5.0	988.9	7.9
P-values	< 0.01	30.0	10.0	0.1	6.2	0.5	5.0	1.1	2.1
	c	0.02		0.01				**Legend**	
Adjusted	≥ 0.01	988.2	9.8	988.2	9.8			**TN**	**FN**
P-values	< 0.01	1.8	0.2	1.8	0.2			**FP**	**TP**

One might wonder why we need TD since simple average over replicates might be able to achieve better performance. To deny this possibility, we compute correlation between simple average $\langle x_{ii'k} \rangle$ and hidden common structure, y_i as well (Figs. 5 and 6). It is obvious that TD based unsupervised FE can outperform simple average. Thus it is worthwhile trying TD based unsupervised FE instead of simple average.

Based upon the above observations, threshold value below which hidden common structure is not a primary structure anymore seems to be about $c = 0.05$. Although this specific c value might not be important, it is important that there is a threshold value above which u_{1i} is always selected and is also highly coincident with hidden common structure. It is something like analogous to phase transition. The important point is that we might be able to make use of this result for quality assessment; i.e., we can trust the results obtained by the integration of multiple profiles only when it is above the threshold value and *we do not have to specify threshold value in advance since threshold value is automatically decided when u_{1i} is not most correlated with the hidden common structure, y_i.*

Nevertheless, in the real applications, we do not know what the hidden common structure is, thus there might not be any ways to estimate threshold values, since we need hidden common structure with which we have to compute correlation coeffcients. In the below, we demonstrate how we can estimate threshold value above which we can trust the obtained results by the integration of multiple profiles without knowing what the hidden common structure is.

2.2 Hi-C dataset

Now we come to the stage where we apply TD based unsupervised FE to real Hi-C datasets and see if the phase transition mentioned above can be observed and can be used for quality assessment, i.e., the selection of the optimal bin in this section, or not.

We have applied TD based unsupervised FE to two Hi-C datasets. Since we do not know what the hidden common structures are at all, we employ known functional sites instead of hidden common structure (Tables 2, 3, 4, and 5). At first, since we

Table 2 Absolute correlations between $u_{\ell_1 i}$ or simple average $\langle x_{ii'k} \rangle_i$ and CTCF binding sites. Bold numbers are the absolute largest correlation coefficients.

CTCF **PC**	**1**	**2**	**5**	$\langle x_{ii'k} \rangle_i$
		GSE260760		
1000000				
Pearson	$\mathbf{4.98 \times 10^{-1}}$	1.23×10^{-1}		3.36×10^{-1}
P-value	1.43×10^{-175}	6.49×10^{-11}		5.64×10^{-75}
Spearman	$\mathbf{5.72 \times 10^{-1}}$	4.99×10^{-1}		5.18×10^{-1}
P-value	8.34×10^{-244}	7.69×10^{-177}		8.30×10^{-193}
150000				
Pearson	5.65×10^{-3}	2.75×10^{-2}	$\mathbf{4.42 \times 10^{-1}}$	3.14×10^{-1}
P-value	4.41×10^{-1}	1.76×10^{-4}	0	0
Spearman	3.66×10^{-1}	5.52×10^{-1}	$\mathbf{5.73 \times 10^{-1}}$	5.14×10^{-1}
P-value	0	0	0	0
		GSE255264		
1000000				
Pearson	$\mathbf{5.09 \times 10^{-1}}$	3.62×10^{-2}		4.45×10^{-1}
P-value	1.24×10^{-184}	5.52×10^{-2}		1.62×10^{-136}
Spearman	$\mathbf{6.87 \times 10^{-1}}$	3.15×10^{-1}		6.35×10^{-1}
P-value	0	1.10×10^{-65}		0
150000				
Pearson	$\mathbf{4.46 \times 10^{-1}}$	4.01×10^{-2}		3.88×10^{-1}
P-value	0	4.38×10^{-8}		0
Spearman	$\mathbf{6.25 \times 10^{-1}}$	2.54×10^{-1}		5.79×10^{-1}
P-value	0	3.40×10^{-271}		0
40000				
Pearson	2.88×10^{-3}	$\mathbf{3.89 \times 10^{-1}}$		3.32×10^{-1}
P-value	4.47×10^{-1}	0		0
Spearman	4.42×10^{-1}	$\mathbf{5.55 \times 10^{-1}}$		5.11×10^{-1}
P-value	0	0		0

observe that $u_{\ell_1 i}$ associated with the absolute laregest correlation coefficient almost always has more correlation with functional site than simple average $\langle x_{ii'k} \rangle_i$ excluding dELS for GSE255264, it is reasonable to consider TD based unsupervised FE as an advanced criteria for quality assessment of Hi-C datasets to check if Hi-C data is good enough to capture hidden common structure. Then we found that we could observe phase transition-like behaviour as expected for TD based unsupervised FE applied to real Hi-C dataset. For GSE260760, u_{1i} has the largest correlation with functional sites regardless to the types of functional sites for bin size of 1000000, but does not anymore for bin size 150000. For GSE255264, u_{1i} has the largest correlation with functional sites regardless to the types of functional sites for bin sizes of 1000000 and 150000, but does not anymore for for bin size 40000. Thus, phase transition-like phenomena is supposed to take place between 1000000 and 150000 for GSE260760 whereas it is so between 150000 and 40000 for GSE255264. Thus, bin sizes less than or equal to 150000 for GSE260760 and those less than or equal to 40000 for GSE255264 cannot be regarded good enough to capture hidden common structures anymore.

Table 3 Absolute correlations between $u_{\ell_1 i}$ or simple average $\langle x_{ii'k} \rangle_i$ and promoter-like signatures (PLS). Bold numbers are the absolute largest correlation coefficients.

PLS **PC**	**1**	**2**	**5**	$\langle x_{ii'k} \rangle_i$
		GSE260760		
1000000				
Pearson	$\mathbf{4.51 \times 10^{-1}}$	7.08×10^{-2}		3.25×10^{-1}
P-value	1.06×10^{-140}	1.75×10^{-4}		4.59×10^{-70}
Spearman	$\mathbf{6.33 \times 10^{-1}}$	4.86×10^{-1}		6.00×10^{-1}
P-value	0	3.88×10^{-166}		6.31×10^{-274}
150000				
Pearson	4.41×10^{-3}	7.91×10^{-3}	$\mathbf{3.32 \times 10^{-1}}$	2.72×10^{-1}
P-value	5.47×10^{-1}	2.81×10^{-1}	0	0
Spearman	2.74×10^{-1}	4.74×10^{-1}	$\mathbf{5.19 \times 10^{-1}}$	4.77×10^{-1}
P-value	0	0	0	0
		GSE255264		
1000000				
Pearson	$\mathbf{6.39 \times 10^{-1}}$	1.36×10^{-3}		5.19×10^{-1}
P-value	0	9.43×10^{-1}		2.43×10^{-193}
Spearman	$\mathbf{7.64 \times 10^{-1}}$	4.50×10^{-1}		7.45×10^{-1}
P-value	0	5.65×10^{-140}		0
150000				
Pearson	$\mathbf{5.15 \times 10^{-1}}$	1.50×10^{-3}		4.17×10^{-1}
P-value	0	8.38×10^{-1}		0
Spearman	$\mathbf{5.87 \times 10^{-1}}$	2.90×10^{-1}		5.70×10^{-1}
P-value	0	0		0
40000				
Pearson	6.69×10^{-4}	$\mathbf{3.60 \times 10^{-1}}$		2.87×10^{-1}
P-value	8.60×10^{-1}	0		0
Spearman	2.75×10^{-1}	$\mathbf{3.93 \times 10^{-1}}$		3.68×10^{-1}
P-value	0	0		0

2.3 Cluster structure detection of Hi-C datasets

Since TD based unsupervised FE can select bins associated with functional site, it is expected to detect common clustering structure within multiple Hi-C datasets. At first, we plot 2424 and 447 $x_{ii'k}$s selected by u_{1i} and u_{2i} with bin size of 40000 for GSE255264, respectively (Fig. 7). Since in this bin size for GSE255264, u_{1i} is not associated with the highest absolute correlation with functional sites, u_{1i} is not expected to associated with (unknown) hidden common structure whereas u_{2i} is expected to associated with (unknown) hidden common structure since u_{2i} is associated with the highest absolute correlation with functional sites (Tables 2, 3, 4, and 5). As expected, 2424 $x_{ii'k}$s selected by u_{1i} have more sample dependence than 447 $x_{ii'k}$s selected by u_{2i}, since Mean vs SD ratio of 447 $x_{ii'k}$s selected by u_{2i} is less than that of 2424 $x_{ii'k}$s selected by u_{1i} (Table 6). This means, for the bin size of 40000, (unknown) hidden common structure is not a primary part within datasets anymore and it is only the secondary contribution. In actual, although 2424 $x_{ii'k}$s selected by u_{1i} are widely distributed along whole genome, the distribution of 447 $x_{ii'k}$s selected by u_{1i} is restricted

Table 4 Absolute correlations between $u_{\ell_1 i}$ or simple average $\langle x_{ii'k} \rangle_i$ and proximal with enhancer-like signatures (pELS). Bold numbers are the absolute largest correlation coefficients.

pELS **PC**	**1**	**2**	**5**	$\langle x_{ii'k} \rangle_i$
		GSE260760		
1000000				
Pearson	$\mathbf{5.43 \times 10^{-1}}$	8.50×10^{-2}		3.76×10^{-1}
P-value	2.89×10^{-215}	6.47×10^{-6}		4.77×10^{-95}
Spearman	$\mathbf{7.08 \times 10^{-1}}$	5.46×10^{-1}		6.75×10^{-1}
P-value	0	2.00×10^{-217}		0
150000				
Pearson	4.63×10^{-4}	1.84×10^{-2}	$\mathbf{4.08 \times 10^{-1}}$	3.23×10^{-1}
P-value	9.50×10^{-1}	1.20×10^{-2}	0	0
Spearman	3.32×10^{-1}	5.61×10^{-1}	$\mathbf{6.19 \times 10^{-1}}$	5.78×10^{-1}
P-value	0	0	0	0
		GSE255264		
1000000				
Pearson	$\mathbf{7.01 \times 10^{-1}}$	3.06×10^{-3}		5.82×10^{-1}
P-value	0	8.71×10^{-1}		1.07×10^{-253}
Spearman	$\mathbf{8.22 \times 10^{-1}}$	4.95×10^{-1}		8.03×10^{-1}
P-value	0	3.74×10^{-173}		0
150000				
Pearson	$\mathbf{5.78 \times 10^{-1}}$	1.65×10^{-3}		4.79×10^{-1}
P-value	0	8.22×10^{-1}		0
Spearman	$\mathbf{6.70 \times 10^{-1}}$	3.56×10^{-1}		6.54×10^{-1}
P-value	0	0		0
40000				
Pearson	9.76×10^{-4}	$\mathbf{4.17 \times 10^{-1}}$		3.40×10^{-1}
P-value	7.97×10^{-1}	0		0
Spearman	3.35×10^{-1}	$\mathbf{4.74 \times 10^{-1}}$		4.48×10^{-1}
P-value	0	0		0

to the limited region along whole genome (Fig. 7). This is nothing but the evidence that in the bin size of 40000, the present dataset for GSE255264 does not have ability to capture common structure within the dataset.

On the other hand, For GSE260760 where bin size of 150000 is far below the phase transition-like point since even u_{2i} is not associated with the highest absolute correlation with functional sites anymore but u_{5i} is (Tables 2, 3, 4, and 5). Fig. 8 shows the 225, 171 and 187 $x_{ii'k}$s selected by u_{1i}, u_{2i}, and u_{5i}. The 187 $x_{ii'k}$s selected by u_{5i} associated with the highest absolute correlation with functional sites cannot be regarded as more independent of samples than others, since Mean vs SD ratios are almost equivalent between u_{1i}, u_{2i}, and u_{5i} (Table 6). The fact that the number of selected bins are small (see Fig. 8, 225, 171 and 187 $x_{ii'k}$s that are less than 1,000) also suggests that the integration is substantially degradated.

Based upon the above observations, optimal bin sizes when integrating GSE260760 and GSE255264 are 1000000 and 150000, respectively, within the tested bin sizes. Since less bin sizes are usually employed to investigate genome structures, we need more

Table 5 Absolute correlations between $u_{\ell_1 i}$ or simple average $\langle x_{ii'k} \rangle_i$ and distal with enhancer-like signatures (dELS). Bold numbers are the absolute largest correlation coefficients.

dELS **PC**	**1**	**2**	**5**	$\langle x_{ii'k} \rangle_i$	
		GSE260760			
1000000					
Pearson	$\mathbf{8.51 \times 10^{-1}}$	6.84×10^{-2}		6.59×10^{-1}	
P-value	0	2.89×10^{-4}		0	
Spearman	$\mathbf{8.62 \times 10^{-1}}$	6.45×10^{-1}		8.44×10^{-1}	
P-value	0	0		0	0
150000					
Pearson	6.69×10^{-3}	4.10×10^{-2}	$\mathbf{6.06 \times 10^{-1}}$	5.69×10^{-1}	
P-value	3.61×10^{-1}	2.10×10^{-8}	0	0	
Spearman	4.54×10^{-1}	7.32×10^{-1}	$\mathbf{8.15 \times 10^{-1}}$	8.02×10^{-1}	
P-value	0	0	0	0	
		GSE255264			
1000000					
Pearson	7.12×10^{-1}	1.48×10^{-2}		$\mathbf{7.93 \times 10^{-1}}$	
P-value	0	4.33×10^{-1}		0	
Spearman	8.22×10^{-1}	5.88×10^{-1}		$\mathbf{8.26 \times 10^{-1}}$	
P-value	0	2.63×10^{-260}		0	
150000					
Pearson	6.05×10^{-1}	9.10×10^{-3}		$\mathbf{6.63 \times 10^{-1}}$	
P-value	0	2.14×10^{-1}		0	
Spearman	7.38×10^{-1}	5.02×10^{-1}		$\mathbf{7.45 \times 10^{-1}}$	
P-value	0	0		0	
40000					
Pearson	3.61×10^{-3}	5.11×10^{-1}		$\mathbf{5.55 \times 10^{-1}}$	
P-value	3.40×10^{-1}	0		0	
Spearman	4.80×10^{-1}	6.50×10^{-1}		$\mathbf{6.51 \times 10^{-1}}$	
P-value	0	0		0	

Table 6 Mean vs SD ratios. Smaller values suggest more independent of samples (ks). Bold ℓ_1s are most correlated with functional sites

$u_{\ell_1 i}$ used for selection	GSE255264		GSE260760		
ℓ_1	1	**2**	1	2	**5**
Mean vs SD ratio	0.629	0.302	0.472	0.436	0.475

number of reads and have to have smaller bin sizes as threshold (i.e., to make phase transition-like phenomena occur at smaller bin sizes). The advantages of this strategy are that we can check if the number of reads is enough or not before the detailed investigation of the obtained contact map. If we cannot suppress the appearance of phase transition-like phenomena even after increasing the number of mapped short reads, it might mean that there are no common structures at all any more in finer

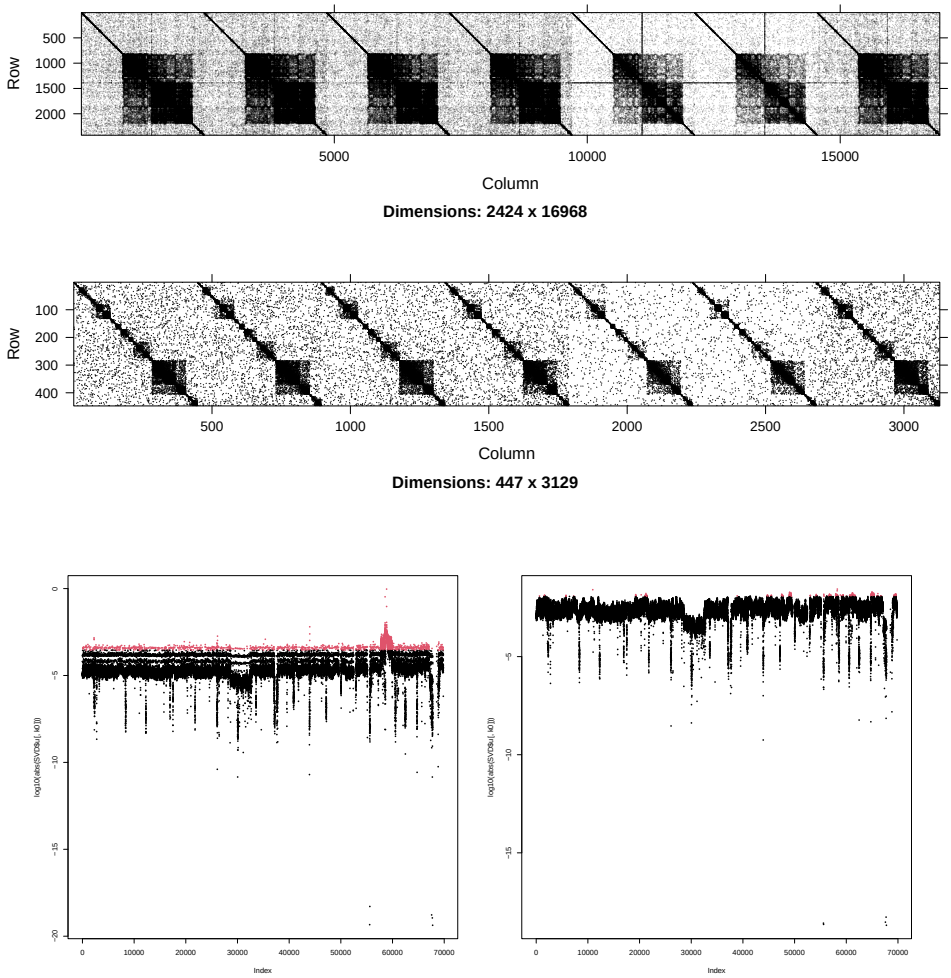

Fig. 7 Top: 2424 $x_{ii'k}$s selected by u_{1i}, middle: 447 $x_{ii'k}$s selected by u_{2i} for GSE255264. From left to right, seven datasets are aligned horizontally as ascending order of GSM IDs. Bottom left: $\log_{10} u_{1i}$, bottom right $\log_{10} u_{2i}$, red ones are selected ones. Bin sizes are 40000.

resolution; in this case, we have to give up integrating multiple Hi-C datasets in finer resolutions. Our strategy also allows us to perform this kind of decision making, too.

In conclusion, with replacing the correlation toward hidden common structure with that toward functional sites, we got the followings:

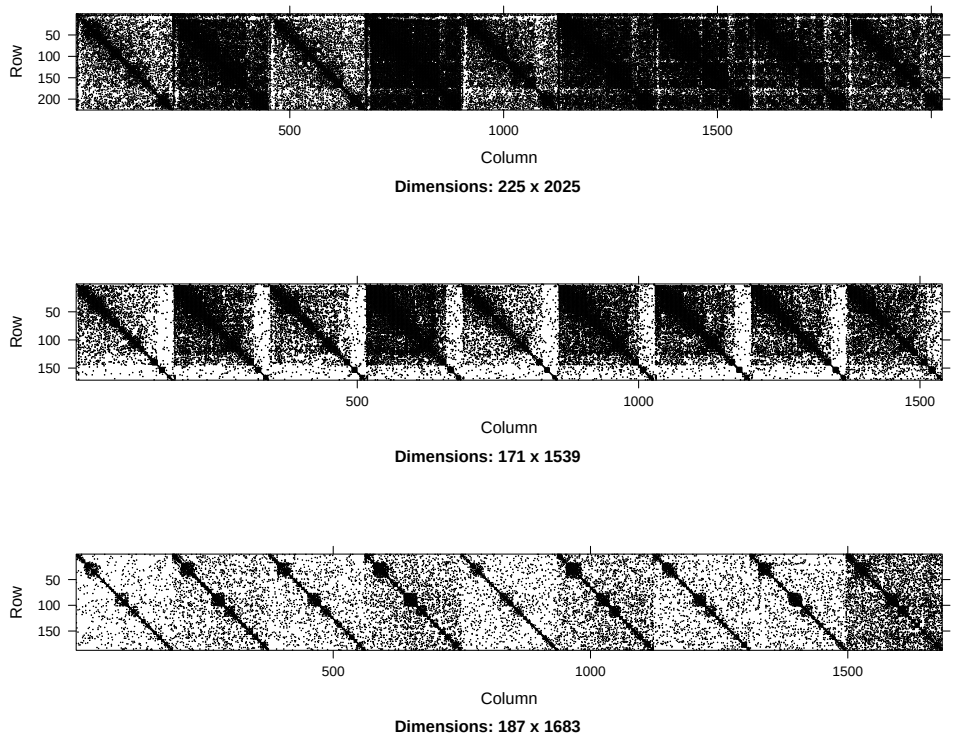

Fig. 8 Top: 255 $x_{ii'k}$s selected by u_{1i}, middle: 171 $x_{ii'k}$s selected by u_{2i}, bottom: 187 $x_{ii'k}$s selected by u_{5i} for GSE260760. Bin sizes are 150000. From left to right, nine datasets are aligned horizontally. From left to right, nine datasets are aligned horizontally as ascending order of GSM IDs.

1. Realization of "the phase transition-like" phenomena below which we cannot expect that common structures are primary factors within dataset (since u_{1i} is not associated with the highest absolute correlation with functional sites).
2. More significant correlation with functional sites using TD based unsupervised FE than that with simple average $\langle x_{ii'k} \rangle_i$.

Thus, we can conclude that our methods work very well.

3 Discussion

One might wonder why changing bin sizes in Hi-C data is analogous to change of c that represents the contribution of hidden common structure represented by y_i in synthetic

data. Since the number of bins increases as the size of bin decreases, the number of reads per individual bins decreases as well under the condition of the fixed total number of short reads. This results in the short of the number of reads in individual bins and the inaccurate estimation of the number of reads in individual bins. Thus, as the size of bins decreases, the fluctuation of the number of reads in bins increases, which results in the divergence among individual datasets. Thus, the decreases of bin size should be analogous to that of the parameter "c" in synthetic data. This suggests that it is not surprising even if the phase transition-like phenomena is observed as bin sizes decreases as it is observed as "c" decreases in the synthetic data.

In addition to this, TD based unsupervised FE has ability of quality assessment of Hi-C measurement. Since the number of reads downloaded for individual datasets is fixed as 100000000, the number of expected reads in individual bins is also constant. Nevertheless, if the number of interaction differs from datasets to datasets, it becomes difficult to capture the hidden common structure within multiple Hi-C datasets. It is the reason why the decrease of bin sizes corresponds to decrease of c in synthetic data. Thus, the phase transition-like phenomena can take place as in the case of synthetic dataset.

4 Conclusion

In this study, we have applied TD based unsupervised FE to multiple Hi-C datasets. We have found phase transition-like phenomena that is a boundary if u_{1i} is most correlated with functional site or not, as observed in synthetic data. We can specify the optimal bin size as the smallest bin size as long as u_{1i} is most correlated with functional site. As a bi-product, u_{1i} is usually more correlated with functional site than simple average over datasets. Thus, we can make use of it as a representative profile. In conclusion, we have developed the strategy by which we can select optimal bin size when we integrate multiple Hi-C datasets. As a bi-product, we can have representative profiles, $u_{\ell_1 i}$s, which are more correlated with functional sites than simple average, $\langle x_{ii'k} \rangle_i$.

5 Methods

5.1 Synthetic data

The synthetic data used in this study is a tensor $x_{ii'k} \in \mathbb{R}^{1,000 \times 1,000 \times 10}$. Initially, all $x_{ii'k}$s are filled with 0. Then 100 $x_{ii'k}$s for $1 \leq i, i' \leq 10$ are filled with 1. In addition to this, 1 is added to $x_{ii'k}$ for 200 randomly selected (i, i') pairs in individual ks (this means that random selected pairs are not common between distinct ks). The thirty three $x_{ii'k}$s (replicates) were generated with distinct random seeds. The reason why we have only 33 replicates is because 33 among tried 100 replicates are associated with successful HOSVD (others are terminated because of singularity).

5.1.1 Hidden common structure

We define y_i as

$$y_i = \begin{cases} 1, & i \leq 10 \\ 0, & i > 10 \end{cases} \tag{1}$$

y_i is called as hidden common structure in the analysis of the synthetic data in this study since it is distinct between the is inter-connected with each other regardless to ks and the other is.

5.1.2 Correlation with hidden common structure

For the synthetic data, the correlations between $u_{\ell_1 i}$ associated with the largest absolute correlation or simple average

$$\langle x_{ii'k} \rangle_i = \frac{1}{33000} \sum_{i'=1}^{1000} \sum_{k=1}^{33} x_{ii'k} \tag{2}$$

and y_i are computed.

5.2 Hi-C datasets

There are two Hi-C datasets analyzed in this study.

5.2.1 GSE260760

We have downloaded nine human knee chondrocytes Hi-C profiles [10].

5.2.2 GSE255264

We have downloaded seven AK1 cell line Hi-C profiles [11], which were originally taken from [12]. Although there are nine AK1 datasets, only seven (GSM8067773, GSM8067774, GSM8067775, GSM8067777, GSM8067778, GSM8067779, and GSM8067780) were actually downloadble.

5.2.3 How to download fastq file

To have same number of reads for all datasets, "-X 100000000" option, which limits the number of downloaded reads to as many as 10^8, was added to fastq_dump [13].

5.2.4 How to map reads to human genome

HiC-Pro [14] was used to map reads towards hg38 human genome assuming bin sizes, 1000000, 500000, 150000, 40000, and 20000.

5.2.5 Preprocessing

"raw" matrix files are used for the analyses. $x_{ii'k}$ is normalized as

$$x_{ii'k} \leftarrow \frac{x_{ii'k}}{\frac{1}{N^2} \sum_{i=1}^{N} \sum_{i'=1}^{N} |x_{ii'k}|} \tag{3}$$

5.3 TD based unsupervised FE applied to the synthetic data and Hi-C data

Higher order singular value decomposition (HOSVD) [9] is applied to $x_{ii'k} \in \mathbb{R}^{N\times N\times K}$ and we get

$$x_{ii'k} = \sum_{\ell_1=1}^{N}\sum_{\ell_2=1}^{N}\sum_{\ell_3=1}^{K} G(\ell_1\ell_2\ell_3)u_{\ell_1 i}u_{\ell_2 i'}u_{\ell_3 k} \tag{4}$$

where $G \in \mathbb{R}^{N\times N\times K}$ is a core tensor that represent the contribution of $u_{\ell_1 i}u_{\ell_2 i'}u_{\ell_3 k}$ to $x_{ii'k}$, $u_{\ell_1 i} = u_{\ell_2 i'} \in \mathbb{R}^{N\times N}$ and $u_{\ell_3 k} \in \mathbb{R}^{K\times K}$ are singular value matrices and orthogonal matrices. HOSVD is performed by applying `irlba` [15] to unfolded tensor $x_{i(i'k)} \in \mathbb{R}^{N\times(NK)}$ in sparse matrix format using Matrix package [16], since $x_{ii'k}$ is too large to be stored as dense matrix format. Since HOSVD is a set of singular value decomposition (SVD) applied to unfolded tensor [9], HOSVD is easily replaced with SVD.

$u_{\ell_1 i}$ of interest is selected either based upon the correlation with y_i (for the synthetic data) or that with known functional site (for Hi-C dataset). Once the $u_{\ell_1 i}$ of interest is selected, P-values are attributed to is as

$$\langle u_{\ell_1 i}\rangle = \frac{1}{N}\sum_{i=1}^{N} u_{\ell_1 i} \tag{5}$$

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{\ell_1 i} - \langle u_{\ell_1 i}\rangle}{\sigma_{\ell_1}}\right)^2\right] \tag{6}$$

where $P_{\chi^2}[> x]$ is the cumulative χ^2 probability distribution that the argument is larger than x. σ_{ℓ_1} is the standard deviation of $u_{\ell_1 i}$ computed by the following optimization.

The histogram of P_i is supposed to be flat if $u_{\ell_1 i}$ obeys Gaussian. Optimization of σ_{ℓ_1} is performed such that the histogram of P_i is as flat as possible. At first, frequency of P_i in the sth bin, $h_s, 1 \le s \le S$, must be computed with excluding is whose associated adjusted P-values (see below) less than the threshold value since is whose associated adjusted P-values less than the threshold value are expected not to obey Gaussian. Without exclusion of is whose associated adjusted P-values less than the threshold value, σ_{ℓ_1} is over estimated and P_i is over estimated as well. Then, the number of is whose associated adjusted P-values less than the threshold value is wrongly reduced and that of the selected is as well.

Then the standard deviation of h_s, σ_h, is computed as

$$\langle h_s\rangle = \frac{1}{S}\sum_{s=1}^{S} h_s \tag{7}$$

$$\sigma_h = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(h_s - \langle h_s\rangle\right)^2} \tag{8}$$

Then σ_{ℓ_1} that minimizes σ_h is computed and the P_i is computed by the σ_{ℓ_1}. Please note that it is a sort of self-consistency computation since adjusted P-values must be computed in order to exclud is associated with adjusted P-values less than the threshold value in order to exclude these is from the computation of h_s. This process is iterated until the minimization of σ_h is finalized. P_is are finally adjusted by BH criterion [9] and is associated with adjusted P-values less than the threshold value are selected. In the case that the readers are interested in these processes included in TD based unsupervised FE, please read my text book published recently [9] or vignettes of two Bioconductor Packages [17, 18].

5.3.1 Mean vs SD ratio

To quantize the difference of the selected $x_{ii'k}$ between dataset (k) within a set of selected is, Ω_i, we define Mean vs SD ratio as

$$\langle x_{ii'k}\rangle_{ii'} = \frac{1}{K}\sum_{k=1}^{K} x_{ii'k} \tag{9}$$

$$\langle x_{ii'k}\rangle = \frac{1}{N(\Omega_i)^2}\sum_{i,i'\in\Omega_i}\langle x_{ii'k}\rangle_{ii'} \tag{10}$$

$$\mathrm{SD} = \frac{1}{N(\Omega_i)^2}\sum_{i,i'\in\Omega_i}\sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(x_{ii'k}-\langle x_{ii'k}\rangle_{ii'}\right)^2} \tag{11}$$

$$\text{Mean vs SD ratio} = \frac{SD}{\langle x_{ii'k}\rangle}. \tag{12}$$

The above summation of i, i' is taken in Ω_i and $N(\Omega_i)$ is the number of is in Ω_i.

5.4 Functional sites used to compute correlation with $u_{\ell_1 i}$s or $\langle x_{ii'k}\rangle_i$ for Hi-C datasets

5.4.1 Overlap with bins

Overlap with bins is computed by `findOverlaps` function in IRanges [19].

5.4.2 CTCF

CTCF profile was retrieved from AH104727 data in AnnotationHub [20].

5.4.3 PLS, pELS, and dELS

PLS, pELS, and dELS profiles were recovered from 41586_2020_2493_MOESM12_ESM.txt [21].

Declarations

Ethics approval and consent to participate

Not applicable

Competing interests

The authors declare no competing interests.

Funding

This work was supported by KAKENHI Grant Number 24K15168.

Data availability

All data used in this study are available in GEO as indicated in the paper.

Code availability

The code used to carry out our method is publicly available in the GitHub repository at https://github.com/tagtag/TDbasedUFEHiC.

Author contribution

Y.-H.T. planned the study and performed the analyses. Y.-H.T. and T.T. evaluated the results, discussions, and outcomes and wrote and reviewed the manuscript. All the authors have read and agreed to the published version of this manuscript.

References

[1] Yardımcı, G.G., Ozadam, H., Sauria, M.E.G., Ursu, O., Yan, K.-K., Yang, T., Chakraborty, A., Kaul, A., Lajoie, B.R., Song, F., Zhan, Y., Ay, F., Gerstein, M., Kundaje, A., Li, Q., Taylor, J., Yue, F., Dekker, J., Noble, W.S.: Measuring the reproducibility and quality of Hi-C data. Genome Biology **20**(1), 57 (2019) https://doi.org/10.1186/s13059-019-1658-7

[2] Yang, T., Zhang, F., Yardımcı, G.G., Song, F., Hardison, R.C., Noble, W.S., Yue, F., Li, Q.: Hicrep: assessing the reproducibility of hi-c data using a stratum-adjusted correlation coefficient. Genome Research **27**(11), 1939–1949 (2017) https://doi.org/10.1101/gr.220640.117 http://genome.cshlp.org/content/27/11/1939.full.pdf+html

[3] Ursu, O., Boley, N., Taranova, M., Wang, Y.X.R., Yardimci, G.G., Stafford Noble, W., Kundaje, A.: GenomeDISCO: a concordance score for chromosome conformation capture experiments using random walks on contact map graphs. Bioinformatics **34**(16), 2701–2707 (2018) https://doi.org/10.1093/

bioinformatics/bty164 https://academic.oup.com/bioinformatics/article-pdf/34/16/2701/48917757/bioinformatics_34_16_2701.pdf

[4] Yan, K.-K., Yardımcı, G.G., Yan, C., Noble, W.S., Gerstein, M.: HiCspector: a matrix library for spectral and reproducibility analysis of Hi-C contact maps. Bioinformatics **33**(14), 2199–2201 (2017) https://doi.org/10.1093/bioinformatics/btx152 https://academic.oup.com/bioinformatics/article-pdf/33/14/2199/50314921/bioinformatics_33_14_2199.pdf

[5] Sauria, M.E., Taylor, J.: Quasar: Quality assessment of spatial arrangement reproducibility in hi-c data. bioRxiv (2017) https://doi.org/10.1101/204438 https://www.biorxiv.org/content/early/2017/11/14/204438.full.pdf

[6] Zou, C., Zhang, Y., Ouyang, Z.: HSA: integrating multi-track Hi-C data for genome-scale reconstruction of 3D chromatin structure. Genome Biology **17**(1), 40 (2016) https://doi.org/10.1186/s13059-016-0896-1

[7] Stansfield, J.C., Cresswell, K.G., Dozmorov, M.G.: multiHiCcompare: joint normalization and comparative analysis of complex Hi-C experiments. Bioinformatics **35**(17), 2916–2923 (2019) https://doi.org/10.1093/bioinformatics/btz048 https://academic.oup.com/bioinformatics/article-pdf/35/17/2916/50719885/bioinformatics_35_17_2916.pdf

[8] Li, A., Zeng, G., Wang, H., Li, X., Zhang, Z.: Dedoc2 identifies and characterizes the hierarchy and dynamics of chromatin tad-like domains in the single cells. Advanced Science **10**(20), 2300366 (2023) https://doi.org/10.1002/advs.202300366 https://onlinelibrary.wiley.com/doi/10.1002/advs.202300366

[9] Taguchi, Y.-h.: Unsupervised Feature Extraction Applied to Bioinformatics: A PCA Based and TD Based Approach, 2nd edn. Unsupervised and Semi-Supervised Learning. Springer, Switzland (2024)

[10] Bittner, N., Shi, C., Zhao, D., Ding, J., Southam, L., Swift, D., Kreitmaier, P., Tutino, M., Stergiou, O., Cheung, J.T.S., Katsoula, G., Hankinson, J., Wilkinson, J.M., Orozco, G., Zeggini, E.: Primary osteoarthritis chondrocyte map of chromatin conformation reveals novel candidate effector genes. Annals of the Rheumatic Diseases **83**(8), 1048–1059 (2024) https://doi.org/10.1136/ard-2023-224945 https://ard.bmj.com/content/83/8/1048.full.pdf

[11] Nolan, B., Harris, H.L., Kalluchi, A., Reznicek, T.E., Cummings, C.T., Rowley, M.J.: Hicrayon reveals distinct layers of multi-state 3d chromatin organization. bioRxiv (2024) https://doi.org/10.1101/2024.02.11.579821 https://www.biorxiv.org/content/early/2024/02/12/2024.02.11.579821.full.pdf

[12] Harris, H.L., Gu, H., Olshansky, M., Wang, A., Farabella, I., Eliaz, Y., Kalluchi, A., Krishna, A., Jacobs, M., Cauer, G., Pham, M., Rao, S.S.P., Dudchenko, O., Omer, A., Mohajeri, K., Kim, S., Nichols, M.H., Davis, E.S., Gkountaroulis,

D., Udupa, D., Aiden, A.P., Corces, V.G., Phanstiel, D.H., Noble, W.S., Nir, G., Pierro, M.D., Seo, J.-S., Talkowski, M.E., Aiden, E.L., M, J.R.: Chromatin alternates between A and B compartments at kilobase scale for subgenic organization. Nature Communications **14**(1), 3303 (2023) https://doi.org/10.1038/s41467-023-38429-1

[13] Leinonen, R., Sugawara, H., Shumway, o.b.o.t.I.N.S.D.C. Martin: The Sequence Read Archive. Nucleic Acids Research **39**(suppl_1), 19–21 (2010) https://doi.org/10.1093/nar/gkq1019 https://academic.oup.com/nar/article-pdf/39/suppl_1/D19/7624335/gkq1019.pdf

[14] Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.-J., Vert, J.-P., Heard, E., Dekker, J., Barillot, E.: HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biology **16**(1), 259 (2015) https://doi.org/10.1186/s13059-015-0831-x

[15] Baglama, J., Reichel, L., Lewis, B.W.: Irlba: Fast Truncated Singular Value Decomposition and Principal Components Analysis for Large Dense and Sparse Matrices. (2022). R package version 2.3.5.1. https://CRAN.R-project.org/package=irlba

[16] Bates, D., Maechler, M., Jagan, M.: Matrix: Sparse and Dense Matrix Classes and Methods. (2024). R package version 1.7-0. https://CRAN.R-project.org/package=Matrix

[17] Taguchi, Y.-h.: TDbasedUFE: Tensor Decomposition Based Unsupervised Feature Extraction. (2023). https://doi.org/10.18129/B9.bioc.TDbasedUFE . R package version 1.0.0. https://bioconductor.org/packages/TDbasedUFE

[18] Taguchi, Y.-h.: TDbasedUFEadv Advanced Package of Tensor Decomposition Based Unsupervised Feature Extraction. (2023). https://doi.org/10.18129/B9.bioc.TDbasedUFEadv . R package version 1.0.0. https://bioconductor.org/packages/TDbasedUFEadv

[19] Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M., Carey, V.: Software for computing and annotating genomic ranges. PLoS Computational Biology **9** (2013) https://doi.org/10.1371/journal.pcbi.1003118

[20] Morgan, M., Shepherd, L.: AnnotationHub: Client to Access AnnotationHub Resources. (2023). https://doi.org/10.18129/B9.bioc.AnnotationHub . R package version 3.8.0. https://bioconductor.org/packages/AnnotationHub

[21] Abascal, F., Acosta, R., Addleman, N.J., Adrian, J., Afzal, V., Ai, R., Aken, B., Akiyama, J.A., Jammal, O.A., Amrhein, H., Anderson, S.M., Andrews, G.R., Antoshechkin, I., Ardlie, K.G., Armstrong, J., Astley, M., Banerjee, B., Barkal, A.A., Barnes, I.H.A., Barozzi, I., Barrell, D., Barson, G., Bates, D., Baymuradov,

U.K., Bazile, C., Beer, M.A., Beik, S., M, A.B., Bennett, R., Bouvrette, L.P.B., Bernstein, B.E., Berry, A., Bhaskar, A., Bignell, A., Blue, S.M., Bodine, D.M., Boix, C., Boley, N., Borrman, T., Borsari, B., Boyle, A.P., Brandsmeier, L.A., Breschi, A., Bresnick, E.H., Brooks, J.A., Buckley, M., Burge, C.B., Byron, R., Cahill, E., Cai, L., Cao, L., Carty, M., Castanon, R.G., Castillo, A., Chaib, H., Chan, E.T., Chee, D.R., Chee, S., Chen, H., Chen, H., Chen, J.-Y., Chen, S., J, M.C., Chhetri, S.B., Choudhary, J.S., Chrast, J., Chung, D., Clarke, D., Cody, N.A.L., Coppola, C.J., Coursen, J., D'Ippolito, A.M., Dalton, S., Danyko, C., Davidson, C., Davila-Velderrain, J., Davis, C.A., Dekker, J., Deran, A., DeSalvo, G., Despacio-Reyes, G., Dewey, C.N., Dickel, D.E., Diegel, M., Diekhans, M., Dileep, V., Ding, B., Djebali, S., Dobin, A., Dominguez, D., Donaldson, S., Drenkow, J., Dreszer, T.R., Drier, Y., Duff, M.O., Dunn, D., Eastman, C., Ecker, J.R., Edwards, M.D., El-Ali, N., Elhajjajy, S.I., Elkins, K., Emili, A., Epstein, C.B., Evans, R.C., Ezkurdia, I., Fan, K., Farnham, P.J., Farrell, N.P., Feingold, E.A., Ferreira, A.-M., Fisher-Aylor, K., Fitzgerald, S., Flicek, P., Foo, C.S., Fortier, K., Frankish, A., Freese, P., Fu, S., Fu, X.-D., Fu, Y., Fukuda-Yuzawa, Y., Fulciniti, M., Funnell, A.P.W., Gabdank, I., Galeev, T., Gao, M., Giron, C.G., Garvin, T.H., Gelboin-Burkhart, C.A., Georgolopoulos, G., Gerstein, M.B., Giardine, B.M., Gifford, D.K., Gilbert, D.M., Gilchrist, D.A., Gillespie, S., Gingeras, T.R., Gong, P., Gonzalez, A., Gonzalez, J.M., Good, P., Goren, A., Gorkin, D.U., Graveley, B.R., Gray, M., Greenblatt, J.F., Griffiths, E., Groudine, M.T., Grubert, F., Gu, M., Guigó, R., Guo, H., Guo, Y., Zheng, Y., Gursoy, G., Gutierrez-Arcelus, M., Halow, J., Hardison, R.C., Hardy, M., Hariharan, M., Harmanci, A., Harrington, A., Harrow, J.L., Hashimoto, T.B., Hasz, R.D., Hatan, M., Haugen, E., Hayes, J.E., He, P., He, Y., Heidari, N., Hendrickson, D., Heuston, E.F., Hilton, J.A., Hitz, B.C., Hochman, A., Holgren, C., Hou, L., Hou, S., Hsiao, Y.-H.E., Hsu, S., Huang, H., Hubbard, T.J., Huey, J., Hughes, T.R., Hunt, T., Ibarrientos, S., Issner, R., Iwata, M., Izuogu, O., Jaakkola, T., Jameel, N., Jansen, C., Jiang, L., Jiang, P., Johnson, A., Johnson, R., Jungreis, I., Kadaba, M., Kasowski, M., Kasparian, M., Kato, M., Kaul, R., Kawli, T., Kay, M., Keen, J.C., Keles, S., Keller, C.A., Kelley, D., Kellis, M., Kheradpour, P., Kim, D.S., Kirilusha, A., Klein, R.J., Knoechel, B., Kuan, S., Kulik, M.J., Kumar, S., Kundaje, A., Kutyavin, T., Lagarde, J., Lajoie, B.R., Lambert, N.J., Lazar, J., Lee, A.Y., Lee, D., Lee, E., Lee, J.W., Lee, K., Leslie, C.S., Levy, S., Li, B., Li, H., Li, N., Li, S., Li, X., Li, Y., Li, Y., Li, Y., Li, Y., Lian, J., Libbrecht, M.W., Lin, S., Lin, Y., Liu, D., Liu, J., Lu, A., Liu, T., X, S.L., Liu, Y., Liu, Y., Long, M., Lou, S., Loveland, J., Lu, A., Lu, Y., Lécuyer, E., Ma, L., Mackiewicz, M., Mannion, B.J., Mannstadt, M., Manthravadi, D., Marinov, G.K., Martin, F.J., Mattei, E., McCue, K., McEown, M., McVicker, G., Meadows, S.K., Meissner, A., Mendenhall, E.M., Messer, C.L., Meuleman, W., Meyer, C., Miller, S., Milton, M.G., Mishra, T., Moore, D.E., Moore, H.M., Moore, J.E., Moore, S.H., Moran, J., Mortazavi, A., Mudge, J.M., Munshi, N., Murad, R., Myers, R.M., Nandakumar, V., Nandi, P., Narasimha, A.M., Narayanan, A.K., Naughton, H., Navarro, F.C.P., Navas, P., Nazarovs, J., Nelson, J., Neph, S., Neri, F.J., Nery, J.R., Nesmith, A.R., J, S.N., Newberry, K.M., Ngo, V., Nguyen, R., Nguyen, T.B., Nguyen,

T., Nishida, A., Noble, W.S., Novak, C.S., Novoa, E.M., Nuñez, B., O'Donnell, C.W., Olson, S., Onate, K.C., Otterman, E., Ozadam, H., Pagan, M., Palden, T., Pan, X., Park, Y., E, C.P., Paten, B., Pauli-Behn, F., Pazin, M.J., Pei, B., Pennacchio, L.A., Perez, A.R., Perry, E.H., Pervouchine, D.D., Phalke, N.N., Pham, Q., Phanstiel, D.H., Plajzer-Frick, I., Pratt, G.A., Pratt, H.E., Preissl, S., Pritchard, J.K., Pritykin, Y., Purcaro, M.J., Qin, Q., Quinones-Valdez, G., Rabano, I., Radovani, E., Raj, A., Rajagopal, N., Ram, O., Ramirez, L., Ramirez, R.N., Rausch, D., Raychaudhuri, S., Raymond, J., Razavi, R., Reddy, T.E., Reimonn, T.M., Ren, B., Reymond, A., Reynolds, A., Rhie, S.K., Rinn, J., Rivera, M., Rivera-Mulia, J.C., Roberts, B.S., Rodriguez, J.M., Rozowsky, J., Ryan, R., Rynes, E., Salins, D.N., Sandstrom, R., Sasaki, T., Sathe, S., Savic, D., Scavelli, A., Scheiman, J., Schlaffner, C., Schloss, J.A., Schmitges, F.W., See, L.H., Sethi, A., Setty, M., Shafer, A., Shan, S., Sharon, E., Shen, Q., Shen, Y., Sherwood, R.I., Shi, M., Shin, S., Shoresh, N., Siebenthall, K., Sisu, C., Slifer, T., Sloan, C.A., Smith, A., Snetkova, V., Snyder, M.P., Spacek, D.V., Srinivasan, S., Srivas, R., Stamatoyannopoulos, G., Stamatoyannopoulos, J.A., Stanton, R., Steffan, D., Stehling-Sun, S., J, S.S., Su, A., Sundararaman, B., Suner, M.-M., Syed, T., Szynkarek, M., Tanaka, F.Y., Tenen, D., Teng, M., Thomas, J.A., Toffey, D., Tress, M.L., Trout, D.E., Trynka, G., Tsuji, J., Upchurch, S.A., Ursu, O., Uszczynska-Ratajczak, B., Uziel, M.C., Valencia, A., Biber, B.V., Velde, A.G.v.d., Nostrand, E.L.V., Vaydylevich, Y., Vazquez, J., Victorsen, A., Vielmetter, J., Vierstra, J., Visel, A., Vlasova, A., Vockley, C.M., Volpi, S., Vong, S., Wang, H., Wang, M., Wang, Q., Wang, R., Wang, T., Wang, W., Wang, X., Wang, Y., Watson, N.K., Wei, X., Wei, Z., Weisser, H., Weissman, S.M., Welch, R., Welikson, R.E., Weng, Z., Westra, H.-J., Whitaker, J.W., White, C., White, K.P., Wildberg, A., Williams, B.A., Wine, D., Witt, H.N., Wold, B., Wolf, M., Wright, J., Xiao, R., Xiao, X., Xu, J., Xu, J., Yan, K.-K., Yan, Y., Yang, H., Yang, X., Yang, Y.-W., Yardımcı, G.G., Yee, B.A., Yeo, G.W., Young, T., Yu, T., Yue, F., Zaleski, C., Zang, C., Zeng, H., Zeng, W., Zerbino, D.R., Zhai, J., Zhan, L., Zhan, Y., Zhang, B., Zhang, J., Zhang, J., Zhang, K., Zhang, L., Zhang, P., Zhang, Q., Zhang, X.-O., Zhang, Y., Zhang, Z., Zhao, Y., Zheng, Y., Zhong, G., Zhou, X.-Q., Zhu, Y., Zimmerman, J., Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., Kaul, R., Halow, J., Nostrand, E.L.V., Freese, P., Gorkin, D.U., Shen, Y., He, Y., Mackiewicz, M., Pauli-Behn, F., Williams, B.A., Mortazavi, A., Keller, C.A., Zhang, X.-O., Elhajjajy, S.I., Huey, J., Dickel, D.E., Snetkova, V., Wei, X., Wang, X., Rivera-Mulia, J.C., Rozowsky, J., Zhang, J., Chhetri, S.B., Zhang, J., Victorsen, A., White, K.P., Visel, A., Yeo, G.W., Burge, C.B., Lécuyer, E., Gilbert, D.M., Dekker, J., Rinn, J., Mendenhall, E.M., Ecker, J.R., Kellis, M., Klein, R.J., Noble, W.S., Kundaje, A., Guigó, R., Farnham, P.J., J, M.C., Myers, R.M., Ren, B., Graveley, B.R., Gerstein, M.B., Pennacchio, L.A., Snyder, M.P., Bernstein, B.E., Wold, B., Hardison, R.C., Gingeras, T.R., Stamatoyannopoulos, J.A., Weng, Z.: Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature **583**(7818), 699–710 (2020) https://doi.org/10.1038/s41586-020-2493-4