Reports

BioTechniques

For reprint orders, please contact: reprints@future-science-group.com

Comparison of normalization methods for Hi-C data

Hongqiang Lyu*[,1], Erhu Liu[1] & Zhifang Wu[1]

ABSTRACT

Hi-C has been predominately used to study the genome-wide interactions of genomes. In Hi-C experiments, it is believed that biases originating from different systematic deviations lead to extraneous variability among raw samples, and affect the reliability of downstream interpretations. As an important pipeline in Hi-C analysis, normalization seeks to remove the unwanted systematic biases; thus, a comparison between Hi-C normalization methods benefits their choice and the downstream analysis. In this article, a comprehensive comparison is proposed to investigate six Hi-C normalization methods in terms of multiple considerations. In light of comparison results, it has been shown that a cross-sample approach significantly outperforms individual sample methods in most considerations. The differences between these methods are analyzed, some practical recommendations are given, and the results are summarized in a table to facilitate the choice of the six normalization methods. The source code for the implementation of these methods is available at https://github.com/lhqxinghun/bioinformatics/tree/master/Hi-C/NormCompare

METHOD SUMMARY

Six normalization methods for Hi-C data were compared comprehensively in terms of multiple considerations, including heat map texture, statistical quality, influence of resolution, consistency of distance stratum and reproducibility of topologically associating domain architecture. Among these considerations, the quality of statistics was investigated in depth from three aspects, comprising distribution of interaction frequency, correlation of replicates and comparability of replicates between contexts. The performance of these methods is compared.

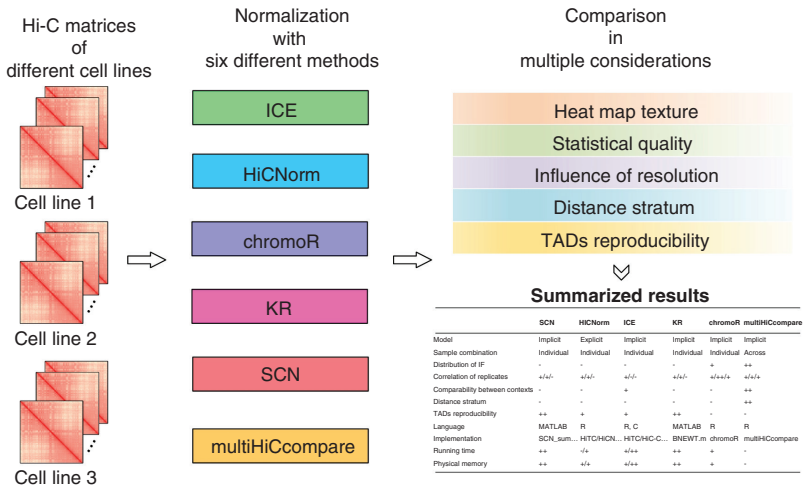
KEYWORDS

comprehensive comparison • Hi-C data • normalization methods

[1]*School of Electronic & Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China; *Author for correspondence: hongqianglv@mail.xjtu.edu.cn*

BioTechniques 68: 56-64 (February 2020) 10.2144/btn-2019-0105

GRAPHICAL ABSTRACT



	SCN	HiCNorm	ICE	KR	chromoR	multiHiCcompare
Model	Implicit	Explicit	Implicit	Implicit	Implicit	Implicit
Sample combination	Individual	Individual	Individual	Individual	Individual	Across
Distribution of IF	-	-	-	-	+	++
Correlation of replicates	+/+/-	+/+/-	+/+/-	+/+/-	+/++/+	+/+/+
Comparability between contexts	-	-	+	-	-	++
Distance stratum	-	-	-	-	-	++
TADs reproducibility	++	+	+	++	-	-
Language	MATLAB	R	R, C	MATLAB	R	R
Implementation	SCN_sum...	HiTC/HiCN...	HiTC/HiC-C...	BNEWT.m	chromoR	multiHiCcompare
Running time	++	-/+	+/++	++	+	-
Physical memory	++	+/+	+/++	++	+	-

Hi-C technology allows for genome-wide profiling of chromatin interactions in space [1]. It is well known that spatial organization of chromatin is non-random and is crucial for deciphering how the 3D architecture of DNA affects genome functionality and transcription [2,3]. As a chromosome conformation capture (3C)-based method [4], Hi-C provides a deeper insight into the 3D organization of chromatin by comprehensive detection of spatial interactions between genomic regions [1,5]. Compared with other 3C-based technologies, such as chromosome conformation capture-on-chip (4C), chromosome conformation capture carbon copy (5C) and ChIA-PET, Hi-C technology combines DNA proximity ligation with deep sequencing and is not dependent on associated proteins. These advances give it the power to implement genome-wide mapping of chromatin interactions [6–8].

Hi-C has predominately been used to study the genome-wide interactions of genomes [9]. Hi-C experiments usually produce hundreds of millions of paired-end sequencing reads, and this enormous amount of genomic data presents great opportunities and challenges. A series of *ad hoc* algorithms, computational and statistical methods, as well as bioinformatics tools are available for the exploration and interpretation of Hi-C data. These pipelines cover all current aspects of Hi-C analysis workflow, ranging from preprocessing of sequencing reads to normalization and inference of genome structure [3,10]. The preprocessing pipeline consists of read mapping, fragment assignment, filtering and binning, and we are left with a symmetrical contact matrix. Each entry in the matrix reflects the interaction frequency (IF) observed between the corresponding pair of loci (also known as bins). The two loci are separated by a fixed size genomic interval, which is conveyed as resolution [11,12]. It is found that the value of matrix entries exhibits an exponential decay in signal as the distance between loci increases [1], which is consistent with the expectation that 3D interactions mostly occur within chromosomes (*cis*) rather than between chromosomes (*trans*) [3]. Following preprocessing, normalization is carried out to correct systematic biases, making Hi-C samples more comparable and downstream analysis reliable [9]. The

Table 1. Summary of datasets used in the paper.

Species	Cell type	Restriction enzyme			Resolution	Replicate Samples (n)	Ref.
		HindIII	MboI	DpnII			
Human	GM12878	✓	✓	✓	2.5M/1M/500K/250K/100K/50K/10K/5K	14	[8]
	IMR90	✓	✓		2.5M/1M/500K/250K/100K/50K	8	[8,13]
	K562	✓	✓		2.5M/1M/500K/250K/100K/50K	8	[8,25]
Mouse	CH12-LX	✓	✓		2.5M/1M/500K/250K/100K/50K/10K/5K	10	[8]
	Liver	✓			2.5M/1M/500K/250K/100K/50K	2	[26]
	mES	✓			2.5M/1M/500K/250K/100K/50K	2	[13]

inference of genome architecture can then be investigated at different levels, such as topologically associating domains (TADs) [8,13].

Normalization is one of the most important pipelines in Hi-C data analysis [9]. Comparison between normalization methods benefits their choice as well as the downstream analysis. The raw outputs of many genomic technologies are influenced by technical biases and biological factors [14]. In Hi-C experiments, the interaction frequencies of contact matrix contain many unwanted biases that are derived from different systematic deviations in experimental procedures and driven by DNA sequence and technical variation, including library size, fragment length, GC content, sequence mappability, copy number variations and other unknown factors. It is believed that these biases lead to extraneous variability among raw samples and affect the reliability of downstream interpretations [15,16]. In Hi-C analysis workflow, normalization pipeline attempts to remove the unwanted systematic biases, so that the interaction frequencies reflecting the underlying architecture can be preserved as far as possible [17]. A number of normalization methods for correcting Hi-C data are available. These methods can be roughly grouped in different ways. According to model assumptions, they can be divided into explicit and implicit approaches [11]. Explicit approaches assume that the systematic biases, such as fragment length, GC content and sequence mappability, are known and accounted for in the statistical model [9]. A representative explicit method is HiCNorm [18]. Alternatively, implicit approaches assume that the cumulative effect of bias is captured in the sequencing coverage of each bin [9]. Typical methods are sequential component normalization (SCN) [19], iterative correction and eigenvector decomposition (ICE) [14], Knight-Ruiz (KR) [20] and chromoR [21]. This difference in model assumption makes the two groups of methods have their own characteristics in terms of strategy and implementation. For explicit approaches, probabilistic models are commonly used and their implementation relies on a variety of additional information, such as restriction site, genome sequence and mappability score. On the contrary, the strategies of implicit approaches are dominated by matrix balancing, spatial transformation and local regression, with fewer parameters and less additional information. In terms of sample combination, normalization methods can also be grouped into individual-sample and across-sample approaches. The difference between them is that the latter determines biases with the help of fusion of data from multiple samples, while the former does not. Thus far, most of the existing normalization methods are individual-sample approaches, with only a few belonging to across-sample ones, such as multiHiCcompare [22]. With the continuous accumulation of Hi-C data and the emergence of various normalization methods, understanding how normalization methods impact downstream analysis and how to choose them is a valuable tool. However, there has been no comprehensive comparison of Hi-C normalization methods. Recently, Forcato *et al.* quantitatively evaluated the performance of 13 Hi-C computational methods [3]. These methods cover the identification of TADs and interaction peaks [23], with normalization approaches not included. Thus, it is essential to comprehensively assess the performance of existing Hi-C normalization methods.

In this paper focusing on various datasets involving different species, cells and experimental designs, six Hi-C normalization methods including SCN, HiCNorm, ICE, KR, chromoR and multiHiCcompare were compared in terms of multiple considerations, including heat map texture, statistical quality, influence of resolution, consistency of distance stratum and reproducibility of TAD architecture. Among these considerations, the quality of statistics was investigated in depth from three aspects comprising distribution of IF, correlation of replicates and comparability of replicates between contexts. In regard to comparison of results, the differences between these methods were analyzed, some practical recommendations were given, and the itemized results for these considerations and representative implementations of these methods are summarized in Tables 1 & 2 to facilitate comparison of the six Hi-C normalization methods.

MATERIALS & METHODS

Datasets

A total of 44 real Hi-C samples involving different species, cells and experimental design were used to investigate how well the various Hi-C normalization methods can correct for the system biases in them. Hi-C contact matrices were prepared in two steps:

1. First, a highly compressed binary format file .hic was created with Pre command provided by Juicer [24] based on the paired-end reads from multiple ▶

Table 2. Summary of comparison results and representative implementations for the six Hi-C normalization methods.

	SCN	HiCNorm	ICE	KR	chromoR	multiHiC-compare
Model	Implicit	Explicit	Implicit	Implicit	Implicit	Implicit
Sample combination	Individual	Individual	Individual	Individual	Individual	Across
Distribution of IF	-	-	-	-	+	++
Correlation of replicates	+/+/-	+/+/-	+/-/-	+/+/-	+/++/+	+/+/+
Comparability between contexts	-	-	+	-	-	++
Consistency of distance stratum	-	-	-	-	-	++
Reproducibility of TADs	++	+	+	++	-	-
Language	MATLAB	R	R, C	MATLAB	R	R
Representative implementation	SCN_sumV2.m	HiTC /HiCNorm.R	HiTC/Hi-Corrector	BNEWT.m	chromoR	multiHiCcompare
Running time	++	-/+	+/++	++	+	-
Physical memory	++	+/+	+/++	++	+	-

'–' indicates that the method provided unsatisfactory results for the given consideration, while '+' indicates satisfactory results. For the item of correlation of replicates, */*/* represents the satisfactions of results from low to high resolutions in the range of 2.5M to 50K, and for the last two items, */* indicates the satisfactions of results produced by the previous corresponding implementation. The running time and physical memory were obtained in the same computing environment (Supplementary Table S1). ICE: Iterative correction and eigenvector decomposition; IF: Interaction frequency; KR: Knight-Ruiz; SCN: Sequential component normalization; TAD: Topologically associating domain.

resources. Herein, the Hi-C data derived from four Hi-C studies [8,13,25,26] were considered. A total of 44 .hic files were obtained, including 30 files for three hg19 human cells (GM12878, IMR90 and K562) and 14 files for three mm9 mouse cells (CH12-LX, Liver and mES). These files can also be downloaded from the Juicer data archive at https://bcm.app.box.com/v/aidenlab/ and Gene Expression Omnibus [27] database at https://www.ncbi.nlm.nih.gov/geo/.

2. Hi-C contact matrices were extracted from these .hic files using the Dump command provided by a java-based program juicer_tools [24].

Finally, the contact matrices for six cell contexts of human and mouse were prepared at eight different resolution levels, including 2.5M, 1M, 500K, 250K, 100K, 50K, 10K and 5K. All the datasets above are summarized in Table 1.

Normalization

Six different Hi-C normalization methods, including SCN, HiCNorm, ICE, KR, chromoR and multiHiCcompare, were considered in the comparative analysis. The categories and strategies are outlined as follows:

1. SCN is an implicit individual-sample approach for Hi-C normalization. It attempts to make contact matrix double stochastic using a matrix-balancing strategy. In SCN, all rows and columns of a contact matrix are successively scaled by dividing by the corresponding sums. This process is usually repeated until convergence.
2. HiCNorm is an explicit individual-sample approach for Hi-C normalization. It introduces a Poisson regression model to correct contact matrix. In HiCNorm, the systematic biases, including fragment length, GC content and sequence mappability, are estimated as a Poisson offset, and the residuals of the regression are regarded as the normalized matrix.
3. ICE is an implicit individual-sample approach for Hi-C normalization. It attempts to make all bins of contact matrix equally visible using a matrix-balancing strategy. In ICE, the systematic biases between two bins are considered as the product of their individual biases and the maximum likelihood solution for the individual biases is obtained by an iterative correction procedure, yielding a normalized matrix.
4. KR is a fast algorithm for balancing of square nonnegative matrices. It is widely used to correct Hi-C contact matrices as an implicit individual-sample normalization approach. In KR, a scaling vector is calculated using an inexact Newton iteration with conjugate gradient, and the matrix is made double stochastic by diagonal scaling.
5. chromoR is an implicit individual-sample approach for Hi-C normalization. It attempts to correct contact matrix by means of decomposition, de-noising and reconstruction procedures. In chromoR, IF is meant to be a Poisson random variable. Thus, Haar-Fisz transform can be applied to decompose the Poisson distributed frequency into Gaussian-distributed coefficients in multiple scales, followed by a de-noising procedure via wavelet shrinkage. A corrected matrix is then reconstructed using the inverse Haar-Fisz transform.
6. multiHiCcompare is an implicit across-sample approach for Hi-C normal-

ization. It allows for data-driven joint normalization based on locally weighted regression (loess). In multiHiCcompare, the difference in IF between two matrices with respect to bin distance, named difference versus distance (MD), is fitted using loess on a log scale. This estimated difference is equally segmented and compensated into the two original matrices in the opposite direction to remove systematic biases, and an anti-log transformation is employed to obtain the final normalized matrices.

During implementation, default parameter options were used in all function calls unless stated otherwise. The details of the implementations of these normalization methods are given in Supplementary Method 1.

logCPM transformation

The value of interaction frequencies of contact matrices normalized by the aforementioned methods are obviously on different scales due to the various strategies employed in these methods, even at the same resolution level. In order to adjust the value to the same scale across different methods, a transformation of log counts per million (logCPM) was used before comparative analysis. The definition of logCPM is given by:

$$Y_{i,j} = \log 10\left(\frac{X_{i,j}}{L}10^6 + s\right) \quad \text{(Equation 1)}$$

Where $X_{i,j}$ and $Y_{i,j}$ are the interaction frequencies of row *i* and column *j* in contact matrices before and after the transformation, respectively, L denotes the library size estimated by the sum of the lower triangular matrix, and *S* was set to 1 to ensure that $Y_{i,j}$ is non-negative.

Comparative design

The Hi-C normalization methods mentioned earlier were compared in terms of multiple considerations, including heat map texture, statistical quality, influence of resolution, consistency of distance stratum and reproducibility of TADs architecture. Herein, only the *cis* contact matrix is taken into account, since these methods have no ability to handle asymmetric *trans* contact matrix, apart from ICE and HiCNorm, and chromosomal matrix can be extended on a genomic level to examine *trans* interactions in implementation.

1. As mentioned earlier, Hi-C normalization methods are devoted to the elimination of unwanted systematic biases so that the interaction frequencies reflecting the underlying architecture can be preserved as far as possible. Thus, before any other investigation, it is important to verify that these methods have not caused excessive damage to the biological structure represented by raw data. Since heat map is the primary means of overall graphical presentation of Hi-C data, an intuitive comparison of texture between the heat maps of raw and normalized contact matrices is fundamental.
2. Comparison of the statistical quality is necessary, as it is the guarantee of consistency and authenticity of the results of subsequent procedures, such as architecture identification and differential analysis. To do this, first the IF distribution of raw and normalized contact matrices is examined. Second, the correlation of replicates is scored by both coefficient of variation (CV) and Spearman coefficient, where the CV is calculated per matrix element across replicates. Finally, the comparability of replicates between contexts is investigated via MA plot (Supplementary Method 2). There are some differences between the three considerations. The first is position independent, while the last two are position dependent and can be further distinguished depending on whether the comparison occurs within the same context or between two different contexts.
3. Resolution may have a considerable impact on the performance of Hi-C normalization methods, since it determines the dimension of contact matrix and is crucial to matrix sparsity. In order to determine the difference between these methods in response to changes in resolution, an investigation of some concerns, especially the considerations under which the comparison results are sensitive to resolution in the previous section, can be further conducted at multiple resolution levels.
4. It is known that the IF of contact matrix follows an exponential decay in signal as bin distance increases. For the consistency of this type of distance stratum, the IF of raw and each normalized contact matrix is fitted against the distance with the help of loess on a log scale, that is, loess(Yi,j log10(|i – j| + 1)). These loess-fitted curves are then checked to determine which normalization method can ensure a consistent decay pattern across samples. In addition, the stratum-adjusted correlation coefficient [28] within context is compared with that across contexts, to check the difference of variation between normalized contact matrices of intracontext and intercontexts. Furthermore, this variation difference can also be verified on each bin distance with the help of average deviation (AD) (Supplementary Method 2).
5. It is assumed that the contact matrices within the same context share similar TAD structures. For a comparison of TAD architecture reproducibility between normalization methods, the TADs of raw and normalized matrices are compared using TopDom [29] (Supplementary Method 2), and the Jaccard Index for concordance of TAD boundaries between replicates is computed as the measure of TAD reproducibility. Besides, the Jaccard Index of TADs boundaries between two different contexts is also calculated to check whether the reproducibility of TADs across contexts is lower than that within context after normalization.

RESULTS & DISCUSSION

Heat map

To conduct an intuitive comparison between the results of these six different methods, heat maps of the raw and normalized contact matrices for chromosome 1 (0–100,000,000) and 18 (0–75,000,000) of the Hi-C sample GM12878–001 and CH12-LX-104 at resolutions of 1M, 500K and 100K were given (Figure 1 & Supplementary Figures S1–S11). It can be seen that, at the three resolution levels, the matrices normalized by the different methods maintain a texture similar to the corresponding raw matrices, except for those by chromoR, whose details are not clear at 1M and 500K resolutions. This can be explained by the de-nosing procedure in chromoR, which helps to achieve higher correlation of replicates but at the same time blurs the details of contact matrix, especially at low resolutions. ▶

Statistical quality

The statistical quality of normalized contact matrices was compared from three different aspects, including distribution of IF, correlation of replicates and comparability of replicates between contexts. This analysis was conducted based on the contact matrices for chromosome 1 and 18 of GM12878, IMR90 and K562 contexts at five resolution levels, including 1M, 500K and 100K, 10K and 5K.

1. For distribution of IF, box plots of raw and normalized contact matrices across contexts were presented without outliers (Figure 2A & Supplementary Figures S12A–S16A). It can be seen that there are differences in the IF distribution between raw samples from multiple sources. These differences are expected to be removed by normalization procedures. Considering the six normalization methods, multiHiCcompare aligns the median and two quantiles of IF much better than any other method in all cases, and the stabilization of distribution is significantly improved. Regarding the other methods, chromoR usually has a smaller interquartile range (IQR) thanks to its de-noising procedure, leaving SCN, HiCNorm, ICE and KR with almost the same distribution as that of raw.
2. For correlation of replicates within context, box plots of CV and Spearman coefficient of contact matrices for GM12878 replicates were shown without outliers (Figure 2B, Supplementary Figure S12B–S16B & Figure S23). At resolutions of 1M and 500K, in most cases chromoR and multiHiCcompare can achieve a smaller CV and larger Spearman coefficient than raw and the other methods, especially in comparison to ICE; that is, the correlation of replicates by chromoR and multiHiCcompare is generally higher at low-resolution levels while the correlation by ICE is weaker, even weaker than that of raw in some cases. As for 100K, 10K and 5K resolutions, no significant difference can be observed in the box plots of CV and Spearman coefficient between these methods. Thus, the results of the comparison between replicate correlations vary at different resolution levels.
3. For comparability of replicates between contexts, an MA plot between contact matrices of GM12878 and IMR90 was generated to detect and visualize the IF difference (Figure 2C & Supplementary Figure S12C–S16C). Generally, the loess line fitted to the MA plot is improved or similar after normalization. In most cases, the loess fitted line of multiHiCcompare runs much closer to the zero line than those of the other methods. The fitted line of ICE is usually better than those of SCN, HiCNorm, KR and chromoR, and in some cases even preferable than that of multiHiCcompare (Supplementary Figure S16C). In addition, the line quality of chromoR varies greatly compared with the other methods. Beyond the loess-fitted line, the median and IQR of the M value were also counted. The median value of multiHiCcompare can be equal to zero in five out of six cases due to its strategy of differential segmentation and compensation on a log scale. chromoR has the smallest IQR in five out of six cases, which is consistent with the results of previous distribution analyses. In addition, the remaining normalization methods including SCN, HiCNorm, ICE and KR sometimes have smaller medians than chromoR in absolute value and lower IQRs than multiHiCcompare.

In addition to human data, the contact matrices for samples of three mouse contexts, including CH12-LX, Liver and mES, were also involved in exactly the same way, and consistent comparison results were received (Supplementary Figure S17–S22 & Figure S24).

Resolution

The level of resolution has a more significant impact on the results of comparison in terms of replicate correlation. Thus, the average CV and Spearman coefficients of all the contexts with more than two replicates were computed at six resolution levels, including 2.5M, 1M, 500K, 250K, 100K and 50K, for further analysis. The curves of the two coefficients versus resolution for chromosome 1 and 18 of GM12878, IMR90, K562 and CH12-LX contexts were given (Figure 3 & Supplementary Figure S25). Obviously, the average CV generally increases and the Spearman coefficient decreases as contact matrices step up to a higher resolution. That is to say, the correlation of replicates weakens with the increase in matrix dimension and sparsity. Considering the different normalization methods, the changes of these curves are roughly similar to the raw, except for those corresponding to chromoR, multiHiCcompare and ICE. In general, chromoR and multiHiCcompare can achieve a higher correlation of replicates compared with the other methods. At resolutions of 500K and 250K, the replicate correlation by chromoR is usually significantly higher. ICE has a slightly weaker correlation than the other methods at resolutions of 1M and 500K in most cases.

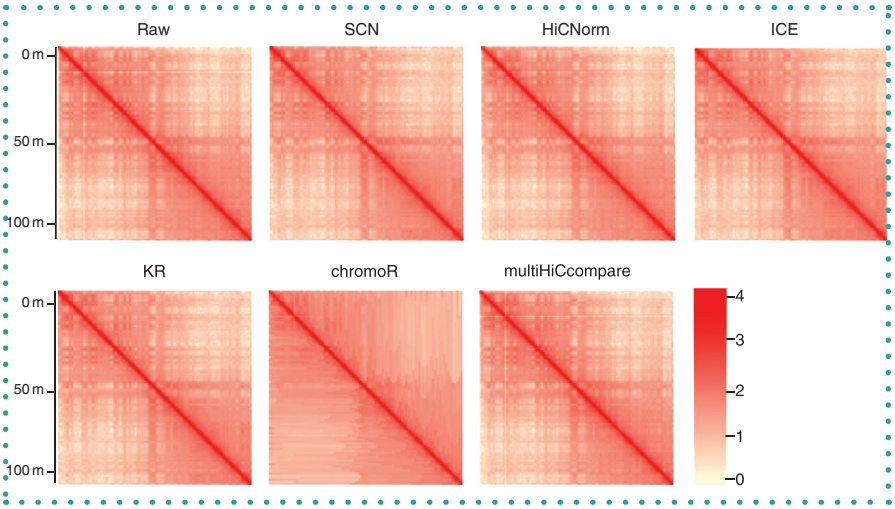


Figure 1. Comparison of heat maps between different methods. These heat maps correspond to raw and normalized contact matrices for chromosome 1 (chr1: 0–100,000,000) of the Hi-C sample GM12878–001 at 1M resolution.

ICE: Iterative correction and eigenvector decomposition; KR: Knight-Ruiz; SCN: Sequential component normalization.

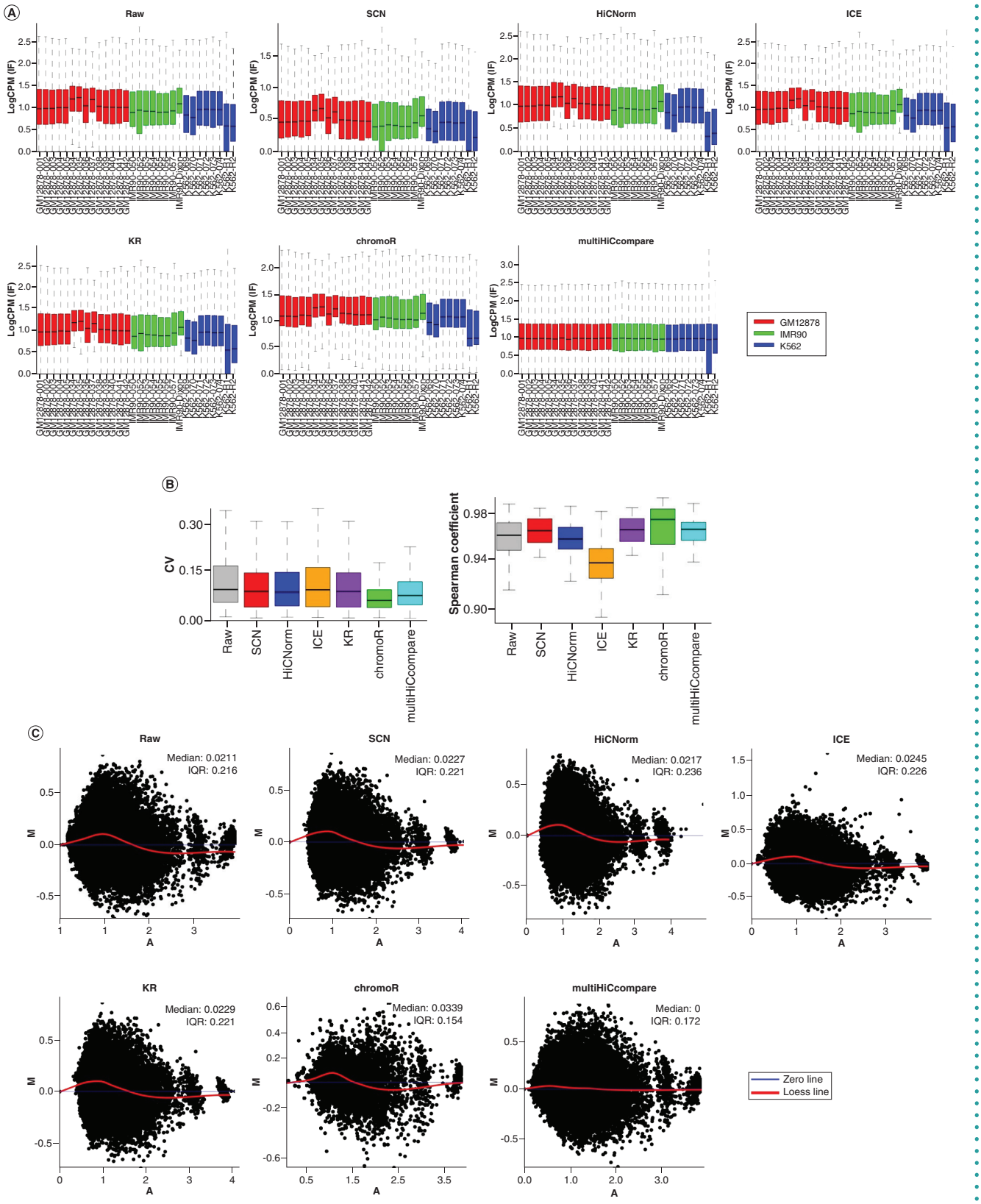


Figure 2. Comparison of statistical quality between different methods. This comparison is conducted on raw and normalized contact matrices of human chromosome 1 at 1M resolution. (A) Box plots of the interaction frequency for replicates of GM12878, IMR90 and K562. (B) Box plots of CV and Spearman coefficient for GM12878 replicates. (C) MA plot between replicates of GM12878 and IMR90.

CV: Coefficient of variation; ICE: Iterative correction and eigenvector decomposition; IQR: Interquartile range; KR: Knight-Ruiz; SCN: Sequential component normalization.

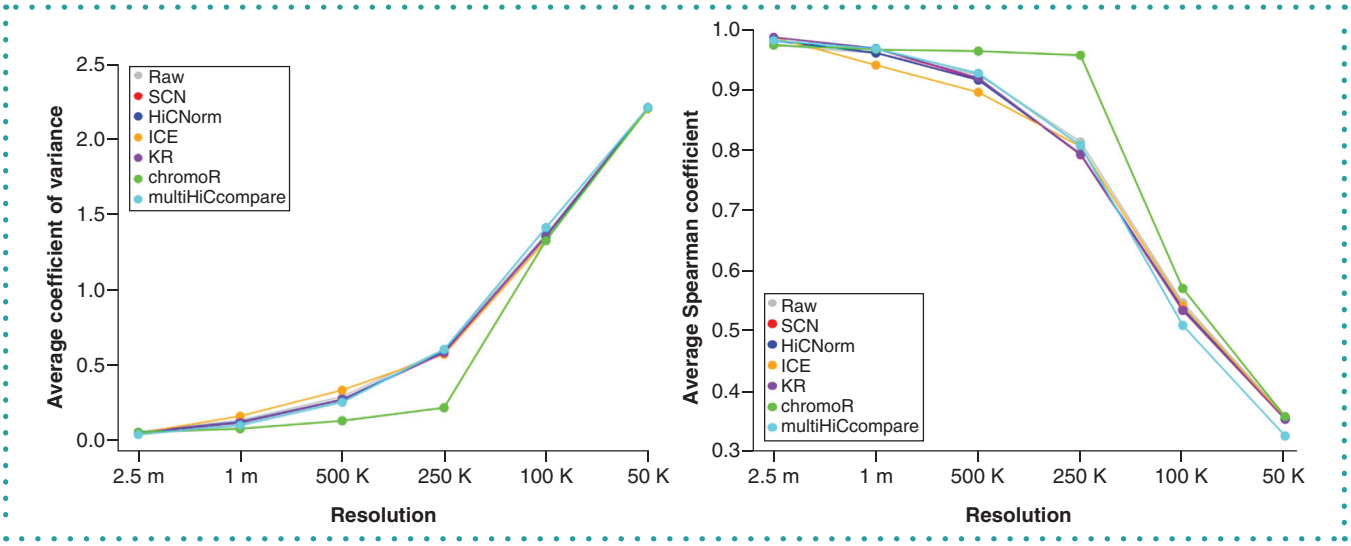


Figure 3. Comparison of average coefficients of variation and Spearman coefficients of replicates between different methods at multiple resolutions. The comparison is based on the raw and normalized contact matrices for chromosome 1 of GM12878 at six resolution levels, including 2.5M, 1M, 500K, 250K, 100K and 50K.

ICE: Iterative correction and eigenvector decomposition; KR: Knight-Ruiz; SCN: Sequential component normalization.

Distance stratum

To investigate how well the different normalization methods can produce a consistent exponential decay pattern across samples, the IF of raw and each normalized contact matrix within the three human contexts, including GM12878, IMR90 and K562, was fitted against the distance using loess on a log scale. The curves corresponding to chromosome 1 and 18 at two resolution levels, including 1M and 500K, were visualized (Figure 4 & Supplementary Figures S26–S28). Although the raw loess curves exhibit a similar decay pattern across replicates and even across contexts, they differ from each other in detail. In the reduction of such differences, the effectiveness of these normalization methods involved in the paper is not significant, except for multiHiCcompare, which presents a much better consistency of curve changes in all cases. It is no surprise that multiHiCcompare can achieve a much smaller variation between fitted curves, since it takes into account the distance stratum by means of a novel MD plot concept in the normalization design, while the other methods do not. This investigation was also conducted on the three mouse contexts, including CH12-LX, liver and mES, in exactly the same way, with identical comparison results received (Supplementary Figures S29–S32). In addition, the SCC between replicates within GM12878 and IMR90 were compared with that across the two contexts. As shown in box plots for

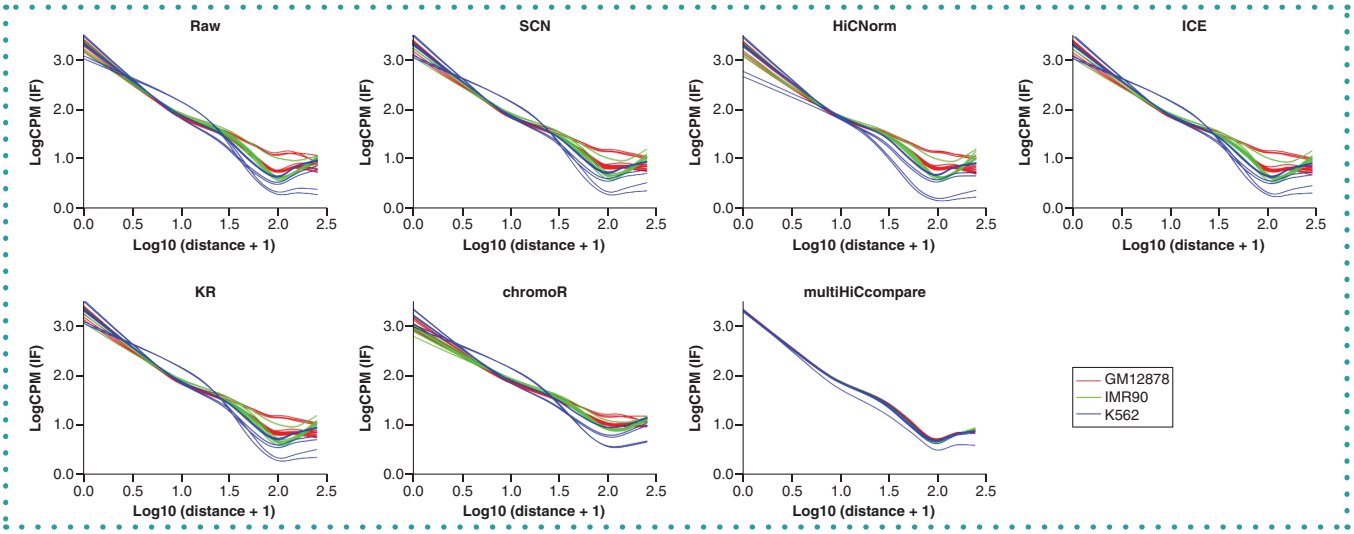


Figure 4. Comparison of loess-fitted curves of the interaction frequency versus bin distance between different methods. These curves are derived from raw and normalized contact matrices for chromosome 1 of GM12878, IMR90 and K562 at 1M resolution.

ICE: Iterative correction and eigenvector decomposition; IF: Interaction frequency; KR: Knight-Ruiz; SCN: Sequential component normalization.

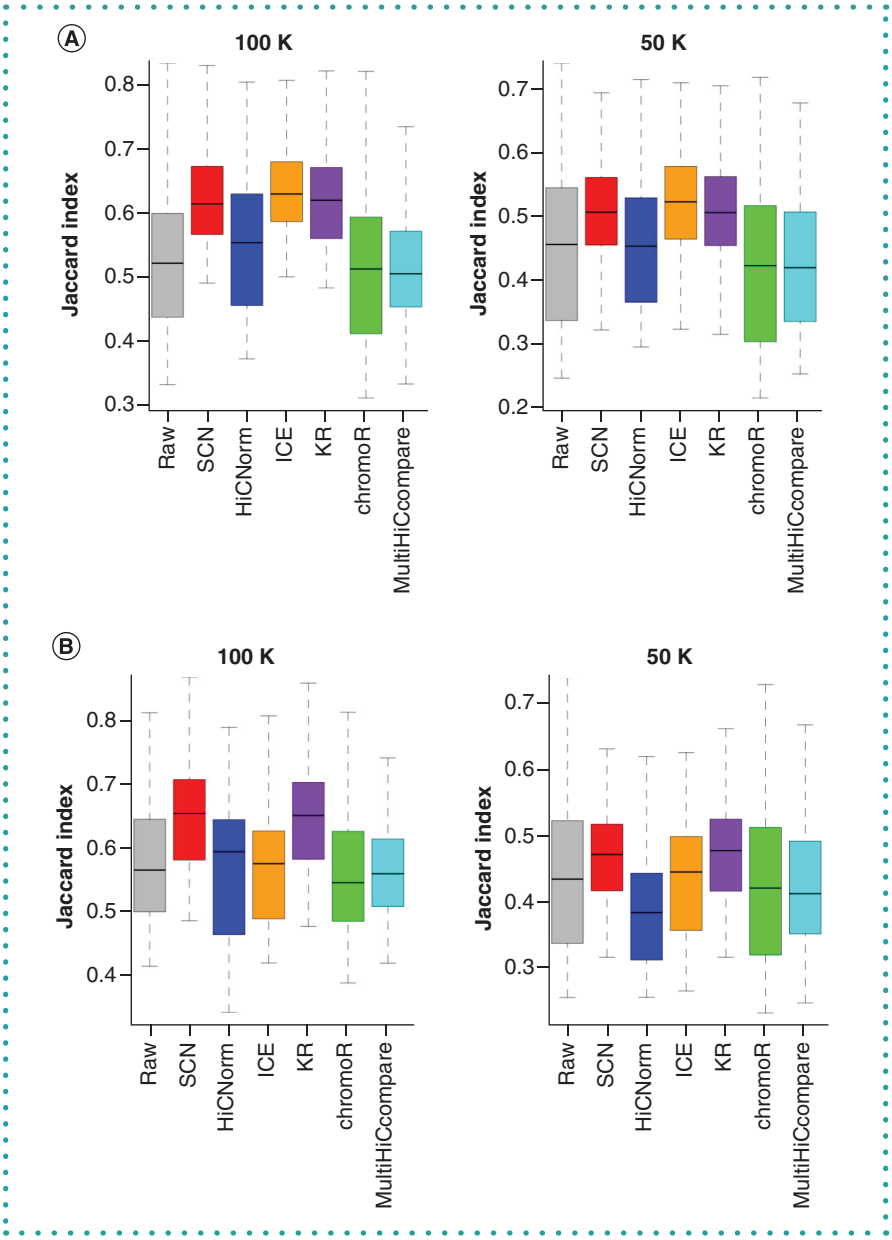


Figure 5. Comparison of topologically associating domain architecture reproducibility of replicates between different methods. Box plots of Jaccard Index for topologically associating domains between replicates are derived from raw and normalized contact matrices for (A) chromosome 1 and (B) chromosome 18 of GM12878 at 100K and 50K resolutions.

ICE: Iterative correction and eigenvector decomposition; KR: Knight-Ruiz; SCN: Sequential component normalization.

chromosome 1 at resolutions of 100K and 50K (Supplementary Figure S33), it can be seen that ICE and multiHiCcompare have a slightly higher correlation level within contexts compared with the other methods. All the methods produce a lower inter-context correlation level relative to their corresponding intracontext correlation levels in all cases. Furthermore, this difference in variation was also checked for each bin distance with the help of AD (Supplementary Figures S34 & S35). Generally, multiHiCcompare still shows a smaller variation, especially near the diagonal, and the AD ratio of intracontext to intercontext is less than 1 at most bin distances for all the methods; that is, the AD variation within context is usually smaller than that across contexts. This is in line with expectations.

TAD architecture

To compare the reproducibility of TAD architecture between different methods, the Jaccard Index for concordance of TAD boundaries within the contexts with more than two replicates was calculated, and the corresponding box plots for chromosome 1 and 18 of GM12878, IMR90, K562 and CH12-LX at resolutions of 100K and 50K were separately represented without outliers (Figure 5 & Supplementary Figure S36). It was shown that SCN and KR had a similar Jaccard Index distribution in almost all the cases, and their Jaccard Index was higher than that of any other method in nine out of 16 cases, especially compared with chromoR and multiHiCcompare. Thus, SCN and KR had the best TAD reproducibility in comparison to chromoR and multiHiCcompare. For HiCNorm and ICE, the distribution of the Jaccard Index varies greatly compared with the other methods – in some cases the Jaccard Index of ICE is even higher than that of SCN and KR. In addition, the Jaccard Index of TAD boundaries between two different contexts were also computed, and the corresponding box plots for chromosome 1 of GM12878 and IMR90 at resolutions of 100K and 50K were given (Supplementary Figure S37). Obviously, the reproducibility of TADs across contexts is lower than that within context in all the cases. This is consistent with our expectations, since the normalized matrices in the same context should share a more similar TAD architecture compared with those spanning different contexts.

The present comparison deals with six Hi-C normalization methods to determine the difference between them. Although there is no method that can be considered the gold standard to correct for systematic biases of Hi-C data, there are still some inferences that are beneficial to the choice of these methods. Different from the other methods, multiHiCcompare is an across-sample approach and takes into account the decay pattern of IF versus distance. These advances give multiHiCcompare the power to achieve an obviously better performance in most considerations, including distribution of IF, comparability of replicates between contexts and consistency of distance stratum. On the contrary, for such cross-sample approaches, larger memory is needed in order to allow all the matrices ▶

to be loaded at the same time. Among the other methods that are individual-sample approaches, SCN, KR and ICE are all based on the matrix-balancing strategy, and have been widely used in recent studies due to their conceptual simplicity and parameter-free nature. SCN and KR show a better reproducibility of TAD architecture. ICE usually produces a preferable comparability of replicates between contexts than the other methods except for multiHiCcompare, but has a slightly weaker correlation of replicates at resolutions of 1M and 500K in most cases. HiCNorm exhibits good performance in all the considerations, although some additional genomic features, such as genome sequence and mappability information, need to be specified during implementation since it is an explicit approach. As for chromoR, its de-nosing procedure helps to achieve smaller IQR and higher correlation of replicates, but at the same time blurs the details of contact matrix, especially at low resolutions. Thus, chromoR is preferably used for the normalization of Hi-C data at higher resolution levels. To facilitate selection of one of the six Hi-C normalization methods, the comparison results for these considerations, the representative implementations of these methods and their consumption of computational resources are summarized in Table 2.

FUTURE PERSPECTIVE

In light of the comparison results in this paper, it has been shown that the cross-sample normalization approach demonstrates significantly better performance than individual-sample methods in most considerations. Regarding the pace of Hi-C data production and the lack of such approaches at present, it is expected that some new competitive cross-sample normalization methods involving joint analysis of multiple replicates in different contexts will be designed in the near future. This will pave the way for optimal selection of Hi-C normalization methods and advancement of downstream procedures, such as architecture identification and differential analysis.

SUPPLEMENTARY DATA

To view the supplementary data that accompany this paper please visit the journal website at: www.future-science.com/doi/suppl/10.2144/btn-2019-0105

AUTHOR CONTRIBUTIONS

HL conceived and designed the study. HL and EL carried out acquisition of the data. HL, EL and ZW performed the analysis and interpretation of data. EL and ZW drafted the manuscript. HL revised it critically for important intellectual content, and gave final approval of the version to be published.

ACKNOWLEDGMENTS

The authors would like to thank Wenjun Yang for his assistance in sample collection.

FINANCIAL & COMPETING INTERESTS DISCLOSURE

This work has been supported by the National Natural Science Foundation of China under Grant 61602367. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

OPEN ACCESS

REFERENCES

1. Lieberman-Aiden E, Van Berkum NL, Williams L *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326(5950), 289–293 (2009).
2. Ethier SD, Miura H, Dostie J. Discovering genome regulation with 3C and 3C-related technologies. *Biochim. Biophys. Acta* 1819(5), 401–410 (2012).
3. Forcato M, Nicoletti C, Pal K, Livi CM, Ferrari F, Bicciato S. Comparison of computational methods for Hi-C data analysis. *Nat. Methods* 14(7), 679–685 (2017).
4. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. *Science* 295(5558), 1306–1311 (2002).
5. Belton J-M, Mccord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi–C: a comprehensive technique to capture the conformation of genomes. *Methods* 58(3), 268–276 (2012).
6. De Wit E, De Laat W. A decade of 3C technologies: insights into nuclear organization. *Genes Dev.* 26(1), 11–24 (2012).
7. Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.* 14(6), 390 (2013).
8. Rao SSP, Huntley MH, Durand NC *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159(7), 1665–1680 (2014).
9. Schmitt AD, Hu M, Ren B. Genome-wide mapping and analysis of chromosome architecture. *Nat. Rev. Mol. Cell Biol.* 17(12), 743–755 (2016).
10. Ay F, Noble WS. Analysis methods for studying the 3D architecture of the genome. *Genome Biol.* 16 (2015).
11. Lajoie BR, Dekker J, Kaplan N. The Hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods* 72, 65–75 (2015).
12. Le TBK, Imakaev MV, Mirny LA, Laub MT. High-resolution mapping of the spatial organization of a bacterial chromosome. *Science* 342(6159), 731–734 (2013).
13. Dixon JR, Selvaraj S, Yue F *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485(7398), 376 (2012).
14. Imakaev M, Fudenberg G, Mccord RP *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* 9(10), 999–1003 (2012).
15. Callister SJ, Barry RC, Adkins JN *et al.* Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J. Proteome Res.* 5(2), 277–286 (2006).
16. Vandernoot VA, Langevin SA, Solberg OD *et al.* cDNA normalization by hydroxyapatite chromatography to enrich transcriptome diversity in RNA-seq applications. *Biotechniques* 53(6), 373–380 (2012).
17. Shavit Y, Merelli I, Milanesi L, Lio P. How computer science can help in understanding the 3D genome architecture. *Brief. Bioinformatics* 17(5), 733–744 (2016).
18. Hu M, Deng K, Selvaraj S, Qin ZH, Ren B, Liu JS. HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics* 28(23), 3131–3133 (2012).
19. Cournac A, Marie-Nelly H, Marbouty M, Koszul R, Mozziconacci J. Normalization of a chromosomal contact map. *BMC Genomics* 13(1), 436 (2012).
20. Knight PA, Ruiz D. A fast algorithm for matrix balancing. *IMA J. Numer. Anal.* 33(3), 1029–1047 (2013).
21. Shavit Y, Lio P. Combining a wavelet change point and the Bayes factor for analysing chromosomal interaction data. *Mol. Biosyst.* 10(6), 1576–1585 (2014).
22. Stansfield JC, Cresswell KG, Dozmorov MG. multiHiCcompare: joint normalization and comparative analysis of complex Hi-C experiments. *Bioinformatics* 35(17), 2916–2923 (2019).
23. Cavalli G, Misteli T. Functional implications of genome topology. *Nat. Struct. Mol. Biol.* 20(3), 290–299 (2013).
24. Sureka R, Wadhwa R, Thakur SS, Pathak RU, Mishra RK. Comparison of nuclear matrix and mitotic chromosome scaffold proteins in ***Drosophila*** S2 Cells – transmission of hallmarks of nuclear organization through mitosis. *Mol. Cell. Proteomics* 17(10), 1965–1978 (2018).
25. Naumova N, Imakaev M, Fudenberg G *et al.* Organization of the mitotic chromosome. *Science* 342(6161), 948–953 (2013).
26. Rudan MV, Barrington C, Henderson S *et al.* Comparative Hi-C Reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep* 10(8), 1297–1309 (2015).
27. Barrett T, Wilhite SE, Ledoux P *et al.* NCBI GEO: archive for functional genomics data sets-update. *Nucleic Acids Res.* 41(D1), D991–D995 (2013).
28. Yang T, Zhang F, Yardimci GG *et al.* HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Res.* 27(11), 1939–1949 (2017).
29. Shin HJ, Shi Y, Dai C *et al.* TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res.* 44(7), (2016).