Briefings in Bioinformatics, 2022, **23**, 1–16

https://doi.org/10.1093/bib/bbab509

Problem Solving Protocol

HiC1Dmetrics: framework to extract various one-dimensional features from chromosome structure data

Jiankang Wang and Ryuichiro Nakato

Corresponding author. Ryuichiro Nakato, Institute for Quantitative Biosciences, The University of Tokyo, Tokyo 1130023, Japan. Tel.: +81-3-5841-1471; Fax: +81-3-5841-7308; E-mail: rnakato@iqb.u-tokyo.ac.jp

Abstract

Eukaryotic genomes are organized in a three-dimensional spatial structure. In this regard, the development of chromosome conformation capture methods has enabled studies of chromosome organization on a genomic scale. Hi-C, the high-throughput chromosome conformation capture method, can reveal a population-averaged, hierarchical chromatin structure. The typical Hi-C analysis uses a two-dimensional (2D) contact matrix that indicates contact frequencies between all possible genomic position pairs. Oftentimes, however, such a 2D matrix is not amenable to handling quantitative comparisons, visualizations and integrations across multiple datasets. Although several one-dimensional (1D) metrics have been proposed to depict structural information in Hi-C data, their effectiveness is still underappreciated. Here, we first review the currently available 1D metrics for individual Hi-C samples or two-sample comparisons and then discuss their validity and suitable analysis scenarios. We also propose several new 1D metrics to identify additional unique features of chromosome structures. We highlight that the 1D metrics are reproducible and robust for comparing and visualizing multiple Hi-C samples. Moreover, we show that 1D metrics can be easily combined with epigenome tracks to annotate chromatin states in greater details. We develop a new framework, called HiC1Dmetrics, to summarize all 1D metrics discussed in this study. HiC1Dmetrics is open-source (github.com/wangjk321/HiC1Dmetrics) and can be accessed from both command-line and web-based interfaces. Our tool constitutes a useful resource for the community of chromosome-organization researchers.

Keywords: Hi-C, chromosome organization, chromatin states, linear score

Introduction

In recent decades, researchers have made great efforts to discover how chromosomes are organized in three-dimensional (3D) space within a eukaryotic genome. Although experimental approaches such as fluorescence *in situ* hybridization and early-stage chromosome conformation capture (3C) can be used to investigate a handful of targeted regions [1], the high-throughput 3C technology Hi-C [2] provides a genome-wide assessment of 3D chromosome organization. The typical Hi-C analysis segments a genome into fixed-sized bins (e.g. 10 kb), estimates contact frequencies between all possible bin pairs and generates a two-dimensional (2D) contact matrix. The matrix is usually visualized as a square or triangular heatmap, then analyzed to identify chromatin structures such as chromatin loops and topologically associating domains (TADs), the sub-megabase (average 880 kb) structures [3, 4]. As the cost of next-generation sequencing has plummeted in recent years, numerous Hi-C datasets have been generated for a wide variety of species, with high resolution (up to hundreds of bp [5] and keep improving [6]). Many analysis tools also have been developed [7]. However, Hi-C analysis using a 2D matrix has several limitations. First, the file for 2D-matrix analysis often becomes large because the linear increase of resolution requires the quadratic increase of sequencing reads [8], leading to the heavy computational task. Second, the visualization and comparison of multiple Hi-C samples using 2D heatmaps can be problematic. Although multiple 2D heatmaps can be displayed side by side, human visual perception of differences in color intensity is inherently inaccurate [9]. Third, whereas it is important to integrate Hi-C data with other epigenomic data such as chromatin immunoprecipitation sequencing (ChIP-seq), it is not trivial to directly combine a 2D heatmap with linear tracks, i.e. combine contact frequency between each pair of genomic loci with values at each genomic locus of a particular epigenomic track. This potential inconsistency can confound the quantitative analysis of Hi-C data. Consequently, it is necessary and important to utilize well-designed compressed information extracted from

Jiankang Wang is a PhD student at the Graduate School of Medicine, The University of Tokyo. He is interested in epigenomics and chromosome organization.
Ryuichiro Nakato is a Lecturer at the Institute for Quantitative Biosciences, The University of Tokyo. His research interest is data-driven analysis using large-scale next-generation sequencing data.

Received: August 23, 2021. **Revised:** November 3, 2021. **Accepted:** November 5, 2021

Hi-C data in addition to the 2D matrix-based analysis to analyze chromatin data in detail.

Much effort has been devoted to developing one-dimensional (1D, linear) metrics for Hi-C analysis. Lieberman-Aiden *et al.* [2] applied principal component analysis (PCA) to the normalized 2D matrix and used the first eigenvector (PC1) to divide the whole genome into compartment A (open chromatin, positive PC1 value) and B (closed chromatin, negative PC1 value) [10]. This 'Compartment PC1' was the first 1D metric for Hi-C analysis. The sign of PC1 is arbitrary and it is usually fixed by correlating with gene density [11]. At the finer level of chromatin organization, TADs are highly conserved across cell lines and species, and associated with development and diseases [1]. Various 1D metrics have been proposed for analysis of TAD structure [12]. For example, Dixon *et al.* [3] defined a 'directionality index' (DI) to describe the bias in contact frequency between regions upstream and downstream of a TAD boundary; Crane *et al.* [13] developed an 'insulation score' (IS) to quantify interactions passing across each genomic locus; Ramirez *et al.* [14] suggested a 'separation score' (SS) to represent the degree of TAD separation; Van *et al.* [15] applied a 'contrast index' (CI) to assess the strength of TAD boundaries. Except for these 1D metrics designed for TAD calling, Heinz *et al.* [16] also tried to measure intra- and inter-chromosomal compaction for every genomic locus using 'distal-to-local ratio' (DLR) and 'inter-chromosomal fraction of interactions', respectively. 1D metrics are also useful for comparing two Hi-C samples by representing the changes of interactions at each locus. For example, the 'insulation score change' (ISC) was designed to reflect local differences in chromosomal organization [17]. The 'correlation difference' (CD) was designed to correlate locus-specific interaction profiles between two Hi-C samples [16]. These 1D metrics have the advantage [18] of being easily visualized with a conventional genome browser, which facilitates data interpretation and analysis of public datasets. Also, 1D metrics are especially helpful for quantitative comparisons of multiple samples and integration with other linear tracks.

Despite the usefulness of 1D metrics, there is no thorough tool and review that covers all these 1D metrics, which has restricted the feasibility of using 1D metrics for Hi-C analysis. Moreover, recent Hi-C studies have suggested more specific chromatin structures such as architectural stripes [19] and chromatin hubs [20], which cannot be captured by current metrics. Here, at first, we review the currently available 1D metrics for individual Hi-C sample or two-sample comparisons and then discuss their validity and suitable analysis scenarios. Next, we introduce the new 1D metrics we have developed, namely 'intra-TAD score' (IAS) and 'adjusted interaction frequency' (IF), to explore chromatin stripes and hubs, respectively. We give examples to demonstrate the biological relevance of these new metrics. We also developed another 1D metric, 'directional relative frequency' (DRF), for two-sample comparisons. Using DRF, we introduce the novel chromatin structure 'directional TAD' (dTAD), which depicts an asymmetric event of inter-TAD interactions. Then, through analysis of publicly available Hi-C datasets, we highlight that the 1D metrics-based approach is reproducible and robust for comparison and visualization of multiple Hi-C samples. Finally, we show that the linear tracks of 1D metrics can be combined with other epigenome data to annotate chromatin states in greater detail, using ChromHMM [21] as an example.

To increase the usability of 1D metrics, we developed a new framework, 'HiC1Dmetrics', to summarize all 1D metrics discussed in this study. HiC1Dmetrics can simultaneously deal with various types of 1D metrics and multiple Hi-C samples. HiC1Dmetrics is an open-source software (https://github.com/wangjk321/HiC1Dmetrics) and can be accessed via both command-line and web-based interfaces. We believe that our study will enhance the value of the Hi-C assay and facilitate studies of chromosome organization.

Results

Existing 1D metrics for a single Hi-C sample

Our HiC1Dmetrics can calculate multiple published metrics including IS, SS, DI, CI, DLR and PC1 (Figure 1A and B). As a representative dataset, we downloaded a high-resolution (5 kb) Hi-C data for HCT-116 cells [22] from the Gene Expression Omnibus (GEO) under accession number GSE104334. Other datasets used in this study were summarized in Supplementary Table S1. An example region is shown in Figure 1B (the same plots for GM12878 and K562 cells are shown in Supplementary Figure S1A). In this section, we will briefly review the existing metrics that are included in HiC1Dmetrics. We also summarized Table 1 to better illustrate the candidate scenarios for each 1D metric.

IS is defined for each bin as the average number of interactions that occur across the bin [13]. Therefore, IS has local minimums at the highly insulated regions, which also represent TAD boundaries. SS represents inter-TAD interactions normalized by intra-TAD interactions, and it reaches a local minimum between TADs [14]. Compared with IS, the additional information for SS enables the evaluation of the degree of TAD separation, implying that adjacent TADs with more contacts between them (less separation) receive a larger score. DI is used to quantify the degree of downstream or upstream interaction bias for each genomic position [3]. DI should have a local minimum at the 3′ end of each TAD and a local maximum at the 5′ end of each TAD. Quantitative comparison of DI values between two Hi-C samples is less interpretable because the meaningful information of DI is the sharp changes at TAD boundaries [23]. CI is defined as the local contrast, for which a high value corresponds to enriched intra-domain contact relative to inter-domain contacts, which are generally exhibited by TAD boundaries [15]. CI also

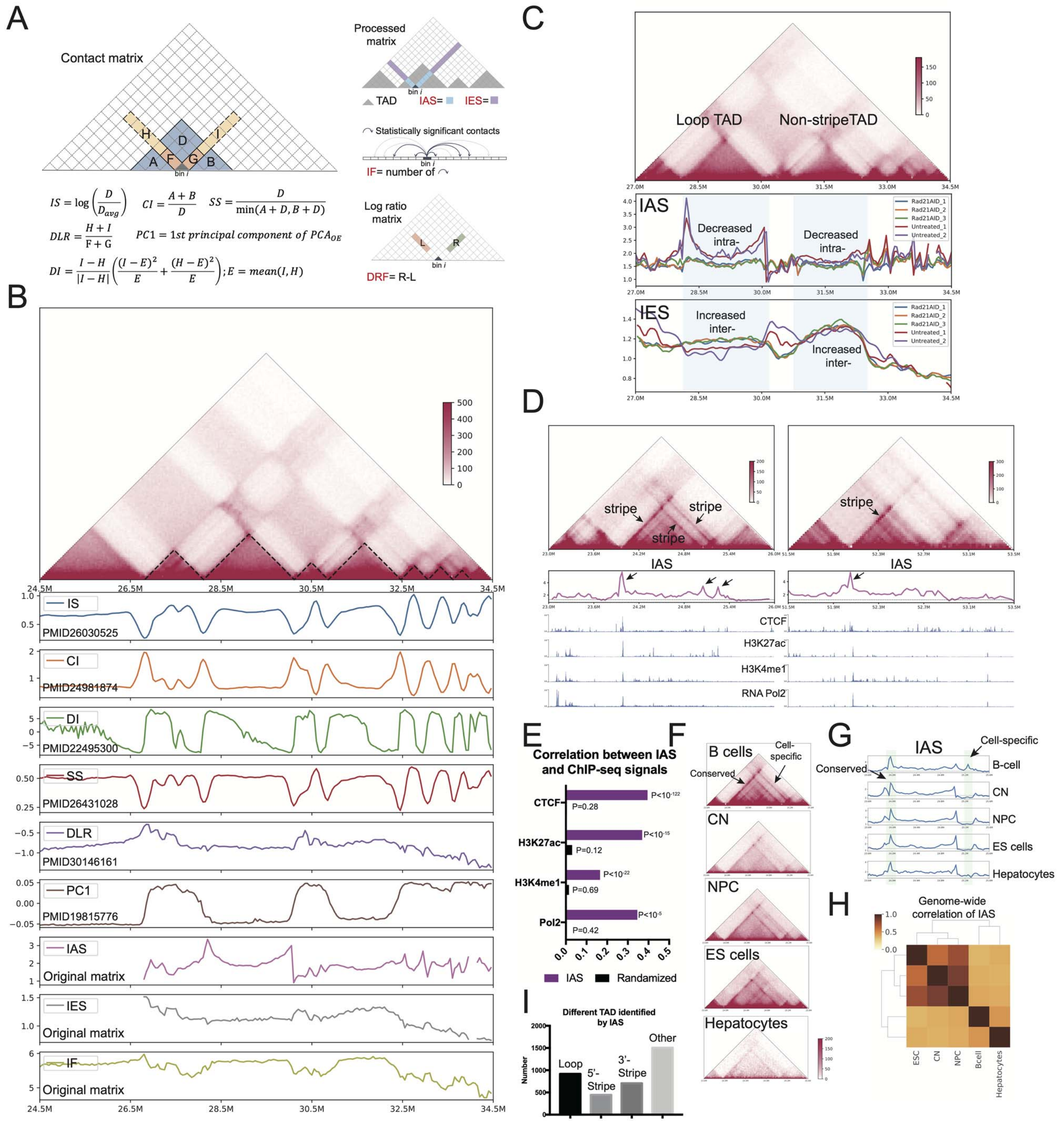

Figure 1. 1D metrics designed for a single Hi-C sample, and the novel metric IAS. (**A**) Concept of published 1D metrics and our original metrics. $\mathrm{PCA_{OE}}$ represents the PCA of observed/expected matrix. (**B**) Visualization of 1D metrics applied to Hi-C data acquired with non-treated HCT116 cells (chromosome 21, 24.5–34.5 Mb). (**C**) IAS and IES metrics for Rad21AID-treated or non-treated Hi-C samples (chromosome 21, 27.0–34.5 Mb). Blue shading denotes regions with decreased intra-TAD interactions and increased inter-TAD interactions. (**D**) IAS values for two candidate regions reveal that a stripe can be indicated by the peak IAS value. The lower panel shows the ChIP-seq signals for the indicated factors. B cells dataset from GSE98119 was used. (**E**) Spearman correlations between IAS values (or randomized IAS values) and ChIP-seq signals in B cells. (**F**) 2D heatmap for various cell types. The arrows indicate a conserved or cell type-specific stripe. (**G**) IAS signals for different cell lines. The plotted region is the same as in panel C. (**H**) Spearman correlation of genome-wide IAS values reveals the overall cell-type specificities of stripes. (**I**) Number of the various types of TADs identified by IAS.

can be used to assess the strength of boundaries. SS and CI are more appropriate than IS for comparing different TAD boundaries within a genome because they are normalized by local chromatin [15, 24].

Whereas the aforementioned metrics were designed for TAD calling with additional information, DLR helps quantify the degree of localized chromatin compaction for each genomic locus [16]. For example, a larger DLR value reflects lesser compaction of local chromatin (i.e. distance <3 Mb), enabling more interactions with distal sites (i.e. >3 Mb). DLR is related to the binding of cohesin, a key protein complex that mediates

Table 1. Candidate scenarios for various 1D metrics

	Raw	Existing 1D metrics						Novel 1D metrics		
	2D matrix	IS	CI	DI	SS	DLR	PC1	IAS/IES	IF	DRF
Local visualization	○	○	○	○	○	○	○	○	○	○
Genome-wide visualization	△	○	○	○	○	○	○	○	○	○
Multiple sample visualization	△	○	○	○	○	○	○	○	○	○
Quantitative comparison	△	○	○	×	○	○	×	○	○	○
Integrate other 1D track (e.g. ChIP-seq)	△	○	○	○	○	○	○	○	○	○
TAD analysis	△	○	○	○	○	×	×	○	×	○
Chromatin interaction analysis	△	△	△	×	△	○	○	○	○	○
Hi-C reproducibility	×	○	○	○	○	○	△	○	○	×
Expenditure (file size, running time)	△	○	○	○	○	○	△	○	○	○
'Stripe'	×	×	×	×	×	×	×	○	×	×
'Hubs'	×	×	×	×	×	×	×	×	○	×
'Directional TAD'	×	×	×	×	×	×	×	×	×	○
Information loss	○	△	△	△	△	△	△	△	△	△

Note: ○ Good △ Moderate ×Bad.

chromatin architecture [16]. In fact, we observed negative correlations between DLR and cohesin binding (Supplementary Figure S1B), implying that cohesin is required for compaction of local chromatin. PC1 is used to classify a genome into compartments A and B but is not appropriate for quantitative comparisons because PCA is an unbiased method [16]. In practice, the 'saddle plot'-based compartment analysis [25], or binary comparisons using compartment A/B information (called 'compartment switching') are often used. Overall, compare with 2D matrix, these 1D metrics offer a rapid way to quantitatively analyze particular chromatin structure [12, 18, 26].

The novel metric IAS captures architectural stripes

Considering TAD structures obtained by 2D-based analysis, all chromatin contacts can be divided into two classes: intra-TAD, i.e. the rectangular region along the diagonal with high-density interactions, and inter-TAD, i.e. the region surrounding TADs which usually have fewer inter-TAD interactions [27]. As Hi-C is inherently skewed towards short-range interactions that represent the vast majority of chromatin interactions, it is difficult to directly compare contact frequencies within or between TADs [28]. Although CI and SS represent the pattern of intra-TAD and inter-TAD interactions based on their ratio, the information of each is lost. In addition, CI and SS use fixed window size for calculations, which is unfair for comparing TADs of different size. To overcome these aspects, we propose two new metrics, namely 'intra-TAD score' (IAS) and 'inter-TAD score' (IES), that reflect the ratio of measured interactions to the expected contact frequency within and between TADs, respectively (Figures 1A and B and Methods). IAS and IES are normalized by both contact distance and TAD size.

To demonstrate the benefits of IAS and IES, we implemented a comparative Hi-C analysis before and after the auxin-mediated depletion of the cohesin subunit Rad21 in HCT116 cells (GSE104334). As reported in the original study [22], Rad21 depletion resulted in decreased intra-TAD interactions and increased inter-TAD interactions (Supplementary Figure S1C). This tendency was captured by a lower IAS value yet higher IES value (Figure 1C). Moreover, IAS could distinguish loop-TAD (28.0–30.0 Mb), according to the peaks at the 5′ and 3′ end of the TAD. IAS also enabled the discovery of the 'stripe' structure, which indicates high-level interactions between a certain locus and a contiguous region [19]. The stripe structures tether highly active regulatory regions, i.e. tissue-specific enhancers, suggesting their physiological impact [29]. We also used the original dataset of the stripe study [19] (GSE98119) and observed clear IAS peaks at stripe loci (Figure 1D). The IAS peaks overlapped with the ChIP-seq peaks for the insulator factor CTCF and active chromatin marks (H3K27ac, H3K4me1 and RNA Pol2). Genome-wide analysis also revealed that IASs were significantly correlated with those marks (Figure 1E), suggesting the importance of these marks for stripes. Consequently, IAS provides a quantitative means of studying stripes together with other 1D tracks (ChIP-seq), which is challenging using a conventional 2D matrix.

To further evaluate the applicability of IAS, we compared Hi-C samples from different cell types to explore the cell-type specificity of stripes, which has not been fully investigated previously. As shown in Figures 1F and G, IAS clearly represented the conserved stripes based on common IAS peaks, as well as the cell-type specific stripes that can only be observed in B cells. Over the entire genome, the stripes are more conserved among embryonic stem (ES) cells, cortical neurons (CN) and neural progenitor cells (NPC) but not B cells and hepatocytes (Figure 1H). We again emphasize that it is difficult to quantitatively evaluate such differences among multiple Hi-C samples using 2D heatmaps directly. IAS has merits for integrating with other 1D tracks and when comparing among Hi-C samples. To benefit Hi-C researchers, HiC1Dmetrics also uses IAS to provide the function for extracting stripes (Methods). Even though the demand of detecting stripes has been proposed [30], HiC1Dmetrics is the first tool that can automatically call stripes.

Because IAS is based on TADs, the results may be affected by TAD-calling tools and resolution. We confirmed that the result of IAS was consistent at different resolutions (Supplementary Figure S1D) and TAD-calling tools (Supplementary Figures S2A–C). Although different tools may yield a different TAD list, the TADs obtained with our custom method (IS based [13]) overlapped substantially with those obtained with Juicer [31] (Supplementary Figure S2A). We also observed moderately high correlations (Spearman >0.8) among the IAS obtained with different TAD-calling tools (Supplementary Figure S2B and C). HiC1Dmetrics also supports a parameter that specifies the external TAD list. In addition, as described by Barrington *et al*. [29], TADs can be classified as 5′-stripe, 3′-stripe, loop or other TADs, with distinct epigenomic and genomic profiles. For example, stripe-TADs exhibit asymmetric cohesin binding and regulatory machinery towards the anchored sites (3′ or 5′). Stripe-TADs are more susceptible than loop-TADs to structural changes during differentiation. HiC1Dmetrics applies a Z-test based on IAS and classifies all TADs as 5′-stripe, 3′-stripe, loop or other TADs (Figure 1I and Methods).

The novel metric IF identifies chromatin hubs

It has been reported that chromatin hubs, i.e. chromatin regions having a relatively high frequency of contacts, are extensively correlate with transcriptional regulation and histone modifications [20]. Hubs are highly conserved and tend to serve as super enhancers in gene regulatory programs that are enriched during development and in disease-associated variants. Here we defined ‘adjusted interaction frequency’ (IF), a normalized 1D metric to quantify the relative frequency of statistically significant chromatin interactions for each genomic locus (concept in Figure 1A, example in 1B and also see Methods). A larger IF value indicates relatively more interactions anchored from the locus. HiC1Dmetrics could detect 1348 hubs within a whole genome by extracting genomic loci in the top 10% of IF values (Figure 2A).

To demonstrate the biological relevance of IF, we compared Hi-C data [32] for MCF-7 cells (MCF_0h), MCF7 cells treated with estrogen (MCF7_24h) and tamoxifen-resistant MCF7 cells (MCF7_TamR), which allowed comparison of the dynamics of chromosome organization during endocrine resistance (Figure 2B). The 2D heatmap did not reveal any clear differences, but the IF plot strongly suggested that a hub was established in early-response cells (MCF7_24h) and drug-resistant cells (MCF7_TamR; Figure 2C). Notably, transcriptome analysis (RNA-seq) revealed that the expression of genes nearby the newly established hub increased significantly in TamR cells. Moreover, those genes constitute a cluster of oncogenes, for which high expression is associated with reduced survival, underscoring the biological relevance of the newly established hub. Genome-wide analysis also showed similar results (Supplementary Figures S2D–E). Thus, IF can provide new biological insights that cannot be easily elucidated based on the typical 2D heatmap. Because IF is calculated based on the number of significant interactions, we downsampled the Hi-C data to test the changes of IF values. Figure 2D showed that IF is changed slowly and gradually as downsampling rate increased.

In summary, 1D metrics for Hi-C data provide unique information pertaining to chromosome structures. Figure 2E depicts the correlation across metrics. SS is positively correlated with IS but negatively correlated with CI, suggesting that they provide similar but not identical information for boundaries. The other 1D metrics had relatively lower coefficients between each other, suggesting that each metric provides distinct information. Supplementary Figure S2F summarizes the distribution of the 1D metrics, revealing that most metrics, except DI and PC1, exhibited a unimodal distribution without obvious skew, for which the IF value was zero for a small subpopulation owing to centromeres or other regions with no detected interactions. These distributions agreed with the feasibility of quantitative comparisons.

1D metrics designed for two-sample comparisons

The 1D metrics are also powerful approaches for comparing two Hi-C samples under different conditions. HiC1Dmetrics also includes several published metrics that are designed for two-sample comparisons. Here, we used the same dataset for Figure 1B to compare Hi-C between Rad21-depleted and control HCT116 cells (Figure 3A). Because IS is scaled by its average value, the IS values for treated and control samples can be compared among samples. As proposed by Viny *et al*. [17], a change in IS (i.e. ISC) reflects changes in local chromosomal contacts. A positive ISC value indicates a decrease in insulation level. Similarly, we defined the change in CI (CIC) and the change in SS (SSC), both of which revealed local structure changes, especially for TAD boundaries. Delta-DLR was defined previously [16], and a larger value indicates more interactions with distal sites, suggesting the ‘decompaction’ of the locus. Delta-DLR correlates well with changes in transcription of large genes [16]. In contrast, PC1 is not suitable for direct comparison of Hi-C samples. Instead, the CD metric estimates the correlation between interaction profiles of two Hi-C samples for each locus [33]. A high CD value implies that two samples have a similar interaction pattern for a given locus.

As shown in Figure 1C, our two defined metrics IAS and IES can be used to quantify the relative levels of intra- and inter-TAD contacts. We also defined two additional 1D metrics, IASC and IESC, to determine the change (log

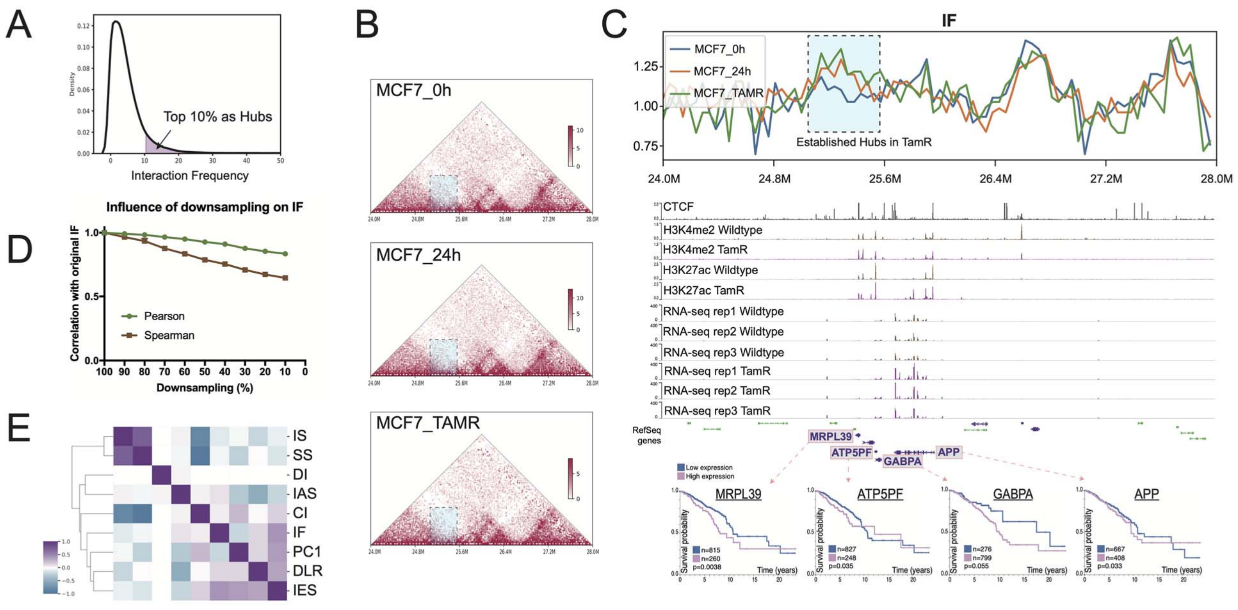

Figure 2. Novel metric IF. (**A**) Distribution of IF and the definition of hubs. (**B**) Hi-C contact heatmap for MCF-7 cells. MCF7_0h: untreated cells; MCF7_24h: cells treated with estrogen for 24 h; MCF7_TAMR: tamoxifen-resistant MCF-7 cells. (**C**) IF values for MCF-7 cells treated as noted for panel C. The middle panel shows the ChIP-seq and RNA-seq signals. The bottom panel shows Kaplan–Meier curves for genes nearby the newly established hubs. (**D**) Correlation between the original IF value and the IF values of downsampled Hi-C dataset. (**E**) Correlation heatmap for genome-wide 1D metrics for the same Hi-C sample.

ratio) of IAS and of IES, respectively (Figure 3A, Methods). For instance, a negative IASC value reflects a decrease in intra-TAD interactions, whereas a positive IESC value reflects an increase in inter-TAD interactions, each of which was found to be an important phenomenon in a cohesin depletion analysis [22]. We also defined a change in IF (IFC) to evaluate the changes in significant interactions as described in Methods (Figure 3A). To test the validity of IASC, IESC and IFC, we compared those metrics with published tools (FIND [34] and diffHiC [35]) that detect differential chromatin interactions (DCI). Genomics loci with DCI are expected to have larger values (either positive or negative) for changes in 1D metrics. Indeed, loci with DCI exhibited statistically higher values for IASC, IESC and IFC compared with non-DCI (Figure 3B).

The correlation and distribution of these two-sample metrics are shown in Figure 3C and Supplementary Figure S3A, respectively. IESC was highly correlated with delta-DLR owing to the consistency between inter-TAD contacts and distal interactions. Not surprisingly, the changes in IS, SS and CI (ISC, SSC and CIC in Figure 3C) still had close relationships. CD shared partial common information with IFC for contact patterns at each locus. Overall, these metrics represent various aspects of changes in chromosome organization.

The novel 1D metric DRF identifies a directional change of inter-TAD interactions

Although the aforementioned metrics for two Hi-C samples are effective, we noticed that certain TADs (or domains) exhibited inconsistent changes on the 5′ (left) and 3′ (right) sides when comparing two Hi-C samples. For example, in the TAD region shown in Figure 4A, the number of contacts clearly increased (red) on the 5′ side but decreased (blue) on the 3′ side (same dataset for Figure 3A). Chromatin simulation analysis demonstrated that the candidate TAD region of the Rad21AID-treated sample had more interactions with the 5′ side and fewer interactions with the 3′ side (Figure 4B), compared with control. Considering this observation, we defined ‘directional TAD’ (dTAD) as TADs (or domains) with asymmetric changes in inter-TAD interactions and proposed the new 1D metric ‘directional relative frequency’ (DRF, see Methods) to identify them. We could classify all TADs as ‘5′-dTAD’, ‘3′-dTAD’ or ‘non-dTAD’ in various cell types (Supplementary Figure S3B and C). For example, all loci within the 5′-dTAD have negative DRF values, suggesting that a TAD become oriented to the 5′ side after treatment.

To investigate the new chromosome structure termed dTAD, we utilized the precision nuclear run-on sequencing data obtained from the same study (GSE104334) [22]. We tested whether inter-dTAD directionality correlated with gene expression. Indeed, expression of genes upstream of 5′-dTADs was significantly higher than that measured for genes upstream of non-dTADs, whereas expression of genes downstream of 5′-dTADs was significantly lower (Figure 4C). For the 3′-dTADs, these two trends were exactly opposite. In addition, genes within either 5′-TADs or 3′-dTADs were more highly expressed than those in non-dTADs. These results suggested that dTADs have distinct transcriptional profiles compared with other types of TADs, even in non-treated samples.

Next, we assessed changes in gene expression in regions nearby dTADs, since dTADs were defined based

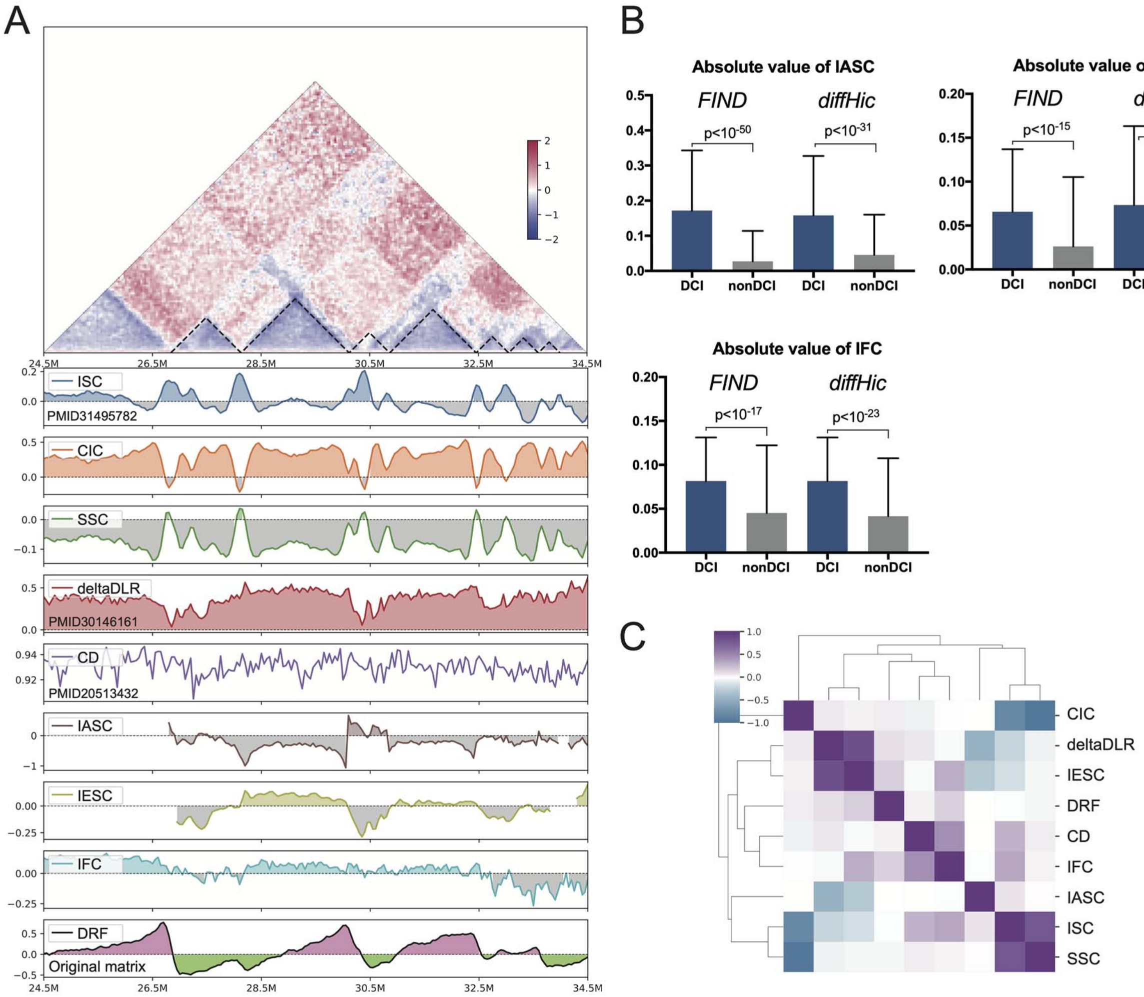

Figure 3. 1D metrics designed for two Hi-C samples. (**A**) Visualization of published and our original 1D metrics for comparison between RAD21AID-treated and non-treated Hi-C samples in HCT116 cells. The visualized region is the same as in Figure 1B. (**B**) Absolute values of IASC, IESC and IFC for DCI or non-DCI. DCI were computed using FIND or diffHic. (**C**) Correlation heatmap for genome-wide 1D metrics for the same comparison of Hi-C samples.

on the comparison of two Hi-C samples. Figure 4D shows the percentages of differentially expressed genes (false-discovery rate, FDR < 0.05) in response to Rad21-depletion for the indicated regions proximal to dTADs. Whereas genes upstream of 5′-dTADs tended to be upregulated compared with non-dTADs, the downstream genes tended to be relatively downregulated. A similar, yet opposite tendency was observed for regions upstream or downstream of 3′-dTADs. The genes within dTADs were more likely to be upregulated. In short, for dTADs, which represent asymmetric changes in inter-TAD interactions, the upstream- and downstream-proximal regions tended to exhibit opposite trends in gene expression (Figure 4E).

1D metrics are good indicators for evaluating similarity across samples

Previous studies have primarily used 1D metrics to identify particular chromatin structures. Here, however, we attempted to take advantage of the ‘one-dimensional’ nature of Hi-C metrics and expanded their usage. We first evaluated the performance of 1D metrics when dealing with multiple Hi-C samples. To estimate the reproducibility of 1D metrics under different conditions, we utilized three publicly available Hi-C datasets for three murine studies: Data1 (GSE93431) contains Hi-C data for the cohesin loader Nipbl knockout (two replicates) and non-treated hepatocytes (four replicates) [25]; Data2 (GSE96107) contains Hi-C data for non-treated ES cells (seven replicates), NPC (four replicates) and CN (six replicates) [36] and Data3 (GSE98671) contains Hi-C data for non-treated ES cells (two replicates), control ES cells (two replicates) and insulator factor Ctcf-depleted ES cells (two replicates) [37]. We applied 1D metrics to Hi-C samples from these three datasets and calculated the pairwise Pearson correlation coefficient among them, followed by hierarchical clustering. The 1D metrics could classify Hi-C samples into reasonable clusters (Figure 5A and B), whereas raw contact vectors generated poor clustering (Supplementary Figure S4A).

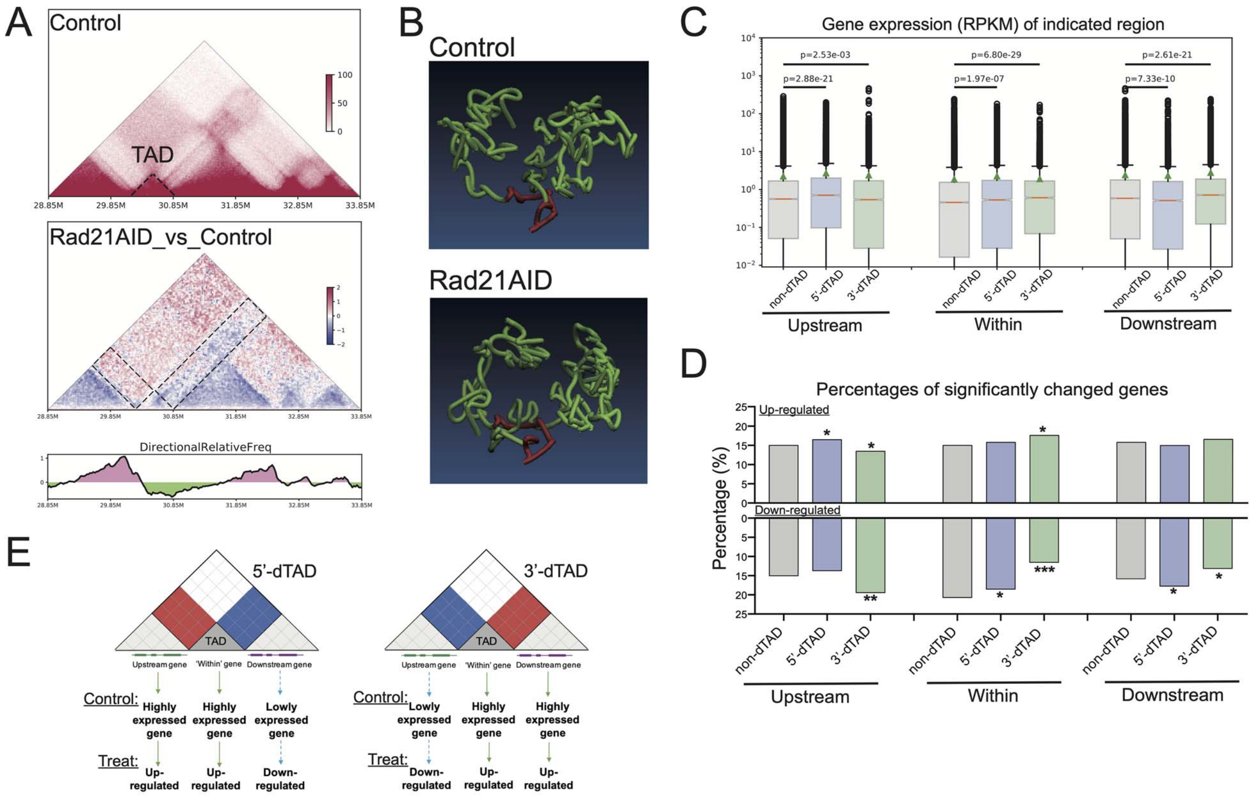

Figure 4. Novel 1D metric for a directional change in inter-TAD interactions. (**A**) Contact matrix for the non-treated sample and differential matrix for RAD21AID-treated versus non-treated samples. The corresponding DRF value was negative for the 5′-dTAD. The dashed rectangles denote increased (red) or decreased (blue) inter-TAD interactions. (**B**) Chromosome simulation of the same loci shown in panel C. Red color indicates the region of the 5′-dTAD in panel C. (**C**) Assessment of gene expression for each indicated region. Statistical comparison between dTADs (5′ or 3′) and non-dTADs was accomplished with the Mann–Whitney U test. (**D**) The percentage of significantly changed genes (FDR < 0.05) is shown for the indicated region. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$; Fisher's exact test. (**E**) Distinct transcription patterns on the two sides of dTADs. Highly expressed genes were identified based on a comparison with non-dTADs.

Using CI and IAS as examples, biological replicates for the same condition always gave the highest correlation. CI and IAS could also correctly group the same cell types of non-treated samples from different studies (ES cells of Data2 and Data3), whereas the treated samples (depletion of Nipbl or Ctcf, each an essential factor for chromatin structure) were separated into different clusters. IAS produced even better clustering than CI, as use of IAS resulted in fewer breaks in terms of studies and cell types (Figure 5B, upper panel). These results suggested that the relative contact frequency within TADs—as indicated by IAS—is a better indicator of sample similarity than the boundary information alone. In addition, ES cells can be grouped separately from NPC and CN, whereas fewer differences exist between NPC and CN, possibly owing to dynamic changes in chromatin structure that occur during cell differentiation that yields NPC and CN [36].

To further evaluate the clustering performance of 1D metrics, we carried out the same clustering analysis but applied SCC (stratum-adjusted correlation coefficient [38]), a designed score (not 1D metric) for measuring similarities between Hi-C interaction matrices. Unexpectedly, the clustering results obtained with SCC were less biologically appropriate than results obtained with CI or IAS (Supplementary Figure S4B). We also defined a reproducible score (see Methods) to further determine whether 1D metrics can capture similarities and dissimilarities among samples. We evaluated the reproducibility score for Hi-C datasets from five studies (Figures 5C and Supplementary Figure S4C). Although clustering performance varied among the various 1D metrics because they are not specifically designed to evaluate sample similarity, some of them—especially IAS and IES—performed better than SCC. This result suggested that 1D metrics are also suitable for comparing multiple Hi-C samples, thus enabling fast and reliable estimation of the reproducibility of Hi-C samples among replicates and under different conditions across studies.

In addition, we applied the metrics CI and PC1 as examples for the application of 1D metrics to multiple Hi-C samples. By definition, CI values can be directly compared, and therefore CI values of multiple Hi-C samples can be visualized as a heatmap (Figure 5D). We evaluated the variation in CI among samples (analysis of variance, ANOVA-like test) and could identify statistically significant changes ($P < 10^{-14}$) in TAD boundaries after Rad21 depletion. Because PC1 is not suitable for direct comparisons, we converted PC1 into discrete values (i.e. compartment A or B) and constructed a plot using those converted data (Figure 5E). The comparison clearly showed

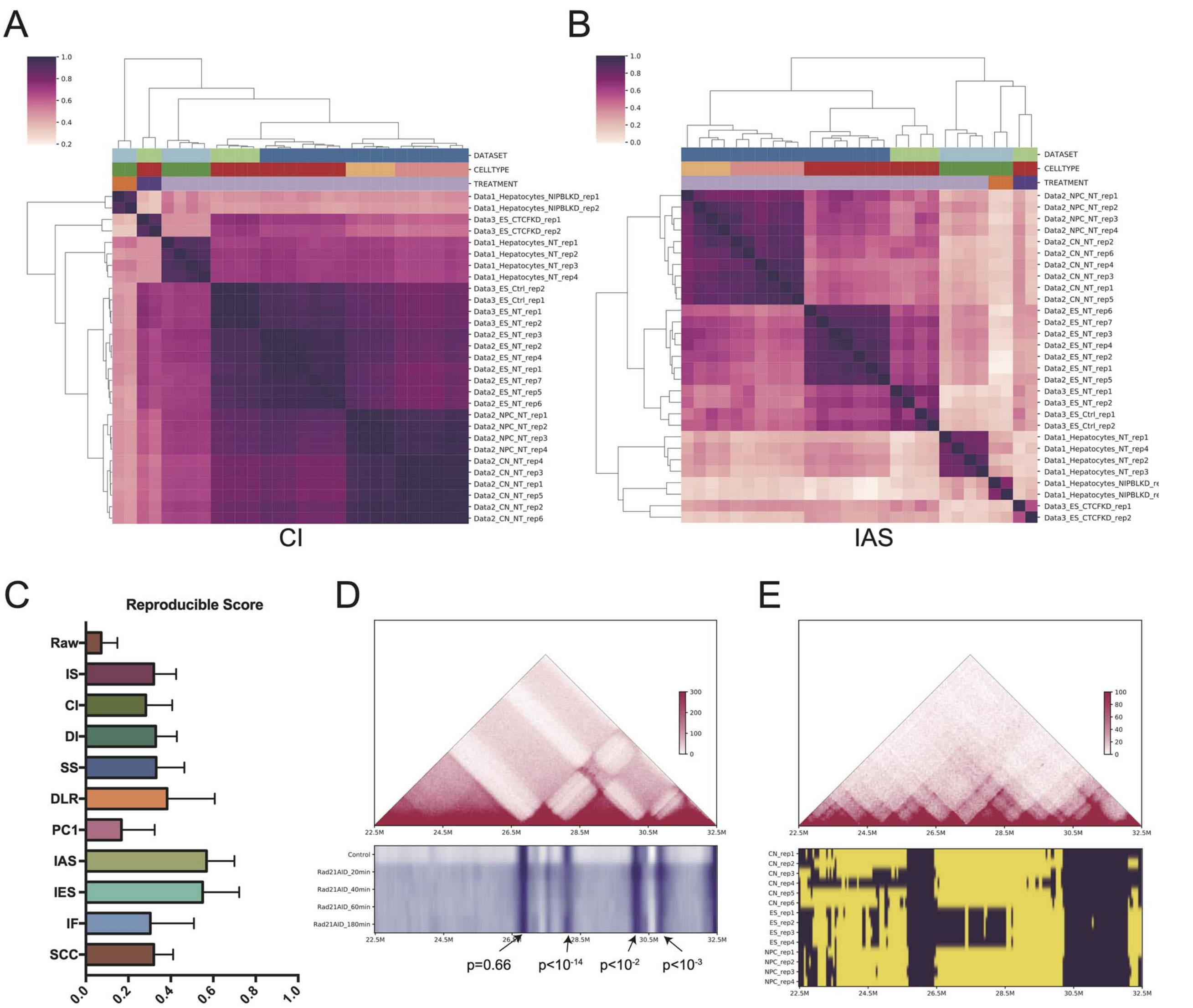

Figure 5. Application of 1D metrics to multiple samples. (**A**, **B**) Pairwise Pearson correlation heatmap followed by hierarchical clustering of CI and IAS for various Hi-C samples. The top panel shows the different datasets, cell types, and treatments for each sample. Data1: Nipbl knockout and non-treated hepatocytes; Data2: non-treated ES cells, NPC and CN, and Data3: non-treated ES cells, control ES cells, and Ctcf-depleted ES cells. (**C**) Reproducibility scores obtained by different 1D metrics and SCC for five different Hi-C datasets. (**D**) Heatmap based on CI for multiple samples in HCT116 cells. The *P*-value for TAD boundaries is indicated (ANOVA-like test). (**E**) Example of PC1 for multiple samples for mouse cells. The yellow and purple colors indicate compartments A and B, respectively.

that a switch from compartment B to A occurred from ES cells to CN and NPC. All these visualization and statistical analyses are included in HiC1Dmetrics.

Chromatin state annotation using ChromHMM with 1D metrics information

Because 1D metrics are one-dimensional signals along genomic loci, we wondered whether they could be used to segment and annotate entire genomes into distinct ‘chromatin states’, as is commonly done in epigenetics studies [39]. Here we used ChromHMM, the most popular tool for chromatin-state annotation, which explicitly models the observed combination of chromatin marks based on a multivariate hidden Markov model [21]. We used Hi-C (GSE99541, GSE64525) and ChIP-seq (CTCF, RNA Pol2, H3K27ac, H3K27me3, H3K4me1, H3K4me3 and Input) data for MCF-7 cells [40] and human mammary epithelial cells (HMEC) [41] (GSE23701, GSE57498). We trained ChromHMM in two ways: only ChIP-seq markers (conventional model), and ChIP-seq markers with three 1D metrics (IS, PC1 and IF; with-1D model; see Methods).

Using the conventional model, we annotated all genomes into nine commonly depicted states. The addition of 1D metrics to ChromHMM enabled the identification of more detailed chromatin states based on both ChIP-seq and Hi-C information for both MCF-7 and HMEC (Figure 6A, Supplementary Figure S5A). First, the with-1D model could separate the annotation of active promoters into compartment A or B. We observed that compartment A and even B (closed, gene-poor region with less transcription) could be classified into several chromatin states based on epigenomic patterns, making it easier to focus on specific genomic structures. Second, the with-1D model could also identify a state termed 'active promoters on boundaries'. It has been reported that genes at TAD boundaries are transcribed at significantly higher levels [42]. Ramirez *et al.* [24] also reported that boundaries with promoters are different from non-promoter boundaries. Third, as described above, hubs have been suggested to be physically associated with the gene regulatory machinery and involved in disease progression [20]. The with-1D model could identify strong enhancers with hubs. We compared the conventional model with the with-1D model and observed a high correlation between corresponding chromatin states (Figure 6B, Supplementary Figure S5B), suggesting that the inclusion of 1D metrics did not substantively affect the nine chromatin states in the conventional model. Moreover, three new states identified using the with-1D model were just derived from the 'Active Promoter' and 'Strong Enhancer' states in the conventional model (Figure 6B, Supplementary Figure S5B), indicative of the importance of structural information when investigating these active states.

To show the effectiveness of chromatin annotation by the with-1D model, we visualized the three new states with the raw Hi-C 1D metrics and ChIP-seq signals for MCF-7 cells and HMEC (Figure 6C and D). These two cell lines differed in their repertoires of states derived from Hi-C contact frequency and ChIP-seq signals. We focused on 'Strong Enhancer on Hubs' (SEH) to investigate whether the difference was biologically meaningful. Notably, genes near SEH (red dashed rectangle) in MCF-7 or HMEC had distinct profiles for breast cancer. PDP1 and TMEM67, which were identified based on the SEH state for the breast-cancer line MCF-7, are overexpressed in breast cancer (Supplementary Figure S6A, data from Oncomine), whereas NECAB1 and SLC26A7, identified based on the SEH state for normal HMEC, are downregulated in breast cancer. These cell-specific transcriptions were also shown by RNA-seq tracks (Figure 6C and D, lower panel). We also assessed the overall survival of patients with abnormal expressions in these genes. Kaplan–Meier curves revealed that overexpression of PDP1 and TMEM67 was significantly associated with poorer overall survival, whereas elevated expression of NECAB1 and SLC26A7 was significantly associated with better overall survival (Supplementary Figure S6B). Indeed, PDP1 and TMEM67 have been reported to act as oncogenes [43, 44], whereas NECAB1 and SLC26A7 are putative tumor-suppressor genes [45, 46]. Considering the roles of hubs in the gene regulatory machinery and disease [20], it is reasonable to conclude that SEH information can distinguish cancer cells from non-cancer cells. Therefore, the combination of Hi-C (1D metrics) and ChIP-seq (signal) in ChromHMM can recognize chromatin patterns with critical functions.

Discussion

The 3D folding of chromosomes in eukaryotic genomes is key to understanding DNA-dependent processes [47]. Recent advances in next-generation sequencing and Hi-C technology have enabled us to glimpse chromatin contacts in a genome-wide manner. Hi-C datasets are often large, and it is difficult to compare multiple samples. To address this challenge, several studies have used the 1D metrics of Hi-C to derive compressed information for desired chromosome structures [2, 3, 13–16]. Here, we focused on current deficiencies of 1D metrics areas and made improvements in three ways. First, we introduced new 1D metrics, including IAS, IF and DRF, to help elucidate newly identified chromosome structures. We demonstrated the biological relevance of these new metrics. Second, we expanded the utility of 1D metrics, emphasizing that 1D metrics offer a promising means of measuring Hi-C reproducibility and comparing multiple samples. 1D metrics can also be included in ChromHMM to achieve more detailed annotations. Third, we provide a framework, called HiC1Dmetrics, for calculating and analyzing various types of 1D metrics, including both published metrics and our newly developed metrics, for multiple samples. To facilitate the development of 1D metrics, HiC1Dmetrics offers both command-line and web-based interfaces.

With the Hi-C technique [48], compartments A/B and TADs have been successively revealed based on a low- and enhanced-resolution map [2, 3], respectively. Well-designed 1D metrics have proved to be powerful tools for defining those structures. More recent research has shown that additional specific features at the TAD level could be observed, such as loops [49] and stripes [19]. From the view of 1D metrics, we propose the first linear score (i.e. IAS) that can simultaneously identify loops and stripes of each TAD, as they can be reflected by unique linear patterns along the genome. Although we did not consider nested TADs for the computation, our IAS provide a rapid and intuitive way to not only classify various TAD-level structures but also evaluate the relative levels of interactions. It is important to note that IAS is dependent on the used TAD list. The influence of TAD segmentations should be carefully considered when comparing multiple samples. First, the same TAD calling method should be applied to minimize the variability of the TAD calling algorithm. Second, in the case of comparing intra-TAD interactions, for example, cell lines under different conditions (such as Figure 1C), the same TAD segmentations (i.e. the one of control) could be used. Third, in the case of comparing stripes among samples that might have distinct TADs (such as Figure 1H), the users should also consider the variability from different TAD segmentations. Our other efforts focused on genomic loci anchored by enriched interactions [20], which can be represented by our 1D metric IF to calculate the relative frequency of interactions at each locus. When comparing two Hi-C samples, 1D metrics also can describe changes in genome folding [16, 17, 33]. In particular, we describe a new structure, namely dTAD, which reflects asymmetric changes in inter-TAD contacts, based on

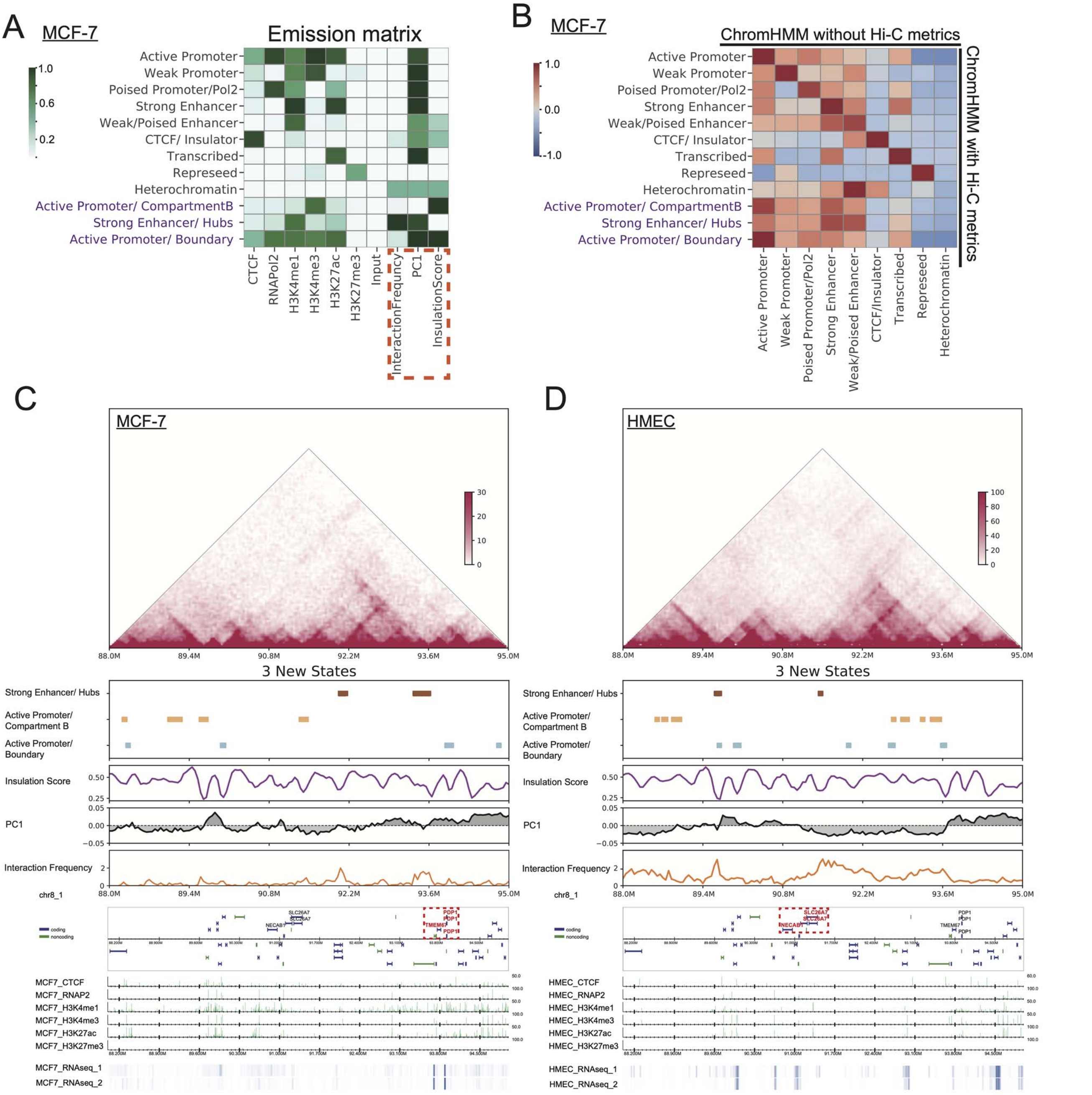

Figure 6. Incorporation of 1D metrics into ChromHMM. (**A**) Emission matrix for the ChromHMM model trained with seven ChIP-seq signals and three 1D metrics for MCF-7 cells. (**B**) Correlation between the conventional model trained with only ChIP-seq and the new model with information from 1D metrics. The result of MCF-7 cells is presented. (**C**, **D**) The candidate region for MCF-7 cells or HMEC. Selected states (bar) and 1D metrics as well as ChIP-seq signals are also illustrated. The lower panel indicates the RNA-seq signals. The red dashed box indicates the genes PDP1 and TMEM57 as well as SLC26A7 and NECAB1.

a new 1D metric, DRF. We showed that dTADs are accompanied by distinct transcriptional events. Considering the continuing increase in our knowledge of asymmetric chromosome architectures [19, 29, 50, 51], a possible explanation for dTADs is that a particular biological perturbation may differentially impact the two sides of a TAD. Regarding the elevated level of transcription of dTADs in control samples, dTADs may be more involved in transcription-driven genome organization. Owing to the hierarchies of TADs, it is also possible that dTADs are nested in other 'meta-TADs' or that dTADs are actually 'sub-TADs' of some larger TADs, which may also explain the inconsistency between the two sides of a dTAD. Whereas dTADs were found among different cells and conditions (Supplementary Figure S3C), the biological mechanism of dTADs remains to be elucidated. Our newly defined 1D metric DRF and the newly defined structure dTAD will contribute to future research on asymmetric changes in inter-TAD interactions.

Although previous studies have used 1D metrics merely to identify particular chromosome structures [3, 13–15], we propose that 1D metrics themselves can be directly compared among samples to provide important architectural information for each locus. Our results show that 1D metrics can reliably determine similarities between replicates and dissimilarities between conditions, which can also be utilized to evaluate Hi-C reproducibility. Because 1D metrics reflect certain chromosomal attributes, different metrics are relevant to different aspects of similarities between Hi-C samples. Moreover, by utilizing the 'one-dimension' trait, we propose that 1D metrics could be incorporated into ChromHMM to obtain annotations with information of chromosome organization. Although much effort has gone into integrally analyzing Hi-C and ChIP-seq data [52, 53], 1D metrics provide an opportunity for new and promising strategies. Therefore, our study greatly expands the potential utility of 1D metrics, yet there is abundant room for further progress.

When raw contact matrices of Hi-C contain chromatin interactions for every possible pair of genomic loci, 1D metrics merely represent partial information of Hi-C. Therefore, 1D metrics also have limitations. For example, 1D metrics only provide information pertaining to particular aspects of chromatin contacts, whereas other information can be lost; this shortcoming could possibly be rectified by incorporating several 1D metrics simultaneously [18]. Still, it must be noted that, fundamentally, 1D metrics are computational scores, the values of which might depend on the algorithm or parameters used. This could potentially produce misleading results. For instance, the calculation of IS under different window sizes yields different outcomes [54]. In addition, the calculation of 1D metrics requires the segmentation of a genome into bins with certain resolutions (e.g. 10 kb), leading to difficulties in determining exact matches of the coordinates of 1D metrics with other linear tracks. Despite these shortcomings, the conversion from 2D contact matrices to 1D metrics offers a fast and easy way to extract and disseminate detailed information for particular types of chromosome structures. The improvement of Hi-C resolution [5] and the increasing data size also raise the necessity of using 1D metrics and our HiC1Dmetrics. Table 1 summarizes various scenarios that are applicable to 1D metrics and 2D matrices. Although other information is inevitably lost from 2D matrices, 1D metrics can be effective when the primary focus is on identifying specific chromatin structures (e.g. stripes), comparing among multiple Hi-C samples, or integrating Hi-C with other 1D tracks such as ChIP-seq. With the explosive growth of Hi-C research, further efforts to design new 1D metrics and expand their usage will still be required for ongoing investigations into chromatin organization.

In summary, we present a framework, named HiC1-Dmetrics, to calculate and analyze 1D metrics for Hi-C samples. We improved certain aspects of 1D metrics by introducing our original metrics and expanding the usage of 1D metrics. Our study will facilitate research concerning the 3D genome.

Materials and Methods

Calculation of one-sample metrics

As described in [13], IS was calculated for each bin i as the sum of interactions that occur across that bin, which can be visualized by sliding a l bins $\times l$ bins square along the matrix diagonal:

$$\mathrm{IS}_{\mathrm{i}} = \sum_{j=1}^{l}\sum_{k=1}^{l} \mathrm{C}_{\mathrm{i}+j,\mathrm{i}-k}$$

where C is the contact number of the indicated bin and l is the bin distance from bin i. The obtained score was then normalized by $\mathbf{log}\left(\frac{\mathrm{IS}_{\mathrm{i}}}{\mathrm{IS}_{\mathrm{avg}}}\right)$, where $\mathbf{IS}_{\mathbf{avg}}$ is the average insulation of a given chromosome. For DI [3]:

$$\mathbf{DI}_{\mathrm{i}} = \frac{I-H}{|I-H|}\left(\frac{(H-E)^2}{E} + \frac{(I-E)^2}{E}\right)$$

where $H = \sum_{j=1}^{l} \mathrm{C}_{\mathrm{i},\mathrm{i}-j}$, $I = \sum_{j=1}^{l} \mathrm{C}_{\mathrm{i},\mathrm{i}+j}$, and the expected number of reads $E = (A+B)/2$. The concept of SS is described in [14] as a ratio:

$$\mathbf{SS}_{\mathrm{i}} = \frac{D}{\mathbf{min}\,(A+D, B+D)}$$

and the CI is calculated according to [15] and can be represented by:

$$CI_{\mathrm{i}} = \frac{A+B}{D},$$

where $D = \sum_{j=1,k=1}^{l} \mathrm{C}_{\mathrm{i}+j,\mathrm{i}-k}$; $A = \left(\sum_{j=1,k=1}^{l} \mathrm{C}_{\mathrm{i}-j,\mathrm{i}-k}\right)/2$; $B = \left(\sum_{j=1,k=1}^{l} \mathrm{C}_{\mathrm{i}+j,\mathrm{i}+k}\right)/2$. The DLR was used in [16] and can be formulated by:

$$\mathbf{DLR}_{i} = \mathbf{log}\left(\frac{\sum_{j=1}^{l} \mathrm{C}_{i,i-j} + \sum_{j=1}^{l} \mathrm{C}_{i,i+j}}{\sum_{j=1}^{i-l} \mathrm{C}_{i,j} + \sum_{j=i+1}^{L} \mathrm{C}_{i,j}}\right),$$

where l is the bin distance from bin i and L is the bin number of the indicated chromosome.

Compartment PC1 was established as described [2]. First, for each possible distance d between two bins, the expected matrix was calculated by $\mathbf{Expect}_{i,j} = \frac{\sum \mathbf{Observed}_{i,j}}{\mathbf{len}\left(\mathbf{Observed}_{i,j}\right)}$, where $|i-j| = d$. The expected matrix was smoothed as indicated in [49]. Second, the observed matrix was divided by the expected matrix to obtain an observed/expected matrix. Then, pairwise Pearson correlation coefficients were calculated for each row to generate a correlation matrix, on which PCA is used. The eigenvector of the first component is selected as PC1. Because the sign of the eigenvector is arbitrary, we modified it with a gene density file to ensure that regions of high gene density were assigned as compartment A.

For our metrics IAS and IES, we first called the contact domain using our custom TAD-calling method. Then, the distance-normalized matrix E was generated according to [49]. For each bin within a TAD from bin s to e, IAS and IES could be calculated (Figure 1A) by the following formulas:

$$\mathrm{IAS}_i = \left(\sum\nolimits_{j=1}^{i-s} \frac{C_{i,i-j}}{E_{i,i-j}} + \sum\nolimits_{k=1}^{e-i} \frac{C_{i,i+k}}{E_{i,i+k}}\right) / \left(e - s\right); i \in \mathrm{TAD}_{s,e}$$

$$\mathrm{IES}_i = \left(\sum\nolimits_{j=1}^{s} \frac{C_{i,j}}{E_{i,j}} + \sum\nolimits_{k=e}^{L} \frac{C_{i,k}}{E_{i,k}}\right) / \left(L - e + s\right); i \in \mathrm{TAD}_{s,e},$$

where L is the bin number of the indicated chromosome. For the metric IF, we used FitHiC2 [55] to obtain statistically significant interactions with threshold FDR < 0.05 (or custom threshold by users). The number of significant interactions (N_{SI}) of each bin was assigned by coverage command of Bedtools [56]. The obtained value was normalized by:

$$\mathrm{IF} = \frac{\log N_{\mathrm{SI}}}{\mathrm{mean}_{\mathrm{nonzero}}\left(\log N_{\mathrm{SI}}\right)},$$

where $\log N_{\mathrm{SI}}$ is the log value of the number of significant interactions.

Calculation of two-sample metrics

The ISC is described in [17] to represent the difference of IS: $\mathrm{ISC} = \mathrm{IS}_{\mathrm{treat}} - \mathrm{IS}_{\mathrm{control}}$. Delta-DLR is described in [16] as: $\mathrm{DLR} = \mathrm{DLR}_{\mathrm{treat}} - \mathrm{DLR}_{\mathrm{control}}$. CD is described in Homer software [33] to correlate the interaction profile of a locus in one Hi-C sample to the same locus in another sample. So, for each bin: $CorrD_i = \mathrm{Pearson}\left(C_i^{\mathrm{treat}}, C_i^{\mathrm{control}}\right)$. The changes in each of CI, SS and IF were calculated by subtracting the score for the control sample from the score of the treated sample. The changes in IAS and IES were achieved by $\log\left(\mathrm{treat}/\mathrm{control}\right)$.

To generate the differential matrix for DRF, we scaled the Knight–Ruiz normalized Hi-C contact matrices to the total number of reads. The differential matrix was: $T = \log\left(C_{\mathrm{treat}}\right) - \log\left(C_{\mathrm{control}}\right)$. Then the DRF, which represents the inconsistency of relative contacts between the left side (3′) and right side (5′), could be formulated by:

$$DRF_i = \sum\nolimits_{j=a}^{b} T_{i,i+j} - \sum\nolimits_{j=a}^{b} T_{i,i-j},$$

where a and b represent the bin distance from bin i.

Hi-C data processing

The Hi-C data for human and mouse were aligned to hg38 and mm10 reference genomes, respectively. Only reads with a high mapping quality (MAPQ $\geq$30) were retained to generate a so-called '.hic' file, which is described in Juicer software [31]. The intra-chromosomal contact matrices were extracted from the .hic file with Knight-Ruiz normalization. For TAD calling, IS was calculated along a chromosome as described in [13]. The local minima of normalized IS indicated potential boundaries. Only loci for which IS was <10% were considered. Then, a delta vector of IS was calculated for each bin to extract only those boundaries for which the 'strength' was greater than a threshold value, as described by Crane *et al.* [13]. The TADs generated with this custom method overlapped substantially with TADs discovered by Juicer software (Supplementary Figure S2A). Chromosome simulation was accomplished with PHi-C [57] and visualized with VMD software. The reproducible score was defined as $\max\left(P_{\mathrm{replicate}}\right) - \min\left(P_{\mathrm{nonreplicate}}\right)$, where P is the Pearson correlation coefficient between two samples.

ChIP-seq and gene expression analysis

ChIP-seq data for MCF-7 cells and HMEC were downloaded from GEO (Supplementary Table S1). The reads were aligned to the hg38 reference genome using Bowtie [58] with '-n2 -m1' parameters. We used MACS2 [59] to call peaks and DROMPA3 [60] for visualization, where reads were normalized to total read number. Other ChIP-seq data were visualized in the WashU Epigenome Brower with the internal public data hubs. Gene expression analysis of Rad21AID-treated and non-treated HCT116 cells was carried out using the precision nuclear run-on sequencing of GSE104334. Differential expression analysis was achieved by DESeq2 with internal normalization.

ChromHMM with 1D metrics

For conventional ChromHMM, we used seven ChIP-seq tracks (binary 1/0 represent peaks or not) including CTCF, RNA Pol2, H3K4me1, H3K4me3, H3K27ac, H3K27me3 and Input. For ChromHMM with 1D metrics, we used three additional tracks: IS was binarized to boundary/non-boundary; PC1 was binarized to compartment A or B; IF was binarized to hubs or non-hubs. So, the observed sequence was $y = \left(y_1, y_2, \cdots y_n\right)^T$, where $y_n = \left(y_{n,1}, y_{n,2}, \cdots, y_{n,t}\right)$. The binary value of track n at chromatin locus t was $y_{n,t} \in (0, 1)$. We applied ChromHMM [14] software to solve this discrete multivariate hidden Markov model problem, which assumes the observed sequences are independent Bernoulli random variables:

$$P\left(y_t \middle| z_t\right) = \prod\nolimits_{n=1}^{N} P\left(y_{t,n} | z_{t,n}\right),$$

where z_t is the hidden state of locus t. We trained ChromHMM to partition all genomes into 15 chromatin states (200-bp resolution) and collapsed the redundant states into 9 or 12 distinct states. The merged states were manually annotated by comparison with the published ChromHMM [21, 61]. For the correlation between emission matrix of conventional ChromHMM and ChromHMM with 1D metrics, only seven ChIP-seq tracks were considered. Gene expression data for cancer and normal tissues were collected from the Oncomine database. Data for survival curves were collected from the Human Protein Atlas database.

Extracting structural information by HiC1Dmetrics

HiC1Dmetrics provides various functions to identify chromosomal structures. For the stripe calling, after extracting the local maximum positions of IAS, only positions with IAS > IASmean is retained. Then, similar to [13], we calculate a delta vector of IAS for each bin to only extract strong IAS peaks. To avoid clustered small peaks, the IAS value of a 'stripe' position should be higher than any position around 100 kb. For the chromatin hubs, genomic loci in the top 10% of IF values are extracted and merged as the hub region.

For the classification of TAD, the enrichment of IAS was calculated by the Z-test to compare the IAS on a TAD corner to the background model, which was obtained by randomly sampling all IAS values. The corner was defined by the left/right 100 kb of a TAD. A TAD with either left or right enrichment (FDR < 0.05, Bonferroni) was classified as a left/right stripe-TAD, respectively, whereas a TAD with enrichment of both the left corner and right corner was classified as a loop-TAD. Then the other TADs were classified as 'non-stripe TAD'.

For the classification of directional TAD, a TAD with all negative values for DRF was identified as a 5′ directional TAD (5′-dTAD). In contrast, a TAD with all positive values for DRF was identified as a 3′ directional TAD (3′-dTAD). The default parameters for DRF calculation were a = 500 000 and b= 2 000 000.

Key Points

- We review published 1D metrics and present new 1D metrics. We present a framework, named HiC1Dmetrics, to calculate and analyze various 1D metrics for Hi-C samples.
- We propose the novel 1D metrics IAS and IF to describe the recently proposed structures stripe and chromatin hubs, respectively. We propose the original 1D metric DRF to introduce dTAD, a novel asymmetric event of inter-TAD interactions. We demonstrated the biological relevance of our new metrics.
- Our results highlight that the 1D metrics-based approach is reproducible and robust for comparison and visualization of multiple Hi-C samples.
- We show that the linear tracks of 1D metrics can be combined with other epigenome data to annotate chromatin states in greater details.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Authors' contributions

J.W. prepared scripts and performed the bioinformatics analysis. R.N. and J.W. conceived the project. J.W. and R.N. drafted the manuscript.

Software and data availability

HiC1Dmetrics (https://github.com/wangjk321/HiC1Dmetrics) was written with Python3 and released in https://pypi.org/project/h1d/. Detailed requirements and usage are introduced in https://h1d.readthedocs.io/. HiC1-Dmetrics uses acceptable memory (~ gigabytes) on high-resolution (i.e. ≥5 kb) Hi-C, but the fine-resolution data (e.g. <1 kb) may use much more memory. We also provide a Streamlit-based web application (http://hic1d.herokuapp.com) that is easier and more convenient for analyzing 1D metrics from Hi-C samples. The datasets used in this study are listed in Supplementary Table S1 and available at the Gene Expression Omnibus.

Funding

This work was supported by Japan Society for the Promotion of Science KAKENHI Grants-in-aid for Scientific Research (grant number 17H06331) and the Japan Agency for Medical Research and Development (grant number JP21gm6310012h0002).

Abbreviations

IS: insulation score; CI: contrast index; DI: directionality index; SS: separation score; DLR: distal-to-local ratio; PC1: principal component 1; IAS: intra-TAD score; IES: inter-TAD score; IF: adjusted interaction frequency; DRF: directional relative frequency; GEO: Gene Expression Omnibus; TAD: topologically associating domain; dTAD: directional TAD; ES: embryonic stem cells; NPC: neural progenitors; CN: cortical neuron; SCC: stratum-adjusted correlation coefficient; SEH: strong enhancer on hubs.

References

1. Mota-Gomez I, Lupianez DG. A (3D-nuclear) space odyssey: making sense of Hi-C maps. *Genes (Basel)* 2019;**10**:415.
2. Lieberman-Aiden E, van Berkum NL, Williams L, *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 2009;**326**:289–93.
3. Dixon JR, Selvaraj S, Yue F, *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 2012;**485**:376–80.
4. Eagen KP. Principles of chromosome architecture revealed by Hi-C. *Trends Biochem Sci* 2018;**43**:469–78.
5. Wang Q, Sun Q, Czajkowsky DM, *et al.* Sub-kb Hi-C in D. melanogaster reveals conserved characteristics of TADs between insect and mammalian cells. *Nat Commun* 2018;**9**:188.
6. Krietenstein N, Abraham S, Venev SV, *et al.* Ultrastructural details of mammalian chromosome architecture. *Mol Cell* 2020;**78**:554–565.e7.
7. Forcato M, Nicoletti C, Pal K, *et al.* Comparison of computational methods for Hi-C data analysis. *Nat Methods* 2017;**14**:679–85.
8. Zhang Y, An L, Xu J, *et al.* Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus. *Nat Commun* 2018;**9**:750.
9. Cleveland WS, Mcgill R. Graphical perception – the visual decoding of quantitative information on graphical displays of data. *J R Stat Soc Ser A* 1987;**150**:192–229.

10. Nuebler J, Fudenberg G, Imakaev M, *et al.* Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proc Natl Acad Sci U S A* 2018;**115**:E6697–706.
11. Liu Y, Nanni L, Sungalee S, *et al.* Systematic inference and comparison of multi-scale chromatin sub-compartments connects spatial organization to cell phenotypes. *Nat Commun* 2021;**12**:2439.
12. Zufferey M, Tavernari D, Oricchio E, *et al.* Comparison of computational methods for the identification of topologically associating domains. *Genome Biol* 2018;**19**:217.
13. Crane E, Bian Q, McCord RP, *et al.* Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* 2015;**523**:240–4.
14. Ramirez F, Lingg T, Toscano S, *et al.* High-affinity sites form an interaction network to facilitate spreading of the MSL complex across the X chromosome in drosophila. *Mol Cell* 2015;**60**:146–62.
15. Van Bortle K, Nichols MH, Li L, *et al.* Insulator function and topological domain border strength scale with architectural protein occupancy. *Genome Biol* 2014;**15**:R82.
16. Heinz S, Texari L, Hayes MGB, *et al.* Transcription elongation can affect genome 3D structure. *Cell* 2018;**174**:1522–1536.e22.
17. Viny AD, Bowman RL, Liu Y, *et al.* Cohesin members Stag1 and Stag2 display distinct roles in chromatin accessibility and topological control of HSC self-renewal and differentiation. *Cell Stem Cell* 2019;**25**:682–696.e8.
18. Ing-Simmons E, Vaquerizas JM. Visualising three-dimensional genome organisation in two dimensions. *Development* 2019; **146**:dev177162.
19. Vian L, Pekowska A, Rao SSP, *et al.* The energetics and physiological impact of Cohesin extrusion. *Cell* 2018;**173**:1165–1178.e20.
20. Huang J, Marco E, Pinello L, *et al.* Predicting chromatin organization using histone marks. *Genome Biol* 2015;**16**:162.
21. Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc* 2017;**12**:2478–92.
22. Rao SSP, Huang SC, Glenn St Hilaire B, *et al.* Cohesin loss eliminates all loop domains. *Cell* 2017;**171**:305–320.e24.
23. Lajoie BR, Dekker J, Kaplan N. The Hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods* 2015;**72**:65–75.
24. Ramirez F, Bhardwaj V, Arrigoni L, *et al.* High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat Commun* 2018;**9**:189.
25. Schwarzer W, Abdennur N, Goloborodko A, *et al.* Two independent modes of chromatin organization revealed by cohesin removal. *Nature* 2017;**551**:51–6.
26. Stansfield JC, Tran D, Nguyen T, *et al.* R tutorial: detection of differentially interacting chromatin regions from multiple Hi-C datasets. *Curr Protoc Bioinformatics* 2019;**66**:e76.
27. Ron G, Globerson Y, Moran D, *et al.* Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains. *Nat Commun* 2017;**8**:2237.
28. Matthews BJ, Waxman DJ. Computational prediction of CTCF/cohesin-based intra-TAD loops that insulate chromatin contacts and gene expression in mouse liver. *Elife* 2018;**7**:e34077.
29. Barrington C, Georgopoulou D, Pezic D, *et al.* Enhancer accessibility and CTCF occupancy underlie asymmetric TAD architecture and cell type specific genome topology. *Nat Commun* 2019;**10**:2908.
30. Matthey-Doret C, Baudry L, Breuer A, *et al.* Computer vision for pattern detection in chromosome contact maps. *Nat Commun* 2020;**11**:5795.
31. Durand NC, Shamim MS, Machol I, *et al.* Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst* 2016;**3**:95–8.
32. Zhou Y, Gerrard DL, Wang J, *et al.* Temporal dynamic reorganization of 3D chromatin architecture in hormone-induced breast cancer and endocrine resistance. *Nat Commun* 2019;**10**:1522.
33. Heinz S, Benner C, Spann N, *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 2010;**38**:576–89.
34. Djekidel MN, Chen Y, Zhang MQ. FIND: difFerential chromatin INteractions detection using a spatial Poisson process. *Genome Res* 2018;**28**:412–22.
35. Lun AT, Smyth GK. diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics* 2015;**16**:258.
36. Bonev B, Mendelson Cohen N, Szabo Q, *et al.* Multiscale 3D genome rewiring during mouse neural development. *Cell* 2017;**171**:557–572 e524.
37. Nora EP, Goloborodko A, Valton AL, *et al.* Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell* 2017;**169**:930–944.e22.
38. Yang T, Zhang F, Yardimci GG, *et al.* HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Res* 2017;**27**:1939–49.
39. Nakato R, Sakata T. Methods for ChIP-seq analysis: a practical workflow and advanced applications. *Methods* 2021;**187**:44–53.
40. Kong SL, Li G, Loh SL, *et al.* Cellular reprogramming by the conjoint action of ERalpha, FOXA1, and GATA3 to a ligand-inducible growth state. *Mol Syst Biol* 2011;**7**:526.
41. Taberlay PC, Statham AL, Kelly TK, *et al.* Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer. *Genome Res* 2014;**24**:1421–32.
42. Ulianov SV, Khrameeva EE, Gavrilov AA, *et al.* Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. *Genome Res* 2016;**26**: 70–84.
43. Fan J, Shan C, Kang HB, *et al.* Tyr phosphorylation of PDP1 toggles recruitment between ACAT1 and SIRT3 to regulate the pyruvate dehydrogenase complex. *Mol Cell* 2014;**53**:534–48.
44. He J, McLaughlin RP, van der Beek L, *et al.* Integrative analysis of genomic amplification-dependent expression and loss-of-function screen identifies ASAP1 as a driver gene in triple-negative breast cancer progression. *Oncogene* 2020;**39**:4118–31.
45. Nakaoka HJ, Hara T, Yoshino S, *et al.* NECAB3 promotes activation of hypoxia-inducible factor-1 during Normoxia and enhances tumourigenicity of cancer cells. *Sci Rep* 2016;**6**: 22784.
46. Gorbatenko A, Olesen CW, Boedtkjer E, *et al.* Regulation and roles of bicarbonate transporters in cancer. *Front Physiol* 2014;**5**:130.
47. Szabo Q, Bantignies F, Cavalli G. Principles of genome folding into topologically associating domains. *Sci Adv* 2019;**5**:eaaw1668.
48. Bonev B, Cavalli G. Organization and function of the 3D genome. *Nat Rev Genet* 2016;**17**:661–78.
49. Rao SS, Huntley MH, Durand NC, *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 2014;**159**:1665–80.
50. Rowley MJ, Poulet A, Nichols MH, *et al.* Analysis of Hi-C data using SIP effectively identifies loops in organisms from C. elegans to mammals. *Genome Res* 2020;**30**:447–58.
51. Vietri Rudan M, Barrington C, Henderson S, *et al.* Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep* 2015;**10**:1297–309.
52. Chicco D, Bi HS, Reimand J, *et al.* BEHST: genomic set enrichment analysis enhanced through integration of chromatin

long-range interactions. *bioRxiv* 2019;168427January 15, 2019. https://doi.org/10.1101/168427 preprint: not peer reviewed.

53. Lan X, Witt H, Katsumura K, *et al.* Integration of Hi-C and ChIP-seq data reveals distinct types of chromatin linkages. *Nucleic Acids Res* 2012;**40**:7690–704.
54. Kruse K, Hug CB, Hernandez-Rodriguez B, *et al.* TADtool: visual parameter identification for TAD-calling algorithms. *Bioinformatics* 2016;**32**:3190–2.
55. Kaul A, Bhattacharyya S, Ay F. Identifying statistically significant chromatin contacts from Hi-C data with FitHiC2. *Nat Protoc* 2020;**15**:991–1012.
56. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;**26**: 841–2.
57. Shinkai S, Nakagawa M, Sugawara T, *et al.* PHi-C: deciphering Hi-C data into polymer dynamics. *NAR Genom Bioinform* 2020;**2**:lqaa020.
58. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;**9**:357–9.
59. Zhang Y, Liu T, Meyer CA, *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 2008;**9**:R137.
60. Nakato R, Itoh T, Shirahige K. DROMPA: easy-to-handle peak calling and visualization software for the computational analysis and validation of ChIP-seq data. *Genes Cells* 2013;**18**: 589–601.
61. Ernst J, Kheradpour P, Mikkelsen TS, *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 2011;**473**:43–9.