communications biology

Article

https://doi.org/10.1038/s42003-024-06904-0

Crosslinking intensity modulates the reliability and sensitivity of chromatin conformation detection at different structural levels

Bingxiang Xu [1,2,3,4,8,9] ✉, Xiaomeng Gao[3,4,5,8], Xiaoli Li[6], Feifei Li [7,9] ✉ & Zhihua Zhang [3,4,5,9] ✉

Formaldehyde (FA) is a chemical that facilitates crosslinking between DNA and proteins. It is widely used in various biochemical assays, such as chromosome conformation capture (3C) and Chromatin Immunoprecipitation (ChIP). While the concentration and temperature of FA treatment are recognized as crucial factors in crosslinking, their quantitative effects have largely remained unexplored. In this study, we employed 3C as a model system to systematically assess the impacts of these two factors on crosslinking. Our findings indicate that the strength of crosslinking significantly influences chromatin conformation detection at nearly all known structural levels. Specifically, a delicate balance between sensitivity and reliability is required when detecting higher-level structures, such as chromosome compartments. Conversely, intense crosslinking is preferred when targeting lower-level structures, such as topologically associated domains (TADs) or chromatin loops. Based on our data, we propose a conceptual molecular thermal motion model to elucidate the roles of these two factors in restricting FA crosslinking. Our results not only shed light on the previously overlooked confounding factor in FA crosslinking but also highlight the need for caution in new technology developments that rely on FA crosslinking.

In mammalian cells, the vast expanse of genomic DNA, which can extend for meters in length, is meticulously condensed and organized into a complex three-dimensional (3D) structure to fit within the confines of the micrometer-sized nucleus. This intricate 3D architecture of the genome plays a crucial role in numerous fundamental biological processes that occur within the nucleus[1]. Consequently, elucidating the precise 3D structure of the genome is essential for understanding the functional mechanisms it orchestrates, such as gene regulation[2], and for interpreting GWAS risk loci[3]

The advent of chromosome conformation capture (3C) and its derivatives, such as Hi-C, has been a primary driving force in revolutionizing the exploration of genome architecture over recent decades[4]. This has led to the revelation of a hierarchical organization of the genome, where chromosomes predominantly occupy mutually exclusive territories, known as chromosome territories[5]. Each chromosome can be broadly divided into active and inactive compartments, which may further consist of distinct structural units, such as topologically associating domains (TADs)[6,7]. The loop extrusion model has been proposed to explain the formation of chromatin loops and the consequential domain structure, suggesting the functional independence of these DNA sequence modules[8]. In addition to 3D genome studies, 3C-based technologies have also found their extensive applications in diverse fields, including complex genome assembly[9], haplotyping[10], and species clustering in metagenomics [11].

In the 3C protocols, DNA-protein crosslinking, typically facilitated using formaldehyde (FA), constitutes the initial phase, proficiently

[1]Key Laboratory of Hebei Province for Molecular Biophysics, Institute of Biophysics, School of Health Science & Biomedical Engineering, Hebei University of Technology, Tianjin, 300130, China. [2]School of Exercise and Health, Shanghai University of Sport, Shanghai, 200438, China. [3]China National Center for Bioinformation, Beijing, China. [4]Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China. [5]School of Life Science, University of Chinese Academy of Sciences, Beijing, China. [6]Department of Cell Biology and Genetics, Core Facility of Developmental Biology, Chongqing Medical University, Chongqing, 400016, China. [7]Division of Cell, Developmental and Integrative Biology, School of Medicine, South China University of Technology, Guangzhou, 510006, China. [8]These authors contributed equally: Bingxiang Xu, Xiaomeng Gao.[9]These authors jointly supervised this work: Bingxiang Xu, Feifei Li, Zhihua Zhang.
✉e-mail: xubingxiang@sus.edu.cn; liff@scut.edu.cn; zhangzhihua@big.ac.cn

preserving the chromosome conformation at DNA-protein interaction sites. This foundational step allows subsequent enzymatic digestion and ligation procedures to accurately delineate the in situ contact patterns. The intricate chemical processes of crosslinking, digestion, and ligation are well-documented in existing literatures. For example, the chemical principle of formaldehyde-induced crosslinking was elucidated as far back as 70 years ago[12]. Notably, alterations in formaldehyde concentration and incubation temperature can result in markedly distinct chemical outcomes[13,14]. Consequently, the judicious selection of crosslinking conditions and the choice of digestion enzymes can profoundly influence the results of 3C-based experiments. Nevertheless, a thorough optimization of 3C experimental parameters remains largely unexplored. Job Dekker and his team assessed various crosslinking scenarios, incorporating formaldehyde followed by EGS and DSG, in conjunction with four different nucleases (MNase, DdeI, DpnII, and HindIII). Their study indicated that the inclusion of additional chemical agents during crosslinking could significantly modulate the interpretations of Hi-C experiments[15]. Yet, the ideal concentration and temperature for formaldehyde treatment, which are pivotal parameters for crosslinking, remain unaddressed.

In the existing body of literature, there is a noted variability in the temperature at which FA is added and the subsequent crosslinking process occurs. For example, when cells are handled on a clean bench, FA is typically introduced at room temperature, approximately 25 °C. In contrast, for fluorescence-activated cell sorting (FACS)-sorted cells, FA may be directly applied at a much lower temperature[16–18]. Furthermore, the crosslinking temperature can be elevated in scenarios where samples undergo a heat shock or when FA is administered immediately post-incubation. The concentration of FA utilized for crosslinking also varies; recent investigations commonly employ either 1% or 2% FA[19–24]. This discrepancy in crosslinking parameters can markedly influence the resulting chromatin conformation map. For instance, studies utilizing *Drosophila* cells have demonstrated that divergent crosslinking methodologies can yield contradictory conclusions regarding chromatin conformation alterations subsequent to heat shock[25,26]. Consequently, a comprehensive evaluation of crosslinking conditions is imperative in 3C-based research endeavors.

In this investigation, we conducted a comprehensive assessment of the quantitative influences that fluctuations in FA concentration and alterations in cross-linking temperature exert on chromosome conformation profiles. Furthermore, we introduced a theoretical model of molecular thermal motion to elucidate the roles these parameters play in modulating FA crosslinking.

Results

In this study, we investigated the impact of two pivotal parameters in crosslinking-based 3C library generation: crosslinking temperature and formaldehyde (FA) concentration. We conducted Hi-C experiments on two widely studied human cell lines, K562 and GM12878, to evaluate four combinations of low (4 °C), medium (25 °C), and high (37 °C) temperatures with 1% and 2% FA concentrations. Additionally, for K562 cells, we tested the combinations of 4 °C and 37 °C with a 0.5% FA concentration to illustrate conditions of extremely low crosslinking strength (Fig. 1a). In total, we generated 12 Hi-C libraries for K562 cells and 8 libraries for GM12878 cells. For each crosslinking condition, we obtained two biological replicates with approximately 1 billion sequenced reads per library (Supplementary Data 1), aligning with the recommendations for bin sizes ≥5 kbp by HiCRes[27] and consistent with sequencing depths in contemporary studies. All subsequent steps in Hi-C library generation strictly followed the standard in situ Hi-C protocol[28]. The resulting sequencing data were analyzed using the HiC-Pro pipeline[29] with default parameter settings (Methods).

Crosslinking conditions may substantially affect the global preferences of DNA fragmentation and ligation

We utilized well-established attributes of 3C libraries to assess the variability and flexibility of ligating free DNA ends influenced by crosslinking conditions[29]. Our findings suggested that both temperature and FA concentration might significantly impact these properties.

Firstly, the restriction enzymatic digestion was biased towards the open chromatin regions with the increase of crosslinking temperature and FA concentration. This was shown by comparing the enzyme cutting frequencies in open regions (ATAC-seq peaks, Supplementary Table 1) against closed ones (H3K27me3 ChIP-seq peaks, Supplementary Table 1). After normalizing the sequencing depths (see details in Supplementary text), we observed a clear increasing trend of cutting frequency in open regions and a more enhanced enrichment compared to those in closed regions. The differences, measured by the probability of superiority (PS) of open regions over closed ones, monotonically increased from 0.46 ($p \approx 1.0$, Man Whitney U test) to 0.82 ($p \approx 0.0$) in K562 cells, and from 0.58 ($p \approx 1.0$) to 0.80 ($p \approx 0.0$) in GM12878 cells, in response to increased crosslinking temperature or FA concentration (Fig. 1b and Supplementary Fig. 1a).

Secondly, we observed a monotonic increase in the proportion of "religation" fragments, i.e., ligations between genomically neighboring ends, with the increase of crosslinking temperature and FA concentration. This proportion increased by 15-fold and 6-fold in K562 and GM12878 cells, respectively, from the smallest to largest conditions (Fig. 1c and Supplementary Fig. 1b).

Thirdly, within a FA fixed complex, fragment ends became increasingly prone to ligate to genomically proximal ends compared to other end pairs. This was evidenced by the growing enrichment for forward-reverse (FR) ligations, compared to FF, RR, and RF ligation directions in libraries, with the increase of crosslinking temperature or FA concentration (Fig. 1d and Supplementary Fig. 1c, with χ^2 statistics monotonically increasing from 2.29×10^3 in 4 °C / 0.5% FA to 1.03×10^7 in 37 °C /2% FA in K562 cells, and from 6.44×10^4 in 4 °C/1% FA to 2.12×10^7 in 37 °C/2% FA in GM12878 cells, all with $p < 10^{-10}$, Pearson's χ^2 tests). This enrichment led to a monotonic increase in the coefficients of association, measuring the unevenness of the four ligation directions, from 0.002 to 0.149 in K562 cells and from 0.010 to 0.212 in GM12878 cells.

At last, ligation with higher temperature or FA concentration exhibited enrichment for short-range cis (≤ 20 kbp) contacts, while it was depleted for distal (>20 kbp) cis and trans contacts (Fig. 1e and Supplementary Fig. 1d). This resulted in a decreased slope of contact frequency decay curves (or p(s) curves, Fig. 1f and Supplementary Fig. 1e). These observations indicated that the global ligation preferences and library structures differed essentially according to the crosslinking conditions. We also included a BL Hi-C library for K562 cells[30], which adopted a two-step ligation for more efficient capture of structural and regulatory chromatin interactions, to see if the advancement achieved by these sophisticated protocols could be mimicked by simply altering crosslinking conditions (Fig. 1f). The BL Hi-C achieved a contact frequency decay slope between that of 37 °C/1% FA and 37 °C/2% FA, with the exception of an increased number of contacts at medium distances (100 ~ 500 kbp) in the BL Hi-C library (Fig. 1f).

These observations indicated that the global cutting preference, ligation preferences, and library structures differed essentially according to the crosslinking conditions. Assuming the above indicators we examined roughly represented the crosslinking strength, we can rank the conditions in terms of crosslinking strength as follows: 4 °C/ 0.5% FA ≤ 37 °C / 0.5% FA ≤ 4 °C/1% FA ≤ 25 °C/1% FA ≤ 37 °C/1% FA ≤ BL Hi-C ≤ 37 °C/2% FA.

Contact maps under varied crosslinking conditions should not be considered equivalent as biological replicates

We subsequently evaluated the similarity of 3C contact profiles, as represented by the Hi-C contact maps, across the various assessed crosslinking conditions. A discernible distinction was readily apparent through simple visual inspection (Fig. 2a and Supplementary Fig. 2a). Employing the GenomeDISCO scores[31] chromosome-wisely to quantify similarity and their mean value for measurement of genome-wide similarity, we observed a monotonic decrease in the score between replicates with increasing crosslinking strength (Fig. 2b and Supplementary Fig. 2b), with the exception of

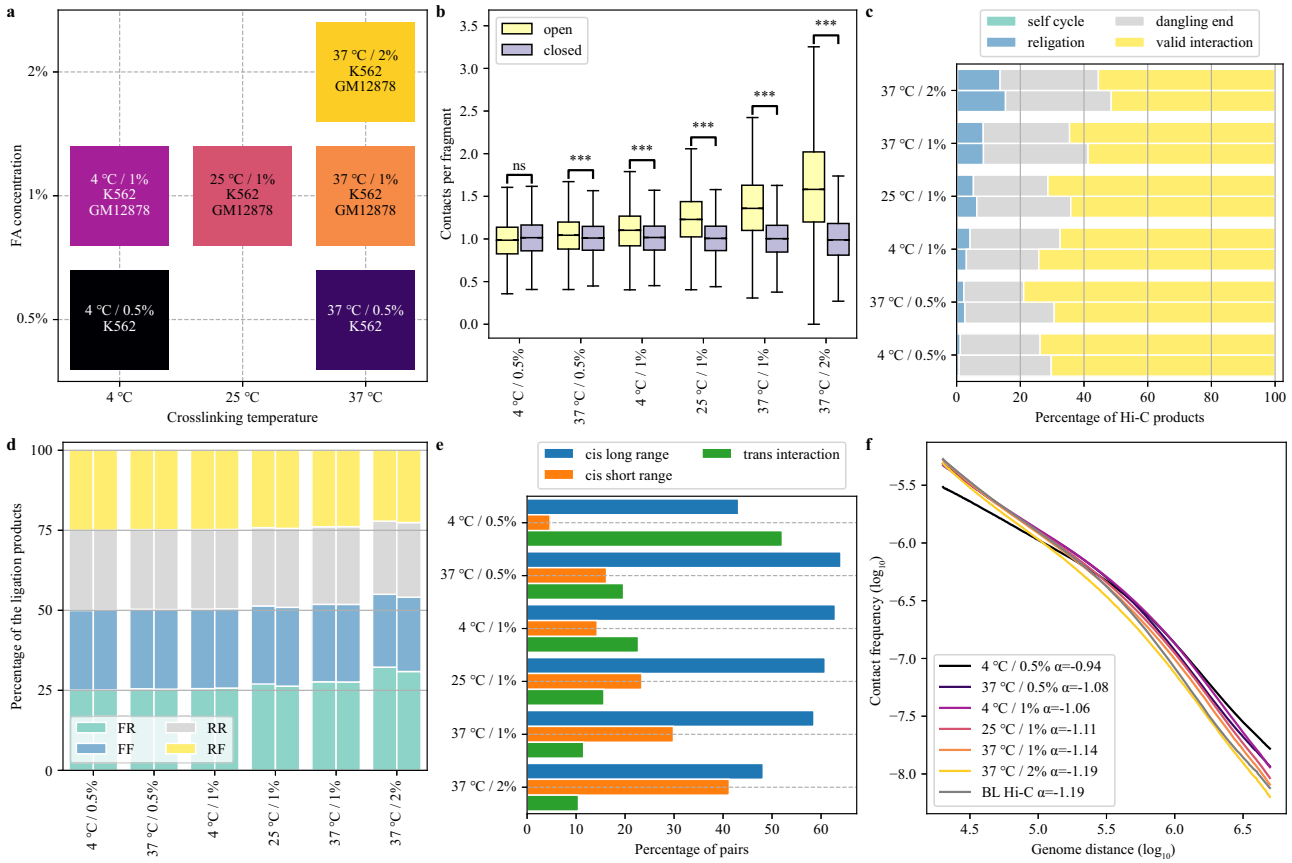

Fig. 1 | The global potential significant impact of crosslinking conditions on DNA ligation preferences. a The schematic representation of the experimental design employed in this study. **b** Boxplots showing normalized enzyme cutting frequencies of fragments in open (ATAC-seq peaks, 117855 fragments in total) and closed (H3K27me3 peaks, 116991 fragments in total) regions for libraries crosslinked with varying temperatures and FA concentrations in K562 cells. A Mann-Whitney test was conducted for significant detection with ns (p>0.1), * (0.01<p≤0.1), ** (0.001<p≤0.01), and *** (p≤0.001). **c** The proportions of sequencing product types under each crosslinking condition in K562 cells, including dangling ends, self-circles, religations, and valid pairs, which represent sequencing reads with no ligation, ligation between two ends of a single fragment, ligation between adjacent ends in neighboring fragments, and ligation between non-adjacent fragments, respectively. Biological replicates are displayed side by side, with self-circles constituting no more than 0.1% in each library. **d** The distributions of the four ligation directions in each library for K562 cells. **e** The proportions of short-range, long-range, and trans interactions among valid interaction pairs in each library for K562 cells. **f** The contact frequency decay curves for each crosslinking condition in K562 cells.

the extremely lowly crosslinked condition at 4 °C/0.5% FA. However, these similarities between replications remained larger than inter-condition comparisons (all $p < 0.05$, paired t-tests). This can be visually demonstrated through a hierarchical clustering plot of the genome-wide similarities (Fig. 2c), where all replicates were initially aggregated. Furthermore, only a small fraction of inter-condition comparisons exhibited a sufficiently high score, i.e., a score ≥0.9, specifically, only 2 out of 15 in K562 cells and 1 out of 6 in GM12878 cells. While these instances were consistently observed in intra-condition comparisons. Interestingly, at the extremely weak crosslinked condition (4 °C/0.5% FA), the contact matrices of K562 cells were even mis-clustered with GM12878 (Fig. 2c). Given this failure to capture the cell type specific chromatin conformation features in this condition, it was removed from further discussion.

Such differences can hardly be eliminated by data normalization. Even when clustering the libraries with the genome-wide essential distance, i.e., the distance that could not be canceled by any existing normalization procedure (Methods for the definition), we still observed the initial aggregation between replicates in the clustering in each cell type (Fig. 2d and Supplementary Fig. 2c). Moreover, akin to GenomeDISCO scores, only 2 and 1 out of 15 and 6 inter-condition comparisons achieved an essential distance ≤1, in K562 and GM12878 cells (Fig. 2d and Supplementary Fig. 2c), respectively, while all essential distances of intra-condition comparisons were ≤1.

Next, we aimed to examine how the aforementioned parameters of FA crosslinking could influence the identification of chromosome topology architecture across various levels of chromosome structure hierarchy.

Crosslinking strength may determine the balance between reliability and sensitivity in compartment identification

First, compartment assignment can be essentially influenced by crosslinking strength. In the principle component analysis (PCA) visualization of the PC1 vectors calculated from the autocorrelations of contact maps which determined the compartment assignments (see details in Methods) in all single libraries, a discernible trend is evident, wherein the "PC1" vectors align nearly perfectly according to the crosslinking strengths in both cell types (Fig. 3a). This alignment illustrated an almost linear effect of crosslinking strength on compartment separation, leading to merely 8.56% and 5.07% of genome regions changing their compartment assignments under different crosslinking conditions in K562 and GM12878 cells, respectively. Despite occupying a relatively small fraction of the genome, compartment assignments in these regions were found to be more reasonable under over-crosslinking conditions. In those regions displaying inconsistencies in compartment assignments, we observed that an increase in crosslinking strength could enhance the reliability of compartment identification. In other words, higher crosslinking strength led to a greater enrichment of active and repressive epigenomic marks in bins assigned as A or B

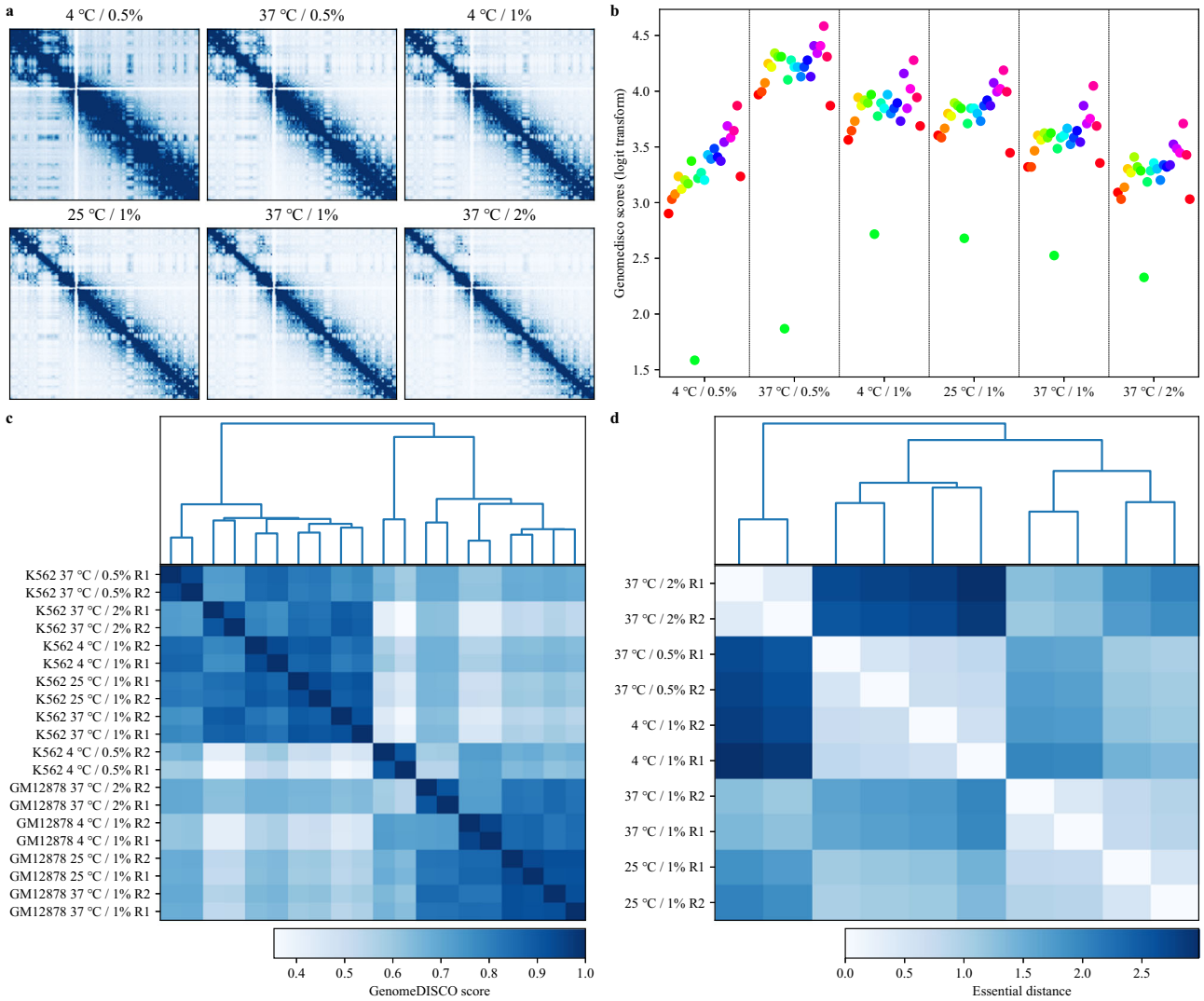

Fig. 2 | Contact maps generated under varying crosslinking conditions should not be considered as biological replicates. a An example of contact maps derived from different crosslinking conditions in K562 cells. The contact maps for chromosome 6 are shown at a 50 kbp resolution, with contact frequencies rescaled such that the row sums of intra-chromosome contact frequencies equal one. Frequencies ranging from 0 to 2.16×10^{-4} are mapped to colors ranging from white to blue. **b** The chromosome-wise GenomeDISCO scores between biological replicates of the intra-chromosome contact maps across all crosslinking conditions in K562 cells, calculated at a 50 kbp resolution. Chromosomes 1 through 22 and X are presented from left to right for each condition. **c** The clustering of GenomeDISCO scores across all libraries in K562 and GM12878 cells, with scores calculated at a 50 kbp resolution. **d** The genome-wide essential distances between libraries in K562 cells, with scores calculated at a 50 kbp resolution.

compartments, respectively. The active marks employed included gene expression level (the PS of A regions over B regions monotonously increased from 0.41 to 0.61 in K562 cells, and from 0.38 To 0.58 In GM12878 cells, with the increased crosslink strengths), DNA hypomethylation (DNA methylation rates, with PS values monotonously decreased from 0.76 to 0.21 in K562 cells, and from 0.56 to 0.46 In GM12878 cells) and chromatin accessibility (ATAC-seq, with PS values monotonously increased from 0.42 to 0.62 in K562 cells, and from 0.60 To 0.65 In GM12878 cells), while the repressive marks comprised DNA hypermethylation and H3K27me3 (ChIP-seq, with PS values monotonously decreased from 0.68 to 0.34 in K562 cells, and from 0.82 To 0.19 In GM12878 cells, Fig. 3b and Supplementary Fig. 3a, Supplementary Table 1 and Supplementary text). Interestingly, under extremely weak crosslinked conditions (37 °C / 0.5% FA in K562 and 4 °C/1% FA in GM12878), the direction of the enrichment was even inverted.

Second, in addition to reliability, excessive crosslinking may render the compartment assignment less reproducible and distinct. We observed a monotonic decrease in both the chromosome-wise correlation coefficients of the PC1 values (Fig. 3c and Supplementary Fig. 3b), and the fractions of the genome assigned to the same compartments (Fig. 3d) between biological replicates, with an increase in crosslinking strength across both cell types. Furthermore, the scales of the autocorrelation matrices of contact maps, encompassing both the positive values between identical compartments and negative values between different compartments, were found to decrease monotonically with the increase of crosslinking strength (see an example in Supplementary Fig. 3c, d). The relative strengths between inter- and intra-compartment contacts, as depicted in the back-diagonal corners of the saddle plots (Fig. 3e and Supplementary Fig. 3e, see definitions in Methods), also increased monotonically, resulting in increasing compartment scores (see definitions in Methods, Supplementary Fig. 3f, g). Collectively, these data suggest that an increase in crosslinking strength may diminish compartment separation.

Thirdly, the compartment assignments were globally kept but locally refined in the BL Hi-C library. There were 93.43% and 92.80% of the

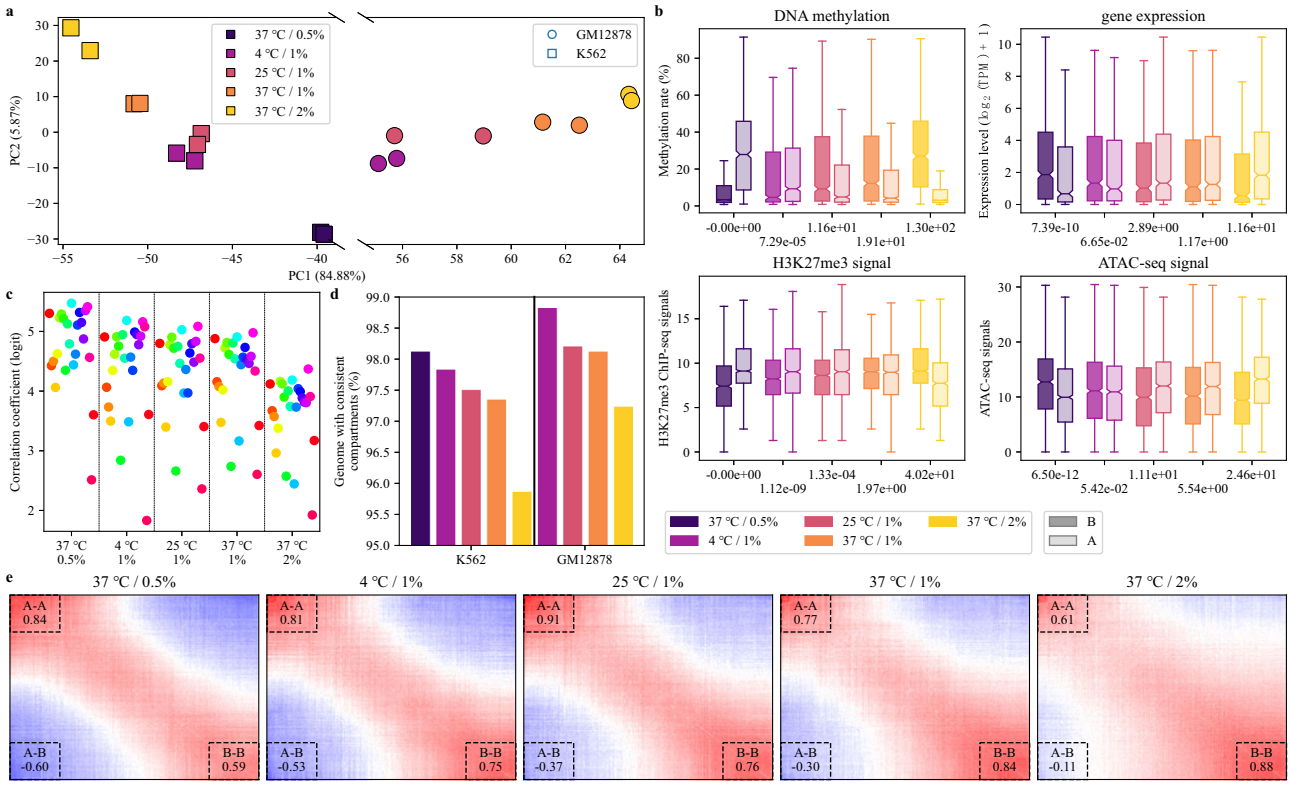

Fig. 3 | The strength of crosslinking may play a pivotal role in determining the balance between reliability and sensitivity in compartment identification. a The principal component analysis (PCA) results from the "PC1" values of all libraries in K562 (left) and GM12878 (right) cells. Only the first (PC1) and second (PC2) principal components are displayed, along with their respective ratios of explained variances. **b** The DNA methylation rates, gene expression levels, H3K27me3 ChIP-seq signals, and ATAC-seq signals in regions designated as compartment B (left in each pair) and compartment A (right in each pair) under each crosslinking condition in genome regions with inconsistent compartment assignments in K562 cells. The p-values of Mann-Whitney U tests comparing values in each pair are indicated below (after $-\log_{10}$ transformation). Group-wise sample sizes (from left to right): DNA methylation: 1501, 903, 1142, 1262, 1243, 1161, 1247, 1157, 959, 1445; gene expression: 779, 675, 797, 657, 572, 882, 650, 804, 564, 890; H3K27me3 signal: 903, 1501, 1262, 1142, 1161, 1243, 1157, 1247, 1445, 959; ATAC-seq signal: 903, 1501, 1262, 1142, 1161, 1243, 1157, 1247, 1445, 959. **c** The chromosome-wise Spearman correlation coefficient (SCC) values of the "PC1" vectors determining compartment assignments between replicates of all crosslinking conditions in K562 cells. Chromosomes 1 through 22 and X are marked from left to right in each condition. **d** The percentage of the genome with consistent compartments in biological replicates for each crosslinking condition. **e** The genome-wide saddle plots demonstrating the separation of compartments A and B in K562 cells.

genome kept in the same compartments when comparing BL Hi-C with those in 37 °C/1% and 2% FA. However, BL Hi-C improved the compartment separation, achieving more A-A and B-B interactions while maintaining A-B interaction frequencies (see Supplementary Fig. 3h). In addition, BL Hi-C gives more reliable compartment assignments. For regions with inconsistent assignment, the enrichment of active or repressed epigenetic signals in A or B compartment was more pronounced for BL Hi-C (except for H3K27me3, see Supplementary Fig. 3i).

At last, the difference in compartment detection we observed cannot be attributed exclusively to the bias in enzyme cleavage preferences. The preferred enzyme cutting in compartment A yielded shorter enzymatic fragments, making it difficult to form A-A contacts in over crosslinked libraries[32]. This explained the dissolved compartment separation in these libraries. However, in regions with inconsistent compartment assignments, the cutting frequencies did not exhibit a monotonically increasing trend in A compartment (Supplementary Fig. 3j). This implied that enhancement of the reliability of compartment assignment in over-crosslinking libraries cannot be explained by the enzyme cutting preferences.

All above results were not due to the differences in library complexities among the crosslinking conditions. All the aforementioned findings, the more reliable compartment assignments (Supplementary Fig. 3k), the dissolved compartment separations (Supplementary Fig. 3l, m) and the increased compartment scores (Supplementary Fig. 3n, o) with the increased crosslinking strength, were all kept when the effective library sizes were made the same by random down sampling to equal number of effective contacts (see method for detail, Supplementary Fig. 3k, o).

Therefore, the appropriate selection of crosslinking level is crucial for striking a balance between the identification of reliable and distinct chromosome compartments.

Elevating the crosslinking level may improve the convergence of TAD boundaries detection toward functional insulation sites

To quantitatively assess the impact of crosslinking on the detection of topological associated domains (TADs[7],), we employed the commonly used insulation score (IS) profiles[33] to delineate the TAD structures at a 5 kbp resolution (see details in Methods).

Initially, the identified TAD structures exhibited rough consistency across the various crosslinking conditions we tested (examples in Fig. 4a and Supplementary Fig. 4a). If we define two TADs as approximately identical when their genomic regions overlap by more than 80% of both their lengths, and correspondingly define the split, merge, and shift events of the TADs (referring to Methods), we observed a minimal occurrence of shift events (Fig. 4b and Supplementary Fig. 4b).

The proportions of shifted TADs did not significantly differ from those observed between biological replicates in most cross-condition comparisons (Fig. 4b and Supplementary Fig. 4b). Notably, only two out of the 20 comparisons in K562 cells (between 37 °C/ 0.5% FA and 37 °C/2% FA, and between 4 °C/1% FA and 37 °C/ 2% FA) and one out of 12 comparisons in GM12878 cells (between 4 °C/1% FA and 37 °C/2% FA) demonstrated

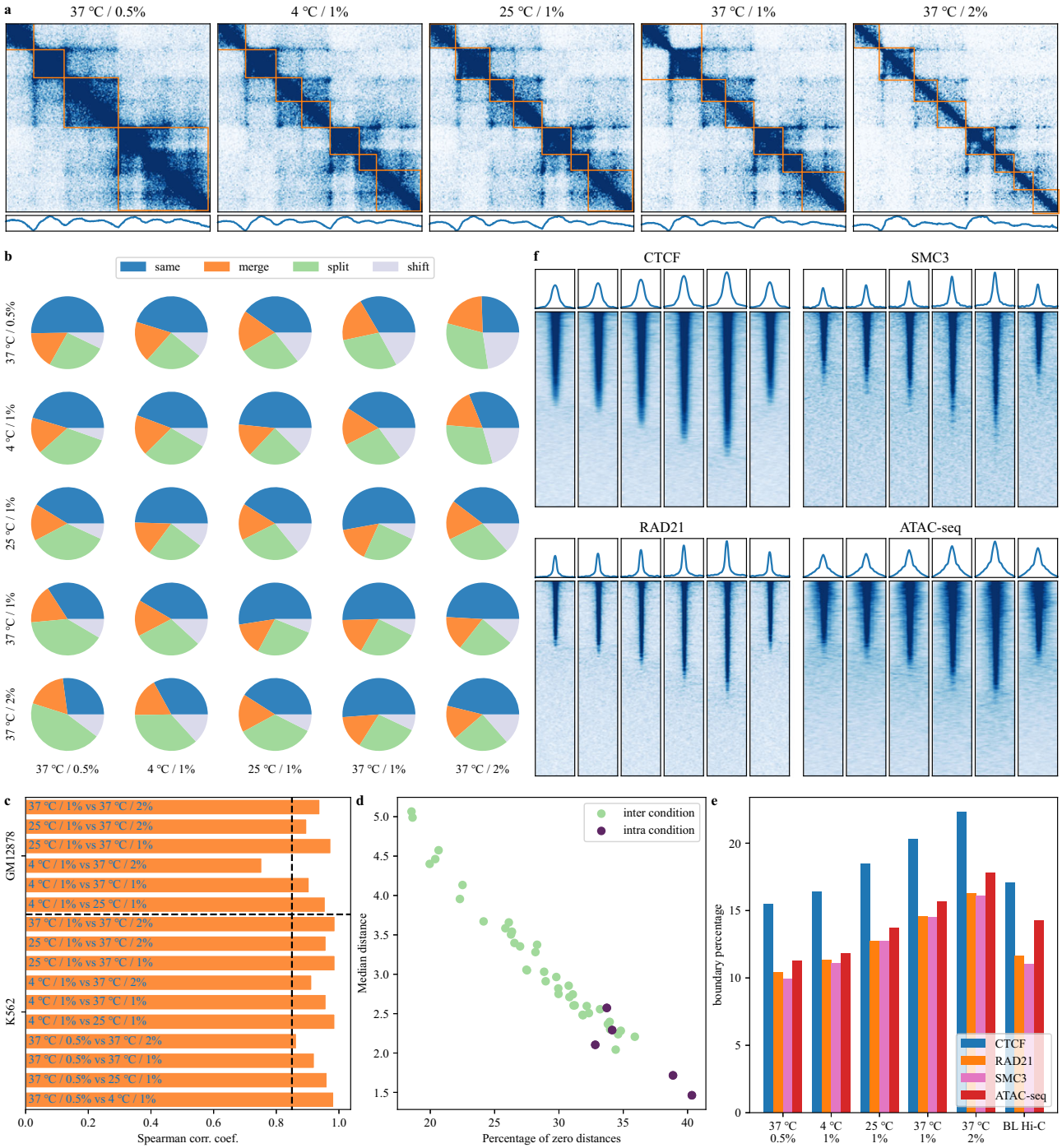

Fig. 4 | Increasing the crosslinking level may enhance the accuracy of TAD boundary detection, aligning it more closely with functional insulation sites. **a** An example illustrating the global similarity of TAD structures identified under different crosslinking conditions. The contact maps for the region chr20:56.59-58.13 Mb in K562 cells are shown alongside the TADs detected in this region marked by the orange boxes. The IS profiles are displayed below the contact maps for comparison. Color scales of all contact maps were set uniformly from 0 to 5×10^{-4} along all contact maps. **b** Pie charts comparing TAD positions between different crosslinking conditions in K562 cells. Comparisons are made from rows to columns, with average results of the two comparisons between biological replicates displayed along the diagonal. **c** Spearman correlation coefficient (SCC) values of IS profiles between every pair of crosslinking conditions in K562 and GM12878 cells. The conditions being compared in each bar are marked accordingly. **d** The proportions of zero boundary distances and median values in boundary distances between inter- and intra-condition comparisons in K562 cells. Values are calculated on a library-wise basis. **e** The percentages of TAD boundaries in each crosslinking condition that harbored at least one signal peak in CTCF, RAD21, SMC3 ChIP-seq data, and ATAC-seq data in K562 cells. **f** Heatmap of CTCF, SMC3, RAD21 ChIP-seq, and ATAC-seq signals at TAD boundaries in each crosslinking condition in K562 cells. Conditions ranging from 37 °C/0.5% FA to 37 °C/2% FA are shown from left to right in each panel. BL Hi-C boundaries are displayed on the far right. Signals are aligned according to the highest value in the boundary bins. Average signal profiles among all boundaries are shown at the top of each heatmap. In each heatmap, only the top 3500 boundaries with the highest average signals are displayed for comparability.

significantly larger proportions of shifted TADs compared to the involved biological replicates (t tests, $\alpha = 0.01$, after Bonferroni adjustment), and these differences were only observed between the furthest conditions.

This consistency of TADs was further supported by high global correlation coefficients of IS profiles between crosslinking conditions in both cell types (all with Spearman correlation coefficient, SCC >0.85, Fig. 4c), with one exception of comparing two most distinct conditions in GM12878 cells (with SCC = 0.75). In addition, IS profiles around the small portions of condition-specific TAD boundaries, defined as a boundary in one condition with no boundary in the other located no more than 50 kbp away, also support the consistency of TADs. IS values around these boundaries were either globally elevated or with reduced contrasts in pair-wise comparisons, illustrating the low reliability (Supplementary Fig. 5a, b).

Despite the rough consistency in TADs, boundary positions exhibited substantial variations between crosslinking conditions, surpassing the variation observed between biological replicates. This variation was quantified as the minimum genome distance between the two closest boundaries in each comparison, revealing a clear shift of the distance distributions toward larger distances in inter-condition comparisons than intra-condition ones (Fig. 4d and Supplementary Fig. 4c, see details in methods). As expected, biological replicates were preferentially clustered when only IS values at boundaries (at any condition) were considered (Supplementary Fig. 4d).

Furthermore, boundaries identified under stronger crosslinking conditions exhibited greater functional supports from epigenomic data at a finer scale compared to weaker crosslinking conditions. Given that the peak length of TF ChIP-seq and ATAC-seq is typically less than 1 kbp, we examined the consistency between known functional characteristics and the boundary identified at such a fine scale. Plots of CTCF, SMC3, RAD21 binding, and chromosomal accessibility profiles in both cell types revealed that with increased crosslinking strength, more boundaries were occupied by peaks of these epigenetic marks in their respective bins (Fig. 4e and Supplementary Fig. 4e, Supplementary Table 1, Supplementary text[34,35],). Moreover, the strength of binding, as well as the accessibilities, demonstrated a monotonous increase with crosslinking strength (Fig. 4f, Supplementary Fig. 4e, f). All results were retained and not due to the differences among library complexities (Supplementary Fig. 4g, i).

The enhancement of TAD boundary positioning in over crosslinked conditions was largely independent of enzyme cutting preferences (Supplementary text). As shown in Supplementary Fig. 4j, the refined boundaries in over crosslinked conditions did not exhibit higher cutting frequencies after normalization.

For BL Hi-C, the number of TADs and boundaries detected were significantly fewer than other conditions in the data we tested (Supplementary Table 2). Despite this, TADs still exhibited consistency when compared to other libraries (Supplementary Fig. 4k). It worth mentioned that the occupancies of functional characteristics in boundaries of BL Hi-C were only comparable to crosslinking conditions of 4 °C or 25 °C and 1% FA (Fig. 4e, f). These imply that both the sensitivity and accuracy of BL Hi-C detected boundaries were compromised compared to over crosslinking.

In summary, although the global TAD structures remained largely stable across different crosslinking strengths, the refinement of domain boundaries in stronger crosslinking conditions revealed greater functional significance.

Excessive crosslinking augmented the reliability of chromatin loop detection without significantly compromising its sensitivity

The sensitivity of loop detection remained minimally affected by crosslinking. We identified loops at a 5 kbp resolution using the highly sensitive caller mustache[36] within distances ranging from 30 kbp to 2 mbp. Since the number of detected loops showed a monotonic increase with effective library sizes (Supplementary Table 2, Fig. 5a and Supplementary Fig. 6a), and contacts in "condition-specific" loops were still enriched in conditions where they were not called (Supplementary Fig. 7a, b), it was evident that the libraries had not been saturated with sequencing. We utilized the ratio between the number of loop anchors involved and the number of loops called as the sensitivity index for loop detection[15] instead of direct counting of the loop numbers. The ratios (averaging 1.6057 in K562, including the BL Hi-C library, and 1.6781 in GM12878 cells, respectively) remained nearly identical across all conditions, even after merging the replicated libraries to roughly double the sequencing depth (Fig. 5a and Supplementary Fig. 6a). Crosslinking conditions were well-fitted by a single linear model without intercept in each cell type (both with $R^2 > 0.99$) both before and after merging biological replicates, without any significant outliers (with p-value threshold 0.1, detected using standardized residuals). These observations indicated that crosslinking minimally affected the sensitivity of loop detection. This consistency of the sensitivity was also maintained in the BL Hi-C library, though the number of loops detected in BL Hi-C was fewer than in other conditions (Supplementary Table 3).

Loops detected in highly crosslinked libraries exhibited more functional support. We utilized CTCF/SMC3, together with histone marks H3K4me3 and H3K27ac for promoters and enhancers, to infer the potential functionality of the loops. First, the enrichment of functional signals in the loop anchors gradually increased with crosslinking strength (Fig. 5b and Supplementary Fig. 6b). Second, more interactions between functional genome elements, such as enhancers and promoters, were found enriched in highly crosslinked libraries. An over-enrichment of both enhancer-promoter interactions (EPI), promoter-promoter interactions (PPI), and enhancer-enhancer interactions (EEI) were revealed in over crosslinked libraries in the two cell types (Fig. 5c and Supplementary Fig. 6c, all with $p < 10^{-50}$, χ^2 tests, see Methods for the details of the enrichment analysis) after taking the annotation of cell type-specific enhancers from the Enhancer Atlas database[37]. This was also observed for CTCF-mediated loops, i.e., those loops with anchors decorated by CTCF binding peaks with significant binding motifs (Fig. 5d and Supplementary Fig. 6d, all with $p < 10^{-50}$, χ^2 tests).

Interestingly, BL Hi-C demonstrated advantage in the detection of only CTCF-mediated loops, not the enhancer/promoter involved ones, compared to over crosslinking. BL Hi-C yielded a higher enrichment of CTCF and SMC3 signals in loop anchors and a higher enrichment of CTCF-mediated loops than other conditions (Fig. 5b, d). The enrichment of enhancer/promoter signals and enhancer/promoter involved loops were similar between BL Hi-C and crosslinking under 37 °C/2% FA (Fig.5b, c).

Furthermore, the advantage of over crosslinking in detecting enhancer/promoter-involved loops was further validated by the increased expression of genes which are involved in enhancer/promoter loops in over crosslinked libraries. Genes were roughly classified into three groups based on whether their TSSs were connected by loops to distal regulatory elements, i.e. no loop, trivial loop (TSSs connected to non-enhancer or promoter regions), E / P loop (TSSs connected to at least one promoter or enhancer region), respectively. The expression differences between the groups increased with crosslinking strength, as measured by Kruskal-Wallis tests (Fig. 5e and Supplementary Fig. 6e). There was also a monotonically increasing trend of the expression levels of genes with increased crosslinking strengths in the "trivial loop" (with PCC of 0.84 in both K562 and GM12878 cells between crosslinking strengths and median expression levels, $p = 0.04$ and 0.08, respectively, t tests for the PCC values) and "E / P loop" groups (with PCC values 0.92 and 0.87, $p = 0.03$ and 0.06 in K562 and GM12878 cells, Fig. 5e and Supplementary Fig. 6e), which was not evident in the no-loop group (with $p > 0.1$ in both cell types). Expression levels of genes in the "trivial loop" and "E/P loop" groups detected by BL Hi-C were not significantly higher than corresponding groups in 37 °C/2% FA (Mann-Whitney U test, both with $p > 0.1$, with PS values 0.48 and 0.49, respectively, Fig. 5e).

The improvement in loop detection with excessive crosslinking can hardly be explained by the enzyme cutting preference. The normalized cutting frequency in each anchor (see the definition in Supplementary text) did not exhibit a monotonically increasing trend with increased crosslinking strength (Supplementary Fig. 6f). Enhancement can also hardly be attribute to common biases such as shorter loop lengths (Supplementary Fig. 6g) and reduced library complexities in overly crosslinked libraries. The results remained when loops were re-weighted to make the loop length

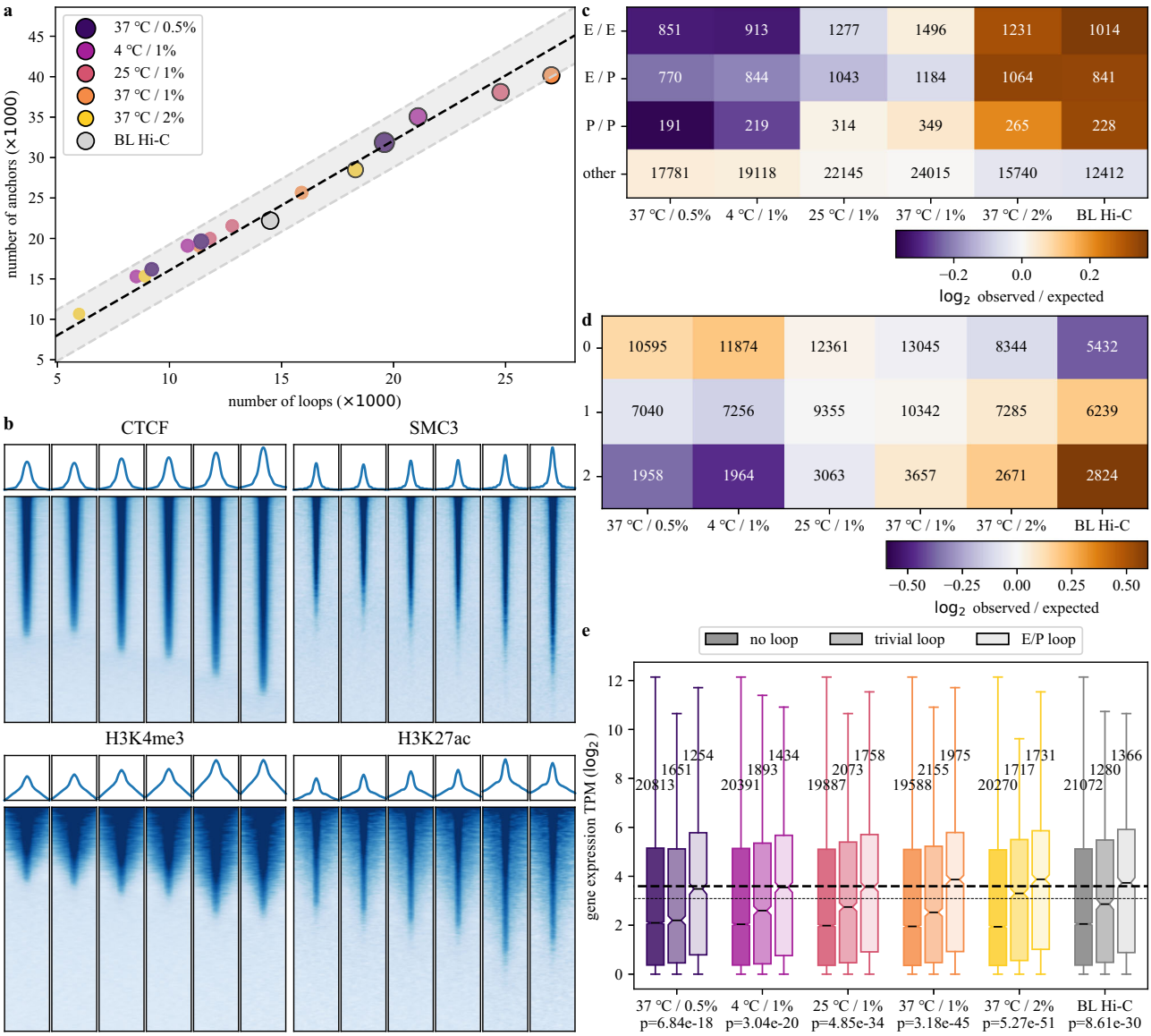

Fig. 5 | Excessive crosslinking enhances the reliability of chromatin loop detection without significantly impacting its sensitivity. a The relationship between the number of loops and the number of anchors involved across all crosslinking conditions in K562 cells. Dots without edges represent individual biological replicates, while those with edges represent their merged data. The dashed line and shadow region indicate the fitted linear model of these two numbers and their corresponding confidence intervals. **b** The heatmap of CTCF, SMC3, H3K4me3, and H3K27ac ChIP-seq signals at loop anchors in each crosslinking condition in K562 cells. Conditions ranging from 37 °C/0.5% FA to 37 °C/2% FA are displayed from left to right in each panel. The BL Hi-C library is placed on the far right. Signals are aligned according to the maximum value in the anchor bins. Average signal profiles among all anchors are shown at the top of each heatmap. In each heatmap, only the top 50% (for CTCF and SMC3) and top 30% (for H3K4me3 and H3K27ac) anchors with the highest average signals are displayed for comparability. **c** The enrichment of enhancer (E) and promoter (P) involved loops in each crosslinking condition. Colors indicate the observed/expected number of loops, where the expected numbers are calculated using a contingency table assuming independence between enhancer/promoter involvement and crosslinking conditions. **d** The enrichment of CTCF involved loops in each crosslinking condition. The number of loops with 0, 1, and 2 binding CTCF motifs at occupied anchors is marked. The structure of this panel is identical to **(c)**. **e** The expression levels of three gene groups classified based on the elements they interact with through loops in each crosslinking condition. The *p*-values of Kruskal-Wallis tests measuring the differences between the three groups in each condition are indicated below. Thin and thick horizontal lines mark the mean values of average expression levels among crosslinking conditions for the "trivial loop" and "E/P loop" groups, respectively. On each box plot, the number of genes involved in the corresponding groups (sample sizes) is marked.

distributions identical (see Methods for details, Supplementary Fig. 6h, i) or when libraries were randomly down sampled to equalize the effective library sizes (Supplementary Fig. 6j–n) across different crosslinking strengths.

A conceptual model for FA crosslinking and advances for condition choice

One conceptual model derived from our data proposes the following scenario: FA crosslinking of DNA and proteins induces a restriction of molecular thermal motion. In an ideal extreme case, complete crosslinking of DNA and proteins results in the fixation of chromatin conformation, rendering the restriction fragments capable of only re-ligation (FR) with no valuable Hi-C reads obtained. As the crosslinking level decreases, the freedom of post-crosslinking molecular thermal motion increases, expanding the search space for fragment ends. In accordance with the principles of polymer physics[38], where distal monomers in a polymer chain exhibit a longer Euclidean distance than proximal ones in 3D physical space, fragment ends in proximity become ligatable with decreased crosslinking levels. With the expanding search space, more distal fragments become accessible,

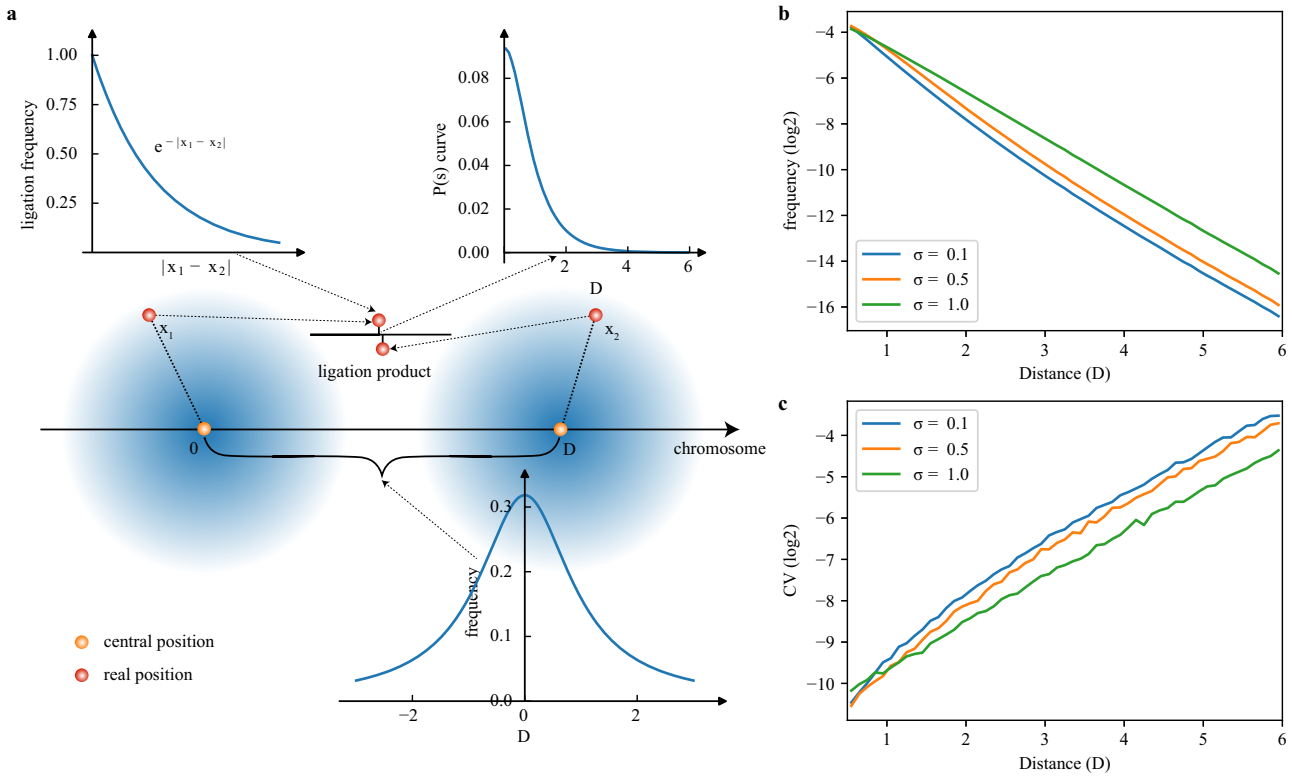

Fig. 6 | The conceptual model for FA crosslinking and insights into the selection of optimal conditions. a A schematic diagram of the model, where σ represents the standard deviation that regulates the strength of crosslinking. **b** The mean contact frequency decay curves as functions of central distance D under varying crosslinking strengths, based on 500 simulation runs. **c** The coefficient of variance values, which indicate the signal-to-noise ratio, as functions of central distance D under different crosslinking strengths, calculated from the 500 simulation runs.

reflected in the flatter contact frequency decay curves observed in our weaker crosslinking libraries.

To assess the feasibility of this model, we conducted a simple quantitative simulation (see details in Methods and Fig. 6a) by considering the crosslinking as a diffusion process of two ends that stochastically float in a 3D spherical space. The occurrence of ligation was modeled by a Bernoulli distribution with a ligation probability that exponentially decreased with the increase of the distance between two ends. The positions of the ends were modeled by two independent Gaussian distributions centering at their central positions and with identical standard deviations (SD). The distances between the central positions (termed as central distance) were driven by a standard Cauchy distribution (Fig. 6a). The crosslinking strengths were negatively modeled by the SDs, i.e., a larger SD resulted in a smaller crosslinking strength. The contact frequency decay curves of ligated ends (Fig. 6b) were calculated by simulating 1×10^7 independent end pairs at a time. The coefficient of variance (CV) of these curves was then calculated as the CV values of contact frequencies at each central distance in 500 rounds of simulation (Fig. 6c). The CV values, as a function of central distances, measure the reproducibility of independent libraries (biological replicates) under each crosslinking strength. As expected, the slope of the contact frequency decay curve decreased with the increase of crosslinking strength (Fig. 6b), reflecting our experimental data (Fig. 1e and Supplementary Fig. 1d). Similarly, the CV values also monotonically increased with the increase of crosslinking strength in most parts of the central distance interval we examined (Fig. 6c), which revealed the defection of reproducibility between biological replicates observed (Fig. 2b and Supplementary Fig. 2b).

Discussion

Our investigation into the impact of crosslinking temperature and FA concentration on the Hi-C readout has unveiled substantial implications for the enzyme cutting profile, ligation profile, and accuracy in detecting different layers of chromosomal conformation. The enzyme cutting preference might be the primary cause of the dissolved chromatin compartment separation induced by over crosslinking. However, it can hardly explain the improvement in reliability on feature detection. Our study cannot exhaustively identify all potential source of bias driven by crosslinking, more elaborate experimental design with sophisticated technology, e.g., super-resolution live imaging, at the moment of crosslinking, shall always be a key way for further investigation. Nevertheless, the results we reported in this study represent the most comprehensive survey on this topic to date.

Our results demonstrate that intensive crosslinking yields more robust, and possibly more biologically associated compartmental assignments, TAD boundaries, and chromatin loops. According to our working model (Fig. 6), this result seemingly indicates that the approximate chromatin interactions were more functional than distal ones, which was widely accepted but has recently been challenged[39]. On the other hand, heavy crosslinking captures more flexible chromatin interactions, providing a more heterogeneous representation of chromatin interactions. This attenuates the reproducibility in highly intense crosslinking conditions. Based on our conceptual model, the effects of crosslinking strengths on the results of 3C-based assays can be analogized to taking photos with different shutter speeds. Increasing crosslinking strength is like using a higher speed, which not only captures more meaningful decisive moments but also preserves more noise caused by external factors such as camera vibration. It should be mentioned that our working model treated the chromatin fiber as a homopolymer, and the effects of enzyme cutting preferences were overlooked. Since the effects of enzyme cutting preferences could be efficiently canceled by current data normalization algorithms, this approach did not detract the main findings of this study.

Recent studies have demonstrated that additional crosslinkers such as DSG and EGS may enhance loop and compartment detection[15]. Our results

complement these findings by highlighting that increased crosslinking improves accuracy for TAD boundary and loop detection. Nonetheless, the influence on compartment detection remains nuanced. One explanation for this inconsistency might be as follows. When elevate crosslinking level by temperature or FA concertation, as discussed above, the ligation reaction on restricted chromatin fibers resulted in relative lower distal connections. On the other hand, when elevating crosslinking levels by adding extra linkers, the chromatin fibers could be captured by farer distal ones, as the extra linker may make a larger crosslinking radius compared with FA [40,41].

The bias introduced by crosslinking strengths cannot be canceled by the current data normalization framework (Fig. 2d and Supplementary Fig. 2c). Moreover, this bias can hardly be normalized even beyond the current bilinear framework without introducing new bias. Firstly, the second-order feature of the data, which is vital for determining the folding model of the chromatin[6,42], was incorrectly altered when normalization. Secondly, the quantitative effects of crosslinking strength are cell type-specific. As a result, the crosslinking conditions should always remain exactly the same among all libraries in whole studies. For experiments aiming at detection of chromatin features reliably, it is advisable to increase the crosslinking strength by elevating the temperature and concentration, if the relatively reduction of effective contacts is acceptable.

In conclusion, our study underscores the critical influence of varying crosslinking temperatures and FA concentrations in protein-DNA fixation experiments. As such, careful consideration is warranted when comparing results from studies utilizing different crosslinking conditions. Additionally, the development of novel analytical methodologies capable of normalizing the effects of diverse crosslinking conditions is imperative for comprehensive integrated analyses in this field.

Methods

Cell culture and crosslinking

K562 and GM12878 cells, which were purchased respectively from the American Type Culture Collection and the Coriell Institute, were maintained at 37 °C under 5% CO2 in RPMI-1640 medium (Sigma-Aldrich) supplemented with 10% FBS, 2mM L-glutamate, and streptomycin/penicillin. Crosslinking temperature stands for the temperature when the FA were added to the cells. For 37 °C/0.5%, 37 °C/1% or 37 °C/2%, the FA with final concentration of 0.5%, 1% or 2% were immediately added to cells after removal from incubator. For 25 °C/1%, the cells were mixed with equal volume of 4 °C RPMI-1640 medium to adjust to 25 °C, and then 1% FA were added. For 4 °C / 0.5% and 4 °C/1%, the cells were washed and resuspended with PBS at 4 °C, simulating the operation of FACS, and then the FA was added. The crosslinking process lasts for 10 min in room temperature for all conditions.

In situ Hi-C library preparation

In situ Hi-C was conducted according to the literature[28]. Briefly, after crosslinking for 10 min, 0.125 M glycine was added to quench the reaction. Then, cells were lysed and digested with the MboI restriction enzyme (NEB, R0147). Biotin-14-dATP was used to mark the DNA ends, followed by proximity ligation in intact nuclei. After crosslink reversal, samples were sheared to a length of ~300 bp and then treated with the End Repair/dA-Tailing Module (NEB, E7442L) and Ligation Module (NEB, E7445L), following the manufacturer's instructions. Then, biotin-labeled fragments were pulled down using Dynabeads MyOne Streptavidin C1 beads (Invitrogen, 65602). Finally, the Hi-C library was amplified for about 10 cycles of PCR with the Q5 master mix (NEB, M0492L), following the manufacturer's instructions. Size selection was performed with AMPure XP beads, quantified and sequenced on an Illumina HiSeq X Ten instrument with 2 × 150 bp reads.

Hi-C data processing

Prior to analysis, all Hi-C libraries were rigorously assessed for quality using FastQC (https://github.com/s-andrews/FastQC) to verify sequencing standards. Subsequently, adapters and sequences with inadequate sequencing qualities (< 10) were discarded employing fastp[43]. Additionally, reads of length ≤ 20 bp were eliminated from further consideration. The retained, high-quality reads were then processed utilizing the standard HiC-Pro pipeline[29], incorporating the hg19 genome assembly—a selection aimed at maintaining consistency with other publicly accessible datasets. Notably, the selection of the genome assembly would not impact the final conclusions drawn from the study. Reads mapping to chromosome Y or M, or those with a MAPQ score of 0, were systematically excluded. The contact frequency decay curves were derived from genome distances ranging from 20 kbp to 5 mbp, segmented into 500 logarithmically spaced bins. Finally, the contact data were transformed into the .hic format through the utilization of Juicer tools [44].

The determination of A/B compartments was carried out on each observed versus expected (oe) normalized, KR balanced intra-chromosome contact matrix at a 100 kbp resolution. Before further calculations were performed, centromeres and other regions that could not be mapped were systematically masked. Compartment assignments were ascertained by calculating the first eigenvector (PC1) of the auto-correlation matrix chromosome-wisely using spectral factorization. Subsequently, these annotations were juxtaposed with the v19 gene annotation—the most recent version for hg19—from GENCODE[45], ensuring that regions with a higher gene density (average number of TSSs per bin) corresponded to positive PC1 values. In instances where this was not the case, the PC1 values were inverted by multiplying them by -1. Bins exhibiting positive PC1 values were designated as "A" compartments, while those displaying negative PC1 values were categorized as "B" compartments. To render the PC1 values from different chromosomes comparable, they were scaled by multiplying them with the square root of their respective chromosome lengths. The principal component analysis (PCA) visualization and clustering of these PC1 values were applied to the sequence of rescaled, chromosome-specific PC1 vectors.

Definition of the essential distance measurement of two contact maps

For an arbitrary chromosome's contact map C and a vector $W = \{w_i\}$ with all $w_i>0$, the current normalization algorithms, both matrix-balancing based, and probability model based, adopt the form

$$\hat{C} = diag(W) \times C \times diag(W)$$

The function diag(∗) transform a vector to a diagonal matrix. For two contact maps C_1 and C_2 of same chromosome, the distance:

$$d(C_1||C_2) = \min_W \frac{||C_1 - diag(W)C_2 diag(W)||}{||C_1||}$$

measured the minimum distance between C_1 and any possible normalization of C_2. The denominator $||C_1||$ is used to cancel out the effects of sequencing depth differences. The distance d was then symmetrized to define the essential distance for a chromosome:

$$E(C_1||C_2) = \frac{d(C_1||C_2) + d(C_2||C_1)}{2}$$

In this study, we first calculated E for each chromosome by taking the 2-norm and finding the minimum value using stochastic gradient descent. Subsequently, a genome-wide essential distance was computed by averaging the E values across each chromosome.

Hierarchical clustering

All distance matrices utilized in this study for hierarchical clustering, except for the essential distances which were directly employed, were either obtained by computing the Euclidean distances between the values to be clustered or were derived from similarity matrices (here denoted as S) using the formula $1 - S$. The clustering processes were all executed in

a bottom-up manner, wherein the minimum distances between points in the newly formed clusters were used to represent their distances.

Generating the genome-wide saddle plots and calculating the AA, AB and BB contact intensities

The saddle plots were initially computed for each chromosome using the KR-normalized, observed over expected contact maps. Prior to these calculations, centromeres and other unmappable regions were excluded. Subsequently, the contact frequencies were log-transformed and rearranged in descending order based on their PC1 values. The chromosomal AA, BB, and AB contact intensities were determined by averaging the frequencies located in the top 20% left, bottom right, and bottom left quadrants of the saddle plots, respectively. Chromosomal compartment scores were then derived using the formula $(AB) - \frac{(AA)+(BB)}{2}$. To generate the genome-wide saddle plot, each chromosomal saddle plot was scaled to a 1000×1000 matrix, followed by calculating the average across all chromosomes. The global AA, AB, and BB contact intensities were ascertained by taking the medians of the respective chromosomal AA, AB, and BB contact intensities.

Determination of TADs and their boundaries based on insulation score profiles

The insulation score profiles were computed following the methodology outlined in[33], utilizing a sliding window size of 250 kbp. To ensure comparability across profiles from different chromosomes, each profile was normalized by dividing it with the respective chromosome's average values, followed by a log transformation. The insulation score values were subsequently modeled using a three-component Gaussian mixture model to categorize them into strong boundaries, weak boundaries, and TAD interiors. Model parameters were estimated using the Expectation-Maximization (EM) algorithm. The role of each bin was then determined based on the role with the highest posterior probability. Bins inferred as boundaries (both strong and weak) that were also local minimum points of the insulation scores within their 21-bin sized neighborhoods were designated as candidate boundaries. Adjacent boundaries were merged to form a TAD if more than 50% of the bins between them were classified as weak boundaries or interiors.

The consistency between TADs across different conditions was determined based on their positions. For a TAD labeled as A_1 in condition A, if there existed a unique TAD labeled as B_1 that satisfied the following criteria:

$$\begin{cases} \frac{|A_1 \cap B_1|}{|A_1|} > 0.8 \\ \frac{|A_1 \cap B_1|}{|B_1|} > 0.8 \end{cases}$$

Then A_1 and B_1 were considered the same. Otherwise, if there was a series of domains $B_1, B_2, \cdots, B_n$ in B that satisfied

$$\frac{|A_1 \cap (\bigcup_{i=1}^{n} B_i)|}{|A_1|} > 0.8$$

Then A_1 was considered as a merged domain in condition B. Furthermore, if there was a series of domains $A_2, A_3, \cdots, A_n$ in condition A and a domain B_1 in condition B that satisfied:

$$\frac{|B_1 \cap (\bigcup_{i=1}^{n} A_i)|}{|B_1|} > 0.8$$

Then A_1 was considered as a split event in condition B. TADs that did not fit into any of the above scenarios were classified as shift events in condition B. Here, $|*|$ denoted the length of the corresponding genome region.

Loop aggregated peak analysis

To perform the APA analysis, we first filtered out loops in the given list that had anchor distances less than or equal to 75 kbp (15 bins) to minimize the influence of extremely high contact frequencies near the diagonals of contact maps. For each loop involving bin a and bin b (where $a < b$), we calculated local contact frequencies by taking a 21×21 sub-matrix $M[(a-10):(a+10),(b-10):(b+10)]$ from the log-transformed, observed/expected normalized, KR-balanced contact matrix M. The average APA profile was then computed by averaging these local contact frequencies. Next, we calculated the APA z-scores for each loop using the following formula:

$$z = \frac{M[a,b] - mean(M[(a+6):(a+10),(b-10):(b-6)])}{std(M[(a+6):(a+10),(b-10):(b-6)])}$$

After obtaining the loop-wise z-scores, we calculated a global APA z-score by averaging all the individual z-scores.

Determination of CTCF involved loops

To identify loops associated with CTCF, we first identified significant CTCF motif hits using FIMO, a tool within the MEME suite[46]. We utilized a position weight matrix obtained from the JASPAR database[47] and set a p-value threshold of 1×10^{-5}. Hits that were not covered by a CTCF ChIP-seq binding peak were deemed not to be bound by CTCF in the corresponding cell type and were excluded from further analysis.

Loops were then categorized based on the number of anchors (0, 1, or 2) occupied by at least one significant CTCF hit. To assess the relationship between these classifications and crosslinking conditions, we employed a contingency table χ^2 test. This statistical approach allowed us to verify the association between CTCF occupancy and loop formation under different crosslinking conditions.

Enrichment of numbers of and E/P loops and CTCF mediated loops

Loops involving E / P (Enhancer / Promoter) were identified as those with anchors occupied by intergenic enhancers, as cataloged in EnhancerAtlas[37], or promoters of genes that are expressed with a TPM (Transcripts Per Million) greater than 1. These E / P involved loops were further categorized into P / P, E / P, and E / E classes, depending on whether both anchors were occupied by promoters, one by a promoter and the other by an enhancer, or both by enhancers, respectively. As mentioned above, CTCF-mediated loops were identified based on the occupation of loop anchors by a significant CTCF binding motif hit. The motif was sourced from[47], and the hits were scanned using FIMO within the MEME suite[46]. Loops were classified according to the number of anchors occupied by such qualified CTCF binding sites. To analyze the data, the counts of loops across all crosslinking conditions were organized into a contingency table. In this table, loop statuses—whether involved in E/P or not, or mediated by CTCF or not—were listed as rows, while the crosslinking conditions were presented as columns. The independence between loop status and crosslinking condition was assessed using Pearson's χ^2 test. Expected frequencies were computed assuming the independence of rows and columns and then divided by the observed counts to obtain an observed/expected ratio, which indicates the directionality of enrichment.

The conceptual model for FA crosslinking

To conceptually elucidate the impact of crosslinking strength on the outcomes of chromatin conformation capture assays, we employed a simplified model. This model considered only two enzyme cutting ends and constrained their positions to a one-dimensional space. The central position of the first end was fixed at the origin, while the second end was randomly positioned at coordinate D following a Cauchy distribution with density

function:

$$f(D) = \frac{1}{\pi(1 + D^2)}.$$

Once D was determined, the actual positions of the two ends after crosslinking (denoted as x_1 and x_2) were determined by two independent Gaussian distributions with equal standard deviations:

$$\begin{cases} x_1 \sim N(0, \sigma^2) \\ x_2|D \sim N(D, \sigma^2) \end{cases}.$$

The common standard deviation σ represented the crosslinking strength, where smaller values of σ indicated higher crosslinking strength. Whether the two ends were ligated (denoted as δ) was then determined by a Bernoulli distribution with probability $e^{-|x_1 - x_2|}$. The contact frequency decay curve could thus be expressed as the conditional distribution of D given $\delta = 1$.

To simulate the contact frequency decay curve and assess its stability, we configured the model with three σ values: 0.1, 0.5, and 1.0, representing high, medium, and low crosslinking scenarios, respectively. For each σ value, we continued simulations until 1×10^7 ligations ($\delta = 1$) were obtained. Contact frequency decay curves were then estimated using Gaussian kernel density estimation as a function of D. This simulation procedure was repeated 500 times. The coefficient of variation (CV) values for the 500 contact frequencies at each D were calculated to measure stability (signal-to-noise ratio).

Public data analysis

The public ChIP-seq, ATAC-seq, polyA+ RNA-seq and bisulfite sequencing datasets were all listed in (Supplementary Table 1). The data processing pipelines of them were all listed in the (Supplementary text).

Statistics and reproducibility

The statistical tests utilized and the significance threshold values are delineated in detail within the "Methods" section. For each cell type and crosslinking condition, two biological replicates were incorporated. These replicates were characterized by high reproducibility scores of contact maps, as computed by GenomeDISCO.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Hi-C data were deposited into the Genome Sequence Archive database in the National Genomics Data Center under accession numbers HRA003237 and HRA006154 and are available at the following URL: https://ngdc.cncb.ac.cn/gsa-human/browse/HRA003237 and https://ngdc.cncb.ac.cn/gsa-human/browse/HRA006154. The source data for graphs were deposited into the gitee repository at https://gitee.com/matrix_evolution/crosslinking-paper.

Code availability

The code used in this study are freely accessible at https://gitee.com/matrix_evolution/crosslinking-paper.

Received: 27 March 2024; Accepted: 16 September 2024;
Published online: 30 September 2024

References

1. Denker, A. & de Laat, W. The second decade of 3C technologies: detailed insights into nuclear organization. *Genes Dev.* **30**, 1357–1382 (2016).
2. Smallwood, A. & Ren, B. Genome organization and long-range regulation of gene expression by enhancers. *Curr. Opin. Cell Biol.* **25**, 387–394 (2013).
3. Baxter, J. S. et al. Capture Hi-C identifies putative target genes at 33 breast cancer risk loci. *Nat. Commun.* **9**, 1028 (2018).
4. Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nat. Rev. Genet.* **17**, 772 (2016).
5. Meaburn, K. J. & Misteli, T. Cell biology: chromosome territories. *Nature* **445**, 379–781 (2007).
6. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
7. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
8. Fudenberg, G. et al. Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep.* **15**, 2038–2049 (2016).
9. Kaplan, N. & Dekker, J. High-throughput genome scaffolding from in vivo DNA interaction frequency. *Nat. Biotechnol.* **31**, 1143–1147 (2013).
10. Kronenberg, Z. N. et al. Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C. *Nat. Commun.* **12**, 1935 (2021).
11. Xu, Z. & Dixon, J. R. Genome reconstruction and haplotype phasing using chromosome conformation capture methodologies. *Brief. Funct. Genomics* **19**, 139–150 (2020).
12. Fraenkel-Conrat, H. & Olcott, H. S. Reaction of formaldehyde with proteins; participation of the guanidyl groups and evidence of crosslinking. *J. Am. Chem. Soc.* **68**, 34–37 (1946).
13. French, D. & Edsall, J. T. The Reactions of Formaldehyde with Amino Acids and Proteins. *Adv. Protein Chem.* **2**, 277–335 (1945).
14. Sutherland, B. W., Toews, J. & Kast, J. Utility of formaldehyde cross-linking and mass spectrometry in the study of protein-protein interactions. *J. Mass Spectrom.* **43**, 699–715 (2008).
15. Oksuz, Akgol et al. Systematic evaluation of chromosome conformation capture assays. *Nat. Methods* **18**, 1046–1055 (2021).
16. Nagano, T. et al. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* **547**, 61–67 (2017).
17. Yang, H. et al. A map of cis-regulatory elements and 3D genome structures in zebrafish. *Nature* **588**, 337–343 (2020).
18. Xia, Y. et al. Capturing 3D Chromatin Maps of Human Primary Monocytes: Insights From High-Resolution Hi-C. *Front Immunol.* **13**, 837336 (2022).
19. Krijger, P. H. L., Geeven, G., Bianchi, V., Hilvering, C. R. E. & de Laat, W. 4C-seq from beginning to end: A detailed protocol for sample preparation and data analysis. *Methods* **170**, 17–32 (2020).
20. Nagano, T. et al. Single-cell Hi-C for genome-wide detection of chromatin interactions that occur simultaneously in a single cell. *Nat. Protoc.* **10**, 1986–2003 (2015).
21. Jager, R. et al. Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat. Commun.* **6**, 6178 (2015).
22. Melo, U. S. et al. Hi-C Identifies Complex Genomic Rearrangements and TAD-Shuffling in Developmental Diseases. *Am. J. Hum. Genet.* **106**, 872–884 (2020).
23. Nagano, T. et al. Comparison of Hi-C results using in-solution versus in-nucleus ligation. *Genome Biol.* **16**, 175 (2015).
24. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
25. Li, L. et al. Widespread rearrangement of 3D chromatin organization underlies polycomb-mediated stress-induced silencing. *Mol. Cell* **58**, 216–231 (2015).
26. Ray, J. et al. Chromatin conformation remains stable upon extensive transcriptional changes driven by heat shock. *Proc. Natl. Acad. Sci. USA* **116**, 19431–19439 (2019).

27. Marchal, C., Singh, N., Corso-Díaz, X. & Swaroop, A. HiCRes: a computational method to estimate and predict the genomic resolution of Hi-C libraries. *Nucleic Acids Res*. **50**, e35 (2022).
28. Ke, Y. et al. 3D Chromatin Structures of Mature Gametes and Structural Reprogramming during Mammalian Embryogenesis. *Cell* **170**, 367–381.e20 (2017).
29. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
30. Liang, Z. et al. BL-Hi-C is an efficient and sensitive approach for capturing structural and regulatory chromatin interactions. *Nat. Commun.* **8**, 1622 (2017).
31. Ursu, O. et al. GenomeDISCO: a concordance score for chromosome conformation capture experiments using random walks on contact map graphs. *Bioinformatics* **34**, 2701–2707 (2018).
32. Belaghzal, H. et al. Liquid chromatin Hi-C characterizes compartment-dependent chromatin interaction dynamics. *Nat. Genet.* **53**, 367–378 (2021).
33. Crane, E. et al. Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* **523**, 240–244 (2015).
34. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
35. Xu, B. et al. Cell cycle arrest explains the observed bulk 3D genomic alterations in response to long-term heat shock in K562 cells. *Genome Res.* **32**, 1285–1297 (2022).
36. Roayaei Ardakany, A., Gezer, H. T., Lonardi, S. & Ay, F. Mustache: multi-scale detection of chromatin loops from Hi-C and Micro-C maps using scale-space representation. *Genome Biol.* **21**, 256 (2020).
37. Gao, T. & Qian, J. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res.* **48**, D58–d64 (2020).
38. Mirny, L. A. The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res.* **19**, 37–51 (2011).
39. Ghavi-Helm, Y. et al. Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nat. Genet.* **51**, 1272–1282 (2019).
40. Tian, B., Yang, J. & Brasier, A. R. Two-step cross-linking for analysis of protein-chromatin interactions. *Methods Mol. Biol.* **809**, 105–120 (2012).
41. Li, X. et al. Long-read ChIA-PET for base-pair-resolution mapping of haplotype-specific chromatin interactions. *Nat. Protoc.* **12**, 899–915 (2017).
42. Benedetti, F., Dorier, J., Burnier, Y. & Stasiak, A. Models that include supercoiling of topological domains reproduce several known features of interphase chromosomes. *Nucleic Acids Res*. **42**, 2848–2855 (2014).
43. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
44. Durand, N. C. et al. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst.* **3**, 95–98 (2016).
45. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
46. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
47. Castro-Mondragon, J. A. et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **50**, D165–d173 (2022).

Acknowledgements

This work was supported by the Beijing Natural Science Foundation (Z200021 to Z.Z.H.), the Strategic Priority Research Program of CAS (XDA24020307 to Z.Z.H.), the National Natural Science Foundation of China (32200515 to X.B.X and 32341011 to Z.Z.H.), the Science and Technology Innovation 2030 – Major Project (2022ZD04017 to Z.Z.H.) and the National Key R & D Program of China (2020YFA0509500 to Z.Z.H.).

Author contributions

X.B.X. and Z.Z.H. conceived this project. G.X.M. and L.X.L. performed the experiments. X.B.X. and L.F.F. analyzed data. X.B.X., L.F.F. and Z.Z.H. prepared the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-024-06904-0.

Correspondence and requests for materials should be addressed to Bingxiang Xu, Feifei Li or Zhihua Zhang.

Peer review information *Communications Biology* thanks Sarah G. Swygert and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Joanna Timmins and Laura Rodríguez Perez. A peer review file is available.