Published online 3 October 2017 *Nucleic Acids Research, 2017, Vol. 45, No. 22* ***12739–12751***
doi: 10.1093/nar/gkx885

Characteristic arrangement of nucleosomes is predictive of chromatin interactions at kilobase resolution

Hui Zhang[1,2,†], Feifei Li[1,†], Yan Jia[1], Bingxiang Xu[1,2], Yiqun Zhang[1,2], Xiaoli Li[1,2] and Zhihua Zhang[1,2,*]

[1]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China and [2]University of Chinese Academy of Sciences, Beijing 100049, China

Received February 22, 2017; Revised September 17, 2017; Editorial Decision September 18, 2017; Accepted September 24, 2017

ABSTRACT

High-throughput chromosome conformation capture (3C) technologies, such as Hi-C, have made it possible to survey 3D genome structure. However, obtaining 3D profiles at kilobase resolution at low cost remains a major challenge. Therefore, we herein present an algorithm for precise identification of chromatin interaction sites at kilobase resolution from MNase-seq data, termed chromatin interaction site detector (CISD), and a CISD-based chromatin loop predictor (CISD_loop) that predicts chromatin–chromatin interactions (CCIs) from low-resolution Hi-C data. We show that the predictions of CISD and CISD_loop overlap closely with chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) anchors and loops, respectively. The validity of CISD/CISD_loop was further supported by a 3C assay at about 5 kb resolution. Finally, we demonstrate that only modest amounts of MNase-seq and Hi-C data are sufficient to achieve ultrahigh resolution CCI maps. Our results suggest that CCIs may result in characteristic nucleosomes arrangement patterns flanking the interaction sites, and our algorithms may facilitate precise and systematic investigations of CCIs on a larger scale than hitherto have been possible.

INTRODUCTION

The 3D genome architecture underlies many cellular processes in the nucleus (1–3). Proximity ligation-based chromosome conformation capture (3C) and its variations (4–6) constitute a major engine driving the exploration of the 3D genome architecture (7,8). Using one genome-wide version of the 3C technology, Hi-C, it has been possible to explore the global 3D architecture of the human (9,10), mouse (11,12), fly (13) and yeast genomes (14). Mediator-specific 3D chromatin interaction maps have also been produced by the ChIA-PET method in mammals for such proteins as CTCF (15,16), Pol II (16,17), cohesin (18) and histone modifications (19). With these mappings, genomes were found to be physically separated into two compartments (A and B), one active and the other inactive (9). Higher resolution mapping could reveal finer structures. For example, with increased Hi-C resolution, the so-called 'topologically associated domains' (TADs) (11,13,20,21) and sub-TAD (22) structures in mammals have been identified, along with detailed chromatin fiber looping structures (16,23). However, because most *cis*-regulatory sequences are in the size range of hundreds of base-pairs and may be closely clustered (24), precise definition of individual enhancer-promoter interactions on a genome-wide scale remains beyond the capacity of current Hi-C methodology.

At least four major obstacles hamper the construction of high-resolution chromatin interaction maps by current Hi-C technology. First, it remains prohibitively expensive to substantially increase Hi-C resolution solely by deep sequencing (7). The number of reads needed to increase resolution will necessarily grow exponentially, and this number is already running into the billions at a resolution of 1 kb (23). Second, a theoretical limit is defined by the natural density of restriction enzyme sites in the genomes (25). Third, the efficiency of experimental steps in the 3C protocol varies from site to site along the genome, making it difficult to find optimal conditions and controls for genome-wide assays (26), and this also require further work on the normalization of the data (27–29). Fourth, the ligation step in 3C is always subject to the problem of crosslinking. A recently detailed fluorescence *in situ* hybridization (FISH) and 3C analysis at the murine *HoxD* locus showed that the two technologies do not produce concordant results, imply-

[*]To whom correspondence should be addressed. Tel: +86 10 8409 7249; Fax: +86 10 8409 7720; Email: zhangzhihua@big.ac.cn
[†]These authors contributed equally to the paper as first authors.
Disclaimer: The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

ing that a high density of 3C signal does not always reflect spatial proximity (30).

Attempts have been made to improve the resolution of chromatin–chromatin interaction (CCI) maps (15,16,23,26,31–35). Capture-based methods (15,16,26,31–33), or alternative DNA cutters (34,35), have been described in the literature. For example, Duan *et al*. replaced restriction enzymes with DNase I (34), and obtained better genome coverage and resolution than normal Hi-C. However, as a result of inherent limitations of the current 3C-based protocols, it remains challenging to substantially increase comprehensive mapping resolution to a level beyond 1–2 kb within reasonable cost constraints (23,26).

Chromatin 3D architecture is associated with various epigenetic features. For example, by comparing the ChIA-PET map with DNase-seq, ChIP-seq and RNA-seq datasets, Snyder *et al*. showed that there is a strong association between CCIs and chromatin accessibility (19). Computational models have also been developed to investigate associations between histone marks and A/B compartments (36), CCI hubs and TADs (37). More recently, methods that integrate high-dimensional multi-omics data in multiple cell types to predict CCIs have also been reported (38–41). However, the massive cross-cell types and multi-omics data required by these methods have made it difficult to detect cell-type specific interactions, or to elucidate the underlying mechanisms linking CCIs to chromatin dynamics.

As nucleosomes are the basal structural units of chromatin, the arrangement of nucleosomes could carry fundamental information about the chromatin dynamics. Although nucleosomes have been shown to have strong DNA sequence preference (42), their arrangement is highly dynamic, i.e. subject to either active remodeling by adenosine triphosphate-dependent remodeling enzymes (43), or passive remodeling by stochastically aligning to bound transcription factors (TFs) (44). Genomic events or features, e.g., stably bound TFs, the end of a heterochromatin domain, or simply a nucleosome-free DNA region (NFR), are sufficient to cause statistical phasing of a considerable portion of the nucleosomes (44,45). Moreover, the phasing patterns of the nucleosomes vary considerably among bound TFs (46,47). For example, Sun and colleagues found that the nucleosome profiles of TF binding sites could be classified into tens of clusters (47). These differences may reflect the local chromatin context among the TFs. Based on these facts we propose a hypothesis that physical CCIs could significantly alter the local chromatin context, resulting in characteristic nucleosomes arrangement patterns flanking the CCI sites. To test the hypothesis, we examined nucleosome arrangements flanking CCI sites. Distinct arrangements of nucleosomes flanking the binding sites of CCI-associated factors, e.g., CTCF, have been well documented (48). Moreover, we found distinct differences in the nucleosomes arrangement patterns between the interacting allele compared to the non-interacting allele at allele-specific CCI sites.

Based on this hypothesis, we developed two computational algorithms, named CISD (chromatin interaction site detector) and CISD_loop (CISD-based chromatin loop predictor) that respectively predict CCI sites and CCIs at kilobase resolution. CISD and CISD_loop only require low-resolution MNase-seq data and low-resolution Hi-C input, respectively. We show that the predictions of CISD and CISD_loop are enriched for ChIA-PET anchors and loops, respectively. We performed 3C experiments at 5kb resolution and validated CISD_loop predictions that have not been reported in the ChIA-PET data. Because the algorithms trained in one cell type can be applied to other cell types with high accuracy, the association between the characteristic nucleosomes arrangement pattern and the CCI sites may be universal in human cells. The power of characteristic nucleosome arrangement patterns to predict CCIs further supports our initial hypothesis. Finally, by saturation analysis, we show that only moderate amounts of MNase-seq and Hi-C data are sufficient to achieve ultra-high resolution CCI maps.

MATERIALS AND METHODS

Data

Data that were downloaded from public domain are listed in the Supplementary Text.

Chromatin interaction site detector (CISD)

Basically, CISD determines whether the nucleosomes arrangement pattern in a given genome locus is a pattern that is characteristic of chromatin interactions. The algorithm can be largely separated into a data preparation section and two model training sections (Figure 2A).

Data preparation. Here, we convert MNase-seq data into the frequency spectrum by fast Fourier transform (FFT). CISD first smooths the input MNase-seq reads. For any given genomic region (1 kb-long in this study), the mapped reads are binned into 10 bp-long bins, resulting in an n-dimensional vector V. V is then fed into the iNPS for denoising and smoothing (49). The iNPS is an improved version of NPS (50); both iNPS and NPS denoise and smooth the wave-form signal by Laplacian of Gaussian convolution (LoG) (49,50). After the LoG, the frequency feature of the original data is preserved, while its direct current component is substantially reduced. The denoised and smoothed V (denoted as V') is further normalized by dividing the standard deviation of V' over the whole genome, and it was denoted as $\tilde{V} = \{\tilde{v}_j\}$, $j = 0, 2, ..., n-1$. After data denoising, smoothing and normalization, CISD converts the linear data into frequency space. To do this, for any given $\tilde{V}$, FFT is applied to retrieve the frequency information. In general, FFT is a fast computational method for Discrete Fourier Transformation (DFT). The DFT converts an n-dimensional vector $\tilde{V}$ of complex numbers into a complex number vector $C = \{c_j\}$, $j = 0, 2, 3, ..., n-1$,

$$c_j = \sum_{t=1}^{n} \tilde{v}_t e^{-it\omega_j},$$

where i is the basic unit of the imaginary number, and $\omega_j = 2\pi j/n$. Since (i) C is conjugate symmetric and (ii) we are only interested in the modulus of C, we discarded that half of the elements in C where the real parts are negative. We thus arrive at the definition of the FFT profile (denoted as F) of the nucleosomes arrangement in a genome segment such that

$$F = \{F_j = \|c_j\|\},\ j = 0, 2, 3, \ldots, \lceil n/2 \rceil\ ,$$

where $\|c_j\|$denotes the modulus of c_j.

Model training one: periodic region detection. We define a metric for the periodicity of a given genomic region by a logistic regression model (LRM). To train the LRM, we constructed a positive and a negative dataset containing randomly selected 10 000 CTCF/cohesin co-binding of the ChIA-PET anchors and the randomly selected 10 000 control genomic regions, respectively, in the GM12878 and K562 cell types. In the positive dataset, the co-binding of CTCF and RAD21 is inferred by the ChIA-PET anchors. CTCF and cohesin are considered to be co-binding if the two ChIA-PET anchors overlap by more than one base pair. The negative set is composed of genomic segments that are located at least 5kb away from the CTCF, cohesin and ZNF143 binding sites and are also not in promoter regions, i.e. at least 5 kb away from UCSC annotated transcription start sites. We chose F_0, F_5 and F_6 in the FFT profiles (F_j) as the features with which to train the LRM. The LRM was trained by R. An artificial threshold of LRM score (**periodic score**) was then chosen to determine if the input genomic segment is carrying a periodic nucleosome pattern. The LRM was assessed by receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC), which are both drawn and calculated by the R package 'pROC'. In this paper, we took 0.5 as the threshold and applied the LRM to the whole genome, denoting the determined periodic nucleosomal regions as high score peaks (**HSPeaks**) to be used as input for the next step.

Model training two: interaction site detection. As we have shown above, not all periodic nucleosomal regions are associated with chromatin interactions. Accordingly, we trained a support vector machine (SVM) to further distinguish interactive loci from the remaining periodic nucleosomal regions. We took the full frequency spectrum (F) as the feature. To train the SVM, we constructed a positive and a negative dataset from the HSPeaks. The positive sets consisted of overlapping CTCF and cohesin ChIA-PET anchors, while the negative set was randomly sampled from a subset of HSPeaks that did not overlap any ChIA-PET anchors. The SVM model was implemented by R-package 'e1071' with default parameter settings.

CISD-based Chromatin loop predictor (CISD_loop)

CISD_loop is a method for determination of intra-TAD chromatin loops between CISD sites. In addition to CISD, TAD annotation and raw Hi-C reads are required for the operation of CISD_loop. The TADs annotated in hESC were used as the reference for all human cells, as the TAD structure is believed to be largely consistent among tissue types (11). CISD_loop was trained according to the following procedure. First, we constructed a 'total' dataset composed of all intra-TAD CISD site pairs. Then, the training and testing datasets were drawn from the 'total' dataset so that the positive set (5000 data points in this work) was composed of the CISD pairs that overlapped with ChIA-PET loops, and the negative set was an identical number of randomly sampled CISD pairs from the remaining data. A voting system with four decision tree based voters, GBDT (51), random forest (51), ExtraTrees (51) and CART (51) were composed to form CISD_loop. The rationale of our modeling choices is that physically contacting genome loci may be experiencing similar biology processes, e.g. located in the same transcription factory, so that their biophysical or biochemical characters may be covariant (38,39), while the decision tree based methods are convenient to capture such covariant. A CISD site pair is predicted as a CCI loop only if all four models predict it is positive. This is because we found that the more votes a prediction get, the more likely it is to be supported by ChIA-PET data (Supplementary Figure S1). The models were implemented by the Python-package 'sklearn' (51).

There are three categories of features been used by CISD_loop. The first category is the FFT profiles of the nucleosome pattern. These were selected to capture the characteristic biophysical features that may be covariant between interacting CISD site pairs. There are 50 FFT profile components for each CISD site, so in total, there are 100 features in this feature category for a pair of CISD sites. We compared the correlation between positive and negative CISD pairs, and indeed, the Pearson's correlation coefficients of the FFT profiles in the positive set are higher than that in negative set (Supplementary Figure S2). These data indicate that the covariant of the FFT profiles indeed carries information about the CCIs.

The second feature category is the Hi-C map. The Hi-C experiment directly measures the contact frequency between genome loci. Although, high resolution Hi-C maps remain prohibitively expensive, low resolution Hi-C is now becoming more accessible to ordinary laboratories. Thus, we included Hi-C maps into the feature list of CISD_loop. The two most prominent characters of a chromatin loop in a Hi-C map are (i) that the absolute contact frequencies between the looping anchors are high, and (ii), that the relative contact frequencies between the looping anchors are higher compared to their flanking regions. Thus, we created a Hi-C contact index and their differences to model characters (i) and (ii), respectively. The Hi-C contact index was defined as the sum of the normalized Hi-C reads counts between a given pair of genome loci. For any given pair of CISD sites, various levels of the Hi-C contact index can be defined with different ranges. For example, as shown in Supplementary Figure S3, the sum of normalized Hi-C reads counts in the yellow, green and blue diamond boxes were defined as the Hi-C contact indices at levels 1, 2 and 3, respectively, for pairs of loci B_i and B_j. The default Hi-C map used was at 5kb resolution, however, this resolution is not mandatory. The features used in CISD_loop were Hi-C contact indices at levels 1, 2 and 3, and their differences.

The third category feature is the genomic linear distance between two given loci. The distance is critical to CCIs. Both empirical data and theoretical polymer models have shown that the probability of interaction is a function of the genomic distance, i.e. the closer the loci the higher the chance of contact between them. Moreover, the majority of

CCIs are found to be intra-TAD. These facts motivated us to include the genomic distance as a feature for CISD_loop. Finally, we only considered the intra-TAD CISD pairs in CISD_loop. In total, there are 106 features that were employed in CISD_loop.

Determination of allele-specific MNase-seq reads

We downloaded the most updated phased SNPs of the GM12878 cell line in the 1000 Genomes Project (52) from Gerstein's lab (53). By overlapping the mapped reads with phased SNPs, we identified 8 382 815 and 8 456 093 paternal-specific and maternal-specific MNase-seq reads, respectively.

Determination of allele-specific CCIs

We took the ChIA-PET loop anchors in the GM12878 cell line as the gold standard CCI sites (54). However, the anchor lengths were too short (mean = 424 bp) to identify a sufficient number of SNPs. Therefore, instead of using ChIA-PET data, we pooled all Hi-C reads that mapped within a 5 kb genome region flanking each ChIA-PET loop anchor (23). We used the ratio of the maternal-specific reads number to the paternal-specific reads number to index the allele specificity of the CCIs in each region. CCIs in chromosome X were omitted. As the index has a bell shape distribution (Supplementary Figure S4), we took the first and last 10% as maternal- and paternal-specific CCIs anchors, respectively. We further filtered out CCIs with abnormally high numbers of reads.

Aggregation analysis on allele-specific MNase-seq data

For each individual interaction loci, there are, on average, only about 6.5 reads carrying SNPs that can be used to determine their allelic origin. It is nearly impossible to calculate a meaningful periodic score from only 6.5 reads. The purpose of this analysis is to aggregate information from sporadic allele-specific MNase-seq reads to reveal possible general patterns of nucleosome arrangements at allele-specific CCI sites. The rationale behind this procedure is that, if the periodic nucleosome arrangement pattern we discovered at CCI sites is indeed universal, we shall expect to see the pattern also in all allele-specific CCI sites. Thus, if we pool all MNase-seq reads which carrying an SNP from, for example, the paternal-specific CCI loci, and align them together according to the relative distance to the middle point of each loci, the pooled reads form a single 'virtual' locus. We expect to see the periodic 'virtual' nucleosome pattern in the paternal allele but not in maternal allele. Using the sequencing depth in Rao *et al.*'s data, which is about 5500 reads per 10 kb region, we sampled 5500 allele-specific reads for a 'virtual' locus. Because the total number of allele-specific reads is much larger than 5500, we can repeat this process to calculate periodic scores for a series of 'virtual' loci, in order to draw the plot Figure 1E.

As an example, we describe how the information of nucleosome arrangements in the maternal allele at paternal-specific CCI sites is aggregated. We first collected all maternal-specific reads in the 10 kb flanking regions of all the paternal-specific CCI sites, denoted as $ASR(m, p)$. By sampling 5500 reads with replacement from the $ASR(m, p)$, we could generate a 'virtual' allele-specific sequencing dataset, denoted as VD. We then set the CTCF motifs found in each CCI site as the 'virtual' origin, and map each of the reads in VD into the 'virtual' locus according to the relative distance from the 'virtual' origin. After the mapping, the periodic score could then be calculated as described below. We generated 200 virtual datasets for $ASR(m, p)$ and drew boxplots of the periodic scores as in Figure 1E.

Hi-C data normalization

Our normalization method is composed of the following two steps. The first step is to make the matrix balanced (23). In this step, the raw contact matrix was balanced by the KR normalization algorithm used by Rao *et al.* (23). The step corrects possible experimental bias also considered by other normalization methods, such as HiCNorm (29). The second step is to rescale the normalized matrix so that the matrices are comparable between different experiments. This was achieved by point division of an expectation matrix, i.e., each element of the contact matrix is divided by the corresponding element of the expectation matrix. The expectation matrix was prepared as described by Rao *et al.* (23).

Code availability. Source code for the CISD and CISD_loop can be found at https://github.com/huizhangucas/CISD.

Evaluation of the models. In all complete datasets, we had identical numbers of entries in the positive and negative sets. The accuracy of a model was defined as (TP + TN)/(TP + TN + FP + FN), where TF, TN, FF and FN are true positives, true negatives, false positives and false negatives, respectively. Standard 5-fold cross-validation was performed according to the following procedure. The original total data sample was randomly partitioned into five subsets of equal size. Of the five subsets, a single subset was retained as the testing data, and the remaining four subsets were used as training data. We repeated this procedure five times with each of the five subsets used exactly once as the testing data. The five results were averaged for the final evaluation.

RESULTS

Distinct nucleosome arrangement patterns at CCI sites

It is herein proposed that physical CCIs may alter the local chromatin context, which, in turn, causes rearrangement of the nucleosomes around the interaction sites and a resulting distinct pattern of nucleosomes. To test this hypothesis, we plotted MNase-seq signals flanking the binding sites of two CCI mediating TFs (Rad21, CTCF) and two randomly picked TFs (NFYB and KAP1) in K562 cells (Figure 1A). The nucleosome pattern varies among the TFs, with CTCF and RAD21 show a strong periodic arrangement. Then we extended our analysis to all 99 TFs for which ChIP-seq data are available for K562 cells in the ENCODE project (55). Principal component analysis (PCA) of MNase-seq signals flanking the ChIP-seq peaks of the 99 TFs resulted in isolating five TFs (CTCF, RAD21, SMC3, ZNF143 and

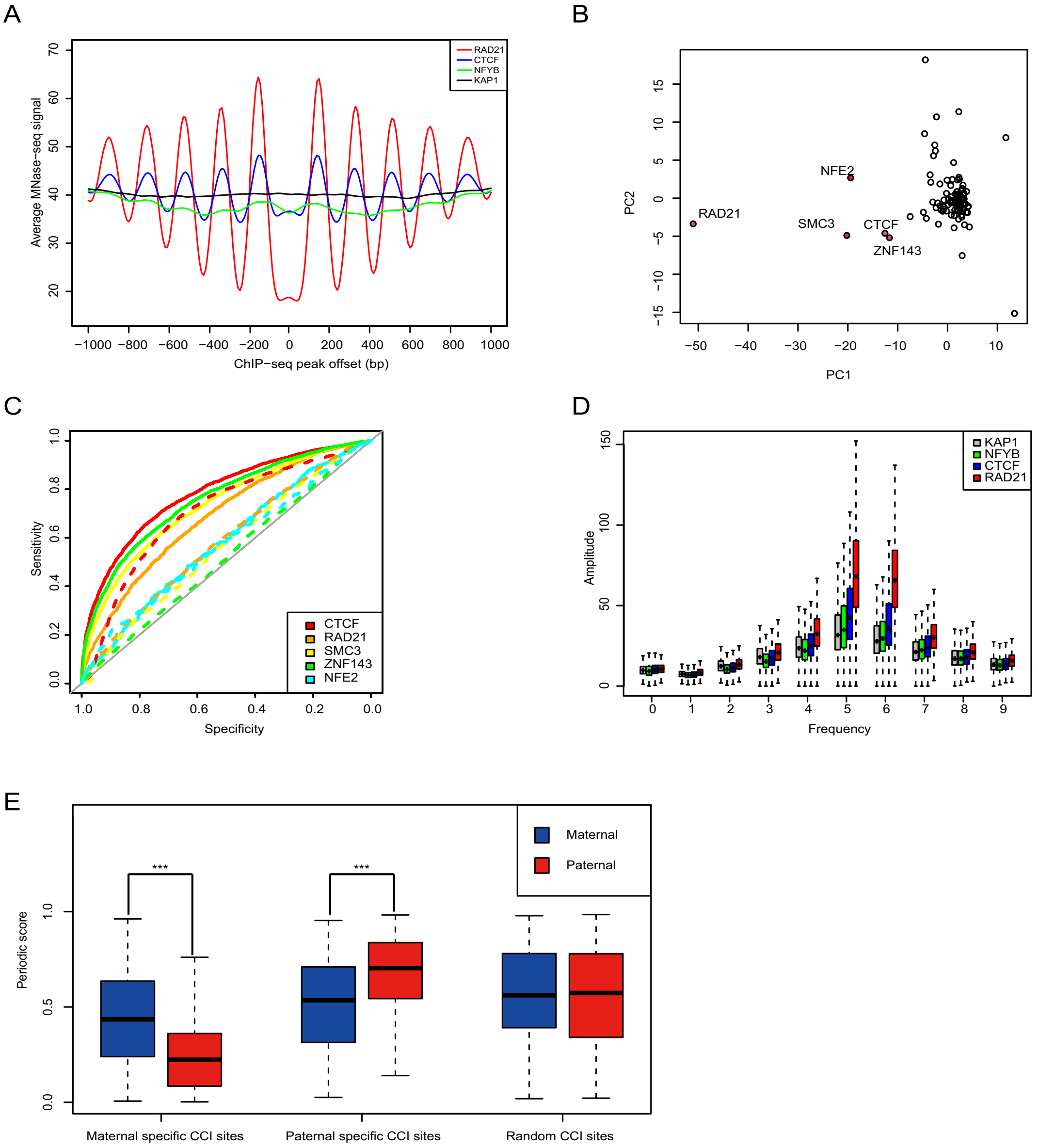

Figure 1. Characteristic nucleosomes arrangement patterns flank the binding sites of major CCI-mediating proteins. (**A**) Distribution of MNase-seq reads flanking the ChIP-seq peaks of Rad21, CTCF, NFYB and KAP1 in K562. (**B**) PCA analysis of MNase-seq signals flanking ChIP-seq peaks of 99 TFs. Five TFs separated from the remaining TFs are marked in red. (**C**) ROC curves for CCI predictions using the ChIP-seq signal strength and the periodic score, represented as dashed and solid lines, respectively. The ChIA-PET loop anchors were taken as the golden standard to assess the two prediction methods. (**D**) The FFT profiles of MNase-seq signals flanking the ChIP-seq peaks of the four proteins in (A). (**E**) Aggregation analysis of allele-specific MNase-seq data at allele-specific CCI sites. Each boxplot represents the distribution of periodic scores from 200 virtual datasets. ***: rank sum test *P*-value < 1e-10.

NFE2) from the others (Figure 1B). Notably, CTCF, cohesin (RAD21 and SMC3) and ZNF143 are well known CCI-associated proteins (18,19,56). NFE2 is a chromatin remodeler and also reported to be engaged in enhancer–promoter interactions (57). However, the binding strength of the five TFs alone is not sufficient to predict CCI sites (Figure 1C). For example, we collected ChIA-PET data of CTCF, RNAP II, Rad21, H3K4me1/2/3 and H3K27ac in k562 cells, and found that, for any given TF, there was always a considerable portion of ChIA-PET anchors that did not overlap with its ChIP-seq peaks in K562. Overall, 16.3% (5765 of 35 323) of the ChIA-PET anchors did not overlap with any ChIP-seq peaks arising from these five TFs in K562.

To quantify the differences in the nucleosome arrangement patterns (Figure 1B), we compared the fast FFT frequency spectrum of MNase-seq data flanking the ChIP-seq peaks of the four TFs (Figure 1D see 'Materials and Methods' section) and found that the amplitude of the fifth and sixth frequencies was most significantly different between two CCI-associated TFs (CTCF, Rad21) and the others. Thus, by the fifth, sixth and direct components of the frequencies (see 'Materials and Methods' section), we defined a score, termed as the periodic score, to index nucleosome periodicity. Then, we compared the ROC curves of the two predictive indices, the strength of TF bindings and the periodic score of the ChIP-seq peaks, for the CCI prediction (Figure 1C). The ChIA-PET loop anchors were taken as the golden standards for the prediction assessment. We found that the periodic score is a better predictor for ChIA-PET anchors than the binding strength of the TFs (Figure 1C).

To further test our hypothesis, we compared the nucleosome periodic scores between two alleles at the allele-specific CCI sites. Because two sister chromosomes occupy largely mutually exclusive territories in the nucleus (12,58), if the hypothesis is true, we shall expect that allele-specific CCI sites will also have allele-specific nucleosomes arrangement patterns. Accordingly, we examined this prediction in GM12878 cells, which have well-phased, high-density SNP data (52), and 1707 and 1712 maternal- and paternal-specific Rad21 ChIA-PET anchors with CTCF motifs were identified, respectively. However, because of limited SNP density flanking most of allele-specific CCI sites (±1 kb), allele origins could only be assigned to an average of 6.5 MNase-seq reads, making it necessary to perform an aggregation analysis (see 'Materials and Methods' section). Indeed, at the maternal-specific CCI sites, we found that nucleosomes are more periodically arranged in the maternal allele compared to the paternal allele, and *vice versa* for paternal-specific CCI sites. No significant difference can be found in randomly selected CCI sites (Figure 1E). Taken together, the evidence shown above supports our hypothesis and suggests that nucleosome arrangement patterns are reflective of the local chromatin environment and might be utilized for the detection of CCIs.

Detecting chromatin interaction sites at kilobase resolution with CISD

Based on the predictive potential of nucleosome arrangements in the context of local chromatin environment as shown above, we developed the CISD to identify CCI sites. For any given genome locus, CISD takes MNase-seq data as input and determines whether the nucleosomes display the assumed arrangement pattern of a CCI site (hereinafter termed CISD sites when predicted by CISD). First, CISD is composed of an LRM that determines whether the input genome locus has a periodic nucleosome arrangement pattern. If so, a second support vector machine (SVM) model then determines whether the locus has a nucleosome pattern characteristic of CCIs (Figure 2A). To assess the performance of LRM, we plotted the ROC curves and calculated AUC. To draw the ROC, we took the golden standard of the positive data consisting of periodical scores calculated from the overlapping ChIA-PET anchors of CTCF and Rad21, while the negative dataset consisted of the periodical scores calculated from randomly selected sites that were located at least 5 kb away from TSS and ChIP-seq peaks of CTCF, cohesin and ZNF143.The AUC for the LRM was 0.97 and 0.92 in K562 and GM12878 cells, respectively, and 5-fold cross-validations of the accuracy of the SVM model were above 80% in both cell types. We also trained the model in one cell types and tested in the other cell type, and obtained similar results (Supplementary Text and Table 1). The resolution of CISD is defined as the length of the genome segment needed to make a credible prediction. In this work, we took 1 kb as the default resolution. Using the default threshold for the periodic score (0.5), we applied CISD to K562 and GM12878 cells and predicted 22 112 and 26 801 CISD sites, respectively. Several canonical CCI sites were successfully identified by CISD, such as the human β-globin locus in K562 (Figure 2D). The genome-wide distributions of the CISD sites between the two cell types were found to be similar (Supplementary Figure S5a).

To evaluate the performance of CISD, we took ChIA-PET loop anchors as the golden standard as ChIA-PET is believed to identify CCIs at high resolution and confidence (54). We found that CISD can accurately predict ChIA-PET loop anchors, i.e. most of the CISD sites overlapped with ChIA-PET loop anchors (60.0 and 72.3% for K562 and GM12878 cells, respectively). Compared to DNase I hypersensitive sites (DHSs), which were reported to be predictive of chromatin looping anchors (19,41,59), and CTCF ChIP-seq peaks, CISD sites were 2.4-fold and 1.64-fold more enriched for ChIA-PET anchors in K562 cells, respectively (Figure 2B). Results largely corresponding to these were also seen in GM12878 cells (Supplementary Figure S5). The predictions were also well supported by ChIA-PET anchors when we applied CISD to MNase-seq datasets with different degrees of MNase digestion (Supplementary Text and Table S3). The enrichment of ChIA-PET anchors in CISD sites is also true when compared with ChIP-seq peaks of other TFs in both cell types (Supplementary Text and Figure S5). Because the DHSs and CTCF ChIP-seq peaks are ubiquitous in the genome, it is not surprising to find that the total numbers of DHSs and CTCF ChIP-seq peaks overlapping ChIA-PET anchors are larger than that of CISD sites (Figure 2B and Supplementary Figure S5b). We also compared CISD sites to the Hi-C loops reported by Rao et.al (23), and similarly, we found that the accuracy of predicting Hi-C loops anchors by CISD is higher than that of using ChIP-seq peaks of CTCF and DHS (Figure 2B).

To examine the extent to which CISD sites lack support from ChlA-PET data, termed as nonsupported (ns) CISD sites, might be involved in CCIs, we compared the number of Hi-C reads around the nsCISD sites to those around genomic regions that have a high periodic score, but were not predicted as CISD sites (as control sites). Indeed, the nsCISD sites were significantly more enriched for Hi-C reads than were control sites (Figure 2C, *P*-value<2.2e-16, rank sum test). Finally, we chose four nsCISD sites with sufficient high densities of restriction sites, on which to perform 3C experiments. Two of the four CISD sites were verified at 5kb resolution (Supplementary Table S1). Whether the other two CISD sites were involved in a CCI are remains undetermined, as a single 3C experiment only examines one pair of interaction anchors, while there are much more possible interaction targets in the genome.

CISD can detect canonical chromatin interaction sites

The human β-globin locus consists of a locus control region (LCR) comprising five β-globin-like genes (HBE; HBG1 and HBG2; HBD; and HBB) and one pseudogene (HBpsi). Five DNase I hypersensitive sites (HS1-5) have been reported in the LCR and are believed to be required for tissue-

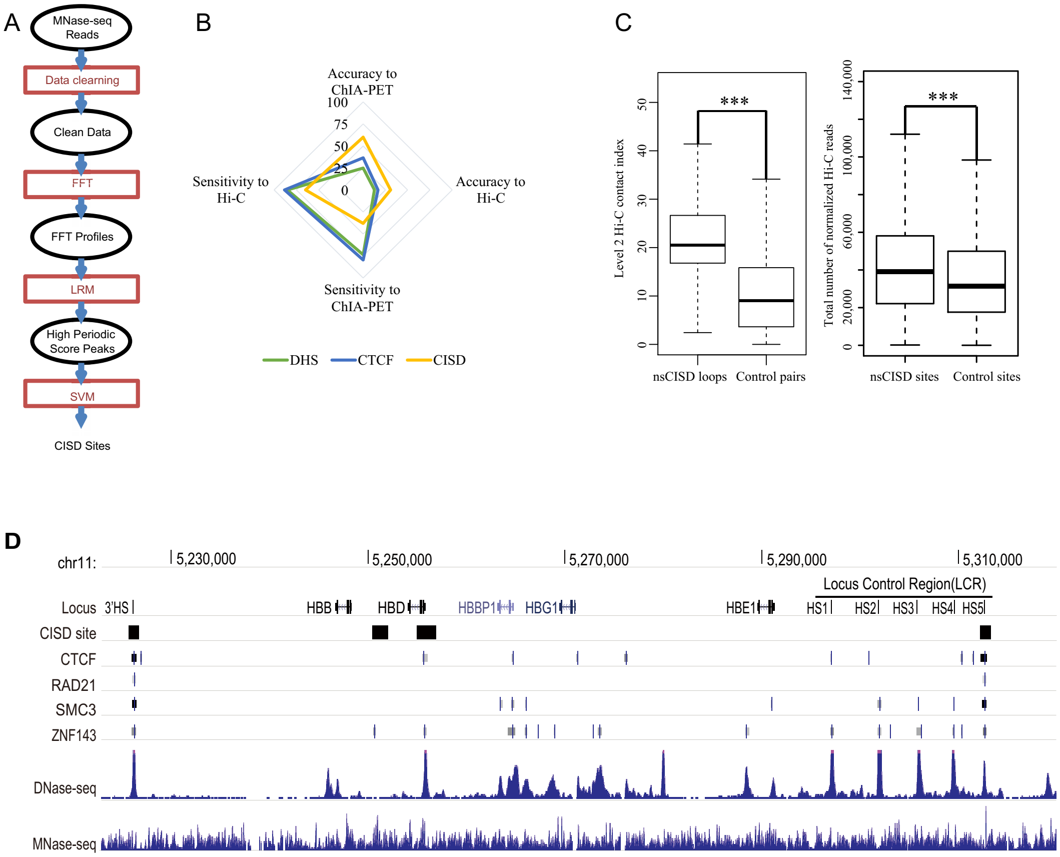

Figure 2. CISD workflow and performance. (**A**) The flow chart of CISD. Ovals represent datasets that were used or generated, and square boxes represent data processing steps. (**B**) Radar chart for the performance of by DHS, CTCF and CISD as predictors of CCIs in K562 cells. Sensitivity and accuracy were calculated using Hi-C loop anchors and ChIA-PET loop anchors. Sensitivity was defined as the percentage of predicted ChIA-PET anchors or HI-C anchors, while the accuracy was defined as the percentage of correct predictions, including true positives and true negatives, over all predictions. (**C**) Distribution of Hi-C reads counts (23) around nsCISD and control sites (right) and the distribution of the level 2 Hi-C contact index (see 'Materials and Methods' section) between nsCISD loops and between randoml loops from control sites (left) in K562 cells. For each site, the reads count is calculated from the normalized Hi-C contact matrix. The 5 kb resolution matrix was used in this figure. (***: rank sum test *P*-vale < 2.2e-16). (**D**). CISD predictions in the LCR and β-globin locus. DNase-seq and ChIP-seq peaks are also shown.

Table 1. Intra and inter cell-type performance of CISD and CISD_loop

Training set	Testing set	# of CISD sites	% of ChIA-PET anchors	# of CISD loops	% of ChIA-PET loops
K562	K562	22 112	62.30%	16 726	30.75%
K562	GM12878	20 881	78.9%	11 240	52.9%
GM12878	GM12878	26 801	72.30%	25 098	44.78%
GM12878	K562	31 633	48.6%	17 374	24.6%

specific and developmental expression of the downstream β-globin genes (60). In K562 cells, previous analyses have revealed looping interactions between the LCR and the globin gene region, roughly spanning from HBD to HBG (61,62), as well as interactions between the LCR and HS elements located downstream of the region (3′-HS) (63). In the β-globin locus, CISD detected four positive sites with about 1 kb resolution, and all of them located in the previously known interaction regions (Figure 2D). Two of the four CISD sites, i.e. those at the HS5 and 3′HS, overlap with ChIA-PET loop anchors and with CTCF, cohesin and ZNF143 ChIP-seq peaks. The CISD site at the HBD promoter region is supported by 3C data (64) and has CTCF and ZNF143 ChIP-seq peaks. The CISD site located upstream of HBB overlapped with 5C-detected loops (62,64) and also has ZNF143 binding. Thus, CISD precisely detected the current LCR–β-globin interaction network.

Prediction of CCIs between CISD sites with CISD_loop

To predict chromatin loops between CISD sites, we developed CISD_loop which takes CISD sites and low-resolution Hi-C data as input. CISD_loop is a method for determination of intra-TAD chromatin loops between CISD sites. A voting system with four decision tree-based voters, GBDT (51), random forest (51), ExtraTrees (51) and CART (51) were composed to form CISD_loop (see 'Materials and Methods' section and Figure 3A). The 5-fold cross-validation of the accuracy of CISD_loop was 80.1 and

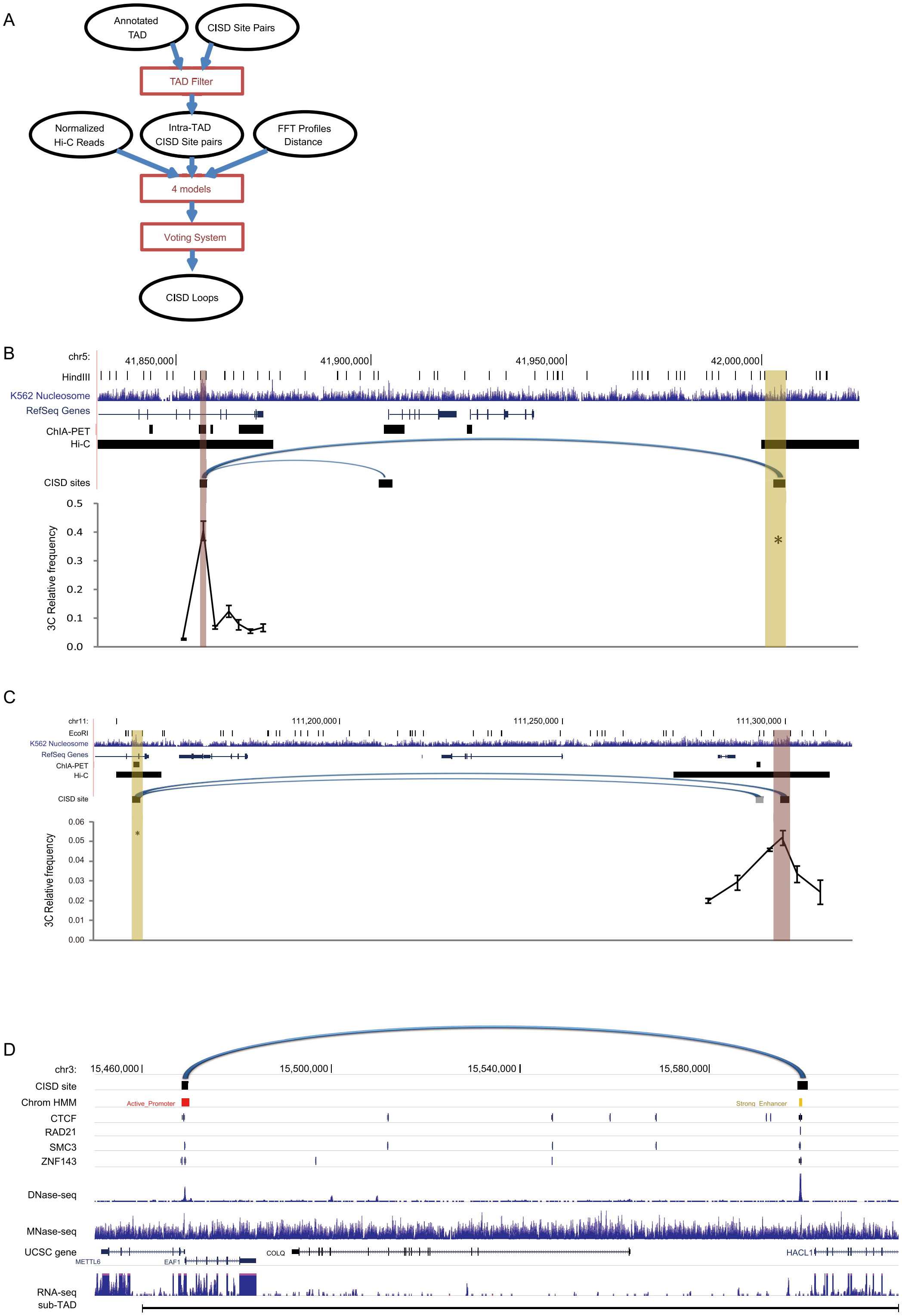

Figure 3. CISD_loop workflow and performance. (**A**) Flowchart of CISD_loop. For explanation, see Figure 2A. (**B**) 3C experiment *at loci* between chr5:42,000,785-42,006,207 (anchor, nsCISD site) and chr5:41,856,243-41,857,738 (target), and (**C**) 3C experiment *at loci* between Chr11:111,153,407-111,155,719 (anchor) and Chr11:111,297,381-111,301,207 (target, nsCISD site). Each data point represents mean ± SD of three technical replicates and three normalized biological replicates. The CCIs predicted by CISD_loop are marked as arches. The third CISD site was marked in gray, indicating that it could not be validated in this experiment because a restriction site appears in the site. (**D**) MNase-seq reads, ChIP-seq peaks, DNase-seq reads and CISD_loop predictions within a sub-TAD region on chr3:15,460,000-156,200,00. A putative chromatin interaction predicted by CISD_loop is marked as an arch.

78.5% in K562 and GM12878 cells, respectively. We applied CISD_loop on Hi-C data from K562 and GM12878 cells and predicted 16 726 and 25 098 interactions, respectively. The CISD_loop has successfully predicted 51.2 and 41.5% of the intra-TAD ChIA-PET loops connecting CISD sites in K562 and GM12878, respectively. Compared to random intra-TAD CISD site pairs, CISD_loop predictions have 6.4- and 4.9-fold higher enrichment for ChIA-PET loops in K562 and GM12878 cells, respectively. To examine the extent to which CISD loops lacking support from ChlA-PET data, termed as non-supported (ns) CISD loops, might be involved in CCIs, we compared the Hi-C contact index between the nsCISD loops to those of randomly connected CISD sites. Indeed, the nsCISD loops were significantly more enriched for paired Hi-C reads than were controls (Figure 2C, *P*-value < 2.2e-16, rank sum test).

To further assess the reliability of the nsCISD loops, we performed 3C experiments to validate at ultra-high resolution. However, the resolution of 3C experiments cannot exceed the length of the restriction fragments, and most of the CISD sites are located in restriction fragments longer than 10 kb (for commonly used HindIII, EcoRI, BglII, PstI and BamHI enzymes). We only found four CISD loops of which both anchors are in restriction fragments <7 kb (group I, Supplementary Table S1). Thus, we extended our search to include predictions with less rigid parameter settings in CISD_loop, and found three more examples (group II, Supplementary Table S1). Three out of four (75%) predictions in group I were validated (Figure 3B and C; Supplementary Table S1), i.e. the highest crosslinking frequencies were detected between restriction fragments with the predicted CISD sites, and all quantitative polymerase reaction (qPCR) products were confirmed by Sanger sequencing to be the expected ligation products. In comparison to the more than 10 kb length of the Hi-C segments in which the selected CISD sites were locPCRated, the 3C experiments confirmed that our predictions are at a much higher resolution (Figure 3B and C). One out of three (33.3%) predictions in group II was validated (Supplementary Table S1). The substantially lower validation rate compared to group I indicates that the rigid parameter settings used in current CISD_loop has increased its specificity with the cost of its sensitivity as predictor for CCIs. In other words, given that the predictions we tested by the 3C assays were for the loops without ChIA-PET data support, an empirical validation rate of 75% suggests that the majority of the CISD_loop predictions may be true.

Evidence for the reliability of CISD_loop predictions also comes from transcriptome data. For example, in a sub-TAD on chromosome 3 (chr3: 15 460 000-15 620 000; Figure 3D) (23) a strong enhancer and three genes (METTL6, EAF1 and COLQ) are annotated in the UCSC genome browser. However, RNA-seq data show that only METTL6 and EAF1 are actively transcribed in K562 cells. The promoter of the transcriptionally silent gene COLQ is much closer to the enhancer than the common bidirectional promoter of METTL6 and EAF1. Our CISD_loop prediction showed direct contact between the strong enhancer and bidirectional promoter of METTL6 and EAF1, while bypassing COLQ (Figure 3D). Thus, our prediction provides a plausible explanation for this apparent anomaly.

Modest amounts of data are sufficient for CISD/CISD_loop to achieve ultrahigh resolution predictions

To examine how many MNase-seq reads are necessary to obtain highly accurate CISD site predictions, we composed testing datasets by randomly sampling descending numbers of mapped MNase-seq reads, e.g. one half, one quarter and one eighth, from the original 1.4 billion mapped reads in K562 cells (46). The original sequencing depth was about 16-fold, and the testing datasets simulated sequencing depths of about 8-, 4- and 2-fold, with read density equivalent to about 472, 236, 118 and 59 thousand reads per million base-pair region (RPM), respectively. Even with the lowest number of reads, CISD could successfully identify periodic nucleosome regions (Figure 4A), and over a third of the predictions overlapped with currently available ChIA-PET loop anchors (Figure 4B). However, because the proportion predictions overlapping ChIA-PET loop anchors dropped substantially when read density was less than 118 thousand RPM, we recommend a density of 118 thousand RPM or higher for CISD site prediction.

We next investigated how many MNase-seq and Hi-C reads would be needed to obtain high accuracy of CISD loops. We composed sets of testing data containing all, 10 and 1% of the current Hi-C reads in K562 cells (23), corresponding to read densities equivalent to about 331, 33 and 3 thousand RPM, respectively, and examined the performance of CISD_loop for all combinations of the three Hi-C and the four MNase-seq testing datasets (Figure 4C). Exponential reduction of Hi-C reads number did not substantially affect the validation rate. With only 10% of Hi-C reads, nearly half of the ChIA-PET loops between the CISD sites could still be predicted, similar to the performance achieved with the full data, and the number only dropped to about 40% when the Hi-C data were reduced to 1%. Thus, we may set the sequencing depth according to the desired prediction rate. Finally, we also trained CISD/CISD_loop in one cell type, either K562 or GM12878, and tested it in the other cell type and obtained similar results (Supplementary Figure S6a and b; Table 1), suggesting that CISD/CISD_loop can be widely applied in human cell lines with modest amounts of input data.

The characteristic nucleosome pattern flanking CCI sites may be concordant in different cell types

To test whether the characteristic nucleosome patterns identifying CCIs is concordant across cell types, we trained CISD and CISD_loop with data from one cell type and tested them on data from another cell type (Table 1). The training ChIA-PET data size in the two cell types are not equivalent, i.e. the ChIA-PET data in the K562 cells are more concentrated for Rad21 while in the GM12878 cells are more concatenated for Pol II. Therefore, to make a fair comparison, we retrained the CISD/CISD_loop using ChIA-PET data for Rad21, which have similar amounts of data in both cells types and were also generated from the same laboratory (46). After retraining CISD/CISD_loop, this resulted in 11 240 and 17 374 predicted loops in GM12878 and K562 cells, respectively, of which 52.9 and 24.6% were validated by ChIA-PET loops. The apparent difference in validation rates can be explained by the total

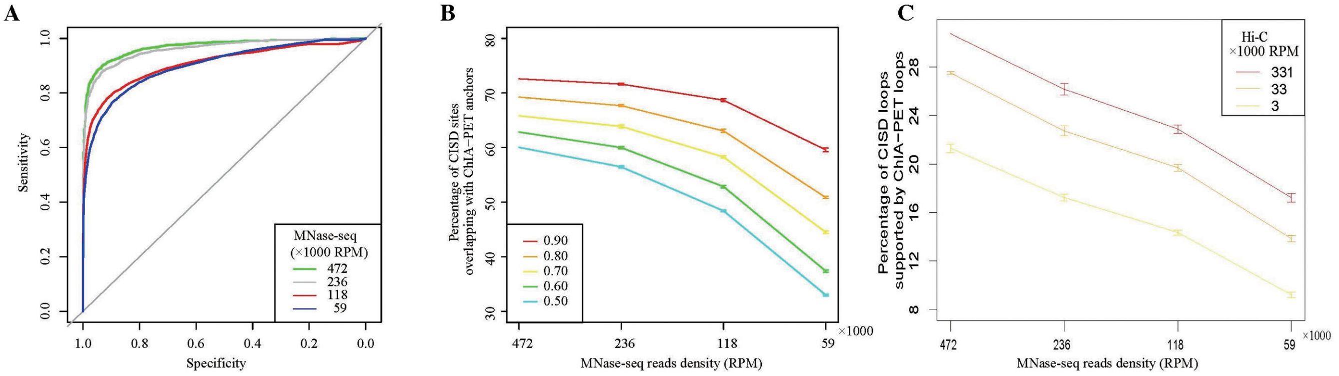

Figure 4. MNase-seq and Hi-C data requirements for high-resolution prediction by CISD and CISD_loop. (**A**) ROC curves for LRM predictions with different portions of MNase-seq data. (**B**) Percentages of CISD sites in K562 supported by ChIA-PET loop anchors under different periodic score thresholds and different portions of MNase-seq data. (**C**) Percentages of CISD loops in K562 supported by ChIA-PET loop under different densities of MNase-seq and Hi-C reads. The threshold of the periodic score in the LRM step was set at 0.5. Each data point represents mean ± SD of 10 technical replicates.

number of ChIA-PET loops available for validation (226 450 and 29 070 in GM12878 and K562 cells, respectively) differing substantially between the two cell types. On average, the validation rates of 52.9% (compared to 25.0%) and 24.6% (compared to 1.4%) were still nearly 2- and 17-fold higher than random assignments for CISD and CISD_loop, respectively. These results suggested that the characteristic nucleosome patterns flanking CCI sites may be concordant in cell types.

DISCUSSION

In this paper, we developed the CISD/CISD_loop algorithms for genome-wide identification of potential CCI sites and loops at kilobase resolution. This ultrahigh resolution can be achieved because the number of nucleosomes with distinct arrangement pattern flanking barriers is not large (65), and current MNase-seq data are sufficient to detect such patterns. In addition to ultrahigh resolution, CISD/CISD_loop also make 3D genome profile exploration more economical than ultra-deep sequencing, as only MNase-seq and low-resolution Hi-C data are used.

As a complement to current methods, e.g. EpiTensor (39) and TargetFinder (40), which detect consistent CCIs across cell types, CISD/CISD_loop can predict CCIs in a cell type-specific manner, essentially because CISD/CISD_loop do not rely on data from other cell types. Cell type-specific CCI prediction is an important advance, given the highly dynamic nature of chromatin interactions and the prevalence of cell-type specificity in promoter–enhancer interactions (3,43,44).

Epigenetic features of the genome have been used for computational modeling TF bindings (66,67) and chromatin architecture (19,36,37,39,40). The success of such models suggests the presence of profound links between CCI sites, TF binding and the dynamics of the chromatin (40). However, data integration-based methods can only provide limited insights toward the elucidation of such links. The predictive power of CISD/CISD_loop suggests that CCIs may serve as special barriers that alter the local chromatin context and may be associated with rearrangements of the nucleosomes. This, in fact, may be a potential mechanism linking CCI sites and the dynamic behavior of nucleosomes.

The presence of phased nucleosomes flanking CTCF binding sites has been well documented (48), and many other phased nucleosome arrays have been seen in mammalian genomes (45). Recently, DNase I hypersensitive sites (DHSs) have also been suggested to predict chromatin interactions (19,41). The binding of CTCF alone was not sufficient to predict chromatin interactions because (i) not all CTCF binding sites have high periodic scores or overlap with ChIA-PET anchors (Supplementary Figure S7a), and (ii) only 55.38% of the CTCF ChIA-PET anchors are found in other ChIA-PET data. Compared to the predictive power of DHSs, CISD sites are more enriched for ChIA-PET loop anchors, as shown in the main text (Figure 2B); thus, it is less likely that the predictive power of CISD simply relies on the presence of open chromatin. Finally, periodic nucleosome arrangements alone are not predictive of ChIA-PET loop anchors. For example, in lymphoblastoid cells, a 76 kb-long region is reported to have a very well-phased nucleosome array caused by the DNA sequence (45). Therefore, we examined the ChIA-PET data on this region from K562 cells. Indeed, the nucleosomes are regularly arranged, as indicated by MNase-seq data; however, only one ChIA-PET loop anchor was found in the region, which is also one of the two CISD-predicted sites in the same region (Supplementary Figure S7b). Thus, neither the binding of TFs nor simple arrays of well-phased nucleosomes is sufficient to indicate stable chromatin contacts.

One may ask what characteristics of nucleosome pattern make CISD predictive? Two prominent characteristics of the CISD sites immediately presented themselves by plot the distribution of MNase-seq data flanking the CISD sites.

First, in the CISD sites, there is a large gap situated at the middle of the phased nucleosome array. To show this pattern, we compared the length distribution of the longest intervals in the nucleosome arrays in the CISD sites, ChIA-PET anchors, TSSs and the 76 kb phased nucleosome region (Supplementary Figure S8). Both CISD sites and ChIA-PET anchors peaked at about 300 bp, while the 76 kb region showed an additional peak at about 220 bp, and the average length of the NFR was even shorter. Second, the distribution of nucleosome arrays is symmetric at the CISD sites (Supplementary Figure S9). It is well known that the phased nucleosomes flanking the NFR at TSS is asymmetric (45), while at CISD sites, the nucleosomes are symmetric.

One may ask if the two characteristics we have outlined above are sufficient to distinguish the CISD sites from rest of the phasing regions. Our answer is that it is not, because only a small number of such sites overlap with the ChIA-PET anchors. To show this, we binned the genomic loci according to their symmetry level of the nucleosome arrangement, i.e. the Pearson's correlation coefficient between the forward and reversed MNase-seq signals in the loci. Indeed, with increasing symmetry, the validation rate (as supported by the ChIA-PET data) also increases (Supplementary Figure S10); however, the total number of sites declined substantially, suggesting that additional information in the FFT profile is critical for the identification of CISD sites. Taken together, while we believe that an extended gap flanked by symmetric arrangements is the most prominent feature of the CISD sites, the predictive power of CISD stems is dependent on more than this single feature.

The biochemical mechanisms underlying chromatin interactions are not well understood, and are likely to depend on complex process with many factors involved at multiple levels. Thus, any algorithm relying on a single datum is likely to yield an incomplete result, and indeed, the sensitivity of CISD is limited (Figure 2B). Therefore, it is possible that further improvement may be achieved by marrying data integration and hypothesis-driven modeling. Furthermore, as suggested by Whalen *et al*. (40), the information relevant to looping interactions is not just limited to the interacting loci (40). Thus, taking data from outside the interacting loci into account also merits further investigation.

The substantial attention in recent literature directed toward 3D genome studies reflects the importance of such knowledge. CISD and CISD_loop provide an approach that facilitates the expansion of the field of 3D genome research by allowing the exploration of more cell types, tissues and species.

DATA AVAILABILITY

Data that were downloaded from public domain are listed in the Supplementary Text. Source code for the CISD and CISD_loop can be found at https://github.com/huizhangucas/CISD.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr Zhihu Zhao for the help in setting up the 3C experiment. We thank Dr Qianfei Wang and Dr Changqing Zeng for the help in accessing experimental equipment, and we appreciate helpful discussions with Dr Cheng Li, Dr Guoliang Li, Dr Chengqi Wang, Lijia Yu, Lili Xia and Guangyu Wang. We thank Dr Geir Skogerbo and Mr David Martin for the language correction on the manuscript. We also acknowledge the ENCODE Consortium and the ENCODE production laboratories that generated the datasets provided by the ENCODE Data Coordination Center used in this manuscript.

Author contributions: Z.Z. conceived this project. H.Z. and Z.Z. developed CISD and CISD_loop. F.L. designed the 3C experiment. F.L., Y.J. and X.L. performed the 3C experiment. H.Z., B.X., Y.Z. analyzed data, and Z.Z., H.Z. and F.L. prepared the manuscript. All authors read and approved the final manuscript.

FUNDING

National Basic Research Program of China [2014CB542002]; National Nature Science Foundation of China [31671342, 91540114 to Z.Z. and 31401113 to F.L.]; Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (the second phase) (to Z.Z.). Funding for open access charge: National Basic Research Program of China [2014CB542002].

Conflict of interest statement. None declared.

REFERENCES

1. Roy,A.L., Sen,R. and Roeder,R.G. (2011) Enhancer-promoter communication and transcriptional regulation of Igh. *Trends Immunol.*, **32**, 532–539.
2. Li,G., Ruan,X., Auerbach,R.K., Sandhu,K.S., Zheng,M., Wang,P., Poh,H.M., Goh,Y., Lim,J., Zhang,J. *et al.* (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, **148**, 84–98.
3. Zhang,Y., Wong,C.H., Birnbaum,R.Y., Li,G., Favaro,R., Ngan,C.Y., Lim,J., Tai,E., Poh,H.M., Wong,E. *et al.* (2013) Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature*, **504**, 306–310.
4. de Wit,E. and de Laat,W. (2012) A decade of 3C technologies: insights into nuclear organization. *Genes Dev.*, **26**, 11–24.
5. de Laat,W. and Dekker,J. (2012) 3C-based technologies to study the shape of the genome. *Methods*, **58**, 189–191.
6. Simonis,M., Kooren,J. and de Laat,W. (2007) An evaluation of 3C-based methods to capture DNA interactions. *Nat. Methods*, **4**, 895–901.
7. Belton,J.M., McCord,R.P., Gibcus,J.H., Naumova,N., Zhan,Y. and Dekker,J. (2012) Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods*, **58**, 268–276.
8. Zhang,Y., McCord,R.P., Ho,Y.J., Lajoie,B.R., Hildebrand,D.G., Simon,A.C., Becker,M.S., Alt,F.W. and Dekker,J. (2012) Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell*, **148**, 908–921.
9. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
10. Naumova,N., Imakaev,M., Fudenberg,G., Zhan,Y., Lajoie,B.R., Mirny,L.A. and Dekker,J. (2013) Organization of the mitotic chromosome. *Science*, **342**, 948–953.
11. Dixon,J.R., Selvaraj,S., Yue,F., Kim,A., Li,Y., Shen,Y., Hu,M., Liu,J.S. and Ren,B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.
12. Selvaraj,S., R Dixon,J., Bansal,V. and Ren,B. (2013) Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat. Biotechnol.*, **31**, 1111–1118.
13. Sexton,T., Yaffe,E., Kenigsberg,E., Bantignies,F., Leblanc,B., Hoichman,M., Parrinello,H., Tanay,A. and Cavalli,G. (2012) Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*, **148**, 458–472.

14. Duan,Z., Andronescu,M., Schutz,K., McIlwain,S., Kim,Y.J., Lee,C., Shendure,J., Fields,S., Blau,C.A. and Noble,W.S. (2010) A three-dimensional model of the yeast genome. *Nature*, **465**, 363–367.
15. Handoko,L., Xu,H., Li,G., Ngan,C.Y., Chew,E., Schnapp,M., Lee,C.W., Ye,C., Ping,J.L., Mulawadi,F. *et al.* (2011) CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat. Genet.*, **43**, 630–638.
16. Tang,Z., Luo,O.J., Li,X., Zheng,M., Zhu,J.J., Szalaj,P., Trzaskoma,P., Magalska,A., Wlodarczyk,J., Ruszczycki,B. *et al.* (2015) CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell*, **163**, 1611–1627.
17. Sandhu,K.S., Li,G., Poh,H.M., Quek,Y.L., Sia,Y.Y., Peh,S.Q., Mulawadi,F.H., Lim,J., Sikic,M., Menghi,F. *et al.* (2012) Large-scale functional organization of long-range chromatin interaction networks. *Cell Rep.*, **2**, 1207–1219.
18. Demare,L.E., Leng,J., Cotney,J., Reilly,S.K., Yin,J., Sarro,R. and Noonan,J.P. (2013) The genomic landscape of cohesin-associated chromatin interactions. *Genome Res.*, **23**, 1224–1234.
19. Heidari,N., Phanstiel,D.H., He,C., Grubert,F., Jahanbani,F., Kasowski,M., Zhang,M.Q. and Snyder,M.P. (2014) Genome-wide map of regulatory interactions in the human genome. *Genome Res.*, **24**, 1905–1917.
20. Nora,E.P., Lajoie,B.R., Schulz,E.G., Giorgetti,L., Okamoto,I., Servant,N., Piolot,T., van Berkum,N.L., Meisig,J., Sedat,J. *et al.* (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, **485**, 381–385.
21. Hou,C., Li,L., Qin,Z.S. and Corces,V.G. (2012) Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. *Mol. Cell*, **48**, 471–484.
22. Phillips-Cremins,J.E., Sauria,M.E., Sanyal,A., Gerasimova,T.I., Lajoie,B.R., Bell,J.S., Ong,C.T., Hookway,T.A., Guo,C., Sun,Y. *et al.* (2013) Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, **153**, 1281–1295.
23. Rao,S.S., Huntley,M.H., Durand,N.C., Stamenova,E.K., Bochkov,I.D., Robinson,J.T., Sanborn,A.L., Machol,I., Omer,A.D., Lander,E.S. *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.
24. Shen,Y., Yue,F., McCleary,D.F., Ye,Z., Edsall,L., Kuan,S., Wagner,U., Dixon,J., Lee,L., Lobanenkov,V.V. *et al.* (2012) A map of the cis-regulatory sequences in the mouse genome. *Nature*, **488**, 116–120.
25. Mozziconacci,J. and Koszul,R. (2015) Filling the gap: Micro-C accesses the nucleosomal fiber at 100–1000 bp resolution. *Genome Biol.*, **16**, 169.
26. Naumova,N., Smith,E.M., Zhan,Y. and Dekker,J. (2012) Analysis of long-range chromatin interactions using Chromosome Conformation Capture. *Methods*, **58**, 192–203.
27. Yaffe,E. and Tanay,A. (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.*, **43**, 1059–1065.
28. Imakaev,M., Fudenberg,G., McCord,R.P., Naumova,N., Goloborodko,A., Lajoie,B.R., Dekker,J. and Mirny,L.A. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods*, **9**, 999–1003.
29. Hu,M., Deng,K., Selvaraj,S., Qin,Z., Ren,B. and Liu,J.S. (2012) HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics*, **28**, 3131–3133.
30. Williamson,I., Berlivet,S., Eskeland,R., Boyle,S., Illingworth,R.S., Paquette,D., Dostie,J. and Bickmore,W.A. (2014) Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization. *Genes Dev.*, **28**, 2778–2791.
31. Kolovos,P., van de Werken,H.J., Kepper,N., Zuin,J., Brouwer,R.W., Kockx,C.E., Wendt,K.S., van,I.W.F., Grosveld,F. and Knoch,T.A. (2014) Targeted Chromatin Capture (T2C): a novel high resolution high throughput method to detect genomic interactions and regulatory elements. *Epigenetics Chromatin.*, **7**, 10.
32. Dryden,N.H., Broome,L.R., Dudbridge,F., Johnson,N., Orr,N., Schoenfelder,S., Nagano,T., Andrews,S., Wingett,S., Kozarewa,I. *et al.* (2014) Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Res.*, **24**, 1854–1868.
33. Hughes,J.R., Roberts,N., McGowan,S., Hay,D., Giannoulatou,E., Lynch,M., De Gobbi,M., Taylor,S., Gibbons,R. and Higgs,D.R. (2014) Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat. Genet.*, **46**, 205–212.
34. Ma,W., Ay,F., Lee,C., Gulsoy,G., Deng,X., Cook,S., Hesson,J., Cavanaugh,C., Ware,C.B., Krumm,A. *et al.* (2015) Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. *Nat. Methods*, **12**, 71–78.
35. Hsieh,T.H., Weiner,A., Lajoie,B., Dekker,J., Friedman,N. and Rando,O.J. (2015) Mapping nucleosome resolution chromosome folding in yeast by micro-C. *Cell*, **162**, 108–119.
36. Fortin,J.P. and Hansen,K.D. (2015) Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biol.*, **16**, 180.
37. Huang,J., Marco,E., Pinello,L. and Yuan,G.C. (2015) Predicting chromatin organization using histone marks. *Genome Biol.*, **16**, 162.
38. Chen,Y., Wang,Y., Xuan,Z., Chen,M. and Zhang,M.Q. (2016) De novo deciphering three-dimensional chromatin interaction and topological domains by wavelet transformation of epigenetic profiles. *Nucleic Acids Res.*, **44**, e106.
39. Zhu,Y., Chen,Z., Zhang,K., Wang,M., Medovoy,D., Whitaker,J.W., Ding,B., Li,N., Zheng,L. and Wang,W. (2016) Constructing 3D interaction maps from 1D epigenomes. *Nat. Commun.*, **7**, 10812.
40. Whalen,S., Truty,R.M. and Pollard,K.S. (2016) Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.*, **48**, 488–496.
41. He,C., Wang,X. and Zhang,M.Q. (2014) Nucleosome eviction and multiple co-factor binding predict estrogen-receptor-alpha-associated long-range interactions. *Nucleic Acids Res.*, **42**, 6935–6944.
42. Pennings,S., Muyldermans,S., Meersseman,G. and Wyns,L. (1989) Formation, stability and core histone positioning of nucleosomes reassembled on bent and other nucleosome-derived DNA. *J. Mol. Biol.*, **207**, 183–192.
43. Wang,G.G., Allis,C.D. and Chi,P. (2007) Chromatin remodeling and cancer, Part II: ATP-dependent chromatin remodeling. *Trends Mol. Med.*, **13**, 373–380.
44. Kornberg,R.D. and Stryer,L. (1988) Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. *Nucleic Acids Res.*, **16**, 6677–6690.
45. Gaffney,D.J., McVicker,G., Pai,A.A., Fondufe-Mittendorf,Y.N., Lewellen,N., Michelini,K., Widom,J., Gilad,Y. and Pritchard,J.K. (2012) Controls of nucleosome positioning in the human genome. *PLoS Genet.*, **8**, e1003036.
46. Kundaje,A., Kyriazopoulou-Panagiotopoulou,S., Libbrecht,M., Smith,C.L., Raha,D., Winters,E.E., Johnson,S.M., Snyder,M., Batzoglou,S. and Sidow,A. (2012) Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome Res.*, **22**, 1735–1747.
47. Nie,Y., Cheng,X., Chen,J. and Sun,X. (2014) Nucleosome organization in the vicinity of transcription factor binding sites in the human genome. *BMC Genomics*, **15**, 493.
48. Fu,Y., Sinha,M., Peterson,C.L. and Weng,Z. (2008) The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet.*, **4**, e1000138.
49. Chen,W., Liu,Y., Zhu,S., Green,C.D., Wei,G. and Han,J.D. (2014) Improved nucleosome-positioning algorithm iNPS for accurate nucleosome positioning from sequencing data. *Nat. Commun.*, **5**, 4909.
50. Zhang,Y., Shin,H., Song,J.S., Lei,Y. and Liu,X.S. (2008) Identifying positioned nucleosomes with epigenetic marks in human from ChIP-Seq. *BMC Genomics*, **9**, 537.
51. Pedregosa,F., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V. and Vanderplas,J. (2012) Scikit-learn: machine learning in python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
52. Genomes Project, C., Auton,A., Brooks,L.D., Durbin,R.M., Garrison,E.P., Kang,H.M., Korbel,J.O., Marchini,J.L., McCarthy,S., McVean,G.A. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
53. Rozowsky,J., Abyzov,A., Wang,J., Alves,P., Raha,D., Harmanci,A., Leng,J., Bjornson,R., Kong,Y., Kitabayashi,N. *et al.* (2011) AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.*, **7**, 522.

54. Fullwood,M.J. and Ruan,Y. (2009) ChIP-based methods for the identification of long-range chromatin interactions. *J. Cell Biochem.*, **107**, 30–39.
55. Consortium,T.E.P. (2013) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **488**, 57–74.
56. Davey,C. and Allan,J. (2003) Nucleosome positioning signals and potential H-DNA within the DNA sequence of the imprinting control region of the mouse Igf2r gene. *Biochim. Biophys. Acta*, **1630**, 103–116.
57. Gasiorek,J.J. and Blank,V. (2015) Regulation and function of the NFE2 transcription factor in hematopoietic and non-hematopoietic cells. *Cell Mol. Life Sci.*, **72**, 2323–2335.
58. Hubner,M.R. and Spector,D.L. (2010) Chromatin dynamics. *Annu. Rev. Biophys.*, **39**, 471–489.
59. He,B., Chen,C., Teng,L. and Tan,K. (2014) Global view of enhancer-promoter interactome in human cells. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, E2191–E2199.
60. Li,Q., Peterson,K.R., Fang,X. and Stamatoyannopoulos,G. (2002) Locus control regions. *Blood*, **100**, 3077–3086.
61. Dostie,J., Richmond,T.A., Arnaout,R.A., Selzer,R.R., Lee,W.L., Honan,T.A., Rubio,E.D., Krumm,A., Lamb,J., Nusbaum,C. *et al.* (2006) Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.*, **16**, 1299–1309.
62. Sanyal,A., Lajoie,B.R., Jain,G. and Dekker,J. (2012) The long-range interaction landscape of gene promoters. *Nature*, **489**, 109–113.
63. Tolhuis,B., Palstra,R.J., Splinter,E., Grosveld,F. and de Laat,W. (2002) Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol. Cell*, **10**, 1453–1465.
64. Bailey,S.D., Zhang,X., Desai,K., Aid,M., Corradin,O., Cowper-Sal Lari,R., Akhtar-Zaidi,B., Scacheri,P.C., Haibe-Kains,B. and Lupien,M. (2015) ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nat. Commun.*, **2**, 6186.
65. Chevereau,G., Palmeira,L., Thermes,C., Arneodo,A. and Vaillant,C. (2009) Thermodynamics of intragenic nucleosome ordering. *Phys. Rev. Lett.*, **103**, 188103.
66. Xu,T., Li,B., Zhao,M., Szulwach,K.E., Street,R.C., Lin,L., Yao,B., Zhang,F., Jin,P., Wu,H. *et al.* (2015) Base-resolution methylation patterns accurately predict transcription factor bindings in vivo. *Nucleic Acids Res.*, **43**, 2757–2766.
67. Pique-Regi,R., Degner,J.F., Pai,A.A., Gaffney,D.J., Gilad,Y. and Pritchard,J.K. (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.*, **21**, 447–455.