Analysis of high-resolution 3D intrachromosomal interactions aided by Bayesian network modeling

Xizhe Zhang[a], Sergio Branciamore[a], Grigoriy Gogoshin[a], Andrei S. Rodin[a], and Arthur D. Riggs[a,1]

[a]Department of Diabetes Complications and Metabolism, Diabetes and Metabolism Research Institute, City of Hope, Duarte, CA

Contributed by Arthur D. Riggs, October 2, 2017 (sent for review December 16, 2016; reviewed by Peter N. Cockerill and Leonid A. Mirny)

Long-range intrachromosomal interactions play an important role in 3D chromosome structure and function, but our understanding of how various factors contribute to the strength of these interactions remains poor. In this study we used a recently developed analysis framework for Bayesian network (BN) modeling to analyze publicly available datasets for intrachromosomal interactions. We investigated how 106 variables affect the pairwise interactions of over 10 million 5-kb DNA segments in the B-lymphocyte cell line GB12878. Strictly data-driven BN modeling indicates that the strength of intrachromosomal interactions (hic_strength) is directly influenced by only four types of factors: distance between segments, Rad21 or SMC3 (cohesin components), transcription at transcription start sites (TSS), and the number of CCCTC-binding factor (CTCF)–cohesin complexes between the interacting DNA segments. Subsequent studies confirmed that most high-intensity interactions have a CTCF–cohesin complex in at least one of the interacting segments. However, 46% have CTCF on only one side, and 32% are without CTCF. As expected, high-intensity interactions are strongly dependent on the orientation of the ctcf motif, and, moreover, we find that the interaction between enhancers and promoters is similarly dependent on ctcf motif orientation. Dependency relationships between transcription factors were also revealed, including known lineage-determining B-cell transcription factors (e.g., Ebf1) as well as potential novel relationships. Thus, BN analysis of large intrachromosomal interaction datasets is a useful tool for gaining insight into DNA–DNA, protein–DNA, and protein–protein interactions.

DNA reeling | DNA looping | enhancers | chromatin

Mammalian chromosomes are very complex structures, containing approximately 10^8 bp of DNA highly organized in 3D space, compacted by coiling around nucleosomes and then folded into various size loops. Due to coiling and folding, many distant genomic segments, even 1 Mb or more apart, frequently contact each other because they actually are in close spatial proximity (1). As a prime example, promoters can interact with distal enhancers sometimes 1 Mb or more upstream or downstream, and this interaction is required for correct timing and level of gene transcription (2–4). Genome-wide chromosomal interactions are now being successfully investigated by chromosome conformation capture (3C)-based techniques (5), such as Hi-C and chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) (6, 7). ChIA-PET can ascertain high-resolution interactions, but this method depends on enrichment by chromatin immunoprecipitation (ChIP) followed by paired-end sequencing and so can address only interactions mediated by a prespecified protein. Hi-C, which depends on the ligation of formaldehyde-fixed, sheared chromatin and then massive sequencing to detect ligation of distant DNA fragments, can in theory obtain the interaction frequency between any two genomic fragments. A high-resolution Hi-C dataset requires high sequencing depth, so the limit of resolution of Hi-C is currently about 5 kb. At this resolution a Hi-C dataset contains information on pairwise interactions of about 1 million fragments (8).

Bayesian network (BN) modeling (9–12) is an established systems biology method aimed at optimizing, visualizing, and analyzing biological network models reconstructed from "big data" such as generated by Hi-C studies. However, BN modeling has not been applied to chromatin interaction data before, even though advantages of the BN approach over other comparative secondary data analysis methods are numerous, including flexibility of model visualization and interpretation, ability to incorporate different variables and biological entities into a single model, and straightforward statistical and/or biological follow-up. A primary aim of this study was to see whether BN modeling could be applied to the large datasets generated by combining chromosomal conformation capture data (e.g., Hi-C) with Encyclopedia of DNA Elements (ENCODE) data on protein binding and transcription.

Recent Hi-C and other studies (reviewed by ref. 13) have revealed several chromosomal substructures in which interaction frequencies between distant DNA fragments are higher than would be expected if interactions were due to random diffusion. As the resolution of Hi-C experiments has increased, substructures of smaller size have emerged. The structure named chromosomal compartment was identified at 1-Mb resolution (6). These relatively large compartments, which show variability between cell types, adopt two states, either transcriptionally inactive, with closed chromatin, or active with open chromatin and corresponding histone signatures. At a resolution level of tens of kilobases, a structure named topological associated domain (TAD) appears (14, 15). TADs, which are megabase sized, are highly conserved across different cell types, although the disruption of TAD boundaries has been found to cause developmental

Significance

We report here that a recently developed Bayesian network (BN) methodology and software platform yield useful information when applied to the analysis of intrachromosomal interaction datasets combined with Encyclopedia of DNA Elements publicly available datasets for the B-lymphocyte cell line GM12878. Of 106 variables analyzed, interaction strength between DNA segments was found to be directly dependent on only four types of variables: distance, Rad21 or SMC3 (cohesin components), transcription at transcription start sites, and the number of CCCTC-binding factor (CTCF)–cohesin complexes between interacting DNA segments. The importance of directionally oriented ctcf motifs was confirmed not only for loops but also for enhancer–promoter interactions. Purely data-driven BN analyses also identified known critical, lineage-determining transcription factors (TFs) as well as some potentially new dependencies between TFs.

Author contributions: X.Z., S.B., G.G., A.S.R., and A.D.R. designed research; X.Z., S.B., G.G., and A.S.R. performed research; X.Z., S.B., G.G., A.S.R., and A.D.R. analyzed data; and X.Z., S.B., A.S.R., and A.D.R. wrote the paper.

Reviewers: P.N.C., University of Birmingham; and L.A.M., Massachusetts Institute of Technology.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. Email: ariggs@coh.org.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1620425114/-/DCSupplemental.

anomalies and activate proto-oncogenes (16, 17). A recent Hi-C study using in situ formaldehyde fixation has determined an interaction matrix at 5-kb resolution in several cell lines, and this study has revealed new smaller subdomain structures interpreted as loops (referred to as "Hi-C loops" in the following). Hi-C loops average 185 kb and, importantly, and in contrast to TADs, vary between cell lines (8). The functions of Hi-C loops have not been fully addressed, but they are likely to be involved in cell-type–specific gene regulation (13).

In this paper the term "anchor" is used to designate the nonoverlapping genomic bins involved in Hi-C interactions. Let H be a symmetric matrix with

$$H_{ij} = h(b_i, b_j), \quad h : C \times C \to \mathbb{R}, \qquad [1]$$

where $h(b_i, b_j)$ is the value from the Hi-C dataset for each pairwise bin interaction. In the context of this paper we consider only the upper triangular matrix U defined by H_{ij} such that $i < j$. We also refer to b_i as the left and b_j as the right anchor, respectively.

How is 1D information in DNA converted into a 3D interphase chromosome? How are loops with resultant loop anchors formed? To explain *cis* action in X chromosome inactivation, DNA reeling was proposed in 1990 as a mechanism for forming loops, with the DNA strand being pulled toward a protein complex fixed in position by sequence-specific DNA binding (18). According to this model, as DNA is extruded from this site, a loop is formed. This DNA reeling/loop extrusion process also would bring distant elements, such as enhancers and promoters into close proximity. More recently it has been proposed that the cohesin complex, containing RAD21, SMC1, and SMC3, is involved in chromosome loop formation and chromosome condensation (19, 20). A variation of these reeling/extrusion models (Fig. 1*A*), with the termination of reeling often being fixed by CCCTC-binding factor (CTCF) sites, can in large part explain the pattern of interactions seen by Hi-C experiments as well as changes in chromosomal interactions and gene function as a result of deletion or inversion of CTCF sites (21–23).

ChIP-seq experiments have established that TAD and Hi-C loop anchors are enriched for CTCF and for a complex of CTCF and cohesin (8, 15). Since the concensus ctcf sequence motif to which CTCF binds (5′-CCACNAGGTGGCAG-3′) is not palindromic, one can distinguish "forward" (F) and "reverse" (R) motif directions (8). Thus, each pair of ctcf motifs (and CTCF–cohesin complexes) falls into one of four categories: (*i*) convergent, F–R; (*ii*) divergent, R–F; (*iii*) tandem plus, F–F; and (*iv*) tandem minus, R–R. A striking finding about CTCF–cohesin complexes locating in the anchors of various chromosome structures, including contact domains, Hi-C loops, and TADs, is that they are highly enriched in the convergent (F–R) pattern (8, 24, 25), with RAD21 located on the 3′ side of the ctcf sequence (24, 26). Genome-wide, segments containing the convergent CTCF–cohesin complex are known to interact at higher strength compared with other combinations, but their role in the formation and regulation of chromosomal interactions is not yet well understood.

Important remaining questions are, How are high-intensity Hi-C interactions formed? And what proteins are involved? In this study, we address these questions by exploring how various transcription factors and other chromatin proteins affect the formation or function of these interactions (27). If one extends the analyses to account for higher-order interactions (two or more factors acting together in a nonadditive fashion), addressing the above questions directly by experimental or standard bioinformatics methods becomes intractable due to the sheer combinatorial complexity. And yet, such higher-order interactions are biologically very likely. To complicate matters, the analyses are limited by the resolution of the interaction maps.

For BN modeling (9–12), our strategy encompassed simultaneous "(relevant) variable selection" (28), construction, and

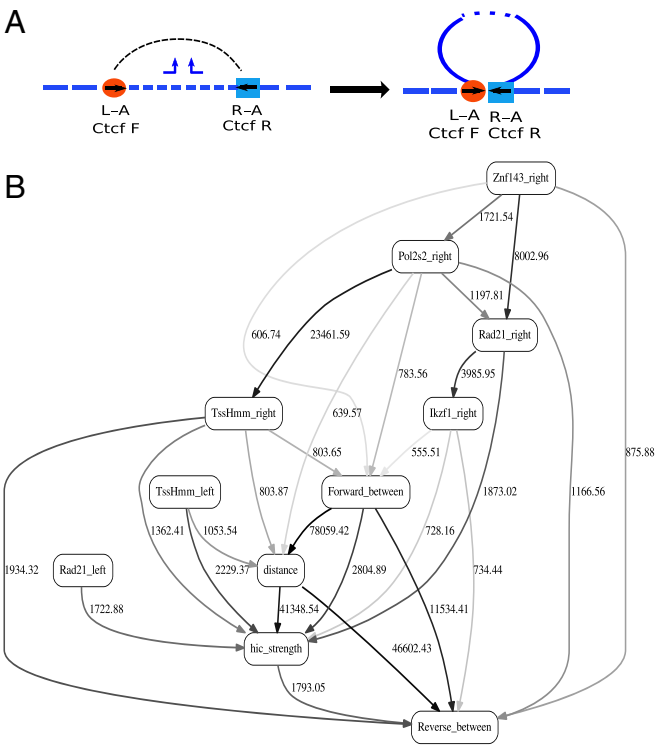

Fig. 1. (*A*) Loop formation due to DNA reeling. A DNA reeling machine binds between two CTCF–cohesin complexes and initiates DNA reeling. Reeling is stopped when an appropriately oriented CTCF–cohesin complex is reached. As a result of this process, two convergent CTCF–cohesin complexes are often pulled close to each other. Paired bent arrows represent a bidirectional reeling machine pulling in DNA from both sides. Red circle is left anchor (L-A) segment with an F ctcf motif. Blue square is a right anchor (R-A) segment containing an R-oriented ctcf motif. (*B*) BN analysis for chromosome 1. See *Results* for an introduction to BN analysis. Shown is the MN for the variable "hic_strength." This is a part of the complete BN shown in *SI Appendix*, Fig. S1. This BN is derived from the chromosome 1 dataset containing interactions wherein both anchors are located within Hi-C loops. Nodes in the network correspond to the variables, and edges to the dependencies between the variables. Directionality of the edge (arrow) is for mathematical convenience only and does not imply causation. "Boldness" of the edge is proportional to the dependency strength, also indicated by the number shown next to the edge. See *SI Appendix*, section 5, Tables S1 and S2.

visualization of biological network models of the above relationships and interactions (a "systems biology" approach) (29). Consequently, we built BNs to elucidate and visualize the effect of various protein factors and other known variables on chromosomal interactions. The primary data we used include binding information for 64 transcription factors as well as several other variables collected from the ENCODE Data Coordination Center (30, 31). Our BN analysis encompasses the high-resolution (5-kb) dataset for over 10 million intrachromosomal interactions (8).

We report here that our purely data-driven (without human expert input) BN analyses suggest that strength of intrachromosomal interactions (hic_strength) is directly dependent on only 4 of the 106 variables included in our datasets. As expected, distance and cohesin (RAD21/SMC3) stand out. However, in addition, two other variables emerged: active transcription starting sites and the number of CTCF–cohesin peaks between anchors. Subsequent studies, stimulated by, but not dependent on, the Bayesian analyses, confirmed the importance of transcription for Hi-C interaction strength. By a type of analysis that to our knowledge has not been previously used, we not only confirm

the expected effect of ctcf sequence motif directionality on Hi-C loop interactions, but also clearly show a similar directionality effect on enhancer–promoter (EP) interactions. Of potential importance, we note that most Hi-C loops do not have CTCF at both anchors.

BNs can also be used to gain information on relationships between proteins in the ENCODE database. For example, the known interactions between CTCF and RAD21(part of the cohesin complex) and between ZNF143 and RAD21 were clearly revealed. Active TSS activity was found to be dependent on several transcription factors, such as Ebf1 and Ikzf1, both of which are known to be important for B-cell development. Subsequent analysis of ENCODE ChIP-seq data indicated that EBF1 is bound at most active promoter–enhancer pairs in the B-cell lymphoma cell line GM12878. In addition, other transcription factors (TFs) and chromatin proteins were suggested as potential key players for B-cell development and/or chromosome structure. In general, we found BN modeling to be an excellent methodology for the secondary data analysis of the large-scale chromatin interaction datasets, on both computational and interpretation/follow-up levels. Consequently, we have built a specialized software analysis pipeline that is directly applicable to such data. It is freely available from the authors.

Results

Bayesian Network Reconstruction. The primary goal of this study was to see whether BN analysis could help extract useful information from complex genome-wide chromatin interaction datasets, including Hi-C and ChIP-seq. Toward this aim we investigated the robustness of the BN reconstruction with respect to algorithmic, biological, and dataset-related parameters. Special attention was paid to the integration of different data types within a single analysis framework, specifically both discrete and continuous variables. While BNs are generally well established in several biomedical research areas (genomics, expression data, metabolomics, etc.) (12, 32–34), they have not been used for chromatin interaction analysis. Therefore, a brief introduction is in order.

Traditional statistical techniques are ill suited for analyzing large-scale, multidimensional data with higher-order interactions between the variables of different types. In this study we have up to hundreds of heterogeneous variables (TFs, chromatin variables, etc.), and one way to coalesce them is via BN modeling. Statistically speaking, the BN is a sparse graphical model of a joint multivariate probability distribution of random (both continuous and discrete) variables that reflects relationships of dependence and conditional independence among them. Its primary goal is reverse engineering (from the "flat" datasets) of the biological relationships (pathways) between variables, with an eye toward devising a compact descriptive/predictive model to guide further analysis and experimentation. The principal output is a network-looking graph with nodes (variables) connected by edges reflecting significant statistical dependencies with accompanying numbers quantifying dependency strengths [see Fig. 1*B* for an example; variable (node) names and explanations are in *SI Appendix*, section 5, Tables S1 and S2, and section 7, methods which also detail organization of the primary flat datasets from which the BNs are derived]. The origins of BN methodology go back to the seminal path analysis work of Sewall Wright (35); however, due to the computational complexity of the model selection process, application of BN modeling to the nontrivial datasets has become feasible only recently. We have developed open-source, publicly available BN reconstruction software that scales up to at least hundreds of thousands of heterogeneous variables and data points, thus making it a perfect fit for the present study (12, 36) (BNOmics, at https://bitbucket.org/uthsph/bnomics/).

We built a series of BNs following different parameters, variable combinations, and visualization shortcuts for the chromatin states and potentially influencing factors and interpreted their structures (topologies), using standard criteria to get hypothesis-generating insights into the underlying mechanistic system. An example of such insight would be a direct influence of a factor or factors on a chromatin interaction (dependence, depicted as a network edge) or absence thereof (conditional independence).

The reader is referred to refs. 37 and 38 for a formal treatment of conditional independence; for our purposes a simplified concept of Markov neighborhood (MN), similar but not precisely equivalent to a formal concept of "Markov blanket" (38), of a network node (variable) is useful. A primary MN refers to a subset of BN nodes directly connected to the node representing a variable of interest. An extended MN might include a subset of nodes directly connected to the variables in the primary MN ("2 degrees of separation," so to speak). The obvious usefulness of the MN approach lies in visualization and variable selection. The latter broadly implies that the variables in the MN or extended MN are suggested by the BN to directly or conditionally influence the variable of interest, and the remaining variables outside of the MN are of little to no interest in this regard. Therefore, MN-contained variables are candidates for further biological (analytical, literature, or experimental) follow-up. Reconstructing BNs from flat data is computationally demanding. A typical BN analysis of a dataset in this study required 1–4 d on a modern workstation. There are also memory limits. For these reasons full BNs for only chromosomes 1 and 2 are presented (*SI Appendix*, Fig. S1 *C* and *D*). However, the BN analysis is of course vastly less time and effort consuming than experimental methods.

When interpreting BNs, edge (dependency) strength is important and is designated by line thickness and the number next to the edge. The number is similar to a basic likelihood-ratio test statistic, in that it is proportional to the ratio of the model fit of the BN with an edge in question to the one without it. It is difficult to evaluate in absolute terms (e.g., generate a *P* value). However, the numbers within the network (and across the networks, in this study) can be directly compared with each other, with a higher number indicating a stronger dependency (or, in other terms, nonparametric statistical correlation). Consequently, if the investigator knows that the link between two certain variables is indeed strong, corresponding edge strength can be used as a benchmark. The edge directionality ("arrows") in the presented results is strictly arbitrary, necessary for mathematical tractability only, and should be essentially ignored.

It is important to stress that in its pure form, as done here, the BN approach is strictly data driven and independent of the investigators' input; for example, selection of the variable of interest does not make that variable different from the others ("dependent variable" in a regression or classification sense) and in the complete BN such selection is basically for visualization and convenience purposes only. Most of our BN analyses were carried out using the full list of variables (64 transcription factors; 100+ variables in total, depending on the analysis and actual primary variable of interest).

We used three-bin maximum-entropy–based discretization for continuous variables, including interaction strength. Previously we have shown (12) that such discretization is optimal with respect to preserving the existing biological signal (dependency, correlation) while minimizing spurious noise. We have experimented with other sensible binnings, and the network topology was robust to changes in discretization mechanism.

Because the complete BN is difficult to visualize, in Fig. 1*B* only the MN of the hic_strength variable is shown. However, it is important to understand that this MN is a subset of the complete BN, not just a smaller BN built from selected variable sets. The complete BN is visualized in *SI Appendix*, Fig. S1 *C* and *D* and is also available directly from the authors as a pdf file and a source code (dot format) compatible with many network and graph

visualization software packages. Interested readers can parse the file or enlarge the figure (using any standard pdf viewer) to thoroughly investigate MNs of specific nodes/variables. Analyses done so far have been chromosome dataset specific [chromosome 1 (chr1) in Fig. 1*B* and *SI Appendix*, Fig. S1*C* and chromosome 2 (chr2) in *SI Appendix*, Fig. S1 *A–D*]. This brings up the issue of scalability in terms of data points (approximately 10 million intrachromosomal interactions). More data points are available (hundreds of millions), but using them would substantially complicate BN reconstruction implementation (predominantly due to computer memory issues).

An important feature of Hi-C datasets is the location-dependent "geographic structure" of the data. Therefore, it is possible, for example, to limit analysis to the interactions that are less than a predefined distance or interactions located within the Hi-C loops. In addition to making biological sense, the advantage of such restrictions is a decrease in memory requirements and computational time without sacrificing sensitivity and specificity. Given the above nuances, numerous BNs can be inferred from the same primary datasets. For example, *SI Appendix*, Fig. S1*A* depicts the MN of the hic_strength variable in chr2 derived from the dataset containing interactions within Hi-C loops only, whereas the MN in *SI Appendix*, Fig. S1*B* reflects the unconstrained dataset. *SI Appendix*, Fig. S1 *C* and *D* shows full BNs for chr1 and chr2, respectively, derived from unconstrained datasets. The BN for the MN shown in Fig. 1*B* was derived from the dataset containing interactions within Hi-C loops only. In general, our results appear to be robust to the algorithmic variations, thus suggesting that the differences between the BNs reflect true biological differences.

BN Analysis Suggests That Intrachromosomal Interaction Strength Directly Depends Only on Four Types of Variables. We first asked whether useful chromosomal structure–function information could be derived just by data-driven BN modeling of a combination of ENCODE protein-binding data and Hi-C DNA–segment interaction data. All TF and nonhistone protein-binding data in the publicly available ENCODE database (30) (ENCODE Data Coordination Center) for the cell line GM12878 were included. In addition to the presence or absence of TFs in 5-kb anchor segments, we included some additional variables such as Tss and other related features. In total, 106 variables were included in our BN analysis (*SI Appendix*, section 5, and Tables S1 and S2). Orientation of the ctcf motif was considered only for the variables forward_between and reverse_between.

The Hi-C dataset used in this study is at a 5-kb resolution and is for the dataset previously used to identify Hi-C loops (8). Only interactions locating within a Hi-C loop smaller than 750 kb in chr1 or chr2 are included. Our primary variable of interest for this study was hic_strength, representing the interaction strength between two genomic loci (anchors) as determined by Hi-C. Of note, as others have done (6, 8), we use O/E (raw observed interaction strength normalized by the expected interaction strength) as the value for the variable hic_strength. The resulting full BN is shown in *SI Appendix*, Fig. S1 *C* and *D* and the MN for hic_strength is shown in Fig. 1*B*.

Fig. 1*B* shows that hic_strength is directly dependent on only 4 of the 106 variables: (*i*) distance between interaction anchors, (*ii*) presence or absence of the protein RAD21 or SMC3 (two components of the cohesin complex) in the interaction anchors, (*iii*) presence or absence of active transcription (TssHmm) in the interaction anchors, and (*iv*) the number of CTCF–cohesin complexes between anchors (reverse or forward_between), which may reflect smaller, internal loops within larger encompassing loops.

For the MN shown in Fig. 1*B*, and the complete BN shown in *SI Appendix*, Fig. S1 *C* and *D*, each component of the cohesin complex (RAD21, SMC3) was treated as a separate variable, and the orientation of the ctcf motif was not considered. We did this for two reasons. First, we wanted to minimize user intervention. Second, consideration of ctcf motif orientation leads to a single variable with four states (left_anchor_forward or reverse and right_anchor forward or reverse), but these states are not independent and thus the relationships are not necessarily resolved optimally by the BN algorithm, given the limited amount of data. BN analysis with orientation included does, however, generate convenient local conditional probability tables that are stratified and sorted for each state. It is one of the principal advantages of BN treatment that this information can be used for subsequent analysis.

Active TSS Are Linked to Stronger Hi-C Interactions. The variable "Tss" designates whether active TSS (CAGE signal, ENCODE) are found within the interaction anchors. For all 5-kb anchors in which the TSS activity is detected, about 35% of them are at more than 1 read per kilobase per million (rpkm). These anchors either overlap with the annotated gene promoter regions or are active enhancers [identified by the coexistence of histone 3 lysine 4 monomethylation (H3K4me1) and histone 3 lysine 27 acetylation (H3K27ac)], which is consistent with previous reports (39, 40). In the dataset used for BN analysis, 2% of total genomic interactions occur between two Tss sites. Among these Tss–Tss interactions, about 40% occur between a gene's promoter region and an active enhancer and 19% are between two different promoter regions.

BN analysis strongly suggested that TSS activity within the interaction anchors is an important variable that influences interaction strength (Fig. 1*B*), and this is consistent with previous reports based on high-resolution analysis of specific chromosomal subregions (13, 27, 41). We found that high-intensity interactions (O/E > 3) are enriched in compartment A, which is the transcriptionally active chromosomal compartment (8, 41); there are 29,940 such interactions mapping to compartment A of chr1 but only 8,222 mapping to compartment B. We next did genome-wide analysis using Hi-C data to study the relationships between the TSS activity in the left and right anchors and their corresponding interaction strength (Fig. 2 *A* and *B*). We found that interaction strength clearly is positively associated with Tss level (Fig. 2*B*). However, it is interesting to note that a higher level of interaction strength is not observed when only one anchor has TSS activity (Fig. 2*A*).

TFs. We found that the hic_strength variable, which was the original focus of this study, is not directly dependent on most TFs or the other variables included in this study (*SI Appendix*, section 5, Tables S1 and S2). This does not mean that other TFs have no influence, but just means that, given Tss information, additional information about TFs is superfluous for hic_strength. In BN parlance, Tss "shields" hic_strength from the TF. Fig. 2*C* shows for several TFs that their binding in an anchor segment affects hic_strength differently, dependent on whether Tss activity is detected. Additional inspection of MNs for hic_strength as well as full BNs for chr1 and chr2 (Fig. 1*B* and *SI Appendix*, section 1) did in fact suggest several other potentially important relationships, such as the connection between Rad21 and Znf143 (Fig. 3), the connections with Ebf1 (Fig. 3 and *SI Appendix*, Fig. S1*B*), and the connection between Rad21 and Ikzf1 (Fig. 1*B*). Both Ebf1 and Ikzf1 are known to be important, lineage-determining factors for B cells, and GM12878 is a B-cell lymphoma. It is worth noting that cell type came to our attention only after BN modeling results identified Ebf1 as a potentially important TF. The interaction between Rad21 and Znf143 is a previously known interaction, serving to validate our BN analysis.

To further illustrate the use of the BN, a MN centered on Ebf1 was generated (Fig. 3 *A* and *B*), and several interesting

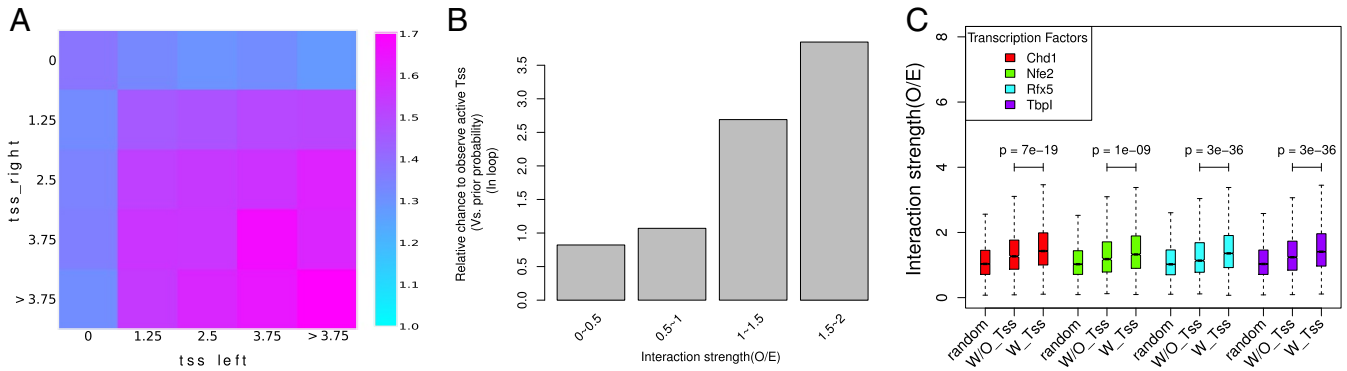

Fig. 2. TSS activity affects interaction strength. (*A*) Heatmap showing that TSS activity within the two interacting anchors is positively associated with interaction strength. The color gradient represents the average interaction strength. Tss level is in units of rpkm. (*B*) Interaction strength affects the chance to observe Tss–Tss interactions. *y* axis represents the relative chance of observing the TSS activity (>5 rpkm) at one anchor given the corresponding interaction strength (*x* axis) and Tss (>5 rpkm) in the other anchor. (*C*) Interaction strength between two anchors with at least one occupied by a TF decreases if no TSS activity is associated with these anchors. W/O_Tss: without Tss. W_Tss: with Tss. *P* values were calculated by a Kolmogorov–Smirnov test. The "random" sample had a similar "distance" distribution to the target sample but was sampled randomly from the whole population.

known and potentially new interactions emerge. For example, the known strong interaction of Rad21 and Smc3 is clear. Also for both left and right anchors (Fig. 3 *C* and *D*) there is a three-way dependency between Rad21, Znf143, and Ebf1. These relationships are not addressed in any detail here as they are beyond the scope of this study, but it should be noted that they were identified by unbiased purely data-driven BN analysis and thus may warrant additional investigation both in silico and on an experimental level. The potential dependencies between Ikzf1, Rad21, and Ebf1 are also of interest. In BN analyses, although conditional independence relationships are often equivocal, dependencies are usually meaningful. With this in mind, several other potentially interesting relationships are revealed in Fig. 3 *A* and *B*. Chd2 and Maz are strongly and consistently clustered near Smc3. Chd2 is a helicase with chromatin-remodeling activity (42), and Maz is a well-known TF sometimes involved in transcriptional pausing (43). Also, Ebf1 shows a strong dependency on Bhlhe40, a helix–loop–helix TF known to be involved in immune function (44). These relationships each could, and perhaps should be, addressed in future studies. However, it is noteworthy that these potential interactions were identified by unbiased BN analysis without any input from us.

The Interaction Between Two Convergent CTCF Pairs Is Stronger Than in Other Combinations. BN analysis clearly shows the dependence of hic_strength on RAD21 or SMC3 (Figs. 1*B* and 3). BN analysis also consistently shows a strong dependency of CTCF on RAD21 and SMC3, two proteins known to be major components of the cohesin complex (*SI Appendix*, Fig. S1 *C* and *D*) (45). In subsequent analysis described next, we further analyzed the CTCF–cohesin complex, that is, sites that have all three proteins bound. Importantly, a CTCF–cohesin complex at a ctcf motif has two directions, with the ctcf motif either in the F or in the R direction (8). As mentioned earlier, a pair of CTCF–cohesin complexes have four different orientation patterns: F–R (convergent), R–F (divergent), F–F, and R–R.

Prior ChIA-PET data obtained after enrichment using anti-CTCF antibodies led to the conclusion that loops enriched for the convergent ctcf pairs have a higher frequency than the other

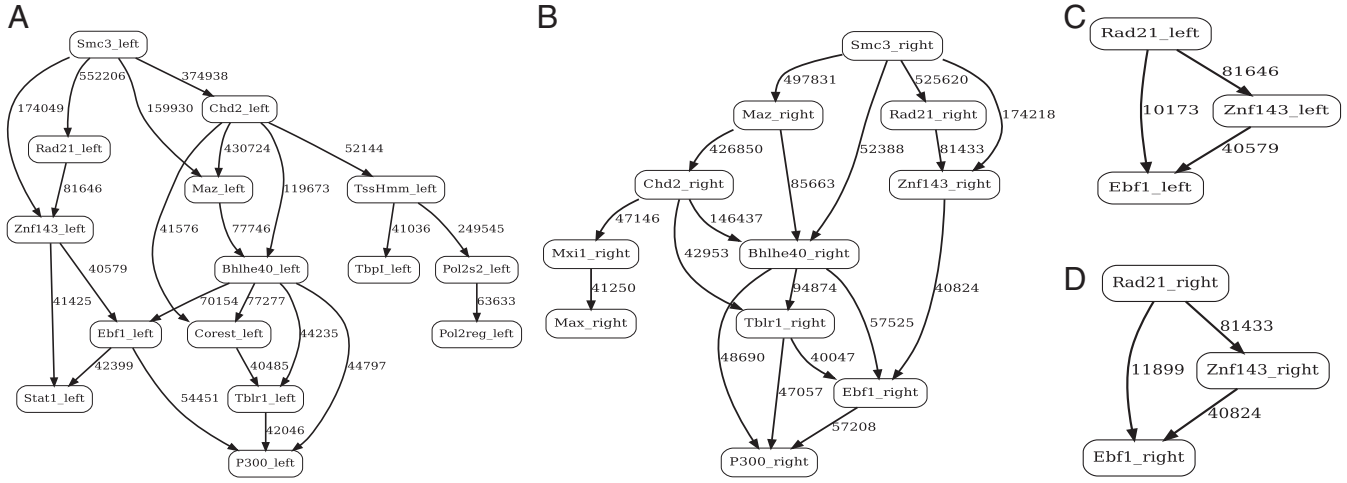

Fig. 3. MNs of Ebf1 variable node, separated into left and right anchors. (*A*) Extended MN of the Ebf1 left variable in the BN derived from the chr1 unrestricted dataset. (*B*) Same as in *A*, for Ebf1 right. (*C*) Visualization of a trivariate interaction between Ebf1, Rad21, and Znf143 variables, left anchor. (*D*) Same as in *C*, right anchor. See Fig. 1*B* and *SI Appendix*, section 5, Tables S1 and S2 for general BN designations and principal variable descriptions. Note that dependency strength is shown as a number (proportional to the likelihood ratio, see text for details) next to the corresponding edge in the network. Only edges above 40,000 in strength are shown in Fig. 3 *C* and *D* for easier network readability.

combinations (23, 24). In our study, we addressed this same question by first identifying ctcf motifs from the regions co-occupied by CTCF, RAD21, and SMC3 and then combining this information with published, genome-wide Hi-C data (8). We found that convergent CTCF–cohesin complexes indeed have higher than average interaction strength (Fig. 4*A*). Second, we found that the interactions between two CTCF–cohesin complexes are stronger for those located within a Hi-C loop than for those crossing a Hi-C loop boundary. Moreover, as Hi-C loops are generally smaller and located within TADs, some convergent CTCF–cohesin complexes are found within the same TAD but crossing the Hi-C loop boundaries. In detail, for the 304 convergent CTCF–cohesin complex pairs in chr2 that cross the boundary of Hi-C loops, 199 are located within a single TAD. Those CTCF–cohesin complex pairs that cross a Hi-C loop boundary have significantly lower contact intensity than those not crossing. Overall, the numbers of the four orientation combinations for CTCF–cohesin complexes are approximately equal, suggesting random orientation. However, if one examines CTCF pairs restricted to Hi-C loop regions, convergent CTCF–cohesin complexes are highly enriched, which may indicate a clustering of same-orientation CTCF–cohesin complexes within the Hi-C loops (Fig. 4*B*).

CTCF–Cohesin Complexes Specify the Direction and Distribution of Long-Range High-Intensity Genomic Interactions. Since convergent CTCF–cohesin complexes are overrepresented in high-intensity interactions, they are likely to affect the distribution of other high-intensity interactions. Thus, we next examined whether the anchors of high-intensity interactions in the neighborhood of a CTCF–cohesin complex show a nonrandom spatial relationship with respect to the orientation of the ctcf motif in the CTCF–cohesin complex. For this study, high-intensity interactions are defined as those that have an O/E value greater than 96% of total interactions; for chr2 this is O/E $>$ 3. We define a neighboring region as 25 kb upstream or downstream of a CTCF–cohesin complex, binned into 5-kb segments, with upstream or downstream being determined by the standard chromosomal DNA sequence numbering system. First, we found that the anchors of high-intensity interactions are indeed enriched within regions at or near CTCF–cohesin complexes, with a peak centered at the CTCF–cohesin site (Fig. 5*A*). We then categorized all these high-intensity interactions into three classes according to their anchors' relationship with the neighboring regions of CTCF–cohesin complexes. For class 1 interactions (22% of total high-intensity, O/E $>$ 3 interactions), the left anchor is located within the neighboring regions of a CTCF–cohesin complex with an F motif and the right anchor is located within the neighboring region of a CTCF–cohesin complex with an R motif. For class 2 interactions (46% of total high-intensity interactions), either the left anchor is located within the neighboring regions of a CTCF–cohesin complex with an F motif or the right anchor is located within the neighboring regions of a CTCF–cohesin complex with an R motif, but not both. Class 3 interactions are the remaining high-intensity interactions with neither anchor in a CTCF–cohesin neighboring region. We find that the first two classes constitute 68% of all high-intensity interactions. It should be kept in mind that 78% of high-intensity interactions (O/E $>$ 3) are not between two CTCF–cohesin complexes; the majority of these have a CTCF–cohesin complex in only one anchor. For all annotated Hi-C loops (8), not just those with high-intensity (O/E $>$ 3) interactions, also about 22% (2,857/12,903) are class 1, with both anchors containing ctcf motifs in convergent orientation. We obtained a similar ratio from the chr2 dataset. Only 24% (178/706) of annotated Hi-C loops in chr2 have a unique F motif in the left anchor and a unique R motif in the right anchor. We note that many high-intensity interactions, as well as annotated loops, do not have a convergent ctcf motif in both anchors. Also up to 62% of total identified CTCF–cohesin complexes are not associated with the anchor regions of a Hi-C loop.

We next investigated the effect of ctcf motif orientation on the distribution of high-intensity interactions. Fig. 5*B* shows, for 5-kb bins near a forward CTCF–cohesin complex, the probability of the bin containing either a left anchor (red curve) or a right anchor (blue curve) of a high-intensity interaction (O/E $>$ 3), with the other anchor being at any distance. Fig. 5*C* shows a similar plot for an R CTCF–cohesin complex. Note that ctcf orientation and left or right anchor designation are based on the standard chromosomal nucleotide base numbering convention, not their relative orientation. Using Fig. 5*B* as an example, a left anchor located in a 5-kb segment containing a ctcf F motif (and a bound CTCF–cohesin complex) indeed does have the highest probability of interacting with a downstream anchor at high intensity. This is consistent with Fig. 5*A*. Importantly, however, the profiles seen in Fig. 5*B* are strongly dependent on the orientation of the ctcf motif. In Fig. 5*B*, the probability of finding a segment containing the left anchor of a high-intensity interaction near an F motif is much higher than the probability of finding a segment containing the right anchor. This dramatically different pattern, which is seen on all chromosomes (*SI Appendix*, section 6), cannot be easily explained by interactions resulting from random diffusion, but, as illustrated in Fig. 5*D*, is consistent with DNA-reeling/extrusion models with an appropriately

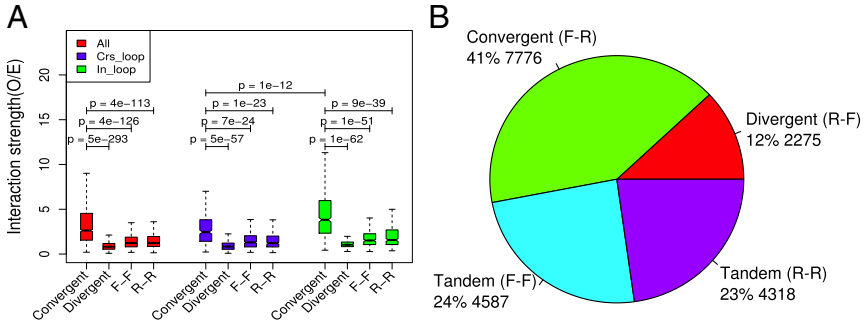

Fig. 4. Orientation of convergent CTCF–cohesin complexes affects interaction strength. F: The CTCF–cohesion complex is in the forward orientation. R: The CTCF–cohesin complex is in the R orientation. (*A*) Convergent CTCF–cohesin pairs (F_R) interact more strongly compared with the other orientations. In_loop: The two anchors (containing CTCF–cohesin complexes) of an interaction are in the same loop. Crs_loop: The two anchors cross the loop boundaries. (*B*) Genome-wide, convergent CTCF–cohesin complex pairs that are within loops (8) are more frequent than the other orientation combinations. Overall, if Hi-C loops are not selected, the four categories of ctcf pairs occur in about equal numbers: F–R, 23,836, 24%; R–F, 25,935, 26%; F–F, 24,709, 25%; R–R, 24,283, 25%.

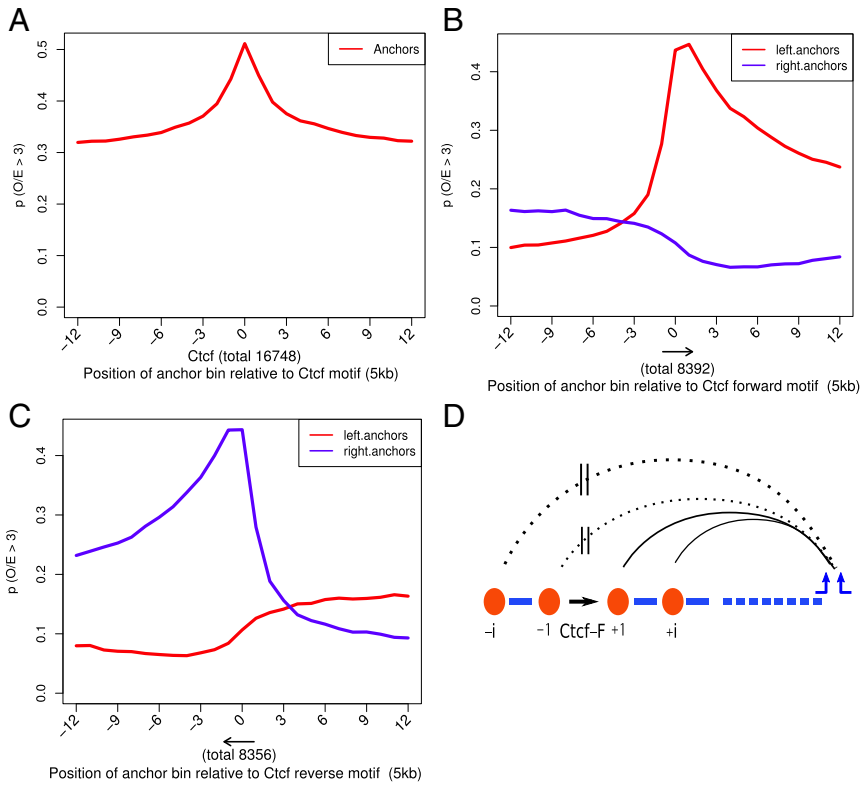

Fig. 5. CTCF–cohesin complexes affect the distribution and direction of high-intensity intrachromosomal interactions. (*A*)The probability profile for 5-kb segments neighboring CTCF–cohesin complexes containing the anchors of high-intensity interactions. The *y* axis has the same meaning in *A*–*C*. (*B*) The probability profile for 5-kb segments neighboring the CTCF–cohesin complexes with F motifs containing the left or right anchors of a high-intensity interaction. (*C*)The probability profile for 5-kb segments neighboring the CTCF–cohesin complexes with R motifs containing the left or right anchors of a high-intensity interaction. (*D*) The formation of an asymmetrical distribution in *B* can be explained by a DNA-reeling/extrusion model. In this model, reeling and loop formation initiated downstream will be terminated by an F CTCF–cohesin complex.

oriented CTCF–cohesin complex acting as a strong barrier to reeling/extrusion that begins downstream.

CTCF–Cohesin Complexes Specify the Choice of Targets in EP Interactions. EP interactions with high interaction strength are of particular interest. Using Hi-C and Encode datasets (30), we selected a group of EP interactions based on the following criteria (39, 41): (*i*) The enhancer is occupied by the transcriptional activator P300 and is marked with H3k4me1 and H3k27ac, (*ii*) the gene promoter is active (Tss > 0), (*iii*) the interaction strength between the enhancer and promoter is three times higher than the expected background level (O/E > 3), and (*iv*) the EP interactions locate within a Hi-C loop. Since the enhancer can be either upstream or downstream of the promoter, we named these two interaction sets "EP_upstream" and "EP_downstream," respectively. ("Upstream" or "downstream" in the name denotes the position of the enhancer with respect to the promoter in the EP and is not related to transcription direction. The complete list of promoter–enhancer interactions can be found in Dataset S1.) Based on previous studies, these enhancers are active and cell-type specific (39). Of interest, we found more than 95% of EP interactions have Ebf1 in either the enhancer or the promoter. For EPs in chr1 (defined above), about 79% (709/899) of promoters and 84% (751/899) of enhancers have Ebf1 binding.

We found that the enhancers in the EP_upstream set do not significantly overlap with the enhancers in the EP_downstream set. If an enhancer could sometimes choose an upstream promoter and sometimes a downstream promoter to interact at high intensity, one would expect considerable overlap. This is not the case, so clearly EP interactions are not random but are directionally biased. In more detail, we found that in chr2, of 214 and 156 enhancers interacting with downstream and upstream promoters, respectively, only 8 interact with both, thus making the vast majority of enhancers directional.

Importantly, EP interactions genome-wide clearly show directional bias related to the CTCF–cohesin complex orientation. Fig. 6*A* shows that enhancers located upstream of promoters ("En left") are enriched within the neighboring regions of an F CTCF–cohesin complex, whereas for downstream enhancers ("En right") (Fig. 6*B*) the enrichment is within the R CTCF–cohesin complex. Of note, our criteria for identification of these EP interactions do not include the presence of either ctcf sequence motifs or CTCF–cohesin complexes. After identifying the EP interacting segments, they were then interrogated for CTCF–cohesin complexes. These results are consistent with appropriately oriented CTCF–cohesin complexes strongly influencing the formation and/or stability of EP loops. It is also noteworthy that the asymmetry seen in Fig. 6*A*, for example, is similar to that seen in Fig. 5*B* and can be explained similarly. A promoter downstream of an enhancer is blocked from "crossing" a forward CTCF–cohesin complex to participate in a high-intensity EP interaction. These results are thus consistent with ctcf-containing elements being able to act as insulators by being a barrier to EP loop formation, perhaps by terminating DNA reeling (13, 21, 22).

Convergent CTCF–Cohesin Complexes Increase Interaction Strength by Forming Intermediate Structures, Probably Loops. Our BN analysis (Fig. 1*A*) showed that the interaction strength variable

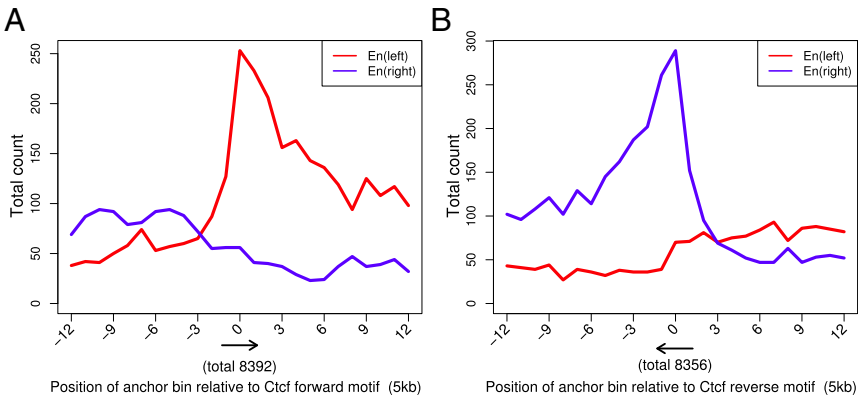

Fig. 6. CTCF–cohesin complexes affect EP interactions. (*A*) The genome-wide occurrence of upstream enhancers (En left) and downstream enhancers (En right) within the 5-kb segments neighboring CTCF–cohesin complexes with F motifs. (*B*) The Same as in *A* but with the R CTCF–cohesin complex. Note that upstream and downstream follow the standard chromosomal base-numbering convention relative to the interacting promoters, not related to the transcription direction. A similar finding is obtained by targeting the high-intensity interactions crossing loop boundaries (*SI Appendix*, section 3 and Fig. S3 *A*–*C*).

(hic_strength) is dependent on the variables Forward_between and Reverse_between (representing the number of F and R CTCF–cohesin complexes between the interaction anchors). How can this be explained? A large fraction of CTCF–cohesin complexes (62% of total) are not associated with the endpoints of the Hi-C loop structures, TADs, or other so far identified chromosomal substructures (8, 15), and our analysis of Hi-C datasets revealed that 78% of high-intensity interactions contain between the anchors one or more CTCF–cohesin complexes. Thus, many loops contain within them additional CTCF–cohesin complexes. Whether the interaction strength between two genomic segments, for example those identified as loop (Hi-C loops) anchors, is affected by convergent CTCF–cohesin complexes located between them was not known. Given that convergent CTCF–cohesin complexes are likely to form or stabilize loops, we hypothesized that these intermediate structures may bring two bracketing anchor segments into closer proximity, thereby increasing the probability of interaction. For this reason, we introduced a variable that we name reduced distance (RD) to model effects caused by the convergent CTCF–cohesin complexes. RD is the distance remaining after subtracting the length of DNA in potential loops demarcated by the convergent CTCF–cohesin complexes. In the example shown in Fig. 7*A*, two genomic loci i and j encompass one pair of convergent CTCF–cohesin complexes. We found that the interaction strength between two genomic anchors whose RD is more than half of the linear distance is significantly higher than the background (Fig. 7*B*).

Discussion

BN modeling has been widely used for analyzing the complex relationships within the "big data" repositories generated in modern biomedical research, but to the best of our knowledge BNs have not been applied to a study of intrachromosomal 3D structure. Here, we used newly developed software (12, 36) and asked the questions, What factors influence chromosomal 3D structure as measured by the probability of contact between distantly located segments of DNA? And can BNs help identify these factors? We applied BN modeling to analyze the relationships between the Hi-C–derived intrachromosomal interaction strength and various genetic elements, interacting proteins, and other variables for which detailed information is available in the publicly available Encode database. A primary conclusion is that the BN analysis works well in this application.

The advantages of BN modeling over more traditional analysis methods (such as univariate statistical approaches, regression, classification, and clustering) are fivefold: (*i*) The entire biological network underlying the observed data is reconstructed, allowing one to model and visualize mechanistic underpinnings of the chromatin biology; (*ii*) such networks are immediately useful for both testing existing hypotheses and automatically generating novel ones; (*iii*) heterogeneous variables can be incorporated within the same analysis framework (a single network) without information loss due to type conversion and violated distributional assumptions; (*iv*) investigators can "switch" from scrutinizing one variable/node within the network to another without carrying out the analysis de novo; and (*v*) resulting networks (and generated hypotheses) can be validated, and compared, using simple built-in instruments and criteria (statistical resampling, localized likelihood-ratio tests).

Our BN analysis identified only four categories of factors directly related to interaction strength. These categories are distance, cohesin complex components (e.g., Rad21), TSS activity, and the number of CTCF–cohesin complexes between anchors. Finding that baseline BN modeling highlights CTCF and Rad21 as important variables serves to validate the approach. Inspection of Fig. 1*B* also shows that given the above categories, interaction strength is conditionally independent of TFs, with only one exception, Izkf1. Ikzf1 is known to be an important TF for hematopoietic cell differentiation (46), so it may not be just a coincidence that the Hi-C and Encode data we used were obtained from a B-cell lymphoma. We are not aware of studies implicating a relationship between Ikzf1 and RAD21 (or the cohesin complex), so this is an example of insight obtained by BN analysis. In Figs. 1*B* and 3 *C* and *D* a connection is seen between Znf143 and Rad21, which is a known interaction (27), again serving to validate our approach. As shown in *SI Appendix*, Fig. S1 *C* and *D*, we also observe a connection between Tss and Ebf1, a known lineage-determining B-cell TF. It is likely that Tss is "shielding" hic_strength from most TFs because given Tss information additional information about TFs is superfluous for hic_strength. Nevertheless, when the MN is refocused on a TF, dependency relationships for that factor emerge. For example, if one examines the MN around Ebf1 (Fig. 3 *C* and *D*), a three-way dependency is seen between Rad21, Znf143, and Ebf1. The dependencies between Rad21 and Znf143 and between Ebf1 and Znf143 are stronger than the dependency between Ebf1 and Rad21, so it is likely that the effect of Rad21 on Ebf1 is mostly indirect, through its effect on the interaction of Znf143 with Ebf1. This illustrates the type of nuanced insight that can be obtained from BN analysis of ChIP-seq datasets for numerous TFs. As another example, there is

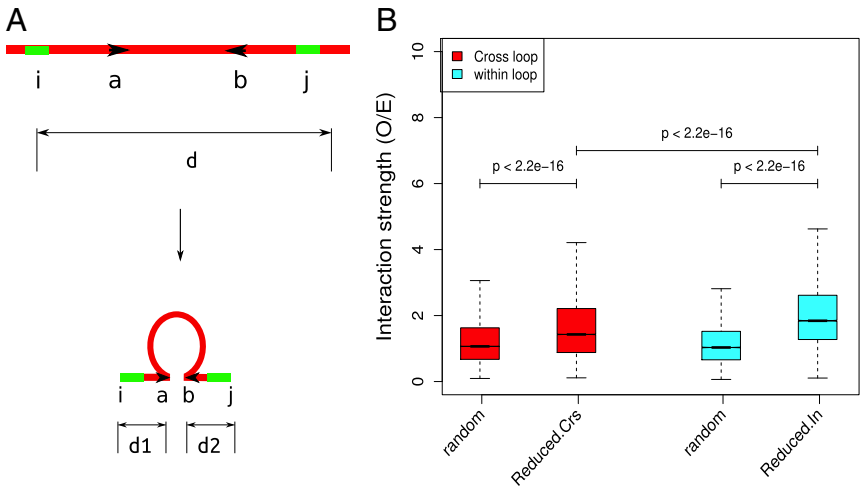

Fig. 7. Convergent CTCF–cohesin complex pairs affect interaction strength via a "reduced distance" (RD) effect. (*A*) An example to show the principle of RD. The genomic anchors i and j are separated by a pair of CTCF–cohesin complexes with convergent direction. If the loop is formed between a and b, the distance between i and j changes from d to d1 + d2. (The calculation of the RD is shown in detail in *SI Appendix*, section 4). (*B*) The interactions that are affected by the RD have higher interaction strength than average.

also a three-way dependency between Maz, Chd2, and Bhlehe40 (Fig. 3 *A* and *B*).

With regard to BN software development, so far we have constructed full BNs for only two chromosomes, each taking several days of computer processing time. Furthermore, resampling can also be used (together with relative edge strengths) to evaluate and validate the BNs via cross-validation or bootstrapping, at the cost of computational efficiency. Finally, one can simulate a number of artificial variables with gradually increasing known dependency strength to calibrate all of the edges in the BN. This approach, although time consuming and domain specific, is known to be effective ("artificial positive controls") (36).

Numerous chromatin interaction studies, including Hi-C, have established that chromatin is highly organized in 3D space by distant DNA segments being brought physically close together, thus necessarily looping out the DNA between them (13). Moreover, primarily based on the directionality of CTCF–CTCF and EP interactions (24), it is becoming increasingly evident that DNA looping does not result from specific interactions formed by random diffusion in 3D space but rather from a mechanism that acts in one dimension along the DNA (13, 21, 22). Our BN results shown in Fig. 1 highlight the importance of the CTCF–cohesin complex, and results shown in Fig. 5 provide strong evidence that directionality of loop formation is determined by the orientation of the ctcf motif in a CTCF–cohesin complex. As others have discussed and reviewed (13, 21, 22), such directionality is very difficult to explain by diffusion-based mechanisms. More likely a molecular motor is pulling the DNA segments together, that is, reeling in DNA and extruding a loop between them (18, 21, 22). Sanborn et al. (22) and Fudenberg et al. (21) have developed computer models assuming that cohesin serves as a molecular motor, with a cohesin dimer binding between convergent ctcf motifs and the CTCF complex serving to terminate reeling/extrusion. With these assumptions, computer-derived interaction patterns match well with experimentally determined Hi-C interactions. As illustrated in Fig. 5*D*, the results of our analyses are consistent with an appropriately oriented CTCF complex acting as a barrier to loop formation.

Our BN analysis highlighted the dependence of hic_strength on active transcription at start sites (TSS), and this was confirmed by our subsequent analysis showing that interaction strength does increase with transcription level at TSS (Fig. 2 *A* and *B*). Previously, Tang et al. (24) used ChIA-PET to study RNA polymerase (Pol II)-containing chromatin interaction anchors and compared these data with Hi-C data. They found that Pol II is frequently associated with CTCF at the base of Hi-C loops, with housekeeping genes located near the base of CTCF–cohesin loops and cell-type–specific genes more centrally located within the loops. They also found that clusters of CTCF complexes, called CTCF foci, colocalize with foci of Pol II, which have been called transcription factories (21, 47). As a result, they suggested that transcription by Pol II selectively draws genes into these CTCF foci. Fig. 6 very clearly shows that strong EP interactions are sensitive to the orientation of CTCF–cohesin complexes, with directionality very similar to that seen for the high-intensity interactions in Fig. 5. We thus speculate, as have others (24, 48), that transcription may sometimes be part of the process bringing distant DNA elements into close physical contact. Whatever the motor driving loop formation, Fig. 6 shows that the directionality of EP interactions is strongly influenced by the orientation of the CTCF complex. We interpret this finding to be consistent with models in which an appropriately oriented CTCF complex terminates reeling/loop extrusion (21, 22), whether derived from cohesin complex reeling, transcription, or other mechanisms. Another, not mutually exclusive, model is that the CTCF complex initiates reeling near the ctcf site, as proposed by Nichols and Corces (49). In either case, the CTCF complex can act as a barrier to reeling/extrusion in the "wrong" direction, thereby serving to insulate a promoter from the influence of an enhancer.

It should be noted that only 22% of O/E > 3 anchor pairs have a CTCF–cohesin complex at both anchors; 32% do not have a CTCF–cohesin complex at either anchor. Thus, most loops do not have CTCF at both anchors, suggesting that reeling/extrusion can be stopped or paused by protein complexes or structures that do not contain CTCF. Earlier studies in yeast found that cohesin was found mostly between sites of convergent transcription and it was suggested that transcription can push cohesin, causing it to redistribute on the chromosome (50). Very recently, while our manuscript was under review, Busslinger et al. (51) and Haarhuis et al. (52) reported that the distribution and directionality of movement of cohesin in the mouse genome is influenced by transcription. These results were interpreted as supporting loop/extrusion models (21, 22), with CTCF being a boundary element limiting the movement of cohesin. It thus seems likely that some high-intensity interactions may be due to direct or indirect consequences of transcription.

Finally, BN analysis indicated that hic_strength is directly dependent on the number of right- or left-oriented CTCF–cohesin complexes located between the interaction anchors (forward_between and reverse_between in Fig. 1). How can this be explained? We suggest, as have others (53), that convergent CTCF-cohesin pairs located between the anchors can form internal loops, thereby reducing the apparent distance.

In conclusion, we have used recently developed BN methodology and software for an investigation of how various factors affect interaction strength between distant chromosomal anchors. BN results highlighted the importance of several factors, some of which were expected, others not. These findings generated hypotheses that were used to guide further data analysis.

Materials and Methods

Constructing BNs and Mapping Encode Data. BN analysis was performed as in ref. 12. We selected the Hi-C datasets at 5-kb resolution from Encode datasets for the cell line GM12878. A total of 64 TFs (Encode project) were mapped to 5-kb–sized bins corresponding with the Hi-C map. The data were arranged into a table in which each row represents the state of a specific interaction between any two anchors and each column (variable) represents whether a specific protein binds at the anchors (upstream or downstream). We set "resolution" at delta = 150 bins (equaling 750 kb).

CTCF–Cohesin Complexes. We used HOMER (54) to search for the ctcf motifs in 5-kb loci co-occupied by CTCF, RAD21, and SMC3 ChIP-seq signals. Specific loci with more than one ctcf motif (only 2% of total) were labeled "forward" or "reverse," depending on which orientation was more frequent. Additional details are in *SI Appendix*, Fig. S2 and section 2.

Normalization of Hi-C Data and the Variables for Interaction Strength. Two values, termed O/E and raw observed, were used to represent the genomic interaction strength, following previously published methods (8). BN analysis was based on O/E. Other results were generated using both O/E and raw-observed values (*SI Appendix*, section 6), but the latter is shown in the main text.

ACKNOWLEDGMENTS. This work was supported by the Susumu Ohno Chair in Theoretical and Computational Biology (held by A.S.R.), a Susumo Ohno Distinguished Investigator fellowship (to G.G.), and City of Hope funds (to A.D.R.).

1. Dekker J, Marti-Renom MA, Mirny LA (2013) Exploring the three-dimensional organization of genomes: Interpreting chromatin interaction data. *Nat Rev Genet* 14:390–403.
2. Deng W, et al. (2012) Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* 149:1233–1244.
3. Shi J, et al. (2013) Role of SWI/SNF in acute leukemia maintenance and enhancer-mediated Myc regulation. *Genes Dev* 27:2648–2662.
4. Levine M, Cattoglio C, Tjian R (2014) Looping back to leap forward: Transcription enters a new era. *Cell* 157:13–25.
5. Dekker J, Rippe K, Dekker M, Kleckner N (2002) Capturing chromosome conformation. *Science* 295:1306–1311.
6. Lieberman-Aiden E, et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326:289–293.
7. Zhang J, et al. (2012) ChIA-PET analysis of transcriptional chromatin interactions. *Methods* 58:289–299.
8. Rao SS, et al. (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159:1665–1680.
9. Friedman N, Linial M, Nachman I, Pe'er D (2000) Using Bayesian networks to analyze expression data. *J Comput Biol* 7:601–620.
10. Rodin AS, Mosley TH, Clark AG, Sing CF, Boerwinkle E (2005) Mining genetic epidemiology data with Bayesian networks application to APOE gene variation and plasma lipid levels. *J Comput Biol* 12:1–11.
11. Pe'er D (2005) Bayesian network analysis of signaling networks: A primer. *Sci STKE* 2005:pl4.
12. Gogoshin G, Boerwinkle E, Rodin AS (2016) New algorithm and software (BNOmics) for inferring and visualizing Bayesian networks from heterogeneous "big" biological and genetic data. *J Comput Biol* 23:1–17.
13. Dekker J, Mirny L (2016) The 3D genome as moderator of chromosomal communication. *Cell* 164:1110–1121.
14. Sexton T, et al. (2012) Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* 148:458–472.
15. Dixon JR, et al. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485:376–380.
16. Lupianez DG, et al. (2015) Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* 161:1012–1025.
17. Hnisz D, et al. (2016) Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* 351:1454–1458.
18. Riggs AD (1990) DNA methylation and late replication probably aid cell memory, and type I DNA reeling could aid chromosome folding and enhancer function. *Philos Trans R Soc Lond B Biol Sci* 326:285–297.
19. Alipour E, Marko JF (2012) Self-organization of domain structures by DNA-loop-extruding enzymes. *Nucleic Acids Res* 40:11202–11212.
20. Siomos MF, et al. (2001) Separase is required for chromosome segregation during meiosis I in Caenorhabditis elegans. *Curr Biol* 11:1825–1835.
21. Fudenberg G, et al. (2016) Formation of chromosomal domains by loop extrusion. *Cell Rep* 15:2038–2049.
22. Sanborn AL, et al. (2015) Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci USA* 112:E6456–6465.
23. Guo Y, et al. (2015) CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. *Cell* 162:900–910.
24. Tang Z, et al. (2015) CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell* 163:1611–1627.
25. Hnisz D, et al. (2015) Convergence of developmental and oncogenic signaling pathways at transcriptional super-enhancers. *Mol Cell* 58:362–370.
26. Uuskula-Reimand L, et al. (2016) Topoisomerase II beta interacts with cohesin and CTCF at topological domain borders. *Genome Biol* 17:182.
27. Heidari N, et al. (2014) Genome-wide map of regulatory interactions in the human genome. *Genome Res* 24:1905–1917.
28. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. *J Mach Learn Res* 3:1157–1182.
29. Rodin AS, Gogoshin G, Boerwinkle E (2011) Systems biology data analysis methodology in pharmacogenomics. *Pharmacogenomics* 12:1349–1360.
30. Dunham Iea (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74.
31. Landt SG, et al. (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* 22:1813–1831.
32. Liu CC, et al. (2014) DiseaseConnect: A comprehensive web server for mechanism-based disease-disease connections. *Nucleic Acids Res* 42:W137–W146.
33. Lo LY, Wong ML, Lee KH, Leung KS (2015) High-order dynamic Bayesian network learning with hidden common causes for causal gene regulatory network. *BMC Bioinformatics* 16:395.
34. Li R, et al. (2016) Identification of genetic interaction networks via an evolutionary algorithm evolved Bayesian network. *BioData Min* 9:18.
35. Wright S (1934) The method of path coefficients. *Ann Math Stat* 5:161–215.
36. Rodin AS, Gogoshin G, Litvinenko A, Boerwinkle E (2012) Exploring genetic epidemiology data with Bayesian networks. *Handbook Stat* 28:479–510.
37. Pearl J (2009) *Causality: Models, Reasoning and Inference* (Cambridge Univ Press, Cambridge, UK).
38. Heckerman D (1995) Tutorial on learning with Bayesian networks (Microsoft Research, Redmond, WA), Technical Report MSR-TR-95-06.
39. Heintzman ND, et al. (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* 459:108–112.
40. Kim TK, et al. (2010) Widespread transcription at neuronal activity-regulated enhancers. *Nature* 465:182–187.
41. Jin F, et al. (2013) A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* 503:290–294.
42. Liu JC, Ferreira CG, Yusufzai T (2015) Human CHD2 is a chromatin assembly ATPase regulated by its chromo- and DNA-binding domains. *J Biol Chem* 290:25–34.
43. Gromak N, West S, Proudfoot NJ (2006) Pause sites promote transcriptional termination of mammalian RNA polymerase II. *Mol Cell Biol* 26:3986–3996.
44. Yang XO, et al. (2009) Requirement for the basic helix-loop-helix transcription factor Dec2 in initial TH2 lineage commitment. *Nat Immunol* 10:1260–1266.
45. Ong CT, Corces VG (2014) CTCF: An architectural protein bridging genome topology and function. *Nat Rev Genet* 15:234–246.
46. Somasundaram R, Prasad MA, Ungerback J, Sigvardsson M (2015) Transcription factor networks in B-cell differentiation link development to acute lymphoid leukemia. *Blood* 126:144–152.
47. Feuerborn A, Cook PR (2015) Why the activity of a gene depends on its neighbors. *Trends Genet* 31:483–490.
48. Papantonis A, Cook PR (2011) Fixing the model for transcription: The DNA moves, not the polymerase. *Transcription* 2:41–44.
49. Nichols MH, Corces VG (2015) A CTCF code for 3D genome architecture. *Cell* 162:703–705.
50. Lengronne A, et al. (2004) Cohesin relocation from sites of chromosomal loading to places of convergent transcription. *Nature* 430:573–578.
51. Busslinger GA, et al. (2017) Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl. *Nature* 544:503–507.
52. Haarhuis JHI, et al. (2017) The cohesin release factor WAPL restricts chromatin loop extension. *Cell* 169:693–707.
53. Doyle B, Fudenberg G, Imakaev M, Mirny LA (2014) Chromatin loops as allosteric modulators of enhancer-promoter interactions. *PLoS Comput Biol* 10:e1003867.
54. Heinz S, et al. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 38:576–589.