Published online 23 August 2017 *Nucleic Acids Research, 2017, Vol. 45, No. 18* ***10403–10414***
doi: 10.1093/nar/gkx738

Computational characterization of chromatin domain boundary-associated genomic elements

Seungpyo Hong and Dongsup Kim[*]

Department of Bio and Brain Engineering, KAIST, Daejeon, Republic of Korea

Received January 03, 2017; Revised July 18, 2017; Editorial Decision August 10, 2017; Accepted August 14, 2017

ABSTRACT

Topologically associated domains (TADs) are 3D genomic structures with high internal interactions that play important roles in genome compaction and gene regulation. Their genomic locations and their association with CCCTC-binding factor (CTCF)-binding sites and transcription start sites (TSSs) were recently reported. However, the relationship between TADs and other genomic elements has not been systematically evaluated. This was addressed in the present study, with a focus on the enrichment of these genomic elements and their ability to predict the TAD boundary region. We found that consensus CTCF-binding sites were strongly associated with TAD boundaries as well as with the transcription factors (TFs) Zinc finger protein (ZNF)143 and Yin Yang (YY)1. TAD boundary-associated genomic elements include DNase I-hypersensitive sites, H3K36 trimethylation, TSSs, RNA polymerase II, and TFs such as Specificity protein 1, ZNF274 and SIX homeobox 5. Computational modeling with these genomic elements suggests that they have distinct roles in TAD boundary formation. We propose a structural model of TAD boundaries based on these findings that provides a basis for studying the mechanism of chromatin structure formation and gene regulation.

INTRODUCTION

The human genome is tightly packed into the cell nucleus, and yet is readily accessible to other cellular components for the maintenance of cellular activity and response to external signals. The physical structure of the genome responsible for this apparent paradox has been sought for a long time. Biochemical and microscopic studies revealed that the genome has different levels of organization, including nucleosomes, chromatin domains, and chromosome territories (1–3). Recent advances in chromatin conformation capture techniques have enabled detailed examination of genomic structures; it was suggested that the human genome has a fractal-like architecture in which the small local structural units of chromatin interact with each other to form larger units (4). The basic structural units of chromatin are nucleosomes, which have a 10-nm fiber structure; their dynamic interaction leads to the formation of larger-scale chromatin domains characterized by high intra-domain and low inter-domain interactions (5). Two classes of chromatin domain have been described; the smaller of these, which are referred to as physical or contact domains, constitute about 185 000 bp (6) and interact with each other to form larger topologically associated domains (TADs) of ∼1 million bp (7).

Although the existence of various genomic structures has been experimentally demonstrated, how they are formed is largely unknown; TADs are of particular interest given that they are implicated in gene regulation (8,9). It is thought that particular genomic elements are located at the boundaries of TADs and are involved in their formation. CCCTC-binding factors (CTCFs) are more frequently observed at TAD boundaries than in other regions, as are transcription starting sites (TSSs) of housekeeping (HK) genes and histone modifications such as H3K36 trimethylation (H3K36me3) (10). In fact, CTCF-binding sites are key features of TAD boundaries since they can interact with CTCF-binding sites of another TAD boundary and thereby segregate the activities of adjacent TADs in association with cohesin complexes (6). However, given that CTCF-binding sites are frequently found outside the TAD boundary, these sites are not in themselves sufficient to demarcate TADs. Hence, in one of previous studies, histone modification marks were incorporated into the prediction of TAD boundaries along with CTCF-binding sites (11). One other previous study examining the association between 76 DNA-binding proteins and chromatin loop anchors located at the ends of contact domains found that the transcription factors (TFs) Zinc finger protein (ZNF)143 and Yin Yang (YY)1 were enriched at TADs (6); the enrichment of ZNF143 at TADs was later confirmed by another group (12). However, these studies did not systematically search for genomic elements associated with TAD boundaries nor did they evaluate their predictive power. The predictability measure could clarify the association between genomic elements and TAD boundaries, since it takes into account both coverage and data enrichment. For example, TSSs of HK genes are the most highly enriched genomic element

[*]To whom correspondence should be addressed. Tel: +82 42 350 4317; Fax: +82 42 350 4310; Email: kds@kaist.ac.kr

© The Author(s) 2017. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

at TAD boundaries, but have weak predictive power since they are present in only a small fraction of TAD boundaries. Conversely, CTCF binding sites are good predictors of the boundaries though they are not particularly enriched at the boundaries than other genomic elements.

In the present study, we investigated the association between TAD boundaries and various genomic elements including TF-binding sites (TFBSs), TSSs, histone modifications, and DNase I-hypersensitive sites. We first identified genomic elements that are enriched at TAD boundaries. We then employed machine-learning methodology to determine those that are important for accurately predicting TAD boundaries (Figure 1). We confirmed that consensus CTCF-binding sites present in multiple cell lines were better predictors of TAD boundaries, and were associated with TFs such as ZNF143. We also searched for genomic elements that represent the structural characteristic of TAD boundaries by measuring their enrichment at and their ability to predict these boundaries, and identified TAD boundary-associated genomic elements such as ZNF143, YY1, MYC complexes, Specificity protein (SP)1, ZNF274 and SIX homeobox (SIX)5 as well as DNase I-hypersensitive sites, H3K36me3 and TSSs. We generated a predictive model based on these genomic elements and analyzed their localization at specific sites in the TAD. Our findings reveal a number of novel TAD boundary-associated genomic elements whose functions can provide insight into the mechanisms of gene regulation.

MATERIALS AND METHODS

TADs and boundaries

A total of 2994 TADs with a length >200 kb were selected from 3062 previously defined TADs (10). The meeting ends of two adjacent TADs were designated as a boundary. When there was a gap between the two ends, these were merged into a boundary if the spacing was <100 kb. There were 1254 gapped cases, of which 718 were merged. Ultimately, 3586 boundaries were analyzed in this study. The genomic locations of TADs and boundaries are shown in Supplementary Tables S1 and S2.

Genomic regions before and after 150 kb of the boundaries were designated as TAD boundary segments (Figure 1A). A total of 3586 segments were collected and the same number of segments were randomly sampled from genomic regions that were not part of a boundary. TAD regions located >100 kb away from a boundary were defined as a non-boundary TAD region; genomic locations in this region were randomly selected and those within ±150 kb were designated as TAD segments (Figure 1A). These were evenly distributed throughout the genome, and both segments together covered 61% of the entire human genome (Figure 1B and Supplementary Figure S1).

Genomic elements

Genomic elements were collected from the University of California, Santa Cruz (UCSC) Genome Browser database (13); most of these were generated by the ENCODE project (14) and included TFBS clusters, which are a collection of transcription factor binding site peaks from multiple experiments in various cell lines (wgEncodeRegTfbsClusteredV3 table). TF signals obtained from the H1-human embryonic stem cell (hESC) cell line were separately considered in order to evaluate the cell type dependency of the TAD structure. Non-TFBS signals such as locations of DNase I-hypersensitive sites and TSSs were also retrieved from the UCSC Genome Browser (13) and were tagged as ‘Other’. The full list of data used in this study is shown in Supplementary Table S3. TSSs were divided into three groups: TSSs of mRNA (TSS), TSSs of mRNA and non-coding RNAs (TSS-ALL), and TSSs of HK genes (TSS-HK). HK genes were identified from a previously published list (15). The center of each signal was taken as its genomic location; for example, a CTCF peak located at 16 110–16 390 on chromosome 1 was assumed to be at 16 250.

Analysis of consensus CTCF-binding sites

We found that genomic elements associated with consensus CTCF-binding sites occupied a narrower region, and we therefore reduced the range of analysis to ±15 kb. Consensus CTCF-binding sites were defined as signals with a peak score ≥600 and supported by 60 or more experiments.

Position-specific linear model (PSLM)

To evaluate the predictive power of each genomic element, we devised a computational model, PSLM, that considers both position and density of genomic elements in genomic segments and classifies them into boundaries or TADs. In the model, the analyzed segments—of boundaries or TADs—were divided into 11 bins and the number of each genomic element in each bin was counted, which are used as 11D feature vectors that represent the positional preference of the genomic element around the boundary or the TAD. For each genomic element, boundary and TAD segments were converted into feature vectors, and a predictive model that segregated boundary and TAD feature vectors was generated. A linear model was developed using the Bayesian Ridge method (16), which resembles linear regression but eliminates abnormally large coefficients by incorporating parameter regularization. To counter the over-fitting problem, we used a 5-fold cross-validation approach. All values presented in this study are results of the test sets.

Population greedy search algorithm (PGSA)

To identify the combination of genomic elements that can best describe TAD boundaries, we devised a heuristic search algorithm, PGSA, which is a simple extension of the greedy search algorithm with genetic algorithm concept. In the greedy search algorithm, a path that maximizes the objective function is selected at each optimization step. Instead, our algorithm keeps a number of paths with highest objective function. In this study, we generated the first PSLM with the consensus CTCF signal, followed by the set of all possible combinations containing the CTCF signal and one additional genomic element. For each combination, a PSLM was generated and the area under the curve (AUC) was determined. Genomic element combinations of the 10 best models were collected, and the next

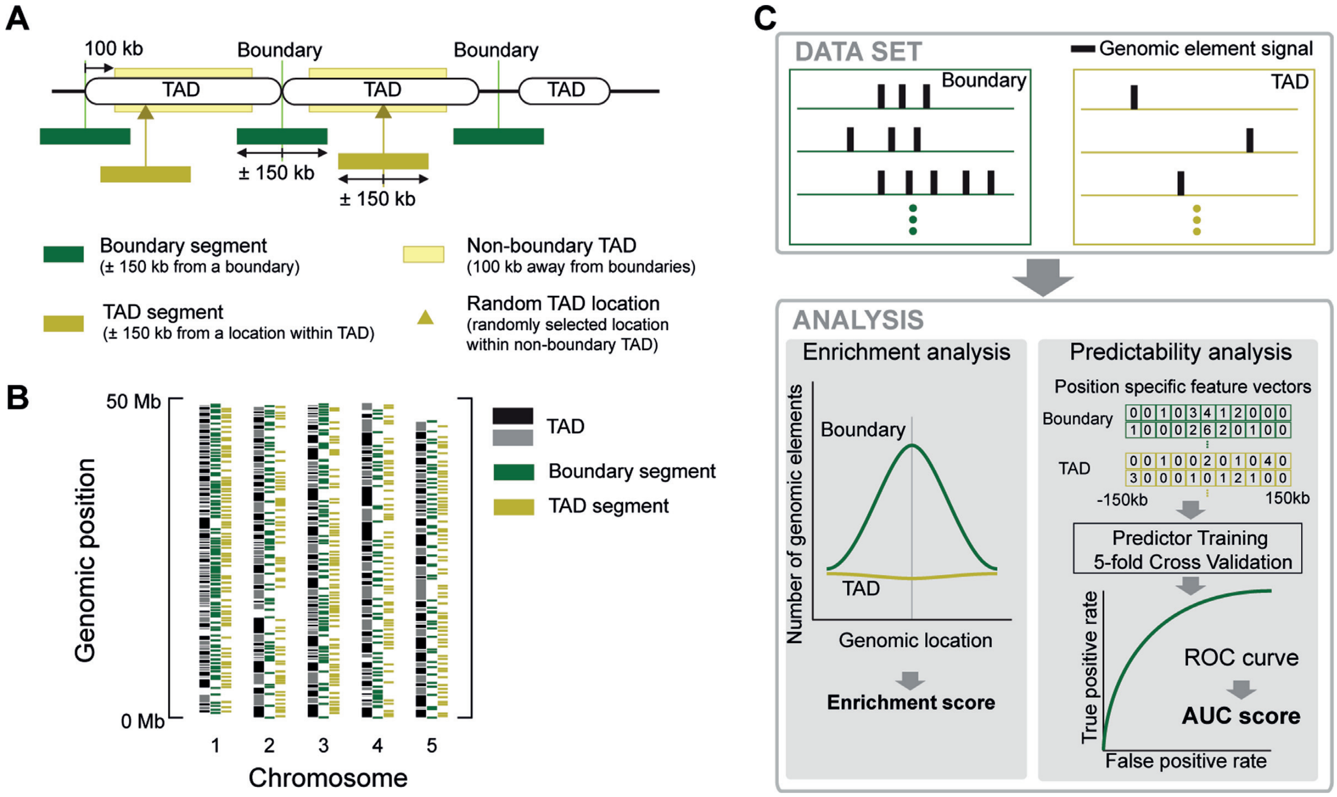

Figure 1. Dataset and schematic illustration of analytical approaches. (**A**) TADs were retrieved from a previously published study (10), and TAD boundaries were defined as the ends of TADs. The ±150-kb region of TAD boundaries were designated as boundary segments, and the same number of genomic regions located in the TAD—designated as TAD segments—were identified by randomly selecting locations within TADs. (**B**) Genomic distribution of boundary and TAD segments (the distribution in the whole region of analysis is shown in Supplementary Figure S1). (**C**) Signals corresponding to each genomic element were mapped onto the boundary and TAD segments. The collective distributions of genomic elements in these segments were compared in order to identify TAD boundary-enriched genomic elements. In another approach, signals in each segment were used to generate a predictive model that distinguishes boundary from TAD segments. The model was used to predict the relative importance and predictive power of genomic elements for TAD boundary formation.

set of combinations was generated by adding one more genomic element to each selected combination. This procedure was repeated until combinations of 10 genomic elements were evaluated. This analysis yielded pseudo-optimal combinations of genomic elements that are highly predictive of TAD boundaries. The analysis also revealed genomic elements that were consistently selected in the best predictive models and would be robust features of the TAD boundary. Genomic elements were incorporated by concatenating 11D feature vectors; thus, combining two genomic elements yielded a 22D feature vector. Our dataset contained some redundant information—for example, both TFBS cluster and H1-hESC-specific data were available for RAD21. Since highly correlated features can cause problems when constructing predictive models, data representing the same genomic element were grouped into a class, and only combinations of different classes were allowed during PGSA optimization steps. The list of genomic elements and their classes is shown in Supplementary Table S1.

With this algorithm, combinations of genomic elements were analyzed to identify those that were consistently used in multiple models. This was measured as the number of top predictive models containing a genomic element, or as the 'selection count'. However, this value was large for genomic elements that were incorporated in the earlier optimization step and the genomic elements that were introduced in a later step but were used repeatedly would be under-weighted. We therefore devised another measure referred to as 'persistence score', which was obtained by dividing the selection count by the number of predictive models in subsequent rounds of optimization.

Boundary enrichment score

For quantitative comparisons of boundary enrichment, a boundary enrichment score was calculated. The value is a weighted integration of the relative preference for boundary over non-boundary segments and is determined by the following equation:

$$S_e = \sum_{i=1}^{m} p(i, BD) \ln \left[\frac{p(i, BD)}{p(i, \text{TAD})} \right]$$

where S_e is the enrichment score; p is the relative frequency of a genomic element at position i observed in boundary (BD) and TAD segments; and m is the number of bins spanning the entire region of analysis (300 kb) ($m = 550$ in this study).

RESULTS

CTCF-binding sites and TSSs in TAD boundaries

CTCF-binding sites and TSSs have been reported to be enriched at TAD boundaries, which was confirmed in our dataset (Supplementary Figure S2). A large number of both elements were found within 50 kb of TAD boundaries; however, CTCF count distributions of boundary and TAD segments were not clearly distinguishable (Supplementary Figure S2A). Similarly, TSS counts could not effectively distinguish between boundary and TAD segments, and ∼800 TAD boundaries lacked TSSs within a 100-kb range (Supplementary Figure S2B). Therefore, CTCF-binding sites and TSSs are not sufficient in themselves to demarcate TAD boundaries, which are likely defined by other genomic elements. We therefore examined the association of various genomic elements with TAD boundaries.

Boundary-enriched genomic elements

Genomic elements enriched or depleted in the vicinity of TAD boundaries would be involved in their formation. Biochemical analysis techniques including next-generation sequencing have identified the locations of various genomic elements. These data were obtained from UCSC Genome Browser database, which contains a large number of data generated by the ENCODE project (13,14). Our compendium mostly included DNA-binding proteins and histone modifications along with other elements such as TSSs and DNase I-hypersensitive sites (Supplementary Table S1). The location of each genomic element was mapped onto boundary and TAD segments (Figure 1A), and the boundary enrichment score was calculated.

CTCF and TSS-HK were reportedly enriched at TAD boundaries (10); this was confirmed by their abundance at the boundary region and slight depletion in TADs (Figure 2A and B). There were approximately 1.7 and 4 times more CTCF-binding sites and TSS-HKs, respectively, near TAD boundaries. For a quantitative comparison, we calculated an enrichment score for each genomic element and found that TSS-HK was the most highly enriched element at the boundary, followed by RNA polymerase II and TFs such as SP4, Sin3A-associated protein (SAP)30, General transcription factor IIF subunit (GTF2F)1, and MYC (Figure 2C).

SP4 is an SP family TF that recognizes GC boxes in many promoters (17), while GTF2F1 is a general TF that binds to the TATA box (18). MYC controls the expression of a wide variety of genes and is an important regulator of cell proliferation and death (19). The enrichment of these factors may be related to the transcriptional activity of TAD boundaries and TSS enrichment. In contrast to these general activating TFs, SAP30 is a component of histone deacetylase complex that is involved in gene repression (20). Therefore, TAD boundary formation may not be solely associated with gene activation, but may involve a more complex regulation that includes changes in chromatin structure. Most DNA-binding proteins were enriched at the boundary (Supplementary Figure S3), possibly due to TSS enrichment. However, some genomic elements were more highly represented at the boundaries than others, suggesting that they have important roles in TAD boundary formation.

The various TAD boundary-associated genomic elements identified by our analysis were not sufficient to characterize or predict TAD boundaries. For example, although some boundaries contained an abundance of TSSs, others lacked them altogether (Supplementary Figure S2B). We therefore used machine-learning techniques to identify genomic elements related to TAD boundaries. Specifically, the positional densities of genomic elements were converted into feature vectors, and linear predictive models that segregated boundary and TAD segments were generated. This type of PSLM was constructed for each genomic element and its predictive power was evaluated based on the average AUC of receiver operating characteristic curves in the 5-fold cross-validation approach.

Boundary-predictive genomic elements

Genomic elements with the highest predictive power differed markedly from boundary-enriched elements (Figures 2C and 3D and Supplementary Figure S3). The model with CTCF had the highest predictive power for the boundary, with an AUC of 0.751 (Figure 3A, D). This model predicted 82% of TAD boundaries with an erroneous prediction rate of 46% of TADs as boundaries when the prediction threshold was set to the maximum F1 score (Figure 3B). The positional preference of a genomic element was inferred from the coefficients of the model. In the case of CTCF, the coefficient was largest at the center, indicating that a higher number of CTCFs were located at the center of the TAD boundary while fewer were present in regions remote from the center (Figure 3C).

In agreement with previous reports (7,10), models for CTCF and cohesin complex proteins including RAD21 and Structural maintenance of chromosomes protein 3 (SMC3) showed the highest prediction accuracy (Figure 3D). Interestingly, CTCF-binding site information from the H1-hESC cell line was more useful for boundary prediction than CTCF information in the TFBS clusters data, which is a collection of CTCF peaks from multiple cell lines. The latter may contain cell type-specific sites, which would lower the prediction accuracy. Since consensus CTCF-binding sites were found to be more closely associated with TAD boundaries (21), CTCFs were classified as consensus or non-consensus and genomic elements associated with the consensus ones were then characterized in the next section.

The TF ZNF143 was found to be a powerful predictor of TAD boundaries, even competing with CTCFs. Although the function of ZNF143 is not well characterized, it was reported to be associated with the regulation of U6 small nuclear RNA transcription and recruitment of chromatin remodelers (22). It was also implicated in distal chromatin interactions (6,12). Thus, ZNF143 may interact with other components of TAD boundaries by acting as a scaffold for other proteins or mediating interactions between TAD boundaries. YY1 participates in transcriptional repression of microRNAs by recruiting Enhancer of zeste homolog 2 and inducing H3K27 trimethylation (23). YY1-binding DNA motifs were detected around CTCF sites in *Tsix*, a non-coding RNA that is involved in X chromosome inactivation, and physical association between YY1 and CTCF has been confirmed (24). Although the structural

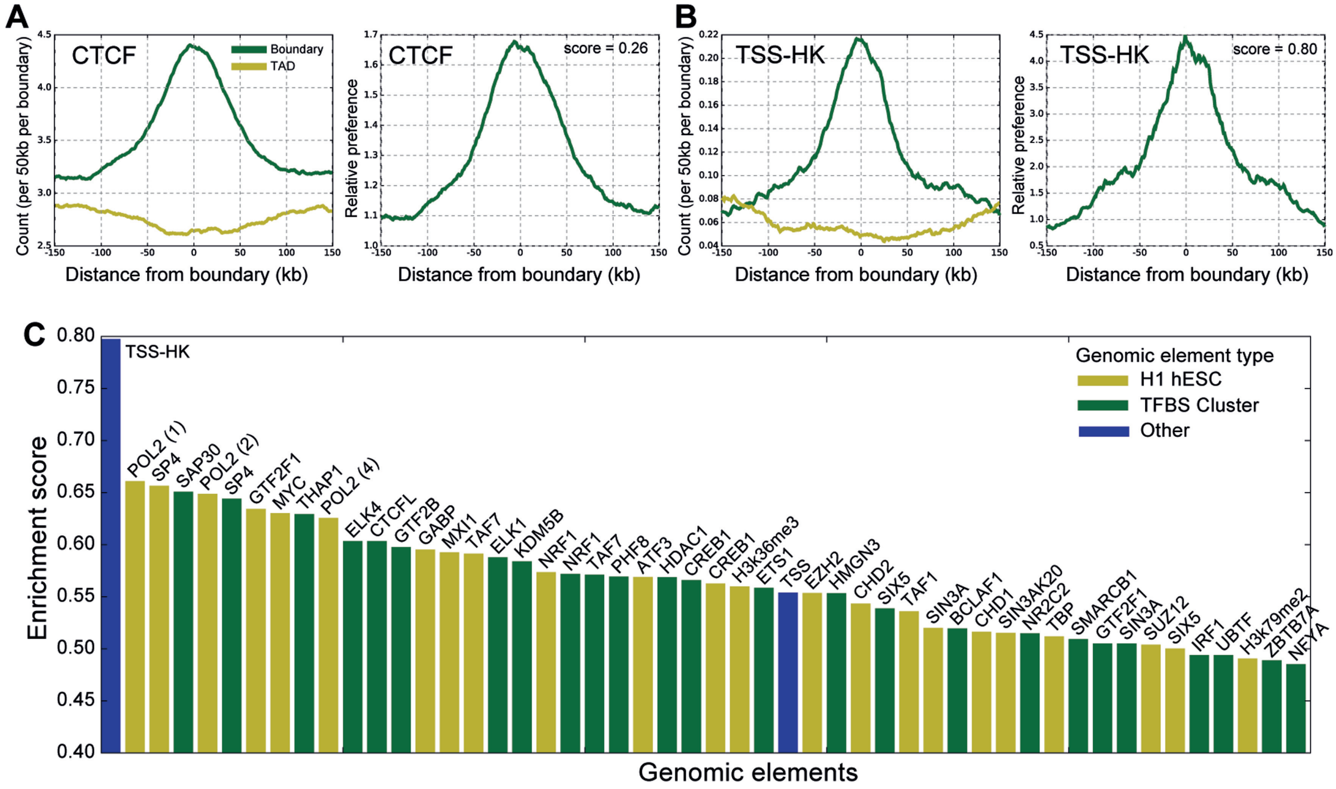

Figure 2. TAD boundary enrichment. Positional distribution of each genomic element and its enrichment at TAD boundaries. (**A**) Enrichment of CTCF-binding sites at TAD boundaries. (**B**) Enrichment of TSS-HK at TAD boundaries. (**C**) Genomic elements with the 50 highest enrichment scores (the full list is shown in Supplementary Table S2).

implications of this interaction are unclear, they may contribute to the shaping of TAD boundaries. Other TFs such as Lysine demethylase 4A (KDM4A), and PHD finger protein 8 (PHF8) were highly predictive of TAD boundaries. KDM4A is involved in the demethylation of trimethylated H3K9 and H3K36, which are associated with heterochromatin regions and transcriptional elongation, respectively (25,26). PHF8 is also a lysine demethylase that catalyzes the demethylation of mono- and di-methylated H3K9 (27). These TFs are mainly related to the histone modification and chromatin conformation. However, our finding that their binding sites were better predictors of TAD boundaries than the corresponding modifications suggests that they have additional functions.

Consensus CTCF-binding sites and associated genomic elements

Given that TAD boundaries are conserved among cell lines and that the consensus CTCF signal is a better predictor of TAD boundaries (8,21), we evaluated the predictive power of a subset of CTCF signals that were supported by multiple experiments. These CTCF signals were more abundant at TAD boundaries (Figure 4A). The predictability was improved with the number of supporting experiments but reached a maximum performance when signals from 60 or more experiment were used (Figure 4B). We also analyzed another subset of CTCF signals with different peak score thresholds that considers the statistical significance and reproducibility of the signal (28). Both enrichment and AUC scores increased as a higher peak score threshold was applied (Supplementary Figure S4A and B). An AUC score of 0.793 was achieved when the predictive model was generated with CTCF signals that were supported by 60 or more experiments and had a peak score larger than 600; these were designated as consensus CTCF signals (Supplementary Figure S4D). Since these signals significantly improved TAD boundary prediction, we hypothesized that they would have characteristics that were distinct from those of non-consensus CTCF binding sites. CTCF is a ubiquitous TF for which the binding site can vary according to cell type; as such, it is thought to bind to specific genomic regions in association with cofactors (29). Thus, specific TFs may be associated with consensus CTCFs. To identify such cofactors, we classified CTCF signals as consensus and non-consensus CTCFs, and searched for genomic elements that were preferentially associated with the former.

ZNF143 was the most abundant genomic element at consensus CTCFs and a good predictor of them (Figure 4C and D). It was also a good predictor of TAD boundaries (Figure 3D). Its enrichment at these boundaries (12) and at chromatin loop anchors (6) contributes to chromatin interactions through associations with CTCF and cohesin (30). SMC3 and RAD21 were also closely associated with consensus CTCFs in terms of both the enrichment score and predictive power (Figure 4C and D). Thus, the association between CTCFs and the cohesin complex, which may in-

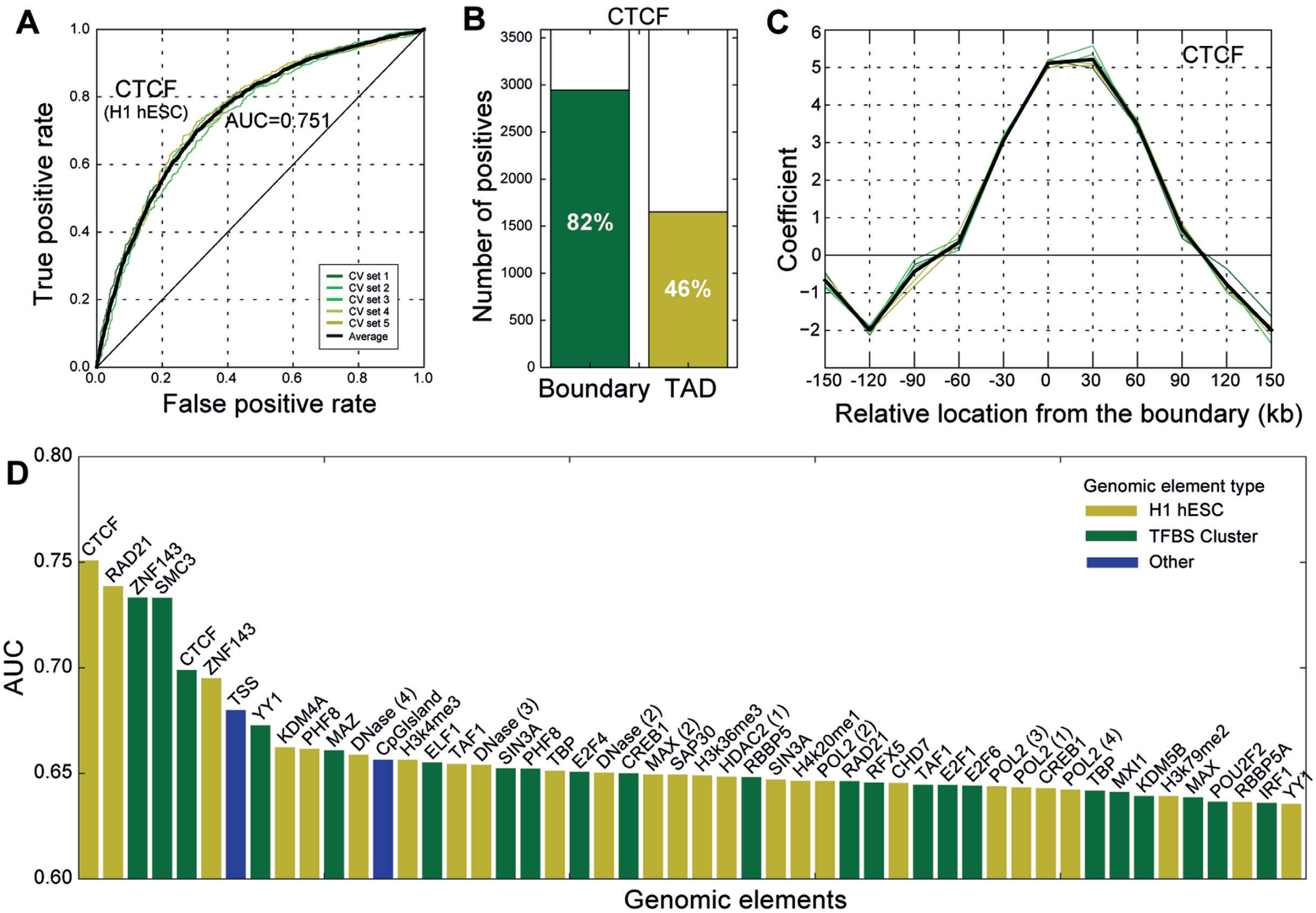

Figure 3. TAD boundary prediction. Predictive power of position-specific linear models. (**A**) ROC curves of CTCF-based models. A 5-fold cross-validation approach was used; curves represent models constructed in each cross-validation procedure. (**B**) True and false positive ratios of CTCF-based models. Levels were measured at the threshold with the highest F1 score. (**C**) Coefficients of the CTCF model. (**D**) Top 50 genomic elements with the highest predictive power (the full list is shown in Supplementary Table S4).

volve ZNF143, is an important factor that discriminates consensus from non-consensus CTCFs. YY1 was likewise a powerful predictor of consensus CTCFs. YY1 can physically associate with CTCF (24) and is enriched at chromatin loop anchors (6). Thus, YY1 may also act as a CTCF cofactor that facilitates distal chromatin interactions. In addition, TFs including GA-binding proteins and MYC were enriched at consensus CTCF-binding sites. The association between MYC and RAD21 has been reported in fission yeast (31); these TFs may contribute to recruitment of cohesin complexes and TAD boundary formation.

TAD boundary prediction with multiple elements

The consensus CTCF was the single most informative genomic element in predicting TAD boundaries, but ~40% of TAD segments were indistinguishable from boundary segments (Supplementary Figure S5B). We therefore hypothesized that other genomic elements also contribute to the formation of TAD boundaries, and we searched for a combination of genomic elements that could improve TAD boundary prediction along with the consensus CTCF. To this end, we devised the PGSA, which increases the predictive power of the model by sequentially incorporating an additional genomic element to the previous model. Up to 10 combinations of genomic elements were evaluated, but the incorporation of three elements—DNase-Cluster, H3K36me3, and TSSs—of both mRNA and non-coding RNA (TSS-ALL) showed the greatest increase in predictive performance (Figure 5A). These elements were consistently detected in later optimization steps (Figure 5B and C), highlighting their importance in TAD boundary prediction. The resultant predictive model was able to predict TAD boundaries with an AUC of 0.826, and 86% of boundaries were successfully predicted with an erroneous prediction rate of 37% of TAD regions as boundaries at the precision of 0.70 (Supplementary Figure S5E).

We examined the coefficients of the PSLM, which was built with these four genomic elements, to determine the locational preference of genomic elements around TAD boundaries. The coefficients were similar to each other irrespective of the choice of training set, suggesting that the predictive models were robust and that the coefficients themselves were informative (Figure 6A). The coefficients indicated that consensus CTCFs were preferentially localized at the center of the TAD boundary, which spans ~100 kb. This is consistent with the enrichment of CTCFs at the bound-

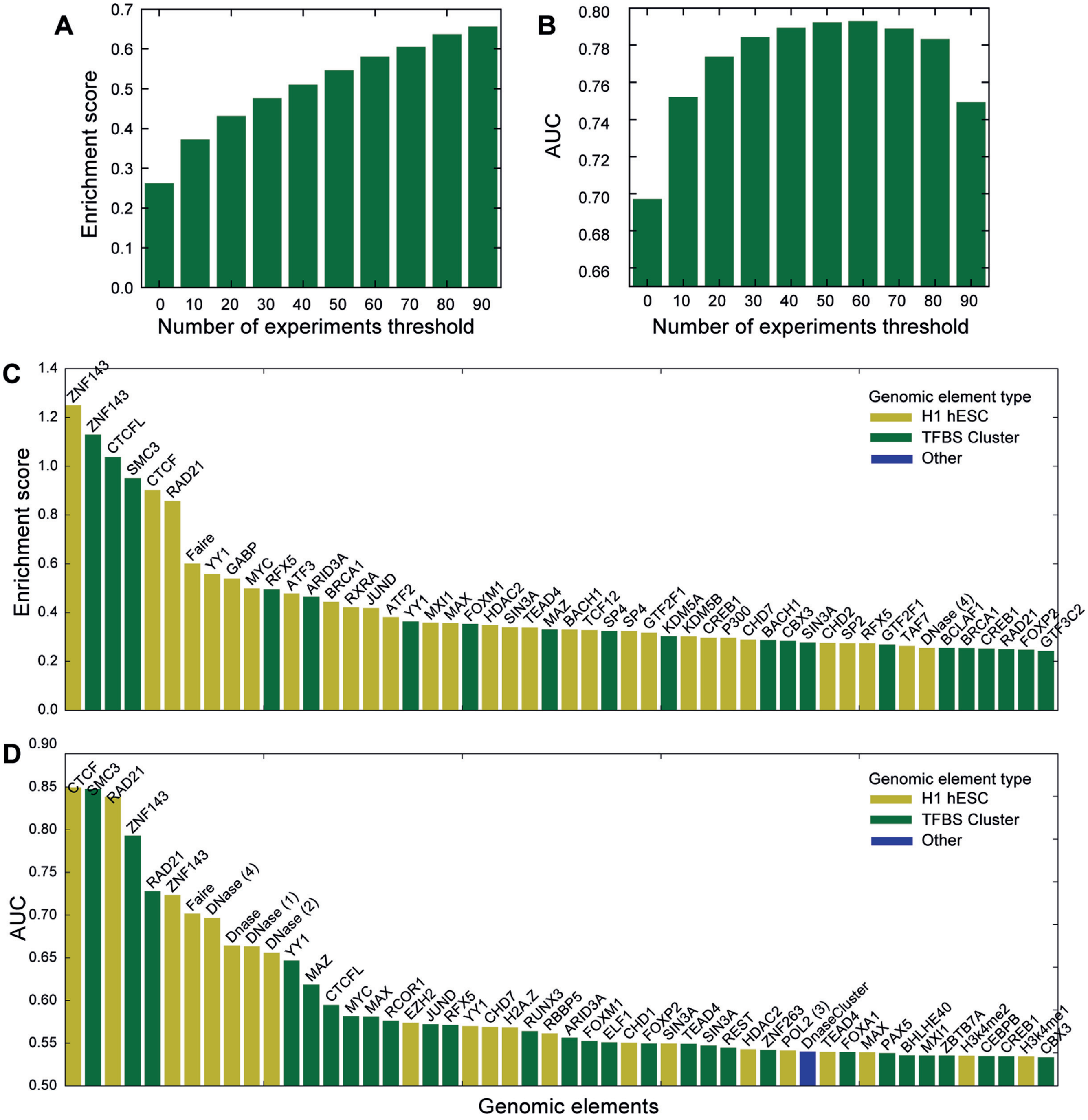

Figure 4. Consensus CTCF-associated genomic elements. (**A**) Dependency of enrichment score on the number of supporting experiments. (**B**) Dependency of predictive power on the number of supporting experiments. (**C**) Genomic elements enriched at consensus CTCFs. (**D**) Genomic elements predictive of consensus CTCFs (the full list is shown in Supplementary Table S5).

ary and also with the fact that multiple CTCF-binding sites are located near TAD boundaries (Supplementary Figure S2A).

H3K36me3 was also clustered at the center of the TAD boundary (Figure 6A). This modification is enriched at contact domains related to compartment A, which is thought to be related to gene activation (6). H3K36me3 is associated with alternative splicing and reduced chromatin accessibility (32), and is recognized by PWWP domain proteins including DNA methyltransferase 3A (DNMT3A) (33). These findings suggest that H3K36me3 can recruit cellular machinery for altering chromatin structure.

TSSs of both mRNA and non-coding RNA (TSS-ALL) were also abundant at the center of the TAD boundary (Figure 6A). The selection of TSS-ALL over the TSS of mRNAs indicates that the binding of large molecular machinery can contribute to the formation of TAD boundaries. Alterna-

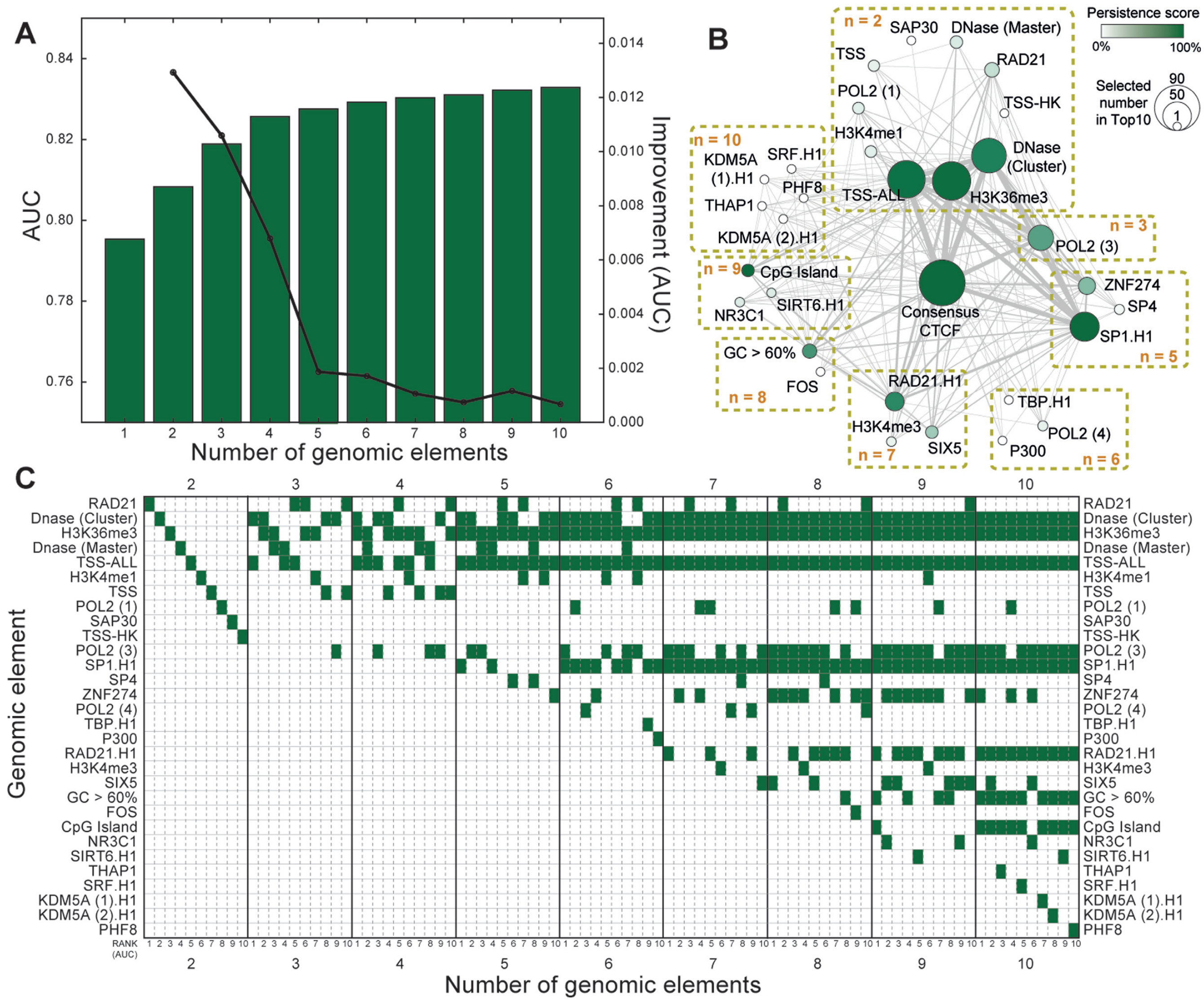

Figure 5. Combination of genomic elements. (**A**) TAD boundary prediction by the best multi-element models identified by PGSA. (**B**) Graphical representation of genomic elements identified by the PGSA approach. Each genomic element is represented as a node whose size and color are the 'selection count' and 'persistence score', respectively. Edges were drawn between two genomic elements used by top predictors and the width is the number of models supporting the relationship. (**C**) Genomic elements used in the top 10 models shown in a matrix-like format. Large columns enclosed by solid lines represent each optimization round, and contain 10 small columns corresponding to the top 10 models. Genomic elements used in a model or column are indicated by a green color. For example, the top-ranked model of the second optimization round, or that with three genomic elements, is composed of DNase and TSS-ALL.

tively, changes in nucleosome position by polymerase binding (34) can influence the structure of these boundaries.

The weights of DNase I-hypersensitive sites were mostly negative around the TAD boundary (Figure 6A), which appears to be contradictory to the fact that the binding of proteins to DNA generally increases chromatin accessibility (5). In fact, the DNase I sites were enriched at the boundaries (Supplementary Table S4). However, DNase I signals are correlated with the CTCF signals, and when comparing a boundary and a TAD region with the same level of CTCF signals the DNase I signals are weaker in the boundary region than in the TAD region (Supplementary Figure S6). Consequently, the coefficient of DNase I is negative in the boundary prediction model. It is unclear why the boundaries are associated with slightly reduced DNase I signal. The chromatin conformation of the boundaries might be different from those of well exposed TAD regions. Recently, it was suggested that the TAD boundary region might be less flexible than other regions and this physical property may contribute to the separation of two adjacent TADs (7). The association of various cellular machineries such as CTCF, RNA polymerases, and DNMT3A could cooperatively induce structural changes in chromatin at TAD boundaries and make the region less flexible. The resultant conformation may reduce interaction between two adjacent TADs, and thus act as a barrier for inter-domain interaction.

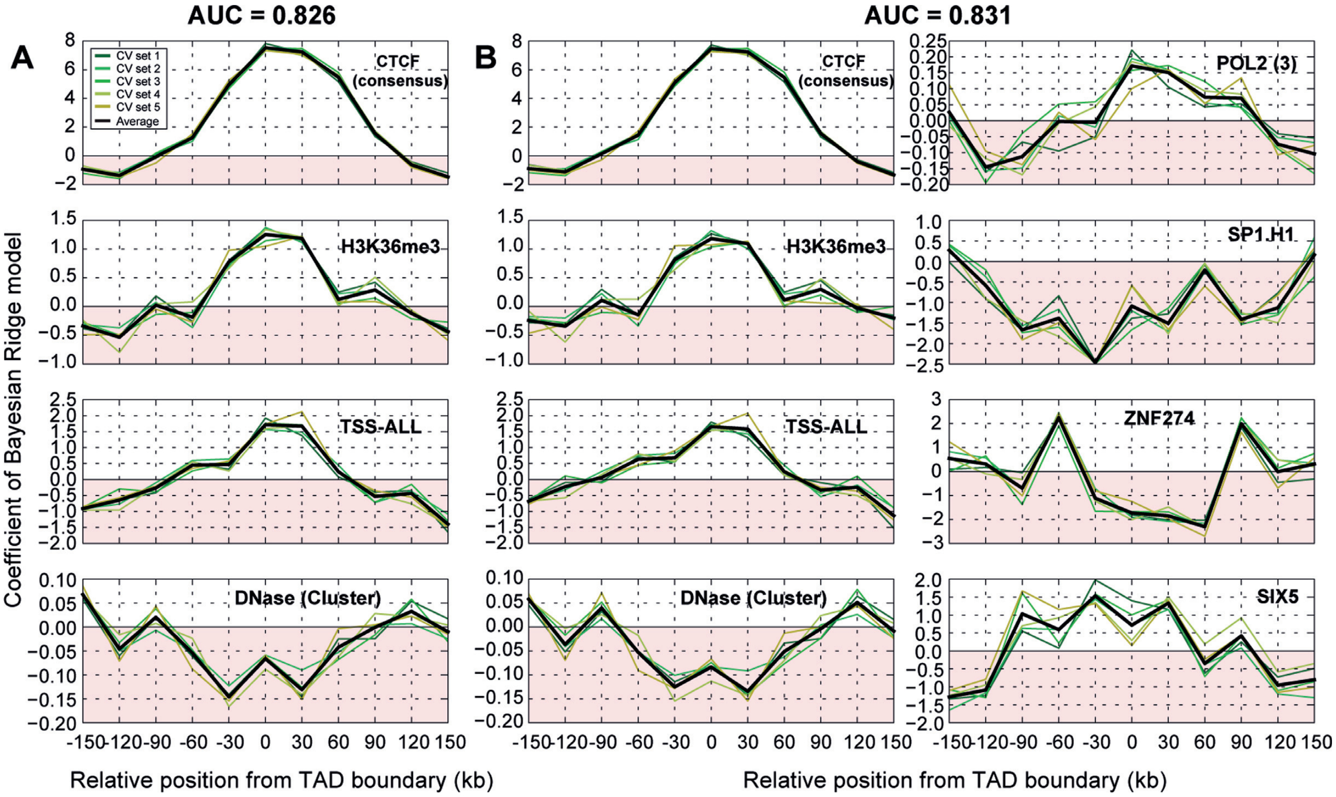

Figure 6. Inference of positional preference. Coefficients of each genomic element in the PSLM. (**A**) Model constructed with four genomic elements (consensus CTCF, H3K36me3, TSS-ALL and DNase-Cluster). (**B**) Model with four additional genomic elements (RNA polymerase II, and TFs SP1, ZNF274 and SIX5).

The incorporation of more genomic elements further improved the predictive power of genomic elements, albeit only modestly (Figure 5A). We analyzed the prediction models with four additional genomic elements—namely, RNA polymerase II, SP1, ZNF274, and SIX5—which were consistently observed in subsequent optimization steps (Figure 5B and C). The coefficients of CTCF, H3K36me3, TSS, and DNase of this model (Figure 6B) were similar to those in the previous model (Figure 6A), suggesting that these genomic elements are robust features of the TAD boundary. RNA polymerase II binding sites of H1-hESC were preferentially associated with TAD boundaries, suggesting that genes that are actively transcribed in hESCs are more closely associated with TAD boundaries than others. SP1 had negative weight in the boundary region. The activity of SP1 is regulated by various post-translational modifications and has been linked to various cellular processes including differentiation, cell growth, and death (35,36). Although the underlying mechanism requires more detailed study, conditional binding is not a feature that favors the constitutive genomic structure of TADs. As such, it is likely that the TFs under cellular regulation are disfavored at TAD boundaries. An intriguing pattern emerged for ZNF274 in that two preferential coefficients flanked the negative coefficient region (Figure 6B). The function of ZNF274 is not fully understood, but it is thought to repress transcription by recruiting the histone–lysine N-methyltransferase SET domain bifurcated 1 and forming heterochromatin (37). Therefore, it could facilitate the formation of and function as a boundary for TADs. To date, there are no known modifications for ZNF274 in UniProt (38), suggesting that it is a constant feature of the genomic structure. SIX5 was positively associated with TAD boundaries; its mutation has been linked to branchio-oto-renal syndrome (39,40), but its role in chromatin structure is mostly unknown. Similar to ZNF274, there have been no regulatory modifications reported for SIX5 (37); it may therefore stably bind to specific DNA sequences, which could contribute to the formation of TAD boundaries.

DISCUSSION

The present study systematically evaluated the enrichment of genomic elements at TAD boundaries and analyzed their ability to predict the location of these boundaries. A model of the TAD boundary is shown in Figure 7. CTCF and cohesin complexes were confirmed as the most important features of TAD boundaries. In particular, consensus CTCF-binding sites were closely associated with these boundaries and are presumed to be involved in distal genomic interactions in conjunction with cohesin complexes, whose binding would be facilitated by cofactors such as ZNF143 and YY1. TAD boundaries are better described as a combination of multiple genomic elements, and a predictive model incorporating these can provide insight into the structural characteristics of boundaries. Genomic elements associated with boundaries were related to larger molecular machiner-

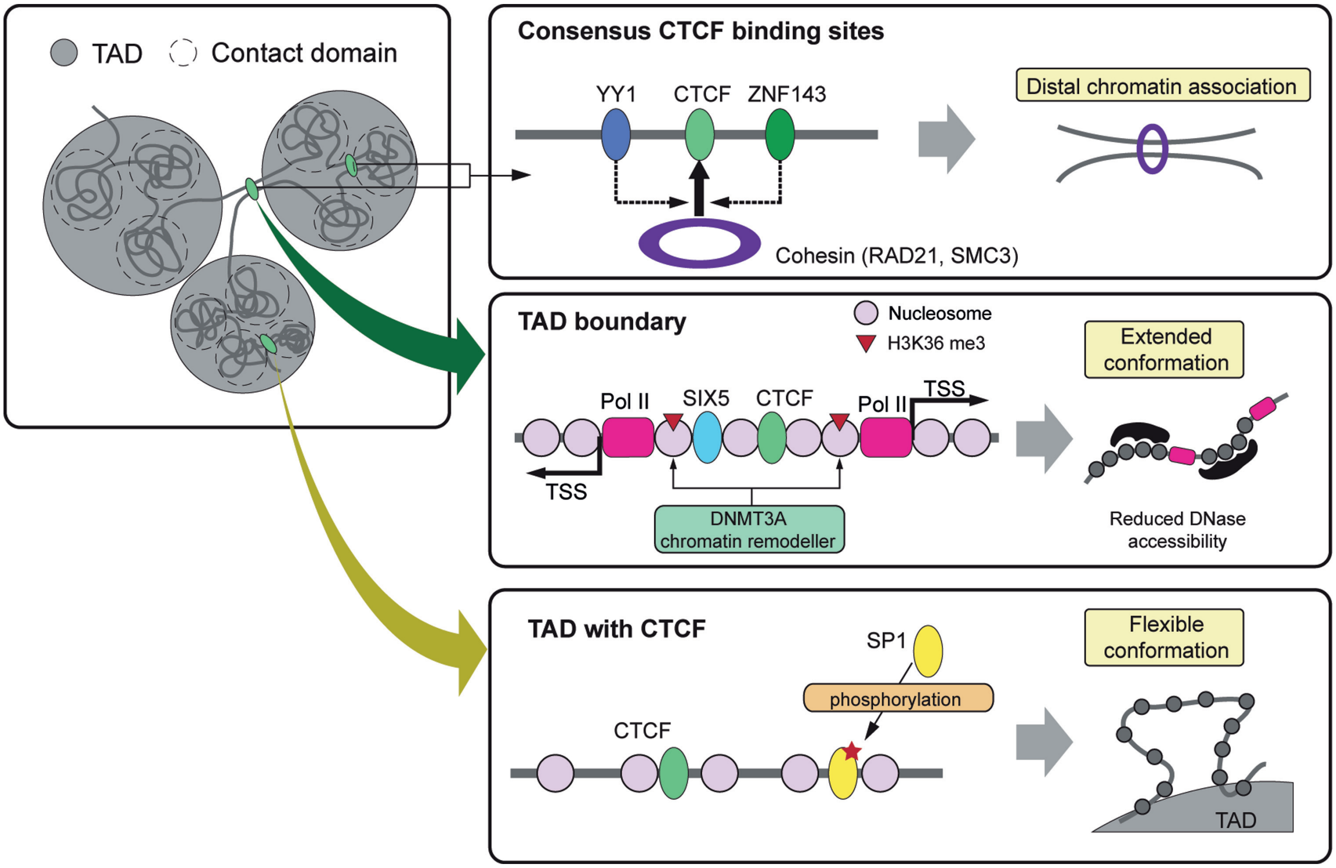

Figure 7. Model of TAD boundary. CTCFs are highly enriched at boundaries of both contact domains and topologically associated domains. In particular, CTCF-binding sites found in multiple cell lines were closely associated with boundaries. Co-factors such as ZNF143 and YY1 may be involved in the association of the cohesin complex to these sites, whereby two distal regions would be in contact. Unlike contact domain boundaries, multiple CTCF-binding sites were observed at TAD boundaries. Other genomic elements including TSS, H3K36me3 and SIX5 were associated with these boundaries, and the chromatin would assume a more closed conformation. The association of TFs such as SIX5 at the boundaries and SP1 in TADs may have different structural effects on chromatin structure. Consequently, the binding of a distinct set of genomic elements and resultant structural changes in the chromatin would lead to the separation of the boundary and TAD regions; the former can then associate with each other, while two adjacent TADs can be more definitively separated from each other.

ies and were distributed over a large genomic region spanning approximately 100 kb. Boundary regions were less accessible to DNase than TAD regions with similar CTCF signals, suggesting that the two regions have distinct chromatin conformations, which may be attributed to H3K36me3. Alternatively, the binding of large molecular machinery may cause slight chromatin condensation at TAD boundaries. The physical structure of these regions is not well understood, but it is thought that boundaries have extended conformation, at least in large scale. Since there are multiple CTCF sites at the boundary, two adjacent boundaries can anneal together instead of forming internal loops. On the other hand, the binding of multiple larger machineries may obstruct dense packing in the region or make the region separated from the condensed region (41), resulting in its protrusion from the domain. The binding of a specific set of proteins can also alter chromatin organization, making the region less flexible and causing it to adopt an extended conformation. In some cases, TFs such as ZNF274 may function as a nucleation site for TAD structures and thereby limit the scope of interactions of adjacent chromatin.

The PSLM model assumes a single boundary structure, but TADs are not homogeneous and their boundaries would be heterogeneous. We evaluated the effect of the heterogeneity by changing the boundary orientation to make a certain genomic element placed more abundantly on the upstream of the boundary. This simple treatment increased the prediction performance of the models (Supplementary Table S6). We think that further details of the TAD boundaries can be revealed by developing computational model that can consider the heterogeneity of TADs. Despite the fact that various genomic elements were observed in the orientation models, the boundary associated genomic elements reported here were recurrently found in the best predictive models, and the models built with these genomic elements were similar to the original model (Supplementary Figures S7 and S8), indicating that the boundary associated elements successfully represent a majority of TAD boundaries.

In conclusion, we revealed a set of genomic elements that are statistically associated with TAD boundaries. Some of these have been previously reported and others remain to be confirmed. Nonetheless, the findings reported here can serve as a basis for studies on the mechanisms of TAD formation and how genomic structure contributes to gene regulation.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Ministry of Health & Welfare, Republic of Korea [HI12C00014]; Ministry of Science and ICT, Republic of Korea [NRF-2012M3A9C4048758]. Funding for open access charge: Ministry of Health & Welfare, Republic of Korea [HI12C00014]; Ministry of Science and ICT, Republic of Korea [NRF-2012M3A9C4048758].

Conflict of interest statement. None declared.

REFERENCES

1. Luger,K., Mäder,A.W., Richmond,R.K., Sargent,D.F. and Richmond,T.J. (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature*, **389**, 251–260.
2. Cremer,T. and Cremer,C. (2001) Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat. Rev. Genet.*, **2**, 292–301.
3. Cremer,T. and Cremer,M. (2010) Chromosome territories. *Cold Spring Harb. Perspect. Biol.*, **2**, a003889.
4. 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.
5. Maeshima,K., Ide,S., Hibino,K. and Sasai,M. (2016) Liquid-like behavior of chromatin. *Curr. Opin. Genet. Dev.*, **37**, 36–45.
6. Rao,S.S.P.S.P., Huntley,M.H.H., Durand,N.C.C., Stamenova,E.K.K., Bochkov,I.D.D., Robinson,J.T.T., Sanborn,A.L.L., Machol,I., Omer,A.D.D., Lander,E.S.S. *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.
7. Dixon,J.R., Gorkin,D.U. and Ren,B. (2016) Chromatin Domains: The Unit of Chromosome Organization. *Mol. Cell*, **62**, 668–680.
8. Dixon,J.R., Jung,I., Selvaraj,S., Shen,Y., Antosiewicz-Bourget,J.E., Lee,A.Y., Ye,Z., Kim,A., Rajagopal,N., Xie,W. *et al.* (2015) Chromatin architecture reorganization during stem cell differentiation. *Nature*, **518**, 331–336.
9. Taberlay,P.C., Achinger-Kawecka,J., Lun,A.T.L., Buske,F.A., Sabir,K., Gould,C.M., Zotenko,E., Bert,S.A., Giles,K.A., Bauer,D.C. *et al.* (2016) Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. *Genome Res.*, **26**, 719–731.
10. 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.
11. Huang,J., Marco,E., Pinello,L. and Yuan,G.-C. (2015) Predicting chromatin organization using histone marks. *Genome Biol.*, **16**, 162.
12. Ye,B.-Y., Shen,W.-L., Wang,D., Li,P., Zhang,Z., Shi,M.-L., Zhang,Y., Zhang,F.-X. and Zhao,Z.-H. (2016) ZNF143 is involved in CTCF-mediated chromatin interactions by cooperation with cohesin and other partners. *Mol. Biol.*, **50**, 431–437.
13. Rosenbloom,K.R., Armstrong,J., Barber,G.P., Casper,J., Clawson,H., Diekhans,M., Dreszer,T.R., Fujita,P.A., Guruvadoo,L., Haeussler,M. *et al.* (2015) The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.*, **43**, D670–D681.
14. Rosenbloom,K.R., Sloan,C.A., Malladi,V.S., Dreszer,T.R., Learned,K., Kirkup,V.M., Wong,M.C., Maddren,M., Fang,R., Heitner,S.G. *et al.* (2013) ENCODE Data in the UCSC Genome Browser: Year 5 update. *Nucleic Acids Res.*, **41**, 56–63.
15. Eisenberg,E. and Levanon,E.Y. (2013) Human housekeeping genes, revisited. *Trends Genet.*, **29**, 569–574.
16. Pedregosa,F., Varoquax,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V. *et al.* (2011) Scikit-learn: machine learning in python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
17. Suske,G. (1999) The Sp-family of transcription factors. *Gene*, **238**, 291–300.
18. Hisatake,K., Ohta,T., Takada,R., Guermah,M., Horikoshi,M., Nakatani,Y. and Roeder,R.G. (1995) Evolutionary conservation of human TATA-binding-polypeptide-associated factors TAFII31 and TAFII80 and interactions of TAFII80 with other TAFs and with general transcription factors. *Proc. Natl. Acad. Sci. U.S.A.*, **92**, 8195–8199.
19. Pelengaris,S., Khan,M. and Evan,G. (2002) c-MYC: more than just a matter of life and death. *Nat. Rev. Cancer*, **2**, 764–776.
20. Zhang,Y., Sun,Z.-W., Iratni,R., Erdjument-Bromage,H., Tempst,P., Hampsey,M. and Reinberg,D. (1998) SAP30, a novel protein conserved between human and yeast, is a component of a histone deacetylase complex. *Mol. Cell*, **1**, 1021–1031.
21. Mourad,R. and Cuvier,O. (2015) Predicting the spatial organization of chromosomes using epigenetic data. *Genome Biol.*, **16**, 182.
22. Yuan,C.-C., Zhao,X., Florens,L., Swanson,S.K., Washburn,M.P. and Hernandez,N. (2007) CHD8 associates with human Staf and contributes to efficient U6 RNA polymerase III transcription. *Mol. Cell. Biol.*, **27**, 8729–8738.
23. Tsang,D.P.F., Wu,W.K.K., Kang,W., Lee,Y.Y., Wu,F., Yu,Z., Xiong,L., Chan,A.W., Tong,J.H., Yang,W. *et al.* (2016) Yin Yang 1-mediated epigenetic silencing of tumour-suppressive microRNAs activates nuclear factor-κB in hepatocellular carcinoma. *J. Pathol.*, **238**, 651–664.
24. Donohoe,M.E., Zhang,L.F., Xu,N., Shi,Y. and Lee,J.T. (2007) Identification of a Ctcf Cofactor, Yy1, for the X Chromosome Binary Switch. *Mol. Cell*, **25**, 43–56.
25. Ng,S.S.S., Kavanagh,K.K.L.K., McDonough,M.M.A., Butler,D., Pilka,E.S., Lienard,B.M.R., Bray,J.E., Savitsky,P., Gileadi,O., von Delft,F. *et al.* (2007) Crystal structures of histone demethylase JMJD2A reveal basis for substrate specificity. *Nature*, **448**, 87–91.
26. Guerra-Calderas,L., Gonz Alez-Barrios,R., Herrera,L.A., Cant,U, De Le On,D. and Soto-Reyes,E. (2015) The role of the histone demethylase KDM4A in cancer. *Cancer Genet.*, **208**, 215–224.
27. Yu,L., Wang,Y., Huang,S., Wang,J., Deng,Z., Zhang,Q., Wu,W., Zhang,X., Liu,Z., Gong,W. *et al.* (2010) Structural insights into a novel histone demethylase PHF8. *Cell Res.*, **20**, 166–173.
28. Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.
29. Zlatanova,J. and Caiafa,P. (2009) CTCF and its protein partners: divide and rule? *J. Cell Sci.*, **122**, 1275–1284.
30. Bailey,S.D., Zhang,X., Desai,K., Aid,M., Corradin,O., Cowper-Sal Lari,R., Akhtar-Zaidi,B., Scacheri,P.C., Haibe-Kains,B., Lupien,M. *et al.* (2015) ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nat. Commun.*, **2**, 6186.
31. Kim,K.-D., Tanizawa,H., Iwasaki,O. and Noma,K. (2016) Transcription factors mediate condensin recruitment and global chromosomal organization in fission yeast. *Nat. Genet.*, **48**, 1242–1252.
32. Simon,J.M., Hacker,K.E., Singh,D., Brannon,A.R., Parker,J.S., Weiser,M., Ho,T.H., Kuan,P.F., Jonasch,E., Furey,T.S. *et al.* (2014) Variation in chromatin accessibility in human kidney cancer links H3K36 methyltransferase loss with widespread RNA processing defects. *Genome Res.*, **24**, 241–250.
33. Qin,S. and Min,J. (2014) Structure and function of the nucleosome-binding PWWP domain. *Trends Biochem. Sci.*, **39**, 536–547.
34. Rhee,H.S. and Pugh,B.F. (2012) Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature*, **483**, 295–301.
35. Jackson,S.P., MacDonald,J.J., Lees-Miller,S. and Tjian,R. (1990) GC box binding induces phosphorylation of Sp1 by a DNA-dependent protein kinase. *Cell*, **63**, 155–165.
36. Marin,M., Karis,A., Visser,P., Grosveld,F. and Philipsen,S. (1997) Transcription factor Sp1 is essential for early embryonic development but dispensable for cell growth and differentiation. *Cell*, **89**, 619–628.

37. Frietze,S., O'Geen,H., Blahnik,K.R., Jin,V.X. and Farnham,P.J. (2010) ZNF274 recruits the histone methyltransferase SETDB1 to the 39 ends of ZNF genes. *PLoS One*, **5**, e15082.
38. Apweiler,R. (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.
39. Hoskins,B.E., Cramer,C.H., Silvius,D., Zou,D., Raymond,R.M., Orten,D.J., Kimberling,W.J., Smith,R.J.H., Weil,D., Petit,C. *et al.* (2007) Transcription factor SIX5 is mutated in patients with branchio-oto-renal syndrome. *Am. J. Hum. Genet.*, **80**, 800–804.
40. Wang,S.H., Wu,C.C., Lu,Y.C., Lin,Y.H., Su,Y.N., Hwu,W.L., Yu,I.S. and Hsu,C.J. (2012) Mutation screening of the EYA1, SIX1, and SIX5 genes in an East Asian cohort with branchio-oto-renal syndrome. *Laryngoscope*, **122**, 1130–1136.
41. Maeshima,K., Kaizu,K., Tamura,S., Nozaki,T., Kokubo,T. and Takahashi,K. (2015) The physical size of transcription factors is key to transcriptional regulation in chromatin domains. *J. Phys. Condens. Matter*, **27**, 64116.