Bioinformatics, 38(3), 2022, 621–630
https://doi.org/10.1093/bioinformatics/btab743
Advance Access Publication Date: 6 November 2021
Original Paper

Genome analysis

preciseTAD: a transfer learning framework for 3D domain boundary prediction at base-pair resolution

Spiro C. Stilianoudakis[1,†], Maggie A. Marshall[2,†] and Mikhail G. Dozmorov [1,*,†]

[1]Department of Biostatistics, Department of Pathology, Virginia Commonwealth University, Richmond, VA 23298, USA and [2]Bioinformatics Program, Virginia Commonwealth University, Richmond, VA 23298, USA

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, all the authors should be regarded as Joint First Authors.

Associate Editor: Peter Robinson

Received on August 10, 2021; revised on October 7, 2021; editorial decision on October 18, 2021

Abstract

Motivation: Chromosome conformation capture technologies (Hi-C) revealed extensive DNA folding into discrete 3D domains, such as Topologically Associating Domains and chromatin loops. The correct binding of CTCF and cohesin at domain boundaries is integral in maintaining the proper structure and function of these 3D domains. 3D domains have been mapped at the resolutions of 1 kilobase and above. However, it has not been possible to define their boundaries at the resolution of boundary-forming proteins.

Results: To predict domain boundaries at base-pair resolution, we developed *preciseTAD*, an optimized transfer learning framework trained on high-resolution genome annotation data. In contrast to current TAD/loop callers, *preciseTAD*-predicted boundaries are strongly supported by experimental evidence. Importantly, this approach can accurately delineate boundaries in cells without Hi-C data. *preciseTAD* provides a powerful framework to improve our understanding of how genomic regulators are shaping the 3D structure of the genome at base-pair resolution.

Availability and implementation: *preciseTAD* is an R/Bioconductor package available at https://bioconductor.org/packages/preciseTAD/.

Contact: mdozmorov@vcu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The advent of chromosome conformation capture (3C) sequencing technologies, and its successor Hi-C, have revealed a hierarchy of the three-dimensional (3D) structure of the human genome such as chromatin loops (Franke *et al.*, 2016; Rao *et al.*, 2014), Topologically Associating Domains (TADs) (Dixon *et al.*, 2012; Nora *et al.*, 2012; Sexton *et al.*, 2012) and A/B compartments (Lieberman-Aiden *et al.*, 2009; Rao *et al.*, 2014), reviewed in Beagan and Phillips-Cremins (2020) and Chang *et al.* (2020). At the kilobase scale, chromatin loops (corner-dot structures on Hi-C chromatin interaction maps) connect gene promoters with distal enhancers and regulate gene expression. At the megabase scale, TADs represent regions on the linear genome that are highly self-interacting. Perhaps the most prominent feature of TADs is that they are demarcated by boundaries constraining enhancer–promoter interactions (Sun *et al.*, 2019), although these constraints can be flexible (Freire-Pritchett *et al.*, 2017). Perturbation of boundaries have been reported to promote human cancers (Hnisz *et al.*, 2016; Taberlay *et al.*, 2016), neurological disorders (Sun *et al.*, 2018) and pathologies of limb development (Franke *et al.*, 2016; Lupianez *et al.*, 2016). Identifying the precise location of boundaries remains a top priority to fully understand the functionality of the human genome.

Several methods have been proposed to identify TAD boundaries [reviewed in Zufferey *et al.* (2018)], and chromatin loops (Ay *et al.*, 2014; Rao *et al.*, 2014; Salameh *et al.*, 2019). They are primarily based on identifying characteristic patterns in Hi-C contact matrices, such as dense inter-TAD contacts, sparse intra-TAD contacts, among other features (Beagan and Phillips-Cremins, 2020; Dixon *et al.*, 2012). Consequently, they are limited by Hi-C data resolution. Resolution refers to the size of genomic regions (bins) used to segment the linear genome and create Hi-C contact matrices. Lower resolution corresponding to larger bin sizes leads to increased uncertainty in domain boundary location.

TAD and loop boundaries are not mutually exclusive. Rao *et al.* demonstrated that boundaries of TADs, referred to as 'contact domains', are enriched in chromatin loops (Rao *et al.*, 2014). Numerous observations demonstrated the presence of hierarchically nested chromatin domains within TADs (Phillips-Cremins *et al.*, 2013; Rao *et al.*, 2014). Domain boundaries are thought to form by the 'loop extrusion' mechanism. During extrusion, the molecular motors (condensin and cohesin) track along the DNA sequence 'extruding' the intervening DNA in an ATP-dependent manner and pausing at the convergent CTCF motifs (Alipour and Marko, 2012; Davidson *et al.*, 2019; Fudenberg *et al.*, 2016; Goloborodko *et al.*, 2016; Hansen *et al.*, 2018; Mirny *et al.*, 2019; Sanborn *et al.*, 2015). Consequently, boundaries are expected to be enriched in CTCF and RAD21/SMC3, members of the cohesin complex (Dixon *et al.*, 2012; Phillips-Cremins *et al.*, 2013; Rao *et al.*, 2014; Tang *et al.*, 2015; Zuin *et al.*, 2014). More recently, ZNF143 has been identified as a cofactor of CTCF–Cohesin complex (Bailey *et al.*, 2015; Wen *et al.*, 2018). Furthermore, distinct patterns of histone modifications have also been shown to be present at boundaries (Dixon *et al.*, 2012; Lieberman-Aiden *et al.*, 2009).

In contrast to low-resolution Hi-C matrices, functional/regulatory genomic annotations [histone modifications, DNAse I hypersensitive sites, DNA methylation and transcription factor binding sites (TFBSs)] have been profiled at a relatively high resolution (10–300 bp) (Dozmorov, 2017; ENCODE Project Consortium, 2012). Genomic annotations have been used to predict functional chromatin contacts [e.g. HiC-Reg (Zhang *et al.*, 2019)], boundaries of chromatin domains [e.g. nTDP (Sefer and Kingsford, 2015), Lollipop (Kai *et al.*, 2018), 3DEpiLoop (Al Bkhetan and Plewczynski, 2018), TAD-Lactuca (Gan *et al.*, 2019)], with 48 methods recently reviewed by Tao *et al.* (2021) [see also Belokopytova and Fishman 2020 for a broader overview (Belokopytova and Fishman, 2020)]. Yet, these methods operate at the resolution of Hi-C data. Because increasing resolution of Hi-C data requires a quadratic increase in sequencing depth (Schmitt *et al.*, 2016) and the associated costs, most currently available Hi-C matrices have relatively low resolution, ranging from 1 to 100 kb. Furthermore, conventional Hi-C relies on a 0.1–10 kb size fragmentation by a restriction enzyme, but the existence of self-ligated products and undigested fragments with sizes 1–10 kb prohibits analysis at higher resolution (Jin *et al.*, 2013; Rao *et al.*, 2014). The association of domain boundaries with genomic annotations suggests these annotations may inform the more precise location of domain boundaries.

We present *preciseTAD*, an optimally tuned transfer learning framework for precise domain boundary detection using genomic annotation data. *preciseTAD* learns the associations between genomic annotations and boundaries detected from low-resolution Hi-C matrices and transfers the learned associations at base-level resolution (predicts the probability of each base being a boundary). This approach circumvents resolution restrictions of Hi-C matrices and allows for the precise detection of domain boundaries. We demonstrate that *preciseTAD*-predicted boundaries are strongly enriched in known molecular drivers of 3D chromatin including CTCF, RAD21, SMC3 and ZNF143. Further, we show that the associations learned in one cell line can be used to predict boundaries in other cell lines using cell-specific genomic annotations only. As such, *preciseTAD* allows for predicting domain boundaries in cells without Hi-C data. We provide domain boundary predictions for 60 cell lines, demonstrating that imputing missing cell-specific genomic annotations with Avocado (Schreiber *et al.*, 2020) is a viable approach to recover domain boundaries. The *preciseTAD* R package and the pre-trained models (*preciseTADhub* ExperimentHub package) are freely available on Bioconductor.

2 Materials and methods

2.1 Data sources

TAD and loop boundaries called by *Arrowhead* (Durand *et al.*, 2016) and *Peakachu* (Salameh *et al.*, 2019) tools were used as training and testing data. The autosomal genomic coordinates in the GRCh37/hg19 human genome assembly were considered. *Arrowhead*-defined TAD boundaries were called from Hi-C data for the GM12878 and K562 cell lines (MAPQ > 0, 5, 10, 25, 50 and 100 kb resolutions) using the *Arrowhead* tool from Juicer (Durand *et al.*, 2016) with default parameters. *Peakachu* chromatin loop boundaries for GM12878 and K562 cell lines were downloaded from the Yue lab website. Experimentally obtained (ChIA-PET) cohesin-mediated chromatin loops, which we refer to as Grubert data, were obtained from the Supplementary Table S4 of Grubert *et al.* (2020) (Supplementary Table S1). Unique boundaries were considered as the midpoints within the coordinates of each chromatin loop anchor. Chromosome 9 was excluded due to the inability to call *Arrowhead* domains at 5 and 10 kb resolutions for the K562 cell line (high sparsity), unless specified otherwise. Cell-line-specific genomic annotations [BroadHMM chromatin states (BroadHMM), histone modifications (HM) and transcription factor binding sites (TFBS)] were obtained from the UCSC Genome Browser Database (Supplementary Table S2).

2.2 Shifted-binning for binary classification

In Hi-C, each chromosome is binned into non-overlapping regions of length *r*, typically, 5 kb and above. The *r* parameter defines the resolution of Hi-C data. Here, we designed a strategy called shifted binning that partitions the genome into regions of the same length *r*, but with middle points corresponding to boundaries defined by the original binning. To create shifted binning, the first shifted bin was set to start at half of the resolution *r* and continued in intervals of length *r* until the end of the chromosome ($mod\ r + r/2$). The shifted bins, referred hereafter as bins for simplicity, were then defined as boundary-containing regions ($Y = 1$) if they contained a TAD (or loop) boundary, and non-boundary regions ($Y = 0$) otherwise, thus establishing the binary response vector (**Y**) used for classification (Supplementary Fig. S1A).

2.3 Feature engineering

Cell line-specific genomic annotations were used to build the predictor space. Bins were annotated by one of either the average signal strength of the corresponding annotation (*Peak Signal Strength*), the number of overlaps with an annotation [*Overlap Count (OC)*], the percent of overlap between the bin and the total width of genomic annotation regions overlapping it [*Overlap Percent (OP)*], or the distance in bases from the center of the bin to the center of the nearest genomic annotation region (*Distance*) (Supplementary Fig. S1B). A (log2 + 1)-transformation of distance was used to account for the skewness of the distance distributions (Supplementary Fig. S2). Models built using a *Peak Signal Strength* predictor space were only composed of histone modifications and transcription factor binding sites as BroadHMM chromatin states lack signal values.

2.4 Addressing class imbalance

To assess the impact of class imbalance (CI), defined as the proportion of boundary regions to non-boundary regions, we evaluated three resampling techniques: *Random Over-Sampling (ROS)*, *Random Under-Sampling (RUS)* and *Synthetic Minority Over-Sampling Technique (SMOTE)*. For ROS, the minority class was sampled with replacement to obtain the same number of data points in the majority class. For RUS, the majority class was sampled without replacement to obtain the same number of data points in the minority class. For SMOTE, under-sampling was performed without replacement from the majority class, while over-sampling was performed by creating new synthetic observations using the $k = 5$ minority class nearest neighbors (Chawla *et al.*, 2002) (implemented in the *DMwR* v.0.4.1 R package). We restricted the SMOTE algorithm to 100% over-sampling and 200% under-sampling to create perfectly balanced classes.

2.5 Establishing optimal data-level characteristics for boundary region prediction

Random forest (RF) classification models [the *caret* v.6.0 R package (Kuhn, 2012)] were compared between combinations of data resolutions, feature engineering procedures and resampling techniques. Following recommendations to evaluate the model on unseen data (Schreiber *et al.*, 2019), a *holdout chromosome* technique was used for estimating model performance. The *i*th holdout chromosome was identified and a data matrix, $A_{N\times(p+1)}$, was constructed by combining the binned genome from the remaining chromosomes (1,2,

$\cdots, i-1, i+1,\cdots,21,22$), where $N=[n_1\ n_2\cdots n_{21}\ n_{22}]'$ and n_k is the length of chromosome k after being binned into non-overlapping regions of resolution r, such that $k \neq i$. The number of annotations, p, and the response vector, $\mathbf{Y}$, defined the column-wise dimension of the matrix A. Re-sampling was then performed on A, and a RF classifier was trained using 3-fold cross-validation to tune for the number of annotations to consider at each node (*mtry*). The number of trees (*ntree*) that were aggregated for each RF model was set to 500. The minimum number of observations per root node (*nodesize*) was set to 0.1% of the rows in the data. The binned data for the holdout chromosome i was reserved for testing. The response vector associated with the testing data ($\mathbf{Y}_{\text{test}}$) was built using Grubert-defined chromatin loops as a ground truth when validating the models. Models were then evaluated using Balanced Accuracy (BA), defined as the average of sensitivity and specificity:

$$\text{BA} = \frac{1}{2}(\text{sensitivity} + \text{specificity}) = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP}+\text{FN}} + \frac{\text{TN}}{\text{TN}+\text{FP}}\right)$$

where values of the confusion matrix, true negatives (TP), false positives (FP), true negatives (TN) and false negatives (FN) were related to genomic bins that contained a Grubert-defined boundary in the test data. That is, TP refers to the number of bins correctly identified as containing a boundary (true positives), FP refers to the number of bins incorrectly identified as containing a boundary (false positives), TN refers to the number of bins correctly identified as not containing a boundary (true negatives) and FN refers to the number of bins incorrectly identified as not containing a boundary (false negatives). Each of these quantities is obtained from the confusion matrix created by validating the model on the test data. The process was repeated for each *i*th holdout chromosome, and performances were aggregated using the mean and standard deviation.

2.6 Feature selection and predictive importance

Many genomic annotations, notably architectural proteins, tend to exhibit an extensive pattern of colocalization (correlation). To suitably reduce the predictor space and improve computational efficiency, while maintaining optimal performance, we utilized recursive feature elimination (RFE). We estimated the near-optimal number of necessary features, ranging from two to the maximum number of features incremented by the power of two. We then aggregated the predictive importance of the union of the optimal set of features across holdout chromosomes using the mean decrease in node impurity among permuted features in out-of-bag samples to determine the most common and top-ranked annotations for predicting boundary regions.

2.7 Evaluating performance across cell lines

We used the same holdout chromosome strategy to evaluate a model trained in one cell line on unseen data from another cell line (Schreiber *et al.*, 2019). Given two cell lines, GM12878 and K562, we first evaluated the performance of cell line-specific models. That is, models trained on cell line-specific data from n - 1 chromosomes were evaluated on the *i*th holdout chromosome data from the same cell line. Second, we evaluated models trained on cell line-specific data from n - 1 chromosomes using the *i*th holdout chromosome data from a different cell line. That is, models trained using K562 cell line-specific data were evaluated on unseen chromosome data from the GM12878 cell line. This process was repeated for each holdout chromosome. To evaluate performance, we constructed receiver operating characteristic (ROC) curves composed of the average sensitivities and specificities at different cutoffs, across each holdout chromosome and reported the corresponding average area under the curve (AUC). As before, the response vector for the test data was derived from cell line-specific Grubert boundaries.

2.8 Boundary prediction at the base-level resolution using *preciseTAD*

To investigate whether we could alleviate the limitations of conventional domain calling tools operating at the Hi-C data resolution, we developed *preciseTAD*. This algorithm leverages an optimized random forest model predicting the probability of each base being a boundary followed by clustering of bases with high boundary probabilities (Supplementary Fig. S3). A random forest model was trained using boundaries at the Hi-C data. resolution on the optimal combination of predictor type (distance to transcription factor binding sites), resampling technique (random undersampling) and top-ranked annotations ($p \in \{CTCF, RAD21, SMC3, ZNF143\}$). To precisely identify boundary locations, we first constructed a base-level resolution predictor space for the chromosome i, $A_{n\times p}$, where n is the length of chromosome i in bases and p is the optimal number of annotations. We then applied the pre-trained model on the base-level predictor space to extract the probability vector, π_n, denoting each base's probability of being a boundary. Bases with the probability $\pi_n \geq t$ (the default boundary probability threshold $t = 1$) were clustered with the DBSCAN algorithm (v.1.1-5) (Hahsler *et al.*, 2019) into *preciseTAD boundary regions* (PTBR). To precisely identify the location of domain boundary, *preciseTAD* implements partitioning around medoids (PAM) on the distance matrix, D_k among bases with $\pi_n \geq t$ within each PTBR. The corresponding medoids were defined as *preciseTAD boundary points* (PTBPs). Intuitively, a PTBP corresponds to the base with the highest density of CTCF, RAD21, SMC3 and ZNF143.

The DBSCAN algorithm has two parameters, *MinPts* and *eps* (ϵ). The *MinPts* parameter, corresponding to the minimum size of a PTBR, was set to 100, an approximate size of ChIP-seq peaks. The maximum size of a PTBR was unconstrained and can be larger than resolution of the original Hi-C data due to the high density of CTCF and other proteins in certain genomic locations. To decide on the optimal value of t and ϵ, we considered the normalized enrichment (*NE*) of flanked boundaries. *NE* was calculated as the average number of overlaps between genomic annotations (CTCF, RAD21, SMC3 and ZNF143) and flanked boundaries, divided by the total number of boundaries. The rationale here is to find a combination of parameters producing the largest number of overlaps between predicted boundaries and genomic annotations. We evaluated *NE* for combinations of t={0.975,0.99,1.0} and ϵ={1000, 5000, 10 000, 15 000, 20 000, 25 000}. The heuristic of ϵ is that density-reachable bases with genomic distances less than ϵ should occupy the same designated cluster. The default combination was set to $t = 1.0$ and $\epsilon = 10\,000$ based on our tests (Supplementary Fig. S4).

2.9 Evaluating called and predicted boundary precision

We assessed the biological significance of our predicted boundaries by their association with the signal of CTCF, RAD21, SMC3 and ZNF143 using *deepTools* (version 2.0) (Ramirez *et al.*, 2016) (*computeMatrix*, *plotProfile* tools). In addition, we compared the median $\log_2$ genomic distances between TAD boundaries and the same top predictive ChIP-seq annotations using Wilcoxon Rank-Sum tests. Furthermore, we compared the overlap between predicted and called boundaries in GM12878 and K562 cell lines. Boundaries were first flanked by resolution, r, and overlaps were visualized using Venn diagrams from the *Vennerable* R package (version 3.1.0). Overlaps were further quantified using the Jaccard index defined as

$$J_{(A,B)} = \frac{A \cap B}{A \cup B}$$

where A and B represent genomic regions created by flanked boundaries. All statistical analyses were performed in R (version 4.0.1). The significance level was set to 0.05 for all statistical tests.

2.10 Predicting boundaries across cell lines

We implemented a strategy to predict domain boundaries across cell lines. To do so, we trained *preciseTAD* models on boundaries and genomic annotation data from one cell line and used it to predict boundaries using genomic annotation data from another cell line. Results were compared by assessing the overlap between flanked same-cell-line and cross-cell-line predicted boundaries using Venn diagrams, Jaccard indices and signal distribution plots.

To predict boundaries in 60 cell lines, we first compiled a set of cell line-specific genomic annotations (CTCF, RAD21, SMC3). For cells lacking some or all genomic annotations, we used Avocado v.0.3.6 (Schreiber *et al.*, 2020) to impute missing annotations. Avocado imputes signal profiles as bigWig files which we converted to bedGraph and called peaks using UCSC's bigwigtobedgraph and MACS2 v.2.2.7.1 with default settings. We note that Avocado can impute data for only three transcription factors, CTCF, RAD21, SMC3. Therefore, we retrained *preciseTAD* models using these three transcription factors in GM12878 cell line and *Arrowhead/Peakachu* boundaries. For Avocado predictions, the following cell line names were manually matched: H1 = H1-hESC, H7 = H7-hESC. Predictions were made using the hg38 genome assembly; therefore, all data were either downloaded as hg38 or lifted over to hg38 genome assembly.

2.11 Data availability

All datasets used in this work are summarized in Supplementary Tables S1, S2 and S6. The predicted domain boundary regions and points for 60 cell lines can be downloaded from https://dozmorovlab.github.io/preciseTAD.

2.12 Code availability

preciseTAD is available on Bioconductor https://bioconductor.org/packages/preciseTAD/ and GitHub https://github.com/dozmorovlab/preciseTAD/ under the MIT license.

3 Results

3.1 *Precisetad* overview

preciseTAD implements the idea of transfer learning across genomic data resolutions. It models the association between chromatin domain boundaries and genomic annotations using low-resolution (5–100 kb) Hi-C genomic regions and applies this model at base-level resolution to predict the probability of each base being a boundary (Fig. 1). Our method utilizes the random forest (RF) algorithm trained on chromatin state (BroadHMM), histone modification (HM) and transcription factor binding site (TFBS) annotation data. Our training/testing framework was used to determine the optimal set of data-level characteristics including resolution (bin size), feature engineering and resampling (Supplementary Fig. S5). We found that spatial associations (linear distance) between boundaries and genomic annotations perform best, transcription factor binding sites outperform other annotations, and a simple random undersampling technique addresses the negative effect of class imbalance. *preciseTAD* uses density-based clustering (DBSCAN) and partitioning around medoids (PAM) to detect annotation-guided boundary regions and summit points with the highest boundary probability. These improved domain boundary locations can provide insight into the association between genomic regulators and the 3D genome organization.

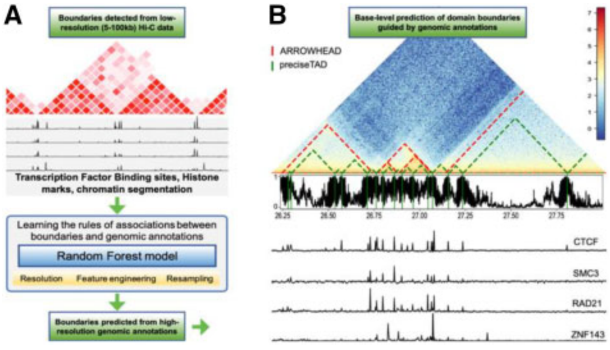
A
Boundaries detected from low-resolution (5-100kb) Hi-C data
Transcription Factor Binding sites, Histone marks, chromatin segmentation
Learning the rules of associations between boundaries and genomic annotations
Random Forest model
Resolution
Feature engineering
Resampling
Boundaries predicted from high-resolution genomic annotations
B
Base-level prediction of domain boundaries guided by genomic annotations
ARROWHEAD
preciseTAD
7
6
5
4
3
2
1
0
1
0
26.25
26.50
26.75
27.00
27.25
27.50
27.75
CTCF
SMC3
RAD21
ZNF143

Fig. 1. Overview of *preciseTAD*. (**A**) The framework of *preciseTAD* includes Random Forest models learning the optimal rules of associations between low-resolution boundaries and genomic annotations, and predicting the probability of each base being a boundary. First, feature engineering steps are applied to capture various rules of boundary-annotation associations. Second, the models are trained on different data resolution, addressing class imbalance. Third, the most predictive features shared across chromosomes and cell lines are selected for the final model. Fourth, a base-level predictor matrix is built, in which each base is annotated with the most optimal features and association rules. Fifth, the probability of each base being a boundary is predicted with the optimal model trained on low-resolution Hi-C data. (**B**) The final step includes clustering bases with high boundary probability into *preciseTAD* boundary regions (PTBRs) using DBSCAN and identifying the most likely boundary points (PTBPs) using PAM. The output of *preciseTAD* includes genomic coordinates of PTBRs, PTBPs and their summary statistics

3.2 Developing a ML framework for optimal boundary prediction

We developed a machine learning (ML) framework for determining the optimal set of data level characteristics to predict boundary regions of Topologically Associating Domains (TADs) and chromatin loops, collectively referred to as domain boundaries. Similar to other boundary prediction methods (Al Bkhetan and Plewczynski, 2018; Kai *et al.*, 2018; Salameh *et al.*, 2019; Wang *et al.*, 2021; Zhang *et al.*, 2019), we chose the random forest (RF) algorithm as our binary classification tool. The reason for it is twofold: (i) to devise a tunable prediction rule in a supervised learning framework that is both robust to overfitting and able to handle multiple correlated predictors, and (ii) to allow for an interpretable ranking of predictors (Boulesteix *et al.*, 2012). Furthermore, previous reports demonstrated robustness of RF to overfitting and superior performance over other machine learning classifiers (Al Bkhetan and Plewczynski, 2018; Gan *et al.*, 2019; Wang *et al.*, 2021).

As an example of TAD boundaries, we derived cell line-specific boundaries at 5–100 kb resolutions using *Arrowhead* (Rao *et al.*, 2014). For chromatin loops, we used loops derived by *Peakachu* (Salameh *et al.*, 2019) (Supplementary Table S1). We chose GM12878 (lymphoblastoid) and K562 (chronic myelogenous leukemia) as cell lines with the most rich and comparable sets of genomic annotations. The choice of *Peakachu* loops was motivated by their good overlap with HiCCUP (Rao *et al.*, 2014) and Fit-Hi-C (Ay *et al.*, 2014) loops and better enrichment in known boundary factors. To ensure the cell line-specific predictions corresponded to experimental domain boundaries, we used cohesin-bound chromatin loop data from ENCODE phase 3 (Grubert *et al.*, 2020), referred hereafter as *Grubert* boundaries (Supplementary Table S1). We found that boundaries *Arrowhead/Peakachu* boundaries showed markedly little enrichment in binding of CTCF and other architectural proteins as compared with *Grubert* boundaries (Supplementary Fig. S6A and C), further motivating the need for more precise boundary detection.

The total number of called TADs, their unique boundaries, and the number of genomic bins expectedly decreased with the decreased resolution of Hi-C data (Table 1, Supplementary Table S2). The number of non-boundary regions highly outnumbered boundary regions. Such a disproportional presence of examples in one class is known as a 'class imbalance' problem that negatively affects predictive modeling (Wei and Dunbrack Jr, 2013). To address class imbalance, we evaluated the three resampling techniques. *Random over-sampling* (ROS) was defined as sampling with replacement from the minority class (boundary regions). *Random under-sampling* (RUS) was defined as sampling with replacement from the majority class (non-boundary regions). Finally, we tested *Synthetic minority over-sampling technique* (SMOTE), which is a combination of both random over- and under-sampling to create balanced classes (Chawla *et al.*, 2002) (see Section 2).

Our models were trained on cell line-specific functional genomic annotation data from ENCODE (ENCODE Project Consortium, 2012). A total of 77 cell line-specific genomic annotations were used to build the predictor space. These included histone modification (HM) data previously shown to be useful for boundary predictions (Al Bkhetan and Plewczynski, 2018; Gan *et al.*, 2019; Sefer and Kingsford, 2015), Broad ChromHMM chromatin segmentation data that captures regions with similar epigenetic activity patterns (BroadHMM), and transcription factor binding sites (TFBS, Supplementary Table S3). Boundary regions were defined as

Table 1. Domain boundary data and class imbalance summaries across resolutions for *Arrowhead*, *Peakachu* and Grubert data in GM12878 cell line

Tool	Resolution/bin size	Total number of called TADs/loops	Total number of unique domain boundaries	Total number of genomic bins	Class imbalance
Arrowhead	5 kb	8052	15 468	535 363	0.03
Arrowhead	10 kb	7676	14 253	267 682	0.05
Arrowhead	25 kb	4670	8363	107 073	0.08
Arrowhead	50 kb	2349	4224	53 537	0.08
Arrowhead	100 kb	1031	1883	26 768	0.07
Peakachu	10 kb	16 185	21 421	267 682	0.14
Grubert	5 kb	16 232	18 455	535 363	0.07

genomic bins containing a called boundary ($Y = 1$), while nonboundary regions were defined as bins that did not contain a called boundary ($Y = 0$, Supplementary Fig. S1, see Section 2).

Four feature engineering procedures were developed to quantify the association between genomic annotations and bins. These included signal strength association (Signal), direct (OC), proportional (OP) and spatial ($\log_2$+1 Distance) relationships (Supplementary Fig. S5, see Section 2). In total, we considered combinations of data from two cell lines $L = \{GM12878, K562\}$, five resolutions $R = \{5\,kb, 10\,kb, 25\,kb, 50\,kb, 100\,kb\}$, four types of predictor spaces $P = \{Signal, OC, OP, Distance\}$ and three re-sampling techniques $S = \{None, RUS, ROS, SMOTE\}$. Once the model inputs were established, a random forest classifier was trained on *n-1* autosomal chromosomes, while reserving the *i*th chromosome for testing. Threefold cross-validation was used to tune the *mtry* hyperparameter, while *ntree* and *nodesize* were fixed at 500 and at 0.1% of the rows in the training data, respectively. Models were validated on the testing data using cell line-specific Grubert-defined boundaries as $\mathbf{Y}_{test}$. Model performance was evaluated by aggregating the mean balanced accuracy (BA) across each holdout chromosome, with additional performance metrics (accuracy, AUROC, AUPRC) shown in Supplementary Table S4. These strategies allowed us to select the best-performing model characteristics (Supplementary Fig. S5, see Section 2).

3.3 Random under-sampling, distance-based predictors and high-resolution Hi-C data provide optimal performance for boundary prediction

When using *Arrowhead* data with class imbalance, the models exhibited low balanced accuracies, with minimal variability among different resolutions (Fig. 2). Similarly, poor performances were found when using ROS. However, RUS and SMOTE resampling led to a drastic improvement in performance, especially at higher resolutions. We found that RUS marginally outperformed SMOTE under most conditions and used it for the subsequent analyses unless noted otherwise. In addition, we found that distance-type predictor space yielded substantially higher balanced accuracy than the peak signal strength, overlap count and overlap percent predictor types.

As with class balancing techniques, this improvement was less evident at lower resolutions, with results consistent for K562 (Supplementary Fig. S7A). Furthermore, 5 kb resolution genomic bins led to the optimal prediction for TAD boundary regions on both cell lines. These observations were replicated when *Peakachu*-defined loop boundary regions were used (Supplementary Fig. S7B and C). Our results indicate that random under-sampling, distance-type predictors and high-resolution Hi-C data provide the optimal set of data characteristics for both TAD and loop boundary prediction.

3.4 Transcription factor binding sites outperform histone- and chromatin state-specific models

We hypothesized that the class of genomic annotation may also affect predictive performance. Using the established optimal settings (RUS, Distance, 5 kb/10 kb resolution), we used histone modifications (HM), chromatin states (BroadHMM) and transcription factor

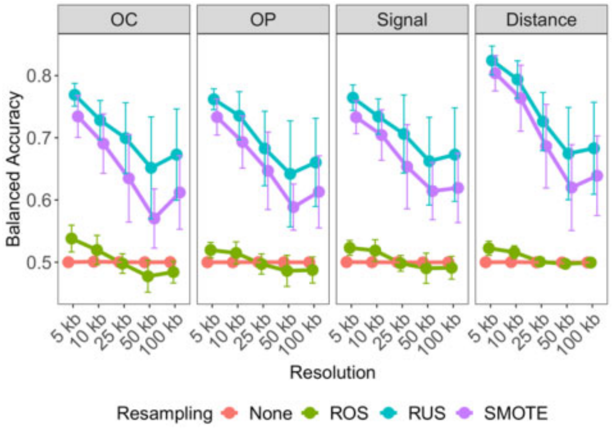

Fig. 2. Determining optimal data-level characteristics for building TAD boundary region prediction models on GM12878. Averaged balanced accuracies are compared across resolution, within each predictor-type: overlap count (OC), overlap percent (OP), average Signal and Distance and across resampling techniques: no resampling (None; red), random over-sampling (ROS; green), random under-sampling (RUS; blue) and synthetic minority over-sampling (SMOTE; purple). Error bars indicate 1 standard deviation from the mean performance across each holdout chromosome used for testing

binding sites (TFBS) to build the predictor space. Despite previous success in using histone modifications for TAD boundary predictions (Al Bkhetan and Plewczynski, 2018; Gan *et al.*, 2019; Zhang *et al.*, 2019), their performance in our settings was least optimal. BroadHMM segmentations also performed less optimally, in agreement with previous observations (Sefer and Kingsford, 2015). We found that TFBSs outperformed other annotation-specific models, with results consistent for loop boundaries, on both cell lines (Fig. 3A; Supplementary Fig. S8A), and the use of all genomic annotations did not significantly improved model performance. These results suggest that transcription factors are the primary drivers of *Arrowhead*/*Peakachu*-defined boundaries in both GM12878 and K562 cell lines.

3.5 Feature importance confirmed the biological role of CTCF, RAD21, SMC3 and ZNF143 for boundary formation

We sought to further optimize our boundary region prediction models. We implemented recursive feature elimination to avoid overfitting and selected only the most influential features across all chromosomes. We were able to obtain near-optimal performance using approximately eight TFBS (Fig. 3B; Supplementary Fig. S8B). However, given that we trained our models on chromosome-specific data, the most significant annotations varied for each chromosome. To determine transcription factors most important for genome-wide boundary prediction, we clustered the predictive importance (mean decrease in accuracy) of the top eight significant TFs across

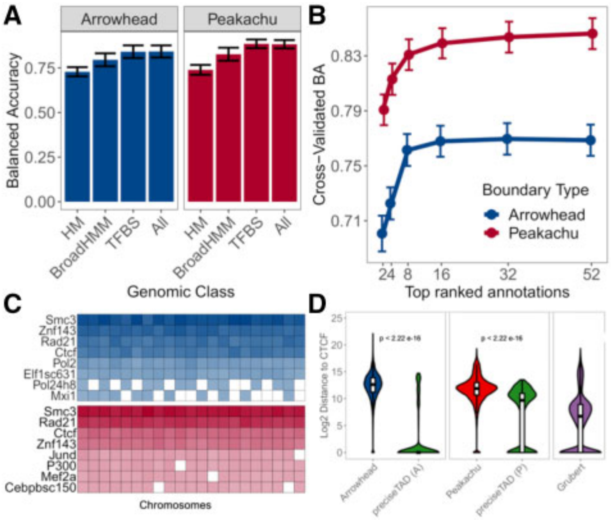

Fig. 3. SMC3, RAD21, CTCF and ZNF143 transcription factors accurately predict TAD and loop boundaries in GM12878. (**A**) Barplots comparing performances of TAD (*Arrowhead*) and loop (*Peakachu*) boundary prediction models using histone modifications (HM), chromatin states (BroadHMM), transcription factor binding sites (TFBS), in addition to a model containing all three classes (ALL). (**B**) Recursive feature elimination (RFE) analysis used to select the optimal number of predictors. Error bars represent 1 standard deviation from the mean cross-validated accuracy across each holdout chromosome. (**C**) Clustered heatmap of the predictive importance for the union of the top 8 most predictive chromosome-specific TFBSs. The columns represent the holdout chromosome excluded from the training data. Rows are sorted in decreasing order according to the columnwise average importance. (**D**) Violin plots illustrating the log_2 genomic distance distribution from original *Arrowhead*/*Peakachu* boundaries versus *preciseTAD*-predicted boundaries to the nearest CTCF sites. The *P*-values are from the Wilcoxon Rank Sum test

chromosomes. We found four transcription factors, CTCF, RAD21, SMC3 and ZNF143, being consistently predictive of TAD and loop boundaries (Fig. 3C; Supplementary Fig. S8C). Although CTCF and cohesin binding are known to colocalize, cohesin peaks are slightly shifted to the 3′ ends of convergently oriented motifs (Fudenberg *et al.*, 2016; Tang *et al.*, 2015) and appear to complement the model's performance. We selected these top four TFBS when building the random forest model, thereby decreasing computational burden while maintaining high predictive power. Comparison of distance-to-nearest-CTCF distribution between original *Arrowhead*/*Peakachu* boundaries and *preciseTAD*-predicted boundaries showed that the latter were located closer to CTCF binding sites (Fig. 3D, Supplementary Fig. S8D), and these results were observed for other transcription factors (Supplementary Fig. S9). In summary, our model was able to yield the known molecular drivers of the loop extrusion model (Alipour and Marko, 2012; Davidson *et al.*, 2019; Fudenberg *et al.*, 2016; Hansen *et al.*, 2018; Mirny *et al.*, 2019; Sanborn *et al.*, 2015), suggesting that TAD and loop boundary formation may be carried out by similar mechanisms (Beagan and Phillips-Cremins, 2020).

3.6 *Precisetad* identifies precise and biologically relevant domain boundaries

Using our optimally built random forest model trained on *Arrowhead*/*Peakachu* boundaries, we attempted to predict the more precise location of boundaries at base-level resolution. Intuitively, instead of bin-level annotations, the predictor-response space was built on a base-level. That is, each base was annotated with the distance to the nearest/overlapping CTCF, RAD21, SMC3 and ZNF143 site. The model trained on a bin-level space was then applied on a base-level space to predict each base's probability of being a boundary. Our method, referred to as *preciseTAD*, uses density-based spatial clustering (DBSCAN) and partitioning around medoids (PAM) to cluster bases with high probability of being a boundary into boundary regions (PTBRs) and summit points (PTBPs, see Section 2; Supplementary Fig. S3). We found that *Arrowhead* PTBRs and *Peakachu* PTBRs were highly overlapping (Jaccard 0.606/0.757, GM12878/K562 cell line, respectively) as compared with the less overlapping original *Arrowhead*/*Peakachu* boundaries (Jaccard 0.227/0.199, Supplementary Fig. S10A). Similarly, *Arrowhead* PTBRs and *Peakachu* PTBRs showed better agreement with experimental Grubert data (e.g. original *Arrowhead*-Grubert Jaccard 0.260 versus *Arrowhead* PTBRs-Grubert Jaccard 0.292, GM12878 cell line), and their results were consistent in K562 cell line (Supplementary Fig. S10B). These results suggest that *preciseTAD* identifies similar boundaries when trained on either *Arrowhead* or *Peakachu* data, and these boundaries better agree with experimentally observed data.

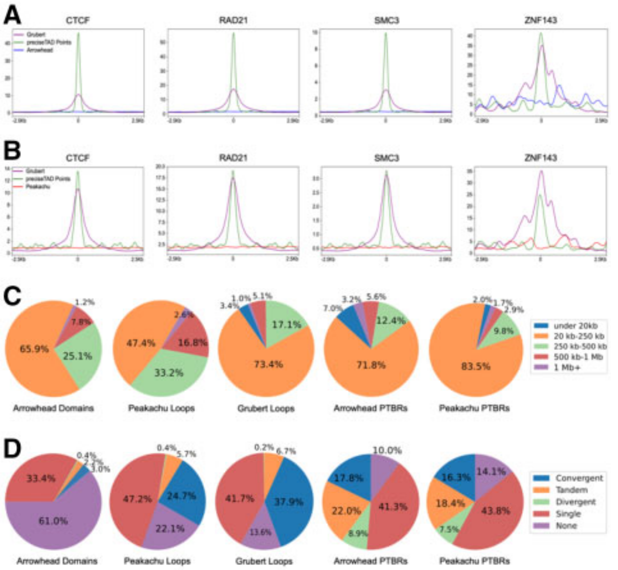

Fig. 4. *preciseTAD* boundaries are more enriched for known molecular drivers of 3D chromatin. Signal enrichment strength of CTCF, RAD21, SMC3 and ZNF143 sites around midpoints of *preciseTAD*-predicted boundaries (green) compared with midpoints of (**A**) *Arrowhead*-called boundaries (blue), (**B**) *Peakachu* loop boundaries (red). Data for midpoints of Grubert cohesin loop boundaries is shown as a proxy for experimental 'ground truth' (purple). Panel insets show signal enrichment around *preciseTAD* boundary points versus Grubert ground truth. (**C**) Domain size distribution and (**D**) CTCF orientation analysis. Data for GM12878 cell line are shown

When trained using *Arrowhead* and *Peakachu* boundaries at 5 and 10 kb, respectively, the *preciseTAD* model predicted a total of 10 990 domain and 14 440 chromatin loop boundaries in GM12878, as well as 9277 domain and 10 896 chromatin loop boundaries in K562 cell line (Supplementary Table S5). To evaluate the biological significance of *preciseTAD* PTBRs, we investigated signal distribution of four known molecular drivers of 3D chromatin (CTCF, RAD21, SMC3 and ZNF143) around boundaries detected by different methods. *preciseTAD*-predicted boundary points (the base-level boundary locations, PTBPs) showed much stronger signal distribution than *Arrowhead* and *Peakachu* boundaries, and frequently outperformed experimentally obtained Grubert data (Fig. 4A and B, Supplementary Fig. S11A and B). Notably, boundaries called by other callers similarly lacked signal distribution specificity (Supplementary Fig. S6A and C). Surprised by the poor performance of domain boundaries detected from Hi-C data, we investigated the overlap of boundaries detected by different callers with CTCF, RAD21, SMC3 and ZNF143 binding sites (TFBSs). We observed less than 30% of *Arrowhead* boundaries and approximately 70% of *Peakachu* boundaries overlapped CTCF and other TFBSs. In contrast, 90–99% of PTBRs detected by *Arrowhead*- and *Peakachu*-trained models overlapped CTCF and other TFBSs (Supplementary Table S7). Furthermore, the locations of CTCF and other TFBSs within *Arrowhead* and *Peakachu* boundary regions were found to be relatively uniform, leading to even signal distribution observed in Supplementary Figure 4A and B. In contrast, *preciseTAD* detects boundary points centered on the strongest CTCF signal. We further compared signal distribution around *preciseTAD* PTBPs with boundaries reported by Lollipop, a method for domain boundary prediction from genome annotation data and various domain characteristics (Kai *et al.*, 2018). Notably, *preciseTAD*- and Lollipop-predicted boundaries showed comparable signal distribution, cementing the importance of genomic annotations for domain boundary prediction (Supplementary Fig. S6B and D). Our results indicate that *preciseTAD*-predicted boundaries (PTBRs) and boundary points (PTBPs) better reflect the known biology of boundary formation when compared with boundaries called solely from Hi-C contact matrices.

We further investigated domain size distribution (Fig. 4C, Supplementary Fig. S11C). We used loop size distribution in experimentally observed Grubert data as ‘ground truth’. We found that the original *Arrowhead* and *Peakachu* algorithms detected higher proportion of large domains (e.g. 25.1%/33.2% of 250–500 kb domains versus 17.1% in Grubert, Figure 4C, GM12878 data). In contrast, size distribution of domains marked by *preciseTAD*-detected PTBRs was similar to that of Grubert. When trained on either *Arrowhead/Peakachu* data, *preciseTAD* identified 71.8%/83.5% 20–250 kb small-sized domains versus 73.4% Grubert, 12.4%/9.8% 250–500 kb mid-sized domains versus 17.1% Grubert, and 5.6%/2.9% large-sized domains versus 5.1% Grubert, with those results consistent in the K562 cell line (Fig. 4C, Supplementary Fig. S11C). Importantly, the larger proportions of 20–250 kb small-sized domains identified by *preciseTAD* are in better agreement with 80–120 kb domain size estimated by microscopy and via modeling of Hi-C data (Goloborodko *et al.*, 2016; Naumova *et al.*, 2013). We should note that, in contrast to *Arrowhead/Peakachu*, *preciseTAD* identifies individual boundaries; consequently, we measured domain size as the distance between consecutive boundaries. Thus, we expect some excessively large domain sizes (e.g. if two PTBRs span centromere region) and very small domain sizes (e.g. PTBRs separated by gaps in poorly organized regions). Despite this limitation, our results suggest that *preciseTAD* identifies boundaries better reflecting experimentally observed domain size distribution.

Directionality of CTCF binding, e.g. convergent orientation of CTCF motifs, is a known defining feature of domain formation (Rao *et al.*, 2014; Tang *et al.*, 2015). We used CTCF orientation distribution in experimentally observed Grubert boundaries as ‘ground truth’. Indeed, Grubert loops contained the largest proportion of convergent CTCF motifs (37.9%) as compared with *Arrowhead* domains (3.0%). *Arrowhead* data contained a large proportion of domains lacking CTCF motifs (61.0%) or domains having single CTCF motifs (33.4%, Fig. 4D). In contrast, *Peakachu* and *preciseTAD*-called domains showed high resemblance to CTCF motif orientation observed in Grubert data. Domains defined by the original *Peakachu* algorithm, *Arrowhead* PTBRs and *Peakachu* PTBRs (*preciseTAD*-defined regions trained on *Arrowhead/Peakachu* data) contained a high proportion of convergent CTCF motifs (24.7%, 17.8%, 16.3%), tandem motifs (5.7%, 22.0%, 18.4% versus 6.7% Grubert) and single CTCF motifs (47.2%, 41.3%, 43.8% versus 41.7% Grubert, Fig. 4D). Domains defined by *preciseTAD* PTBRs contained the largest proportion of divergent CTCF sites (8.9% for *Arrowhead* PTBRs and 7.5 for *Peakachu* PTBRs, Fig. 4D). While this may be attributed to the aforementioned noncontiguous nature of domains defined by PTBRs, these findings may reflect recent observations that domain boundaries contain divergent CTCF motifs while convergent motifs mark the interior of domains at 5–100 kb range (Nanni *et al.*, 2020). Together with strong signal enrichment and domain size distribution results, our observations indicate that *preciseTAD* may identify domain boundaries and boundary points better reflecting known biology of boundary formation.

3.7 Training in one cell line accurately predicts boundary regions in other cell lines

Previous studies suggest that TAD boundaries are relatively similar across cell lines (Dixon *et al.*, 2012; Nora *et al.*, 2012; Sexton *et al.*, 2012). To assess the level of cross-cell-line similarity, we evaluated the overlap between cell line-specific boundaries detected by *Arrowhead* and *Peakachu* methods as well as *preciseTAD*-predicted boundaries trained on *Arrowhead* and *Peakachu* data. Only 24% and 30% of boundaries were overlapping between cell lines for *Arrowhead* and *Peakachu* boundaries ($J = 0.246$ and $J = 0.295$), respectively (Fig. 5A). In contrast, *preciseTAD*-predicted boundaries were more similar between GM12878 and K562 cell lines regardless of which data were used for training (*Arrowhead* PTBRs overlap $J = 0.383$; *Peakachu* PTBRs $J = 0.467$, Fig. 5B). This better agreement between cell type-specific *preciseTAD*-predicted boundaries further supports the notion of their higher biological relevance.

Our observation that *preciseTAD* predicts similar domain boundaries when trained on either *Arrowhead* or *Peakachu* data raises the possibility that boundary-annotation associations learned in one cell line can predict boundaries in another cell line using its genomic annotation data. That is, given that genomic annotations (distance to CTCF, RAD21, SMC3 and ZNF143) are predictive of boundaries in one cell line, their locations in another cell line may be predictive of boundaries in that cell line. Indeed, training and testing using *Arrowhead* boundaries and genomic annotation data from the GM12878 cell line resulted in an average AUC = 0.840 (Fig. 5C). When training on the K562 boundaries/annotations and testing on GM12878, the average AUC was 0.899. Likewise, training and testing using *Peakachu* boundaries and genomic annotation data from the GM12878 cell line was comparable to models trained on K562 boundaries/annotations and testing on GM12878 cell line (Avg. AUC = 0.923 and 0.876, respectively, Fig. 5A). These results were consistent when comparing training/testing strategies on K562 boundaries/annotations with training on GM12878 and testing on K562 data (Supplementary Fig. S12).

In both instances, the average ROC curves were found to be within 1 standard deviation of each other, suggesting that a model trained on data from one cell line performs well when using the data from another cell line. This ability of boundary-annotation associations learned from one cell type to successfully predict boundaries in another indicates that the same underlying forces may drive boundary formation across various cell lines.

We further evaluated biological characteristics of cell type-specific boundaries predicted by models trained on a different cell line. We evaluated two scenarios: (1) training on GM12878 and predicting boundaries on GM12878 (GM on GM) versus training on K562 and predicting on GM12878 (K on GM), and (ii) training on K562 and predicting boundaries on K562 (K on K) versus training on GM12878 and predicting boundaries on K562 (GM on K). Using *Arrowhead*-trained models, 76% ($J = 0.701$) and 81% ($J = 0.751$) of predicted boundaries overlapped in both cross-cell-line prediction scenarios (Supplementary Fig. S13). When using *Peakachu*-trained models, we observed 85% ($J = 0.705$) and 88% ($J = 0.759$) overlap (Supplementary Fig. S14). Furthermore, boundaries predicted on

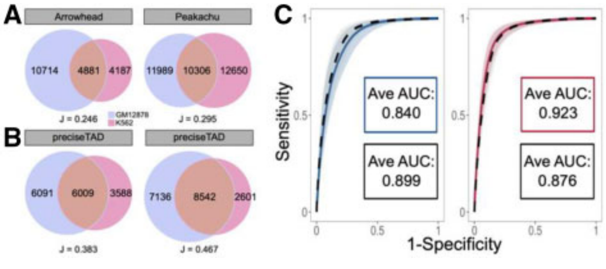

Fig. 5. *preciseTAD* models trained in one cell line can accurately predict boundaries in another cell line. (**A**) Venn diagrams of overlap original *Arrowhead/Peakachu* domain boundaries and (**B**) *Peakachu*-predicted PTBRs for GM12878 (red) and K562 (blue) cell lines. All boundaries were flanked by 5 kb. (**C**) Receiver operating characteristic (ROC) curves and the corresponding average area under the curves (AUCs) when training and testing on GM12878 data (blue, *Arrowhead*; red, *Peakachu*) versus training on K562 and testing on GM12878 data (black, dashed). The curves represent the average sensitivities and specificities across each holdout chromosome. The shaded areas around each curve represent 1 standard deviation from the average

unseen annotation data exhibited a similar level of enrichment for CTCF, RAD21, SMC3 and ZNF143 as did those trained and predicted on the same cell line (Supplementary Figs S13 and S14). These results indicate that *preciseTAD* pre-trained models can be successfully used to accurately predict domain boundaries for cell lines lacking Hi-C data but for which genome annotation data are available.

3.8 Boundary predictions across cell lines

Given the success of *preciseTAD* in predicting domain boundaries across cell lines, we investigated the possibility of predicting boundaries for cell lines with CTCF/RAD21/SMC3/ZNF143 genome annotation data. The ENCODE project provides information about 132 human cell lines; however, only 4 of them have all four annotations and 48 have none (Supplementary Fig. S15A, Supplementary Table S6). Therefore, we considered Avocado, a deep learning method providing pre-trained models to impute missing genomic annotations for 400 cell- and tissue types (Schreiber *et al.*, 2020). Avocado was able to predict the location of three transcription factors (CTCF/RAD21/SMC3) in 60 cell lines available in ENCODE. We found that the signal distributions for all three factors around the midpoints of experimental ENCODE and Avocado-predicted TFBSs are comparable, and the Avocado-predicted CTCF sites are a subset of ENCODE CTCF sites (Fig. 6A). These results suggest that Avocado-predicted genomic annotations can be used when cell type-specific ENCODE data is missing.

We retrained *preciseTAD* models to use CTCF/RAD21/SMC3 annotations. Using three instead of four transcription factors resulted in a non-significant drop in performance (Supplementary Fig. S15B), suggesting ZNF143 is less critical for boundary prediction. We applied those models to predict domain boundaries in 60 cell lines, using Avocado-predicted genomic annotations when ENCODE data was unavailable. To examine the predictive power of Avocado genomic annotations, we compared PTBRs and PTBPs predicted from ENCODE and Avocado annotations (Fig. 6B). We found that the signal distribution around PTBPs predicted using Avocado data was less than that of ENCODE but remained comparable with that observed from Grubert data (Fig. 6B). Similarly, Avocado PTBPs showed less overlap with Grubert data (Fig. 6B). The inferior performance of Avocado-predicted genomic annotations is expected as Avocado has been reported to perform less optimally in imputing transcription factors (Schreiber *et al.*, 2020). In summary, we demonstrate that *preciseTAD* predicts boundaries comparable with those observed experimentally and provides base-level domain boundary predictions for 60 cell lines.

4 Discussion

We present *preciseTAD*, a transfer learning approach for the precise prediction of TAD and chromatin loop boundaries from functional genomic annotations. *preciseTAD* leverages a random forest (RF) classification model built on boundaries obtained from low-resolution chromatin conformation capture data, and high-resolution genomic annotations as the predictor space. *preciseTAD* predicts the probability of each base being a boundary, and identifies the precise location of boundary regions and the most likely boundary points. We performed extensive optimization of our RF model by systematically comparing different Hi-C data resolutions, feature engineering procedures and resampling techniques. Our results demonstrate that distance between boundary regions and genomic annotations coupled with random under-sampling results in the best model performance. We show that binding of four transcription factors (SMC3, RAD21, CTCF, ZNF143) is sufficient for accurate boundary predictions. Compared with ChIA-PET-detected cohesin-mediated loops (Grubert *et al.*, 2020), we showed that *preciseTAD*-predicted boundaries better agree with biological properties of experimental data. Models trained in one cell type can accurately predict boundaries in another cell type without Hi-C data, requiring only cell type-specific genomic annotations. *preciseTAD* is implemented as an R package, while pre-trained models for predicting domain boundaries using genomic annotation data are provided via an ExperimentHub R package *preciseTADhub*. We provide a resource of predicted boundaries for 60 cell lines using Avocado-imputed genomic annotation data for cells lacking experimental data.

preciseTAD allows for predicting any type of 3D features observed in chromatin conformation capture data. Emerging evidence suggests the existence of different types of 3D domains and domain boundaries (Beagan and Phillips-Cremins, 2020; Chang *et al.*, 2020). Consequently, subpopulations of 3D domains can be defined by optimizing different sets of computational and biological characteristics (e.g. enrichment of CTCF binding motifs, high occupancy of CTCF/RAD21/H3K36me3 at boundaries, reproducibility, high intra- versus inter-TAD difference in contact frequencies) (Sauerwald and Kingsford, 2021), using different training data (e.g. CTCF- and RNAPII ChIA-PET) (Al Bkhetan and Plewczynski, 2018), or distinguishing long-range cohesin-dependent and short-range cohesin-independent domains (Phillips-Cremins *et al.*, 2013; Thiecke *et al.*, 2020). Boundaries have also been defined by the patterns of CTCF orientation (Nanni *et al.*, 2020), actively transcribed regions (Harrold *et al.*, 2020) and the level of hierarchy (Cresswell *et al.*, 2020; Fraser *et al.*, 2015; Weinreb and Raphael, 2016). Recent research distinguishes CTCF-associated boundaries, CTCF-negative YY1-enriched boundaries, CTCF- and YY1- depleted promoter boundaries, and the fourth class of weak boundaries largely depleted of all three features (Krietenstein *et al.*, 2020). *preciseTAD* can be trained on boundaries defined by other algorithms and characteristics. Furthermore, the continuous nature of *preciseTAD* predictions may be utilized to quantify boundary strength, as has been done with insulation score and other metrics (Crane *et al.*, 2015; Gong *et al.*, 2018). Our future work will include incorporating the directionality of CTCF binding in predictive modeling, including additional predictor types, developing an algorithm to quantify boundary strength, and defining separate models trained on different boundary types defined by different technologies.

In summary, we demonstrate that domain boundary prediction is a multi-faceted problem requiring consideration of multiple statistical and biological properties of genomic data. Simply considering properties of Hi-C contact matrices ignores the fundamental roles of known molecular drivers of 3D chromatin structures. Instead, we propose *preciseTAD*, a supervised machine learning framework that leverages both Hi-C contact matrix information and genomic annotations. Our method introduces three concepts—*shifted binning*, *distance*-type predictors and *random undersampling*—which we use to build random forest classification models for predicting boundary regions. Our method can bridge the resolution gap between 1D genomic annotations and 3D Hi-C data for more precise and biologically meaningful boundary identification. We introduce *preciseTAD*, an open-source R package for leveraging random forests to predict domain boundaries at base-level resolution, as well as the genomic coordinates of predicted boundaries for 60 cell types. We hope that

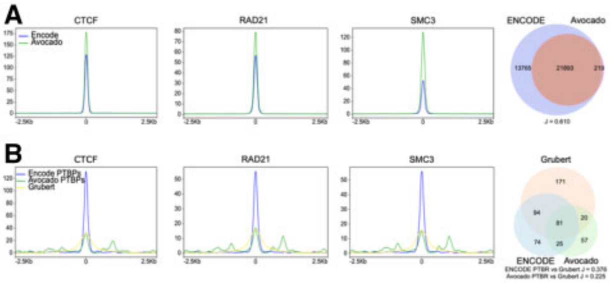

Fig. 6. Avocado-imputed genomic annotations can be used for *preciseTAD* predictions. (**A**) Signal distribution around the midpoint of ENCODE and Avocado TFBSs and the Venn diagram of overlap between ENCODE and Avocado CTCF binding sites. (**B**) Signal distribution around PTBPs predicted from ENCODE and Avocado data and the midpoints of Grubert anchors, and the Venn diagram of overlaps between ENCODE/Avocado-predicted PTBPs and the Grubert data. Data for GM12878 cell line are shown. For the Venn diagrams, regions were flanked by 5 kb

preciseTAD will serve as an efficient and easy-to-use tool to further explore the genome's 3D organization.

Acknowledgements

The authors thank Drs. Jonathan Wren and Cory Giles for helpful comments.

Funding

This work was supported in part by the PhRMA Foundation Research Informatics Award and the George and Lavinia Blick Research Fund scholarship to M.G.D.

Conflict of Interest: none declared.

References

Al Bkhetan,Z. and Plewczynski,D. (2018) Three-dimensional epigenome statistical model: genome-wide chromatin looping prediction. *Sci. Rep.*, **8**, 5217.

Alipour,E. and Marko,J.F. (2012) Self-organization of domain structures by dna-loop-extruding enzymes. *Nucleic Acids Res.*, **40**, 11202–11212.

Ay,F. *et al.* (2014) Statistical confidence estimation for HI-C data reveals regulatory chromatin contacts. *Genome Res.*, **24**, 999–1011.

Bailey,S.D. *et al.* (2015) ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nat. Commun.*, **6**, 6186.

Beagan,J.A. and Phillips-Cremins,J.E. (2020) On the existence and functionality of topologically associating domains. *Nat. Genet.*, **52**, 8–16.

Belokopytova,P. and Fishman,V. (2020) Predicting genome architecture: challenges and solutions. *Front. Genet.*, **11**, 617202.

Boulesteix,A.-L. *et al.* (2012) Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdiscipl. Rev. Data Min. Knowledge Discov.*, **2**, 493–507.

Chang,L.-H. *et al.* (2020) TADs and their borders: free movement or building a wall? *J. Mol. Biol.*, **432**, 643–652.

Chawla,N.V. *et al.* (2002) SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, **16**, 321–357.

Crane,E. *et al.* (2015) Condensin-driven remodelling of x chromosome topology during dosage compensation. *Nature*, **523**, 240–244.

Cresswell,K.G. *et al.* (2020) SpectralTAD: an r package for defining a hierarchy of topologically associated domains using spectral clustering. *BMC Bioinformatics*, **21**, 319.

Davidson,I.F. *et al.* (2019) DNA loop extrusion by human cohesin. *Science*, **366**, 1338–1345.

Dixon,J.R. *et al.* (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376.

Dozmorov,M.G. (2017) Epigenomic annotation-based interpretation of genomic data: from enrichment analysis to machine learning. *Bioinformatics*, **33**, 3323–3330.

Durand,N.C. *et al.* (2016) Juicebox provides a visualization system for hi-c contact maps with unlimited zoom. *Cell Syst.*, **3**, 99–101.

ENCODE Project Consortium. (2012) An integrated encyclopedia of dna elements in the human genome. *Nature*, **489**, 57–74.

Franke,M. *et al.* (2016) Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*, **538**, 265–269.

Fraser,J. *et al.* (2015) Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol. Syst. Biol.*, **11**, 852.

Freire-Pritchett,P. *et al.* (2017) Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem cells. *Elife*, **6**, e21926.

Fudenberg,G. *et al.* (2016) Formation of chromosomal domains by loop extrusion. *Cell Rep.*, **15**, 2038–2049.

Gan,W. *et al.* (2019) A computational method to predict topologically associating domain boundaries combining histone marks and sequence information. *BMC Genomics*, **20**, 980.

Goloborodko,A. *et al.* (2016) Chromosome compaction by active loop extrusion. *Biophys. J.*, **110**, 2162–2168.

Gong,Y. *et al.* (2018) Stratification of tad boundaries reveals preferential insulation of super-enhancers by strong boundaries. *Nat. Commun.*, **9**, 542.

Grubert,F. *et al.* (2020) Landscape of cohesin-mediated chromatin loops in the human genome. *Nature*, **583**, 737–743.

Hahsler,M. *et al.* (2019) Dbscan: fast density-based clustering with r. *J. Stat. Softw.*, **25**, 409–416.

Hansen,A.S. *et al.* (2018) Recent evidence that tads and chromatin loops are dynamic structures. *Nucleus*, **9**, 20–32.

Harrold,C.L. *et al.* (2020) A functional overlap between actively transcribed genes and chromatin boundary elements. *bioRxiv*, 182089. https://doi.org/10.1101/2020.07.01.182089.

Hnisz,D. *et al.* (2016) Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, **351**, 1454–1458.

Jin,F. *et al.* (2013) A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, **503**, 290–294.

Kai,Y. *et al.* (2018) Predicting ctcf-mediated chromatin interactions by integrating genomic and epigenomic features. *Nat. Commun.*, **9**, 4221.

Krietenstein,N. *et al.* (2020) Ultrastructural details of mammalian chromosome architecture. *Mol. Cell*, **78**, 554–565.e7.

Kuhn,M. (2012) *The Caret Package*. R Foundation for Statistical Computing, Vienna, Austria. https://cran. r-project. org/package=caret (2 August 2021, date last accessed).

Lieberman-Aiden,E. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, **326**, 289–293.

Lupianez,D.G. *et al.* (2016) Breaking tads: how alterations of chromatin domains result in disease. *Trends Genet.*, **32**, 225–237.

Mirny,L.A. *et al.* (2019) Two major mechanisms of chromosome organization. *Curr. Opin. Cell Biol.*, **58**, 142–152.

Nanni,L. *et al.* (2020) Spatial patterns of ctcf sites define the anatomy of tads and their boundaries. *Genome Biol.*, **21**, 197.

Naumova,N. *et al.* (2013) Organization of the mitotic chromosome. *Science*, **342**, 948–953.

Nora,E.P. *et al.* (2012) Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature*, **485**, 381.

Phillips-Cremins,J.E. *et al.* (2013) Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, **153**, 1281–1295.

Ramirez,F. *et al.* (2016) DeepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.*, **44**, W160–W165.

Rao,S.S. *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.

Salameh,T.J. *et al.* (2019) A supervised learning framework for chromatin loop detection in genome-wide contact maps. *bioRxiv*, 739698. https://10.1101/739698.

Sanborn,A.L. *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–E6465.

Sauerwald,N. and Kingsford,C. (2021) Capturing the complexity of topologically associating domains through multi-feature optimization. *bioRxiv*, 425264. https://10.1101/2021.01.04.425264.

Schmitt,A.D. *et al.* (2016) Genome-wide mapping and analysis of chromosome architecture. *Nat. Rev. Mol. Cell Biol.*, **17**, 743–755.

Schreiber,J. *et al.* (2020) Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome. *Genome Biol.*, **21**, 81.

Schreiber,J. *et al.* (2019) A pitfall for machine learning methods aiming to predict across cell types. *bioRxiv*, 512434. https://10.1101/512434.

Sefer,E. and Kingsford,C. (2015) Semi-nonparametric modeling of topological domain formation from epigenetic data. Algorithms for Molecular Biology, 14, 1–11.

Sexton,T. *et al.* (2012) Three-dimensional folding and functional organization principles of the drosophila genome. *Cell*, **148**, 458–472.

Sun,F. *et al.* (2019) Promoter-enhancer communication occurs primarily within insulated neighborhoods. *Mol. Cell*, **73**, 250–263.e5.

Sun,J.H. *et al.* (2018) Disease-associated short tandem repeats co-localize with chromatin domain boundaries. *Cell*, **175**, 224–238.e15.

Taberlay,P.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.

Tang,Z. *et al.* (2015) CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell*, **163**, 1611–1627.

Tao,H. *et al.* (2021) Computational methods for the prediction of chromatin interaction and organization using sequence and epigenomic profiles. *Brief. Bioinform.*

Thiecke,M.J. *et al.* (2020) Cohesin-dependent and -independent mechanisms mediate chromosomal contacts between promoters and enhancers. *Cell Rep.*, **32**, 107929.

Wang,Y. *et al.* (2021) TAD boundary and strength prediction by integrating sequence and epigenetic profile information. *Brief. Bioinform.*

Wei,Q. and Dunbrack,R.L. Jr, (2013) The role of balanced training and testing data sets for binary classifiers in bioinformatics. *PLoS One*, **8**, e67863.

Weinreb,C. and Raphael,B.J. (2016) Identification of hierarchical chromatin domains. *Bioinformatics*, **32**, 1601–1609.

Wen,Z. *et al.* (2018) ZNF143 is a regulator of chromatin loop. *Cell Biol. Toxicol.*, **34**, 471–478.

Zhang,S. *et al.* (2019) In silico prediction of high-resolution hi-c interaction matrices. *Nat. Commun.*, **10**, 5449.

Zufferey,M. *et al.* (2018) Comparison of computational methods for the identification of topologically associating domains. *Genome Biol.*, **19**, 217.

Zuin,J. *et al.* (2014) Cohesin and ctcf differentially affect chromatin architecture and gene expression in human cells. *Proc. Natl. Acad. Sci. USA*, **111**, 996–1001.