preciseTAD: A machine learning framework for precise 3D domain boundary prediction at base-level resolution

Spiro C. Stilianoudakis[1] (stilianoudasc@mymail.vcu.edu), Mikhail G. Dozmorov[1*] (mikhail.dozmorov@vcuhealth.org)

[1] Dept. of Biostatistics, Virginia Commonwealth University, Richmond, VA, 23298, USA
[*] To whom correspondence should be addressed

Abstract

The low resolution of high-throughput chromatin conformation capture data limits the precise mapping of boundaries of topologically associating domains and chromatin loops. We developed preciseTAD, an optimized random forest model trained on high-resolution genome annotation data (e.g., CTCF ChIP-seq) to predict the location of domain boundaries at base-level resolution. Distance between boundaries and annotations, random under-sampling, and transcription factor binding sites resulted in best model performance. preciseTAD boundaries were more enriched for CTCF, RAD21, SMC3, and ZNF143, and conserved across cell lines. Using genome annotations, pre-trained models can detect boundaries in cells without Hi-C data. preciseTAD is available at https://bioconductor.org/packages/preciseTAD

Keywords

Hi-C, TAD, machine learning, random forest, DBSCAN

Background

The advent of chromosome conformation capture (3C) sequencing technologies, and its successor Hi-C, have revealed a hierarchy of the 3-dimensional (3D) structure of the human genome such as chromatin loops [1], Topologically Associating Domains (TADs) [2,3], and A/B compartments [4]. At the kilobase scale, chromatin loops connect gene promoters with distal enhancers, thereby regulating gene expression [5,6]. At the megabase scale, TADs represent regions on the linear genome that are highly self-interacting. Disruption of boundaries demarcating TADs and loops promotes cancer [7,8] and other disorders [9–11]. Identifying the precise location of TAD and chromatin loop boundaries remains a top priority in our goal to fully understand the functionality of the human genome.

Several methods have been proposed to identify genomic coordinates of TADs (reviewed in [12,13]), and chromatin loops [1,14–16]. However, a key limitation of them is that they depend on Hi-C data *resolution*. Resolution refers to the size of genomic regions (bins) used to segment the linear genome [17]. Lower resolutions contribute to larger bin sizes, which leads to increased uncertainty in the location of boundaries. Due to high sequencing costs, most available Hi-C datasets have relatively low resolution, ranging from 25 kb to 100 kb. While these low-resolution Hi-C datasets can be used to probe higher-order genomic patterns such as A/B compartments, more refined resolutions are necessary to accurately define TAD and loop boundaries [18]. Because increasing the resolution of Hi-C data requires a quadratic increase in the total sequencing depth [19], obtaining high-resolution HI-C data and boundaries remains difficult.

Another limitation among domain callers is that they disregard prior knowledge about functional genomic annotations associated with TAD and loop boundaries, referred hereafter as domain boundaries. The insulator binding protein, CTCF, and additional cofactors such as SMC3 and RAD21 have been identified as components of the loop extrusion model [18,20–24]. Furthermore, distinct patterns of histone modifications have also been shown to be present at boundaries [3,4,25]. These genomic annotations are obtained using chromatin immunoprecipitation followed by high-throughput DNA sequencing (ChIP-seq). The resolution of ChIP-seq experiments is typically on the order of tens to hundreds of bases [26], well below the resolution of Hi-C data (tens of kilobases; 750 bp is the highest resolution of Hi-C data to date [27]). Therefore, leveraging precisely mapped genomic annotations in a supervised machine learning framework enables more precise prediction of the locations of domain boundaries.

We present *preciseTAD*, an optimally tuned machine learning framework for precise identification of domain boundaries using genome annotation data. Our method utilizes the random forest (RF) algorithm trained on high-resolution chromatin state (BroadHMM), histone modification (HM), and transcription factor binding site (TFBS) data to predict low-resolution boundaries. We introduce a systematic pipeline for building the optimal boundary region prediction classifier. We found that spatial associations (linear distance) between boundaries and annotations perform best, transcription factor binding sites improve prediction performance, and a simple random undersampling technique effectively addresses the negative effect of class imbalance. Translated from Hi-C data resolution level to base level (annotating each base and predicting its boundary probability), *preciseTAD* employs density-based clustering (DBSCAN) and partitioning around medoids (PAM) to detect genome annotation-guided boundary regions and points at a base-level resolution. This approach circumvents resolution restrictions of Hi-C data, allowing for the precise detection of biologically meaningful boundaries. We demonstrate that *preciseTAD* predictions are more enriched for known molecular drivers of 3D chromatin, including CTCF, RAD21, SMC3, and ZNF143. Further, we show that *preciseTAD*-predicted boundaries are more conserved across cell lines, and the model trained in one cell line can predict boundaries in another cell line. This improved precision in the domain boundary location can provide insight into the association between genomic regulators and the 3D genome organization. The *preciseTAD* R package is freely available on https://bioconductor.org/packages/preciseTAD.

Results

Developing an ML framework for optimal TAD 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. We chose the random forest (RF) algorithm as our binary classification tool. The reason for it is two-fold: (1) 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 (2) to allow for an interpretable ranking of predictors [28]. We used Arrowhead-called TAD boundaries [29] and published Peakachu-predicted loop boundaries [16] as ground truth. Data from GM12878 and K562 cell lines at 5-100 kb resolution (Arrowhead) and 10 kb resolution (Peakachu) were used (Additional file 1: Arrowhead Script, Additional file 2: Table S1). We next developed *preciseTAD*, an R package that leverages an optimally built RF to precisely predict genomic coordinates of TAD and loop boundaries at base-level resolution from cell line-specific genome annotation data.

Boundary regions were defined as genomic bins containing a called boundary ($Y = 1$), while non-boundary regions were defined as bins that did not contain a called boundary ($Y = 0$) (Figure 1A, see Methods). 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, Additional file 3: 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 [30,31]. To address the class imbalance, we evaluated the effect of three resampling techniques. *Random over-sampling* (ROS) was defined as sampling with replacement from the minority class (boundary regions). *Random*

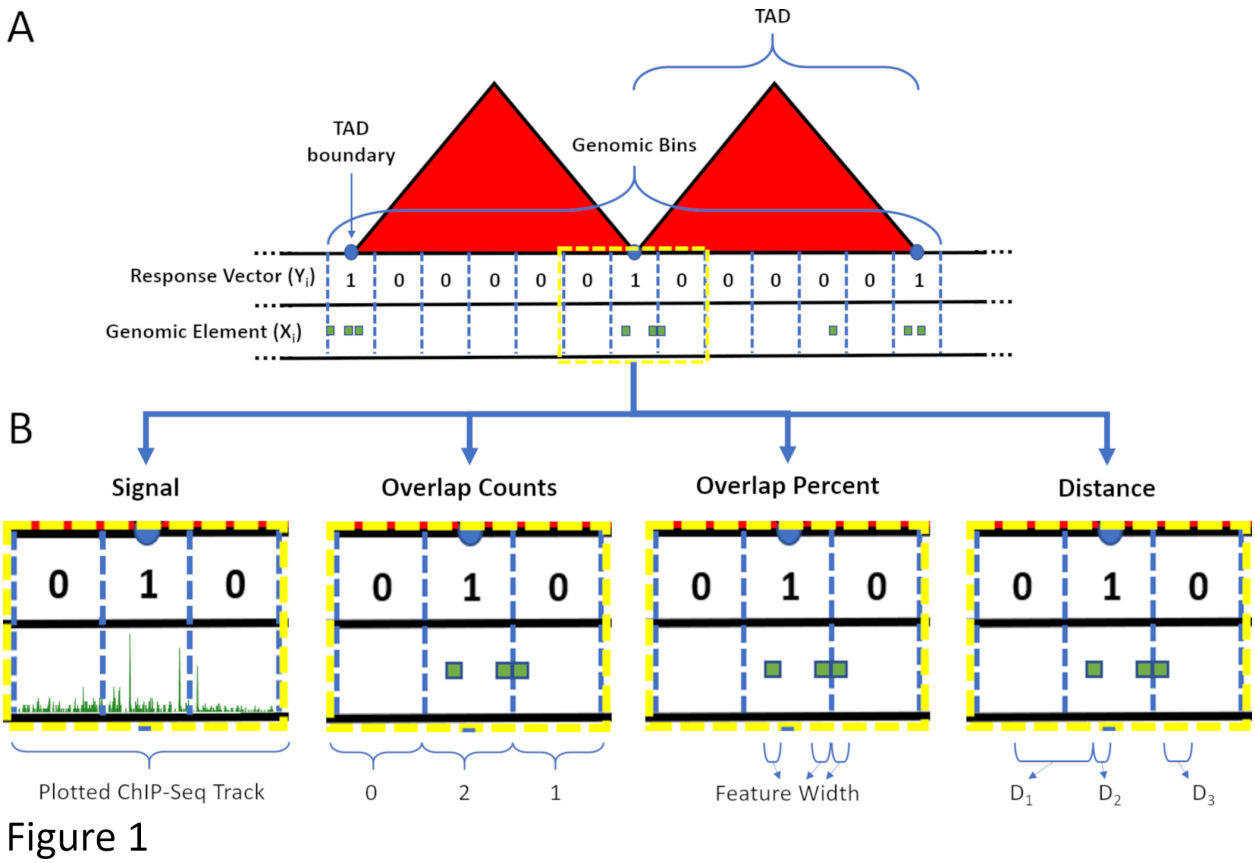

Figure 1: **Resolution-specific data construction and feature engineering for random forest modeling.** (A) The linear genome was binned into non-overlapping resolution-specific intervals using *shifted binning* (see Methods). The response vector **Y** was defined as 1/0 if a genomic bin overlapped/did not overlap with a TAD (or loop) boundary. (B) Four types of associations between bins (blue dashed lines) and genomic annotations (green shapes) were considered to build the predictor space, including Average Peak Signal (Signal), Overlap Counts (OC), Overlap Percent (OP), and log_2 distance (Distance).

under-sampling (RUS) was defined as sampling with replacement from the majority class (non-boundary regions). Lastly, we tested *Synthetic minority over-sampling technique* (SMOTE), which is a combination of both random over- and under-sampling to create balanced classes [32] (see Methods).

Table 1: Domain boundary data and class imbalance summaries across resolutions for Arrowhead and Peakachu on GM12878.

Tool	Resolution	Total number of called TADs/chromatin loops	Total number of unique TAD/chromatin loop boundaries	Total number of genomic bins	Class imbalance
Arrowhead	5 kb	8052	15468	535363	0.03
Arrowhead	10 kb	7676	14253	267682	0.05
Arrowhead	25 kb	4670	8363	107073	0.08
Arrowhead	50 kb	2349	4224	53537	0.08
Arrowhead	100 kb	1031	1883	26768	0.07
Peakachu	10 kb	16185	21421	267682	0.14

A total of 77 cell line-specific genomic annotations were used to build the predictor space. These included BroadHMM chromatin state data, histone modifications (HM), and transcription factor binding sites (TFBS) (Additional file 4: Table S3). Four feature engineering procedures were developed to quantify the association between genomic annotations and bins (Figure 1B). These included signal strength association (Signal), direct (OC), proportional (OP), and spatial (log_2+1 Distance) relationships (see Methods, Additional file 5: Figure S1).

In total, we considered combinations of data from two cell lines L = *{GM12878, K562}*, five resolution 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}* (Figure 2). Once the model inputs were established, a random forest classifier was trained on $n-1$ autosomal chromosomes, while reserving the ith chromosome for testing. Three-fold 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. Model performance was evaluated by aggregating the mean balanced accuracy (BA) across each holdout chromosome (see Methods). These strategies allowed us to select the best performing model in an unbiased manner.

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

Expectedly, when using data with class imbalance present, that is, no resampling, the models exhibited low balanced accuracies, with minimal variability among different resolutions (Figure 3). Similarly, poor performances were found when using ROS. However, RUS and SMOTE re-sampling led to a drastic improvement in performance, especially at higher resolutions. We found that RUS marginally outperformed SMOTE as the optimal class balancing technique for all resolutions and predictor types when predicting TAD boundary regions.

Additionally, we found that using a 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 (Additional file 6: Figure S2A). Furthermore, 5 kb resolution genomic bins led to the optimal prediction for TAD boundary regions on both cell lines. Random forest models built on Peakachu-defined loop boundary regions yielded optimal prediction performance when using a distance-type predictor space, with SMOTE resampling only marginally outperforming RUS, for both cell lines (Additional file 6: Figure S2B, S2C). Our results indicate that random under-sampling, distance-type predictors, and high-resolution Hi-C data provide the optimal set of data level characteristics for both TAD and chromatin loop boundary prediction.

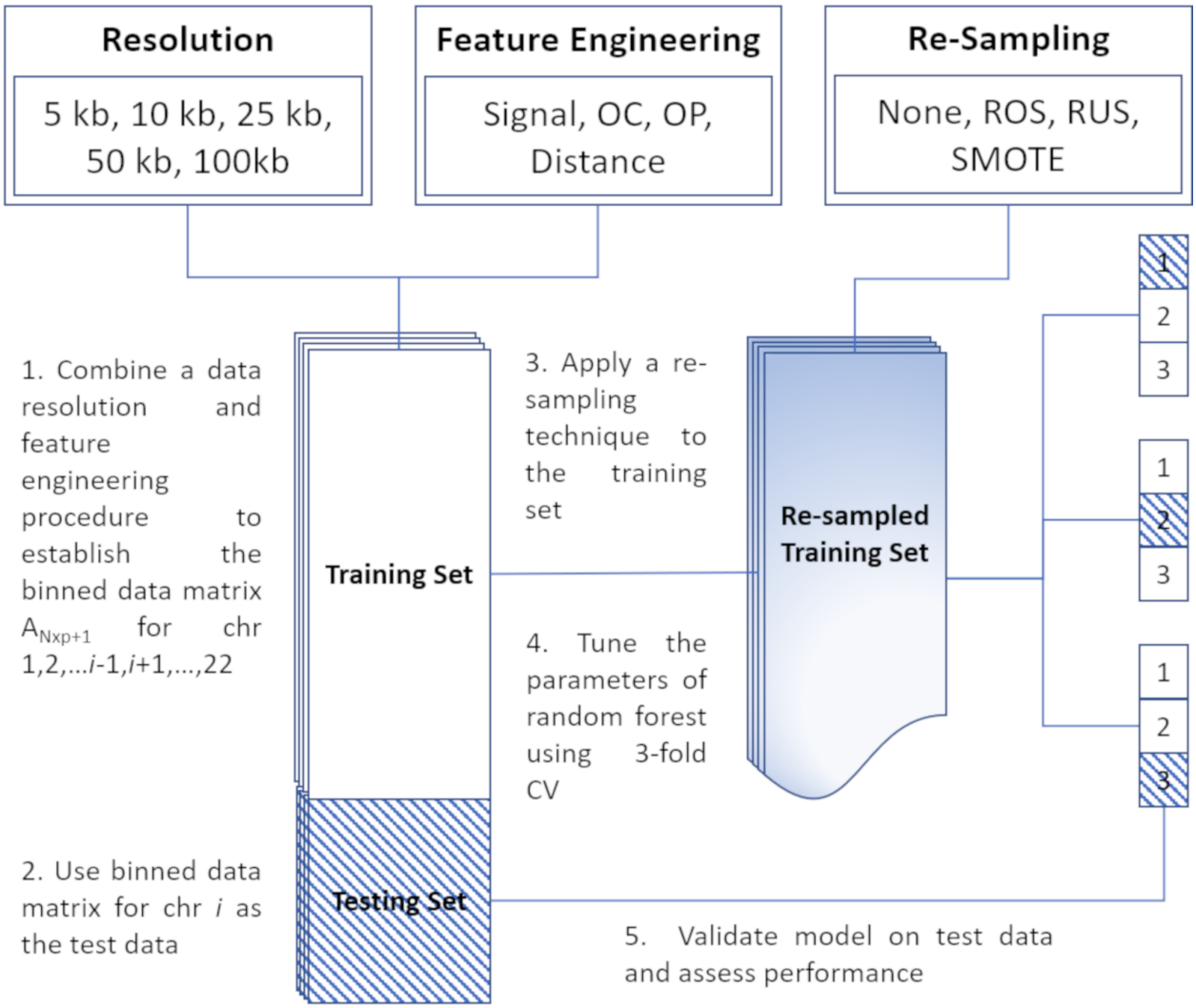

Figure 2

Figure 2: **A machine learning framework for building domain boundary region prediction models.** Step 1 employs a range of feature engineering techniques to define the predictor matrix $A_{N\times(p)}$, where N is the number of genomic bins, p is the number of genomic annotations, i is a holdout chromosome. The response vector Y_N is defined as a boundary region ($Y = 1$) if it overlaps with a genomic bin (else $Y = 0$). Step 2 reserves the predictor-response matrix for the holdout chromosome i as the test data. Step 3 applies a resampling technique to the training data to address the class imbalance. Step 4 trains the random forest model and performs 3-fold cross-validation to tune the mtry parameter. Finally, step 5 validates the model on the separate test data composed of the binned data from the holdout chromosome i and evaluates model performance using balanced accuracy (BA).

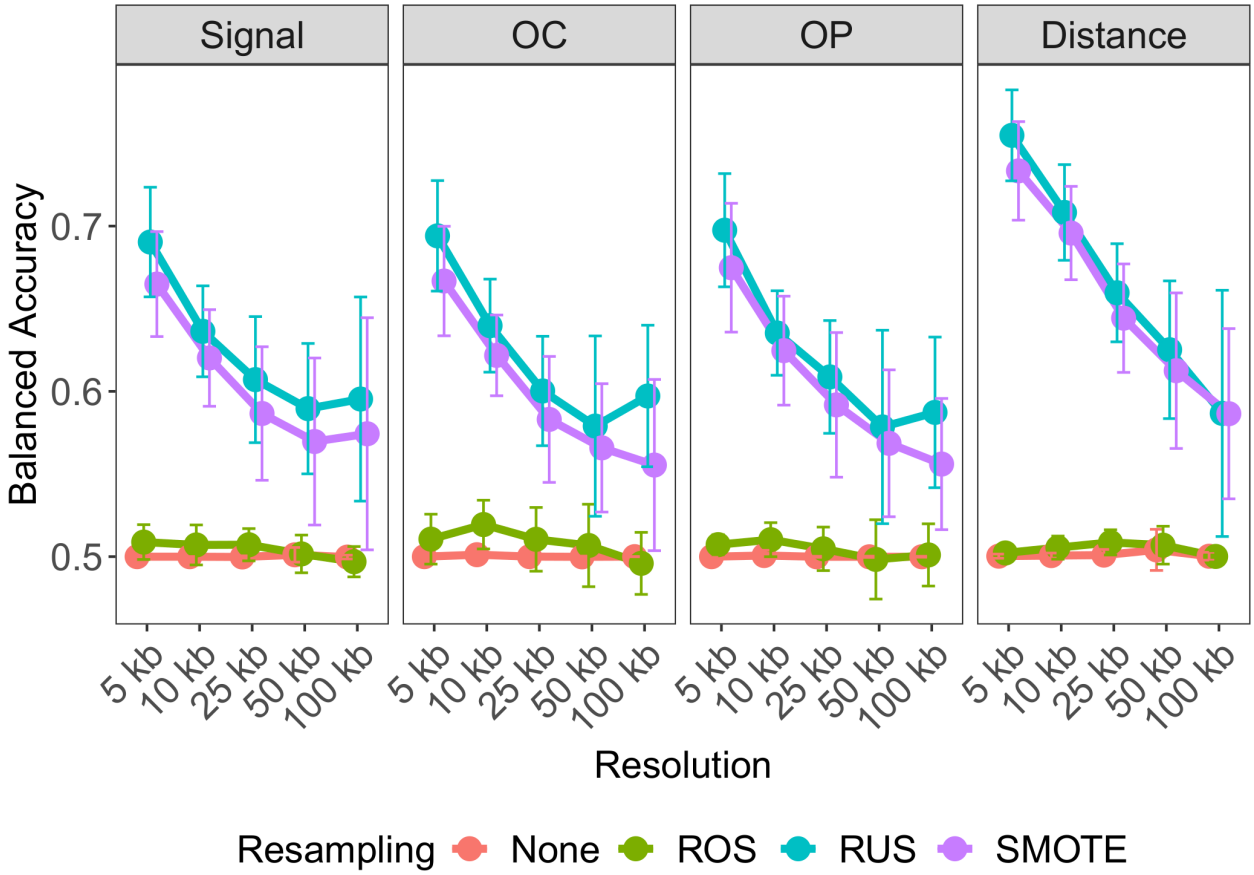

Figure 3

Figure 3: **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: Signal, OC, OP, 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 standard deviation from the mean performance across each holdout chromosome used for testing.

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

We hypothesized that the type of genomic annotations may also affect predictive performance. Using the established optimal settings (RUS, Distance, 5 kb/10 kb (Arrowhead/Peakachu ground truth) genomic bins), we built separate random forest models using histone modifications (HM), BroadHMM chromatin states (BroadHMM), and transcription factor binding sites (TFBS). We found that models built on TFBSs outperformed other annotation-specific models, with results consistent for loop boundaries, on both cell lines (Figure 4A; Additional file 7: Figure S3A). These results suggest that TFBS are the primary drivers of TAD and loop boundary formation in both GM12878 and K562.

Predictive importances confirmed the biological role of CTCF, RAD21, SMC3, and ZNF143 for boundary formation

We sought to further optimize our boundary region prediction models by implementing recursive feature elimination to avoid overfitting and selecting only the most influential TFBS across all autosomal chromosomes. We were able to obtain near-optimal performance using approximately eight TFBS (Figure 4B; Additional file 7: Figure S3B). However, given that we trained our models on chromosome-specific data, the most significant annotations varied for each holdout chromosome. To determine transcription factors most important for boundary prediction across all chromosomes, we clustered the predictive importance (mean decrease in accuracy) of the top eight significant TFs across chromosomes. We found four transcription factors, CTCF, RAD21, SMC3, and ZNF143, being consistently predictive of TAD and loop boundaries (Figure 4C; Additional file 7: Figure S3C). We optimized our model by only considering these top four TFBS when building the random forest model, thereby decreasing computational burden while maintaining high predictive performance. In summary, our model was able to yield the known molecular drivers of the loop extrusion model [18,20–24].

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

Having demonstrated the optimal performance in our framework's ability to predict boundary regions, we next attempted to predict boundary regions in one cell line using the model pre-trained on data from another cell line (See Methods). We found that training and testing using Arrowhead ground truth TAD boundaries and genomic annotation data from the GM12878 cell line resulted in an average AUC=0.792 (Figure 4D). Interestingly, when training on the K562 cell line and testing on GM12878, the average AUC increased slightly to 0.795. Likewise, the average performance of models trained using Peakachu boundaries and genomic annotation data from the GM12878 cell line was comparable to models trained on K562-specific Peakachu boundaries and genomic annotations (Avg. AUC=0.881 and 0.874, respectively). These results were consistent when comparing training/testing strategies on K562 with training on GM12878 and testing on K562 data (Additional file 7: Figure S3D). 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, allowing for the opportunity to predict boundaries for cell lines that do not currently have Hi-C data available.

preciseTAD identifies precise and biologically relevant domain boundaries

Using our optimally built random forest model trained on Arrowhead/Peakachu boundaries, we attempted to predict 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 associated with a genomic annotation. 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)

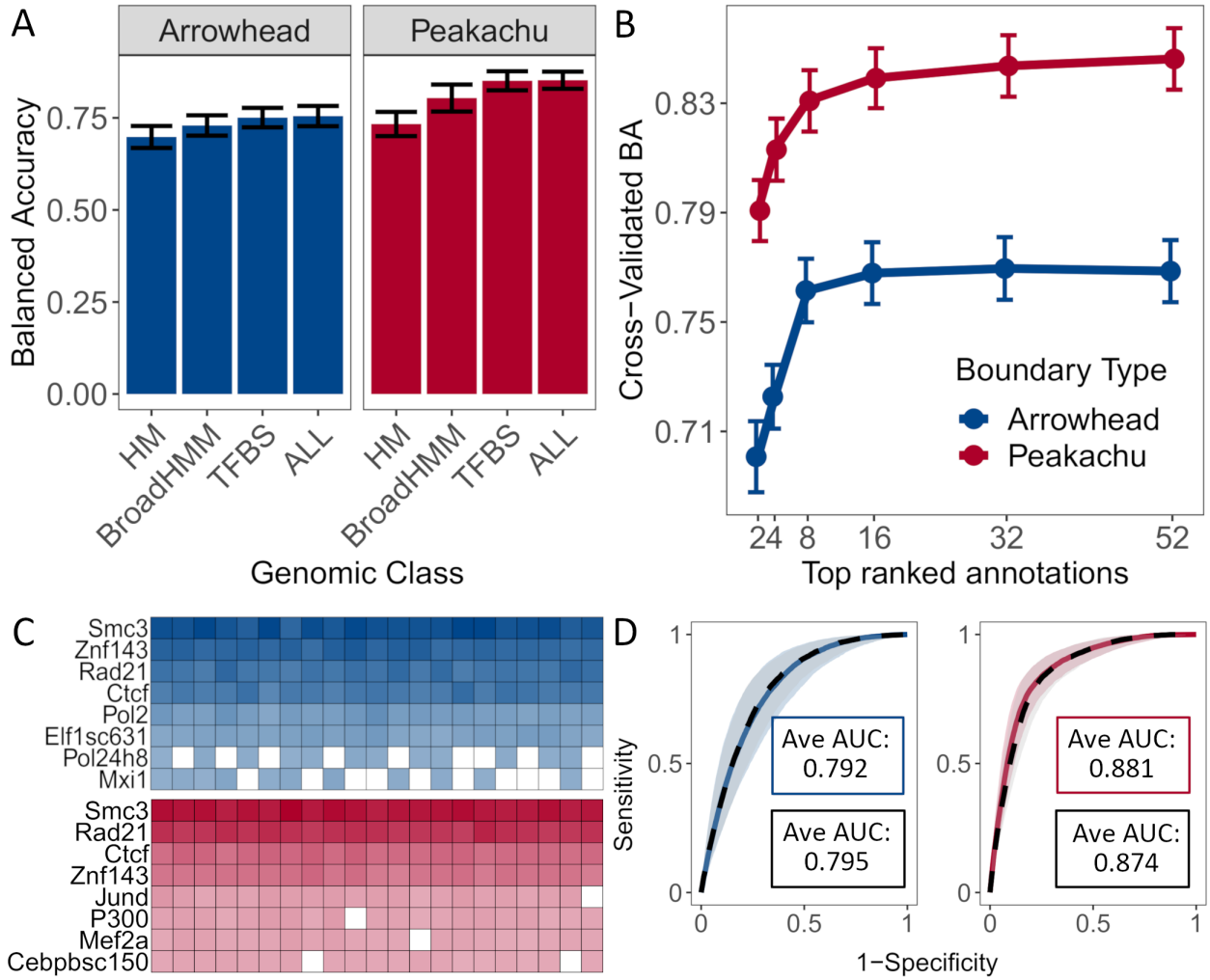

Figure 4

Figure 4: **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) Receiver operating characteristic (ROC) curves and the corresponding average area under the curves (AUCs) when training and testing on GM12878 data (blue, Arrowhead ground truth; red, Peakachu ground truth) 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.

and partitioning around medoids (PAM) to predict the location of boundary regions and summit points (highest boundary probability) at base-level resolution (see Methods; Additional file 8: Figure S4).

When trained using Arrowhead and Peakachu ground truth boundaries at 5 kb and 10 kb, respectively, the *preciseTAD* model predicted a total of 12,258 TAD and 15,707 chromatin loop boundaries in GM12878, as well as 9,603 TAD and 11,154 chromatin loop boundaries in K562 cell line (Additional file 9: Table S4). We reported less predicted TAD boundaries at 5 kb than Arrowhead on both cell lines (Table 1, Additional file 3: Table S2). This can be attributed to Arrowhead's inflation of called TADs at 5 kb, that, when visualized, often do not correspond to domains enriched in internal interactions (Figure 5A) and signal of known drivers of domain boundaries (Figure 5B). *preciseTAD* also predicted fewer chromatin loop boundaries than Peakachu (Table 1, Additional file 3: Table S2). This can be attributed to Peakachu's use of only CTCF sites to call boundaries, while *preciseTAD* leverages four known drivers of 3D chromatin, including CTCF, RAD21, SMC3, and ZNF143. These results suggest that genome annotation-driven predicted boundaries may be more biologically relevant than those detected from Hi-C data alone.

Next, we evaluated the biological significance of *preciseTAD* boundary points. The signal of four known molecular drivers of 3D chromatin (CTCF, RAD21, SMC3, and ZNF143) colocalized more frequently around *preciseTAD*-predicted boundary points, as compared to Arrowhead-called TAD and Peakachu loop boundaries, respectively (Figure 5C, 5D; Additional file 10: Figure S5A, S5B). Surprised by the poor signal distribution around Arrowhead boundaries, we compared signals centered on boundaries called by Arrowhead, Peakachu, and our SpectralTAD [33] TAD caller. We confirmed the poor signal distribution around Arrowhead boundaries, in contrast to the relatively well-performing Peakachu- and SpectralTAD-called boundaries (Additional file 10: Figure S5C, S5D). Signal enrichment heatmaps confirmed that *preciseTAD*-predicted boundaries were more enriched for the same genomic annotations than either Arrowhead or Peakachu boundaries alone (Additional file 11-12: Figure S6-S7). *preciseTAD* boundaries were statistically significantly closer to the top-ranked TFBS (Wilcoxon p-value < 0.001 versus Arrowhead and Peakachu boundaries, respectively, Figure 6A, Additional file 13: Figure S8). These results indicate that *preciseTAD*-predicted boundaries better reflected the known biology of boundary formation.

preciseTAD boundaries are more conserved across cell lines

Previous studies suggest that TAD boundaries are conserved across cell lines [2,3,34,35]. To assess the level of cross-cell-line conservation, we evaluated the overlap between cell line-specific boundaries detected by *preciseTAD*, Arrowhead, and Peakachu. Only 26% and 49% of boundaries were conserved between cell lines for Arrowhead and Peakachu boundaries (J=0.186 and J=0.388), respectively (Figure 6B, 6D). However, 45%/56% of *preciseTAD*-predicted domain boundaries were conserved between GM12878 and K562 cell lines when using models trained on Arrowhead/Peakachu data (J=0.383 and J=0.444, respectively, Figure 6C, 6E). The better conservation of *preciseTAD*-predicted boundaries further supports the notion of their higher biological relevance.

Boundaries predicted by *preciseTAD* models trained on TAD and loop boundaries are highly overlapping

The majority of boundaries predicted by the Arrowhead-trained *preciseTAD* model represented a subset of the larger group of boundaries predicted by the Peakachu-trained model (Additional file 14: Figure S9). This is expected as loop boundaries detected by Peakachu are more abundant, while comparatively wide TAD boundaries detected by Arrowhead likely represent the higher level of the 3D chromatin organization. Together with the results in Figure 4C, the high overlap between boundaries predicted by Arrowhead- and Peakachu-trained models suggests that TAD and loop boundaries may be driven by similar molecular mechanisms.

Guided by the high predictive performance when training and testing on different cell lines (Figure 4D), we opted to evaluate whether models trained using Arrowhead/Peakachu ground truth data in one cell line could be leveraged to predict boundaries using annotation data from another cell line. We evaluated two

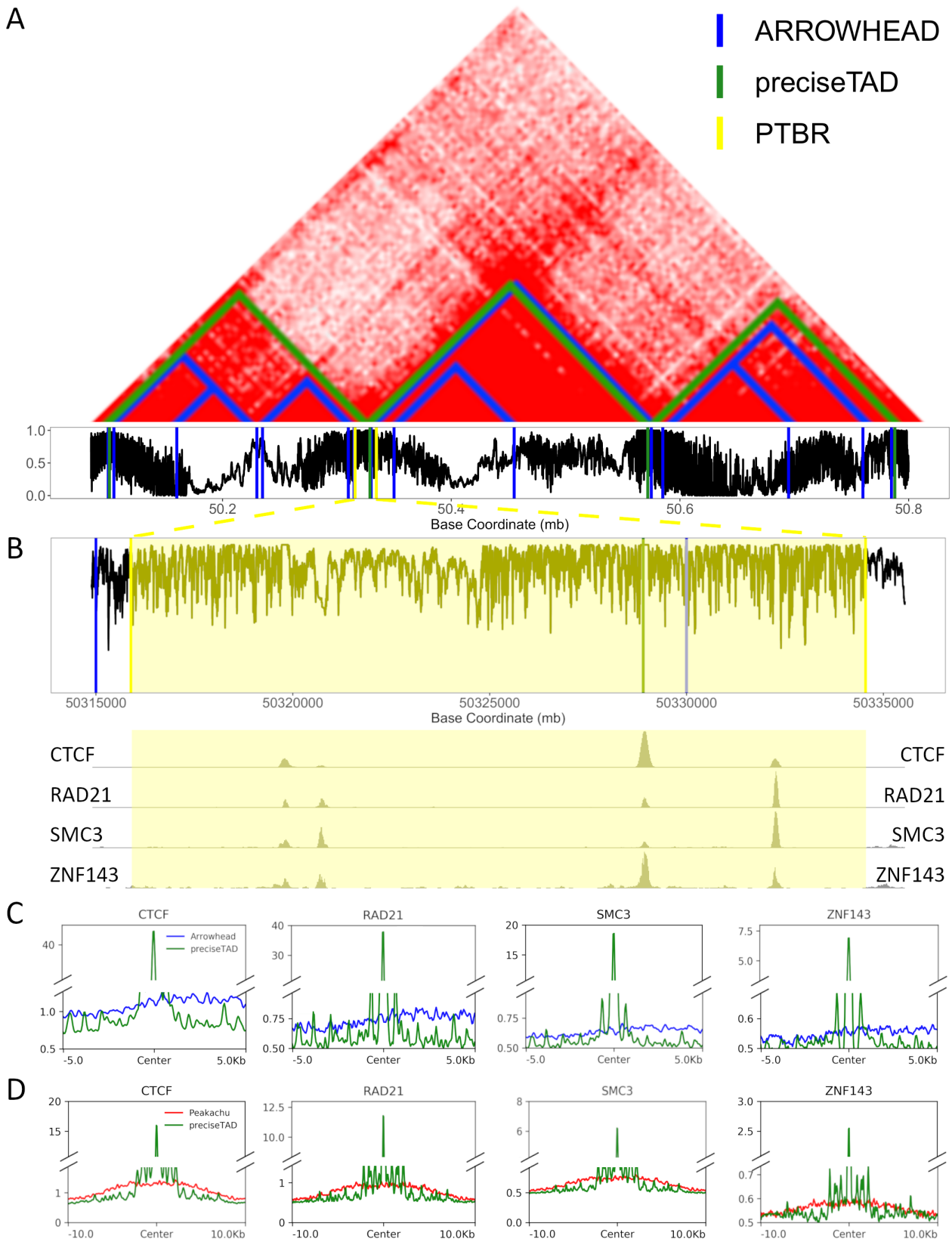

Figure 5

Figure 5: **preciseTAD boundaries are enriched for known molecular drivers of 3D chromatin.** (A) The location of Arrowhead-called TAD boundaries (blue) vs. *preciseTAD*-predicted TAD boundaries (green) on GM12878 data (chr14:50085000-50800000). The black line represents the predicted probability of each base being a TAD boundary. (B) A zoomed-in portion of the genome shows the preciseTAD boundary region (PTBR, highlighted yellow), a cluster of bases with high probability of being a boundary, and the

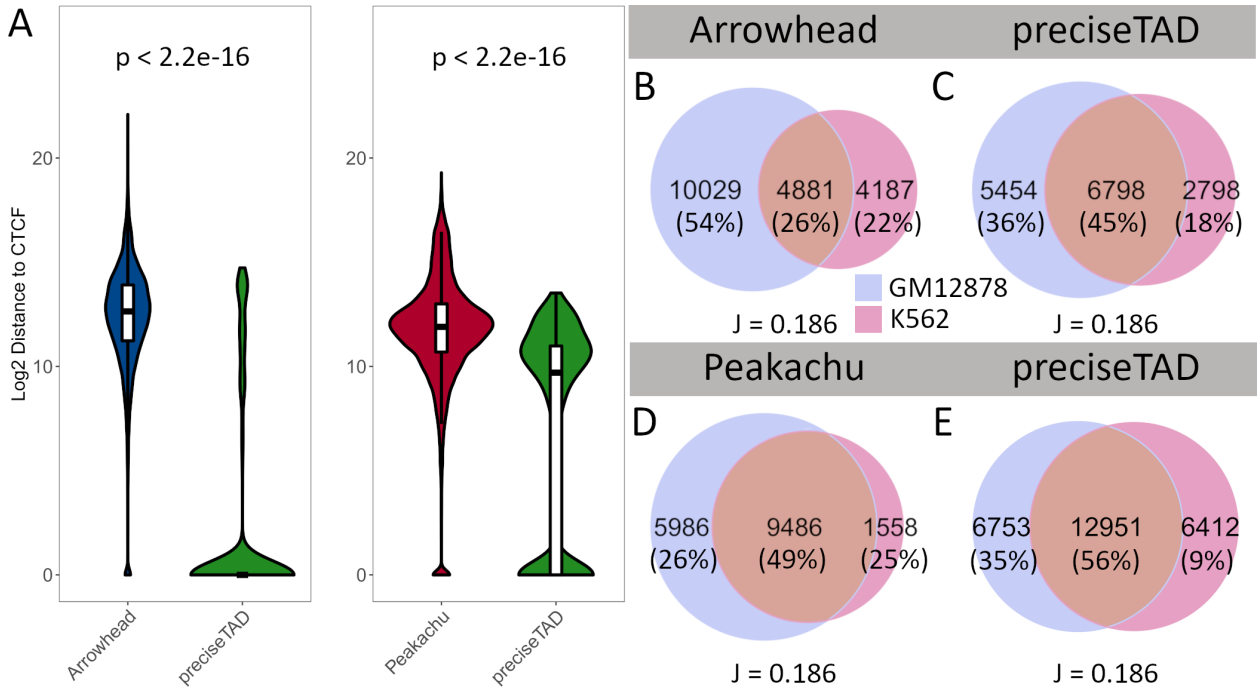

Figure 6

Figure 6: **preciseTAD-predicted boundaries are closer to CTCF sites and more conserved across cell lines**. (A) log_2 genomic distance distribution from called and predicted boundaries to the nearest CTCF sites. The p-values are from the Wilcoxon Rank Sum test. (B-E) Venn diagrams illustrating the levels of conservation (overlap) between domain boundaries for GM12878 (red) and K562 (blue) cell lines identified by Arrowhead (B), Peakachu (D), and *preciseTAD*-predicted boundaries using (C) Arrowhead- and (E) Peakachu-trained models. Boundaries involving Arrowhead/Peakachu were flanked by 5 kb/10 kb, respectively.

scenarios: 1) training on GM12878 and predicting boundaries on GM12878 (GM on GM) vs. training on K562 and predicting on GM12878 (K on GM), and 2) training on K562 and predicting boundaries on K562 (K on K) vs. 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 (Additional file 15: Figure S10A, S10B). Likewise, when using Peakachu-trained models, we observed 85% (J=0.705) and 88% (J=0.759) overlap (Additional file 15: Figure S10C, S10D). Furthermore, boundaries predicted on 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 (Additional file 15: Figure S10). These results indicate that *preciseTAD* pre-trained models can be successfully used to predict domain boundaries for cell lines lacking Hi-C data but for which genome annotation data is available.

Discussion

We present *preciseTAD*, a machine learning approach for precise prediction of TAD or loop boundaries from functional genomic annotations. *preciseTAD* leverages a random forest (RF) classification model built on low-resolution domain boundaries obtained from domain calling tools, 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. *preciseTAD* was benchmarked against two boundary calling tools, Arrowhead [36], an established TAD-caller, and Peakachu [16] a recently published algorithm for predicting chromatin loops. We first optimized our RF model by systematically comparing different combinations of genome binning (resolution), feature engineering procedures, and resampling techniques. Our results demonstrate that genomic 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. *preciseTAD* is implemented as an R package, provides the pre-trained models, and allows for precise prediction of boundaries for any cell line for which the genomic annotations are available.

Our machine learning framework yielded several interesting observations. We first demonstrated that RF models built using *distance*-type predictors outperformed models built on previously published feature engineering techniques, including signal strength, overlap counts, and overlap percents [37–40]. We further demonstrated that class imbalance hinders boundary prediction, but can be effectively addressed by a simple random under-sampling technique, an aspect of boundary prediction unaddressed in previous studies [37–39]. We showed that information about only four transcription factors (CTCF, SMC3, RAD21, ZNF143) is necessary and sufficient for accurate boundary prediction, outperforming histone modification- and BroadHMM-built models [38,39]. These are known components of the loop extrusion model, an established theory of how loops are made by a ring-shaped adenosine triphosphatase-driven complex called cohesin [18,20–24]. Interestingly, the same transcription factors accurately predicted both TAD and loop boundaries, suggesting a similarity of the mechanisms of TAD and loop formation. This suggested that the random forest model, when tuned and feature engineered correctly, is highly effective in predicting biologically relevant domain boundaries.

Besides balanced accuracy (BA), we investigated five other performance metrics, including accuracy, area under the receiver operating characteristic curve (AUROC), precision, F1-score, and area under the precision-recall curve (AUPRC) (Additional file 16: Table S5). Our aim was to have a balanced metric sensitive to class imbalance such that it would not favor one component of the confusion matrix. The accuracy metric can be artificially inflated by true negatives (TN), the set of genomic bins correctly predicted as not containing a ground truth boundary. AUROC captures how a model generally performs across different thresholds. However, it doesn't emphasize one class over the other, so it does not reflect the minority class well. Precision indicates the rate at which positive predictions are correct and can be artificially deflated by low proportions of true positives (TP), the set of genomic bins correctly predicted as containing a ground truth boundary. While F1-score is a composite metric, it can be susceptible to different values for precision and recall. Lastly, AUPRC is insensitive toward class imbalance, preventing us from investigating its effect, and also omits from its calculation TN values. All of these are important considerations to make when choosing a performance metric. For these reasons, we opted to report balanced accuracy (BA). The BA benefits from incorporating all

components of the confusion matrix, while also being sensitive to class imbalance, a necessary characteristic when comparing performances to models built using no data resampling.

Guided by both low-resolution Hi-C data and high-resolution genome annotation data, *preciseTAD* predicts base-level resolution boundaries, alleviating resolution limitations of Hi-C data. However, a natural question is how resolution (width) of boundary regions identified by *preciseTAD* (PTBRs) compares with that of Hi-C data. Our preliminary observations indicate that, under most optimal settings, the width of PTBRs parallels the resolution of Hi-C data. Furthermore, each PTBR is formed by several sub-regions with the probability of being a boundary defined by the threshold t ($t = 1$ in the current study). Yet, the preciseTAD boundary points (PTBPs, medoids identified within each PTBR) had the highest density of CTCF and other transcription factor binding signal (Figure 5). Our results are in line with the emergent view that domain boundaries are dynamic [41], and their well-defined location arises as a consequence of the population average in bulk Hi-C data [18,42].

We show that, unlike Arrowhead, *preciseTAD* does not inflate the number of predicted boundaries, providing only the most biologically meaningful boundaries that demarcate regions of high inter-chromosomal interactions. *preciseTAD* boundaries predicted using models trained on either TAD or loop boundaries (Arrowhead and Peakachu data) were enriched for known architectural transcription factors including CTCF, RAD21, SMC3, and ZNF143, supporting recent observations that TADs and loops may be generated by similar mechanisms [20]. Likewise, *preciseTAD* boundaries were more conserved between GM12878 and K562 cell lines, a known feature among the 3D architecture of the human genome, further highlighting their biological significance.

preciseTAD offers flexibility in controlling both the number of predicted boundaries and the distance between them. The two primary parameters are the probability threshold t and ϵ (referred to as eps; parameter of DBSCAN). The combination of these two quantities changes the resulting number of predicted boundaries from *preciseTAD*. Lower values of t and ϵ will result in more clusters of bases, and therefore, more boundaries. As a heuristic, we evaluated the pairwise combination of 3 different thresholds ($t = (0.975, 0.99, 1.0)$) and 6 different ϵ values ($\epsilon = (1000, 5000, 10000, 15000, 20000, 25000)$)). We found that the normalized overlaps - calculated as the total number of ChIP-seq peaks that overlapped within a given flanked boundary, divided by the number of boundaries - between top TFBS sites and flanked *preciseTAD* boundaries converged for combinations of $t = 1.0$ and $\epsilon = 10000$ (Additional file 17: Figure S11).

A limitation of *preciseTAD* is that it is dependent on the "ground truth" boundaries provided by a domain caller. Given the wide variety of domain callers and their variable performance [12,13], defining "ground truth" boundaries is challenging. Ultra-deep Hi-C sequencing [27] and newer technologies for precise mapping of chromatin interactions (e.g., Micro-C [43]), coupled with more precise technologies for genomic annotation profiling (e.g., CUT&RUN for precise mapping of transcription factor binding sites) will help to refine the location and the genomic signatures of the "ground truth" boundaries. In the current work, we feel that the total number of domain boundaries is sufficient to guide learning of the association between genomic annotations and boundaries for precise boundary predictions. Indeed, models trained on the larger number of Peakachu-predicted boundaries performed better than those trained on Arrowhead boundaries. Although we provide models trained on both boundary types, we recommend Peakachu-trained models for the base-level prediction of domain boundaries.

Another limitation of *preciseTAD* is that it does not distinguish boundary types. The hierarchical nature of TAD boundaries [33,44,45] is not considered by *preciseTAD* due to the lack of gold standard of TAD hierarchy. *preciseTAD* also does not consider the directionality of CTCF binding [46] as it predicts individual boundaries in contrast to pairs of convergent CTCF motifs marking individual domains. 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 [43]. Furthermore, actively transcribed regions can serve as TAD boundaries themselves, independently of CTCF binding [46]. This may lead to some TAD boundaries being undetected by *preciseTAD* despite being detected by domain callers. Our future work will involve incorporating the directionality of CTCF binding in predictive modeling, including additional predictor types, and defining separate models trained on different boundary types.

Using the provided pre-trained models, *preciseTAD* can identify precise domain boundaries using cell line-specific genome annotation information of only four transcription factors, including CTCF, SMC3, RAD21, and ZNF143. Although the ENCODE, NIH Roadmap Epigenomics, FANTOM5, BLUEPRINT, and other members of the International Human Epigenome Consortium (IHEC) [47] have been actively cataloging cell line-specific genome annotation datasets, these data remain very sparse. To date (September 2020), the ENCODE project has only five cell lines (GM12878, K562, H1-hESC, HeLa-S3, HepG2) with all four transcription factor genomic annotations. Efforts to impute missing cell line-specific genome annotation information [48] started to appear. We envision the broader availability of cell line-specific genomic annotations will enable a more systematic analysis of domain boundaries using *preciseTAD*.

Conclusion

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 the 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 ChIP-seq annotations and 3D Hi-C sequencing 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. We hope that *preciseTAD* will serve as an efficient and easy-to-use tool to further explore the genome's 3D organization.

Methods

Data sources

TAD and loop boundaries called by Arrowhead [36] and Peakachu [16] tools were used for training and testing. 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 kb, 10 kb, 25 kb, 50 kb, and 100 kb resolutions) using the default parameters (Additional file 1: Arrowhead Script). Peakachu chromatin loop boundaries called at 10 kb for the GM12878, and K562 cell lines were downloaded from the Yue lab website (Additional file 2: Table S1). Unique boundaries were considered as the midpoints within the coordinate of each chromatin loop anchor. Chromosome 9 was excluded from all downstream analyses due to the sparsity of contact matrices at 5 kb and 10 kb resolutions for the K562 cell line. 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 [49] (Additional file 3: Table S2).

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 ($\mathbf{Y}$) used for classification (Figure 1A).

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*) (Figure 1B). A (log_2+1)-transformation of distance was used to account for the skewness of the distance distributions (Additional file 5: Figure S1). Models built using a *Peak Signal Strength* predictor space were only composed of histone modifications and transcription factor binding sites because BroadHMM chromatin states lack signal values.

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 [32] (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.

Establishing optimal data level characteristics for TAD boundary region prediction

Random forest (RF) classification models (the `caret` v.6.0 R package [50]) were built to compare model performances between combinations of data resolutions, feature engineering procedures, and resampling techniques. Following recommendations to evaluate the model on unseen data [51], a *holdout chromosome* technique was used for estimating model performance. The ith 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 an 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. Models were evaluated using Balanced Accuracy (BA), defined as the average of sensitivity and specificity:

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

where 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 ith holdout chromosome, and performances were aggregated using the mean and standard deviation.

Feature selection and predictive importance

Many genomic annotations, notably architectural proteins, tend to exhibit an extensive pattern of colocalization (correlation) [52]. To suitably reduce the predictor space and improve computational efficiency, while maintaining optimal performance, we implemented recursive feature elimination (RFE). We estimated the near-optimal number of necessary features, ranging from 2 to the maximum number of features incremented by the power of 2. 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.

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 [51]. 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 ith holdout chromosome data from the same cell line. Second, we evaluated models trained on cell line-specific data from $n-1$ chromosomes using the ith 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).

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

To investigate whether we could alleviate the resolution limitations of conventional domain calling tools, we developed *preciseTAD*. This algorithm leverages a random forest model in conjunction with density-based and partitioning techniques to predict boundaries at base-level resolution (Additional file 8: Figure S4).

A random forest model (M) was built on the optimal combination of predictor type (predictor space), resampling technique (class balancing correction), and top-ranked annotations (most informative genomic elements) for a set of binned chromosomes $\{k|k \neq i\}$. 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 and p is the optimal number of annotations. We then evaluate M on the base-level predictor space to extract the probability vector, π_n, denoting each base's probability of being a boundary. A threshold t specifies the probability at which a base with $\pi_n \geq t$ is designated as a potential boundary (the default $t = 1$). Next, density-based spatial clustering of applications with noise (DBSCAN; version 1.1-5) is applied to the matrix of pairwise genomic distances between boundary-annotated bases, D, such that $\pi_n \geq t$. The minimum and maximum coordinates of each cluster, k, of spatially colocalized bases were termed *preciseTAD boundary regions* (PTBR). To precisely identify a single base among each PTBR, *preciseTAD* implements partitioning around medoids (PAM) on the distance matrix, D_k, derived from each cluster. The corresponding cluster medoid was defined as a *preciseTAD boundary point* (PTBP), making it the most representative base coordinate within each clustered PTBR.

The DBSCAN algorithm has two parameters, *MinPts* and *eps* (ϵ). The *MinPts* parameter was set to the recommended value of 3, representing $1 + dim(data)$ [53]. To decide on the optimal value of t and ϵ in *preciseTAD*, we considered the normalized enrichment NE of flanked predicted boundaries, defined as

$$NE = \frac{1}{p}\left[\Sigma_{s=1}^{p}\left[\frac{1}{b}\Sigma_{k=1}^{b} e_{ks}\right]\right]$$

where $e_{ks} = \mathbf{I}\{r_s \in (b_k - f, b_k + f)\}$ is the number of elemental regions r of predictor p that overlap with each flanked boundary. We evaluated NE for combinations of $t = \{0.975, 0.99, 1.0\}$ and $\epsilon = \{1000, 5000, 10000, 15000, 20000, 25000\}$. 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 $\epsilon = 10000$ based on our tests (Additional file 17: Figure S11).

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) [54] (*computeMatrix*, *plotProfile*, and *plotHeatmap* tools). Additionally, 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 called and predicted boundaries. That is, between cell lines, the number of overlapping flanked boundaries divided by the total number of flanked boundaries. Wilcoxon Rank-Sum tests were used to compare chromosome-specific Jaccard indices across cell lines, between *preciseTAD* boundaries and both Arrowhead and Peakachu boundaries, respectively. All statistical analyses were performed in R (version 4.0.1). The significance level was set to 0.05 for all statistical tests.

Additional file legends

Additional File 1: Arrowhead Script. An example script for applying Arrowhead to in situ Hi-C data (HIC001-HIC018) to obtain chromosome-specific TAD boundaries on the GM12878 cell line at 5 kb, 10 kb, 25 kb, 50 kb, 100 kb resolutions.

Additional File 2: Table S1. Data sources for Hi-C matrices used to call topologically associating domains with Arrowhead, as well as loop boundaries obtained by Peakachu.

Additional File 3: Table S2. Domain boundary data and class imbalance summaries across resolutions for Arrowhead and Peakachu on K562.

Additional File 4: Table S3. A complete list of genomic annotations used to build the predictor space for all downstream models. The GRCh37/hg19 human genome assembly was used. "Genomic Class" - broad category of genomic features, "Element" - names of genomic features, "Cell line-Specific Source" - download URL specific to the cell line (not all annotations were provided by the same institutions).

Additional File 5: Figure S1. The log_2 transformation of genomic distances normalizes their distributions. Distances are measured as the number of bases from the center of a genomic bin to the nearest genomic annotation center. Density curves of distances before (red) and after (blue) performing a log_2 transformation across 5 kb, 10 kb, 25 kb, 50 kb, and 100 kb data resolutions for both the (A) GM12878 and (B) K562 cell lines. Each density curve represents an individual genomic annotation (77 total).

Additional File 6: Figure S2. Determining optimal data level characteristics for building TAD boundary region prediction models on K562. (A) Averaged balanced accuracies are compared across

resolution, within each predictor-type: Signal, OC, OP, 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) when using Arrowhead ground truth boundaries for K562. Averaged balanced accuracies are compared for Peakachu-trained models built on (B) GM12878 and (C) K562 within each predictor-type: Signal, OC, OP, and Distance, and across resampling technique: 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.

Additional File 7: Figure S3. SMC3, RAD21, CTCF, and ZNF143 transcription factors accurately predict TAD and loop boundaries in K562. (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) Receiver operating characteristic (ROC) curves and the corresponding average area under the curves (AUCs) when training and testing on K562 data (blue, Arrowhead ground truth; red, Peakachu ground truth) versus training on GM12878 and testing on K5628 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.

Additional File 8: Figure S4. Pseudocode of the *preciseTAD* algorithm.

Additional File 9: Table S4. hg19/GRCh37 genomic coordinates of *preciseTAD*-predicted boundary regions (PTBR) and points (PTBP) for GM12878 and K562 cell lines, using models trained on Arrowhead TAD and Peakachu chromatin loop boundaries as ground truth. For PTBRs, the start and end coordinates define the clusters of spatially proximal bases with the probability of being a boundary equal to 1. For PTBPs, the start and end (start+1) coordinates define the most likely boundary point within each PTBR.

Additional File 10: Figure S5. *preciseTAD* boundaries are enriched for known molecular drivers of 3D chromatin. Signal profile plots comparing the binding strength of top TFBS around flanked (A) Arrowhead called TAD boundaries (blue) and *preciseTAD*-predicted TAD boundaries (green) on K562, (B) Peakachu chromatin loop boundaries (red) and *preciseTAD* predicted loop boundaries (green) on K562, (C) Arrowhead called TAD boundaries (blue), Peakachu chromatin loop boundaries (red), and SpectralTAD called TAD boundaries (green) on GM12878 and (D) on K562.

Additional File 11: Figure S6. *preciseTAD* boundaries are more enriched for known molecular drivers of 3D chromatin, as compared with Arrowhead boundaries. Enrichment heatmaps comparing the signal distribution of CTCF, RAD21, SMC3, and ZNF143 around Arrowhead-called TAD boundaries vs. *preciseTAD*-predicted TAD boundaries for (A) GM12878 and (B) K562 cell lines.

Additional File 12: Figure S7. *preciseTAD* boundaries are more enriched for known molecular drivers of 3D chromatin, as compared with Peakachu boundaries. Enrichment heatmaps comparing the signal distribution of CTCF, RAD21, SMC3, and ZNF143 around Peakachu-predicted chromatin loop boundaries vs. *preciseTAD*-predicted TAD boundaries for (A) GM12878 and (B) K562 cell lines.

Additional File 13: Figure S8. *preciseTAD* boundaries are spatially closer to known molecular drivers of 3D chromatin. Boxplots comparing the log2 genomic distance distributions from predicted and called boundaries to the nearest (A) GM12878-specific and (B) K562-specific CTCF, RAD21, SMC3, and ZNF143 transcription factor binding sites. p-values are derived from the Wilcoxon Rank Sum test.

Additional File 14: Figure S9. The agreement between *preciseTAD*-predicted boundaries using Arrowhead- and Peakachu-trained models. Venn diagrams of boundary overlap using (A) GM12878 and (B) K562 data. Boundaries involving Arrowhead/Peakachu were flanked by 5 kb/10 kb, respectively.

Additional File 15: Figure S10. Pre-trained *preciseTAD* models accurately predict boundaries on cell lines using annotation data only. Venn diagrams and signal profile plots comparing flanked predicted boundaries using Arrowhead (A, B) and Peakachu (C, D) trained models. (A, C) Models trained on GM12878 and predicted on GM12878 (red, GM on GM) vs. models trained on K562 and predicted on GM12878 (blue, K on GM), (B, D) models trained on K562 and predicted on K562 (red, K on K) vs. models trained on GM12878 and predicted on K562 (blue, GM on K). Boundaries involving Arrowhead/Peakachu were flanked by 5 kb/10 kb, respectively.

Additional File 16: Table S5. Additional performance metrics when implementing a random forest using Arrowhead/Peakachu ground truth TAD/loop boundaries. Performances are averaged across each holdout chromosome that was reserved for testing.

Additional File 17: Figure S11. Normalized Enrichment levels suggest t=1.0 and ϵ=10000 as the most optimal parameters for biologically relevant *preciseTAD*-predicted boundaries. Linecharts illustrating the normalized enrichment (NE) between CTCF, RAD21, SMC3, ZNF143 and resolution-flanked *preciseTAD*-predicted boundaries for different combinations of thresholds (t) and epsilon parameter values (eps). NE was calculated as the total number of ChIP-seq peaks that overlapped within a given flanked boundary, divided by the number of boundaries that were predicted, and averaged over the number of annotations included in the model. Data from GM12878 (A) and K562 (B) cell lines, chromosome 22, at 5 kb resolution were used. Error bars indicate 1 standard deviation from the mean.

Availability of data and materials

All data generated or analyzed during this study are publicly available and provided in the supplementary files. The source code of *preciseTAD* version 0.99.6 used for all analyses is available at https://github.com/dozmorovlab/preciseTAD.

Authors' contributions

MD conceptualized and supervised the study. MD and SS wrote the manuscript. SS and MD wrote the software associated with the R package. SS performed the analysis for the final results. The authors read and approved the final manuscript.

References

1. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, others: **A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping**. *Cell* 2014, **159**:1665–1680.

2. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, Berkum NL van, Meisig J, Sedat J, others: **Spatial partitioning of the regulatory landscape of the x-inactivation centre**. *Nature* 2012, **485**:381.

3. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B: **Topological domains in mammalian genomes identified by analysis of chromatin interactions**. *Nature* 2012, **485**:376.

4. Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, others: **Comprehensive mapping of long-range interactions reveals folding principles of the human genome**. *science* 2009, **326**:289–293.

5. Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H, Opitz JM, Laxova R, Santos-Simarro F, Gilbert-Dussardier B, Wittler L, Borschiwer M, Haas SA, Osterwalder

M, Franke M, Timmermann B, Hecht J, Spielmann M, Visel A, Mundlos S: **Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions**. *Cell* 2015, **161**:1012–102510.1016/j.cell.2015.04.004.

6. Franke M, Ibrahim DM, Andrey G, Schwarzer W, Heinrich V, Schöpflin R, Kraft K, Kempfer R, Jerković I, Chan W-L, Spielmann M, Timmermann B, Wittler L, Kurth I, Cambiaso P, Zuffardi O, Houge G, Lambie L, Brancati F, Pombo A, Vingron M, Spitz F, Mundlos S: **Formation of new chromatin domains determines pathogenicity of genomic duplications**. *Nature* 2016, **538**:265–26910.1038/nature19800.

7. Hnisz D, Weintraub AS, Day DS, Valton A-L, Bak RO, Li CH, Goldmann J, Lajoie BR, Fan ZP, Sigova AA, others: **Activation of proto-oncogenes by disruption of chromosome neighborhoods**. *Science* 2016, **351**:1454–1458.

8. Taberlay PC, Achinger-Kawecka J, Lun AT, Buske FA, Sabir K, Gould CM, Zotenko E, Bert SA, Giles KA, Bauer DC, others: **Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations**. *Genome research* 2016, **26**:719–731.

9. Lupiáñez DG, Spielmann M, Mundlos S: **Breaking tads: How alterations of chromatin domains result in disease**. *Trends in Genetics* 2016, **32**:225–237.

10. Sun JH, Zhou L, Emerson DJ, Phyo SA, Titus KR, Gong W, Gilgenast TG, Beagan JA, Davidson BL, Tassone F, Phillips-Cremins JE: **Disease-associated short tandem repeats co-localize with chromatin domain boundaries**. *Cell* 2018, **175**:224–238.e1510.1016/j.cell.2018.08.005.

11. Meaburn KJ, Cabuy E, Bonne G, Levy N, Morris GE, Novelli G, Kill IR, Bridger JM: **Primary laminopathy fibroblasts display altered genome organization and apoptosis**. *Aging Cell* 2007, **6**:139–5310.1111/j.1474-9726.2007.00270.x.

12. Zufferey M, Tavernari D, Oricchio E, Ciriello G: **Comparison of computational methods for the identification of topologically associating domains**. *Genome Biol* 2018, **19**:21710.1186/s13059-018-1596-9.

13. Forcato M, Nicoletti C, Pal K, Livi CM, Ferrari F, Bicciato S: **Comparison of computational methods for hi-c data analysis**. *Nature methods* 2017, **14**:679.

14. Ay F, Bailey TL, Noble WS: **Statistical confidence estimation for hi-c data reveals regulatory chromatin contacts**. *Genome Res* 2014, **24**:999–101110.1101/gr.160374.113.

15. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK: **Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities**. *Mol Cell* 2010, **38**:576–8910.1016/j.molcel.2010.05.004.

16. Salameh TJ, Wang X, Song F, Zhang B, Wright SM, Khunsriraksakul C, Yue F: **A supervised learning framework for chromatin loop detection in genome-wide contact maps**. *bioRxiv* 2019,:739698.

17. Li X, An Z, Zhang Z: **Comparison of computational methods for 3D genome analysis at single-cell hi-c level**. *Methods* 2019,.

18. Fudenberg G, Imakaev M, Lu C, Goloborodko A, Abdennur N, Mirny LA: **Formation of chromosomal domains by loop extrusion**. *Cell reports* 2016, **15**:2038–2049.

19. Schmitt AD, Hu M, Ren B: **Genome-wide mapping and analysis of chromosome architecture**. *Nature reviews Molecular cell biology* 2016, **17**:743.

20. Hansen AS, Cattoglio C, Darzacq X, Tjian R: **Recent evidence that tads and chromatin loops are dynamic structures**. *Nucleus* 2018, **9**:20–32.

21. Sanborn AL, Rao SS, Huang S-C, Durand NC, Huntley MH, Jewett AI, Bochkov ID, Chinnappan D, Cutkosky A, Li J, others: **Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes**. *Proceedings of the National Academy of Sciences* 2015, **112**:E6456–E6465.

22. Alipour E, Marko JF: **Self-organization of domain structures by dna-loop-extruding enzymes**. *Nucleic acids research* 2012, **40**:11202–11212.

23. Mirny LA, Imakaev M, Abdennur N: **Two major mechanisms of chromosome organization**. *Curr Opin Cell Biol* 2019, **58**:142–15210.1016/j.ceb.2019.05.001.

24. Davidson IF, Bauer B, Goetz D, Tang W, Wutz G, Peters J-M: **DNA loop extrusion by human cohesin**. *Science* 2019, **366**:1338–134510.1126/science.aaz3418.

25. Libbrecht MW, Ay F, Hoffman MM, Gilbert DM, Bilmes JA, Noble WS: **Joint annotation of chromatin state and chromatin conformation reveals relationships among domain types and identifies domains of cell-type-specific expression**. *Genome research* 2015, **25**:544–557.

26. Mendoza-Vargas A, Olvera L, Olvera M, Grande R, Vega-Alvarado L, Taboada B, Jimenez-Jacinto V, Salgado H, Juárez K, Contreras-Moreira B, others: **Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in e. Coli**. *PLoS One* 2009, **4**:e7526.

27. Bonev B, Cohen NM, Szabo Q, Fritsch L, Papadopoulos GL, Lubling Y, Xu X, Lv X, Hugnot J-P, Tanay A, others: **Multiscale 3D genome rewiring during mouse neural development**. *Cell* 2017, **171**:557–572.

28. Boulesteix A-L, Janitza S, Kruppa J, König IR: **Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics**. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2012, **2**:493–507.

29. Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL: **Juicer provides a one-click system for analyzing loop-resolution hi-c experiments**. *Cell systems* 2016, **3**:95–98.

30. Prati RC, Batista G, Monard MC: **A survey on graphical methods for classification predictive performance evaluation**. *IEEE Transactions on Knowledge and Data Engineering* 2011, **23**:1601–1618.

31. Wei Q, Dunbrack Jr RL: **The role of balanced training and testing data sets for binary classifiers in bioinformatics**. *PloS one* 2013, **8**:e67863.

32. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP: **SMOTE: Synthetic minority over-sampling technique**. *Journal of artificial intelligence research* 2002, **16**:321–357.

33. Cresswell KG, Stansfield JC, Dozmorov MG: **SpectralTAD: An r package for defining a hierarchy of topologically associated domains using spectral clustering**. *BMC Bioinformatics* 2020, **21**:31910.1186/s12859-020-03652-w.

34. Krefting J, Andrade-Navarro MA, Ibn-Salem J: **Evolutionary stability of topologically associating domains is associated with conserved gene regulation**. *BMC biology* 2018, **16**:1–12.

35. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G: **Three-dimensional folding and functional organization principles of the drosophila genome**. *Cell* 2012, **148**:458–472.

36. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL: **Juicebox provides a visualization system for hi-c contact maps with unlimited zoom**. *Cell systems* 2016, **3**:99–101.

37. Mourad R, Cuvier O: **Computational identification of genomic features that influence 3D chromatin domain formation**. *PLoS computational biology* 2016, **12**:e1004908.

38. Moore BL, Aitken S, Semple CA: **Integrative modeling reveals the principles of multiscale chromatin boundary formation in human nuclear organization**. *Genome Biol* 2015, **16**:11010.1186/s13059-015-0661-x.

39. Al Bkhetan Z, Plewczynski D: **Three-dimensional epigenome statistical model: Genome-wide chromatin looping prediction**. *Sci Rep* 2018, **8**:521710.1038/s41598-018-23276-8.

40. Kai Y, Andricovich J, Zeng Z, Zhu J, Tzatsos A, Peng W: **Predicting ctcf-mediated chromatin interactions by integrating genomic and epigenomic features**. *Nature communications* 2018, **9**:1–14.

41. Chang L-H, Ghosh S, Noordermeer D: **TADs and their borders: Free movement or building a wall?** *Journal of Molecular Biology* 2020, **432**:643–652.

42. Flyamer IM, Gassler J, Imakaev M, Brandão HB, Ulianov SV, Abdennur N, Razin SV, Mirny LA, Tachibana-Konwalski K: **Single-nucleus hi-c reveals unique chromatin reorganization at oocyte-to-zygote transition**. *Nature* 2017, **544**:110–11410.1038/nature21711.

43. Krietenstein N, Abraham S, Venev SV, Abdennur N, Gibcus J, Hsieh T-HS, Parsi KM, Yang L, Maehr R, Mirny LA, Dekker J, Rando OJ: **Ultrastructural details of mammalian chromosome architecture**. *Mol Cell* 2020, **78**:554–565.e710.1016/j.molcel.2020.03.003.

44. Weinreb C, Raphael BJ: **Identification of hierarchical chromatin domains**. *Bioinformatics* 2016, **32**:1601–1609.

45. Fraser J, Ferrai C, Chiariello AM, Schueler M, Rito T, Laudanno G, Barbieri M, Moore BL, Kraemer DC, Aitken S, others: **Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation**. *Molecular systems biology* 2015, **11**.

46. Harrold CL, Gosden ME, Hanssen LLP, Stolper RJ, Downes DJ, Telenius JM, Biggs D, Preece C, Alghadban S, Sharpe JA, Davies B, Sloane-Stanley JA, Kassouf MT, Hughes JR, Higgs DR: **A functional overlap between actively transcribed genes and chromatin boundary elements**. *bioRxiv*,:2020.07.01.18208910.1101/2020.07.01.182089Available: http://biorxiv.org/content/early/2020/07/01/2020.07.01.182089.abstract.

47. Stunnenberg HG, International Human Epigenome Consortium, Hirst M: **The international human epigenome consortium: A blueprint for scientific collaboration and discovery**. *Cell* 2016, **167**:1145–114910.1016/j.cell.2016.11.007.

48. Schreiber J, Durham T, Bilmes J, Noble WS: **Avocado: A multi-scale deep tensor factorization method learns a latent representation of the human epigenome**. *Genome Biol* 2020, **21**:8110.1186/s13059-020-01977-6.

49. Tyner C, Barber GP, Casper J, Clawson H, Diekhans M, Eisenhart C, Fischer CM, Gibson D, Gonzalez JN, Guruvadoo L, Haeussler M, Heitner S, Hinrichs AS, Karolchik D, Lee BT, Lee CM, Nejad P, Raney BJ, Rosenbloom KR, Speir ML, Villarreal C, Vivian J, Zweig AS, Haussler D, Kuhn RM, Kent WJ: **The ucsc genome browser database: 2017 update**. *Nucleic Acids Res* 2017, **45**:D626–D63410.1093/nar/gkw1134.

50. Kuhn M: **The caret package**. *R Foundation for Statistical Computing, Vienna, Austria. URL https://cran. r-project. org/package= caret* 2012,.

51. Schreiber J, Singh R, Bilmes J, Noble WS: **A pitfall for machine learning methods aiming to predict across cell types**. *bioRxiv* 2019,:512434.

52. Van Bortle K, Nichols MH, Li L, Ong C-T, Takenaka N, Qin ZS, Corces VG: **Insulator function and topological domain border strength scale with architectural protein occupancy**. *Genome biology* 2014, **15**:R82.

53. Hahsler M, Piekenbrock M, Doran D: **Dbscan: Fast density-based clustering with r**. *Journal of Statistical Software* 2019, **25**:409–416.

54. Ramirez F, Ryan DP, Gruning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dundar F, Manke T: **DeepTools2: A next generation web server for deep-sequencing data analysis**. *Nucleic acids research* 2016, **44**:W160–W165.