2021

Methods for developing a machine learning framework for precise 3D domain boundary prediction at base-level resolution

Spiro C. Stilianoudakis
Virginia Commonwealth University

Methods for developing a machine learning framework for precise 3D domain boundary prediction at base-level resolution

by

Spiro C. Stilianoudakis

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Biostatistics

Virginia Commonwealth University

Doctoral Committee:

Mikhail G. Dozmorov[1]

Le Kang[1]

Roy T. Sabo[1]

David C. Wheeler[1]

Paul Brooks[2]

Timothy P. York[3]

[1] Dept. of Biostatistics, Virginia Commonwealth University, Richmond, VA, 23298, USA

[2] School of Business, Virginia Commonwealth University, Richmond, VA, 23298, USA

[3] Department of Human and Molecular Genetics, Virginia Commonwealth University, Richmond, VA, 23298, USA

Table of Contents

I. Acknowledgements

Firstly, I want to personally thank my advisor, Dr. Mikhail Dozmorov, for his dedicated mentorship. This dissertation would not be possible without him. During the nearly 4 years working under the guidance of Dr. Dozmorov his door was always open to me. His vast knowledge and support allowed me to provide many significant contributions to the field of genomics in a short amount of time.

I would also like to thank my committee members, Dr. Le Kang, Dr. David Wheeler, Dr. Roy Sabo, Dr. James Brooks, and Dr. Timothy York. Their devoted input allowed me to develop this dissertation into its best possible version. Additionally, I would like to give thanks to Russell Boyle and the members of the Graduate Admissions Committee in the Department of Biostatistics at Virginia Commonwealth University. I am incredibly gracious for their recruitment and impact in my life. I would like to thank my research assistantship advisors, Dr. Caroline Carrico and Dr. Shillpa Naavaal. Their supervision allowed me to develop the tools necessary to become a highly qualified statistician.

Finally, I would like to thank my family and friends whose continued support through graduate school kept me on the track toward success. To my mother, Tawny, father Manoli, and brother, Niko, this dissertation is a reflection of your love and I hope it makes you proud. To John and Kellen, thank you for your help during our time in the department and continued friendship.

It has been a sincere privilege to be a graduate student at VCU. I hope to be able to represent the Department of Biostatistics in the manner of high esteem that it deserves.

II. List of Figures

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 5. 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 1 standard deviation from the mean performance across each holdout chromosome used for testing.

Figure 6. 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 TFBS. The columns represent the holdout chromosome excluded from the training data. Rows are sorted in decreasing order according to the columnwise average importance.

Figure 7. The *preciseTAD* algorithm.

Figure 8. A schematic illustrating how each of the diagnostic summaries are calculated in the preciseTAD algorithm. The illustration depicts blue regions as collections of base coordinates whose predictive probability exceeds a predefined threshold, *t*, organized into two clusters. The summary statistics include the following: PTBRWidth - PTBR width, PTBRCoverage - the ratio of base-level coordinates with probabilities that exceed the threshold to PTBRWidth, DistanceBetweenPTBR - the genomic distance between the end of the previous PTBR and the start of the subsequent PTBR, NumSubRegions - the number of elements in each PTBR cluster, SubRegionWidth - the genomic coordinates spanning the subregion associated with each PTBR, and DistBetweenSubRegions - the genomic distance between the end of the previous PTBR-specific region and the start of the subsequent PTBR-specific region.

Figure 9. *preciseTAD*-predicted boundaries better reflect intra-chromosomal contacts. (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 corresponding signal profiles of CTCF, RAD21, SMC3, and ZNF143.

Figure 10. *preciseTAD*-predicted boundaries are enriched for known molecular drivers of 3D chromatin. Signal profile plots comparing the strength of CTCF, RAD21, SMC3, and ZNF143 binding around Arrowhead-called boundaries (blue, C), Peakachu loop boundaries (red, D) vs. *preciseTAD*-predicted boundaries (green).

Figure 11. *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 (C), and *preciseTAD*-predicted boundaries using (D) Arrowhead- and (E) Peakachu-trained models. Boundaries involving Arrowhead/Peakachu were flanked by 5 kb/10 kb, respectively.

Figure 12. 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.

Figure 14. Training and testing across cell lines performs similarly to within the same cell line. Receiver operating characteristic (ROC) curves and the corresponding average area under the curves (AUCs) when (A) 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), and (B) training and testing on K562 data (blue, Arrowhead ground truth; red, Peakachu ground truth) versus training on GM12878 and testing on K562 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.

Figure 15. Cross-cell-line predicted boundaries strongly overlapped with same-cell-line predicted boundaries. Venn diagrams 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.

Figure 16. Cross-cell-line predicted boundaries were as enriched for known drivers of 3D chromatin as same-cell-line predicted boundaries. Profile plots comparing enrichment levels of CTCF, RAD21, SMC3, and ZNF143 sites around 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.

III. List of Tables

IV. Abstract

Methods for developing a machine learning framework for precise 3D domain boundary prediction at base-level resolution

By Spiro C. Stilianoudakis

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

Virginia Commonwealth University, 2021

Advisor: Mikhail G. Dozmorov, Ph.D.,

Assistant Professor, Blick scholar, Department of Biostatistics

Co-Advisor: Le Kang, Ph.D.

Assistant Professor, Department of Biostatistics

High-throughput chromosome conformation capture technology (Hi-C) has revealed extensive DNA looping and folding into discrete 3D domains. These include Topologically Associating Domains (TADs) and chromatin loops, the 3D domains critical for cellular processes like gene regulation and cell differentiation. The relatively low resolution of Hi-C data (regions of several kilobases in size) prevents precise mapping of domain boundaries by conventional TAD/loop-callers. However, high resolution genomic annotations associated with boundaries, such as CTCF and members of cohesin complex, suggest a computational approach for precise location of domain boundaries.

We developed *preciseTAD*, an optimized machine learning framework that leverages a random forest model to improve the location of domain boundaries. Our method introduces three concepts - *shifted binning*, *distance-type* predictors, and *random under-sampling* - which we use to build classification models for predicting boundary regions. The algorithm then uses density-based clustering (DBSCAN) and partitioning around medoids (PAM) to extract the most biologically meaningful domain boundary from models trained on high-resolution genome

annotation data and boundaries from low-resolution Hi-C data. We benchmarked our method against a popular TAD-caller and a novel chromatin loop prediction algorithm.

Boundaries predicted by *preciseTAD* were more enriched for known molecular drivers of 3D chromatin including CTCF, RAD21, SMC3, and ZNF143. *preciseTAD*-predicted boundaries were more conserved across cell lines, highlighting their higher biological significance. Additionally, models pre-trained in one cell line accurately predict boundaries in another cell line. Using cell line-specific genomic annotations, the pre-trained models enable detecting domain boundaries in cells without Hi-C data.

The research presented provides a unified approach for precisely predicting domain boundaries. This improved precision will provide insight into the association between genomic regulators and the 3D genome organization. Furthermore, our methods will provide researchers with flexible and easy-to-use tools to continue to annotate the 3D structure of the human genome without relying on costly high resolution Hi-C data. The *preciseTAD* R package and supplementary ExperimentHub package, *preciseTADhub*, are available on Bioconductor (version 3.13; https://bioconductor.org/packages/preciseTAD/; https://bioconductor.org/packages/preciseTADhub/).

1. Chapter 1: Introduction

1.1 The 3-dimensional architecture of the human genome

The human genome contains approximately 3 billion nucleotides and, if stretched end-to-end, could reach nearly 2 meters in length. The nucleus, on the other hand, spans approximately 6 micrometers. Therefore, the linear genome must undergo extensive layers of folding and looping to fit inside the nucleus of a cell. This folding does not occur at random, but instead makes up the 3-dimensional (3D) architecture of the human genome. In fact, this 3D architecture is hierarchical in nature (Figure 1). At the smallest scale (nucleosomal), DNA is folded into 11-nm nucleosomes, which in turn wraps around a histone octamer. At the kilobase (kb) scale, *chromatin loops* connect gene promoters with distal enhancers, thereby regulating gene expression [4,5]. At the megabase (Mb) scale, chromatin is organized by spatial domains characterized by preferential contacts between loci in the same domain as opposed to across domain boundaries, referred to as *Topologically Associating Domains* (TADs). Emerging evidence has linked chromatin loops and TADs to critical roles in cell dynamics and cell differentiation. Studies have shown that TADs themselves are highly conserved across species and cell lines [6–11]. Furthermore, TADs have been shown to be divided into sub-chromosomal compartments, referred to as A and B compartments. A compartments are typically gene-rich, DNase I hypersensitive, and transcriptionally active, while B compartments are typically gene-poor and transcriptionally repressed [2,6,10]. Disruption of boundaries demarcating loops and TADs promotes cancer [12,13] and other disorders [14–16]. Therefore, identifying the precise location of TAD and chromatin loop boundaries, referred to as domain boundaries, remains a top priority in our goal to fully understand the functionality of the human genome.

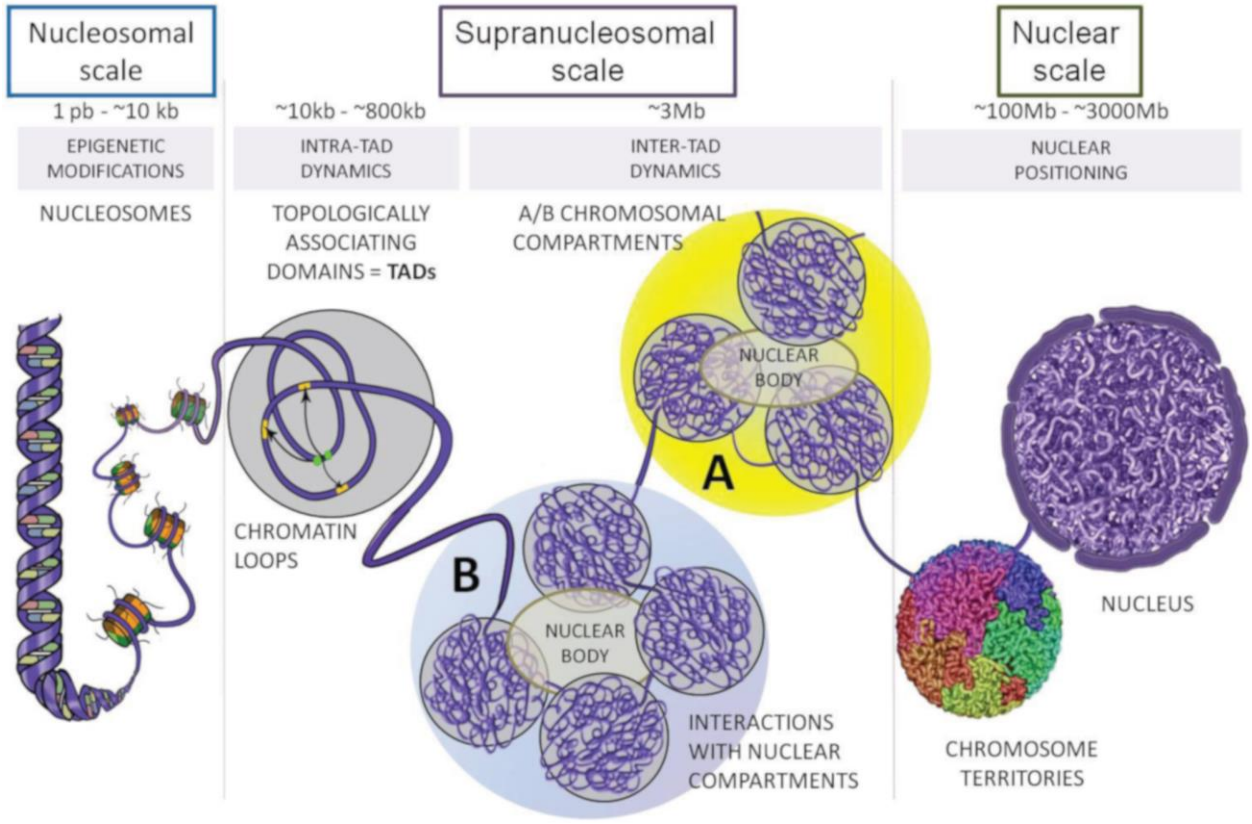


Figure 1. The human genome is organized into a 3D hierarchy*. At the nucleosomal scale (~1 bp - 10 kb), DNA loops around histone octamers, forming nucleosomes which lead to compact chromatin. At the supranucleosomal scale (~10 kb - 800 kb), chromatin loops form regions on the linear genome that are highly self-interacting called Topologically Associated Domains (TADs). TADs themselves organize into epigenomic "compartments" signifying transcriptionally (A) active and (B) inactive chromatin (~3 Mb). At the nuclear scale (~100 Mb - 3000 Mb), chromosomes form "chromosome territories" (obtained from [1]).*

1.2 Domain calling tools and their limitations

In recent years, the development of novel strategies to probe the contacts among higher-order structures of the human genome have enabled analysis of the formation of chromatin loops and TADs. These strategies are based on Chromosome Conformation Capture (3C) sequencing techniques [17]. 3C methods can only capture the structure of a subset of the genome at a

single time (one vs. one). Its extension, 4C, allowed for comparing specific genomic loci with the rest of the genome (one vs. all). A further extension, 5C, allowed for comparing contacts between sets of genomic loci (many vs. many). Finally, the introduction of Hi-C sequencing by Lieberman-Aiden [2] allowed for the capture of all vs. all long-distance chromatin interactions across the entire genome. Generally speaking, Hi-C sequencing involves the following steps: crosslinking cells with formaldehyde, treatment with a restriction enzyme, filling in 5'-overhangs with a biotinylated residue, ligation of the blunt-end fragments, purifying and shearing the DNA, and paired-end sequencing (Figure 2A). The ligated DNA samples produced are the joined fragments of DNA that were in close spatial proximity inside of the nucleus.

The results of a Hi-C sequencing experiment can be visualized via a Hi-C contact matrix, a square and symmetric matrix that measures the pairwise interaction frequencies (IFs) between genomic regions (Figure 2B). The linear genome is binned into non-overlapping regions of fixed width and the matrix entry M_{ij} represents the number of paired-end reads connecting loci i and j. Larger IFs represent pairs of regions that have high levels of interaction when sequenced while low IFs represent pairs of regions with low levels of interaction. The width of the bins is referred to as *resolution* and typically ranges from 5 kb-100 kb. Hi-C resolution is controlled by sequencing depth, with greater sequencing depth leading to smaller genomic bins (i.e., higher resolution). Because increasing the resolution of Hi-C data requires a quadratic increase in the total sequencing depth, obtaining high resolution remains difficult [18].

TADs form as triangular regions along the diagonal of a contact matrix (Figure 2B). Several methods have been proposed to identify genomic coordinates that demarcate TADs (reviewed in [19–21]), and chromatin loops [10,22–24], referred to as *domain-callers*. However, a key limitation of them is that they are heavily reliant on Hi-C data resolution. Conventional domain-callers are restricted to providing genomic coordinates that are divisible by resolution, thereby limiting precise location of boundaries. Another limitation among domain callers is that they

disregard prior knowledge about functional genomic annotations associated with 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, whereby DNA is extruded through cohesin rings forming chromatin loops [25–30]. Furthermore, distinct patterns of histone modifications have also been shown to be present at boundaries [2,6,31]. These genomic annotations are obtained using chromatin immunoprecipitation followed by high-throughput DNA sequencing (ChIP-seq). However, the full list of functional genomic annotations associated with boundary location remains unclear.

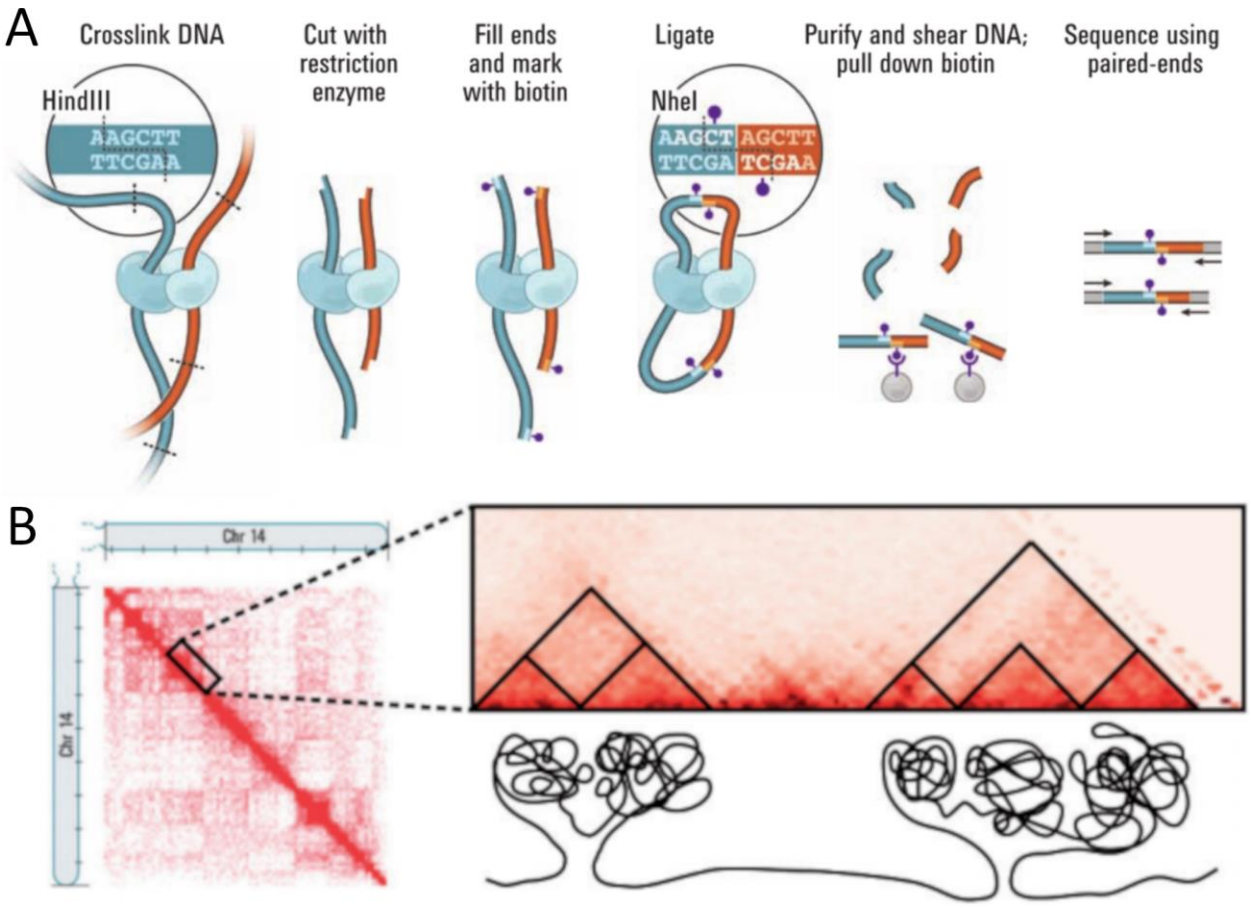


Figure 2. Overview of Hi-C sequencing. *(A) An illustration depicting the steps in the Hi-C sequencing protocol (obtained from [2]). (B) An illustration of the structural formation of TADs.*

The Hi-C contact matrix is shown on the left. TADs and sub-TADs are outlined as triangles, with an example of the corresponding DNA structure depicted below (obtained from [3]).

1.3 Motivation for research

Computational approaches that integrate ChIP-seq data with Hi-C data may be well-suited to identify the key drivers of chromatin architecture. Moreover, the resolution of ChIP-seq experiments is typically on the order of tens to hundreds of bases [32], well below the resolution of Hi-C data (tens of kilobases; 750 bp is the highest resolution of Hi-C data to date [33]). Therefore, leveraging precisely mapped genomic annotations in a supervised machine learning framework enables the possibility for more precise prediction of the locations of domain boundaries. Our research can help bridge the resolution gap between 1D ChIP-seq annotations and 3D Hi-C sequencing data for more precise and biologically meaningful boundary identification.

1.4 Aims

Our goal is to establish a unified approach toward domain boundary prediction using ChIP-seq data. First, we focus on transforming TAD/loop-calling into a prediction problem. We propose a machine learning framework to determine the optimal combination of data-level characteristics necessary for optimal domain boundary prediction performance. Our framework will allow us to be the first to address several impacting factors of Hi-C data on domain boundary location including resolution, feature engineering, and class imbalance. Next, we extend this framework and develop a novel density-based clustering and partitioning technique that precisely predicts biologically meaningful domain boundaries at base-level resolution. Our method will alleviate the resolution limitations of conventional TAD/loop callers. Finally, we develop a novel technique that can be used to predict boundaries for cell lines that do not currently have Hi-C data available. This will circumvent the costly and time-consuming process of performing Hi-C

sequencing at high resolution on many different cell lines. We will apply our methods on two well-studied cell lines, GM12878 and K562, and benchmark them against two popular domain boundary calling tools, Arrowhead [34], an established TAD-caller, and Peakachu [24] a recently published algorithm for predicting chromatin loops. These methods will be developed into an R package that will be freely available for the scientific community to use. The research presented in this dissertation is highlighted by the following aims:

1.4.1 Aim 1: Develop a machine learning framework to establish an optimal domain boundary region prediction model

Domain boundary prediction is a multi-faceted problem requiring consideration of multiple statistical and biological properties of genomic data. It is unclear what the complete set of genomic annotations are most influential to TAD/loop formation. Even more unclear, is if the mechanisms that lead to the formation of TADs and chromatin loops are the same, and how they might differ between cell lines. Thus, there is a need for a unified approach toward domain prediction that can shed light on these mechanisms. Here, we propose a machine learning framework that will transform domain calling into a supervised classification problem by leveraging many different high-resolution 1D ChIP-seq annotations, across multiple cell lines. We will be utilizing the random forest algorithm to predict domain boundary regions. In doing so, we will develop the concept of *shifted binning*, a novel technique for building domain data for predictive modeling. Additionally, we will develop a new technique for feature engineering based on the spatial associations between boundary regions and genomic annotations, known as *distance-type features*. We will compare our method to established feature engineering techniques such as *overlap counts* and *overlap percents*. These methods are, in part, compiled into an R package, *preciseTAD*, and available on Bioconductor. Additionally, pre-trained models will be provided in a public repository via an ExperimentHub R package on Bioconductor, referred to as *preciseTADhub*.

1.4.2 Aim 2: Develop a density-based partitioning technique for precise boundary prediction at base-level resolution

Accurate TAD/loop coordinate mapping remains difficult, as it is strongly reliant on the calling algorithm and Hi-C data resolution. Obtaining genome-wide chromatin interactions at high-resolution is costly. In contrast, the resolution of ChIP-seq experiments remains much higher, and at much lower costs, compared to Hi-C data. Therefore, implementing computational approaches that leverage protein-binding data, as well as other functional genomic elements (histone modifications and cis-regulatory elements) may help to improve the precise location of domain boundaries. Thus, we propose a method that alleviates resolution restrictions by predicting domain boundary coordinates as base-level resolution. We will evaluate the biological significance of our base-level predicted boundaries using the peak signal strength around known molecular drivers of 3D chromatin including CTCF, RAD21, SMC3, and ZNF143 [25–30]. Additionally, we will evaluate the conservation of our predicted boundaries across cell lines and compare our results with Arrowhead and Peakachu. This method is primarily compiled into an R package, *preciseTAD*.

1.4.3 Aim 3: Develop a technique for predicting boundaries on cell lines that do not have publicly available Hi-C data

Currently, there are few cell lines with high-resolution Hi-C data available in the public domain [6,10]. However, high-resolution 1D ChIP-seq data in the form of histone modifications, cis-regulatory elements, and transcription factors, are publicly available for various cell lines [35]. Thus, using genomic annotations most predictive of boundary regions, we will develop a technique that could precisely predict base-level boundary coordinates on one cell line using annotation data from another cell line. We will evaluate two scenarios: 1) training and predicting on the same cell line vs. training and predicting on different cell lines. Our method will expand our knowledge of the cell line specificity of domain boundary formation, while avoiding costly

high-resolution Hi-C sequencing. We will compile this method into an R package and provide pre-trained cell-line specific models in a publicly available repository.

2. Chapter 2: Aim 1 - Develop a machine learning framework to establish an optimal TAD boundary region prediction model

2.1 Introduction

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 [10], Topologically Associating Domains (TADs) [6,8], and A/B compartments [2]. At the kilobase scale, chromatin loops connect gene promoters with distal enhancers, thereby regulating gene expression [4,5]. 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 [12,13] and other disorders [14–16]. Therefore, determining the mechanisms that lead to the formation of TADs and loops is an instrumental step toward precisely identifying their locations throughout the linear genome.

Functional genomic annotations have been shown to be associated with domain boundaries. Among these is the insulator binding protein, CTCF. As a regulator of 3D chromatin, CTCF mediates long-range contacts and the formation of insulated neighborhoods [36,37]. As a transcription factor, CTCF binds to enhancer-promotor regions by co-localizing with other DNA-binding proteins to regulate gene expression [12,38]. These other factors include RAD21 and SMC3, whose recruitment form a cohesin ring under the proposed loop extrusion model, whereby loops are formed during interphase [26–28]. It has also been shown that chromatin interactions are associated with distinct patterns of histone modifications. Specifically, active H3K4 methylation marks have been observed at loop boundaries, likely acting as domain barriers to physically separate active and repressive chromatin domains [31,39,40]. The full list of elements associated with domain formation remains unclear. Moreover, it is unclear if all of these functional genomic elements, or specific combinations of them, play a role in distinguishing between TADs and chromatin loops.

Recent methods have been developed to implement classification models to predict boundary location on the human genome using functional genomic annotations via ChIP-seq data. However, all ignore key characteristics of 3D genomic data that are detrimental to both model performance and precise boundary identification. A method developed by Mourad et al. [41], called *HiCFeat*, used the percentage of overlap between several ChIP-seq defined transcription factor binding site (TFBS) regions and 10 kb genomic bins to build an L1-regularized multiple logistic regression model to predict TAD boundary regions. However, such a high level of resolution is likely to introduce heavily imbalanced classes created from the proportionally much smaller number of TAD boundary regions compared to non-TAD boundary regions. *HiCFeat* did not address the class imbalance. Instead, model performance was evaluated using area under the receiver operating characteristic curve (AUROC) which is known to be insensitive to class imbalance, creating artificially inflated values influenced by the majority class [42,43]. Furthermore, by only considering the percentage of overlap in defining the feature space, *HiCFeat* is limited in its granularity given that many regions on the linear genome will not be overlapped by any TFBS regions. Two additional studies were proposed in 2015 and 2017 respectively, one using histone modifications in a Bayesian Additive Regression Trees (BART) model [40] and the other using combined sets of TFBS, DNase I hypersensitive sites, and histone modifications together in an L2-based regularized linear model [44]. The BART model was built on relatively high-resolution Hi-C data at 5 kb, while the L2-regularized model used much lower 200 kb resolution Hi-C data. Both methods addressed class imbalance by performing random under-sampling (RUS), and used the elemental read count that appeared in each bin to describe the relationship between ChIP-seq regions and genomic bins. Firstly, given the large disparity between the number of TAD vs. non-TAD boundary regions, performing random under-sampling alone is likely to be unstable and introduce bias. Also, it is unclear how other resampling solutions compare to simple random under-sampling such as random over-sampling or a weighted combination of both random under- and over-sampling together.

Secondly, using read count overlap as the feature space suffers from similar limitations as the percentage of overlap. Moreover, it is unclear how either of the currently established feature engineering procedures compares to enumerating the distance in base pairs between genomic elements and genomic bins in TAD boundary prediction, which offers a more spatial measure of association. Likewise, there does not appear to be a clear consensus in the optimal Hi-C data resolution to use when calling TADs for boundary prediction, as can be evidenced by the wide range of resolutions employed in the methods discussed above. Thus, much is left to be investigated and improved regarding boundary prediction to fully identify the complete set of genomic elements that influence domain formation.

Here we propose a unified machine learning framework for optimal prediction of domain boundary regions. Our method utilizes the random forest (RF) algorithm trained on high-resolution and cell-line-specific chromatin state (BroadHMM), histone modification (HM), and transcription factor binding site (TFBS) data. We introduce a systematic pipeline for building the optimal domain 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. We show that binding of four transcription factors (SMC3, RAD21, CTCF, ZNF143) is sufficient for accurate boundary predictions in both TADs and chromatin loops. These methods and models are implemented and stored in publicly available R packages on Bioconductor, *preciseTAD* and *preciseTADhub*.

2.2 Methods

2.2.1 Data sources

TAD and loop boundaries called by Arrowhead [34] and Peakachu [24] 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) at 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 (Table 1). 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 were obtained from the UCSC Genome Browser Database including 15 BroadHMM chromatin states (BroadHMM), 10 histone modifications (HM), and 52 transcription factor binding sites (TFBS) [45] (Additional file 2: Table S1).

2.2.2 Shifted-binning for binary classification

In Hi-C, each chromosome is binned into non-overlapping regions of length *r*. The *r* parameter is defined by 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 ($r - mod(r) + r/2$). 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 (Figure 3A).

2.2.3 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 genomic distance in bases from the center of the bin to the center of the nearest genomic annotation region (*Distance*) (Figure 3B). A ($log_2 + 1$)-transformation of distance was used to account for the skewness of the distance distributions (Additional file 3: 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.

2.2.4 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 [46] (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.

2.2.5 Establishing optimal data level characteristics for TAD boundary region prediction

Random forest (RF) classification models were built to compare performances between combinations of data resolutions, feature engineering procedures, and resampling techniques. Following recommendations to evaluate the model on unseen data [47], 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, **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:

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

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 i^{th} holdout chromosome, and performances were aggregated using the mean and standard deviation.

2.2.6 Feature selection and predictive importance

Many genomic annotations, notably architectural proteins, tend to exhibit an extensive pattern of colocalization (correlation) [48]. To avoid overfitting, improve computational efficiency, and maintain 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.

2.3 Results

2.3.1 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, and (2) to allow for an interpretable ranking of predictors [49]. We used Arrowhead-called TAD boundaries [50] and published Peakachu-predicted loop boundaries [24] 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, Table 1).

Publisher [Source]	Tool	Library	Cell line	Available Resolution(s)
Rao et al [10]	Arrowhead	HIC001-HIC018	GM12878	5 kb, 10kb, 25 kb, 50 kb, 100 kb
Rao et al [10]	Arrowhead	HIC069-HIC074	K562	5 kb, 10kb, 25 kb, 50 kb, 100 kb
Salameh et al [24]	Peakachu		GM12878	10 kb

Salameh et al [24]	Peakachu	K562	10 kb

Table 1: Data sources for Hi-C matrices used to call topologically associating domains with Arrowhead, as well as loop boundaries obtained by Peakachu.

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 3A, 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 2, Additional file 4: 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 [42,43]. 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 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 [46] (see Methods).

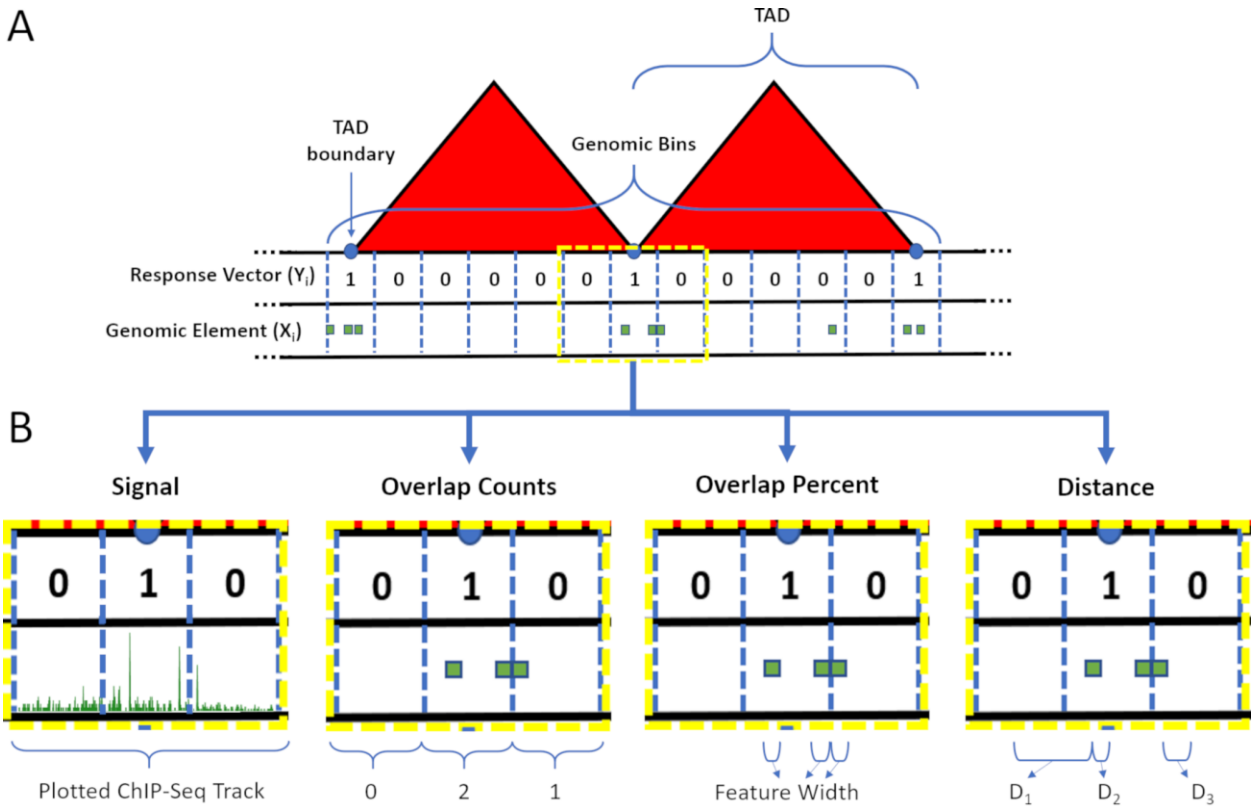


Figure 3. Resolution-specific data construction and feature engineering for random forest modeling. *(A) The linear genome was binned into non-overlapping resolution-specific intervals using shifted binning (see Methods). The response vector* ***Y*** *was defined as 1/0 if a genomic bin overlapped/did 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).*

Tool	Resolution/Bin size	Total number of called domains	Total number of unique domain 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

Table 2: Domain boundary data and class imbalance summaries across resolutions for Arrowhead and Peakachu on GM12878.

A total of 77 cell line-specific genomic annotations were used to build the predictor space. These included 15 BroadHMM chromatin state data, 10 histone modifications (HM), and 52 transcription factor binding sites (TFBS) (Additional file 2: Table S1). Four feature engineering procedures were developed to quantify the association between genomic annotations and bins (Figure 3B). These included signal strength association (Signal), direct (OC), proportional (OP), and spatial ($log_2 + 1$ Distance) relationships. A $log2$ transformation was implemented on the distance feature space to normalize genomic distances (see Methods, Additional file 3: 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 4). 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. 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.

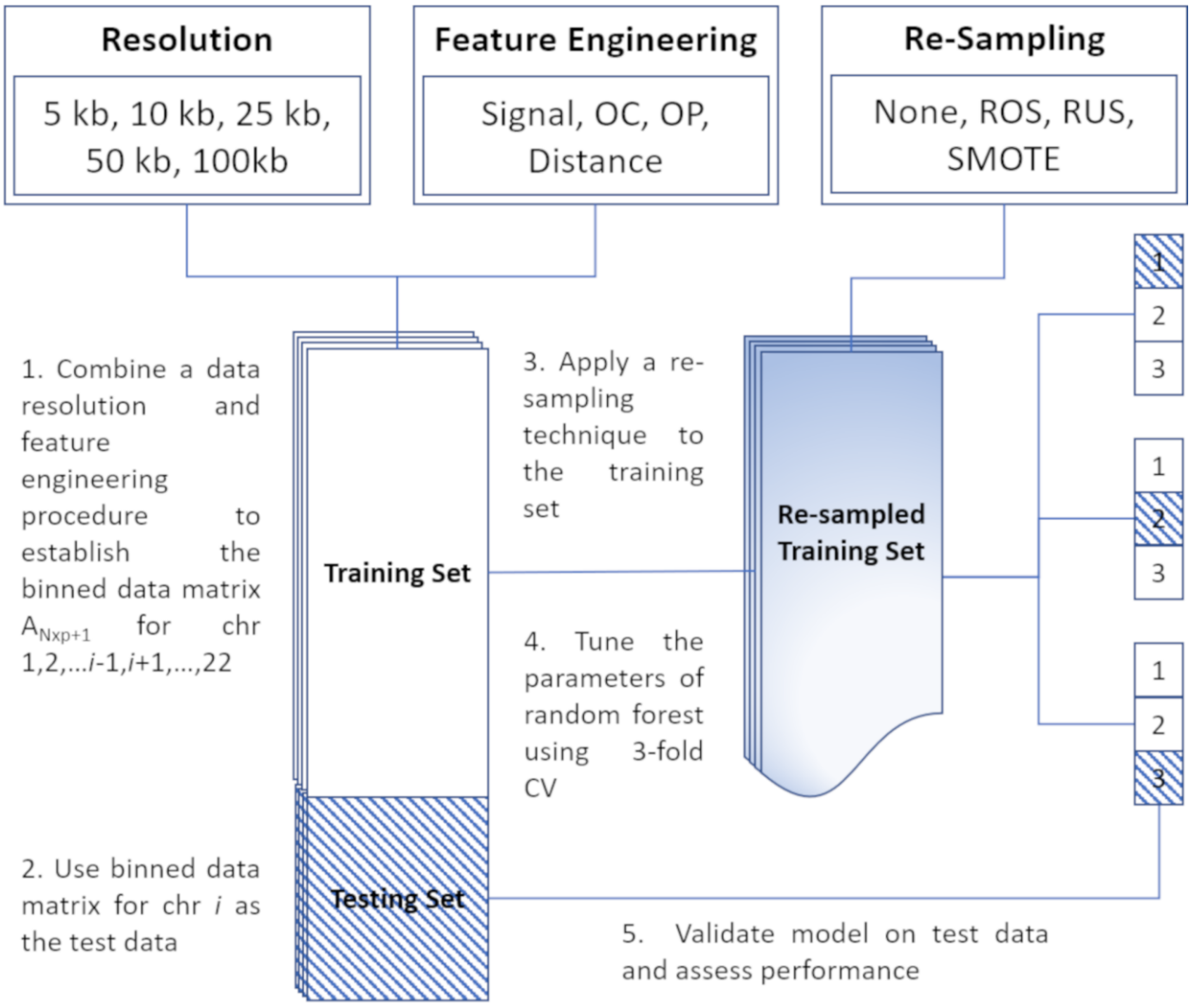


Figure 4. A machine learning framework for building domain boundary region prediction models. *Step 1 employs a range of feature engineering techniques to define the predictor matrix* $A_{N\times(p+1)}$*, where* N *is the number of genomic bins,* p *is the number of genomic annotations,* i *is a holdout 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).

2.3.2 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 5). 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 accuracies 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 5: 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 and RUS performing comparatively similar to each other, for both cell lines (Additional file 5: 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.

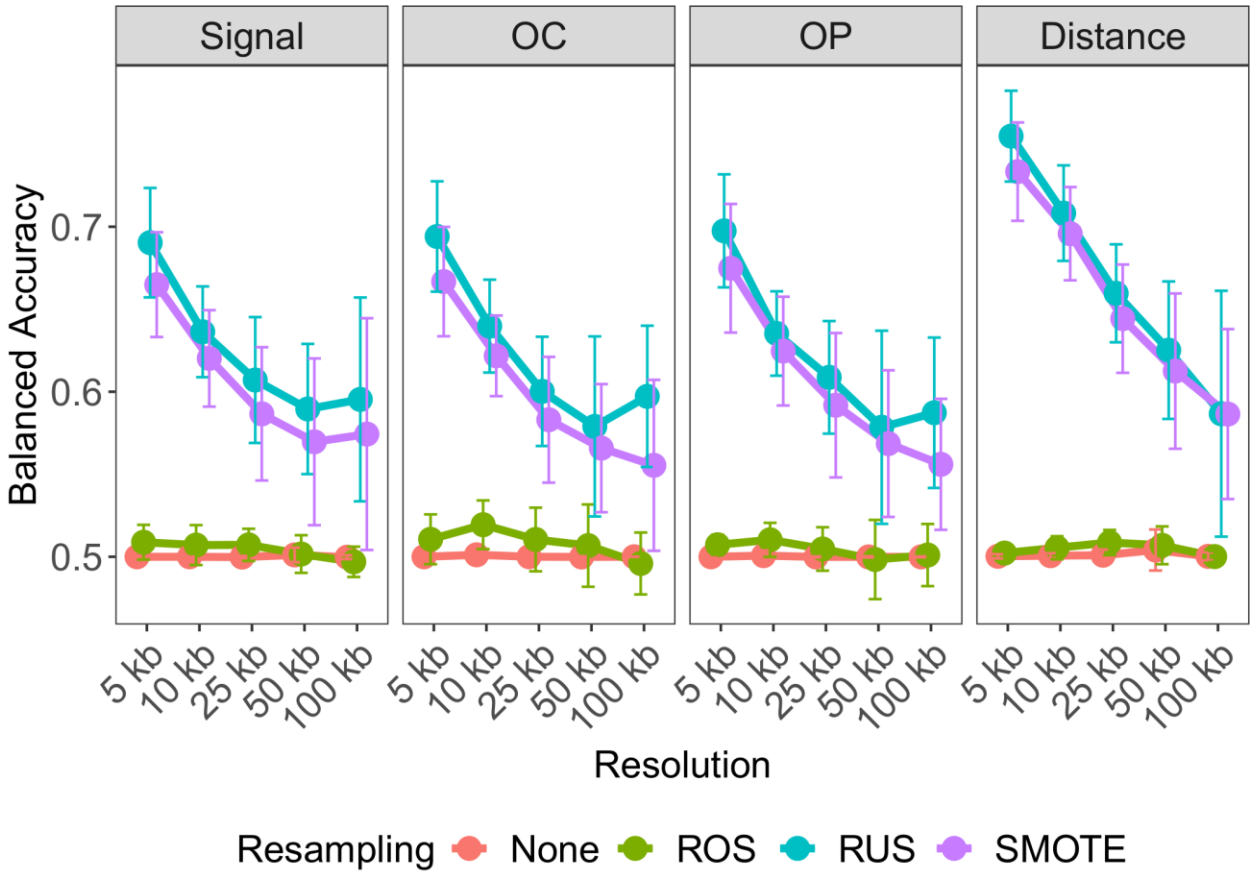


Figure 5. Determining optimal data level characteristics for building TAD boundary region prediction models on GM12878. *Averaged balanced accuracies are compared across resolution, within each predictor-type: Signal, OC, OP, and Distance, and across resampling techniques: no resampling (None; red), 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.*

2.3.3 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 TFBS outperformed other annotation-specific models, with results consistent for loop boundaries, on both cell lines (Figure 6A; Additional file 6: Figure S3A). These results suggest that TFBS are the primary drivers of TAD and loop boundary formation in both GM12878 and K562.

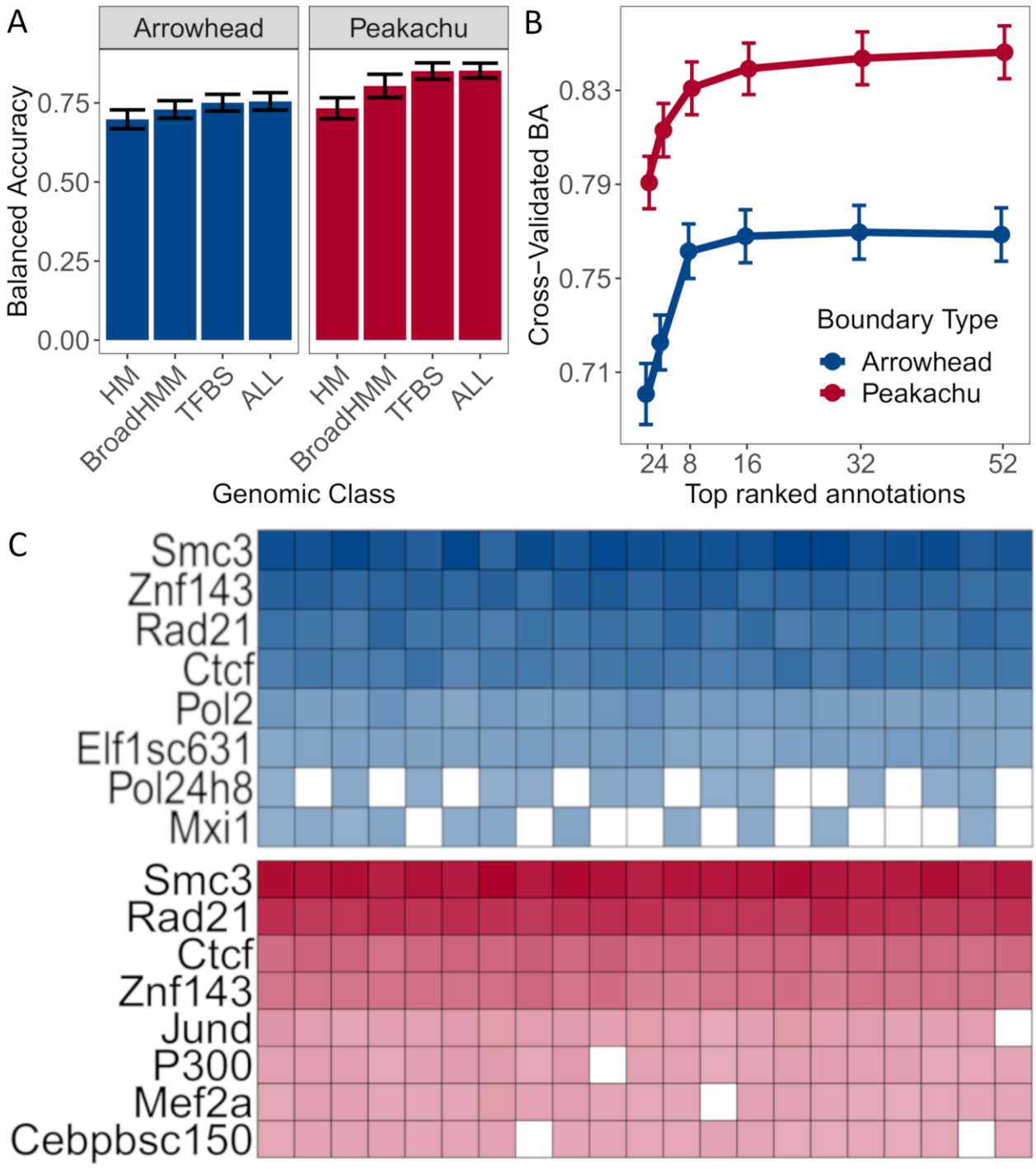


Figure 6. SMC3, RAD21, CTCF, and ZNF143 transcription factors accurately predict TAD and loop boundaries in GM12878. *(A) Barplots comparing performances of TAD (Arrowhead) and loop (Peakachu) boundary prediction models using histone modifications (HM), chromatin states (BroadHMM), transcription factor 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 TFBS. The columns represent the holdout chromosome excluded from the training data. Rows are sorted in decreasing order according to the columnwise average importance.

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

It is known that many elemental proteins colocalize together at binding sites along the linear genome, resulting in a correlated feature space (Additional file 7: Figure S4). Therefore, to avoid overfitting [51], we implemented recursive feature elimination (RFE) to select only the most influential TFBS across all autosomal chromosomes. We obtained near-optimal performance using approximately eight TFBS (Figure 6B; Additional file 6: 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 6C; Additional file 6: Figure S3C). We optimized our model by only considering these top four TFBS when building the random forest classifier, 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 [25–30].

2.4 Discussion

We present a machine learning approach for optimal prediction of TAD or loop boundary regions from functional genomic annotations. Our method 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. We first optimized our RF model by systematically comparing different combinations of genome binning (resolution), feature engineering procedures, and resampling techniques. These methods are implemented in part as an R package, *preciseTAD*, and pre-trained models are available as an ExperimentHub pacakges, *preciseTADhub*, both of which are available on Bioconductor.

During preliminary research we investigated the performance of several different machine learning algorithms including multiple logistic regression (MLR), l_1 and l_2 regularized logistic regression (elastic-net; glmnet version 4.0.0), support vector machines (SVM; e1071 version 1.7.3), gradient boosting machines (GBM; gbm version 2.1.5), and extreme gradient boosting (XGBOOST; xgboost version 1.0.0.2). In each case, the holdout chromosome framework with 3-fold cross-validation was used. Additionally, we considered Arrowhead ground truth TAD boundaries on GM12878, 25 kb genomic bins, distance-type features, and random-undersampling. For elastic-net we tuned the α and λ parameters over values ranging from 0.1-1 and 0-10, respectively. For SVM, we considered a linear kernal, while the cost parameter was tuned over values ranging from 0.25-4. For GBM, $shrinkage$ was set to 1, $n.minobsinnode$ was set to 10, while $interaction.depth$ and $n.trees$ were tuned over values $\{1,2,3,4,5\}$ and $\{50,100,150,200,250\}$ respectively. For XGBOOST, $nrounds$ was set to 50, while $colsampl.bytree$ and $subsample$ were set to 0.8, the learning rate was tuned over values $\{0.025,0.05,0.1,0.3\}$, $max.depth$ was tuned over values $\{2,4,6,8\}$, and $gamma$ was tuned over values $\{0,.3,.5\}$. These models were compared to RF models with $ntree$ and $nodesize$ set to 50 and $.01 \times nrow(data)$, while tuning over 10 values for the $mtry$ parameter. Results indicated

that, on average, RF out-performed all other predictive models (Additional file 16: Figure S10). We were not surprised that RF exhibited greater predictive performance than linear based approaches including MLR, ENET, and SVM, as has been seen in multiple comparative studies in the bioinformatics literature [52–54]. An additional benefit offered by RF is the availability of an interpretable ranking of predictors using variable importance measures [55,56]. A much more comparative performance was seen between bagging- (RF) and boosting-based (GBM & XGBOOST) approaches. There are other potential benefits aside from increased performance of RF. RF has fewer hyperparameters to tune, generates trees very rapidly, and does not need to make and store new training sets, saving time and memory over other methods, making RF the best choice for our purposes [57].

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 [41,52–54]. We further demonstrated that class imbalance hinders boundary prediction, but can be effectively addressed by a simple random under-sampling (RUS) technique, an aspect of boundary prediction unaddressed in previous studies [41,52,53]. We find that random over-sampling (ROS) performed quite poorly compared to the other re-sampling techniques, likely due to models overfitting the data as a result of duplicated minority class samples [55]. Additionally, we found that SMOTE's synthetic observations created from the minority class did not lead to the out-performance of RUS, indicating some residual effects of overfitting. Likewise, instead of creating perfecting balanced classes, some more calibration might be needed between the percentage of over-sampling and under-sampling offered by the algorithm.

We showed that information about only four transcription factors (CTCF, SMC3, RAD21, ZNF143) is necessary and sufficient for accurate TAD and loop boundary region prediction,

outperforming histone modification- and BroadHMM-built models [52,53]. 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 [25–30]. 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 boundary regions.

We opted to only tune the *mtry* hyperparameter in our machine learning framework. This is because the other notable hyperparameters in random forests are not tunable in the classical sense, including *ntrees* and *nodesize* [56–58]. For *ntrees*, evidence suggests that the biggest performance gain is often be achieved after growing the first 100 trees [57,58]. Thus, we were comfortable with the default ntree=500 advised in the *randomForest* R package. For *nodesize*, computation time decreases approximately exponentially with increasing node size [59]. Therefore, we set the default value to 0.1% of the rowise dimension of the training data.

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 8: Table S3). 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.

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 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.

3. Chapter 3: Aim 2 - Develop a density-based partitioning technique for precise boundary prediction at base-level resolution

3.1 Introduction

The introduction of high-throughput chromosome conformation capture sequencing (Hi-C) technologies have allowed researchers to analyze the spatial organization of the human genome. Studies have uncovered non-random 3-dimensional (3D) structures formed by folded genomic DNA [2,17,60]. Among these structures are chromatin loops and topologically associating domains (TADs). Chromatin loops form at kilobase (kb) scale as a result of distal promotors coming into contact with regulatory elements, such as enhancers [4,5]. TADs are higher-order structures that form at megabase (Mb) scale and are characterized by genomic loci with highly self-interacting DNA within a region compared to between regions [6,8,10]. TADs and loops have been reported as being highly conserved across species and cell lines [7,9–11]. The formation of these structural domains has been implicated in cell differentiation and development [61,62]. Importantly, it has been shown that disrupting the boundaries that demarcate both TADs and chromatin loops has been associated with developmental diseases [14–16] and cancer [12,13]. While some important functions of TADs and loops have been identified, their role in the 3D genome remains to be fully understood.

Many different algorithms have been proposed to identify the boundaries of TADs and chromatin loops, referred to as domain boundaries [6,22–24,50,63,64]. However, initial assessments have shown that results vary widely between methods [19–21]. These are due to several impacting factors including the algorithm of choice and Hi-C data resolution. Resolution refers to the size of genomic regions (bins) used to segment the linear genome between which contacts are enumerated [65]. Typical Hi-C experiments are performed in the 10 kb-100 kb resolution range, with higher resolutions necessary to detect hierarchical TADs and loops.

However, it is unclear if higher resolutions lead to improved domain identification due to sparsity and high-dimensionality introduced in Hi-C contact matrices as a result [65].

In contrast to Hi-C data, chromatin immunoprecipitation followed by high-throughput DNA sequencing (ChIP-seq) is performed at much higher resolution, typically on the order of tens to hundreds of bases [32]. Likewise, it has been shown that TADs and chromatin loops are mediated by sets of architectural proteins that colocalize at boundaries. Notably, domain boundaries were enriched in CTCF and cohesin complex (RAD21 and SMC proteins), components of the loop extrusion model [25–30]. Therefore, these enrichment patterns suggest that computational predictions may allow researchers to circumvent the costly resolution restrictions of Hi-C sequencing.

To this end, we have developed *preciseTAD*, a data-driven algorithm for precise domain boundary prediction using key ChIP-seq genomic annotations. Our method utilizes the random forest (RF) algorithm trained on high-resolution CTCF, RAD21, SMC3, and ZNF143 narrow peak sites to predict low-resolution domain boundaries. 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. Further, we show that *preciseTAD*-predicted boundaries are more conserved across cell lines. This improved precision in the domain boundary location can provide insight into the association between genomic regulators and the 3D genome organization. The methods developed are implemented in the *preciseTAD* R package and are freely available on Bioconductor at https://bioconductor.org/packages/preciseTAD.

3.2 Methods

3.2.1 Developing a boundary prediction tool at base-level resolution

To investigate whether we could alleviate the resolution limitations of conventional domain calling tools, we developed *preciseTAD*. This algorithm leverages an optimized random forest model in conjunction with density-based and partitioning techniques to predict boundaries at base-level resolution (Figure 7).

1. Consider an optimized RF model (M) built on the set of autosomal chromosomes $\{k|i \notin k\}$ binned at some resolution r
2. **for** *each chr* i **do**
    3. Construct the base-level resolution predictor space $A_{n\times p}$ where n is the length of chr i and p is the number of predictors
    4. Assign threshold $\{t|0 \leq t \leq 1\}$ and $\{\epsilon|\epsilon > 0\}$
    5. **if** $|t| > 1$ *or* $|\epsilon| > 1$ **then**
        6. **for** *each combination (l) of* t *and* ϵ **do**
            7. Evaluate M on $A_{n\times p}$ to get the probability of each genomic coordinate as being a domain boundary π_n
            8. Subset $\{\pi_n|\pi_n \geq t_l\}$
            9. Construct the pairwise distance matrix D between genomic coordinates where $\pi_n \geq t_l$
            10. Apply DBSCAN on D with $MinPts = 3$ and $eps = \epsilon_l$
            11. **for** *each cluster* k *identified by DBSCAN* **do**
                12. Assign w_k as the number of coordinates that span each cluster of bases in k (PTBR)
                13. Perform PAM on the sub-distance matrix D_k to extract the cluster medoid b_k (PTBP)
                14. **for** *each predictor* p **do**
                    15. Calculate the normalized enrichment (NE) over all predictors

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

                    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
                    16. Determine where NE converges as optimal $\{t, \epsilon\}$ combination
                end
            end
        end
        17. Repeat steps 7-14 on A_{nxp} with optimal $\{t, \epsilon\}$
    else
        18. Perform steps 7-14 on A_{nxp} such that $t = t_0$ and $eps = \epsilon_0$
    end
end

Algorithm 1: Psuedocode for *preciseTAD* implementation.

Figure 7. The preciseTAD algorithm.

First, a random forest classification model, M, is built on cell line-specific CTCF, RAD21, SMC3, and ZNF143 sites, for a set of binned chromosomes $\{k|k \neq i\}$, using $(log_2 + 1)$ genomic distances, ground truth TAD and loop boundaries called from Arrowhead and Peakachu at 5 kb

and 10 kb resolutions respectively, with random under-sampling (See aim 1). A *base-level resolution* predictor space, $A_{n\times p}$, is then constructed for chromosome i, where n is the length of chromosome i and p is the number of annotations. We 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) 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 $1 + dim(data)$ [66]. 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 $e_{ks} = \mathrm{I}\{r_s \in (b_k - f, b_k + f)\}$ 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 9: Figure S5).

3.2.2 Methods for summarizing predicted boundaries and regions

We devised a series of 6 summary measures to assess the quality of our predicted boundaries and the regions of clustered base coordinates that flanked them (Figure 8). The measures included: PTBRWidth - the width spanned by each cluster of bases such that $\pi_n \geq t$, PTBRCoverage - the ratio of base-level coordinates with probabilities that exceed the threshold to PTBRWidth, DistanceBetweenPTBR - the genomic distance between the end of the previous PTBR and the start of the subsequent PTBR, NumSubRegions - the number of elements in each PTBR cluster, SubRegionWidth - the genomic coordinates spanning the subregion associated with each PTBR, and DistBetweenSubRegions - the genomic distance between the end of the previous PTBR-specific region and the start of the subsequent PTBR-specific region.

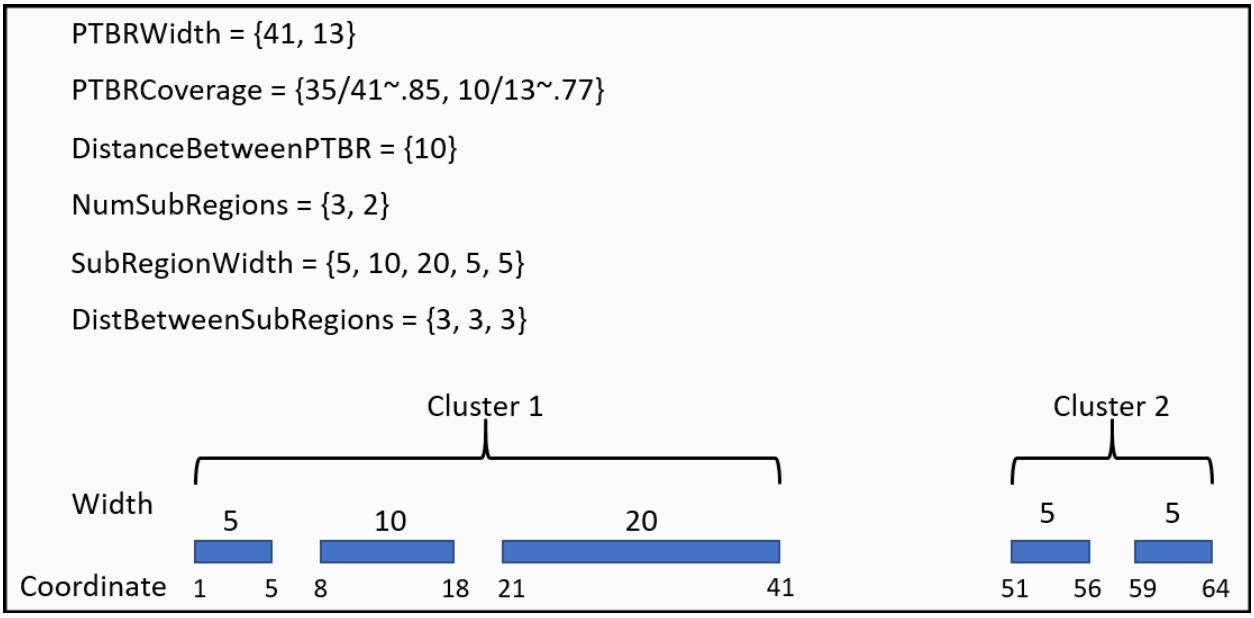


Figure 8. A schematic illustrating how each of the diagnostic summaries are calculated in the preciseTAD algorithm. *The illustration depicts blue regions as collections of base coordinates whose predictive probability exceeds a predefined threshold, t, organized into two clusters. The summary statistics include the following: PTBRWidth - PTBR width, PTBRCoverage - the ratio of base-level coordinates with probabilities that exceed the threshold*

to PTBRWidth, DistanceBetweenPTBR - the genomic distance between the end of the previous PTBR and the start of the subsequent PTBR, NumSubRegions - the number of elements in each PTBR cluster, SubRegionWidth - the genomic coordinates spanning the subregion associated with each PTBR, and DistBetweenSubRegions - the genomic distance between the end of the previous PTBR-specific region and the start of the subsequent PTBR-specific region.

3.2.3 Evaluating signal strength of known molecular drivers of 3D chromatin around predicted vs. called boundaries

We assessed the biological significance of our predicted boundaries by their association with the signal of CTCF, RAD21, SMC3, and ZNF143 using signal profiles and enriched heatmaps from *deepTools* (version 2.0) [67]. To do so, a matrix $M_{r\times c}$, is created where the rows, r, are given by the number of boundaries, either called or predicted. The column dimension, c, is created by flanking each boundary by 5 kb (10 kb for chromatin loop boundaries). The flanking is then broken up into 50 bp segments, for 100 windows on both sides of a given boundary, a total of 200 columns. The cells of the matrix are calculated as a mean coverage value for each window with respect to the signal from the respective ChIP-seq annotation, given by

$$v_c = \frac{\Sigma_j^m \Sigma_i^n x_{ij} w_{ij}}{L}$$

where x_{ij} is the total number of bases of annotation i in window j and w_{ij} is the number of bases that overlap between annotation i and window j. The denominator, L, is the width of the windows (here, $L = 50$). For the profilePlot, the matrix is then summarized by row-wise averages and plotted as a density curve, where the center represents the boundary, and the curve represents the average ChIP-seq peak signal around a flanked region. For the enriched heatmap, the matrix is plotted as a heatmap. Here, the rows of the matrix are first ordered by enriched scores calculated as the sum of coverage values weighted by the distance to the flanked boundary, denoted by

$$s_e = \Sigma_{d=1}^{n_1}\left(\frac{x_{1d} * d}{n_1}\right) + \Sigma_{u=1}^{n_2}\left(\frac{x_{2u} * (n_2 - u + 1)}{n_2}\right)$$

where n_1 and n_2 represent the number of downstream and upstream coverage values to sum over, and $n_1 = n_2 = 100$. Additionally, we compared the median log_2 genomic distances between boundaries and the same top predictive ChIP-seq annotations using Wilcoxon Rank-Sum tests.

3.2.4 Evaluating conservation of predicted vs. called boundaries

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. 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.

3.3 Results

We developed *preciseTAD*, a data-driven algorithm for precise domain boundary prediction using high-resolution ChIP-seq data. Our method employs density-based clustering (DBSCAN) and partitioning around medoids (PAM) to detect genome annotation-guided boundary regions and points at a base-level resolution (see Methods; Figure 7). This approach circumvents

resolution restrictions of Hi-C data, allowing for the precise detection of biologically meaningful boundaries.

3.3.1 *preciseTAD* better reflects intra-chromosomal contacts

When trained using Arrowhead and Peakachu ground truth boundaries at 5 kb and 10 kb resolutions, respectively, *preciseTAD* 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 10: Table S4). We reported less predicted TAD boundaries at 5 kb than Arrowhead on both cell lines (Table 2, Additional file 10: Table S4). 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 9A) and signal of known drivers of domain boundaries (Figure 9B). *preciseTAD* also predicted fewer chromatin loop boundaries than Peakachu (Table 2, Additional file 10: Table S4). 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. In addition to predicting boundary locations, *preciseTAD* provides collections of base coordinates that exhibit high levels of predictability, these are termed *preciseTAD boundary regions* (PTBRs). Our preliminary observations indicated that, under most optimal settings, the width of PTBRs paralleled the resolution of Hi-C data (Table 3; Additional file 11: Table S5).

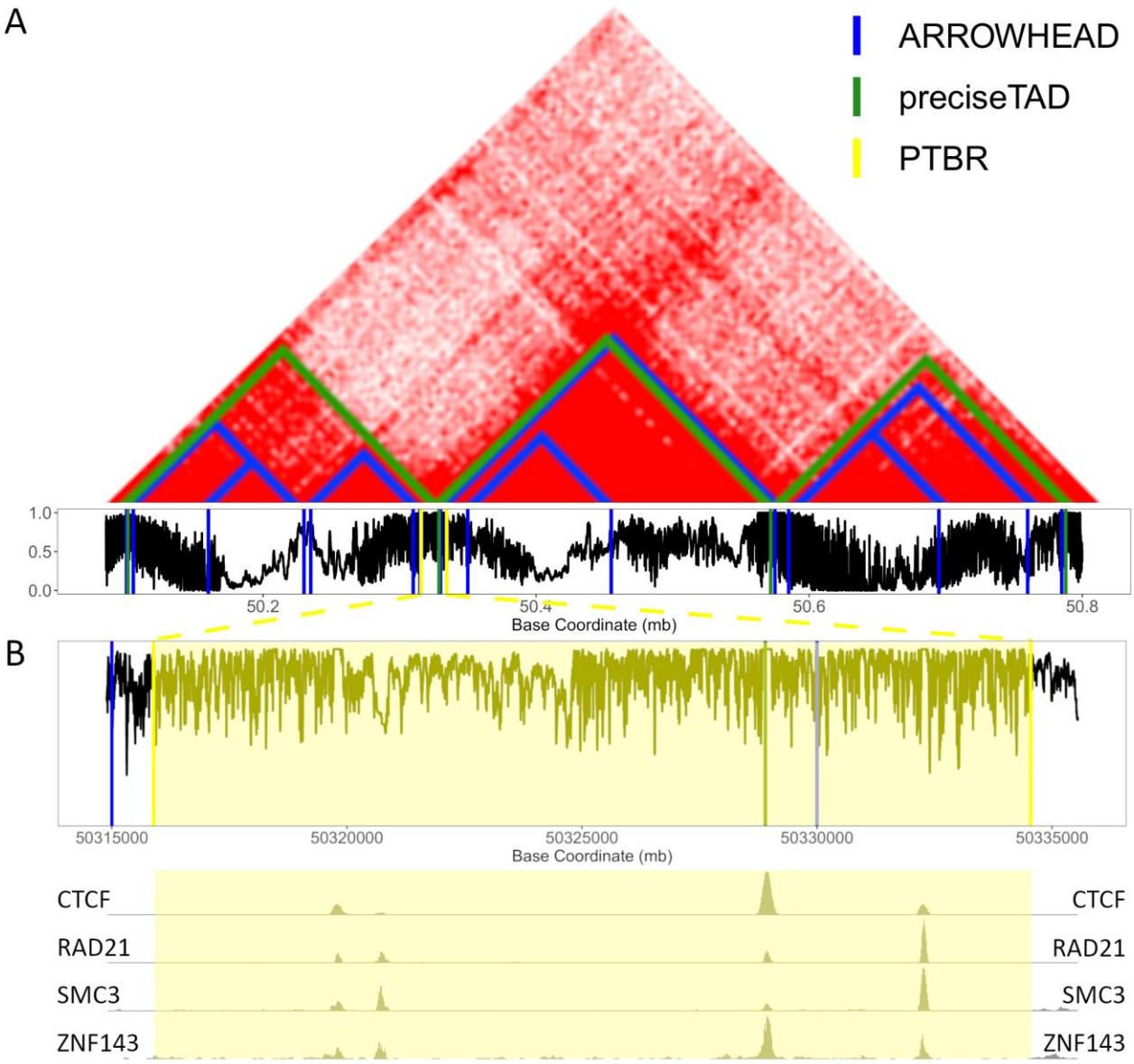


Figure 9. preciseTAD-predicted boundaries better reflect intra-chromosomal contacts. *(A) The location of Arrowhead-called TAD boundaries (blue) vs. preciseTAD-predicted TAD boundaries (green) on GM12878 data (chr14:50085000-50800000). The black line represents the predicted probability of each base 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 corresponding signal profiles of CTCF, RAD21, SMC3, and ZNF143.*

Summary	Predicted TAD boundaries	Predicted loop boundaries
PTBRWidth	13131.4 (10927.7)	14610.2 (10857.7)
PTBRCoverage	0.2 (0.3)	0.1 (0.2)
DistanceBetweenPTBR	205023.8 (440221.8)	153085.4 (344241.3)
NumSubRegions	23.7 (19.6)	193.9 (198.3)
SubRegionWidth	11.0 (30.0)	4.7 (11.0)
DistBetweenSubRegions	572.4 (1191.5)	73.1 (308.3)

Table 3: Summary measures evaluating the quality of *preciseTAD*-predicted TAD and chromatin loop boundaries for GM12878. Summaries are reported as means (standard deviations).

3.3.2 *preciseTAD* identifies precise and biologically relevant domain boundaries

Next, we evaluated the biological significance of *preciseTAD* boundary points (PTBPs). The signal of four known molecular drivers of 3D chromatin (CTCF, RAD21, SMC3, and ZNF143) colocalized more frequently around PTBPs, as compared to Arrowhead-called TAD and Peakachu loop boundaries, respectively (Figure 10A, 10B; Additional file 12: Figure S6A, S6B). Surprised by the poor signal distribution around Arrowhead boundaries, we compared signals centered on boundaries called by Arrowhead, Peakachu, and a recently published TAD-caller, SpectralTAD [64]. We confirmed the poor signal distribution around Arrowhead boundaries, in contrast to the relatively well-performing Peakachu- and SpectralTAD-called boundaries (Additional file 12: Figure S6C, S6D). 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 13-14: Figure S7-S8). *preciseTAD* boundaries were statistically significantly closer to the top-ranked TFBS (Wilcoxon p-value < 0.001 versus Arrowhead and Peakachu boundaries, respectively, Figure 11A, Additional file 15:

Figure S9). These results indicate that *preciseTAD*-predicted boundaries better reflected the known biology of boundary formation.

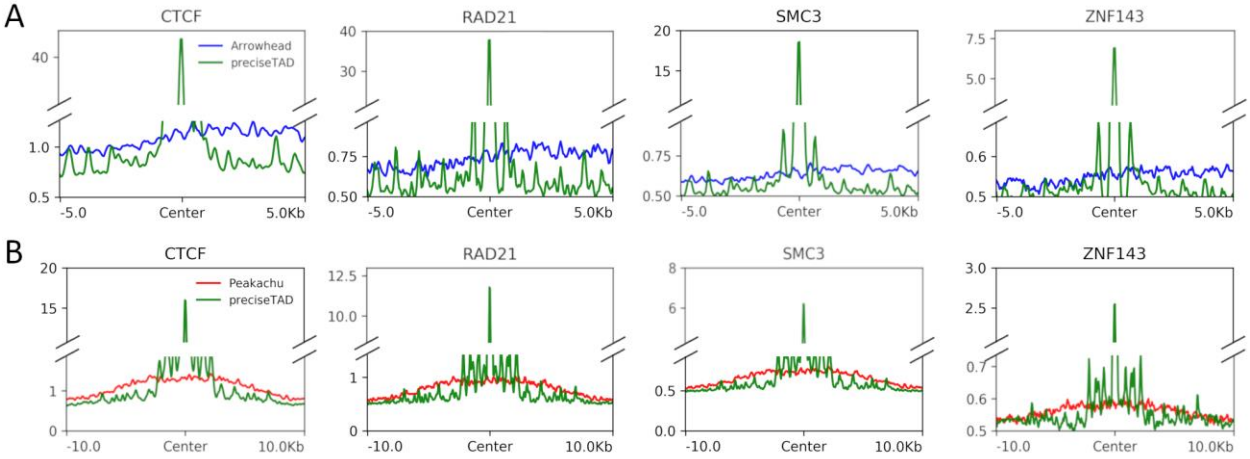


Figure 10. preciseTAD-predicted boundaries are enriched for known molecular drivers of 3D chromatin. *Signal profile plots comparing the strength of CTCF, RAD21, SMC3, and ZNF143 binding around Arrowhead-called boundaries (blue, C), Peakachu loop boundaries (red, D) vs. preciseTAD-predicted boundaries (green).*

3.3.3 *preciseTAD* boundaries are more conserved across cell lines

Previous studies suggest that TAD boundaries are conserved across cell lines [6–8,68]. 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 11B, 11C). 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 11D, 11E). The better conservation of *preciseTAD*-predicted boundaries further supports the notion of their higher biological relevance.

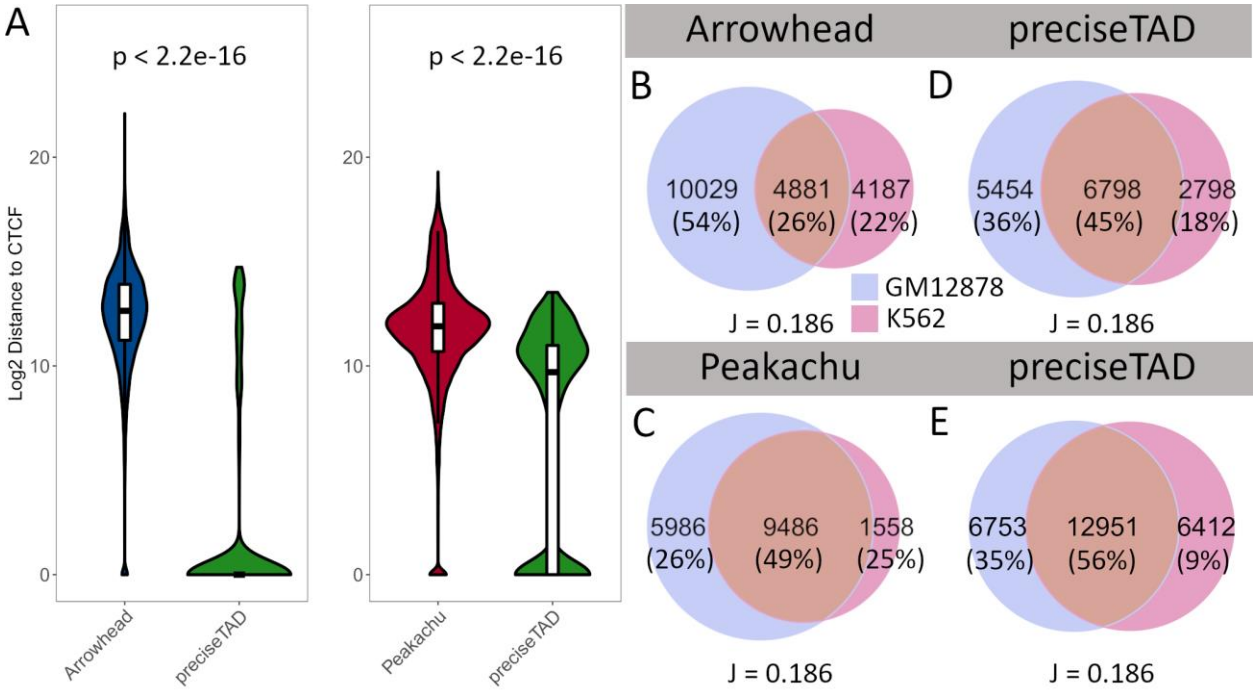


Figure 11. preciseTAD-predicted boundaries are closer to CTCF sites and more conserved across cell lines. *(A) log_2 genomic distance distribution from called and predicted boundaries to the nearest CTCF sites. The p-values are from the Wilcoxon Rank Sum test. (B-E) Venn diagrams illustrating the levels of conservation (overlap) between domain boundaries for GM12878 (red) and K562 (blue) cell lines identified by Arrowhead (B), Peakachu (C), and preciseTAD-predicted boundaries using (D) Arrowhead- and (E) Peakachu-trained models. Boundaries involving Arrowhead/Peakachu were flanked by 5 kb/10 kb, respectively.*

3.3.4 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. We found that 88.8% and 95.8% of our predicted TAD boundaries were overlapped by predicted loop boundaries for GM12878 and K562, respectively (Figure 13). 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. 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.

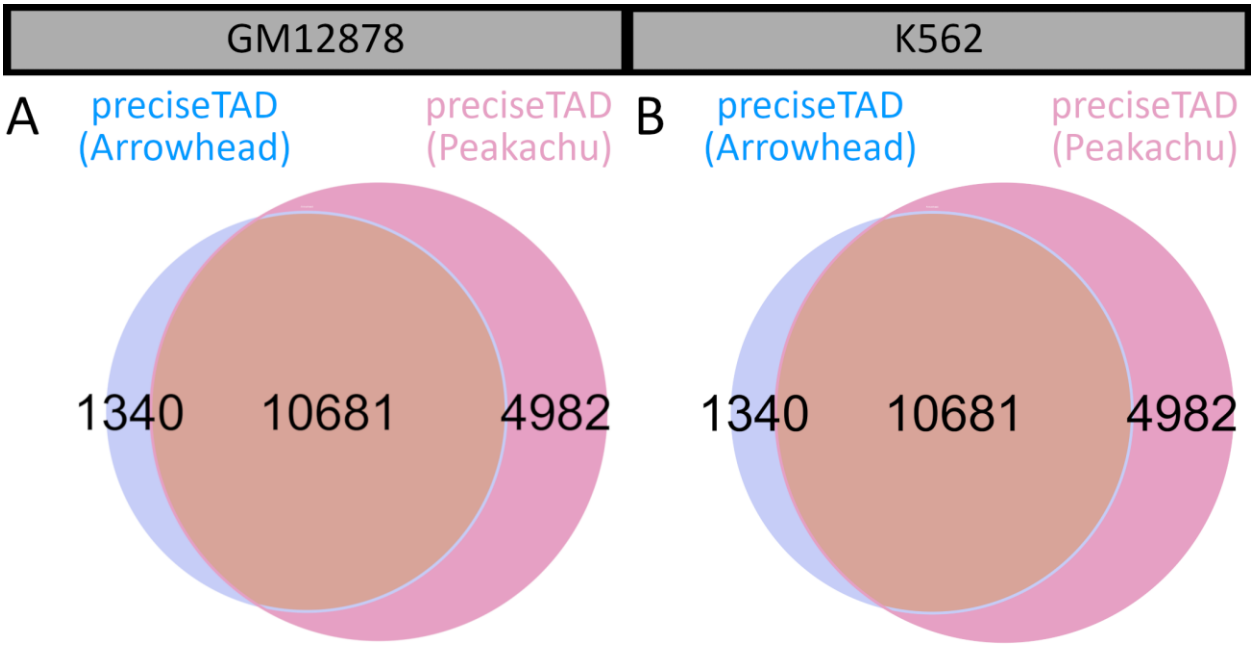


Figure 12. The agreement between preciseTAD-predicted boundaries using Arrowhead- and Peakachu-trained models. *Venn diagrams of boundary overlap using (A) GM12878 and (B) K562 data. Boundaries involving Arrowhead/Peakachu were flanked by 5 kb/10 kb, respectively.*

3.4 Discussion

We present *preciseTAD*, a data-driven approach toward domain boundary prediction. *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 among

them. *preciseTAD* was benchmarked against two boundary calling tools, Arrowhead [34], an established TAD-caller, and Peakachu [24] a recently published algorithm for predicting chromatin loops. *preciseTAD* is primarily implemented as an R package on Bioconductor (https://bioconductor.org/packages/preciseTAD/).

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 (Table 3; Additional file 11: Table S5). 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 signals (Figure 5). Our results are in line with the emergent view that domain boundaries are flexible [69,70], and their well-defined location arises as a consequence of the population average in bulk Hi-C data [27,71].

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 [25]. Additionally, *preciseTAD* is invariant to resolution, resulting in accurate boundary prediction even if implemented on low resolution ground truth boundaries. Furthermore, *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 9: Figure S5).

In summary, we demonstrate that approaching domain calling from a computational and predictive perspective can alleviate resolution restrictions from conventional TAD/loop callers and improve boundary precision. Our method, *preciseTAD*, leverages a random forest classification model built on high-resolution genome annotation data, in addition to density-based clustering (DBSCAN) and partitioning around medoids (PAM) to predict biologically meaningful TAD and loop boundaries. *preciseTAD* is available as an open source R package on Bioconductor. We hope that *preciseTAD* will serve as an efficient and easy-to-use tool to further explore the genome's 3D organization.

4. Chapter 4: Aim 3 - Develop a technique for predicting boundaries on cell lines that do not have publicly available Hi-C data

4.1 Introduction

The 3-dimensional (3D) chromatin architecture of the human genome plays a critical role in cellular homeostasis and gene regulation [9,17,33]. High-throughput sequencing of long-range interactions (Hi-C) in multiple cell lines has revealed that the CCCTC-binding factor (CTCF) and other protein members of cohesin (RAD21 and SMC3) are enriched at boundaries of chromatin loops and topologically associating domains (TADs), suggesting a regulatory role [72,73]. Mechanistically, many CTCF-mediated chromatin loops define insulated neighborhoods that constrain promoter-enhancer interactions within the same TAD [10,74]. Likewise, disruption of individual CTCF-binding sites deregulates the expression of surrounding genes [12].

The CTCF- and cohesin-mediated interaction network has been considered to be largely invariant across cell lines [75]. CTCF- and cohesin-binding sites can be mapped at high resolution 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 [32], well below the resolution of Hi-C data (tens of kilobases) [33]. Despite technical advances, chromosome conformation capture sequencing remains difficult and costly, and few cell lines have been analyzed at high resolutions. However, members of the International Human Epigenome Consortium (IHEC) including ENCODE, NIH Roadmap Epigenomics, FANTOM5, and BLUEPRINT have been actively cataloging cell line-specific genome annotation datasets. Therefore, computational predictions that take advantage of the routinely available ChIP-seq data is a desirable approach to guide the systematic analysis of the 3D structure of the human genome.

Here, we develop a novel technique for leveraging a machine learning model built on cell-line-specific functional genomic annotations to predict the precise locations of 3D domain boundaries across cell lines. Our method relies only on 4 transcription factor binding sites including CTCF, RAD21, SMC3, and ZNF143. We found that prediction performance between same-cell-line models and cross-cell-line models was not significantly different. Furthermore, predicted boundaries made across cell lines exhibited strong levels of conservation when compared to boundaries predicted on same cell line genomic data. Our approach highlights the opportunity for alleviating the costly and imprecise reliance on high-resolution Hi-C sequencing. Moreover, we envision the broader availability of cell line-specific genomic annotations will enable a more systematic analysis of domain boundaries using our method.

4.2 Methods

4.2.1 Framework for training and testing boundary region models across cell lines

As a baseline, we first built and evaluated domain boundary region random forest (RF) classification models using called boundaries from Arrowhead and Peakachu at 5 kb resolution for one cell line. Random under-sampling was implemented to balance the data. Distance-type features were considered for only the top 4 transcription factors including CTCF, RAD21, SMC3, and ZNF143 from the same cell line (i.e., the optimal combination of data level characteristics from Chapter 2). The same holdout chromosome strategy was used for training and testing (Figure 4). 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.

For cross-cell-line training and testing, we adopted a similar strategy as above, except the training set was constructed from another cell line. This included both the ground truth domain boundaries (**Y**) and the distance-type feature space for the $n-1$ chromosomes. The same testing set was used as above. 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.

4.2.2 Evaluating model performance across cell lines

To evaluate performance, we constructed receiver operating characteristic (ROC) curves composed of the average sensitivities and specificities at different cutoffs, across each holdout chromosome. For each holdout chromosome, sensitivities and specificities were obtained at 502 equally spaced cutoff values ranging from 0 to 1 by increments of 0.002 based on model based predicted probabilities for TAD boundary regions, using the *roc* function in the *pROC* R package (version 1.16.2), creating a $500 \times 2 \times 21$ (omitting CHR9) array. The average (and standard deviation) of sensitivies and specificities was aggregated across the holdout chromosome. We reported the corresponding average area under the curve (AUC). These were compared to performances obtained using same cell line training and testing. We compared 2 separate cases for both Arrowhead and Peakachu ground truth: (1) models trained on GM12878 and tested on GM12878 vs. models trained on K562 and tested on GM12878, (2) models trained on K562 and tested on K562 vs. models trained on GM12878 and tested on K562.

4.2.3 Predicting base-level boundaries across cell lines

We extended the strategies outlined above by applying our boundary prediction tool, *preciseTAD*, developed in Chapter 3 (Figure 7). In the cross-cell-line case, the data that the RF model (M) was built on is from a different cell line than the base-level resolution predictor space ($A_{n \times p}$). As before, this strategy was compared to applying *preciseTAD* on models and base level annotation data from the same cell line. The default parameters for *preciseTAD* were set to $t = 1$ and $\epsilon = 10000$.

4.2.4 Comparing boundary location for same cell line prediction vs. cross cell line prediction

We assessed the positional significance of boundaries predicted using the same-cell-line strategy vs. the cross-cell-line strategy using signal profiles and enriched heatmaps from *deepTools* (version 2.0). Additionally, we evaluated the overlap of flanked boundary coordinates predicted by each strategy using Venn diagrams and Jaccard Indices. Boundaries were first flanked by 5 kb and 10 kb on either side for predicted TADs and chromatin loops, respectively.

4.3 Results

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

Having demonstrated the optimal performance in our machine learning framework for boundary region prediction developed in Chapter 2, 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 14A). 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 vs. 0.874, respectively). These results were consistent when comparing training/testing strategies on K562 with training on GM12878 and testing on K562 data (Figure 14B). 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.

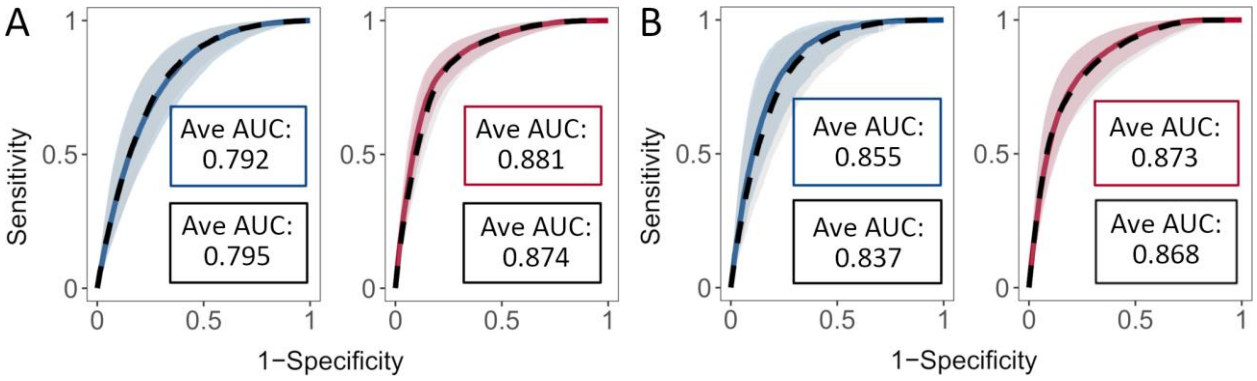


Figure 14. Training and testing across cell lines performs similarly to within the same cell line. *Receiver operating characteristic (ROC) curves and the corresponding average area under the curves (AUCs) when (A) training and testing on GM12878 data (blue, Arrowhead ground truth; red, Peakachu ground truth) versus training on K562 and testing on GM12878 data (black, dashed), and (B) training and testing on K562 data (blue, Arrowhead ground truth; red, Peakachu ground truth) versus training on GM12878 and testing on K562 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.*

4.3.2 Cell line-specific annotation data precisely predict domain boundaries across cell lines

Guided by the high predictive performance when training and testing on different cell lines, 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 using *preciseTAD*. We evaluated two 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 (Figure 15A & B). Likewise, when using Peakachu-trained models, we

observed 85% (J=0.705) and 88% (J=0.759) overlap (Figure 15C & D). 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 (Figure 16). 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.

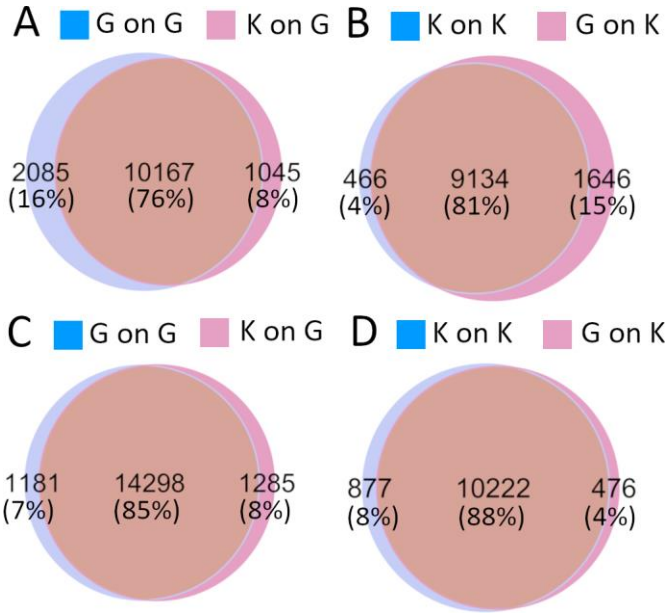


Figure 15. Cross-cell-line predicted boundaries strongly overlapped with same-cell-line predicted boundaries. *Venn diagrams comparing flanked predicted boundaries using Arrowhead (A, B) and Peakachu (C, D) trained models. (A, C) Models trained on GM12878 and predicted on GM12878 (red, GM on GM) vs. 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.*

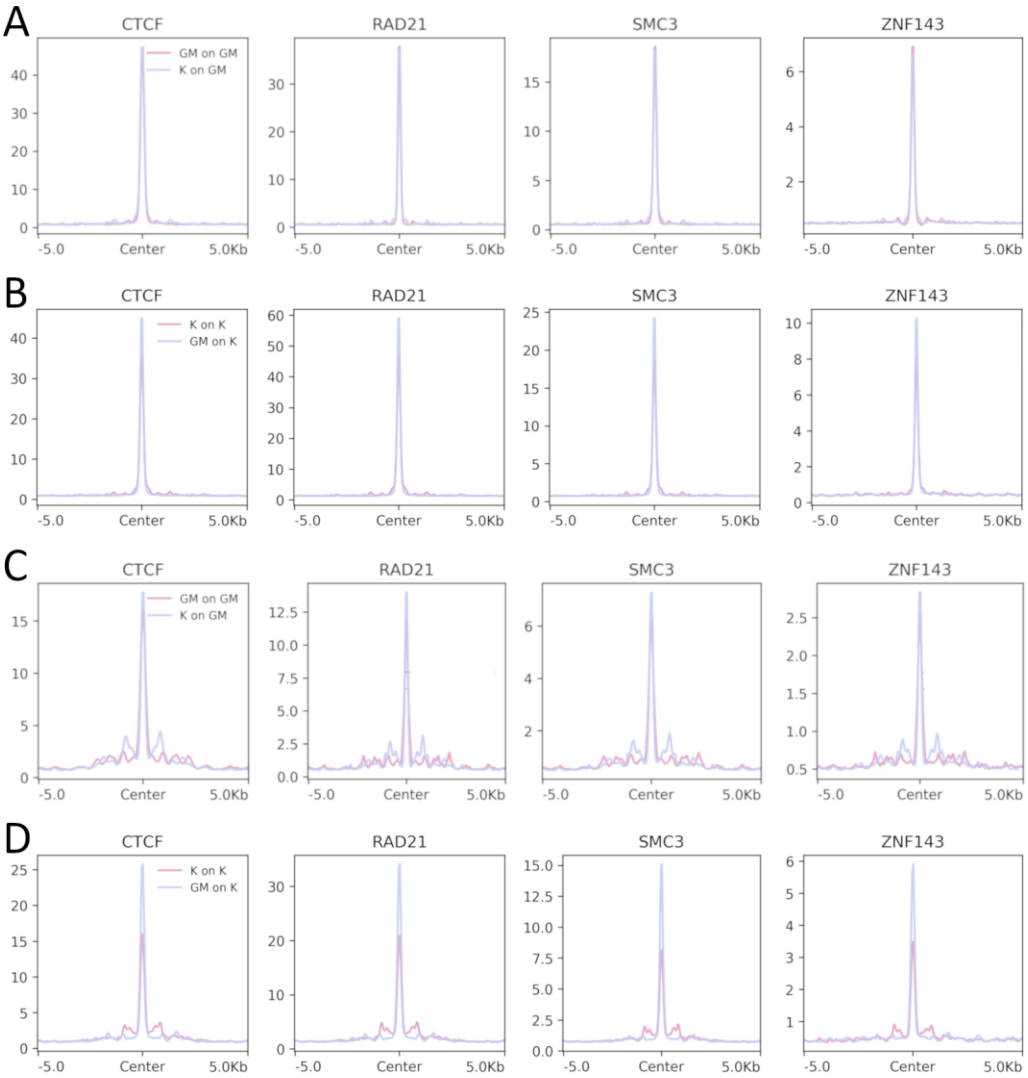


Figure 16. Cross-cell-line predicted boundaries were as enriched for known drivers of 3D chromatin as same-cell-line predicted boundaries. *Profile plots comparing enrichment levels of CTCF, RAD21, SMC3, and ZNF143 sites around flanked predicted boundaries using Arrowhead (A, B) and Peakachu (C, D) trained 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.*

4.4 Discussion

Here we show that cell-line-specific functional genomic annotation data (ChIP-seq) can be used to precisely predict domain boundaries across cell lines. Genomic annotation data, which is sequenced at much higher resolution compared to Hi-C, has been made publicly available for a variety of cell lines. Studies have shown the functional importance of key architectural proteins in organizing the 3D structure of the human genome. These include CTCF and cohesin complex, components of the loop extrusion model, whereby the genome is extruded through cohesin rings forming chromatin loops and higher-order structures such as TADs [9,17,33]. Moreover, CTCF- and cohesin-mediated interaction are largely invariant across cell lines leading to opportunities for computational predictions to aid in the annotation of the differential 3D architecture of multiple cell lines. While previous studies have utilized machine learning to draw conclusions about the most influential genomic elements associated with domain formation [40,41,44], none have extended this to be able to draw conclusions about the cell-line specificity of these binding sites and how they may be used to make predictions across cell lines.

Therefore, we developed a novel technique for leveraging high-resolution cell-line-specific CTCF, RAD21, SMC3, and ZNF143 transcription factor binding sites to predict domain boundaries across cell-lines in a supervised machine learning framework. Using the random forest classifier, we built boundary region prediction models using TAD and chromatin loop ground truth boundaries from Arrowhead and Peakachu, and ChIP-seq data from one cell line, to make predictions on another cell line. Interestingly, we found that cross-cell-line predictive models exhibited average ROC curves within 1 standard deviation of models built using the same-cell-line strategy. We then extended our framework from model performance to precision by comparing the location of *preciseTAD*-predicted boundaries between cross-cell-line versus same-cell-line strategies. We found that there were exceedingly high amounts of overlap

between flanked boundaries predicted by either strategy (>75%). Likewise, cross-cell-line predicted boundaries were found to be equally enriched for known molecular drivers of 3D chromatin as compared to their same-cell-line counterparts.

Detecting 3D domain structures from Hi-C contact matrices continues to be a costly and challenging problem with many cell lines not currently having Hi-C data publicly available. The reasons for this are due to the interplay between data resolution and the proposed TAD-calling algorithm [19–21]. Instead, we propose a method for predicting domain boundaries on one cell line using functional genomic annotations of another cell line. Thus, our method creates new opportunities for predicting domain boundaries on cell lines without using high-resolution Hi-C data for which there is none available. We have deposited pre-trained models in an ExperimentHub R package associated with Amazon Web Services (AWS), *preciseTADhub*, available on Bioconductor (https://bioconductor.org/packages/preciseTADhub/). Our hope is that as cell line-specific genomic annotations become more available, this will enable a more systematic analysis of domain boundaries using our method.

5. Chapter 5: Discussion

5.1 Conclusions and limitations

In this dissertation, we have outlined novel methods for transforming TAD- and chromatin loop-calling into a supervised machine learning framework. Our proposed methods have allowed us to bridge the gap between high-resolution 1D ChIP-seq annotations and much lower resolution 3D Hi-C sequencing data. We have developed and implemented the software associated with our methods as a publicly available R package on Bioconductor at https://bioconductor.org/packages/preciseTAD. Pre-trained models free to download are publicly available as an ExperimentHub package, *preciseTADhub*, on Bioconductor at (https://bioconductor.org/packages/preciseTADhub). Researchers will now have free and easy-to-use tools for exploring the 3D architecture of the human genome.

In Chapter 2, we introduce a novel machine learning framework for building domain boundary region prediction models. We use the random forest classification algorithm to leverage high-resolution ChIP-seq data to find the optimal set of data-level characteristics for boundary prediction. We introduced new techniques for model building and feature engineering including *shifted binning* and *distance-type predictors*. By performing a comprehensive analysis involving two cell lines and both TAD- and loop-called boundaries, we were able to conclude the mechanisms that drive loop formation are similar between TADs and chromatin loops and invariant to cell-line-specificity.

In Chapter 3, we introduce our own novel domain boundary prediction tool, *preciseTAD*. We demonstrate that *preciseTAD* is invariant toward boundary inflation. Likewise, we show that *preciseTAD*-predicted boundaries are more enriched for known molecular drivers of 3D chromatin, including CTCF, RAD21, SMC3, and ZNF143. Moreover, our predicted boundaries are more conserved across cell lines, highlighting their biological significance. *preciseTAD* has

tunable parameters t and ϵ that allow users the flexibility in the number of boundaries that are predicted, as well as the distances between them. We hope that *preciseTAD* will serve as an efficient and easy-to-use tool to further explore the genome's 3D organization.

Lastly, in Chapter 4, we present a technique for circumventing the costly and challenging task of performing high-resolution Hi-C sequencing on the vast number of different cell lines. We demonstrate that, using strategies outlined in Chapters 2 and 3, cell-line-specific functional genomic annotation data (ChIP-seq) can be used to precisely predict domain boundaries across cell lines. Our method capitalizes on the invariance of CTCF- and cohesin-mediated interactions across cell lines. We show that cross-cell-line predictive models exhibited average ROC curves within 1 standard deviation of models built using the same-cell-line strategy. Consequently, when we extended this, flanked *preciseTAD*-predicted boundaries between cross-cell-line versus same-cell-line strategies exhibited over 75% of overlap, and were equally enriched for architectural proteins. Our hope is that as the availability of cell line-specific genomic annotations increases, this will enable a more systematic analysis of domain boundaries on all cell lines using our method.

There are limitations of our proposed methods. First, our methods are dependent on the "ground truth" boundaries provided by a domain caller. Given the wide variety of domain callers and their variable performance [19,20], defining "ground truth" boundaries is challenging. Ideally, we would benchmark *preciseTAD* against simulated boundaries. While methods for simulating Hi-C data sets with boundary annotations exist [20,76], methods for simulating the associated genomic annotations (the main component guiding *preciseTAD* predictions) are lacking. Moreover, simulated Hi-C contact matrices are not performed using any specific underlying chromosomal structure. Thus, building models using both components would only capture noise. However, we feel that the concept of *shifted-binning* is suitable for capturing true signal of domain boundaries while allowing for the underlying variability among domain callers.

Ultra-deep Hi-C sequencing [33] and newer technologies for precise mapping of chromatin interactions (e.g., Micro-C [77]), 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.

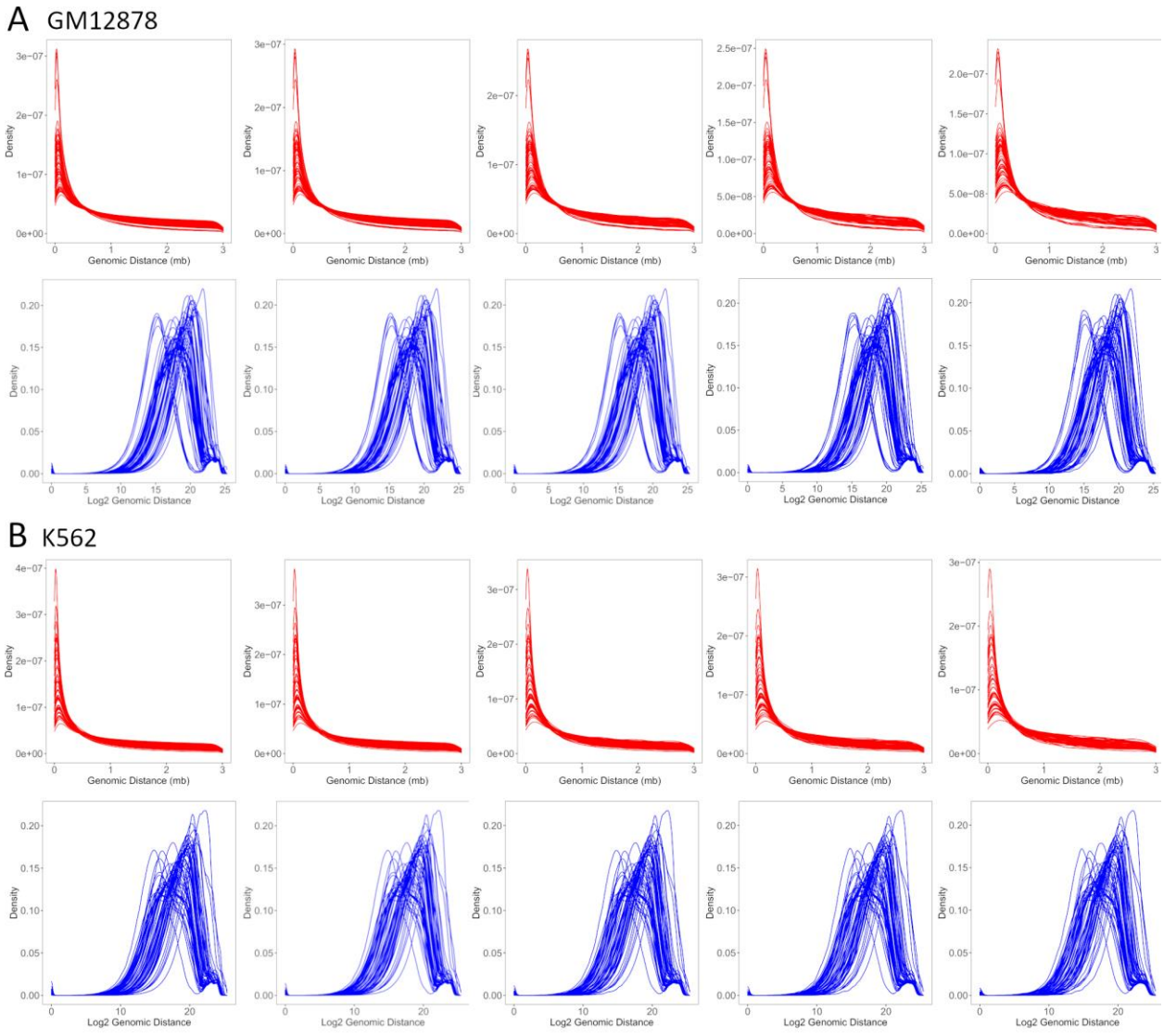
A second limitation is that our methods do not distinguish boundary types. The hierarchical nature of TAD boundaries [3,64,78] is not considered by *preciseTAD* due to the lack of gold standard of TAD hierarchy. *preciseTAD* also does not consider the directionality of CTCF binding [79] as it predicts 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 [77]. Furthermore, actively transcribed regions can serve as TAD boundaries themselves, independently of CTCF binding [79]. 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.

6. Appendix

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. Not included but, available upon request.

Additional File 2: Table S1. 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). Not included due to file size; available upon request.

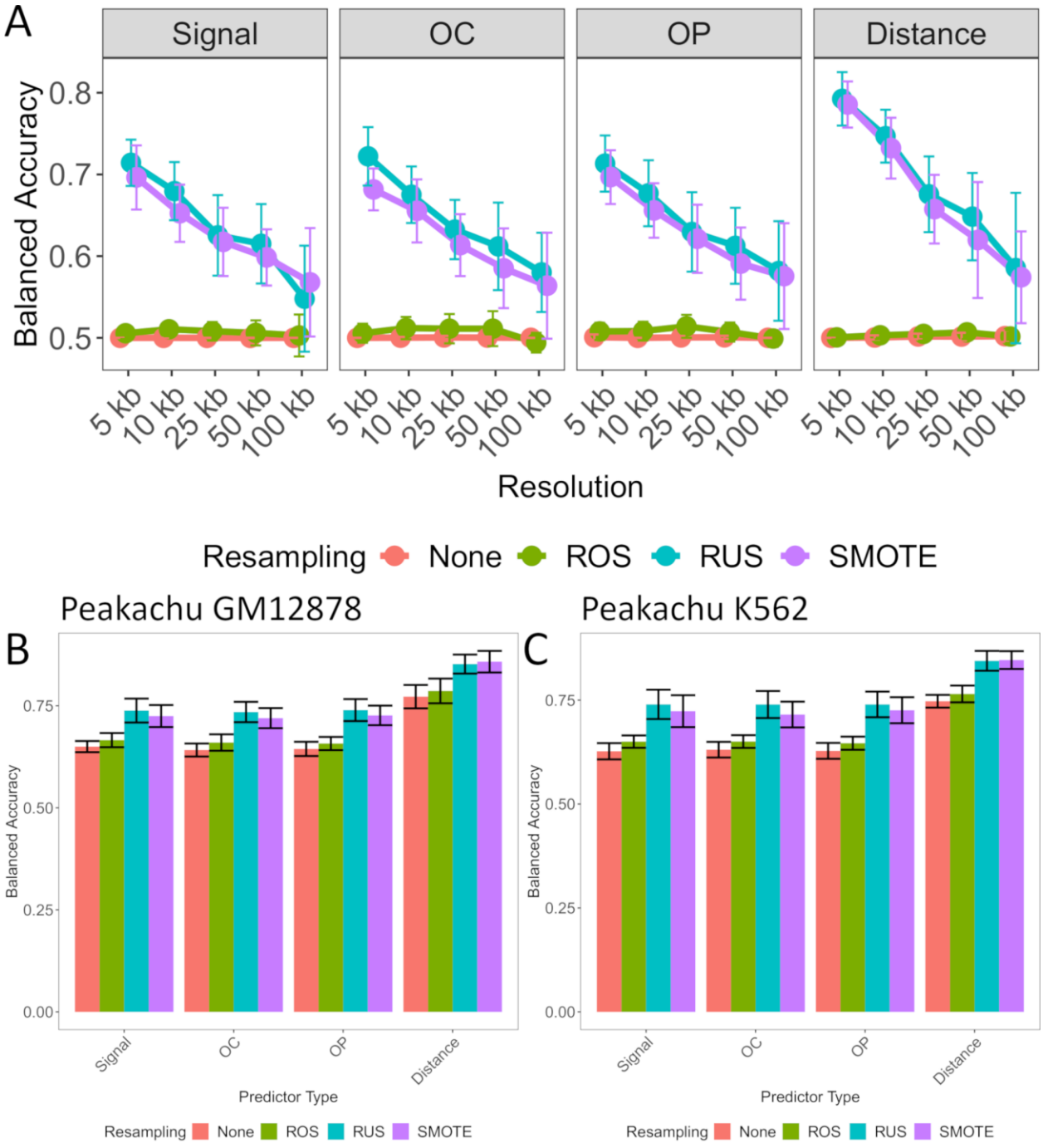
Additional File 3: Figure S1. The log_2 transformation of genomic distances normalizes their distributions. Distances are measured as the number of bases from the center of a genomic bin to the nearest genomic annotation center. Density curves of distances before (red) and after (blue) performing a log_2 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).



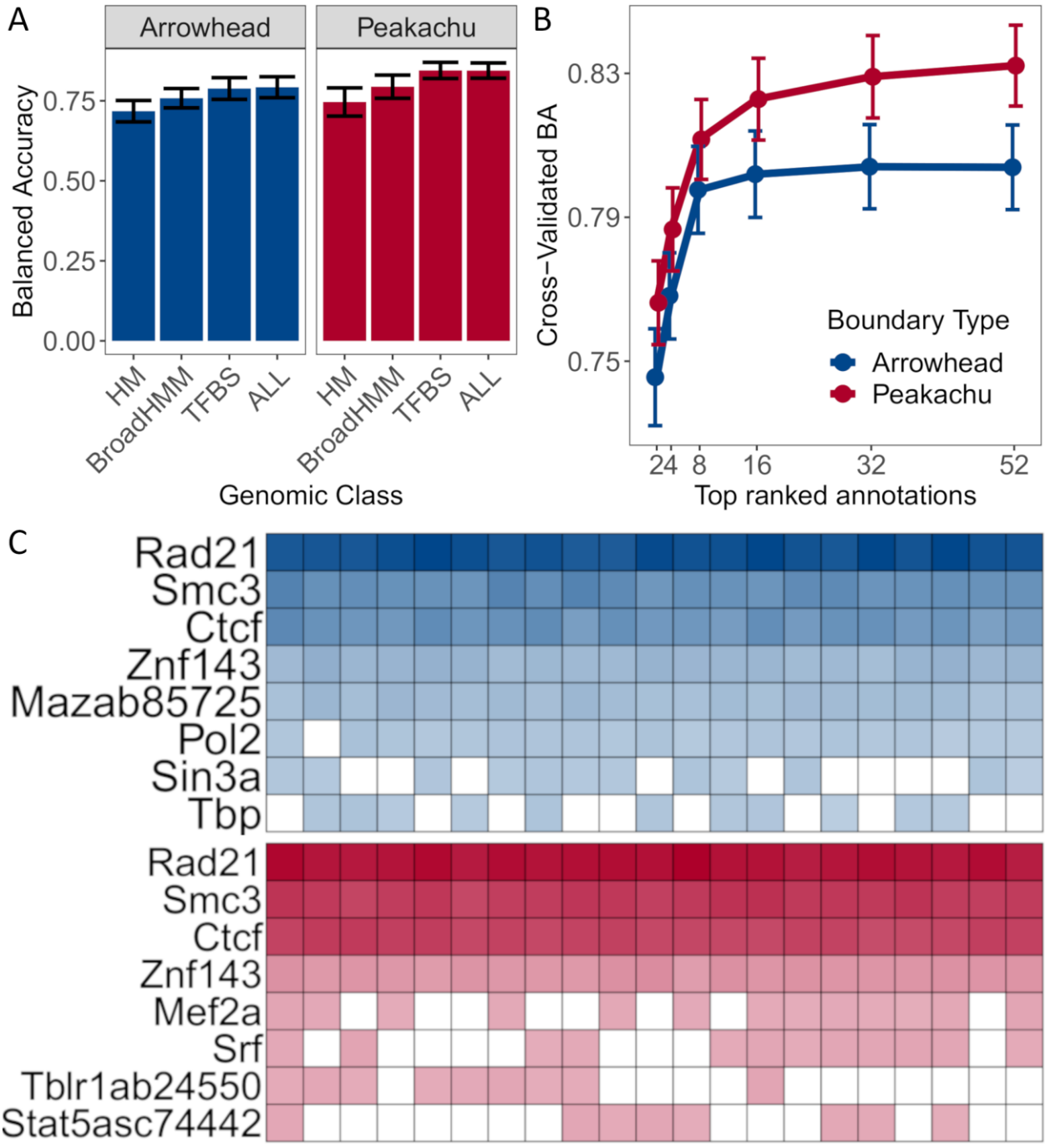
Additional File 4: Table S2. Domain boundary data and class imbalance summaries across resolutions for Arrowhead and Peakachu on K562.

Tool	Resolution/Bin size	Total number of domains	Total number of unique domain boundaries	Total number of genomic bins	Class Imbalance
Arrowhead	5 kb	4751	9316	535363	0.02
Arrowhead	10 kb	5828	10945	267682	0.04
Arrowhead	25 kb	3935	7015	107073	0.07
Arrowhead	50 kb	2115	3808	53537	0.07
Arrowhead	100 kb	945	1759	26768	0.07
Peakachu	10 kb	15651	22073	267682	0.14

Additional File 5: 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.

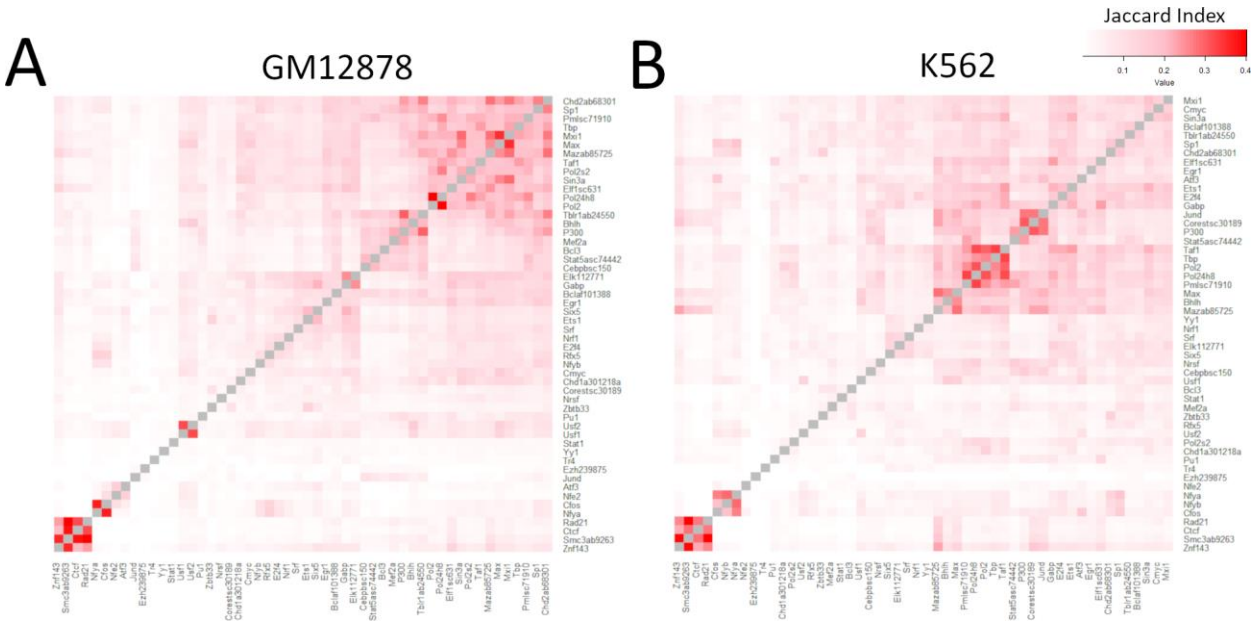


Additional File 6: Figure S3. SMC3, RAD21, CTCF, and ZNF143 transcription factors accurately predict TAD and loop boundaries in K562. (A) Barplots comparing performances of TAD (Arrowhead) and loop (Peakachu) boundary prediction models using histone modifications (HM), chromatin states (BroadHMM), transcription 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 TFBS. The columns represent the holdout chromosome excluded from the training data. Rows are sorted in decreasing order according to the columnwise average importance.



Additional File 7: Figure S4. Transcription factor binding sites are highly correlated.

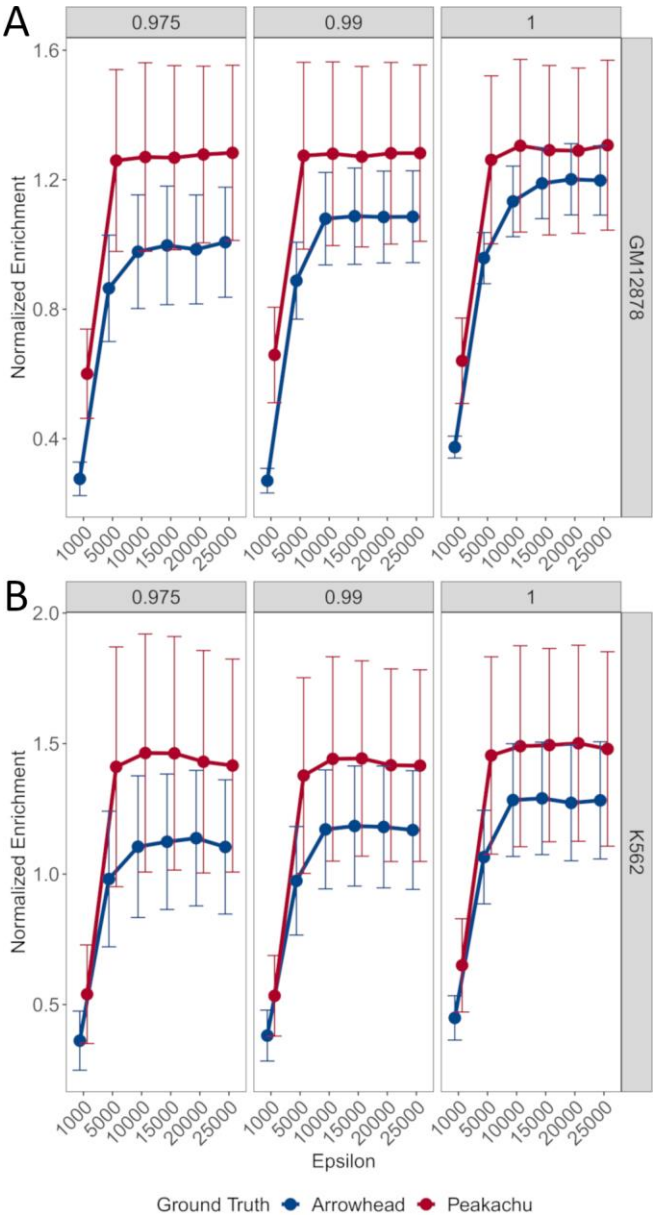
Heatmaps of Jaccard indices illustrate how colocalized cell-line specific transcription factors for (A) GM12878 and (B) K562 are on the linear genome resulting in a correlated feature space.



Additional File 8: Table S4. Additional performance metrics when implementing a random forest using Arrowhead ground truth TAD boundaries at 5 kb for GM12878. Performances are averaged across each holdout chromosome that was reserved for testing. Performances were similar for K562.

Resampling Technique	Predictor Type	Accuracy	AUC	Precision	F1Score	AUPRC
None	Signal	0.970	0.683	NA	NA	0.101
None	OC	0.970	0.649	NA	NA	0.100
None	OP	0.970	0.700	NA	NA	0.106
None	Distance	0.970	0.822	NA	NA	0.100
ROS	Signal	0.937	0.705	0.046	0.051	0.056
ROS	OC	0.875	0.692	0.040	0.061	0.050
ROS	OP	0.927	0.703	0.042	0.051	0.055
ROS	Distance	0.968	0.815	0.090	0.016	0.091
RUS	Signal	0.682	0.740	0.060	0.110	0.080
RUS	OC	0.728	0.752	0.068	0.122	0.095
RUS	OP	0.711	0.753	0.067	0.121	0.094
RUS	Distance	0.730	0.837	0.082	0.148	0.115
SMOTE	Signal	0.727	0.727	0.060	0.109	0.065
SMOTE	OC	0.799	0.742	0.077	0.135	0.071
SMOTE	OP	0.753	0.735	0.066	0.118	0.070
SMOTE	Distance	0.771	0.830	0.087	0.154	0.102

Additional File 9: Figure S5. 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.

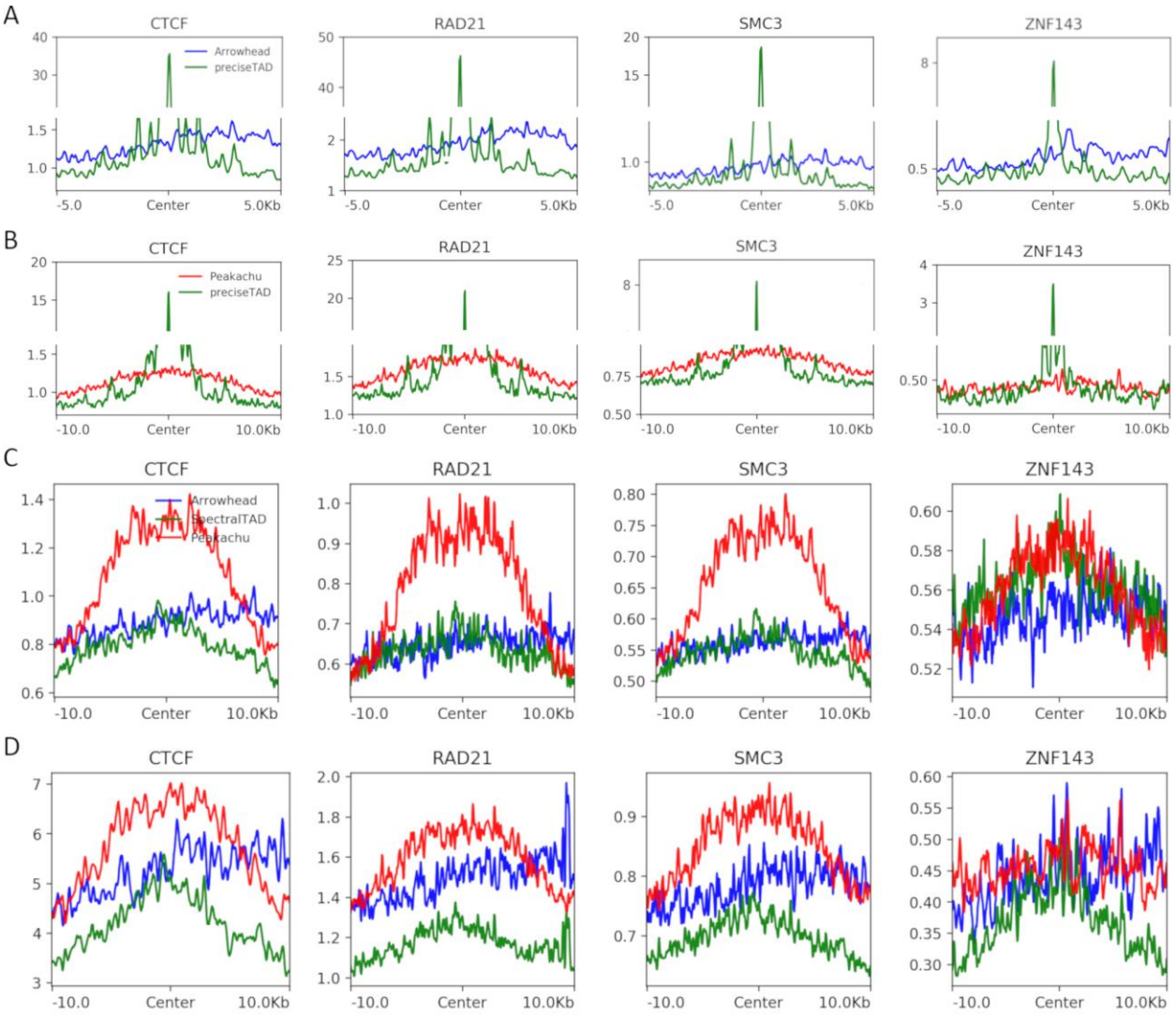


Additional file 10: Table S4. hg19/GRCh37 genomic coordinates of *preciseTAD*-predicted boundary regions (PTBR) and points (PTBP) for GM12878 and K562 cell lines, using models trained on Arrowhead TAD and Peakachu chromatin loop boundaries as ground truth. For PTBRs, the start and end coordinates define 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. Not included due to file size; available upon request.

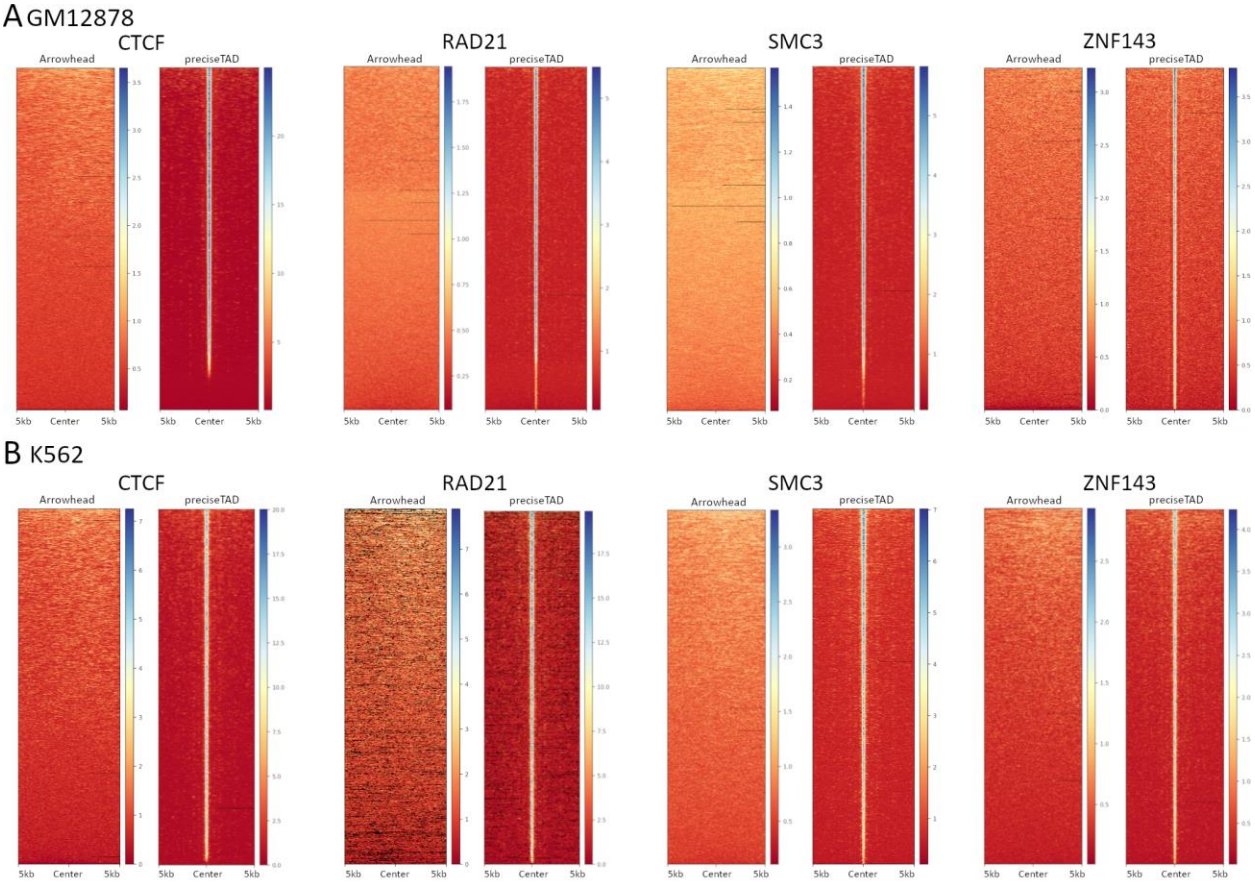
Additional file 11: Table S5. Summary measures evaluating the quality of *preciseTAD*-predicted TAD and chromatin loop boundaries for K562. Summaries are reported as means (standard deviations).

Summary	Predicted TAD boundaries	Predicted loop boundaries
PTBRWidth	14452.1 (9230.3)	17964.6 (9989.5)
PTBRCoverage	0.1 (0.2)	0.1 (0.1)
DistanceBetweenPTBR	289559.6 (641839.6)	231956.8 (543192.5)
NumSubRegions	44.3 (30.4)	216.8 (167.5)
SubRegionWidth	11.3 (30.9)	5.5 (14.9)
DistBetweenSubRegions	326.1 (800.8)	79.3 (287.9)

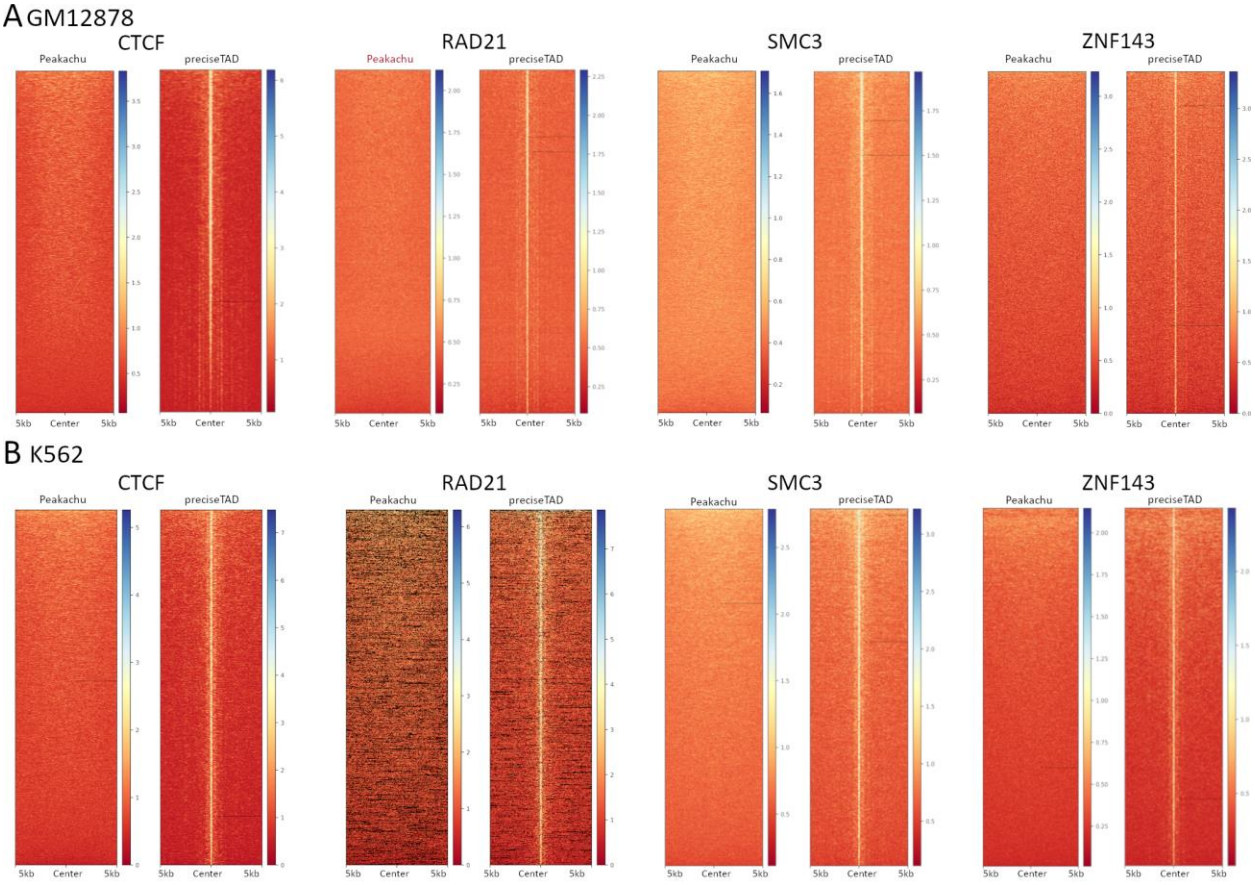
Additional file 12: Figure S6. *preciseTAD*-predicted boundaries are enriched for known molecular drivers of 3D chromatin. Signal profile plots comparing the binding strength of top TFBS around flanked (A) Arrowhead called TAD boundaries (blue) and *preciseTAD*-predicted TAD boundaries (green) on K562, (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.



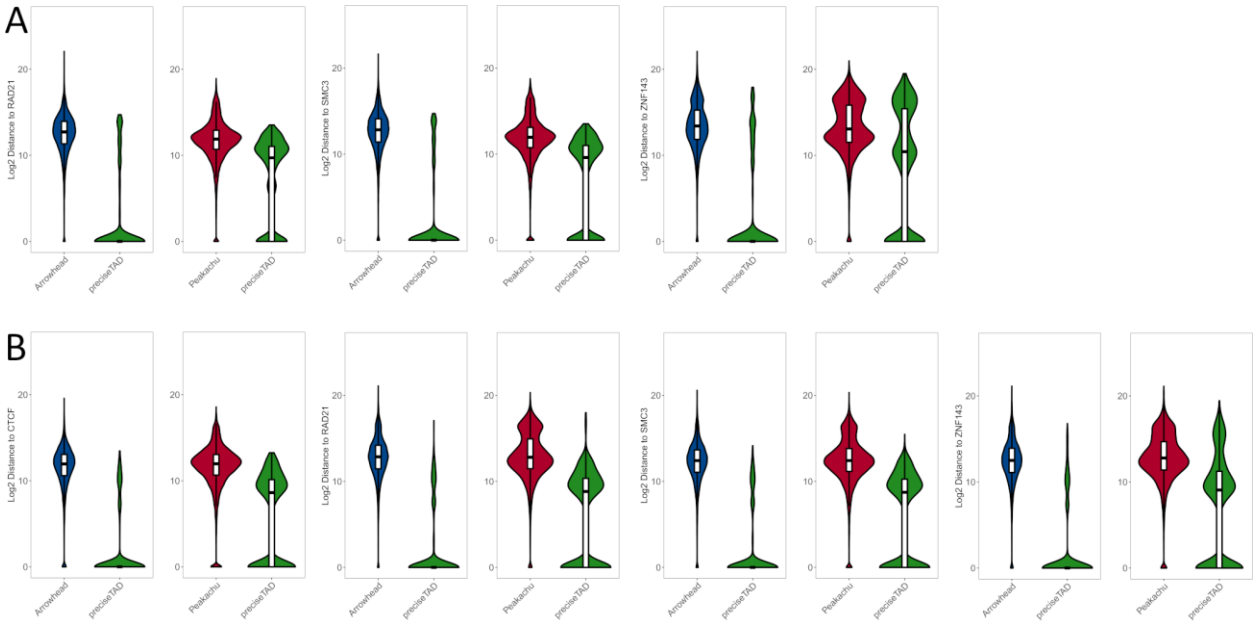
Additional file 13: Figure S7. *preciseTAD*-predicted boundaries are more enriched for known molecular drivers of 3D chromatin, as compared with Arrowhead boundaries. Enrichment heatmaps comparing the signal distribution of CTCF, RAD21, SMC3, and ZNF143 around Arrowhead-called TAD boundaries vs. *preciseTAD*-predicted TAD boundaries for (A) GM12878 and (B) K562 cell lines.



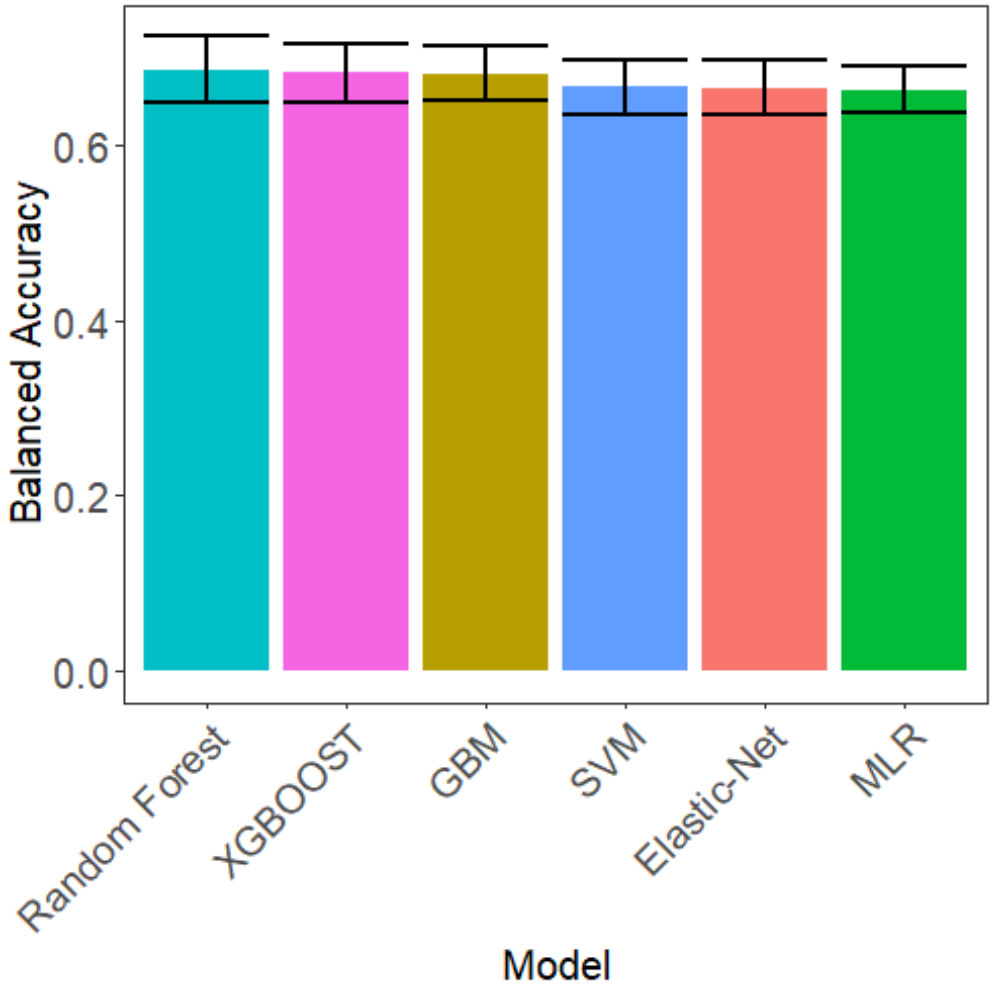
Additional file 14: Figure S8. *preciseTAD*-predicted boundaries are more enriched for known molecular drivers of 3D chromatin, as compared with Peakachu boundaries. Enrichment heatmaps comparing the signal distribution of CTCF, RAD21, SMC3, and ZNF143 around Peakachu-predicted chromatin loop boundaries vs. *preciseTAD*-predicted TAD boundaries for (A) GM12878 and (B) K562 cell lines.



Additional file 15: Figure S9. *preciseTAD*-predicted boundaries are spatially closer to known molecular drivers of 3D chromatin. Boxplots comparing the log_2 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 16: Figure S10. Random forest models more accurately predict TAD boundary regions compared to other machine learning algorithms Barplots comparing the average balanced accuracy when predicting TAD boundary regions on GM12878 at 25 kb using Random Forests, extreme gradient boosting (XGBOOST), gradient boosting machines (GBM), support vector machines (SVM), l_1 & l_2 regularized logistic regression (Elastic-Net), and multiple logistic regression (MLR). Error bars represent 1 standard deviation from the mean balanced accuracy across each holdout chromosome. The models are sorted in decreasing mean performance.



7. Vita

7.1 Education

Doctorate of Philosophy in Biostatistics at Virginia Commonwealth University, 2016-Present

Master of Arts student in Mathematics at Marshall University, 2014-2016

Bachelor of Science with double major in Applied Mathematics & Secondary Education at Marshall University, 2009-2013

7.2 Academic Employment

Graduate Teaching Assistant, Department of Biostatistics, Virginia Commonwealth University, 2016-2018. Responsibilities include: assisting professors with the preparation and presentation of graduate-level statistics courses, grading, and holding office hours.

Research Assistant, School of Dentistry, Virginia Commonwealth University, 2018-Present. Research activities include developing statistical analysis plans for graduate-level faculty research projects, performing statistical analysis, and preparing manuscripts for submission for publication.

7.3 Publications & Other Deliverables

Spiro C. Stilianoudakis, Mikhail G. Dozmorov. 2020. preciseTAD: A machine learning framework for precise 3D domain boundary prediction at base-level resolution. bioRxiv https://doi.org/10.1101/2020.09.03.282186

Stilianoudakis S, Dozmorov M (2020). preciseTADhub: Pre-trained random forest models obtained using preciseTAD. R package version 0.99.8, https://github.com/dozmorovlab/preciseTADhub.

Stilianoudakis S, Dozmorov M (2020). preciseTAD: preciseTAD: A machine learning framework for precise TAD boundary prediction. R package version 1.0.0, https://github.com/dozmorovlab/preciseTAD.

Holz, Magdalena & Naavaal, Shillpa & Stilianoudakis, Spiro & Carrico, Caroline & Byrne, B. & Myers, Garry. (2020). Antibiotics and antimicrobial resistance: Evaluation of the knowledge, attitude, and perception among students and faculty within US dental schools. Journal of dental education. 10.1002/jdd.12445.

Goolsby, Susie & Stilianoudakis, Spiro & Carrico, Caroline. (2020). A pilot survey of personality traits of dental students in the United States. British dental journal. 229. 377-382. 10.1038/s41415-020-2115-4.

Dragon, Carolyn & Shroff, Bhavna & Carrico, Caroline & Stilianoudakis, Spiro & Strauss, Robert & Lindauer, Steven. (2020). The effect of orthognathic surgery on facial recognition algorithm analysis. American Journal of Orthodontics and Dentofacial Orthopedics. 158. 10.1016/j.ajodo.2019.11.013.

Harmon, Evan & Heard, Brittney & Stilianoudakis, Spiro & Mazimba, Sula & Bowman, Brendan & Mehta, Nishaki. (2020). VENTRICULAR ARRHYTHMIAS MOST LIKELY TO OCCUR IN THE IMMEDIATE POST-DIALYSIS PERIOD IN PATIENTS WITH END-STAGE RENAL DISEASE AND CARDIAC IMPLANTABLE ELECTRONIC DEVICES. Journal of the American College of Cardiology. 75. 498. 10.1016/S0735-1097(20)31125-6.

Imbery, Terence & Stilianoudakis, Spiro & Tran, Dan & Bugas, Courtney & Seekford, Karoline. (2020). Is there an association between Perceptual Ability Test scores and color vision acuity?. Journal of dental education. 84. 10.1002/jdd.12111.

Shaw J A, Stiliannoudakis S, Qaiser R, et al. (July 21, 2020) Thirty-Day Hospital Readmissions: A Predictor of Higher All-cause Mortality for Up to Two Years. Cureus 12(7): e9308. doi:10.7759/cureus.9308

Rostami, Soheil & Kang, Balraj & Tufekci, Eser & Stilianoudakis, Spiro & Carrico, Caroline & Laskin, Daniel. (2019). Recognition of the Asymmetrical Smile: A Comparison of Orthodontists, Oral and Maxillofacial Surgeons and Lay Persons. Journal of Oral and Maxillofacial Surgery. 78. 10.1016/j.joms.2019.08.023.

7.4 Presentations

Spiro C. Stilianoudakis. A workshop describing preciseTAD: a machine-learning framework for predicting 3D domain 2020 boundaries using functional genomic elements. European Bioconductor Meeting (EuroBioc2020) 2020. Virtual. https://eurobioc2020.bioconductor.org/workshops.

Spiro C. Stilianoudakis. Developing a computational framework for precise TAD boundary prediction. Statistics & Data Science Symposium (SDSS) 2020. Pittsburgh, Pennsylvania (Virtual). https://ww2.amstat.org/meetings/sdss/2020/PDFs/SDSS2020_Program.pdf.

Spiro C. Stilianoudakis. Developing a computational framework for precise TAD boundary prediction. Eastern North American Region (ENAR) statistical meeting. Nashville, Tennessee (Virtual).

Spiro C. Stilianoudakis. Developing a computational framework for precise TAD boundary prediction. American Statistical Association (ASA) 2019. Virginia Chapter Annual Meeting. Charlottesville, Virginia.

7.5 Professional Membership

American Statistical Association (ASA)

Mathematical Association of America (MAA)

8. References

1. Ea V, Baudement M-O, Lesne A, Forné T: **Contribution of topological domains and loop formation to 3D chromatin organization**. *Genes* 2015, **6**:734–750.

2. 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.

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

4. 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**:1012–1025 10.1016/j.cell.2015.04.004.

5. 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**:265–269 10.1038/nature19800.

6. 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.

7. 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.

8. 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.

9. Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, Ye Z, Kim A, Rajagopal N, Xie W, others: **Chromatin architecture reorganization during stem cell differentiation**. *Nature* 2015, **518**:331.

10. 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.

11. Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, Li Y, Lin S, Lin Y, Barr CL, Ren B: **A compendium of chromatin contact maps reveals spatially active regions in the human genome**. *Cell Rep* 2016, **17**:2042–205910.1016/j.celrep.2016.10.061.

12. 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.

13. 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.

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

15. 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.

16. 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.

17. Dekker J, Rippe K, Dekker M, Kleckner N: **Capturing chromosome conformation**. *science* 2002, **295**:1306–1311.

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

19. 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.

20. 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.

21. Dali R, Blanchette M: **A critical assessment of topologically associating domain prediction tools**. *Nucleic acids research* 2017, **45**:2994–3005.

22. 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.

23. 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.

24. 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.

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

26. 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.

27. 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.

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

29. 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.

30. 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.

31. 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.

32. 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.

33. 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.

34. 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.

35. Consortium EP, others: **The encode (encyclopedia of dna elements) project**. *Science* 2004, **306**:636–640.

36. Ghirlando R, Felsenfeld G: **CTCF: Making the right connections**. *Genes & development* 2016, **30**:881–891.

37. Hnisz D, Day DS, Young RA: **Insulated neighborhoods: Structural and functional units of mammalian gene control**. *Cell* 2016, **167**:1188–1200.

38. Hanssen LL, Kassouf MT, Oudelaar AM, Biggs D, Preece C, Downes DJ, Gosden M, Sharpe JA, Sloane-Stanley JA, Hughes JR, others: **Tissue-specific ctcf–cohesin-mediated**

chromatin architecture delimits enhancer interactions and function in vivo. *Nature cell biology* 2017, **19**:952–961.

39. Handoko L, Xu H, Li G, Ngan CY, Chew E, Schnapp M, Lee CWH, Ye C, Ping JLH, Mulawadi F, others: **CTCF-mediated functional chromatin interactome in pluripotent cells**. *Nature genetics* 2011, **43**:630–638.

40. Huang J, Marco E, Pinello L, Yuan G-C: **Predicting chromatin organization using histone marks**. *Genome biology* 2015, **16**:162.

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

42. 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.

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

44. Hong S, Kim D: **Computational characterization of chromatin domain boundary-associated genomic elements**. *Nucleic acids research* 2017, **45**:10403–10414.

45. 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.

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

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

48. 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.

49. 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.

50. 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.

51. Gregorutti B, Michel B, Saint-Pierre P: **Correlation and variable importance in random forests**. *Statistics and Computing* 2017, **27**:659–678.

52. Yang F, Wang H-z, Mi H, Cai W-w, others: **Using random forest for reliable classification and cost-sensitive learning for medical diagnosis**. *BMC bioinformatics* 2009, **10**:1–14.

53. Díaz-Uriarte R, De Andres SA: **Gene selection and classification of microarray data using random forest**. *BMC bioinformatics* 2006, **7**:1–13.

54. Wu B, Abbott T, Fishman D, McMurray W, Mor G, Stone K, Ward D, Williams K, Zhao H: **Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data**. *Bioinformatics* 2003, **19**:1636–1643.

55. Breiman L: **Random forests**. *Machine learning* 2001, **45**:5–32.

56. Couronné R, Probst P, Boulesteix A-L: **Random forest versus logistic regression: A large-scale benchmark experiment**. *BMC bioinformatics* 2018, **19**:1–14.

57. Banfield RE, Hall LO, Bowyer KW, Kegelmeyer WP: **A comparison of decision tree ensemble creation techniques**. *IEEE transactions on pattern analysis and machine intelligence* 2006, **29**:173–180.

58. Moore BL, Aitken S, Semple CA: **Integrative modeling reveals the principles of multi-scale chromatin boundary formation in human nuclear organization**. *Genome Biol* 2015, **16**:110 10.1186/s13059-015-0661-x.

59. Al Bkhetan Z, Plewczynski D: **Three-dimensional epigenome statistical model: Genome-wide chromatin looping prediction**. *Sci Rep* 2018, **8**:5217 10.1038/s41598-018-23276-8.

60. 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.

61. Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F: *Learning from imbalanced data sets*. Springer; 2018.

62. Oshiro TM, Perez PS, Baranauskas JA: **How many trees in a random forest?** In *International workshop on machine learning and data mining in pattern recognition* Springer; 2012:154–168.

63. Probst P, Boulesteix A-L: **To tune or not to tune the number of trees in random forest.** *J. Mach. Learn. Res.* 2017, **18**:6673–6690.

64. Probst P, Wright MN, Boulesteix A-L: **Hyperparameters and tuning strategies for random forest**. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2019, **9**:e1301.

65. Van Berkum NL, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, Mirny LA, Dekker J, Lander ES: **Hi-c: A method to study the three-dimensional architecture of genomes.** *JoVE (Journal of Visualized Experiments)* 2010,:e1869.

66. Zheng H, Xie W: **The role of 3D genome organization in development and cell differentiation**. *Nature Reviews Molecular Cell Biology* 2019,:1.

67. Chen C, Yu W, Tober J, Gao P, He B, Lee K, Trieu T, Blobel GA, Speck NA, Tan K: **Spatial genome re-organization between fetal and adult hematopoietic stem cells**. *Cell reports* 2019, **29**:4200–4211.

68. Shin H, Shi Y, Dai C, Tjong H, Gong K, Alber F, Zhou XJ: **TopDom: An efficient and deterministic method for identifying topological domains in genomes**. *Nucleic acids research* 2016, **44**:e70–e70.

69. 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**:319 10.1186/s12859-020-03652-w.

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

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

72. 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.

73. 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.

74. 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.

75. Maeshima K, Tamura S, Hansen JC, Itoh Y: **Fluid-like chromatin: Toward understanding the real chromatin organization present in the cell**. *Curr Opin Cell Biol* 2020, **64**:77–8910.1016/j.ceb.2020.02.016.

76. 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.

77. Nora EP, Goloborodko A, Valton A-L, Gibcus JH, Uebersohn A, Abdennur N, Dekker J, Mirny LA, Bruneau BG: **Targeted degradation of ctcf decouples local insulation of chromosome domains from genomic compartmentalization**. *Cell* 2017, **169**:930–944.

78. Zuin J, Dixon JR, Reijden MI van der, Ye Z, Kolovos P, Brouwer RW, Corput MP van de, Werken HJ van de, Knoch TA, IJcken WF van, others: **Cohesin and ctcf differentially affect chromatin architecture and gene expression in human cells**. *Proceedings of the National Academy of Sciences* 2014, **111**:996–1001.

79. Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, Trzaskoma P, Magalska A, Wlodarczyk J, Ruszczycki B, others: **CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription**. *Cell* 2015, **163**:1611–1627.

80. Chen H, Tian Y, Shu W, Bo X, Wang S: **Comprehensive identification and annotation of cell type-specific and ubiquitous ctcf-binding sites in the human genome**. *PloS one* 2012, **7**:e41374.

81. Zheng Y, Keleş S: **FreeHi-c simulates high-fidelity hi-c data for benchmarking and data augmentation**. *Nature Methods* 2020, **17**:37–40.

82. 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.

83. 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**.

84. 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.