Predictive modeling using genomic annotations

Pitfalls and Recommendations

Spiro Stilianoudakis

August 9, 2018

Table of Contents

[ABSTRACT 2](#_Toc531003539)

[INTRODUCTION 3](#_Toc531003540)

[METHODS 5](#_Toc531003541)

[DOMAIN DATA 5](#_Toc531003542)

[GENOMIC ANNOTATIONS 6](#_Toc531003543)

[Feature Engineering 6](#_Toc531003544)

[ENSEMBLE FRAMEWORK 7](#_Toc531003545)

[Training and Testing Sets 7](#_Toc531003546)

[Variable Reduction 7](#_Toc531003547)

[Re-Sampling Techniques 7](#_Toc531003548)

[Classification 8](#_Toc531003549)

[Model Evaluation 9](#_Toc531003550)

[RESULTS 9](#_Toc531003551)

[ENSEMBLE COMPARISONS 9](#_Toc531003552)

[CLASS DISTRIBUTIONS ACROSS RESOLUTIONS 9](#_Toc531003553)

[CLASS IMBALANCE HINDERS PERFORMANCE ACROSS RESOLUTIONS 10](#_Toc531003554)

[CLASS BALANCING IMPROVES MODEL PERFORMANCE IN ALL RESOLUTIONS 10](#_Toc531003555)

[DISTANCE TYPE PREDICTORS OUT-PERFORM OTHER PREDICTOR TYPES 10](#_Toc531003556)

[PERFORMANCES WORSEN AS RESOLUTION INCREASES 10](#_Toc531003557)

[5kb RESOULTION DATA PROVIDES MOST ACCURATE ACCURATE RESULTS 10](#_Toc531003558)

[Discussion 11](#_Toc531003559)

[Abbreviations 11](#_Toc531003560)

[Acknowledgements 11](#_Toc531003561)

[Funding 12](#_Toc531003562)

[Tables 12](#_Toc531003563)

[Table 1. 12](#_Toc531003564)

[Table 2. 12](#_Toc531003565)

[Table 7. 12](#_Toc531003566)

[Figures 12](#_Toc531003567)

[Table Legends 15](#_Toc531003568)

[Table 1: List of Genomic Annoations. 15](#_Toc531003569)

[Table 2: Class Distributions 16](#_Toc531003570)

[Table 7: Model Performances at all Resolutions 16](#_Toc531003571)

[Figure legends 16](#_Toc531003572)

[Figure 1: Model Construction 16](#_Toc531003573)

[Figure 2: Predictor Types 16](#_Toc531003574)

[Figure 3: Ensemble Framework/Model Building Pipeline 17](#_Toc531003575)

[Figure 4: Class Imbalance 17](#_Toc531003576)

[Figure 11: Model Performances for all Resolutions 17](#_Toc531003577)

[Figure 9: Comparing Performances 17](#_Toc531003578)

[Figure 10: Comparing Variable Importances 17](#_Toc531003579)

[Supplementary Tables and Figures 18](#_Toc531003580)

[Supplementary Legends 18](#_Toc531003581)

[References 18](#_Toc531003582)

# ABSTRACT

**Background.** Chromosome conformation capture sequencing technologies have shown that the three-dimensional (3D) structure of the genome is folded into distinct compartments, known as topologically associated domains (TADs) - units of coordinated gene expression. The location of TAD boundaries is highly conserved, suggesting the presence of epigenomic marks aiding in TAD formation. The ability to predict which epigenomic features are most associated with the TAD boundaries will allow to better understand the regulatory role of the 3D structure of the genome.

Existing methods for predicting associations between genomic elements tend to ignore key characteristics of the genomic data. Specifically, the number of TAD boundaries is much less than the number of other genomic regions, leading to heavily imbalanced classes. Furthermore, most methods utilize direct overlap as a means to quantify the association, while distance, the measure of spatial relationships, remains unaccounted for. Consequently, distances on a genomic scale vary widely, leaving uncertainty how the heavily right-tailed distribution of distance measures will affect the model’s performance.

**Methods.** We propose a novel data pre-processing pipeline that addresses those shortcomings. A number of classifier performance metrics were assessed, including the F1 measure and Matthew Correlation Coefficient (MCC), and area under the ROC curve (AUROC).

**Results.** Data preprocessing (log2-transformation and standardization) improves the performance of classification algorithms and allows for the ability to more accurately predict which genomic features are most associated with TAD boundaries.

**Conclusions.** Current methods used to model the epigenomic features associated with TAD boundaries are insufficiently robust to handle properties of genomic data. Models applied to unprocessed data can have poor predictive performances. Focusing solely on typical performance assessment metrics, such as AUROC, can mask poor performance of the models; thus, the use of more balanced metrics, such as F1 and MCC, is warranted. Our model results in better performances and more accurate identification of important features associated with the formation of TAD boundaries.

# INTRODUCTION

The advent of various genome-wide sequencing technologies, such as high-throughput conformation capture (notably Hi-C), have revealed how the spatial organization of the human genome may affect several epigenetic functions (Lieberman-Aiden et al. 2009). Analyses have shown that the genome is tightly compacted into distinct compartments. There exist regions within these compartments that have been shown to be highly conserved and self-interacting, and termed topologically associating domains (TADs) (Dixon et al. 2012). Evidence suggests that regulatory elements and genes tend to interact more frequently within the same TAD (Symmons et al. 2014). This suggests that the boundaries of TADs may play a role in restricting the function of certain genomic elements such as enhancers, thereby impacting the transcription of genes. Furthermore, changing the 3D structure of the DNA, causing disruptions in these domains can lead to adverse outcomes and diseases like cancer. Therefore, it has become increasingly important to be able to identify the key molecular drivers of the formation of TAD boundaries in order to further our understanding of the human genome.

The mechanisms underlying the formation of TADs are a complex and active area of research. Recently, it has been discovered that insulator sequences have a primary role in orchestrating the topological arrangements of higher-order chromatin architecture (Phillips-Cremins and Corces 2013). Insulators are multi-faceted regulatory sequences that moderate a variety of genomic processes including activation, repression, and enhancer blocking. Specifically, the insulator binding protein CTCF has been found to be enriched at the boundary sequences of topologically associating domains in human cells and may therefore act as a mediator of long-range chromatin contacts (Zuin et al. 2014). Other regulatory elements like particular histone modifications such as H3k36 and H3K27 trimethylation, and transcription factors such as ZNF274 and YY1 have been more frequently observed at TAD boundaries than in other regions (Rao et al. 2014). These have been found to be associated with more open chromatin which alter the accessibility of genes for transcription. The distinct patterns of some of these different proteins and functional elements point toward the opportunity of using computational approaches in identifying which epigenomic features are most predictive of the development of TAD boundaries. This will allow us to better understand what leads to their formation.

Most accepted methods for predicting TAD boundaries consider performing a classification algorithm by binning the genome into equal-sized bins and labeling them as either containing a TAD boundary or not. However, a pervasive issue becomes the introduction of highly imbalanced classes. Due to the sparcity of TADs throughout the genome, most bins will not contain a TAD boundary. Therefore, the minority class of genomic bins with TAD boundaries in them becomes proportionally smaller than that of the majority class of genomic bins with no TAD boundaries. It is commonly agreed upon in the machine learning community that imbalanced datasets adversely impact the performance of the classifiers as the learned model is biased towards the majority class to minimize the overall error rate (Estabrooks, Jo, and Japkowicz 2004). However, a clear method for creating balanced classes remains evasive and largely dependent on the problem at hand. For that reason we evaluated 3 widely accepted class balancing techniques, in addition to no balancing. They included random under-sampling, random over-sampling, and synthetic minority over-sampling technique (SMOTE).

Model performance is also contingent upon the resolution of the contact matrix used to identify the location of TADs throughout the genome. This will also affect the size of the genomic bins used in model construction. A larger resolution contact matrix will result in the identification of larger-sized TADs, and subsequently, less TAD boundary points. This will contribute to smaller sample sizes and could effect predictive performances.

Another problem lies in identifying the feature space of subsequent classification algorithms. A variety of positional characteristics of the genomic features can be used to describes the relationship between TAD boundaries and genomic elements, but it is unclear which is best. The use of distances (in base pairs) from genomic bins to regions of functional genomic elements has largely been ignored. However, these distance characteristics offer a more accurate spatial representation of which genomic features are associated with TAD boundaries.

Due to the increasingly large amounts of functional genomic elements to consider, it is a popular approach in genomic studies to use variable reduction techniques to reduce the noise induced by noninformative predictors and increase computational efficiency. Two popular approaches use the norm (LASSO) and a combination of the and norms together (Elastic-Net). While both regularization techniques have merit, LASSO’s inability to handle correlated predictors poses a problem when identifying which element are most predictive of TAD boundaries. The elastic-net, however, is better suited in a situation where multiple genomic elements colocalize around the same TAD boundaries, and could be better suited for this type of alaysis (Zou and Hastie 2005).

Careful consideration must also be taken when deciding which performance metric(s) to present when evaluating predictive models. A common metric to consider is the the area under the ROC curve (AUC), as well as others including accuracy, sensitivity, and specificity.

The popularity surrounding the use of ROC curves is that they do not depend on a threshold. However, when the prevalence of an event is low–that is, a genomic bin containing a TAD boundary–mean probabilities are biased towards the majority class. Likewise, metrics like accuracy and sensitivity and specificity do not incorporate the entirety of the confusion matrix created when validating predictive models on a test data set. Therefore, each of these metrics are likely to be inflated when evaluating model performance on imbalanced data. More robust metrics like the F1-score and Matthew’s Correlation Coefficient are more suitable for this type of analysis and can highlight problems associated with applying predictive models on imbalanced data.

[Mention Previous Methods]

In this study we have developed an ensemble framework that aims to address and compare a variety of data irregularities and characteristics that are associated with the prediction of TAD boundaries. We evaluated domain data obtained from four different resolution contact matrices for the GM12878 cell line including 5kb, 25kb, 50kb, and 100kb. For each resolution, we examined the inclusion of three different predictor types as the feature space for downstream modeling. We coupled each predictor type with two separate variable reduction techniques. Lastly, we performed a series of three different re-sampling techniques aimed at created balanced classes. Random forest classification algorithms were performed to both compare each combination of the above ensemble, and investigate which functional genomic features were most associated with the formation of TAD boundaries. Model performance was evaluated based on a total of four metrics which included test accuracy, area under the ROC curve (AUC), F1 score, and Matthew’s Correlation Coefficient (MCC).

# METHODS

## DOMAIN DATA

Topologically associating domain data for the GM12878 cell line was obtained using the arrowhead algorithm from the Juicer toolbox (<https://github.com/aidenlab/juicer>). Publically available domain data was obtained from in situ Hi-C contact matrices at 5kb resolution (Rao et al. 2014). Additional domain data was obtained from Dali et al. at 25kb, 50kb, and 100kb resolutions (Dali and Blanchette 2017).

Identified TAD boundaries were represented by their location in the genome (hg19 human genome assembly), including chromosome, start and end coordinates for chromosomes 1 through 22. The start and end coordinates were concatenated, sorted, and unique coordinates representing the borders demarcating TADs were obtained.

Figure 1 presents a diagram of how the data was established for modeling purposes. First, the genome was binned into equally sized intervals. This was performed at the same resolution of the contact matrix that the domain data was obtained from. For example, for the domain data obtained from a 5kb resolution contact matrix, the genome was subsequently binned into 5kb intervals. The same processes was repeated for each of the other three resolutions. Next, the boundary coordinates were flanked on either side by a unit corresponding to the resolution of the particular domain data. Again, siting the domain data obtained from a 5kb resolution contact matrix, each uniquely identified TAD boundary was flanked on either side by 5kb for a total width around each boundary point of 10kb. The same process was repeated for the other three resolutions. A genomic bin was labeled as having a TAD boundary within it (Y=1) if it overlapped with a particular flanked boundary. Otherwise, the bin was labeled as not containing a boundary (Y=0).

## GENOMIC ANNOTATIONS

Annotation data, in the form of functional genomic elements from ChIP-seq experiments were obtained from the Encyclopedia of DNA Elements (ENCODE) Consortium. The annotations consisted of a variety of different chromatin states, histone marks, and DNase I hypersensitive sites. Table 1 presents the total list of genomic annotations considered in this study and the sources of downloadable files. Each annotation was represented by their location throughout the genome via chromosomal coordinates. An example of a specific annotation, which is labeled as a functional genomic element, is presented in Figure 2. It is represented by the green segments that overlap with the genomic bins.

## Feature Engineering

Using each genomic element, we created three types of predictors that were used as features to analyze their relationships with TAD boundaries in downstream modelling. Figure 2 presents an illustration of each type of predictor that was considered. The predictors consisted of two different types of overlaps, and a distance type. They are described in detail below:

1. Overlap Counts (OC): For each genomic bin, the total number of overlapped regions between a functional genomic element and a bin was calculated. This was defined as the overlap counts (OC) and is illustrated in the left part of Figure 2. Within each bin (marked by the blue dashed lines), the total number of overlapped regions of the genomic element (represented by the horizontal green segments) are tallied and marked below. In the case of no overlaps the value is given as 0.
2. Overlap Percent (OP): For each genomic bin, the percent overlap between regions defined by a particular functional element and a bin was calculated. From the middle diagram of Figure 2, the feature width was defined as the number of base pairs that overlapped with the genomic bin. The percent was then calculated by dividing the feature width with the bin width. If multiple overlaps existed, the feature width was defined as the sum of the total number of base pairs that overlapped with a particular bin. Note that the bin width remained constant for a given resolution (either 5kb, 25kb, 50kb, or 100kb). Similar to the OC predictor type, bins with no overlaps were given to be 0.
3. Distance: For each genomic bin, the distance (in base pairs) from the center of the bin to the center of the nearest region of a respective functional element was calculated. The right most diagram of Figure 2 depicts an example of how distance was calculated. The vertical green segments represent the centers of the respective genomic element. Each distance, , represented by the blue brackets are the number of base pairs from the center of the bins to the center of the nearest genomic elemental region.

[Mention transformations used]

## ENSEMBLE FRAMEWORK

Figure 3 details the ensemble framework proposed and evaluated in this study. For each resolution and predictor type chosen, the data was split into training to testing sets. A variable reduction technique was then performed. The feature space was reduced and a resampling technique was applied. Lastly, a classification algorithm in the form of a random forest was performed and validated. Details of each step in the framework are provided below.

### Training and Testing Sets

Given a particular resolution, the full data set was composed of genomic bins that either overlapped with a flanked TAD boundary (minority set) or did not (majority set). Figure 3 highlights the severe class imbalance present in both the full data set. The data was then split into a 7:3 ratio of training to testing sets. Each training and testing set was composed of a similar majority to minority class ratio as the full data set, which is represented by the segmented regions in Figure 3. The training set was then used to perform all downstream modeling. The same test set was held separate and used to validate all models.

### Variable Reduction

We evaluated the use of two variable reduction algorithms, the LASSO and Elastic-Net regularizations. Each algorithm was performed prior to re-sampling. For the LASSO, the penalization term was tuned over a grid of 10 values on an exponential scale ranging from 0.01 to 10. The additional term for the Elastic-Net was tuned over 10 equa-distant values from 0 to 1.

### Re-Sampling Techniques

After choosing one of the aforementioned data resolutions, predictor types, and variable reduction algorithms, we then evaluated three re-sampling techniques used to create balanced classes. For each resampling technique, the majority class is defined as the set of genomic bins that did not include a TAD boundary, while the minority class is defined as the set of bins that included a TAD boundary. The following details the re-sampling techniques considered in this paper:

1. No Sampling: All of the data points from the majority and minority class of the training set were used.
2. Random Under-Sampling: All of the minority classes from the training set were used. A random set without replacement from the majority classes of the training set were used to match the number of minority samples. In this paper we performed 50 iterations of random under-sampling. At each iteration a random forest classification algorithm was performed. Performance metrics were aggregated by taking the average across all of the iterations.
3. Random Over-Sampling: All of the majority classes from the training set were used. A random set with replacement from the minority classes of the training set were used to match the number of majority samples. Similar to the under-sampling technique, 50 iterations were performed and results were aggregated by taking the average across iterations.
4. SMOTE (Synthetic Minority Over-Sampling Technique): This method incorporated both random under- and over-sampling. Under-sampling is performed without replacement from the majority class, while over-sampling is performed by creating new synthetic observations using the minority class. The SMOTE algorithm is stored in the DMwR package, available in R (Chawla et al. 2002). The user is able to specify the parameters denoting the percent of over- and under-sampling from the two classes. This allows the user to control the proportion of disbalance in the data. For the purposes of this paper, we used SMOTE to create perfectly balanced data. However further research would be needed to see if varying the proportion of class disbalance could improve predictive performance.

### Classification

A classification algorithm in the form of a random forest was performed and validated using the test set. 10-fold cross validation was used to reduced bias due to random dataset generation. The default number of features to consider at each node of the random forest algorithm was set as the square root of the number of features in the model. Likewise, the number of ensemble trees to aggregate was set at 500. Once the model was implemented, it was assessed using the test set, and performance metrics were then recorded.

Using set notation the formulation of the ensemble framework can be defined as follows:

Set of different resolutions:

R = {5kb, 25kb, 50kb, 100kb}

Set of Predictor Types:

P = {OC, OP, Distance}

Set of Variable Reduction Techniques:

V = {LASSO, Elastic-Net}

Set of Class-Balancing Methods:

C = {None, RUS, ROS, SMOTE}

Therefore, a system produced from the ensemble framework is defined as follows:

E = {r, p, v, c}, where r R, p P, v V, and c C.

Once a system was defined, a random forest algorithm was performed. Letting |X| denote the cardinality of each set, then we evaluated |R| x |P| x |V| x |C| different ensemble systems in this paper.

### Model Evaluation

The efficacy of different ensemble systems was compared using various performance metrics including accuracy, AUC, F1-score, and MCC. These metrics are defined as follows:

Here, TP refers to the number of bins correctly identified as containing a TAD boundary (true positives), FP refers to the number of bins incorrectly identified as containing a TAD boundary (false positives), TN refers to the number of bins correctly identified as not containing a TAD boundary (true negatives), and FN refers to the number of bins incorrectly identified as not containing a TAD boundary. Each of these quantities are obtained from the confusion matrix created by validating the model on the testing set. Accuracy measures the percentage of correct classifications by the model. However, in the case of heavily imbalanced class, as can be seen by the numerator, the metric is biased toward correcly identified majority classes. The F1-score and MCC are more balanced measures. We also considered the AUC for each model, by computing the average trapezoidal approximations for the curve created by the true positive rates (TPR) and false positive rates (FPR)

# RESULTS

## ENSEMBLE COMPARISONS

In this section we provide the details of the comprehensive experiments performed in order to assess each ensemble system. For the task of predicting TAD boundaries using functional genomic elements, we evaluated 4 domain data sets, 3 predictor types, 2 variable reduction techniques, and 4 class balancing techniques. We evaluated each ensemble system using 10-fold cross-validated random forest models. Thus, we generated 96 (=4x3x2x4) separate models.

### CLASS DISTRIBUTIONS ACROSS RESOLUTIONS

Table 2 presents the class distributions across each resolution. After binning the genome at 5kb, 25kb, 50kb, and 10kb intervals there were a total of 44948, 9042, 4541, and 2279 genomic bins created respectively. Figure 4 illustrates the class imbalance problem present in each data resolution. Likewise, we see the apparent decrease in data points (bins) as resolution increases. Furthermore, Table 1 specifically shows the class distribution for each resolution. The overall average perentage of minority classes across all data resolutions was 13.2%, with the most severe case present in the 5kb resolution (6.9%).

### CLASS IMBALANCE HINDERS PERFORMANCE ACROSS RESOLUTIONS

Figure 11 present the results across all of the ensemble systems for each resolution of domain data. The first row in each of the 4 regions shows the performances for models without resampling. The average accuracy and AUC were marginally high across each of the resolutions with values of W%, X%, Y%, and Z% respectively. However, the severity of the model performances were much more apparant when comparing F1-scores and MCCs. The average F1-scores across resolutions for models without resampling were found to be W%, X%, Y%, and Z% respectively, while the MCCs were W%, X%, Y%, and Z% (Table 7). We thus see the necessity for implementing a class balancing technique prior to modeling.

### CLASS BALANCING IMPROVES MODEL PERFORMANCE IN ALL RESOLUTIONS

Moving down the rows in each of the 4 regions of Figure 11 we see that model performance was markedly improved among the F1-scores and MCCs after creating balanced classes compared to the results using no re-sampling. For each of the resolutions, we see that random under-sampling out performed the other class balancing techniques. While SMOTE performed very similarly to RUS, ROS performed poorly in comparison, especially as the resolution of the data increased. Furthermore, we see that the elastic-net reduction technique outperformed the LASSO. Although, the disparity became less evident as the resolution increased.

### DISTANCE TYPE PREDICTORS OUT-PERFORM OTHER PREDICTOR TYPES

Again referencing Figure 11, focusing on the rows corresponding to models using randomly under-sampled data, it was found that models with distance type predictors out performed models with the two other overlap types (counts and percents). The disparity between predictor types was steady across resolutions (Table 7). The greatest difference was seen in the 5kb data. For models with Elastic-Net regularization, the F1-scores were X, Y, and Z, while the MCCs were X, Y, and Z for each predictor type respectively.

### PERFORMANCES WORSEN AS RESOLUTION INCREASES

Figure 9 compares the results of random forest models using distance-type predictors, with elastic-net regularization, and random under-sampling, across each domain data resolution. We see that the results for F1-scores are relatively stable between 5kb, 25kb, and 50kb resolutions with a sharp decrease for the 100kb resolution. Furthermore, the decrease in performance is more evident and steady across accuracy, AUC, and MCC metrics.

### 5kb RESOULTION DATA PROVIDES MOST ACCURATE ACCURATE RESULTS

To identify the functional genomic elements that were most predictive of TAD boundaries, we analyzed the variable importance produced by each random forest algorithm. Figure 10 presents the subsequent scaled importance plots of the top 15 most predictive genomic elements for each model presented in Figure 9. We see that for the 5kb resolution, the most predictive elements include insulator proteins, DNase hypersensitive sites, the transcription factor CTCF, as well as histone modifications H3k9ac, H2az, and H3k4me2. These results align with what is seen in the literature. Although, the insulator protein is featured as a top predictor in 2 of the other 3 resolutions, the rest of the predictors are less consistent among each other and less in agreement with the literature. The results from Figure 9 and 10 confirm that the model results appear to degrade as resolution increases with respect to both prediction and the identification of genomic elements associated with TAD boundaries.

# Discussion

# Abbreviations

# Acknowledgements

*Conflict of Interest.* None.

# Funding

# Tables

## Table 1.

|  |  |  |  |
| --- | --- | --- | --- |
| List of Genomic Elements |  |  |  |
| Genomic Class | Element | Description | Source |
| Genomic Variants | Complex | Regions that are highly non-repetitive | http://dgv.tcag.ca/dgv/app/downloads?ref=GRCh37/hg19 |
|  | Deletion | Regions that are missing nucleotides relative to the reference genome | http://dgv.tcag.ca/dgv/app/downloads?ref=GRCh37/hg20 |
|  | Duplication | Regions that are duplicated relative to the reference genome | http://dgv.tcag.ca/dgv/app/downloads?ref=GRCh37/hg21 |
|  | Gain Loss | Regions that have experienced either gains or losses of genetic material relative to the reference genome | http://dgv.tcag.ca/dgv/app/downloads?ref=GRCh37/hg22 |
|  | Insertion | Regions in which a segment of a chromosome is reversed end to end | http://dgv.tcag.ca/dgv/app/downloads?ref=GRCh37/hg23 |
|  | Inversion | Regions where additional nucleotides have been inserted in a DNA sequence, relative to the reference genome | http://dgv.tcag.ca/dgv/app/downloads?ref=GRCh37/hg24 |
|  | Mobile Element Insertion | Regions that can change its position within a genome, sometimes creating or reversing mutations | http://dgv.tcag.ca/dgv/app/downloads?ref=GRCh37/hg25 |
|  | Novel Sequence Insertion |  | http://dgv.tcag.ca/dgv/app/downloads?ref=GRCh37/hg26 |
|  | Sequence Alteration | Regions where an insertion of a sequence into the donor genome where no subsequence with high similarity to the inserted sequence exists in the reference genome | http://dgv.tcag.ca/dgv/app/downloads?ref=GRCh37/hg27 |
|  | Tandem Duplication | Regions where duplication of exons within the same gene to give rise to the subsequent exon | http://dgv.tcag.ca/dgv/app/downloads?ref=GRCh37/hg28 |
| Evolutionary Constrained Sites | GERP | Regions of constrained elements where there exists a deficit of substitution events | http://mendel.stanford.edu/SidowLab/downloads/gerp/index.html |
| Ultra-Conserved Non-coding Elements | UCNE | Regions of ultra-conserved non-coding elements | https://ccg.vital-it.ch/ |
| Variably Methylated Regions | VMR | Regions with different DNA methylation status relative to a reference genome | https://epigenome.wustl.edu/methylomes/vmr.php |
| Nested Repeats | LINE | identical sequences of DNA that repeat hundreds or thousands of times found at either end of retrotransposons or proviral DNA formed by reverse transcription of retroviral RNA | http://genome.ucsc.edu/ENCODE/ |
|  | LTR | Long terminal repetitive regions | http://genome.ucsc.edu/ENCODE/ |
|  | SINE | Short interspersed nuclear elements | http://genome.ucsc.edu/ENCODE/ |
| Chromatin State Segmentation | Active Promotor | region of DNA that initiates transcription of a particular gene | hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHmm/wgEncodeBroadHmmGm12878HMM.bed.gz |
|  | Heterochromatin | Regions of DNA of different density from normal (usually greater), in which the activity of the genes is modified or suppressed. | hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHmm/wgEncodeBroadHmmGm12878HMM.bed.gz |
|  | Insulator | Regions of DNA that prevent distal enhancers from acting on the promoter of neighbouring genes | hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHmm/wgEncodeBroadHmmGm12878HMM.bed.gz |
|  | Poised Promotor | Promoters of transcriptionally silent developmental regulatory genes | hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHmm/wgEncodeBroadHmmGm12878HMM.bed.gz |
|  | Repetative CNV | Regions of repetitive copy number variations | hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHmm/wgEncodeBroadHmmGm12878HMM.bed.gz |
|  | Repressed | Regions of DNA that inhibit the expression of one or more genes by binding to the operator or associated silencers | hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHmm/wgEncodeBroadHmmGm12878HMM.bed.gz |
|  | Strong Enhancer | Enhancers associated with highly transribed genes | hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHmm/wgEncodeBroadHmmGm12878HMM.bed.gz |
|  | TXN Elongation | Regions of RNA that have additional nucleotides during transcription | hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHmm/wgEncodeBroadHmmGm12878HMM.bed.gz |
|  | TXN Transition | Points in the DNA where a mutation has occurred that changes nucleotides between purines or pyrimidines. | hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHmm/wgEncodeBroadHmmGm12878HMM.bed.gz |
|  | Weak Enhancer | Enhancers associated with weakly transribed genes | hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHmm/wgEncodeBroadHmmGm12878HMM.bed.gz |
|  | Weak Promotor | Promotors associated with highly transribed genes | hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHmm/wgEncodeBroadHmmGm12878HMM.bed.gz |
|  | Weak TXN | Genes that are not highly transcribed | hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHmm/wgEncodeBroadHmmGm12878HMM.bed.gz |
|  | CTCF | CTCF enriched element | https://hgdownload-test.gi.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgSegmentation/ |
|  | E | Predicted enhancer | https://hgdownload-test.gi.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgSegmentation/ |
|  | PF | Predicted promoter flanking region | https://hgdownload-test.gi.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgSegmentation/ |
|  | R | Predicted Repressed or Low Activity region | https://hgdownload-test.gi.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgSegmentation/ |
|  | T | Predicted transcribed region | https://hgdownload-test.gi.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgSegmentation/ |
|  | TSS | Predicted promoter region including TSS | https://hgdownload-test.gi.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgSegmentation/ |
|  | WE | Predicted promoter region including TSS | https://hgdownload-test.gi.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgSegmentation/ |
| Dnase I Hypersensitive Sites | Dnase | regions of chromatin that are sensitive to cleavage by the DNase Ienzyme | http://genome.ucsc.edu/ENCODE/ |
| Histone Modifications | H2az | H2A.Z is a sequence variant of Histone H2A. H2AZ appears to alter nucleosome stability, is partially redundant with nucleosome remodeling complexes, and is involved in transcriptional control | http://genome.ucsc.edu/ENCODE/ |
|  | H3k27ac | Histone H3 (acetyl K27). As with H3K9ac, associated with transcriptional initiation and open chromatin structure. It remains unknown whether acetylation has can have different consequences depending on the specific lysine residue targeted. In general, though, there appears to be high redundancy. Histone acetylation is notable for susceptibility to small molecules and drugs that target histone deacetylases. | http://genome.ucsc.edu/ENCODE/ |
|  | H3k27me3 | Histone H3 (tri-methyl K27). Marks promoters that are silenced by Polycomb proteins in a given lineage; large domains are found at inactive developmental loci. | http://genome.ucsc.edu/ENCODE/ |
|  | H3k36me3 | Histone H3 (tri-methyl K36). Marks regions of RNAPII elongation, including coding and non-coding transcripts. | http://genome.ucsc.edu/ENCODE/ |
|  | H3k4me1 | Histone H3 (mono methyl K4). Is associated with enhancers, and downstream of transcription starts. | http://genome.ucsc.edu/ENCODE/ |
|  | H3k4me2 | Histone H3 (di methyl K4). Marks promoters and enhancers. Most CpG islands are marked by H3K4me2 in primary cells. May be associated also with poised promoters. | http://genome.ucsc.edu/ENCODE/ |
|  | H3k4me3 | Histone H3 (tri methyl K4). Marks promoters that are active or poised to be activated. | http://genome.ucsc.edu/ENCODE/ |
|  | H3k79me2 | H3K79me2 is a mark of the transcriptional transition region - the region between the initiation marks (K4me3, etc) and the elongation marks (K36me3). | http://genome.ucsc.edu/ENCODE/ |
|  | H3k9ac | Histone H3 (acetyl K9). As with H3K27ac, associated with transcriptional initiation and open chromatin structure. It remains unknown whether acetylation can have different consequences depending on the specific lysine residue targeted. In general, though, there appears to be high redundancy. Histone acetylation is notable for susceptibility to small molecules and drugs that target histone deacetylases. | http://genome.ucsc.edu/ENCODE/ |
|  | H3k9me1 | Histone H3 (mono-methyl K9). Is associated with active and accessible regions. NOTE CONTRAST to H3K9me3 which is associated with repressive heterochromatic state. | http://genome.ucsc.edu/ENCODE/ |
|  | H3k9me3 | Histone H3 (tri methyl K9). Is associated with repressive heterochromatic state (silenced chromatin). NOTE CONTRAST to H3K9me1 which is associated with active and accessible regions. | http://genome.ucsc.edu/ENCODE/ |
|  | H4k20me1 | Histone H4 (mono-methyl K20). Is associated with active and accessible regions. In mammals, PR-Set7 specifically catalyzes H4K20 monomethylation. NOTE CONTRAST to H3K20me3 which is associated with heterochromatin and DNA repair. | http://genome.ucsc.edu/ENCODE/ |

## Table 2.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Genomic Bin | Complete | Training | Testing | No Balancing | Random Under-Sampling | Random Over-Sampling | SMOTE |
| 5 kb |  |  |  |  |  |  |  |
| With TAD boundary | 3083 | 2140 | 943 | 2140 | 2140 | 29323 | 4280 |
| Without TAD boundary | 41865 | 29323 | 12542 | 29323 | 2140 | 29323 | 4280 |
| Total | 44948 | 31463 | 13485 | 31463 | 4280 | 58646 | 8560 |
| 25 kb |  |  |  |  |  |  |  |
| With TAD boundary | 1266 | 895 | 371 | 895 | 895 | 5434 | 1790 |
| Without TAD boundary | 7776 | 5434 | 2342 | 5434 | 895 | 5434 | 1790 |
| Total | 9042 | 6329 | 2713 | 6329 | 1790 | 10868 | 3580 |
| 50 kb |  |  |  |  |  |  |  |
| With TAD boundary | 621 | 431 | 190 | 431 | 431 | 2747 | 862 |
| Without TAD boundary | 3920 | 2747 | 1173 | 2747 | 431 | 2747 | 862 |
| Total | 4541 | 3178 | 1363 | 3178 | 862 | 5494 | 1724 |
| 100 kb |  |  |  |  |  |  |  |
| With TAD boundary | 277 | 185 | 92 | 185 | 185 | 1410 | 370 |
| Without TAD boundary | 2002 | 1410 | 592 | 1410 | 185 | 1410 | 370 |
| Total | 2279 | 1595 | 684 | 1595 | 370 | 2820 | 740 |

## Table 7.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Resolution | 5kb |  |  |  |  |  |  | 25kb |  |  |  |  |  |
|  | Variable Reduction | LASSO |  |  | Elastic-Net |  |  | LASSO |  |  | Elastic-Net |  |  |
|  | Predictor Type | OC | OP | Distance | OC | OP | Distance | OC | OP | Distance | OC | OP | Distance |
| Accuracy | None | 0.929 | 0.927 | 0.93 | 0.93 | 0.93 | 0.931 | 0.86 | 0.843 | 0.864 | 0.864 | 0.863 | 0.864 |
|  | RUS | 0.706 | 0.633 | 0.723 | 0.692 | 0.679 | 0.729 | 0.603 | 0.716 | 0.663 | 0.611 | 0.636 | 0.666 |
|  | ROS | 0.81 | 0.718 | 0.933 | 0.907 | 0.922 | 0.934 | 0.816 | 0.775 | 0.867 | 0.853 | 0.861 | 0.868 |
|  | SMOTE | 0.768 | 0.648 | 0.764 | 0.812 | 0.768 | 0.775 | 0.725 | 0.731 | 0.714 | 0.752 | 0.709 | 0.719 |
| AUC | None | 0.669 | 0.577 | 0.888 | 0.734 | 0.737 | 0.902 | 0.606 | 0.539 | 0.791 | 0.626 | 0.638 | 0.804 |
|  | RUS | 0.7 | 0.669 | 0.853 | 0.74 | 0.738 | 0.862 | 0.604 | 0.558 | 0.761 | 0.625 | 0.654 | 0.773 |
|  | ROS | 0.673 | 0.651 | 0.901 | 0.701 | 0.711 | 0.912 | 0.585 | 0.546 | 0.819 | 0.613 | 0.657 | 0.833 |
|  | SMOTE | 0.694 | 0.668 | 0.858 | 0.741 | 0.741 | 0.867 | 0.596 | 0.56 | 0.761 | 0.603 | 0.63 | 0.77 |
| F1-Score | None | 0.018 | 0.03 | 0.031 | 0.004 | 0.002 | 0.055 | 0.016 | 0.09 | 0.042 | 0.021 | NA | 0.047 |
|  | RUS | 0.218 | 0.197 | 0.297 | 0.229 | 0.225 | 0.303 | 0.273 | 0.227 | 0.375 | 0.287 | 0.303 | 0.382 |
|  | ROS | 0.156 | 0.181 | 0.203 | 0.089 | 0.063 | 0.219 | 0.157 | 0.173 | 0.151 | 0.084 | 0.084 | 0.161 |
|  | SMOTE | 0.217 | 0.192 | 0.318 | 0.263 | 0.252 | 0.332 | 0.251 | 0.226 | 0.377 | 0.25 | 0.288 | 0.382 |
| MCC | None | 0.038 | 0.04 | 0.081 | 0.044 | 0.01 | 0.124 | 0.011 | 0.045 | 0.097 | 0.064 | NA | 0.101 |
|  | RUS | 0.166 | 0.143 | 0.3 | 0.189 | 0.186 | 0.309 | 0.11 | 0.07 | 0.274 | 0.133 | 0.159 | 0.285 |
|  | ROS | 0.073 | 0.105 | 0.252 | 0.05 | 0.059 | 0.268 | 0.064 | 0.043 | 0.191 | 0.066 | 0.104 | 0.202 |
|  | SMOTE | 0.153 | 0.132 | 0.313 | 0.209 | 0.205 | 0.331 | 0.1 | 0.075 | 0.265 | 0.109 | 0.142 | 0.271 |
| Resolution | 50kb |  |  |  |  |  |  | 100kb |  |  |  |  |  |
|  | Variable Reduction | LASSO |  |  | Elastic-Net |  |  | LASSO |  |  | Elastic-Net |  |  |
|  | Predictor Type | OC | OP | Distance | OC | OP | Distance | OC | OP | Distance | OC | OP | Distance |
| Accuracy | None | 0.861 | 0.857 | 0.861 | 0.861 | 0.861 | 0.861 | 0.865 | 0.787 | 0.865 | 0.865 | 0.863 | 0.867 |
|  | RUS | 0.565 | 0.208 | 0.639 | 0.556 | 0.573 | 0.64 | 0.563 | 0.559 | 0.619 | 0.561 | 0.482 | 0.61 |
|  | ROS | 0.852 | 0.228 | 0.866 | 0.858 | 0.858 | 0.866 | 0.863 | 0.784 | 0.867 | 0.864 | 0.701 | 0.845 |
|  | SMOTE | 0.734 | 0.206 | 0.734 | 0.741 | 0.68 | 0.727 | 0.699 | 0.61 | 0.709 | 0.708 | 0.518 | 0.667 |
| AUC | None | 0.598 | 0.489 | 0.8 | 0.622 | 0.608 | 0.802 | 0.561 | 0.549 | 0.656 | 0.554 | 0.519 | 0.66 |
|  | RUS | 0.605 | 0.523 | 0.753 | 0.626 | 0.64 | 0.756 | 0.587 | 0.52 | 0.666 | 0.59 | 0.513 | 0.645 |
|  | ROS | 0.602 | 0.531 | 0.821 | 0.642 | 0.644 | 0.818 | 0.592 | 0.527 | 0.712 | 0.595 | 0.492 | 0.663 |
|  | SMOTE | 0.602 | 0.519 | 0.745 | 0.609 | 0.626 | 0.744 | 0.585 | 0.526 | 0.666 | 0.591 | 0.534 | 0.642 |
| F1-Score | None | 0.01 | 0.02 | 0.021 | 0.01 | NA | 0.021 | NA | 0.039 | 0.021 | NA | NA | 0.099 |
|  | RUS | 0.28 | 0.252 | 0.372 | 0.289 | 0.299 | 0.374 | 0.254 | 0.219 | 0.307 | 0.256 | 0.218 | 0.296 |
|  | ROS | 0.029 | 0.256 | 0.128 | 0.029 | NA | 0.13 | NA | 0.051 | 0.06 | NA | 0.194 | 0.113 |
|  | SMOTE | 0.279 | 0.253 | 0.377 | 0.26 | 0.308 | 0.372 | 0.202 | 0.217 | 0.297 | 0.219 | 0.225 | 0.265 |
| MCC | None | 0.04 | 0.019 | 0.071 | 0.067 | NA | 0.071 | NA | -0.077 | 0.058 | NA | -0.021 | 0.142 |
|  | RUS | 0.114 | 0.057 | 0.269 | 0.132 | 0.149 | 0.272 | 0.081 | 0.025 | 0.17 | 0.084 | 0.007 | 0.152 |
|  | ROS | 0.009 | 0.073 | 0.189 | 0.043 | 0.03 | 0.188 | NA | -0.068 | 0.11 | NA | 0.029 | 0.066 |
|  | SMOTE | 0.134 | 0.061 | 0.258 | 0.114 | 0.162 | 0.251 | 0.037 | 0.031 | 0.156 | 0.06 | 0.026 | 0.107 |

# Figures

|  |
| --- |
| **Figure 1.** |
| **Figure 2.** |
| **Figure 3.** |
| **Figure 4.** |
| **Figure 11.** |
| **Figure 9.** |
| **Figure 10.** |

# Table Legends

## Table 1: List of Genomic Annoations.

The full list of genomic annotations used in downstream analyses. These annotations were the functional genomic elements used to predict the formation of TAD boundaries. They were obtained via ChIP-Seq experiments, mapped to human genome assembly hg19, and made available by the ENCODE Consortium.

## Table 2: Class Distributions

Summary of the class distributions for each re-sampling technique, across each resolution. All re-sampling techniques yielded completely balanced classes. That is, there was a 1:1 relationship between the majority and minority classes after sampling took place.

## Table 7: Model Performances at all Resolutions

Performance metrics comparing models using different sampling approaches for each of the 4 resolutions. Withing each resolution, the best value in each column for each performance metric is underlined to compare different sampling approaches, while the highest value in each row is highlighted in bold to compare different predictor types.

# Figure legends

## Figure 1: Model Construction

Diagram of the model construction used for downstream analysis. The linear genome was binned according to the resolution of the respective HiC experiment (either 5kb, 25kb, 50kb, or 100kb intervals). TAD boundaries were then flanked by 1-unit on either side of the boundary point. The unit flanking was indicative of the resolution that the domain data was obtained from (i.e. for 5kb resolution, 1 unit represents 5kb for a total flanking region of 10kb). The response vector Y used for classification was determined by whether or not a genomic bin overlapped with a flanked region. The positional coordinates of each functional genomic element, obtained from ENCODE, were then used to define the feature space of the models.

## Figure 2: Predictor Types

Diagram of the 3 predictor types considered when assessing the relationship between TAD boundaries and functional genomic elements. Each predictor type was used as the feature space in downstream analyses for predicting which functional genomic elements were associated with the formation of TAD boundaries. Featured above are bin-specific examples of the construction of each predictor type. (Left) The overlap count (OC) predictors were calculated by considering the total number of elemental regions that overlapped with each genomic bin. (Middle) The overlap percent (OP) predictors were calculated by dividing the sum of all feature widths within each specific bin and dividing by the total bin width (either 5, 25, 50, or 100 kilobases given resolution of boundary data). (Right) The distance predictors were calculated by measuring the distance (in base pairs) from the center of each genomic bin to the center of the nearest elemental region of interest. The two green segments directly below the rightmost enlarged figure represent the center of the overlapping regions defining the respective functional genomic element.

## Figure 3: Ensemble Framework/Model Building Pipeline

A diagram of the model building pipeline given an a combination of inputs from the ensemble framework. The data was split into a 7:3 training set to testing set ratio, the variable reduction technique of choice was implemented, and a random forest classification algorithm was performed. Each model was then validated on the same testing set.

## Figure 4: Class Imbalance

Barplots illustrating the class imbalance problem featured across each of the four different resolutions that were analyzed. Minority classes represent the number of genomic bins that contain a TAD boundary, while the majority classes represent the number of genomic bins that do not contain a TAD boundary.

## Figure 11: Model Performances for all Resolutions

Comparing model performances for TAD boundary data for all 4 resolutions. The four regions represent (from left to right; top to bottom) data at 5kb, 25kb, 50kb, and 100kb resolutions. Within each region, the rows represent the set of re-sampling techniques, while the columns represent the set of performance metrics that the models were evaluated on. Each plot compares 2 sets of models; one using LASSO regularization and one using Elastic-Net regularization. Within each set, each bar represents the performance of a model with a specific predictor type; either overlap counts (OC) in red, overlap percent (OP) in green, or distance in blue.

# Figure 9: Comparing Performances

Comparing performances of random forest classification algorithms, across TAD boundary data at each resolution for models using distance-type predictors, with elastic-net regularization and random under-sampling. Performances were aggregated by taking the average of each metric across 50 iterations of random under-sampling.

# Figure 10: Comparing Variable Importances

Variable importance plots for the top 15 most predictive functional genomic elements for each of the 4 different resolutions. The x-axis represents the standardized difference between the out-of-bag prediction accuracy after permuting each predictor variable, averaged across all trees. The greater the mean standardized difference, the more importance the predictor is to the model.

# Supplementary Tables and Figures

# Supplementary Legends

# References

Chawla, Nitesh V, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. “SMOTE: Synthetic Minority over-Sampling Technique.” *Journal of Artificial Intelligence Research* 16: 321–57.

Dali, Rola, and Mathieu Blanchette. 2017. “A Critical Assessment of Topologically Associating Domain Prediction Tools.” *Nucleic Acids Research* 45 (6). Oxford University Press: 2994–3005.

Dixon, Jesse R, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. 2012. “Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions.” *Nature* 485 (7398). Nature Publishing Group: 376.

Estabrooks, Andrew, Taeho Jo, and Nathalie Japkowicz. 2004. “A Multiple Resampling Method for Learning from Imbalanced Data Sets.” *Computational Intelligence* 20 (1). Wiley Online Library: 18–36.

Lieberman-Aiden, Erez, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. “Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome.” *Science* 326 (5950). American Association for the Advancement of Science: 289–93.

Phillips-Cremins, Jennifer E, and Victor G Corces. 2013. “Chromatin Insulators: Linking Genome Organization to Cellular Function.” *Molecular Cell* 50 (4). Elsevier: 461–74.

Rao, Suhas SP, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, et al. 2014. “A 3d Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping.” *Cell* 159 (7). Elsevier: 1665–80.

Symmons, Orsolya, Veli Vural Uslu, Taro Tsujimura, Sandra Ruf, Sonya Nassari, Wibke Schwarzer, Laurence Ettwiller, and François Spitz. 2014. “Functional and Topological Characteristics of Mammalian Regulatory Domains.” *Genome Research*. Cold Spring Harbor Lab.

Zou, Hui, and Trevor Hastie. 2005. “Regularization and Variable Selection via the Elastic Net.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2). Wiley Online Library: 301–20.

Zuin, Jessica, Jesse R Dixon, Michael IJA van der Reijden, Zhen Ye, Petros Kolovos, Rutger WW Brouwer, Mariëtte PC van de Corput, et al. 2014. “Cohesin and Ctcf Differentially Affect Chromatin Architecture and Gene Expression in Human Cells.” *Proceedings of the National Academy of Sciences* 111 (3). National Acad Sciences: 996–1001.