Computational and Structural Biotechnology Journal 19 (2021) 2870–2880

ELSEVIER

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

PredTAD: A machine learning framework that models 3D chromatin organization alterations leading to oncogene dysregulation in breast cancer cell lines

Jacqueline Chyr[a], Zhigang Zhang[b], Xi Chen[a], Xiaobo Zhou[a],*

[a] School of Biomedical Informatics, University of Texas Health Science Center, Houston, TX 77054, USA
[b] School of Information Management and Statistics, Hubei University of Economics, Wuhan, Hubei 430205 China

ARTICLE INFO

Article history:
Received 21 February 2021
Received in revised form 4 May 2021
Accepted 4 May 2021
Available online 7 May 2021

Keywords:
Chromatin organization
Topologically associated domains
TAD boundaries
Machine learning
Computational biology
Breast cancer
Epigenetics
Estrogen signaling pathways
3D chromatin architecture

ABSTRACT

Topologically associating domains, or TADs, play important roles in genome organization and gene regulation; however, they are often altered in diseases. High-throughput chromatin conformation capturing assays, such as Hi-C, can capture domains of increased interactions, and TADs and boundaries can be identified using well-established analytical tools. However, generating Hi-C data is expensive. In our study, we addressed the relationship between multi-omics data and higher-order chromatin structures using a newly developed machine-learning model called PredTAD. Our tool uses already-available and cost-effective datatypes such as transcription factor and histone modification ChIPseq data. Specifically, PredTAD utilizes both epigenetic and genetic features as well as neighboring information to classify the entire human genome as boundary or non-boundary regions. Our tool can predict boundary changes between normal and breast cancer genomes. Among the most important features for predicting boundary alterations were CTCF, subunits of cohesin (RAD21 and SMC3), and chromosome number, suggesting their roles in conserved and dynamic boundaries formation. Upon further analysis, we observed that genes near altered TAD boundaries were found to be involved in several important breast cancer signaling pathways such as Ras, Jak-STAT, and estrogen signaling pathways. We also discovered a TAD boundary alteration that contributes to RET oncogene overexpression. PredTAD can also successfully predict TAD boundary changes in other conditions and diseases. In conclusion, our newly developed machine learning tool allowed for a more complete understanding of the dynamic 3D chromatin structures involved in signaling pathway activation, altered gene expression, and disease state in breast cancer cells.

1. Introduction

The human genome consists of more than three billion nucleotides, spanning over two meters in length. In order to fit this vast genomic material in a small nuclear space, the chromatin undergoes multiple levels of organization. The genome is beautifully organized into fractal globular-like structures with extensive but specific looping and folding [1–4]. Furthermore, the chromatin is organized into sub-mega base regions with increased self-interaction frequency and decreased interactions with neighboring domains. The chromatin structure is dynamic and plays a critical role in targeted transcription and regulation of genes [5–7]. Unfortunately, in abnormal or cancer cells, the chromatin structure is altered, leading to aberrant gene regulation and disease states [8–11].

1.1. Chromatin organization and their epigenomic and genomic patterns

High-throughput chromosome conformation capture technologies such as 3C, 4C, ChIA-PET, and Hi-C have been developed to experimentally map long- and short-range chromatin interactions [12–14]. Large-scale 3D chromatin structures such as topologically associating domains (TAD) and TAD boundaries can be detected from analyzing these datatypes [1,15]. TADs are self-associating domains with increased interaction frequencies and TAD boundaries are insulating regions that separate two neighboring TADs [16]. There are a number of well-established TAD prediction tools,

* Corresponding author.
E-mail address: Xiaobo.Zhou@uth.tmc.edu (X. Zhou).

https://doi.org/10.1016/j.csbj.2021.05.013
2001-0370/Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.

such as Arrowhead [17] and TopDom [18]. TAD boundaries are important in gene regulation in that it insulates and prevents neighboring regions from interacting with each other. This region limits the interactions between the two adjacent TADs and may limit the interactions between enhancers and promoters [19–21].

It has been noted that chromatin structures are associated with TF binding, namely architectural-related protein CTCF, and histone modifications [19–21]. There is an enrichment of CTCF binding and transcription start sites of housekeeping genes at TAD boundaries. Previous studies have also reported distinct patterns of histone marks around TAD boundaries in H1 hESC and IMR90 cell lines [15,17,22,23]. Histone modifications (acetylation, methylation, phosphorylation and ubiquitination) serve as distinct markers for transcription regulation, and function to control the accessibility of the chromatin and the recruitment of DNA binding proteins. DNA methylation and DNA accessibility has also been known to affect TF binding and histone methylation [24–26]. Despite an enrichment of CTCF and housekeeping genes at TAD boundaries, distribution of CTCF alone is not sufficient to predict boundaries and housekeeping genes were only enriched in some TAD boundaries [22]. This suggests that computational models using a combination of ChIP-seq and other forms of data are needed to predict and understand large-scale chromatin structures.

1.2. Machine learning and computational tools for multi-omics data analyses

A number of studies have already begun to predict TAD interaction hubs and TAD boundaries using readily available ChIP-seq data [22,23,27]. One group developed a tool called HubPredictor which uses the computational classifier – Bayesian additive regression tree (BART) to successfully predict TAD boundaries in IMR90 with a good prediction accuracy (AUC = 0.77) [23]. Another group used a Bayesian Ridge algorithm to develop a computational model called Position-specific linear model (PSLM) which considered both position and density of various genomic elements. Then, they devised a heuristic search algorithm called Population Greedy Search Algorithm (PGSA) to identify the combination of genomic elements that best describe TAD boundaries. They also classified genomic segments as TAD boundaries or non-TAD boundaries with an AUC of 0.826 [22]. A third group developed a computational method called TAD-Lactuca which uses histone modification and CTCF ChIPseq data, as well as frequency of k-mers, to predict TAD boundaries from non-boundaries (AUC = 0.817 in IMR90 cell line) [28]. However, these studies have not considered a vast combination of both genomic and epigenomic data. They also did not create a universally applicable model and they did not consider alterations found in disease genomes. As an improvement to these previous studies, various epigenomic data such as DNA methylation, transcription factor binding, histone modifications, and chromatin accessibility as well as several genomic data such as transcription start site, gene density, and location information were utilized in our study. Not only that, but previous studies only focused on a very small portion of the genome: TAD boundaries and randomly selected non-TAD boundary regions at a 1:1 ratio. Alternatively, our study interrogates the entire genome by binning it into 10 kb regions. This allowed the most information to be retained while providing a reasonably high resolution. Feature information for each bin's neighbors were also included in our model.

Due to the sheer size and volume of the multi-omics "big" data (e.g. genomics, epigenomics), more powerful data science approaches are needed so that we can obtain deeper insights into living systems. There is a need to develop computational tools to study the 3D chromatin structures in human genomes. Thus, the advanced machine learning algorithm, Gradient Boosting Machine (GBM) [29,30], was applied to multi-omics data to classify a genomic region as TAD boundary or non-TAD boundary. GBM is a powerful classification method because it is based on an ensemble of decision trees. Unlike Random Forest (RF), which builds an ensemble of deep independent trees, GBM builds shallow trees one at a time with each new tree correcting the errors made by previously trained trees [29]. Other advantages of GBM include capability of handling missing data, not requiring pre-processed data, and being a better learner than RF. However, due to its sequential manner, GBM models take longer to train and are more sensitive to overfitting.

1.3. Breast cancer as a global challenge and epigenomic aberrations.

Breast cancer is a globally prevalent disease. The 5-year survival rate for people diagnosed with breast cancer ranges from 90% in high-income countries to about 66% in India and 40% in South Africa [31–34]. In the U.S., about 1 in 8 women will develop breast cancer in her lifetime [35]. It is estimated that there will be more than 280,000 new cases of invasive breast cancer and 49,000 new cases of non-invasive (in situ) breast cancer in 2021 [32]. There are currently more than 3.8 million women with a history of breast cancer in the U.S and more than 7.8 million women worldwide who were diagnosed with breast cancer in the past 5 years [33]. Although the overall death rate from breast cancer is decreasing by about 1% per year due to advances in treatment and early detection through screening [36–40], there is still an estimated 43,600 breast cancer related deaths in U.S. women in 2021 [35].

Only 5–10% of breast cancers are linked to inherited gene mutations with the most common being mutations in the BRCA1 and BRCA2 genes [37,41]. A number of cases rise from accumulated genetic mutations that occur as a result of aging process [42]. Other cases are not explained genetically, but may be a result of epigenetic alterations [43,44]. For example, DNA hypermethylation is significantly enriched in luminal B subtype and DNA hypomethylation is associated with basal-like subtypes [45]. Others have found that changes in histone acetylation and methylation patterns might represent an early sign of breast cancer [46]. Certain chromatin remodeling genes, such as the SWI/SNF gene, was shown to have altered levels of methylation, and these alterations could be involved in breast cancer signaling pathways such as TGF-β pathway silencing and overexpression of MYC [47].

In addition to changes in methylation, other epigenomic aberrations such as alterations 3D chromatin organization plays a role in breast cancer disease and progression [19,48–51]. However, the roles of 3D structure and its dynamic in breast cancer are largely unknown. Previous studies have observed associations between altered expression of ER-regulated genes and dynamic remodeling of ER pathways accompanying the development of endocrine resistance [19]. They have observed that loss of 3D chromatin interactions occurs coincidently with hypermethylation and loss of ER binding. Alterations in active and inactive chromatin compartments are also associated with atypical interactions and gene expression in breast cancer, suggesting that the 3D epigenome remodeling is a key mechanism in endocrine resistance in ER + breast cancer. Moreover, Barutcu et al. also identified changes in the 3D chromosome clustering between normal epithelial breast cells and breast cancer cells [52]. They found that small, gene-rich chromosomes displayed decreased interaction frequency with each other in breast cancer cells compared to chromosomes in normal epithelial cells. They also found that altered compartment regions are associated with up-regulation of genes related to the repression of WNT signaling and other signaling pathways. Furthermore, estrogen-induced chromatin decondensation and nuclear re-organization was linked to local epigenetic regulation in breast cancer [53]. Estrogen is linked to long-range changes in higher-order chromatin organization which suggests that 3D chromatin architecture could be used as a target for breast cancer treatment [50,53,54].

Taken all together, the 3D chromatin structure, namely TADs and TAD boundaries, are important in orchestrating proper enhancer-promoter activity and cell-specific gene expression. Alterations in chromatin structures in breast cancer leads to a dysregulation of enhancer-promoters interactions and promotion of tumor suppressive functions [54–57]. It is particularly important to address the 3D chromatin organization differences between normal and cancer genomes. Therefore, we developed a novel machine learning model called PredTAD to decipher the relationship between epigenomic and genomic features and higher-order chromatin structures. We hypothesize that epigenomic and genomic features from readily accessible data types such as ChIPseq can be used to predict TAD boundaries. We also hypothesize that our tool can uncover large-scale structural changes that contribute to breast cancer disease state. Our analysis of the chromatin organization offers an in-depth understanding of gene regulation, signaling pathways activation, and disease state.

2. Materials and methods

2.1. Hi-C data and TAD boundaries

Normal breast epithelial cell line data (MCF10A) and breast cancer luminal A subtype (ER+/Her2-) cell line data (MCF7 and T47D) were obtained from GEO and ENCODE. MCF10A and MCF7 Hi-C data was obtained from Barutcu et al GSE66733 [56]. Paired-end reads were mapped to hg19. Due to aneuploidy of MCF7 cells, iterative correction and eigenvalue decomposition (ICE) was performed as previously described [56,58]. TAD calling and TAD boundary identification was performed using the insulation square analysis at a 40 kb resolution as previously described [56]. The widths of the boundaries were extended by 80 kb on both side to account for the variation between replicates. The final width of the TAD boundaries spans 200 kb. Weak boundaries with a boundary strength of < 0.15 were excluded. A total number of 3305 and 3273 boundaries were used for MCF10A and MCF7, respectively. T47D TAD data was obtained from ENCODE (ENCFF437EBV). TAD boundaries were defined as the region between two TADs. Each TAD boundary region was standardized to 200 kb width. A total number of 3166 boundaries were used for T47D. High-depth GM12878 Hi-C data was obtained from GSE63525 [17]. Processed TAD data was used. TAD boundaries were defined as 200 kb regions in between two TADs. There are 3097 boundaries for GM12878.

2.2. Epigenomic and genomic elements

GM12878, MCF10A, MCF7, and T47D histone modification and DNA binding protein ChIP-seq data were obtained from the ENCODE project and GEO database [59,60]. DNA chromatin accessibility was determined by ATAC-seq, which was also obtained from GEO [61,62]. Fastq sequence files of ChIP-seq and ATAC-seq data were mapped to hg19 with BWA-MEM and narrow peaks were called with MACS2 [63,64]. If hg38 narrow peak files were used, the location of the peaks was converted to hg19 using UCSC's liftover. Average narrow peaks per 10 kb bin were normalized to 0–1 to allow for application of the trained model across different cell lines. MCF7 and MCF10A DNA methylation were obtained from ENCODE. T47D DNA methylation, also Illumina Methylation 450 K BeadChip array, was obtained from Uehiro, et al. GSE87177 [65]. GM12878 was obtained from GSE62111 [66]. All DNA methylation data were from Illumina Methylation 450 K BeadChip array. Transcription factor binding sites for 161 transcription factors were obtained from UCSC (wgEncodeRegTfbsClusteredV3 table) [60]. Location of coding, noncoding, and housekeeping genes were obtained from UCSC. Gene density was calculated based on the number of TSS of each gene in each genomic bin. Chromosome length and location of centromeres and telomeres were also obtained from UCSC. Distance to centromere is defined as the percentage of the chromosomal arm the region is in, where 0 is at the centromere and 1 is at the telomere.

2.3. PredTAD boundary prediction model

The entire genome is binned into 10 kb regions. Regions in the centromere and telomere were excluded from the study due to lack of Hi-C reads in these regions. Training and test sets were randomly selected at a 7:3 ratio per chromosome. Epigenomic and genomic information of each bin as well as information of neighboring bins were used as the features. Specifically, ten 10 kb bins to the left (or downstream) and ten 10 kb bins to the right (or upstream) were included as features. This method keeps spatial feature information instead of averaging out the signals in the 200 kb region. For ChIP-seq and ATAC-seq data, the mean signal values of narrowPeaks peaks per bin were used. For methylation data, the average methylation signal values of all CpG sites in each bin was used. For transcription factor binding sites and transcription start sites, number of each genomic element in each bin was counted.

The machine learning technique Gradient Boosting Machine or GBM was used to classify each 10 kb genomic bin as either TAD boundary or non-TAD boundary. The number of trees built in the models was set at 500 and the max depth was 10. Although deeper trees may provide better accuracy on a training set, it is prone to overfitting. In total, there are 3305 TAD boundaries for MCF10A, 3273 for MCF7, 3166 for T47D, and 3097 for GM12878. For MCF10A, there were 65,453 positive samples (10 kb regions within a TAD boundary) and 227,588 (10 kb regions not within a TAD boundary). For MCF7, there were 64,547 positive and 228,494 negative samples. For T47D, there were 64,784 positive samples and 231,867 negative samples. For GM12878, there were 63,325 positive samples and 232,714 negative samples. Any 10 kb regions partially located in a TAD boundary were also excluded.

2.4. Boundary alterations prediction

A modification of PredTAD was used to predict TAD boundary alterations between normal breast MCF10A cell line and breast cancer MCF7 or T47D cell lines. Similar to PredTAD, 10 kb regions were used as samples. Each region could be one of three classes: normal, cancer, or conserved. Normal boundaries are boundaries found in normal MCF10A but not in breast cancer MCF7 or T47D cell lines (also considered as MCF10A-specific boundaries). Cancer boundaries are defined as boundaries that are not in normal but are found in breast cancer cell lines (they are also considered as MCF7-specific boundaries or T47D-specific boundaries). Conserved boundaries are defined as boundaries found in MCF10A, MCF7, and T47D.

2.5. Gene expression analysis

MCF10A and MCF7 RNAseq libraries were obtained from Barutcu et al GSE71862 [56]. The libraries were generated with TruSeq Stranded Total RNA with Ribo-Zero Gold Kit and sequenced as 100-bp single-end reads using HiSeq 2000. Gene expression (transcripts per million) was quantified by RSEM v.1.2.7 as previously described [56,67]. T47D RNAseq library was obtained from Lee et al GSE142171. DNA library was generated using TruSeq Stranded Total RNA LT Sample Prep Kit (Gold) and sequenced with NovaSeq 6000 sequencer. Differentially expressed genes were defined as log2FC > 1 with a p-value < 0.05, n = 3 biological replicates. There were 6271 differentially expressed genes between MCF10A and MCF7 cell lines.

Estrogen related gene list was curated from analysis of two independent datasets: Lin CY, et al GSE11352 [68] and Nagarajan, et al. GSE55922 [69]. In the first dataset by Lin CY, et al, MCF7 cells were treated with 10 nM of 17β-estradiol (E2, estrogen) or vehicle control for 12 hrs. RNA was extracted and gene expression was profiled with Affymetrix U133 A and B GeneChips. Data was normalized using the Robust Multichip Average (RMA) normalization method as previously described [68,70]. Differentially expressed genes were defined as log2FC > 1 with an adjusted p-value < 0.2, n = 3 biological replicates. The second dataset is from Nagarajan, et al [69]. Here, MCF7 cells were treated with 10 nM of 17β-estradiol or vehicle control for 2 hrs. RNA was extracted and sequenced with HiSeq 2000 from Illumina. RNAseq was mapped to hg19 using Bowtie2 v.2.1.0 and differential expression was measured using DEseq v.1.14.0 as previously described [71,72]. Differentially expressed genes were defined as log2FC > 1 with a p-value < 0.05, n = 2 biological replicates. A total number of 521 unique genes were considered as estrogen-related genes.

Breast cancer patient RNAseq data was obtained from TCGA cohort. Luminal A (ER+/Her2-) patients were selected because MCF7 and T47D were both a luminal A breast cancer cell line. In total, there are 315 tumor samples and 40 matched normal breast samples. There were 3947 differentially expressed genes (log2FC > 1, p-value < 0.05) between breast cancer tumor samples and matched normal control.

3. Results

3.1. Epigenetic and genetic features in TAD boundaries

Similar to other cell lines, there is also an enrichment of certain TFs, histone modification marks, and TSS of housekeeping genes at TAD boundaries of MCF10A, MCF7, and T47D cells (Fig. 1). DNA methylation has been reported to be negatively correlated with transcription factor binding, namely the TAD boundary associated protein CTCF (Supplementary Fig. 1, p < 0.05) [26]. Thus, it is possible to utilize both epigenomic and genomic data in the modeling and prediction of TAD boundaries.

3.2. PredTAD predicts TAD boundaries in breast cancer cell lines using epigenomic and genomic information

PredTAD model overview is described in Fig. 2. Specifically, a machine learning technique called Gradient Boosting Machine, also known as Generalized Gradient Model or GBM, was built to classify genomic regions as boundaries or non-TAD boundaries. The entire genome was binned into 10 kb sample regions. Telomere and centromeres regions were excluded due to lack of Hi-C reads in these regions. In total, there were 296,651 genomic regions and 70% of the samples was used training while 30% were withheld for testing.

For epigenomic features, nine histone modifications (H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me3, and H4K20me1), 15 DNA binding proteins (CTCF, ELF1, EGR1, EP300, FOXA1, GABPA, GATA3, MAX, PML, POLR2A, RAD21, SIN3A, SMARCA5, SMARCE1, and SRF), DNA accessibility (ATAC-seq), and DNA methylation were used in the model. For genomic features, chromosomal location, relative distance on chromosome (determined by relative distance from the centromere), gene density (number of transcription start sites (TSS) of noncoding, coding, and housekeeping genes), and number of binding sites for 161 transcription factors were included. To improve accuracy, the information of each sample's neighboring regions was also included as features. That is, epigenomic and genomic features for each bin's ten upstream and ten downstream bins were included in PredTAD.

AUCs of 0.7954, 0.7993, and 0.7487 were attained for MCF10A, MCF7, and T47D, respectively. The top features for MCF10A were chromosome number, transcription factor binding sites of SMC3, RAD21, and CTCF, histone modifications (H3K9ac, H3K27me3, and H3K9me3), gene density of noncoding genes, relative distance on chromosome, and DNA methylation. For MCF7, the top features were also chromosome number, ChIP-seq and transcription factor binding sites of CTCF, RAD21, and SMC3, gene density of noncoding genes, DNA accessibility, and DNA methylation. T47D also had similar top features (Fig. 3). These results suggest that there are higher order chromatin structures that is dependent on not just CTCF and other chromatin remodeling proteins such as SMC3 and RAD21, but also the basic properties of the chromosome region (i.e. chromosome number and location), gene density, chromatin accessibility, and DNA methylation.

PredTAD was then re-evaluated without chromosome number as one of the features. Different independent training and testing datasets was created using chromosome random-split strategy. Using this method, training and testing datasets consisted of different chromosomes. Specifically, samples from every chromosome except chromosomes 1, 8, and 19 were used in training and samples from chromosomes 1, 8, and 19 were tested individually. The 5-fold cross validation AUC is 0.7166 for the training set and the test AUC is 0.7338, 0.75172, and 0.6081 for chromosomes 1, 8, and 19, respectively (Supplementary Fig. 2). Without chromosome information there is a slight decrease in accuracy, however, PredTAD is still capable of predicting TAD boundaries.

Next, selected features were used to re-train the model to evaluate whether a subset of features is sufficient in predicting TAD boundaries. Only top 5, 10, and 15 features were used to retrain PredTAD, which yielded a 5-fold cross validation AUC of 0.7101, 0.7860, and 0.7860 and a test AUC of 0.7236, 0.8034, and 0.8036, respectively. ROC curves were plotted, and top importance features were re-ranked (Supplementary Fig. 3). When chromosome number was removed as a feature from the top 15 subset, a 5-fold cross validation AUC of 0.7537 and a test AUC of 0.7615 were obtained.

3.3. Comparison to other models

PredTAD is better at predicting TAD boundaries compared to three other previously published methods: HubPredictor, PGSA, and TAD Lactuca. HubPredictor is based on Bayesian Additive Regression Trees (or BART) which is a "sum of trees" model that averages results from an ensemble of regression trees. However, their model only uses CTCF and histone modification ChIP-seq data. Additionally, they use 300 kb samples with no additional sub-bins. PGSA is a position-specific linear model (or PSLM) with regularized linear regression. It uses DNA binding and histone modification ChIP-seq as well as transcription start sites and transcription factor binding sites. They use 300 kb samples which are further divided into 11 sub-bins and the number of each genomic element in each sub-bin was counted. TAD-Lactuca uses 21 bins of 40 kb widths (840 kb region total). Only CTCF and 8 histone modifications were included as features. When HubPredictor, PGSA, and TAD-Lactuca were applied to the MCF7 cell line, lower AUCs were obtained (Fig. 4).

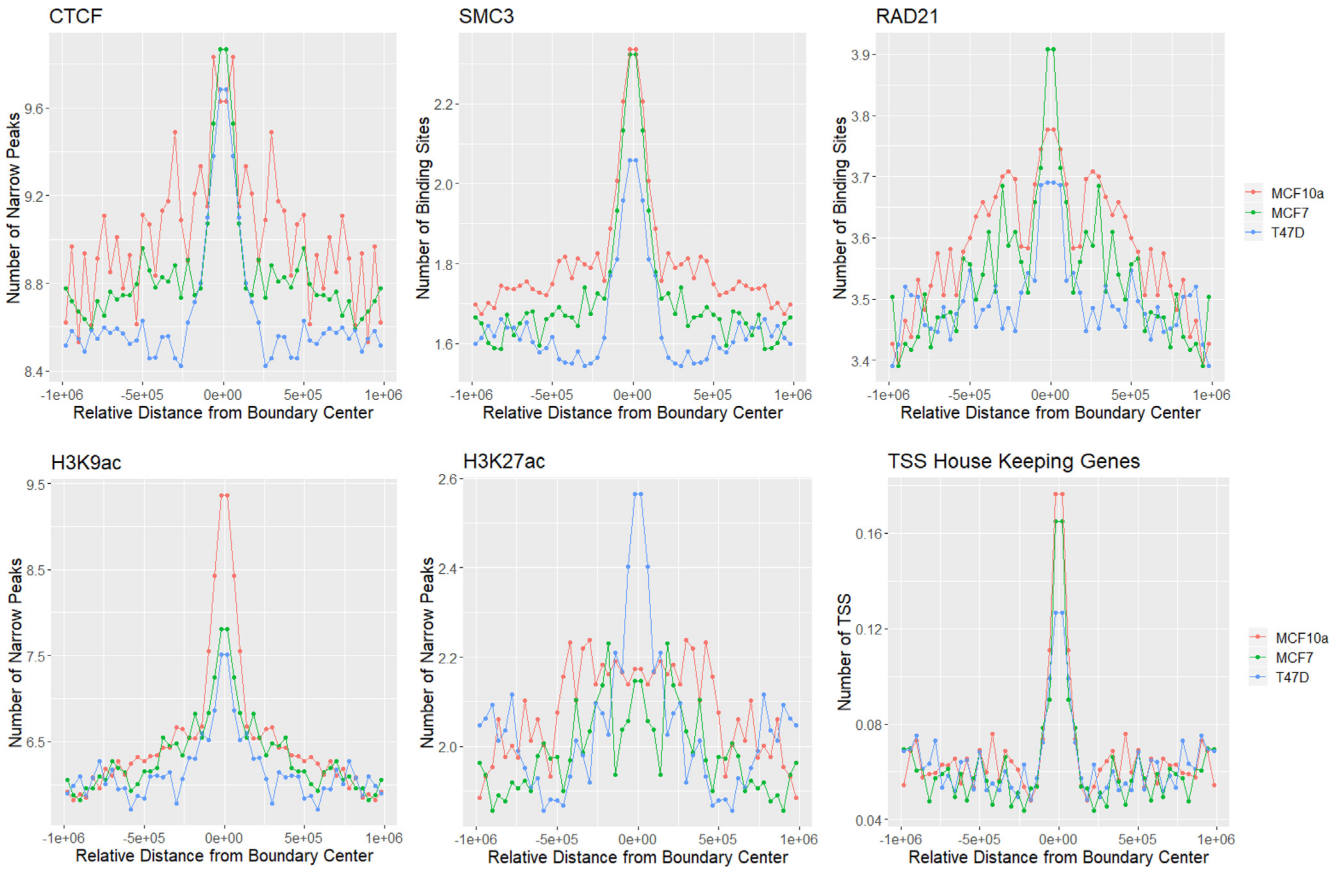


Fig. 1. Epigenomic and genomic features. Distribution of ChIPseq, TFBS, and TSSs centered on TAD boundaries for MCF10A, MCF7, and T47D cell lines (±1 MB).

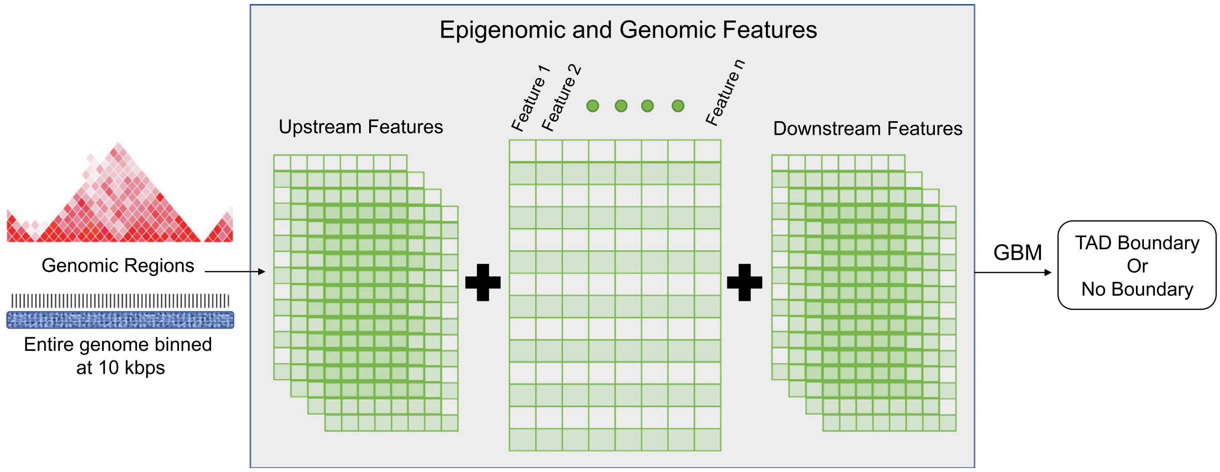


Fig. 2. Overview of PredTAD. The entire genome is binned into 10 kb bins. For each 10 kb bin, 192 epigenomic and genomic features were measured. Feature information for ten upstream and ten downstream bins were also included in the modeling. Gradient Boosting Machine (or GBM) was used to classify each 10 kb samples to either a TAD boundary or non-TAD boundary.

3.4. TAD boundary alterations between normal and breast cancer cells are controlled by CTCF, H3K9ac, RAD21, and SMC3

Higher-order chromatin organization is important for the proper regulation of genes. However, TAD and TAD boundaries are often perturbed in cancer [56,73]. Although majority of TAD boundaries in normal epithelial breast and breast cancer cell line overlapped, roughly 17–21% of the boundaries are different. There are 566–570 new TAD boundaries found in the breast cancer cell lines (MCF7 and T47D) that is not found in the normal epithelial breast cell line MCF10A. There is also a loss of 516 normal TAD boundaries (Fig. 5A). Studying altered TAD boundaries offers great insight on gene regulation and disease state of breast cancer.

To examine which features was most involved in TAD boundary changes between normal MCF10A cells and breast cancer MCF7 and T47D cells, a modified PredTAD model was used. The same epigenomic and genomic features were used to predict if a boundary region is gained, lost, or conserved. A boundary gain is a bound-

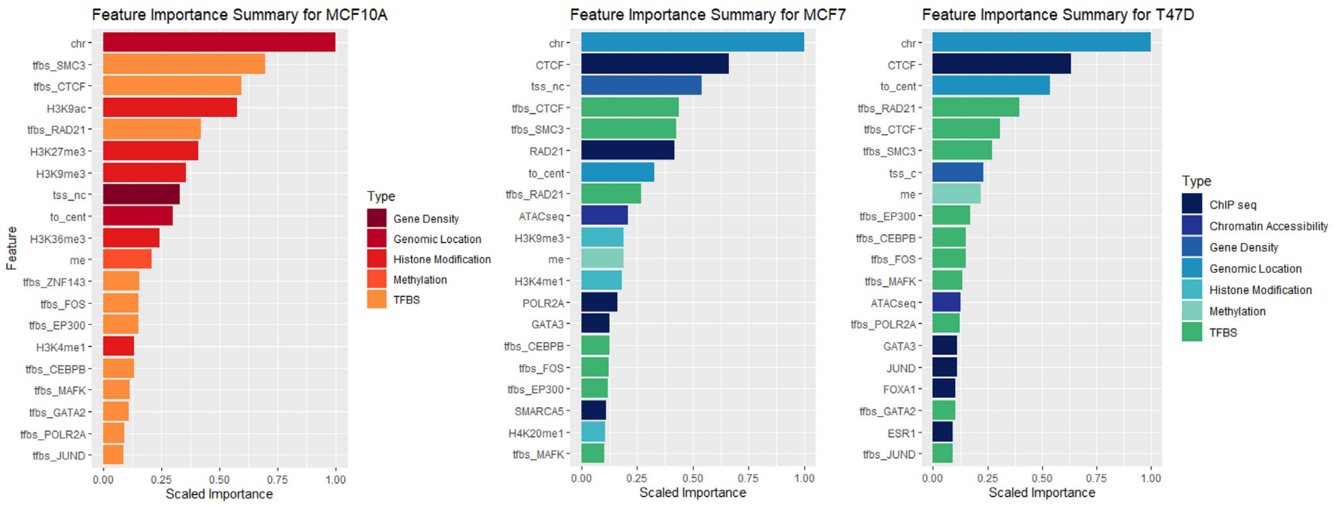
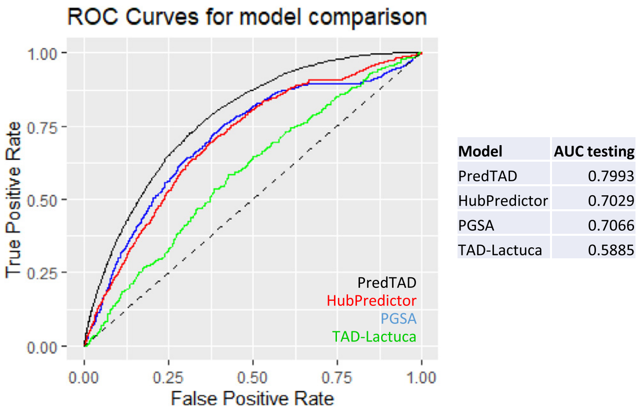


Fig. 3. PredTAD top predictive features. Top 20 most informative epigenomic and genomic features are listed for normal cell line MCF10A (left) and breast cancer cell lines MCF7 and T47D (right). Feature importance is determined by calculating the relative influence of each variable. Feature importance are scaled between 0 and 1.



Model	AUC testing
PredTAD	0.7993
HubPredictor	0.7029
PGSA	0.7066
TAD-Lactuca	0.5885

Fig. 4. Comparison of TAD prediction models. ROC curve is plotted for three TAD prediction models: PredTAD, PSLM, and BART. The AUC for testing sets are shown in the table.

ary found in breast cancer cell lines MCF7 and T47D but not in the breast normal cell line MCF10A. A TAD boundary loss is a boundary found in MCF10A but not in MCF7 or T47D cell lines. A conserved TAD boundary or "no change" is one that is found in all three cell lines. The 5-fold cross validation accuracy is 0.7545. The topmost important features are listed in Fig. 5B. The top most important features include CTCF, H3K4me1, H3K9ac, and H3K27me3. A closer analysis of these top features reveals that a number of these features are significantly enriched in conserved boundaries (Fig. 6). They include CTCF, H3K9ac, RAD21, and SMC3.

3.5. Altered chromatin organization contributes to oncogenic pathway activation in breast cancer cells

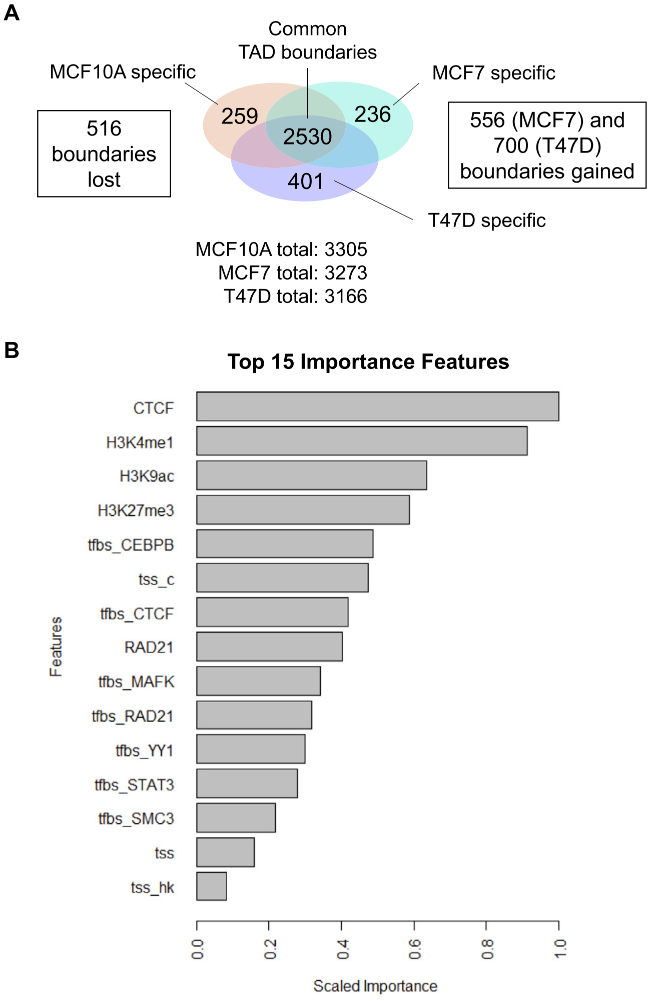


Fig. 5. A: TAD boundaries in normal (MCF10A) and breast cancer (MCF7 and T47D) cell lines. B: Top 15 most informative epigenomic and genomic features for predicting TAD boundary alterations between normal and breast cancer cell lines.

TAD boundary alterations can affect the expression of nearby genes. Genes near (±1 MB) MCF7 boundary gain and loss were examined. There are 6443 genes near gained boundaries. Of those genes, 1375 gene are differentially expressed in cell line RNAseq data (p-value < 0.05) and 108 are estrogen related genes. For lost boundaries (boundaries found in MCF10A but not in MCF7), 7298 genes are affected. T47D did not have RNAseq replicates for statistical analysis. Of those genes, 1551 genes are differentially expressed in cell line (p-value < 0.05) and 106 were estrogen related genes.

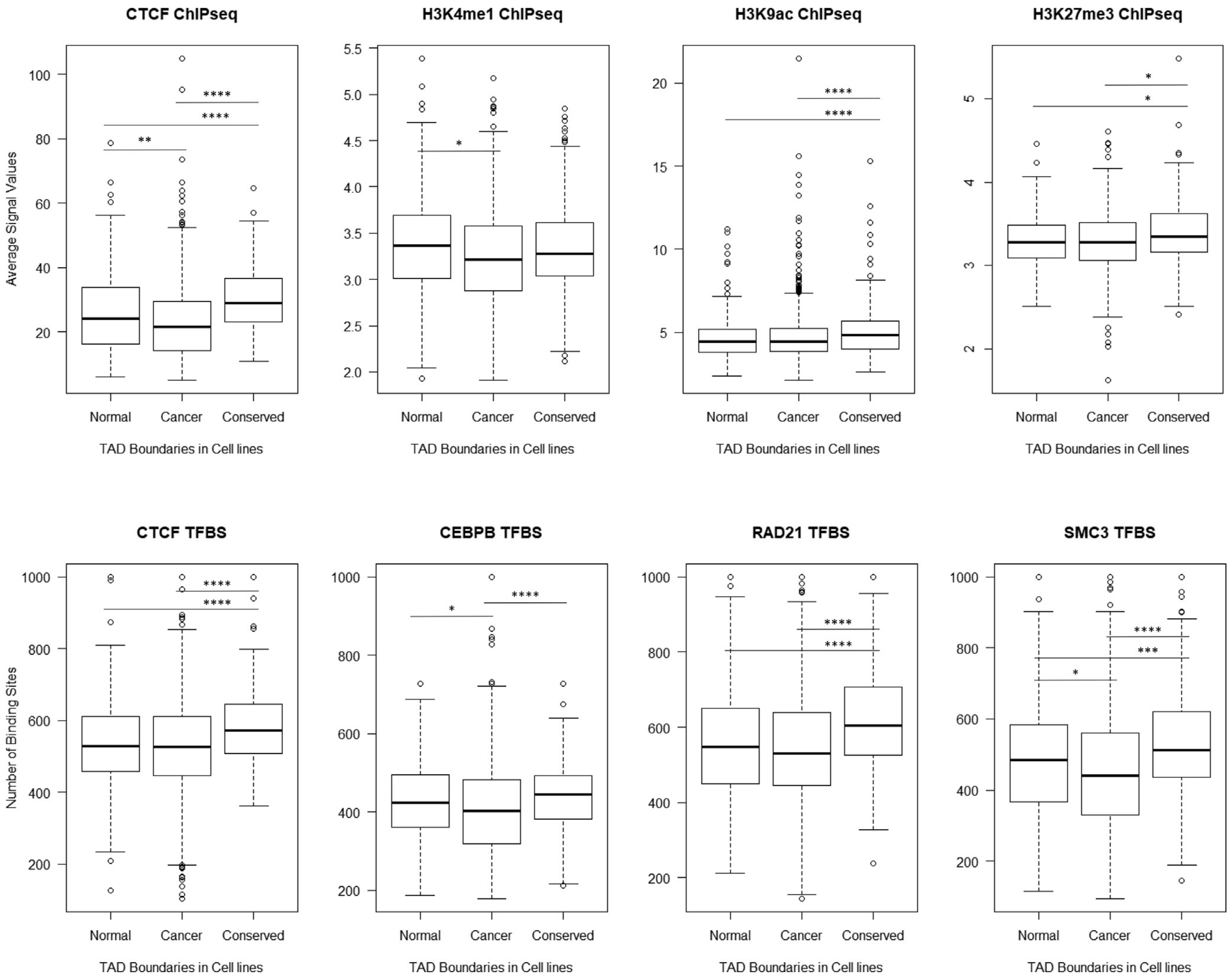


Fig. 6. Expression of features in conserved and perturbed TAD boundaries. (Top) Mean signal values for CTCF, H3K4me1, H3K9ac, and H3K27me3 ChIPseq are shown for boundaries found in MCF10A only (Normal), MCF7 or T47D only (Cancer), and in all three cell lines (Conserved). (Bottom) Transcription factor binding site counts for CTCF, CEBPB, RAD21, and SMC3 are shown for normal, cancer, and conserved TAD boundaries. TAD Boundary regions are 200 kb in length. Two-sided *t*-test was used to evaluate significance. * p < 0.05, ** p < 0.01, *** p < 0.001, **** p < 0.0001.

To examine which KEGG pathways these genes were involved in, David's Gene Ontology online tool was used [74]. KEGG PATHWAY is a collection of pathway maps on the molecular interaction, reaction, and relation networks for: metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, human diseases, and drug development [75]. The differentially expressed genes near gained boundaries were involved in pathways such as hippo signaling pathway, estrogen signaling pathway, and Ras and Jak-STAT signaling pathways. The differentially expressed genes near lost boundaries were involved in estrogen signaling pathway and metabolic pathways. The full list is shown in Table 1.

3.6. Loss of TAD boundary promotes RET oncogene expression and breast cancer

Out of the 1375 differentially expressed genes near boundary alterations, RET may play a vital role in breast cancer. RET is a receptor tyrosine kinase that phosphorylates PTK2/FAK1. PredTAD predicted a boundary loss in chr10:43,600,000–43,800,000 in both MCF7 and T47D cell lines. This boundary is upstream of the RET gene Fig. 7A. Closer analysis indicates low H3K4me1 ChIP-seq signals and DNA hypermethylation. As a result, the TAD boundary is lost and RET is significantly overexpressed in MCF7 cells compared to MCF10A cells (log2FC > 1, p-value < 0.05, n = 3). RET is also overexpressed in T47D cell line. There is also an increase in RET gene expression in TCGA breast cancer patients compared to matched normal breast tissue samples (log2FC > 1, p-value < 0.05) Fig. 7B. Overexpression of RET gene leads to a poor prognosis on breast cancer (Cox p-value = 0.038) Fig. 7C. There are several tyrosine kinase inhibitors that can either enhance sensitivity of breast cancer to tamoxifen therapy and/or re-sensitize tumors that have developed tamoxifen resistance [76].

3.7. Application of PredTAD in predicting TAD boundaries in other cell lines

Finally, PredTAD was trained using high-depth Hi-C data from GM12878 cell line and tested with breast cancer samples. Only commonly available and important epigenomic features (12 TF and 8 histone medication ChIPseq data and DNA methylation) as well as generic genomic features such as gene density and transcription factor binding sites was used in this model. A 5-fold AUC of 0.8082 and a testing AUC of 0.7471 was obtained. This

Table 1
Signaling pathway analysis of differentially expressed genes near boundary alterations.

KEGG Pathways for Genes near Boundaries Gained	p-Value
Hippo signaling pathway	9.7E-2
Estrogen signaling pathway	9.6E-2
Inflammatory mediator regulation of TRP channels	9.1E-2
GABAergic synapse	8.8E-2
Pathways in cancer	8.3E-2
Fc gamma R-mediated phagocytosis	8.3E-2
Ras signaling pathway	8.0E-2
Jak-STAT signaling pathway	7.5E-2
Chagas disease (American trypanosomiasis)	6.4E-2
Ether lipid metabolism	6.3E-2
Focal adhesion	6.3E-2
Alcoholism	5.6E-2
MAPK signaling pathway	5.4E-2
Serotonergic synapse	4.8E-2
Adrenergic signaling in cardiomyocytes	4.3E-2
Circadian entrainment	3.7E-2
Cysteine and methionine metabolism	3.1E-2
Steroid hormone biosynthesis	2.9E-2
KEGG Pathways for Genes near Boundaries Lost	**p-Value**
Estrogen signaling pathway	9.2E-2
Vasopressin-regulated water reabsorption	8.9E-2
Glioma	8.8E-2
Metabolic pathways	8.6E-2
Sphingolipid signaling pathway	8.3E-2
Insulin secretion	7.8E-2
Platelet activation	7.7E-2
Tryptophan metabolism	6.1E-2
Glutamatergic synapse	6.0E-2
Serotonergic synapse	5.0E-2
Hypertrophic cardiomyopathy (HCM)	4.8E-2
Inflammatory mediator regulation of TRP channels	4.4E-2
Regulation of lipolysis in adipocytes	4.3E-2
Non-small cell lung cancer	4.3E-2
RAP1 signaling pathway	4.2E-2
Mineral absorption	4.2E-2
GABAergic synapse	3.7E-2
Focal adhesion	3.5E-2

result suggests that PredTAD can be trained with one cell line and applied to cell lines of other diseases to predict TAD boundaries with high accuracy.

4. Discussion

Breast cancer is a highly prevalent disease, with more than 7.8 million cases worldwide [33]. Fortunately, the overall death rate has been steadily decreasing due to advances in early detection and treatments [36–40]. Breast cancer is a genetically and epigenetically heterogeneous disease. Mutations in the BRCA1 and BRCA2 have been linked to breast cancer, but this is not the sole driver for the development of breast cancer. Other characteristics of the genome, such as hyper- or hypo-methylation in gene-rich regions, abnormal chromatin remodeling, and miRNA-tumor suppressor silencing, can play a major role in breast cancer development and progression.

In our work, we focused on the prediction of 3D chromatin structures in breast cancer cells. The 3D chromatin organization plays an important role in gene regulation. Chromatin structures such as TAD and TAD boundaries regulate gene expression by dictating genomic interactions. To detect these structures, Hi-C data is often analyzed, however the generation of Hi-C and other high-throughput chromatin conformation capturing assays are expensive, time consuming, and not readily available. Previous studies have shown that TAD boundaries are enriched with certain factors such as the core architectural protein CTCF and house-keeping genes. In fact, TAD boundaries can be characterized by a combination of multiple epigenomic and genomic elements. Therefore, we developed a machine learning computational tool that uses the GBM algorithm to predict TAD boundaries in breast cancer cells using readily available and cost-effective ChIP and genetic information. More importantly, we applied our tool to study the epigenetic changes in a pervasive disease.

In our analysis, CTCF and histone modifications were important in the prediction of TAD boundaries. This finding was to be expected. What was a surprise, however, was that chromosome number was found to be one of the top predictive features in MCF10A, MCF7, and T47D TAD boundary predictions. We suspect the reasoning is because genes are not evenly distributed among the 23 pairs of chromosomes. For example, chromosomes 16–22, but not chromosome 18, are considered as small gene-rich chromosomes, and these tend to cluster together in the nucleus. A large portion of these chromosomes are often active and will have different combinations of histone marks compared to gene-poor chromosomes. Therefore, including chromosome number in our model increased the accuracy in our TAD boundary prediction.

We also noticed another one main difference between the top predictive features of MCF10A and MCF7: the type of feature. For MCF10A, it was the transcription factor binding sites of multiple genes, but in MCF7 and T47D, it was a mix of transcription factor binding sites and ChIP-seq mean signal values. This was not a surprise since transcription factor binding sites are generated from their corresponding proteins' ChIP-seq data and offer similar, redundant information. Since MCF10A is a normal cell line, the generic transcription factor binding sites were more informative, whereas for abnormal breast cancer cell lines which have genomic alterations, ChIP-seq signals were a more accurate representation of current status and thus, more informative.

The machine learning algorithm GBM was used in PredTAD because it is a good learner. It trains models in an additive and sequential manner – with each subsequent tree learning from the previous trees' errors. Parameters of 500 trees at a 10 max depth were selected to offer a deep model without overfitting. A higher number of trees increases the run time drastically and does not significantly increase the accuracy. A lower number of trees such as 100 trees decreases the run time but offer an AUC of around 0.77 compared to 0.80 for MCF7. When compared to other TAD boundary prediction methods, PredTAD performed the best. PredTAD utilizes more epigenomic and genomic features than the other prediction models. Moreover, much smaller bin sizes were considered which increases the resolution and prevents averaging high and low values together. Additionally, ten upstream and ten downstream neighboring bin's information were included in the prediction. When predicting TAD boundaries that were gained, lost, or conserved between MCF10A, MCF7, and T47D cell lines, SMC3 and RAD21 were among the most important features. This was not surprising since both SMC3 and RAD21 are part of the cohesin complex and plays important roles in gene regulation, DNA damage repair, and stabilization of the genome [77,78]. H3K9ac marks active enhancers and is highly correlated with active promotors [79,80].

One major advantage of PredTAD is its application to any cell type and conditions. Once the model is trained, we can apply it on other cells lines (using other cell line data) to predict TAD boundary formations. When trained with high depth GM12878 Hi-C data, PredTAD can be used to predict TAD boundaries in breast cancer cells and other cell lines under different conditions or diseases. This allows users to understand changes in chromatin organization without the need of generating Hi-C data. A limitation for our tool is the need for experimentally generated ChIPseq data. Furthermore, with advancements in next generation sequencing such as micro-C or single-cell Hi-C data [81,82], PredTAD can be fine-tuned to incorporate these data types.

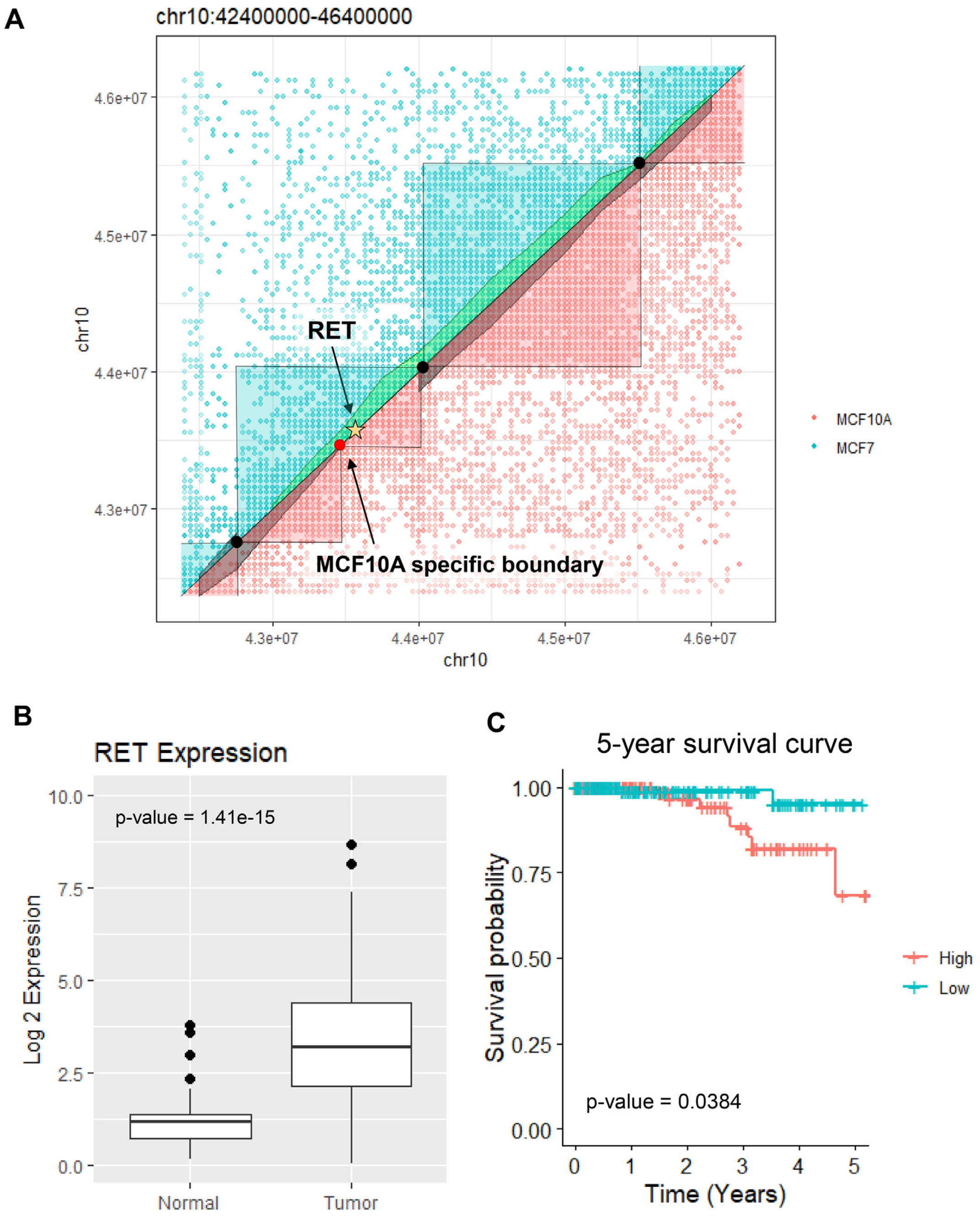


Fig. 7. RET gene in breast cancer. **A:** Hi-C heatmap of MCF10A and MCF7 is shown for the region around RET (chr10:41750000–46500000). Conserved TAD boundaries are indicated with black dots. MCF710A specific boundary is indicated with a red dot. The location of RET is marked with a star. **B:** RET log2 expression for TCGA breast cancer tumor samples and matched normal samples are shown (p-value = 1.4069e−15). **C:** Kaplan-Meier curves for TCGA breast cancer patients are shown (Cox regression p-value = 0.038). Patients were stratified into two groups: high and low expression of RET. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Further analysis revealed that altered chromatin organization is associated with a dysregulation of important genes and pathways. Changes in the 3D epigenome led to a significant up- or down-regulation of over a thousand genes, of which, over 100 were related to estrogen signaling pathway. Estrogen signaling are implicated in breast cancer progression, and many human breast cancers are estrogen-dependent. In fact, other studies have found that estrogen stimulation induces enhancer-promoter looping and a more global higher-order recompartmentalization of chromatin domains [53,83–86]. Endocrine therapy, such as estrogen receptor modulators and other aromatase inhibitors, have been used to treat ER-positive breast cancer and these treatments have substantially improved disease-free survival [87–89]. However, altered chromatin organization pattern and differential chromatin interactions are associated with endocrine resistance [19,85], suggesting that targeting cancer-related chromatin remodeling and 3D architecture may be a new avenue for breast cancer treatment [50,52–54,90].

5. Conclusions

Studying the 3D chromatin architecture of breast cancer genomes allows for a better understanding of cancer progression and cancer treatment resistances. Major characteristics of chromatin structure include TADs and TAD boundaries, and these features influence long-range chromatin interactions, which may regulate the expression of many key genes. The generation of high-throughput chromatin conformation data such as Hi-C is expensive. Fortunately, previous studies have found that chromatin structures are associated with a combination of genetic and epigenetic features such as histone modification and house-keeping genes. Thus, we developed PredTAD, a novel machine learning tool that accurately predicts TAD boundaries from non-boundaries using a diverse set of epigenomic and genomic features. In our work, we found that chromosome number was the most informative feature. When predicting conserved or perturbed boundaries in breast cancer cell lines, features such as CTCF, H3K4me1, H3K9ac, H3K27me3, RAD21, and SMC3 were found to be the most important. Further analysis revealed that genes near perturbed boundaries were involved in a number of oncogenic pathways including hippo signaling pathway, estrogen signaling pathway, and Ras and Jak-STAT signaling pathways. A breast cancer related boundary loss near the oncogene RET was identified. This was associated with an overexpression of RET in cell line and breast cancer patient data. RET has previously been implication in endocrine resistance and has an impact on tumor growth and metastasis [91–95]. In conclusion, studying chromatin organization offers a better understanding of gene regulation, signaling pathways activation, and disease state. Our work offers great insights on targeting 3D chromatin remodeling for breast cancer therapies.

6. Data availability

The datasets analyzed during this study are available on GEO (GSE85158, GSE89013, GSE101736, GSE71862, GSE11352, GSE 63525, GSE62111, and GSE55922), ENCODE, and TCGA. Code and resources can be found at https://github.com/jchyr-sbmi/PredTAD/.

CRediT authorship contribution statement

Jacqueline Chyr: Conceptualization, Methodology, Formal analysis, Writing - original draft, Writing - review & editing. **Zhigang Zhang:** Methodology, Formal analysis. **Xi Chen:** Formal analysis. **Xiaobo Zhou:** Supervision, Funding acquisition.

Acknowledgements

We greatly appreciate Dr. Zhou lab members for their valuable discussions.

Funding

This work was supported by National Institutes of Health [U01AR069395 to X.Z., R01GM123037 to X.Z., and U01CA166886 to X.Z].

Declaration of Competing Interest

The authors declare no conflicts of interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.05.013.

References

[1] Lieberman-Aiden E et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 2009;326(5950):289–93.

[2] Greenwald WW et al. Subtle changes in chromatin loop contact propensity are associated with differential gene regulation and expression. Nat Commun 2019;10(1):1–17.

[3] Xu H et al. Exploring 3D chromatin contacts in gene regulation: the evolution of approaches for the identification of functional enhancer-promoter interaction. Comput Struct Biotechnol J 2020;18:558–70.

[4] Dawson WK, Lazniewski M, Plewczynski D. Free energy-based model of CTCF-mediated chromatin looping in the human genome. Methods 2020;181:35–51.

[5] Hansen AS et al. Recent evidence that TADs and chromatin loops are dynamic structures. Nucleus 2018;9(1):20–32.

[6] Zhang H et al. Chromatin structure dynamics during the mitosis-to-G1 phase transition. Nature 2019;576(7785):158–62.

[7] Fyodorov DV et al. Emerging roles of linker histones in regulating chromatin structure and function. Nat Rev Mol Cell Biol 2018;19(3):192.

[8] Ando M et al. Chromatin dysregulation and DNA methylation at transcription start sites associated with transcriptional repression in cancers. Nat Commun 2019;10(1):2188.

[9] Du G et al. The hierarchical folding dynamics of topologically associating domains are closely related to transcriptional abnormalities in cancers. Comput Struct Biotechnol J 2021;19:1684–93.

[10] Ryu HY, Hochstrasser M. Histone sumoylation and chromatin dynamics. Nucleic Acids Res 2021.

[11] Du G et al. The hierarchical folding dynamics of topologically associating domains are closely related to transcriptional abnormalities in cancers. Comput Struct Biotechnol J 2021.

[12] Dekker J et al. Capturing chromosome conformation. Science 2002;295(5558):1306–11.

[13] Belton J-M et al. Hi–C: a comprehensive technique to capture the conformation of genomes. Methods 2012;58(3):268–76.

[14] Wang H et al. Discover novel disease-associated genes based on regulatory networks of long-range chromatin interactions. Methods 2021;189:22–33.

[15] Dixon JR et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 2012;485(7398):376–80.

[16] Dixon JR, Gorkin DU, Ren B. Chromatin domains: the unit of chromosome organization. Mol Cell 2016;62(5):668–80.

[17] Rao SS et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 2014;159(7):1665–80.

[18] Shin H et al. TopDom: an efficient and deterministic method for identifying topological domains in genomes. Nucleic Acids Res 2016;44(7):e70.

[19] Achinger-Kawecka J et al. Epigenetic reprogramming at estrogen-receptor binding sites alters 3D chromatin landscape in endocrine-resistant breast cancer. Nat Commun 2020;11(1):1–17.

[20] Robson MI, Ringel AR, Mundlos S. Regulatory landscaping: how enhancer-promoter communication is sculpted in 3D. Mol Cell 2019;74(6):1110–22.

[21] Schoenfelder S, Fraser P. Long-range enhancer–promoter contacts in gene expression control. Nat Rev Genet 2019;20(8):437–55.

[22] Hong S, Kim D. Computational characterization of chromatin domain boundary-associated genomic elements. Nucleic Acids Res 2017;45(18):10403–14.

[23] Huang J et al. Predicting chromatin organization using histone marks. Genome Biol 2015;16(1):162.

[24] Fuks F. DNA methylation and histone modifications: teaming up to silence genes. Curr Opin Genet Dev 2005;15(5):490–5.

[25] Tate PH, Bird AP. Effects of DNA methylation on DNA-binding proteins and gene expression. Curr Opin Genet Dev 1993;3(2):226–31.

[26] Wang H et al. Widespread plasticity in CTCF occupancy linked to DNA methylation. Genome Res 2012;22(9):1680–8.

[27] 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(1):110.

[28] Gan W et al. A computational method to predict topologically associating domain boundaries combining histone Marks and sequence information. BMC Genomics 2019;20(13):1–12.

[29] Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot 2013;7:21.

[30] Johnson R, Zhang T. Learning nonlinear functions using regularized greedy forest. IEEE Trans Pattern Anal Mach Intell 2014;36(5):942–54.

[31] McCormack V et al. Breast cancer survival and survival gap apportionment in sub-Saharan Africa (ABC-DO): a prospective cohort study. Lancet Global Health 2020;8(9):e1203–12.

[32] Howlader N et al. SEER cancer statistics review. National Cancer Inst 1975. 2008..

[33] Stoltenberg M et al. The central role of provider training in implementing resource-stratified guidelines for palliative care in low-income and middle-income countries: lessons from the Jamaica Cancer Care and Research Institute in the Caribbean and Universidad Católica in Latin America. Cancer 2020;126:2448–57.

[34] DeSantis CE et al. International variation in female breast cancer incidence and mortality rates. Cancer Epidemiol Prevent Biomarkers 2015;24(10):1495–506.

[35] Cokkinides V et al. American cancer society: Cancer facts and figures. Atlanta: American Cancer Society; 2005.

[36] Ramadan SZ. Using Convolutional Neural Network with Cheat Sheet and Data Augmentation to Detect Breast Cancer in Mammograms. Comput Math Methods Med 2020;2020.
[37] Howlader, N., et al., Lifetime risk (Percent) of dying from cancer by site and race/ethnicity: Female, Total US, 2014-2016. SEER Cancer Statistics Review, 1975. 2016.
[38] Qian S, Golubnitschaja O, Zhan X. Chronic inflammation: key player and biomarker-set to predict and prevent cancer development and progression based on individualized patient profiles. Epma J 2019;10(4):365–81.
[39] Zubor P et al. Why the gold standard approach by mammography demands extension by multiomics? Application of liquid biopsy miRNA profiles to breast cancer disease management. Int J Mol Sci 2019;20(12):2878.
[40] Crigna AT et al. Cell-free nucleic acid patterns in disease prediction and monitoring—hype or hope?. EPMA J 2020:1–25.
[41] Palmer JR et al. Contribution of germline predisposition gene mutations to breast cancer risk in African American women. JNCI: J Natl Cancer Inst 2020;112(12):1213–21.
[42] Antoniou A et al. Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: a combined analysis of 22 studies. Am J Human Genet 2003;72 (5):1117–30.
[43] Jovanovic J et al. The epigenetics of breast cancer. Mol Oncol 2010;4 (3):242–54.
[44] Byler S et al. Genetic and epigenetic aspects of breast cancer progression and therapy. Anticancer Res 2014;34(3):1071–7.
[45] Network CGA. Comprehensive molecular portraits of human breast tumours. Nature 2012;490(7418):61.
[46] Elsheikh SE et al. Global histone modifications in breast cancer correlate with tumor phenotypes, prognostic factors, and patient outcome. Cancer Res 2009;69(9):3802–9.
[47] Rodenhiser DI et al. Epigenetic mapping and functional analysis in a breast cancer metastasis model using whole-genome promoter tiling microarrays. Breast Cancer Res 2008;10(4):1–15.
[48] Huang J et al. BAT Hi-C maps global chromatin interactions in an efficient and economical way. Methods 2020;170:38–47.
[49] Kempf N et al. Analysis of Cellular EMT States Using Molecular Biology and High Resolution FISH Labeling. In: The Epithelial-to Mesenchymal Transition. Springer; 2021. p. 353–83.
[50] Feng Y, Liu X, Pauklin S. 3D chromatin architecture and epigenetic regulation in cancer stem cells. Protein Cell 2021:1–15.
[51] Feng Y, Pauklin S. Revisiting 3D chromatin architecture in cancer development and progression. Nucleic Acids Res 2020;48(19):10632–47.
[52] Barutcu AR et al. Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. Genome Biol 2015;16(1):1–14.
[53] Rafique S et al. Estrogen-induced chromatin decondensation and nuclear re-organization linked to regional epigenetic regulation in breast cancer. Genome Biol 2015;16(1):1–19.
[54] Li L et al. Cancer is associated with alterations in the three-dimensional organization of the genome. Cancers 2019;11(12):1886.
[55] Guo Y et al. CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell 2015;162(4):900–10.
[56] Barutcu AR et al. Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. Genome Biol 2015;16(1):214.
[57] Hnisz D, Day DS, Young RA. Insulated neighborhoods: structural and functional units of mammalian gene control. Cell 2016;167(5):1188–200.
[58] Imakaev M et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods 2012;9(10):999–1003.
[59] Franco HL et al. Enhancer transcription reveals subtype-specific gene expression programs controlling breast cancer pathogenesis. Genome Res 2018;28(2):159–70.
[60] Rosenbloom KR et al. ENCODE data in the UCSC Genome Browser: year 5 update. Nucleic Acids Res 2013;41(Database issue):D56–63.
[61] Liu Y et al. Identification of breast cancer associated variants that modulate transcription factor binding. PLoS Genet 2017;13(9):e1006761.
[62] Porter JR et al. Global inhibition with specific activation: how p53 and MYC redistribute the transcriptome in the DNA double-strand break response. Mol Cell 2017;67(6):1013–25.
[63] Li, H., Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997, 2013.
[64] Zhang Y et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol 2008;9 (9):R137.
[65] Uehiro N et al. Circulating cell-free DNA-based epigenetic assay can detect early breast cancer. Breast Cancer Res 2016;18(1):129.
[66] Moen EL et al. Characterization of CpG sites that escape methylation on the inactive human X-chromosome. Epigenetics 2015;10(9):810–8.
[67] Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinf 2011;12(1):323.
[68] Lin CY et al. Whole-genome cartography of estrogen receptor alpha binding sites. PLoS Genet 2007;3(6):e87.
[69] Nagarajan S et al. Bromodomain protein BRD4 is required for estrogen receptor-dependent enhancer activation and gene transcription. Cell Reports 2014;8(2):460–9.
[70] Bolstad BM et al. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003;19 (2):185–93.
[71] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol 2010;11(10):R106.
[72] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9(4):357–9.
[73] Taberlay PC et al. Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. Genome Res 2016;26(6):719–31.
[74] Dennis Jr G et al. DAVID: database for annotation, visualization, and integrated discovery. Genome Biol 2003;4(5):P3.
[75] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28(1):27–30.
[76] Garcia R et al. Constitutive activation of Stat3 by the Src and JAK tyrosine kinases participates in growth regulation of human breast carcinoma cells. Oncogene 2001;20(20):2499–513.
[77] Strunnikov AV, Jessberger R. Structural maintenance of chromosomes (SMC) proteins: conserved molecular properties for multiple biological functions. Eur J Biochem 1999;263(1):6–13.
[78] Deardorff MA et al. Mutations in cohesin complex members SMC3 and SMC1A cause a mild variant of cornelia de Lange syndrome with predominant mental retardation. Am J Hum Genet 2007;80(3):485–94.
[79] Guertin MJ et al. Accurate prediction of inducible transcription factor binding intensities in vivo. PLoS Genet 2012;8(3):e1002610.
[80] Danko CG et al. Identification of active transcriptional regulatory elements from GRO-seq data. Nat Methods 2015;12(5):433–8.
[81] Ardakany AR et al. Mustache: multi-scale detection of chromatin loops from Hi-C and Micro-C maps using scale-space representation. Genome Biol 2020;21(1):1–17.
[82] Burgess DJ. Chromosome structure at micro-scale. Nat Rev Genet 2020;21(6). 337-337.
[83] Hsu P-Y et al. Estrogen-mediated epigenetic repression of large chromosomal regions through DNA looping. Genome Res 2010;20(6):733–44.
[84] Fullwood MJ et al. An oestrogen-receptor-α-bound human chromatin interactome. Nature 2009;462(7269):58–64.
[85] Zhou Y et al. Temporal dynamic reorganization of 3D chromatin architecture in hormone-induced breast cancer and endocrine resistance. Nat Commun 2019;10(1):1–14.
[86] Mourad R et al. Estrogen induces global reorganization of chromatin structure in human breast cancer cells. PLoS ONE 2014;9(12):e113354.
[87] Bian L, Xu F-R, Jiang Z-F. Endocrine therapy combined with targeted therapy in hormone receptor-positive metastatic breast cancer. Chin Med J 2020;133 (19):2338.
[88] Leary A, Dowsett M. Combination therapy with aromatase inhibitors: the next era of breast cancer treatment?. Br J Cancer 2006;95(6):661–6.
[89] Lewis-Wambi JS, Jordan VC. Treatment of postmenopausal breast cancer with selective estrogen receptor modulators (SERMs). Breast Disase 2006;24 (1):93–105.
[90] Kaur J, Daoud A, Eblen ST. Targeting chromatin remodeling for cancer therapy. Curr Mol Pharmacol 2019;12(3):215.
[91] Andreucci E et al. Targeting the receptor tyrosine kinase RET in combination with aromatase inhibitors in ER positive breast cancer xenografts. Oncotarget 2016;7(49):80543.
[92] Gattelli A et al. Ret inhibition decreases growth and metastatic potential of estrogen receptor positive breast cancer cells. EMBO Mol Med 2013;5 (9):1335–50.
[93] Morandi A, Plaza-Menacho I, Isacke CM. RET in breast cancer: functional and therapeutic implications. Trends Mol Med 2011;17(3):149–57.
[94] Mechera R et al. Expression of RET is associated with Oestrogen receptor expression but lacks prognostic significance in breast cancer. BMC cancer 2019;19(1):1–10.
[95] Plaza-Menacho I et al. Targeting the receptor tyrosine kinase RET sensitizes breast cancer cells to tamoxifen treatment and reveals a role for RET in endocrine resistance. Oncogene 2010;29(33):4648–57.