Fall 2021

Effective Cancer Detection Using Higher-Order Genome Architecture and Chromatin Interactions

My Chung
San Jose State University

EFFECTIVE CANCER DETECTION USING HIGHER-ORDER GENOME ARCHITECTURE AND CHROMATIN INTERACTIONS

A Master Project

Presented to

The Faculty of the Department of Computer Science

San Jose State University

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science

by

My Xuan Chung

December 2021

Project Committee Members

Wendy Lee, Ph.D	Department of Computer Science
Carlos Rojas, Ph.D	Department of Computer Engineering
William Andreopoulos, Ph.D	Department of Computer Science

ABSTRACT

EFFECTIVE CANCER DETECTION USING HIGHER-ORDER GENOME ARCHITECTURE AND CHROMATIN INTERACTIONS

By My Xuan Chung

Cancer is a complex disease which requires interactions between cell-intrinsic alterations and tumor microenvironment. The connection between epigenetics and genomic structure plays a key role in chromatin interaction which promotes enhancer-promoter interactions for transcriptional activities. Alterations of chromatin states in oncogenic signaling pathway potentially cause cancer cell-intrinsic changes and inappropriate instructions to normal cell cycles, leading to abnormal cell growth. Resulting phenotypic changes are correlated to underlying changes in higher-order chromatin structure such as topologically associating domains (TADs) and compartments. In cancer cells, TAD structure is usually altered to facilitate the communication between enhancers and promoters in addition to higher density of histone modification level, thus increasing transcriptional super-enhancer activities within certain boundary strengths. Strong insulation scores and boundaries indicate high boundary strength (boundary IV) which allows more intra-TAD interactions. High level of histone activating mark H3K27ac positioning near promoters increases transcriptional activity and gene expression. Therefore, spatial chromosomal structures by TADs and epigenetic markers are the key regulators of chromatin interactions in oncogenic activities from carcinogenesis to metastasis. The result indicates that XGBoost multi-class classifier has achieved the highest accuracy of 81.13% in classifying normal and cancer cell lines based on chromatin interactions, followed by Random Forest at 73.76% and TabNet classifier at 73.50%. The detection model could be further improved with high quality data sources and meaningful features for clinical applications in early-stage cancer detection and prognosis.

ACKNOWLEDGMENTS

I would like to sincerely thank Dr. Wendy Lee, Dr. Carlos Rojas, and Dr. William Andreopoulos for their thoughtful discussions and valuable feedback on the project during weekly meetings and virtual communication.

TABLE OF CONTENTS

I. Introduction

Chromosome structures at different developmental stages allow certain genes to be expressed by altering the accessibility of DNA segments for transcription. Chromatins fold into 3D organization of higher-order and primary order by organizing linear DNA with protein histones to pack the genome and fits it into the cell nucleus. Epigenetic alterations such as histone modification and changes in chromatin structures can upregulate or downregulate gene expressions. In the higher-order section, AB compartments consist of topologically associating domains (TADs) as active and inactive regions. Active TADs are rich in genes, open chromatin marks and transcription factors while inactive TADs contain few genes. TADs also consist of intra-chromosomal (cis) interactions with functional domains such as regulatory regions including enhancers and promoters (Fig. 1) [1]. Because gene expression and biological functions rely on the chromatin interactions between the higher-order and primary order, disruptions in TADs could lead to improper gene regulation and thus disease formation.

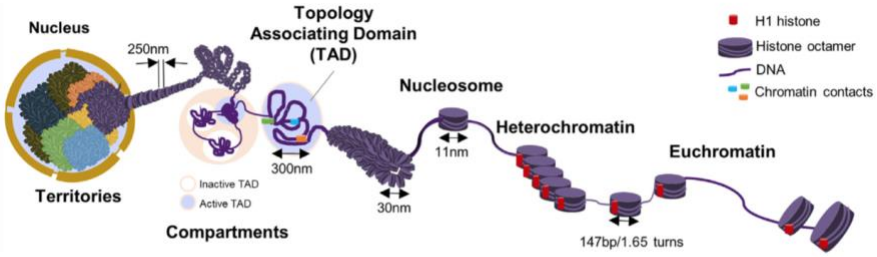

Fig. 1. Genome organization adapted from Chang et al. [1]. Higher-order structure includes compartments and TADs; primary-order structure consists of heterochromatin and euchromatin.

TADs are formed by loop extrusion in which cohesin rings made up of RAD21, SMC1, and SMC3 load onto DNA segment to generate a loop(s) and stop loading at CCCTC-binding factor (CTCF) anchor (Fig. 2). CTCF is a highly conserved zinc finger protein which serves as an insulator in locus control region to allow for

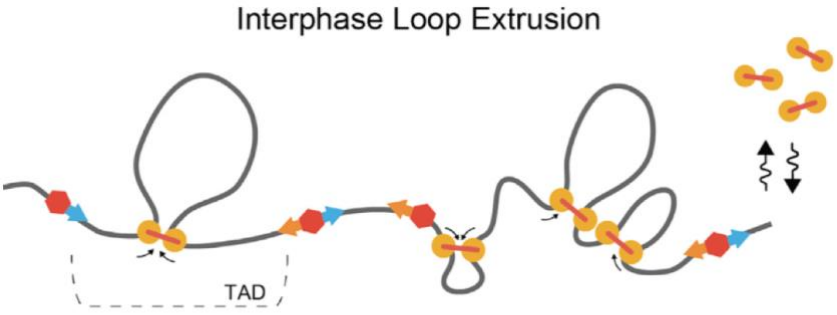

Fig. 2. Loop extrusion mechanism adapted from Fudenberg et al. [2]. Cohesin rings loop through gene segments. TADs are formed in between CTCF anchors.

more intra-TAD interactions. CTCF loop loss is caused by the depletion of either CTCF, cohesin, or both. TAD boundaries are created at the point where cohesin is blocked by CTCF anchor, and within each TAD. There could be one or more sub-TADs identified as loops, part of active and inactive compartmental domains. More enhancer-promoter interactions and increased gene expression are usually found in CTCF loop domains [2,3]. While CTCF protein is found to be stable in different developmental cycles or regulatory events, loops are more dynamic under those cycles or treatment conditions. According to Rao et al. (2017), loops were lost when HCT-116 cells were being treated with auxin, and loops were reformed after auxin removal, indicating their flexibility within genome structure. Loop domains are formed in the presence of cohesin, thus being categorized as cohesin-associated loops (CA-loop) whereas inter-chromosomal links are only detected at cohesin-independent loops (CI-loop). CTCF binding level is at 90% for CA-loops but much lower at 20% for CI-loop. CA-loops increase promoter activation by distal enhancers, and loss of loop causes super-enhancer colocalization to form links. Super-enhancers which are found within CA-loops plus high density of H3K27 acetylation tend to upregulate gene expression, but super-enhancers found at inter-chromosomal links tend to downregulate the expression of nearby genes at the strengthening event of links after cohesin loss [4].

Structures of TADs are made up of compartments, CTCF anchors, and loops. While it is basic to include CTCF and loops to define TADs, compartments play important role in the understanding of TAD interactions in a larger scale. Besides, contact domains, sub-TADs are classified into CTCF loop domains in which CTCF anchors halt the loop extrusion and ordinary domains which are not bound by CTCF proteins but specified by certain histone marks. There are two proposed models in the interrelation between compartments and TADs (Fig. 3). In the current model (left image), there are multiple sub-TADs with higher interaction frequencies (darker red triangles) resided within compartments. A few CTCF loops presented at the TAD boundaries while others presented along the sub-TADs, indicating strong interactions between CTCF sites. The new model expands the chromatin structure to identify TADs using compartmental domains in addition to cohesion-CTCF anchors and

loop extrusion events. The overall triangle in the right image corresponds to the tiny domain marked by the black arrow in the left image using high-resolution binning size of 1-5kb of Hi-C contact map. The directions of CTCF sites are more abundant in this model and define CTCF loop domains from ordinary domains with different orientations of colored arrowheads. Furthermore, the structure of TADs is varied by CTCF loop domains spanning or encompassing active and inactive compartmental domains as shown in the right image [3].

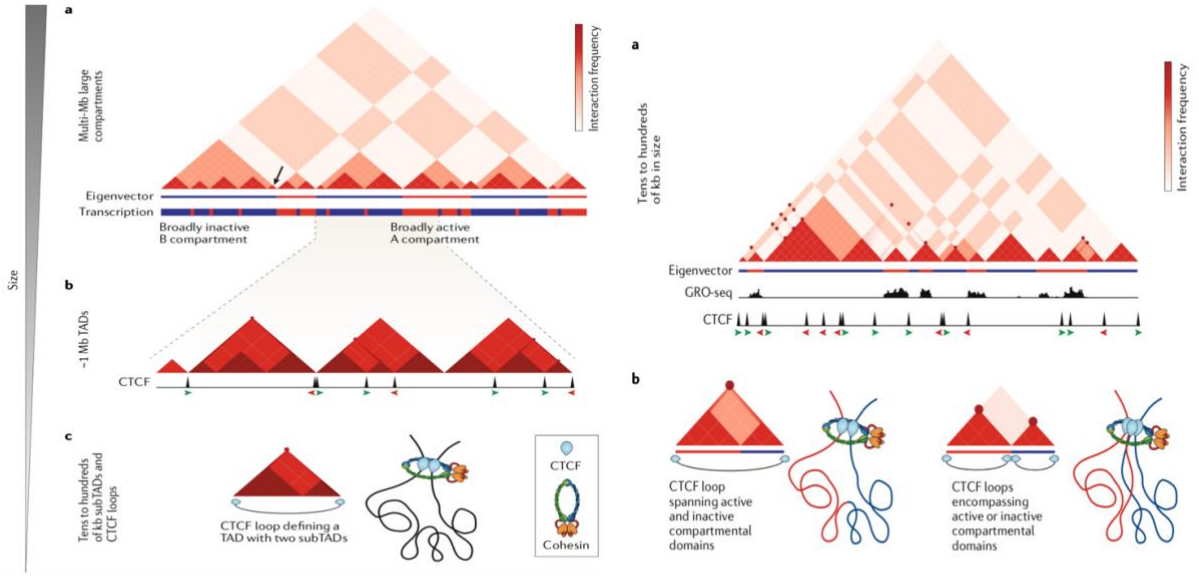

Fig. 3. The current model (left) and the new model (right) of chromatin organization, adapted from Rowley et al. [3]. In the current model, TADs are substructures inside compartment. In the new model, TADs span or encompass compartmental domains.

CTCF and cohesin ring have different effects on TADs structure in conjunction with compartmental domains. In the presence of both CTCF and cohesin, loop extrusion occurs and halted by CTCF to contain active and inactive domains, thus increasing intra-compartmental interactions (Fig. 4a). Loop extrusion continues even in the absence of CTCF protein, and the cohesin ring holds both compartmental domains and thus interaction frequencies remain intact (Fig. 4b). The absence of cohesin ring allows compartmental domains to be segregated so that they can anticipate more interaction with neighboring compartments in the presence of CTCF (Fig. 4c) [3].

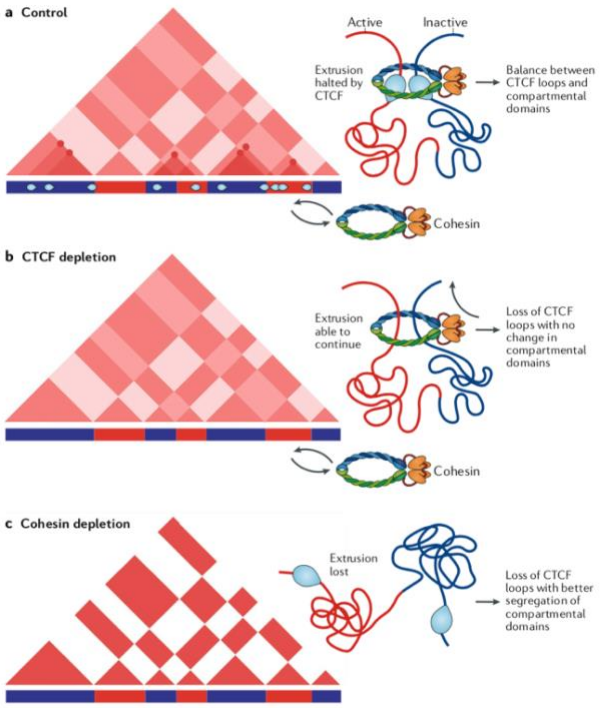

Fig. 4. Effects of CTCF and cohesin in 3D chromatin organization adapted from Rowley et al. [3].

(a) CTCF-cohesin loops spanning active and inactive compartmental domains.

(b) CTCF loop loss but cohesin ring remains to maintain compartmental domains.

(c) Cohesin absence causes extrusion loss and distinct compartmental domains.

While the triangular heat map facilitates the visualization of loops, sub-TADs, and compartmental domains corresponding to CTCF tracks, histone tracks, or other quantitative tracks for boundary strength, insulation scores, and directionality index, the square plots of contact matrices provide broader and extended views of contact domains and sub-compartments in interested chromosomal regions. Hi-C contact maps from *in situ* Hi-C experiment exhibited the maps in three resolutions - 500kb, 50kb, and 5kb, in which 5kb binning size gives the best resolution while the lower resolution map is binned to larger size such as 500kb, 800kb, 1Mb, etc. In other words, contact maps could be binned from 1Mb (lower resolving power) down to 5kb or 1kb (higher resolving power) to improve resolution of contact domains along the axis. The 5kb heat map revealed some CTCF loop domains with tiny dots positioning away from the diagonal. However, types of contact domains became harder to distinguish at lower resolution (50kb and 500kb), but sub-compartments were visually detected in the background of the low-resolution maps. The squares along the diagonal axis

are enriched as contact domains (enrichment) which can be classified into CTCF loop domains and ordinary domains, as shown the drawing version of the heat map (right image) (Fig. 5). Contact depletion occurs at the TAD boundaries or intersections between contact enrichments. Sharp valley at the depletion between contact domains suggests strong TAD border whereas gradual valley with successive tiny triangles in between domains indicates weak TAD border. Loops are presented as dots, and squares with dots are loop domains while squares without dots are ordinary domains [5].

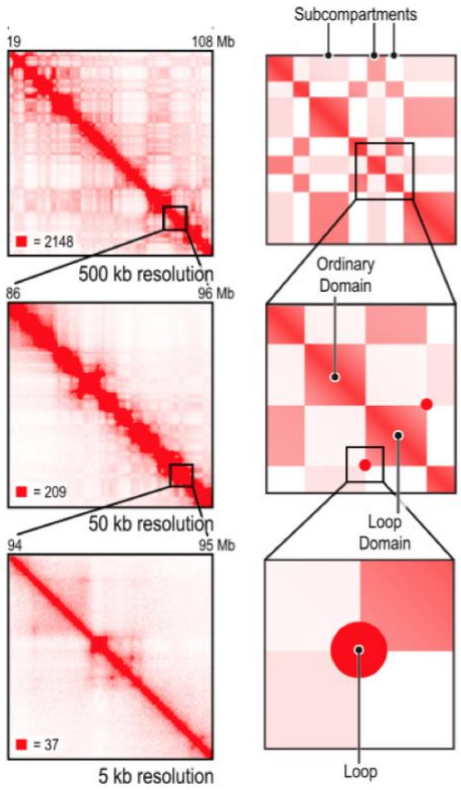

Fig. 5. Hi-C contact matrices of chromatin structure adapted from Rao et al. [5]

Left image: Contact maps of 500kb – 50kb – 5kb binning resolutions in which 5kb bin size displays the very-high-resolution map.

Right image: Drawings of Hi-C maps. Squares along the diagonal axis, loop domains, and sub-compartments are the components of Hi-C contact matrices.

Two basic modes of enhancer-promoter interactions from TAD structures – classical and hubs. In classical communication, CTCF-cohesin loop brings an enhancer closer to a promoter for interacting outside of the loop adjacent to the TAD boundary. Another mode is hub formation within the TAD boundary where both promoters and enhancers communicate inside the loop domain [6]. Besides, strong TAD boundaries allow more intra-TAD interactions while weak boundaries allow more inter-TAD interactions. It has been found that strong insulation of TADs is correlated with high CTCF binding. In addition, a strong TAD boundary insulates super-enhancers in both upstream and downstream directions. In cancer, both strong boundaries and super-enhancers tend to be co-duplicated but strong TAD boundaries have low deletion frequency, probably to protect essential components needed for cancer growth [7].

In addition to bringing an enhancer and a promoter in a close proximity by loop extrusion and CTCF anchor, histone marks are vital for characterizing chromatin states and affecting chromatin accessibility for transcriptional activity. Trimethylation of H3K27 (H3K27me3) is a repression marker, and a region with the presence of H3K27me3 is marked as repressed chromatin region and has less or no transcriptional activities. Acetylated H3K27 (H3K27ac) is a marker of active enhancer, and remarkably high level of H3K27ac marker indicates super-enhancer (Table 1). The presence of enhancers at a gene region activates gene transcription, and super-enhancers increase more transcriptional products than normal enhancers within CTCF loops [8,9,10]. Therefore, alterations in histone modification or structural components such as TADs and loop domains could destabilize chromatin structures, making them susceptible to aberrant gene expression. Oncogenic signaling pathway are connected to a remarkably high density of histone mark modification which signified the presence of super-enhancers at oncogenes, leading to an abnormally high gene expression in cancer [11].

Regulatory element	Activation marks						
	DHS	Methyl	H3K4me1	H3K4me3	H3K27ac	Med	CTCF
Promoter	+	Low	–	+	++	+	+/–
Enhancer	+	Low	+	–	++	+	+/–
Super-enhancer	+	Low	+	–	+++	+++	+/–
Insulator	+	Low	–	+/–	–	–	+

Nature Reviews | Cancer

Table 1. Activating chromatin marks and associated regulatory elements adapted from Sur et al. [7]

The correlation diagram shows the summary of the relationship between different components in genome structure and epigenetics (Fig. 6). As mentioned earlier, CTCF anchors are found in cohesin-associated loops to support intra-TAD interactions while insulating loop domains from interacting with neighboring TADs. Also, the presence of both CTCF loop and cohesin ring shield compartmental domains for intra-compartmental interactions. TAD formation brings enhancers and promoters closer communication, with activating histone marks H3K27ac to promote gene expression and repressing marks H3K27me3 to reduce transcriptional activity. CTCF presence is lower at cohesin-independent loops which allow for more inter-chromosomal interactions and super-enhancer colocalization, thus down-regulating the expression of nearby genes.

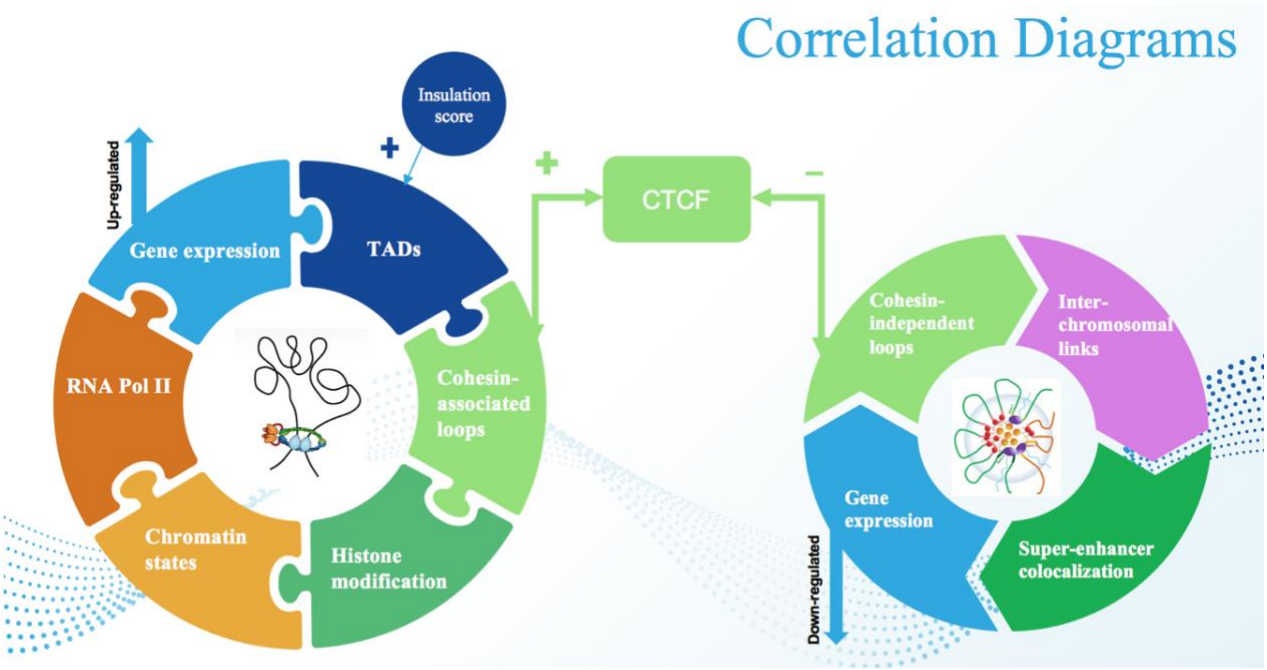

Fig. 6. Correlation diagram shows the relationship between genomic structure and epigenetic events in regulating gene expression

Chromosome conformation can be captured by different techniques. In this project, high-throughput chromosome conformation capture Hi-C technique will be utilized to probe all genomic interactions, both intra- and inter-chromatin contacts in an unbiased "all versus all" approach. This technique starts with formaldehyde crosslinking of DNA, digestion with restriction enzyme, biotinylation for proximity ligation, and library amplification for deep sequencing. A variety of tools will be implemented to obtain and visualize features of 3D chromatin architecture. FAN-C is a comprehensive tool which converts raw sequencing reads from Hi-C experiment to normalized matrices for correcting biases (technical and biological variations) and performs quantitative feature analysis including AB compartments, insulation scores and boundaries, directionality index, and loop domains for TADs [12]. Coolbox generates Hi-C contact matrices at different binning resolutions and with different shapes [13]. JuicerTools analyzes features of TADs such as contact domains and loop domains [5].

To identify cancer from normal cells in a mixed population such as a diagnostic test, a tissue or blood biopsy, or simply a blood test, some signals emitted by cancer cells are potentially important sources to capture cancer happening early in time. Cancer cells acquire super-enhancers which are signified by high density of histone activating markers to cause abnormal changes in oncogenes and thus influencing oncogenic signaling pathway [14]. This is one of important oncogenesis signals for detecting the presence of tumors in body environment. The main goal of this project is to detect cancer signals early for effective treatments and increased survival rate using chromatin architecture and interactions. The first goal of this project is to build a pipeline for understanding higher-order chromatin architecture through Hi-C contact matrices and quantitative features of TADs such as insulation scores, boundaries, and directionality index. Additional features including contact domains, loop domains, cohesin component RAD21, CTCF anchor, and histone marks are collected to support a comprehensive view of chromatin interactions. The additional features plus quantitative features from the first goal provide extensive data sources for the second goal of this project which is to build a machine learning model for classifying normal from cancerous human cell lines. The performance of this model will be evaluated as guided by a conceptual question: How do higher-order genomic structures and epigenetic regulation influence the chromatin interactions in normal and cancerous cell lines? This multi-class classifier with comprehensive features potentially improves prognosis and early detecting cancer for appropriate treatments.

II. Methods

1. Cell lines and *in situ* Hi-C experiment to obtain chromatin contacts

The four cell lines in this project were normal lung fibroblast IMR90, lung adenocarcinoma A549, lymphoblastoid GM12878, and chronic myeloid leukemia K562. Raw FASTQ data of IMR90, GM12878 and K562 were obtained from Rao et al. (2014) [5], and FASTQ data of A549 was from D'Ippolito A. et al. (2018) [15]. Both experiment used *in situ* Hi-C technique to capture chromatin structure. *In situ* Hi-C experiment started with cross-linking DNA and cutting it with a restriction

enzyme. To create DNA-DNA proximity ligation, fragmented loci were ligated with biotin to create chimeric junctions between adjacent segments. Biotin was then purified, and ligated sequences were undergoing paired end sequencing using high-throughput sequencing from Illumina (Fig. 7). This technique generated both intra- and inter-chromosomal contacts to detect loop domains across the entire genome for 3D chromosome conformation [5]. Raw data were obtained from SRA Run Selector and can be found in Table 2.

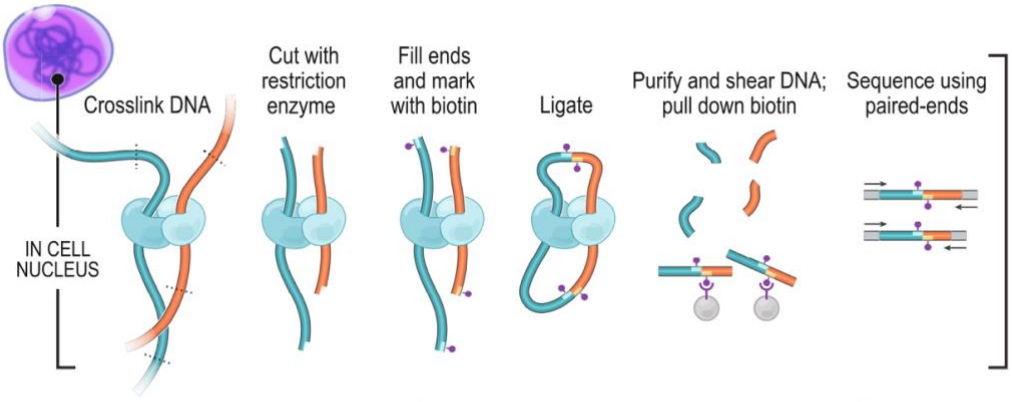

Fig. 7. In situ Hi-C experiment for mapping chromatin contacts in different cell line, adapted from Rao et al. [5]

Table 2: SRA accession IDs for experimental samples

Total: 34 samples

A549	GM12878	IMR90	K562
SRR5129660	SRR1658577	SRR1658672	SRR1658693
SRR5129661	SRR1658580	SRR1658673	SRR1658694
SRR5129662	SRR1658581	SRR1658674	SRR1658695
SRR5129663	SRR1658583	SRR1658675	SRR1658696
SRR5129664	SRR1658586	SRR1658676	SRR1658697
SRR5129665	SRR1658597	SRR1658677	SRR1658698
SRR5129666	SRR1658599	SRR1658678	SRR1658699
SRR5129667	SRR1658647	SRR1658679	SRR1658700
X	X	X	SRR1658701
X	X	X	SRR1658702

2. Hi-C data preprocessing

2.1 *From raw FASTQ data to Hi-C contact matrices*

Hi-C data was preprocessed using FAN-C version 0.9.17, visualized using Coolbox and analyzed using JuicerTools version 1.22.01 (Fig. 8) [5, 12, 13]. Raw FASTQ datasets were retrieved from SRA website and downloaded as fastq.gzip files. Paired-end reads were split at ligation junctions by HindIII or MboI restriction enzyme and then mapped independently to human reference genome hg19 using Bowtie2 to generate mapped reads in SAM format [16]. Unmapped reads were further mapped by iterative mapping function in FAN-C. Aligned reads were then filtered to keep only uniquely aligned reads with a mapping quality of 30 and with the step size of 10 to extend number of base pairs at each round. SAM files were then sorted for the next pairing step such that reads were paired based on read names and assigned to restriction fragments of hg19 reference genome. Unmappable and multimapping read pairs, and PCR duplicates were filtered. Distance filters was applied for filtering self-ligated fragments < 25 kb, and reads mapping more than 1000 bp from a restriction site were removed as well. For strand filters, inward (un-ligated) and outward (self-ligated) read pairs separated less than 1kb were filter because they mostly came from the same fragment and were invalid. Valid read pairs were then binned at different resolutions 100kb and 10kb and normalized using Knight Ruiz (KR) [5]. The entire process from mapping and filtering to binning and normalization was performed using individual *fanc* commands run with bash scripts on HPC SLURM cluster with allocated memory, ntasks, ntasks-per-node, and cpus-per-task.

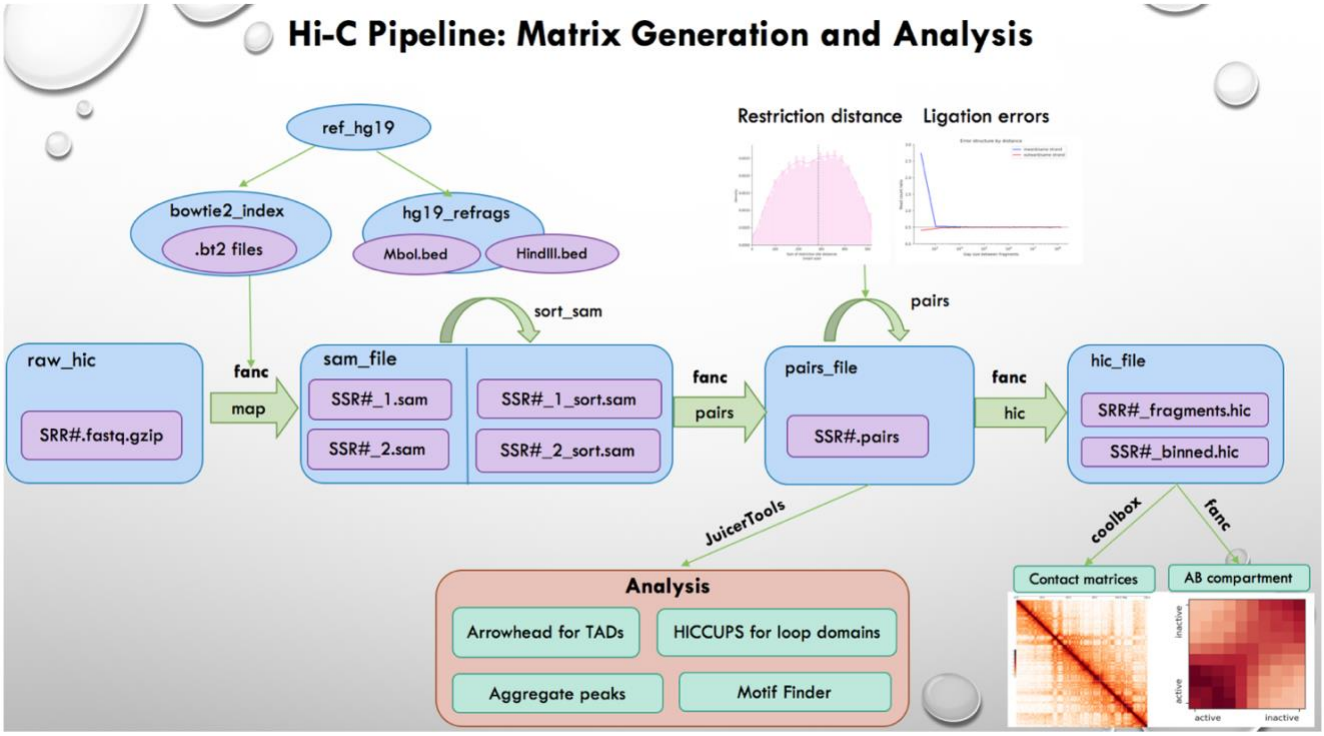

Fig. 8. Summary of steps to process raw data into Hi-C contact maps in Hi-C pipeline

2.2 *Processing statistics for filtering and valid pairs*

Table 3 summarized the processing statistics and showed the percent of valid read pairs generated from mapping and filtering processes in Hi-C processing pipeline. Types of read pairs including inward, outward, self-ligation, PCR duplicates, restriction site distances, multimapping, and ummapable reads were filtered to valid read pairs. The average percent of valid pairs was around 85%.

Table 3: Processing statistics for filtered and valid pairs from total raw reads

Samples	Total reads	Valid pairs	Inward pairs	Outward pairs	Self-ligations	PCR duplicates	Restriction site distance	% Valid pairs	Sum pairs filtered
GM12878_577	87,569,310	75,182,513	9,328,045	2,267,893	4,013,079	910,459	32	85.85	16,519,508
GM12878_580	63,592,550	54,573,765	6,919,379	1,643,387	3,014,084	544,765	27	85.82	12,121,642
GM12878_581	31,678,266	27,537,842	3,225,015	778,275	1,406,609	160,860	22	86.93	5,570,781
GM12878_583	69,592,394	57,017,845	10,278,652	1,122,055	3,426,703	1,439,424	115	81.93	16,266,949
GM12878_586	59,628,563	49,030,434	8,269,787	1,311,858	3,009,318	1,247,139	58	82.22	13,838,160
IMR90_673	107,407,597	90,914,417	12,169,737	2,452,534	4,836,390	2,165,104	14	84.64	21,623,779
IMR90_675	66,649,091	56,971,226	14,619,254	3,856,624	6,670,308	1,016,736	16	85.48	26,162,938
IMR90_676	133,365,970	116,341,749	12,842,701	2,914,095	4,844,483	1,430,733	21	87.23	22,032,033
IMR90_677	118,965,610	94,694,521	15,024,332	5,083,151	9,134,155	4,807,058	45	79.60	34,048,741

2.3 *Matrix Balancing*

Knight Ruiz normalization was used to overcome non-uniformities in coverage due to experimental noises - technical variations such as DNA crosslinking, the number of restriction sites at a locus, or the accessibility to target sites, as well as biological variations such as true loop domains which have both cohesin and CTCF anchors, ordinary domains which contain no cohesin-CTCF anchor, cohesin-independent loops, or sub-TADs spanning one or both compartments. Table 4 demonstrated the underlying formula for KR and ICE normalization.

Table 4: Normalization for balancing of Hi-C contact matrices

Knight-Ruiz	**ICE**
Product of non-negative matrix and diagonal matrices D1 and D2 to obtain singular value P $P = D_1 A D_2$	Expected contact frequency from biases and relative contact probabilities $E_{ij} = B_i B_j T_{ij}$ $T_{ij} = \Sigma_k \lambda_k \cdot E^k{}_i \cdot E^k{}_j + \text{const}$

3. Hi-C contact maps visualized in different chromosome regions using CoolBox

Contact maps for different chromosome regions were visualized using Coolbox [13]. Chromosome regions of interest were selected for each of the four cell lines to generate high-resolution images. This tool is easy to use and obtain high quality images.

4. FANC revealed quantitative features of TADs and compartments

TAD boundaries, insulation scores, and directionality index were calculated for all chromosomes of each cell line using commands provided in FANC [12]. For making Hi-C contact matrices, triangular-shaped images showing the binning size of 10-kb and 100kb KR-normalized maps were assigned using parameter *–p triangle*. Contact intensity was adjusted by minimum and maximum saturation through *–vmin* and *–vmax* parameters, and color scale was tuned from linear to log for more visibly defined loop domains using *–l* parameter. Enrichment profiles of AB compartments was used for visualizing interactions between A and B compartments. They were generated from the average O/E values between regions separated by percentile bins with increment of 10 using compartment eigenvector oriented by GC content from the reference genome hg19.

5. Feature analysis using JuicerTools - Arrowhead and HICCUPS

Contact domains were identified using Arrowhead algorithm, and peak loci as loop domains were called by HICCUPS algorithm from JuicerTools [5]. Arrowhead transformation is a matrix transformation to annotate domains and is defined as:

$$A_{i,i+d} = \frac{M^*_{i,i-d} - M^*_{i,i+d}}{M^*_{i,i-d} + M^*_{i,i+d}}$$

where M^* denotes the normalized contact matrix, $A_{i,i+d}$ is the measurement of the directionality preference of locus i. $A_{i,i+d}$ is close to zero if both locus $i+d$ and locus $i-d$ are both inside or outside a domain. $A_{i,i+d}$ is positive if locus $i+d$ is outside and locus i-d is inside a domain,

and vice versa. Arrowhead was implemented using dynamic programming for efficient calculation. HICCUPS was used for annotating peaks by identifying enriched pixels $M_{i,j}^{*}$ in which contact frequency is higher than expected when number of contacts in a pixel was compared to number of contacts in the surrounding area [5].

6. CTCF, RAD21, and histone tracks using ChIP-Seq data from ENCODE

Contact domains called by Arrowhead were then determined if they are CTCF loop domains or ordinary domains using CTCF, cohesin RAD21 and histone marker information. Processed datasets as ChIP-Seq data of these components were retrieved from ENCODE and accession IDs were provided in Table 5. Histone data were selected based on appropriate false discovery rate FDR less than 1% or 5%, depending on the size of processed datasets.

Table 5: ENCODE processed data for histone markers, CTCF anchor, and RAD21 cohesin

	A549	GM12878	IMR90	K562
CTCF	ENCFF335GSE	ENCFF833FTF	ENCFF453XKM	ENCFF002CEL
RAD21	ENCFF958VNQ	ENCFF002CPK	ENCFF195CYT	ENCFF002CXU
H3K27ac	ENCFF282VMF	ENCFF411MHX	ENCFF899APS	ENCFF931VAQ
H3K27me3	ENCFF046XDC	ENCFF523KGZ	ENCFF741WIY	ENCFF908KJV

7. Data engineering and organization

7.1 *Feature engineering*

Data engineering was performed to combine all features obtained from the analysis in the first pipeline and from ENCODE project site based on scientific background introduced in the Introduction section of this report. The purpose is to generate comprehensive features of genomic structures and epigenetic regulations to represent chromatin interactions which potentially distinguish cell lines through a machine learning model. Different datasets had been converted to .csv files and then read using pandas library. Chromosome locations from ChIP-Seq experiment were trimmed based on FDR less then 1%, and chromosome locations from all datasets were sorted prior to processing. Criteria for engineering feature columns were summarized in the flowchart (Fig. 9). Orange rectangular boxes

represented features of genomic structures and histone marks; orange hexagonal boxes indicated structure effects from CTCF and cohesin ring statuses at TAD boundaries or contact domains. Blue diamond-shaped boxes exhibited the representative datasets of genomic architecture and histone modification. Green boxes referred to TADs and loop domains which are the main contributors in 3D chromatin structures.

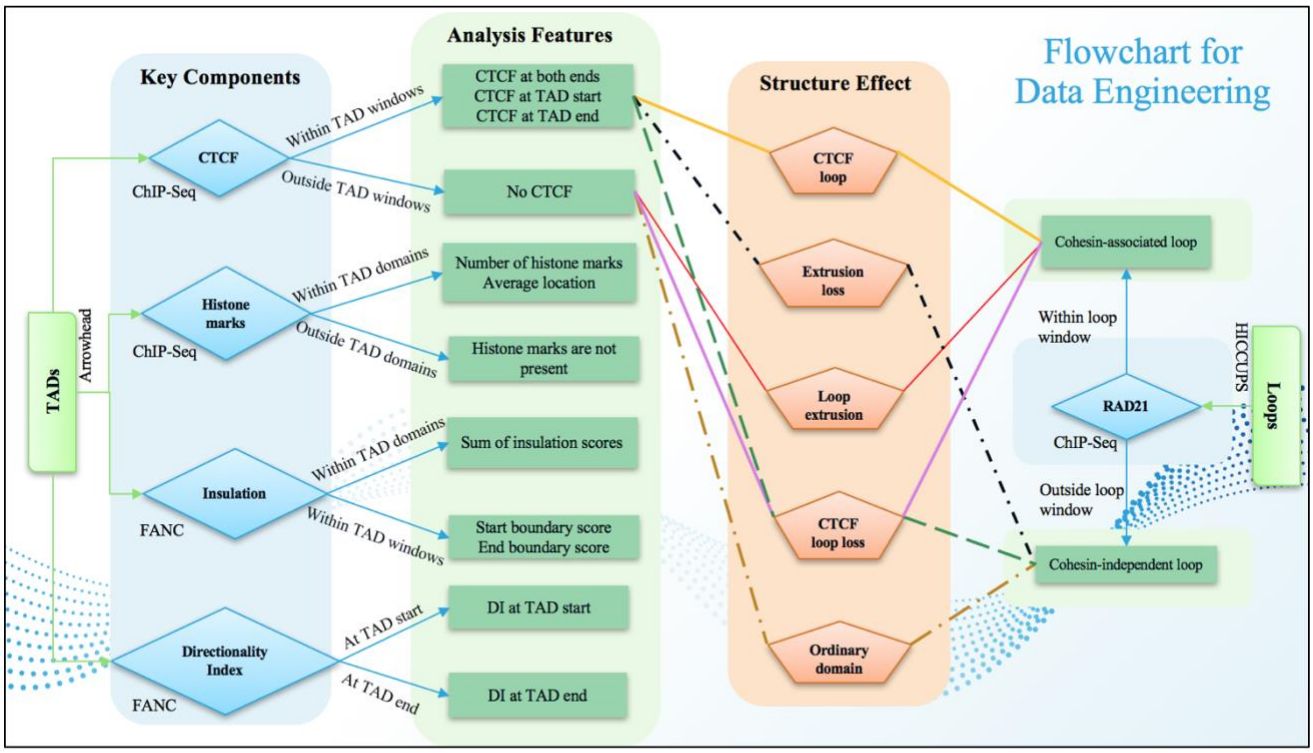

Fig. 9. Flow chart of feature engineering and transformation in building a complete tabular dataset for detecting different cell lines.

7.2 *CTCF presence at TAD boundaries*

The first feature combination was to identify if CTCF anchor presented at TAD boundaries using TAD locations from Arrowhead and CTCF ChIP-Seq data from ENCODE. Both datasets consisted of columns - chromosomes, start and end locations on chromosome. However, TAD dataset also contained variance of upper and lower triangles and the sign of the entries in those triangles whereas CTCF dataset included p-values and false discovery rate (FDR) as q-values to reduce false positives from multiple testing problem. Then, CTCF data was queried with TAD data in which CTCF start and end locations positioning within TAD start and end windows, respectively, would be

classified as ‘both boundaries’, ‘start boundary’, and ‘end boundary’. Besides, mean distance was computed from the distance between CTCF start and TAD start locations, and from the distance between CTCF end and TAD end locations.

7.3 *Quantitative features for TAD boundaries*

The second feature combination aimed to incorporate quantitative features including insulation scores, insulating boundaries, and directionality index corresponding to TAD boundaries from Arrowhead and to assign histone modification status to those boundaries. The intention was to support for the identification of the strength and histone activities around the boundaries. When loop extrusion happens, cohesin ring loops through gene sequence, regulatory elements (enhancers, promoters, repressors), and histone marks, leading to extruding a segment of gene. These factors should have been in close proximity with each other after extrusion event. Scores of insulating boundaries were assigned to each of the start and end TAD locations if they matched criteria of a TAD window. Several insulation scores within each TAD start and end positions were summed to generate the sum of insulation scores with respect to all TAD boundaries. Directionality index was also captured corresponding to the exact TAD start and end locations, creating features of start and end directionality index. For histone mark datasets, activator H3K27ac and repressor H3K27me3 were incorporated along with TAD boundaries, within which the number (count) of histone mark occurrences and distance mean values of those marks were determined.

7.4 *Loop domains and RAD21 cohesin protein*

The third feature combination incoroporated whether a loop domain from HICCUPS was cohesin-associated or cohesin-independent using the cohesin RAD21 ChIP-Seq data from ENCODE. RAD21 data was queried with the loop domain data, and the RAD21 positions corresponding to loop start and end locations were identified. RAD21 cohesin presented within a loop start or end window would be categorized as cohesin-associated loop (‘CA_loop’) whereas others outside the loop window would be assigned as cohesin-independent loop (‘CI_loop’). For cohesin-associated loop, mean

distance RAD21 located from its respective start or end loop position was calculated and assigned to RAD21 from loop location as a feature.

7.5 *Structure effect of CTCF and cohesin*

Lastly, processed loop features were incorporated with processed TAD features to create a complete data source for building a classification model. Loop start locations were checked for their presence within a TAD window based on four conditions: (1) both start and end locations of loops were within an examined TAD window, (2) either loop start or loop end position was within the window, (3) either start or end position was in between TAD windows, and (4) switching occurred such that loop start location was within TAD end window or loop end location was within TAD start window (Table 6). Once loop start locations satisfied a TAD checking criteria, additional features were investigated. For those loop-start locations that satisfied condition 1, the feature structure effect exhibited the effect of CTCF anchor and cohesin ring onto 3D chromatin organization. As mentioned in the Introduction section of this project report, CTCF loop domains were resulted from the presence of both CTCF and cohesin ring at TAD boundaries spanning both active and inactive compartmental domains, and this was a signature of 'CTCF-loop' sub-category. Next, the presence of CTCF and the absence of cohesin at a contact domain caused the loss of CTCF loop and extrusion while segregating those compartmental domains, leading to 'extrusion-loss'. Another case involved CTCF absent and cohesin ring presented near TAD boundaries, causing CTCF loop domain loss but continuing loop extrusion, and the structure effect was assigned as 'loop-extrusion'. If both CTCF and cohesin ring were absent at a contact domain, it was more likely to be 'ordinary-domain' as the structure effect. For loop locations meeting condition 2, 'CTCF-loop-loss' was a sub-category of the structure effect if either CTCF or cohesin ring occurred at a contact domain. If a loop was upstream of the TAD end location, there was no loop found at that current domain causing CTCF loop loss. 'Latent' chromatin state represented cases in which both activators and repressors presented at high density in a contact domain.

Table 6: Structure effects of CTCF anchor and cohesin ring on loop and contact domains

Condition	CTCF status	Loop type	Structure effect
Both loop start and end locations were presented within a TAD domains	CTCF at both TAD ends CTCF at TAD start CTCF at TAD end	CA loop	CTCF loop domain
		CI loop	Extrusion loss
	No CTCF	CA loop	Loop extrusion
		CI loop	Ordinary domain
Either loop start or end location occurred within a TAD domain	CTCF at both TAD ends CTCF at TAD start CTCF at TAD end	CA loop	CTCF loop domain
		CI loop	CTCF loop loss
	No CTCF	CA loop	CTCF loop loss
		CI loop	Ordinary domain
Either loop start or end location occurred outside a TAD or contact domain	CTCF at both TAD ends CTCF at TAD start CTCF at TAD end No CTCF	CA loop or CI loop	Loops were outside TAD domain
Switching occurred – loop start location was in the TAD start window, or loop end location was in the TAD end window	CTCF at both TAD ends CTCF at TAD start CTCF at TAD end No CTCF	CA loop or CI loop	Inversed loop
Loop start location was upstream of TAD end location	CTCF at both TAD ends CTCF at TAD start CTCF at TAD end No CTCF	CA loop or CI loop	CTCF loop loss

7.6 *Boundary strength stratification*

As mentioned in the Introduction, insulation scores have been proved to be a good feature to stratify TAD boundaries [7]. Four different boundaries in which Boundary I as the weakest and Boundary IV as the strongest were assigned to each TAD or contact domain. They were classified based on the quantile ranking of summed insulation score (Table 7).

Table 7: Stratification of TAD boundary strengths

Boundary strength	Insulation score (percent quantile)
Boundary I	Lower than 25% quantile
Boundary II	Between 25% and 50% quantile
Boundary III	Between 50% and 75% quantile
Boundary IV	Higher than 75% quantile

7.7 Chromatin states categorization

Chromatin interactions tend to occur within a TAD domain. Loop extrusion of cohesin ring and CTCF anchor bring enhancers closer to promoters for communication. Histone activating marker H3K27ac and repressing marker H3K27me3 regulate gene expression. Therefore, chromatin states should not be limited with histone modification; it is a combination of spatial genome organization and epigenetic regulators. Table 8 described multiple features involved in classifying chromatin states in a concise but comprehensive aspect.

Table 8: Chromatin states dynamics were classified based on histone marks and chromatin structure

H3K27ac (percent quantile)	H3K27me3 (percent quantile)	CTCF anchor	Loop category	Chromatin states
[0, 25]	[0, 50]	CTCF at both TAD ends CTCF at TAD start CTCF at TAD end	CA loop CI loop Loop loss	Insulator
[0, 25]	[0, 50]	CTCF at both TAD ends CTCF at TAD start CTCF at TAD end No CTCF	CA loop CI loop	Weak enhancer
[75, 100]	[0, 50]	CTCF at both TAD ends CTCF at TAD start CTCF at TAD end	CA loop	Positive SE
[0, 50]	[75, 100]	CTCF at both TAD ends CTCF at TAD start CTCF at TAD end No CTCF	CI loop Loop loss	Negative SE
[50, 75)	[0, 50]	CTCF at both TAD ends CTCF at TAD start CTCF at TAD end No CTCF	CA loop CI loop Loop loss	Strong enhancer
[0, 50]	[50, 100]	CTCF at both TAD ends CTCF at TAD start CTCF at TAD end No CTCF	CA loop CI loop Loop loss	Repressor
[50, 100]	[0, 50]	CTCF at both TAD ends CTCF at TAD start CTCF at TAD end No CTCF	CA loop CI loop Loop loss	Active promoter
Remaining cases	Remaining cases	Remaining cases	Remaining cases	Latent

The entire chromosomes in each cell line underwent the above combinatorial processes, and the processing happened for each chromosome per cell line at a time. The product of this data engineering work was an organized, processed dataset which showed the relationship between the features of genomic structures and epigenetic regulations in chromatin interactions and was ready to be used as feature columns for building a multi-class classifier.

8. Building machine learning detection model

8.1 *Data splitting and transformation*

Processed data from feature engineering task were divided into training and testing sets for Random Forest and XGBoost models and additional validating set for TabNet classifier. Data transformation using KNN imputer was applied for datasets used in Random Forest and TabNet while datasets for XGBoost remained intact as one of model requirements.

8.2 *Training Classifiers*

A. Random Forest Classifier

Originally introduced in 1995, Random Forest has been one of the best machine learning models for classification tasks [17]. Decision Trees are sensitive to dataset and prone to overfitting, and Random Forest is an ensemble model consisting of numerous decision trees. By having many independent decision trees where bootstrap aggregating allows trees to subsample different subsets of the dataset, Random Forest are more resistant to overfitting. Random forest works based on random sampling of data points and splitting nodes based on slightly different subsets of features. It is computationally efficient and can handle large datasets.

In this project, Random Forest is implemented using Sci-kit Learn to classify cell lines. The number of decision trees in the forest, the minimum number of samples at a leaf node, and the maximum depth of the trees were the key parameters affecting model accuracy after several trials (Fig. 10). As a result, leaving out maximum depth gave the highest accuracy since it is best to allow decisions trees to grow unlimitedly for this dataset. For the number of estimators, tuning results

showed that model accuracy was using 150 or more decision trees. For the Random Forest model, final parameter values include 200 estimators, “gini” criterion, random state of 42, and each tree only sampled 85% samples of the dataset.

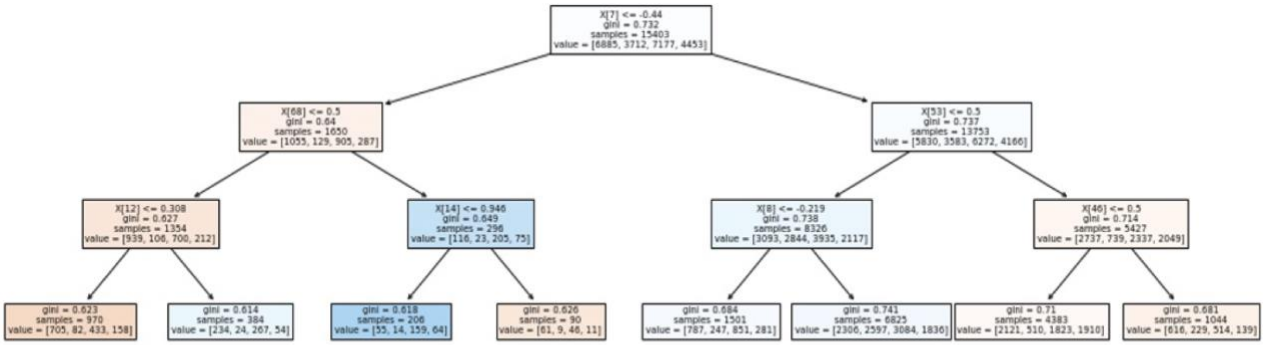

Fig. 10. Sample decision tree from Random Forest with the maximum depth of 3

B. XGBoost Classifier

XGBoost is a novel tree boosting system introduced in 2014 and published in 2016 by Tianqi Chen and Carlos Guestrin with support from the community [18]. XGBoost is scalable, highly efficient when compared to similar machine learning models such as Random Forest and AdaBoost. One notable feature of XGBoost is its awareness for sparsity, and this is particularly helpful in this project. The processed dataset of chromatin interactions undergoes extensive feature engineering, and the dataset has sparse columns resulting from one-hot-encoding and missing values from merging processed TAD and loop datasets. XGBoost can handle sparse columns and missing values natively, which is absent in the Sci-kit Learn implementation of Random Forest. Missing values remain unchanged for running XGBoost to better represent the TAD and loop relationship while in other machine learning algorithms, missing values needs imputation.

Hyperparameter tuning of the final XGBoost classifier suggests the best settings such as 8 for maximum depth of the tree, 0.7 for learning rate, 300 tree estimators, 80% subsample ratio (Table 9). The highest accuracy achieved is 81.13%. Different from Random Forest model, XGBoost benefits from a larger number of trees and limiting the layers of individual trees (Fig. 11). Other regularization

parameters were examined including sub-sampling columns when constructing each tree and regularization lambda, but did not help with accuracy, which indicates that the model is not overfitting.

Table 9: Hyperparameter tuning for XGBoost classifier

Maximum depth	Learning rate	Number of estimators	Subsample	Accuracy (%)
8	0.4	180	0.5	79.70
8	0.1	100	0.7	76.52
8	0.2	150	0.7	79.93
8	0.3	150	0.7	80.33
8	0.5	150	0.7	80.58
8	0.6	150	0.7	80.18
8	0.7	300	0.8	81.13

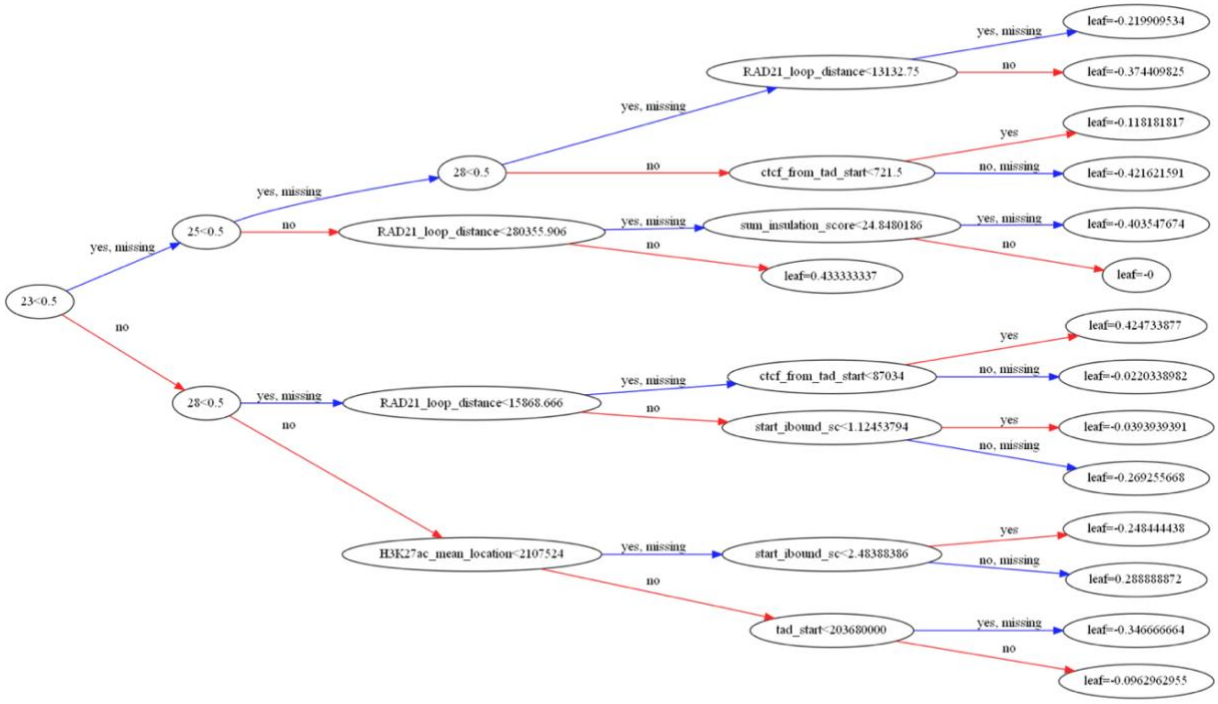

Fig. 11. Sample decision tree from XGBoost classifier with the tree maximum depth of 4

C. TabNet Classifier

Dataset engineered for this project is tabular in nature. Conventionally, classification tasks with tabular data are best handled by machine learning models such as Random Forest. Deep learning models can hardly achieve similar performance despite their excellent performance with image datasets. That was true until Google researchers introduced TabNet in 2019. TabNet is a complex neural network architecture. Users can control the "deepness" of the network by adjusting the number of steps, where each can be viewed as a block of feature transformer, attentive transformer and mask that contain fully connected layers and normalization layers (Fig. 12). TabNet model is capable of selecting different features important for each transformation, mask, and even row in the dataset [19].

As with other neural networks, training and optimization are time consuming and not a straightforward process. After testing different hyperparameters, the default parameter values recommended in the original publication generally produced good results. The final TabNet model parameters include the width of the prediction layer and the width of the attention embedding both at 64, number of steps at 5, gamma coefficient for feature usage at 1.5, number of independent Gated Linear Units at 2, number of shared Gated Linear Units at 2, and the highest validation accuracy is 73.5%.

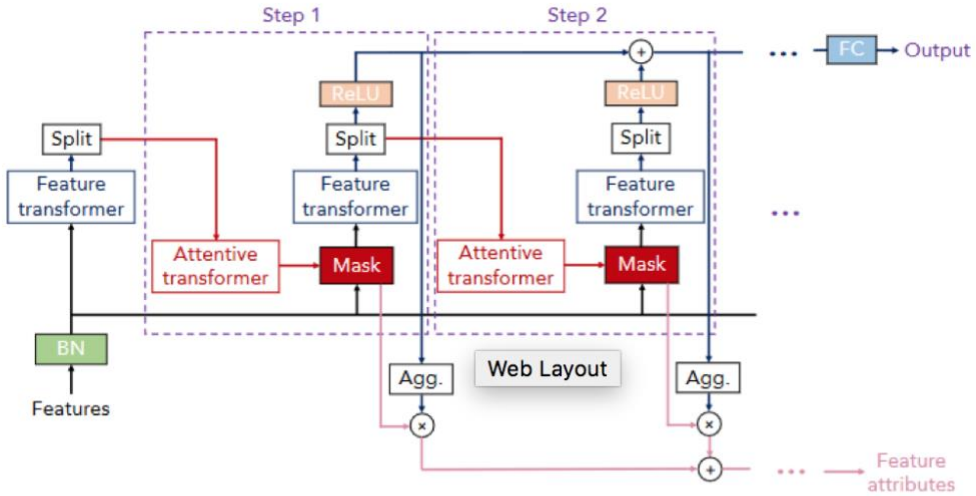

Fig. 12. TabNet classifier model operation adapted from Arik et al. [18]

III. Results

1. Higher resolution images reveal loop domains of chromatin contacts

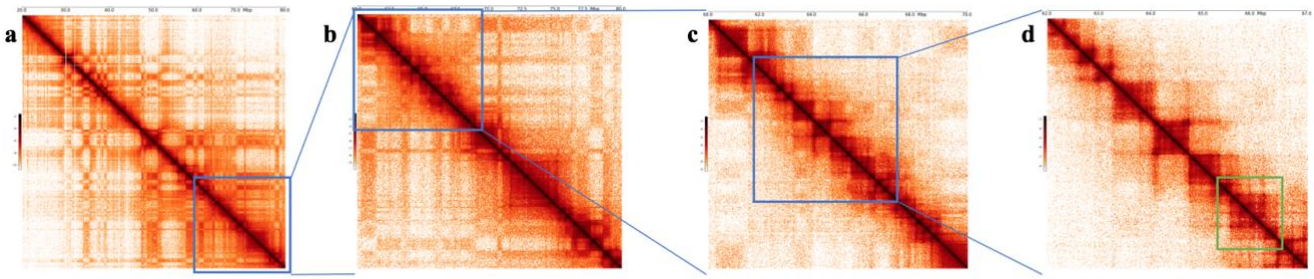

Fig. 13. Chromatin contacts were explored in contact matrices from low to high resolutions. (a) low resolution 60Mb as 20-80Mb. (b) & (c) moderate resolution 20Mb and 10Mb. (d) high resolution 5Mb as 62-67Mb. Blue squares for zooming in and green square for focusing

Contact matrices from chromosome 1 of A549 are exhibited with various resolutions to reveal intricate structure of loop and ordinary domains (Fig. 13). Starting from a lower resolution, chromosome region 20-80Mb (60Mb in distance), small squares were clumped together and made it hard to view the internal structure of TADs, but sub-compartmental domains were visible on either side of the diagonal axis such as the top-right and the bottom-left triangles (Fig. 13a). Projected to the higher resolutions (blue squares), chromosome regions of 60-80Mb and 60-70Mb (20Mb and 10 Mb in distance, respectively) showed the number of contact domains in certain pairs of loci as various sized squares along the axis. Peak pixels as dots located away from the diagonal axis represent loops, and contact domains with these dots are loop domains. Domains without dots are ordinary domains. At the latter resolution 10Mb in distance, loops started to be observable on the map, and squares started to be distinguishable. Squares with stronger intensity (darker color) and multi-layer locations suggest higher number of domains congregated at prime locations on a chromosome (Fig. 13b and Fig. 13c). Due to limitations in processing Hi-C data to 5kb or 1kb resolution, the matrix could be binned down to 3-5Mb in distance to maintain a clear, high-resolution image, further zooming in to less than 3Mb in chromosome distance caused blurred images with indistinguishable points. At 5Mb-distance contact

matrix (62-67Mb), higher number of loops become distinctly visible in the smaller chromosome region as indicated in the green squares.

2. Dynamics in chromatin contacts of the four cell lines are captured on Hi-C contact maps

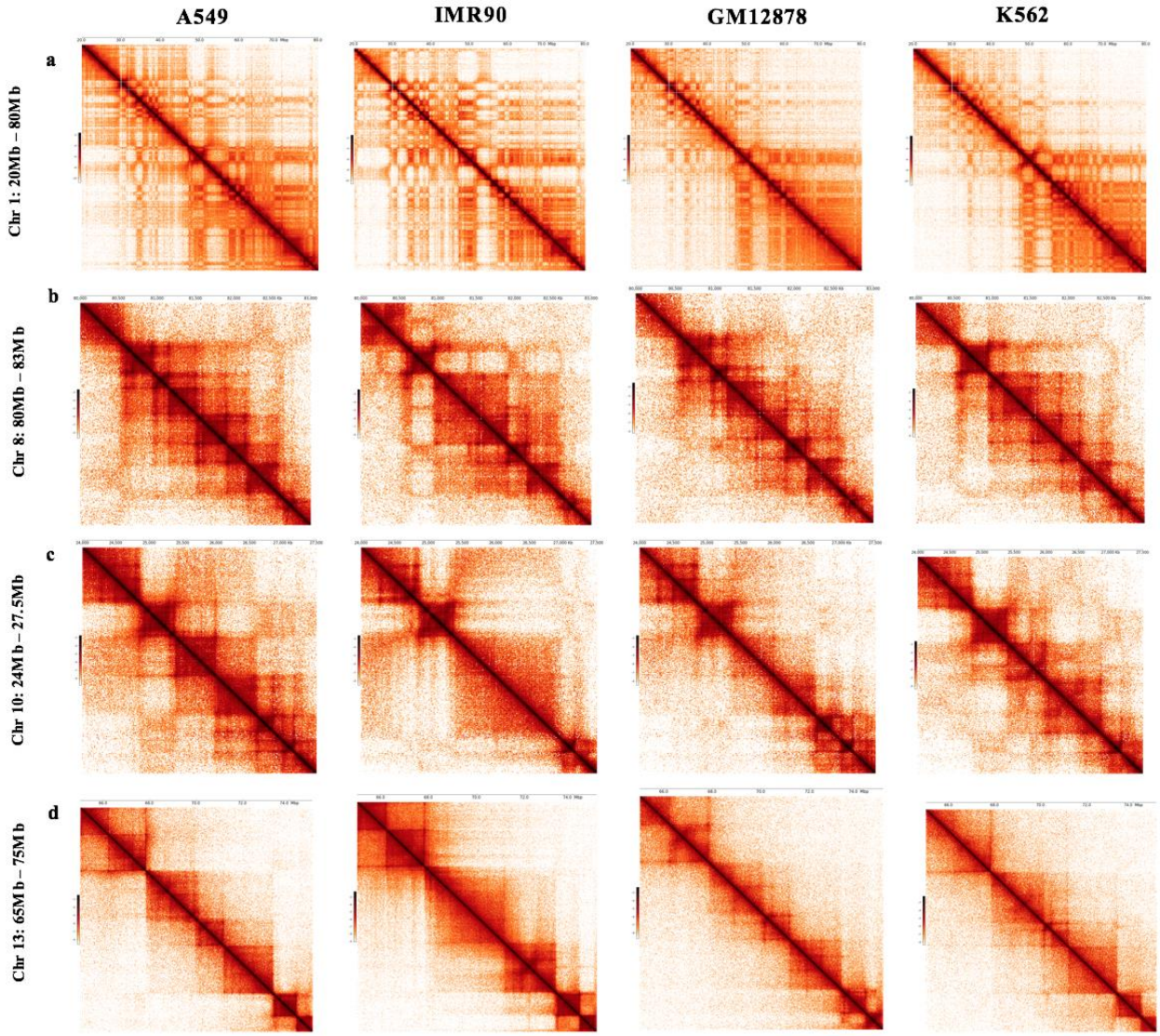

Fig. 14. Chromatin structure dynamics at different genomic regions in the four cell lines. (a) Genomic structure is stable in this region of chromosome 1. (b, c, d) Structural variations were observed at higher resolutions in the specified regions, each having unique architecture for classification

Although TAD boundaries are invariant across different cell types [7], number of chromatin contacts are varied in certain genomic regions to distinguish one cell line from other cell lines. In other words, there are some unique genomic regions that make A549 differentiate from IMR90 or make cancer cells distinguishable from healthy cells (Fig. 14). The chosen region 20-80Mb on chromosome

1 suggested structural similarity of all four cell lines (Fig. 14a). The clarity of sub-compartmental domains in the background were slightly varied potentially due to technical variations or different number of contact domains being called in each cell line. Besides, numbers of chromatin contacts are visibly varied among these cell lines. Normal lung fibroblast cells IMR90 is expected to have genomic regions with different transcription activities from lung adenocarcinoma A549, and this could be observed in certain contact domains in chromosome 8, 10, and 13 (Fig. 14b,c,d). The center of each contact maps of these chromosomes suggested a large difference in the number of chromatin contacts and peak pixels as loops in the chosen chromatin regions between these cell lines, specifically A549 and IMR90. The differences could become more profound if they occur in loop domains as in the region 80-83Mb in chromosome 8 since this would interfere with transcription and gene expression activities. It is more likely to observe ordinary domains in chromosome 10 and 13 as the absence of dots, and thus structural differences in these areas might have lower impacts on genomic activities or subtly alter histone modification status.

3. CTCF and histone marks describe chromatin contact characteristics in A549 and IMR90

CTCF, cohesin ring and histone marks have significant effects on contact domains in which the presence of CTCF and cohesin ring make them loop domains. Ordinary domains are those not bound by CTCF or cohesin but have specific histone modifications. CTCF anchors signify some locations with peak loci as dots present along the diagonal axis. A549 cancer cell line seems to have slightly higher density of CTCF at its contact domains in the 80-83Mb region of chromosome 8 than it is in IMR90 normal cell line (Fig. 15). Both A549 and IMR90 have high repressing markers H3K27me3 and low activating marker H3K27ac at this chromosome region. However, histone profiles for H3K27ac and H3K27me3 are vastly different in both cell lines. For instance, higher level of H3K27ac was expressed at TAD domains in A549 whereas only a few activating marks were present at contact domains in IMR90. The distribution of H3K27me3 was high in the region chr8: 81-81.7Mb in A549 and elevated in the larger range in IMR90 (chr8: 80-81Mb and chr8: 81.2-81.7Mb). These factors

support distinct chromatin profiles of domain structure and histone modification in these two cell lines to make them unique and distinguished for classification and detection.

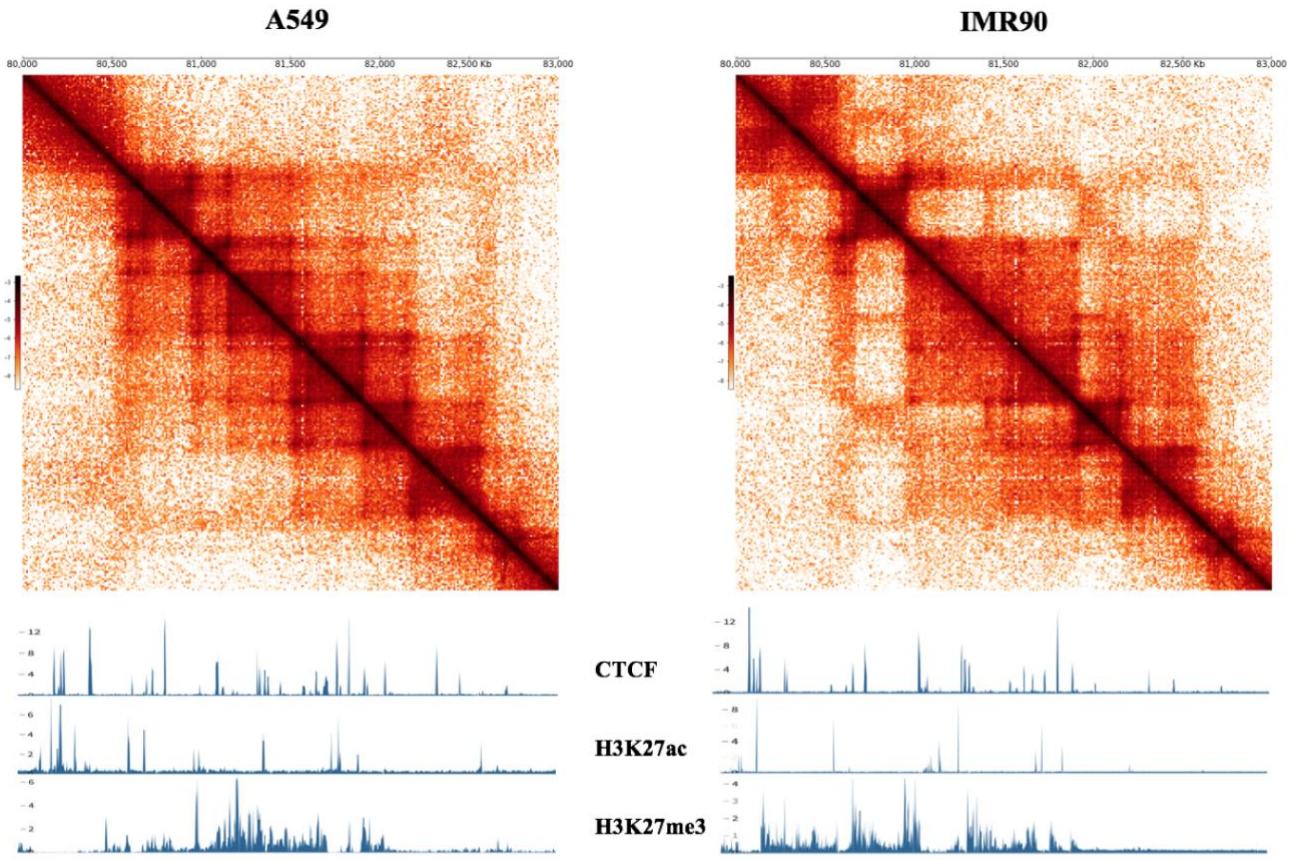

Fig. 15. CTCF and histone marks correspond to chromatin contacts in A549 and IMR90. Left: contact matrix of A549 chr8: 80-83Mb and associated tracks. Right: contact matrix of IMR90 chr8: 80-83Mb and relevant tracks. Contact domains of these cell lines possess a distinctive pattern of genomic architecture and epigenetic features.

4. Triangular Hi-C contact matrices and compressed matrices viewed at different resolutions.

Hi-C contact matrices could be visualized as triangular shape or compressed triangular shape, and they were created with normalized matrix data. The top two plots of 10-kb resolution matrices and the bottom plot of 100-kb resolution matrix exhibited the flexibility of Hi-C data in visualizing chromosome 1 in K562 cell line (Fig. 16). The middle matrix with 1.0-mb region showed zoom-in resolution with saturated, bigger dots dispersed throughout the triangular plot whereas the top plot

revealed an overall picture of multiple loop distribution with various dot intensities in which the small pyramids as loops had higher intensity than regions lacking loops. It was the same trend for the bottom matrix with the lower resolution of 100-kb where individual loops were clearly expressed with bigger squares and more color-defined than 10-kb resolution matrices. This discrepancy indicated FAN-C tool versatile for different matrix binning resolutions. When image saturation for 10-kb data was tuned to higher saturations with –vmax, the colored dots were toned down and blurred, and the lower the saturation (-vmax = 0.05) was tuned, the more defined the image was.

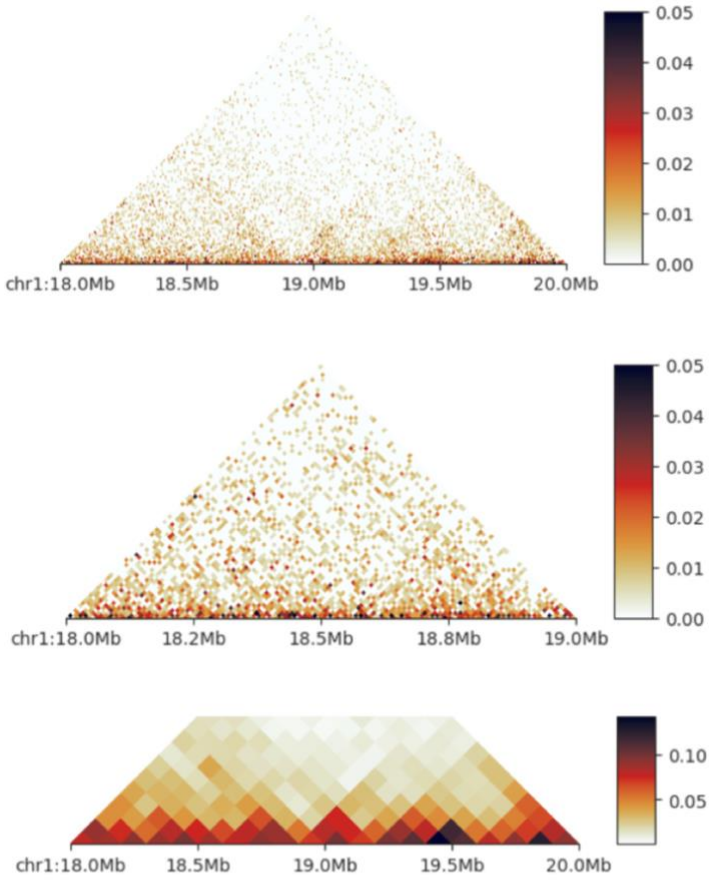

Fig. 16. Triangular contact matrices of different binning resolutions. Top and middle: 2Mb region of 10kb-KR normalized matrix. Bottom: 2Mb region compressed to 1Mb region in 100kb-KR normalized matrix

5. Enrichment profiles revealed interactions between active and inactive compartments.

The aggregate compartment plot known as saddle plot was created from eigenvector (EV) and the average O/E values to show enriched ‘A’ (active and high GC region) and ‘B’ (inactive and low GC content) compartments. The enrichment profile lacked the negative EV percentile cutoffs, thus only the positive cutoffs were shown as lighter to darker red regions on the plot (Fig. 17). The absence of negative EV entries potentially resulted in the lack of ‘B’ compartments or low GC content while

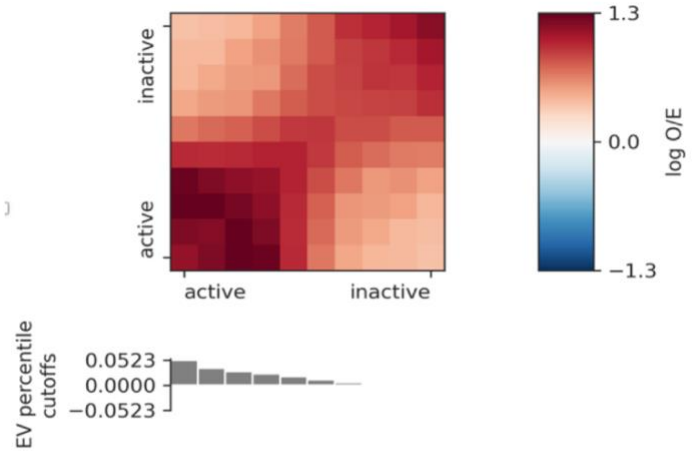

Fig. 17. Enrichment profile of AB compartments from Hi-C contact matrix at 100kb resolution

only 'A' compartments as high GC content were presented on the enrichment profile and shown as positive EV entries. The deficiency of negative EV entries could be due to missing eigenvector in correlation matrix of the normalized 100-kb Hi-C contact matrix.

6. Contact domains along the matrix diagonal were analyzed with insulating scores, boundaries, and directionality index.

Chromatin structure is created from multiple contact domains and expressed in the matrices as pyramids (in triangular heatmap) or squares (in square heatmap) with various sizes along an axis. Although 'peak' pixels could be observed in the square heatmap, triangular map reshapes contact domains so that 'peak' pixels are better annotated on the top of pyramids to differentiate loop domains from ordinary domains. The top image showed the triangular heat map of K562 at 100kb bin size and chromosome region from 163Mb to 173Mb (Fig. 18). Local maxima present contact domain enrichments (both intra-TAD and ordinary domain) in which smaller contact domains or regions between domains have lower insulating scores while larger domains suggest strong domains with high insulating scores. Insulation scores were

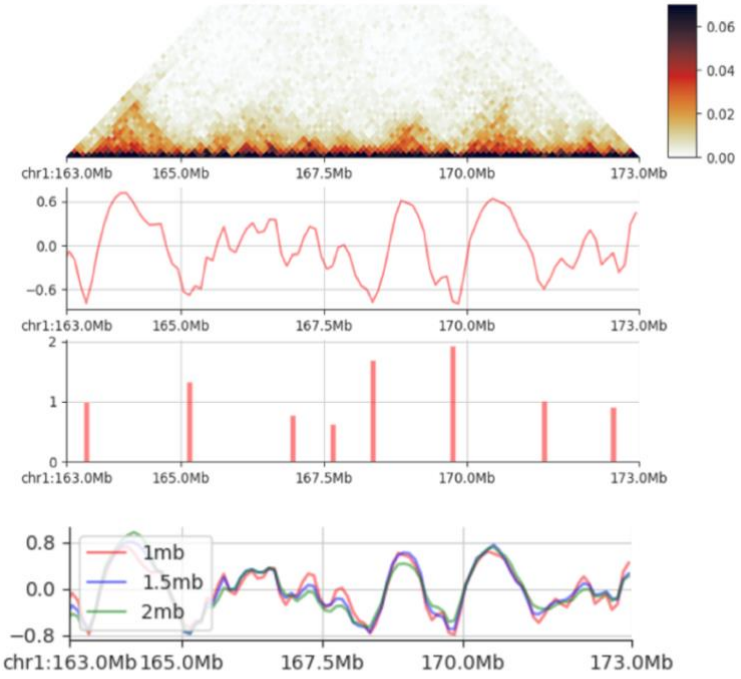

Fig. 18. Triangular contact map and associated quantitative features. Contact map showed contact enrichments with or without loop domains. Line plots represent insulation scores. Bar plot shows corresponding insulating boundaries

presented at a line graph with the window size of 1mb and have multiple local maxima. Local minima in the line plot corresponding to vertical sticks of different heights in the bar plot show regions between TADs as self-interacting domains or boundaries where one contact domain transitions to the next domain. The last line plot shows multiple sliding window sizes of 1mb, 1.5mb and 2mb with very similar trend of local maxima and local minima.

7. Insulation scores and boundaries are direct features to quantify TAD boundary strength

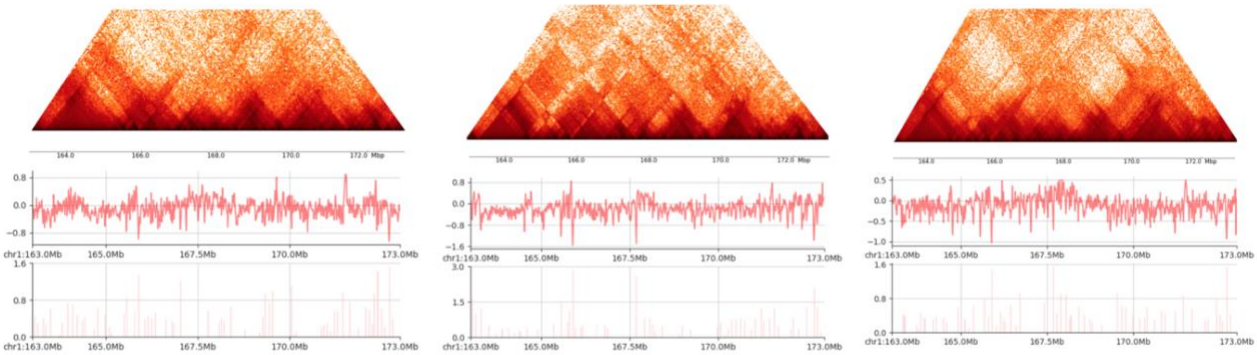

Fig. 19. Contact domains were quantified by insulation scores and boundaries in GM12878, IMR90, and K562. Triangular heat maps showed chromatin contacts as multi-sized pyramids. Quantitative features were plotted corresponding to peaks in the contact domains

The three triangular contact maps present contact enrichments with sub-TADs at 10kb bin size and KR normalization for GM12878, K562, and IMR90 cell lines. Contact depletions also occur at transitioning points between domains as TAD boundaries. The corresponding insulation scores and boundaries showed local maxima and local minima, respectively. Each cell line has trivial differences in TAD quantitative feature profiles due to invariant characteristics of TAD boundaries across cell types. However, cancer causes changes in chromatin structures to some extents depending on the strong or weak TAD boundaries in which the latter is more prone to mutational events, and thus easily disrupting the structure. Figure 19 showed minor differences between the three cell lines, and this indicates the chosen chromosome region does not have much variations among cell lines. There were a few strong boundaries at this chromosome region in each cell line, and these strong boundaries could preserve and protect TAD domains from being changed in cancer events. Since there were no

significant differences between the profile of these cell lines at this chromosome region (163Mb - 173Mb), this region might not be affected and not the best choice for distinguish the cell lines.

8. Uniform distribution of chromatin states across boundary strengths in all chromosomes of the four cell lines

Histone modification and TAD boundary strength are one of the main contributors in chromatin state characterization, and state dynamics are consistent among different cell lines as TAD boundaries are invariant across cell lines. However, different boundary strengths and the level of histone mark modification constitute to maintain or alter chromatin states. Super-enhancer and strong enhancer chromatin states are the representatives of the activating mark H3K27ac, thus explanation in this section will focus on these two states as PSE (super-enhancers with upregulated gene expression) and SE (strong enhancer). In particular, the distributions of positive super-enhancers PSE and the distribution means were varied across boundary strengths I to IV in all four cell lines GM12878, K562, IMR90, and A549 (Fig. 20 top plots). PSE distribution and means were exhibited evenly across all boundary strengths in A549 and higher than the other three cell lines. The only exception is that it was unexpected to see IMR90 (normal lung cell line) had higher PSE distribution and mean value in boundary I than A549 (lung carcinoma line). This might be due to technical variations between experimental datasets – TAD domains from JuicerTools and processed histone ChIP-Seq from ENCODE. In addition, the number of chromatin contacts or TAD domains were varied across cell lines, which contributed more data and wider distribution in IMR90 compared to the other cell lines. GM12878 had lower PSE distribution in all boundary strengths than the other three cell lines. This could be due to the lower number of TAD domains or variations between datasets with different processing techniques. PSE distribution was significantly low in boundary I of GM12878, probably because it had lower number of contact domains and thus less associated features being processed compare to the remaining cell lines, or it could be due to biological variations among boundary strengths of GM12878. Strong enhancers were consistently distributed among cell lines and at lower

density than other chromatin states. Level of negative super-enhancers NSE (super-enhancers with downregulated gene expression) was varied across cell lines and remarkably low due to a few counts of H3K27ac marks around cohesin-independent loops or inter-chromosomal links.

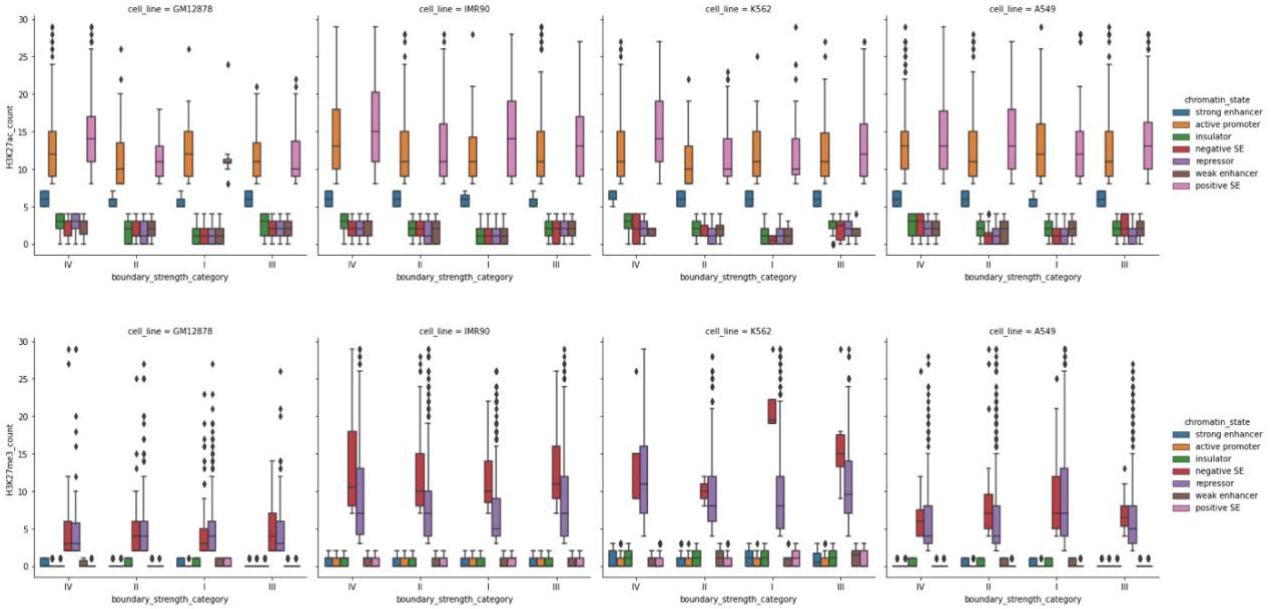

Fig. 20. Chromatin state distribution across TAD boundary strengths. Top plots: state distribution in the presence of H3K27ac activator. Bottom plots: state distribution in the presence of H3K27me3 repressor

Trimethylation of H3K27 (H3K27me3) suggests repressing chromatin marks, and the chromatin state profile with respect to boundary strengths was changed in the opposite direction (Fig. 20 bottom plots). Negative super-enhancer NSE (super-enhancers with downregulated gene expression) and repressor chromatin states are the representatives of the repressor mark H3K27me3. All chromatin states were observed in K562 and IMR90 cell lines whereas only three states of NSE, repressor and latent were predominantly exhibited in GM12878 and A549. The repressor mark H3K27me3 causes NSE and repressor distributions to be high across boundary strengths and varied among the four cell lines. In lung cell line, IMR90 exhibited higher NSE in all boundary strengths but lower repressor distribution in boundary I compared to A549. For the other two cell lines, GM12878 showed lower NSE and repressor distributions than K562. The distributions and means of NSE across boundary strengths in all cell lines were higher than PSE distributions and means. Strong enhancers

and PSE were at a minimal level in all boundary strengths in the context of H3K27me3 because this histone mark represses a chromatin region to downregulate its gene expression.

To sum up, PSE and strong enhancers were present at high density in the context of H3K27ac activator while they were lower in the context of H3K27me3 repressor. The opposite trend is applied to NSE and repressors.

9. Variations in chromatin states of chromosome 1, 13 and 21

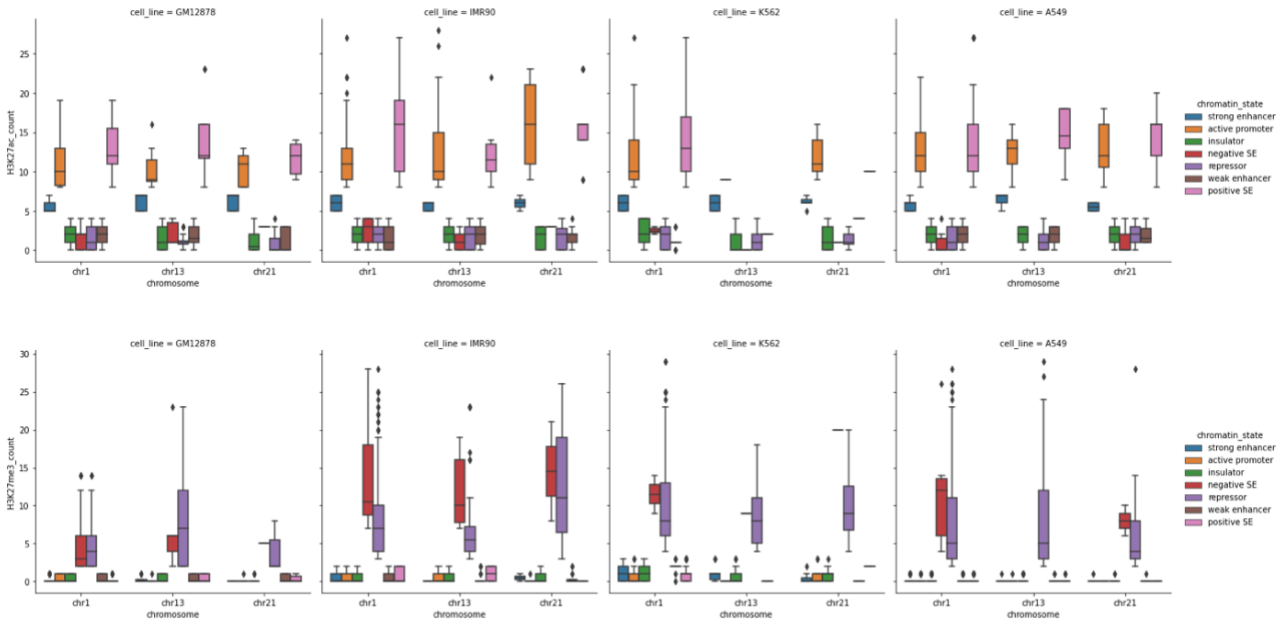

Fig. 21. Chromatin state distribution per chromosomes in the presence of H3K27ac (top plots) and H3K27me3 (bottom plots)

Different chromosomes in each cell line exhibited a unique chromatin profile corresponding to number of activator H3K27ac and repressor H3K27me3 present at specific loci. PSE distributions from the increasing H3K27ac occurrence were higher in chromosome 13 and 21 of A549 whereas IMR90 exhibited higher NSE distributions when H3K27me3 accumulated near certain gene regions in the three chromosomes (Fig. 21). This difference suggests that certain loci in chromosome 1, 13 and 21 potentially expressed more in A549 cancer line than in IMR90 normal line, and the expression level of these chromosomes in IMR90 is halted by downregulation from NSE. In K562 leukemia line, chromosome 13 and 21 had zero PSE distribution but a remarkable presence of repressors, thus

indicating these chromosomes as more likely to be repressed compared to chromosome 1 and other cell lines. All three chromosomes in GM12878 showed PSE distributions but low level of NSE, suggesting that these chromosomes tended to be more active than passive in this cell line. In general, the four cell lines had various level of chromatin states distributed across chromosome 1, 13 and 21. All three chromosomes of IMR90 maintained high NSE state distribution while the remaining cell lines contained a limited number of NSE in the context of trimethylated H3K27. PSE state was mainly absent in K562 but consistently present in chromosome 1, 13, and 21 of the other cell lines in the presence of acetylated H3K27.

10. Histone marks of enhancers accumulate at strong TAD boundary across three chromosomes

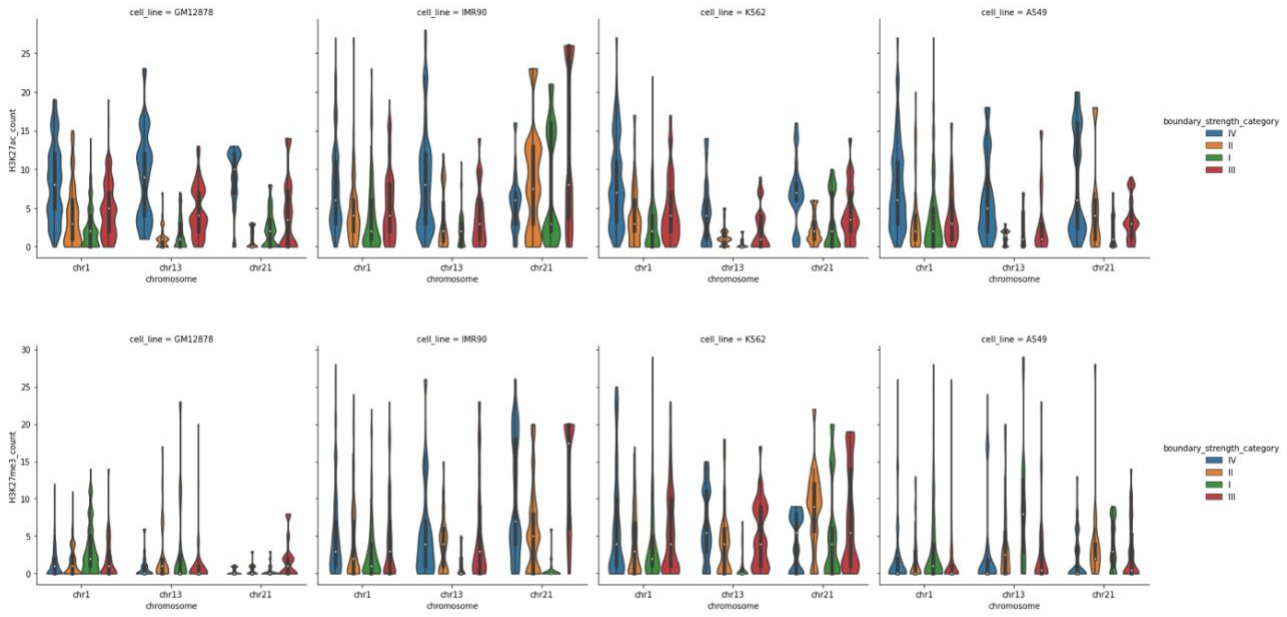

Fig. 22. TAD boundary strengths per chromosomes in the presence of H3K27ac activator and H3K27me3 repressor

Boundary strengths exhibited differently for three chromosomes in the four cell lines (Fig. 22). In this case, boundary IV, the strongest boundary, predominantly existed in chromosome 1, 13, and 21 of A549 when examining boundary strength in the presence of H3K27ac. It became less intense when H3K27me3 occurred at contact domains. TAD boundary strength was less significant in chromosome 13 of K562 with respect to the number of H3K27ac while chromosome 21 in GM12878 had the lowest

intensity corresponding to H3K27me3 count. This phenomenon further supports super-enhancers tend to localize and co-duplicated with strong TAD boundaries for protection from deletion while repressors are more likely to occupy/spread across different boundary strengths.

11. XGBoost Classifier performs better than Random Forest and TabNet for this tabular dataset

For Random Forest model, mean accuracy after 10-fold cross validation was 73.76%. This was achieved with by having 200 decision trees and allowing each tree to sample 85% of the entire dataset. These two parameters were selected after performance optimization, as other parameters are best to remain at default values. Feature importance plot only provided insights to numerical features, as categorical features were converted to many columns after one-hot-encoding (Fig. 23). Normalized confusion matrix showed K562 were best classified at 83% using this model, followed by IMR90 at 79%, A549 at 77%, and GM12878 at 54% (Fig. 24).

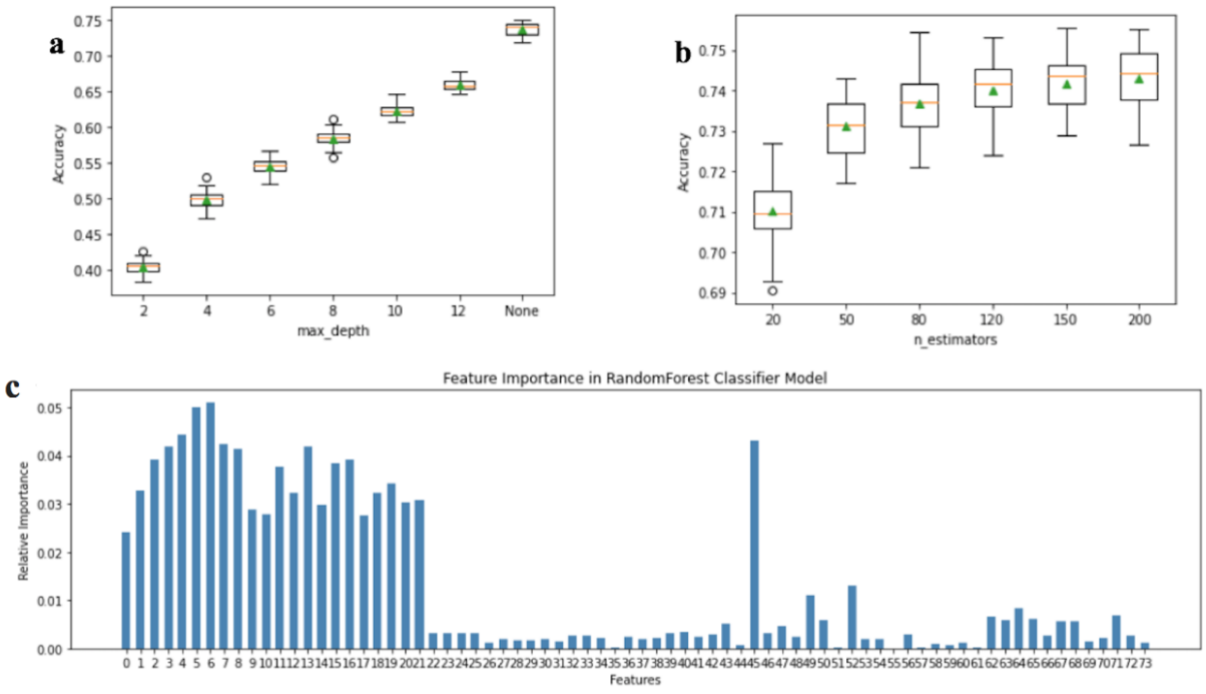

Fig. 23. Performance evaluation of Random Forest model. (a) Tuning max depth. (b) Tuning number of trees. (c) Feature importance for numerical and categorical attributes

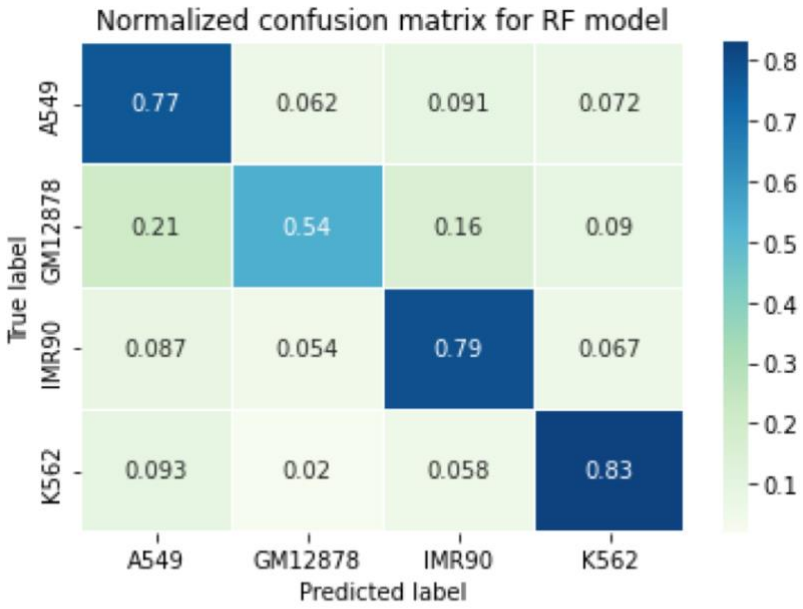

Fig. 24. Normalized confusion matrix for Random Forest classifier.

XGBoost model was implemented and has been found to perform better than Random Forest as expected. The highest model accuracy was recorded at 81.13%, achieved by hyperparameter tuning of max depth, number of trees, learning rate, and subsampling. After optimization, model accuracy was highest when number of tree depth was limited to 10, which was the opposite of Random Forest where unrestrained tree depth was assigned. XGB also benefits from higher number of trees than Random Forest; for XGB diminishing returns were not observed at 300 trees or higher. Since this model consists of both shrinkage from AdaBoost and column subsampling from Random Forest to handle overfitting, its improved generalization enhances classification accuracy in each cell line. Normalized confusion matrix indicated K562 and IMR90 were best classified at 83%, followed by A549 at 82% and GM12878 at 69% (Fig. 25).

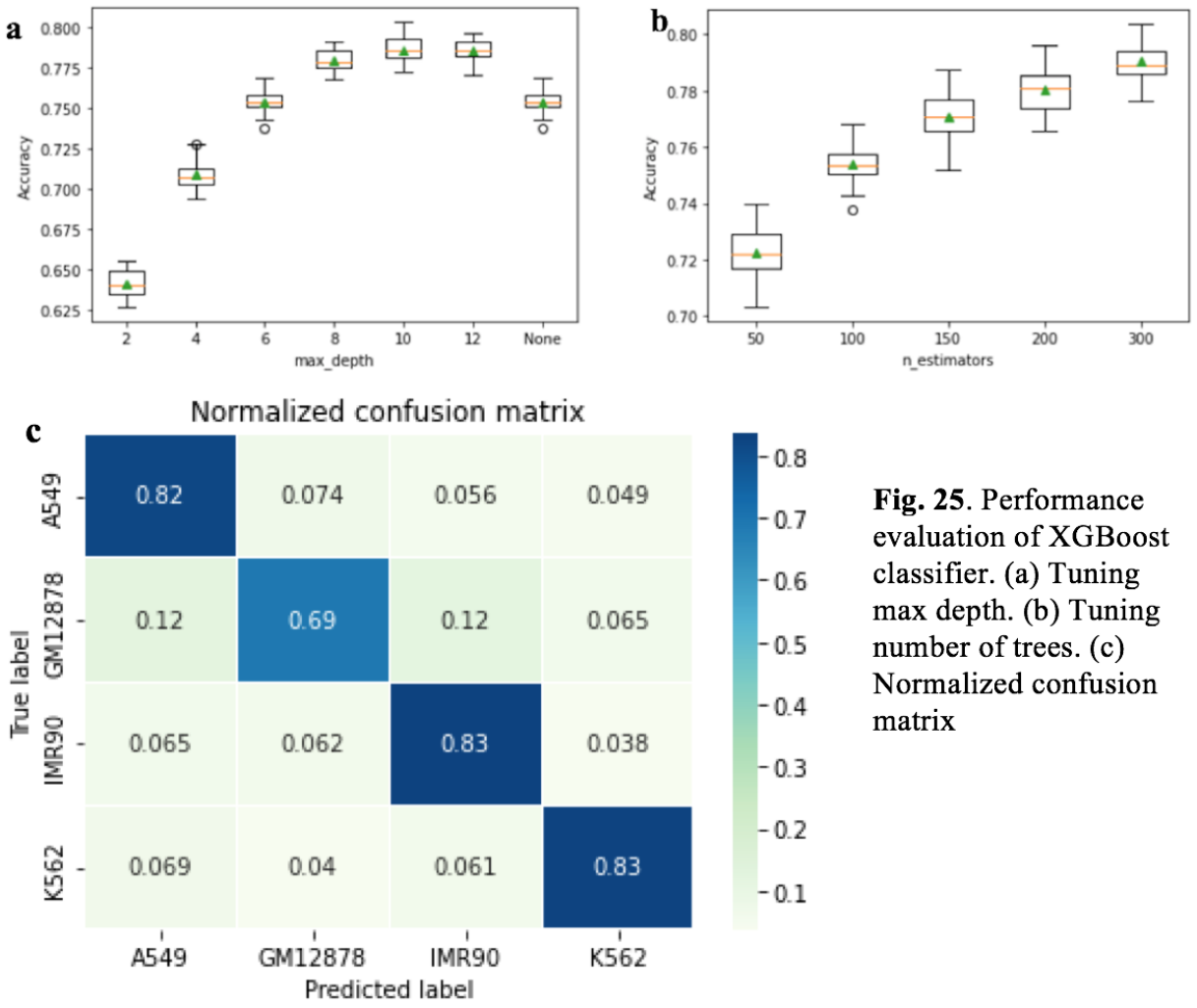

Fig. 25. Performance evaluation of XGBoost classifier. (a) Tuning max depth. (b) Tuning number of trees. (c) Normalized confusion matrix

The third model, TabNet classifier also produces promising results, given that deep neural networks typically do not perform well with tabular data. The highest validation accuracy was 73.50%, achieved with prediction layer width 64, mask width 64, and number of steps 5. These values were close to the default or recommended values in the original publication. All other parameters were included in optimization, but they should be left at default values for the best model accuracy. Normalized confusion matrix indicated IMR90 were best classified at 74%, followed by A549 at 73%, K562 at 72%, and GM12878 at 52% (Fig. 26).

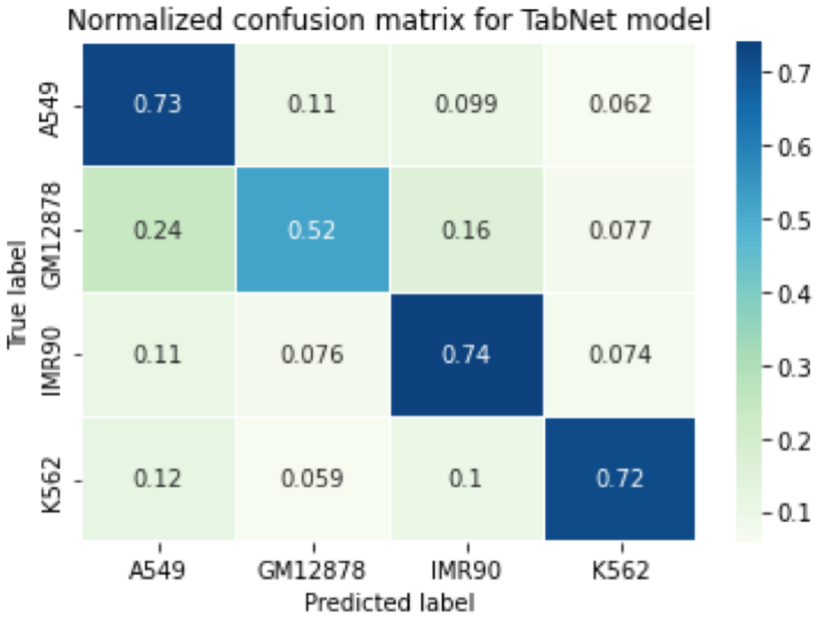

Fig. 26. Normalized confusion matrix for TabNet classifier

IV. Discussion

Chromatin interactions are captured on contact matrices. Important components such as CTCF, cohesin, and histone marks provide better insights to gain the awareness of important roles of genomic structure and epigenetic regulation in sustaining and supporting cell growth, development, and proliferation. Be thankful to the invention of Hi-C technique to capture chromatin structures and better visibility resulted from high-resolution maps. In addition to the key diagonal axis, multiple square patterns in the background of the matrices reflect the intervals of subcompartments in which compartment A harbors activating chromatin marks and is enriched for euchromatin while compartment B harbors inactivating marks and is enriched for heterochromatin. This allows contact maps to contain other information about the surrounding of TADs or contact domains.

Hi-C maps reveal various structures at certain loci on specific chromosomes, and loop domains become visible at higher resolution maps. However, some discrepancies occur between contact maps of different chromosome regions among cell lines. This could be due to biological variations such as dynamics in gene folding, the presence/absence of loop domains in the same regions, or different

expression profiles. Technical variations include experimental biases (different reagents, timing, protocols), or artifacts in capturing images, processing errors, and the use of multiple tools on the same dataset. These variations potentially contribute to the differences in the same genomic regions.

Why were chromatin state distributions, PSE vs. NSE or enhancer vs. repressor, varied at each boundary strength? Histone marks and TAD features such as locations, insulation scores, and boundaries are explored for all chromosomes within a cell. From developmental perspective, different regulatory regions turn on or off certain genes to determine cell fate; in other words, active expression of relevant genes is required for a 'general' cell to be differentiated into a heart cell or skin cell. In this project, each chromosome has different boundary strengths, and certain loci hold more enhancers (high H3K27ac level) or more repressors than other regions so that certain genes would be highly expressed while others would be repressed. Although TAD boundaries are invariant across cell types, the modification of histone marks could be varied depending on gene locations at different cell stages. Number of H3K27ac and H3K27me3 marks are observed across all boundary strengths, and each cell line has a different profile of histone mark distributions. When looking at chromatin state distribution per chromosome, PSE distribution is high in the presence of activators and low in the presence of repressor.

Overall, the differences in chromatin state distributions among cell lines could be due to both biological and technical variations. In term of biological variations, lung cancer cells A549 have vastly different distributions of both PSE and NSE from normal lung cell IMR90, and it is expected to have some genes abnormally upregulated or suppressed compared to normal lung cell line IMR90. Thus, the difference in PSE and NSE distributions might be an indicator of what has gone wrong in cell growth and proliferation. It would be helpful to look at corresponding gene expression level and relevant chromosomes to narrow down the scope of investigation. The results from the chromatin state variations in chromosome 1, 13, and 21 suggest that super-enhancers tend to upregulate instead of downregulate transcriptional activities or gene expression based on the frequent presence of PSE instead of NSE. In other words, transcriptional activation through high PSE and strong enhancers could

be the dominant scheme compared to downregulation from repressors or moderate NSE within the scope of the chosen three chromosomes in this project.

There is low number of state distributions in K562 and GM12878 in certain chromosomes. This might be due to technical errors during Hi-C experiments or processing Hi-C data. Some contact domains might have been lost during TAD calling, causing noticeably low distribution.

Some limitations in processing Hi-C data include out-of-memory error, difference in tools for generating normalized maps and analyzing features, and utilizing processed data from other labs. Besides, the approximation of TAD window to create checking criteria when combining of different features to build tabular dataset contribution to variation or estimation of how all features organized in every single row. Model accuracies around 70% to 85% are potentially indicators of how well data were engineered and organized for classification.

V. References

[1] Chang, P., Gohain, M., et al., Computational Methods for Assessing Chromatin Hierarchy. *Comp Struc Biotech J.*, 16. 2018

[2] Fudenberg G., Imakaev, M., et al., Formation of chromosomal domains by loop extrusion. *Cell Reports*, 15, 2038-2049, 2016

[3] Rowley, M.J., & Victor, G.C., Organization principles of 3D genome architecture. *Nature Reviews: Genetics*, 19, 2018

[4] Rao, S., Huang, S.C., et al., Cohesin loss eliminates all loop domains. *Cell*, vol. 171, no. 2, 2017

[5] Rao, S. S. P., Huntley, M. H., Durand, N. C., et al., A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, 159, 2014

[6] Kyrchanova, O., & Georgiev, P., Mechanisms of enhancer-promoter interactions in higher eukaryotes. *Int J Mol Sci.*, vol. 22, no. 2, 2021

[7] Gong, Y., Lazaris, C, et al., Stratification of TAD boundaries reveals preferential insulation of super-enhancers by strong boundaries. *Nature Communications*, 9, 2018

[8] Sur I., & Taipale, J., The role of enhancers in cancer. *Nat Rev Cancer*, 16, 2016

[9] Ong, C.T., & Corces V. G., Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat Rev Genet.*, vol. 12, no. 4, 2011

[10] Ernst J., Kheradpour, P., et al., Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*., 2011

[11] Hnisz, D., Schuijers, J., et al., Convergence of developmental and oncogenic signaling pathways at transcriptional super-enhancers. *Molecular Cell*, vol. 58, no. 2, 2015

[12] Kruse, K., Hug, C. B., & Vaquerizas, J. M., FAN-C: a feature-rich framework for the analysis and visualization of chromosome conformation capture data. *Genome Biology*, vol. 21, no. 303, 2020

[13] Xu, W., Zhong, Q., et al., CoolBox: A flexible toolkit for visual analysis of genomics data. *BMC: Bioinformatics*, 22, 2021

[14] Hniz, D, Schuijers, J., et al. Convergence of developmental and oncogenic signaling pathways at transcriptional super-enhancers. *Molecular Cell*, vol. 58, no. 2, 2015

[15] D'Ippolito, A.M., McDowell, I.C., et al., Pre-established chromatin interactions mediate the genomic response to glucocorticoid. *Cell Systems* 7, 2018

[16] Bowtie2. Retrieved from http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

[17] Ho, K. T., Random decision forests. *IEEE Xplore*, vol. 1, p. 278-282, 1995

[18] Chen, T. and Guestrin, C., XGBoost: A scalable tree boosting system. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 785–794, 2016

[19] Arik, S. O. and Pfister, T., TabNet: Attentive Interpretable Tabular Learning. *Assoc Adv AI*, 2021

Appendix for feature engineering and corresponding data sources

Feature	Data or Feature Source	Note
Chromosomes	All datasets	Chromosomal locations
TAD locations	JuicerTools - Arrowhead	Start and end locations of TADs
Loop locations	JuicerTools - HICCUPS	Start and end locations of loop domains/peak loci
CTCF status	ENCODE ChIP-Seq (processed)	Query using TAD window. CTCF occurs at both ends or one end, and no CTCF presence
Sum of insulation scores	FANC insulation	Summation of all insulation scores within a contact domains
Insulating boundaries at TAD start location	FANC boundaries	Insulating boundary scores at TAD start locations Query using TAD window.
Insulating boundaries at TAD end location	FANC boundaries	Insulating boundary scores at TAD end locations Query using TAD window.
Directionality index at TAD start location	FANC directionality index	Directionality index score at TAD start location Query using TAD window.
Directionality index at TAD end location	FANC directionality index	Directionality index score at TAD end location Query using TAD window.
Mean location of H3K27ac	ENCODE ChIP-Seq (processed)	Mean location of the activator within a TAD or contact domain

Number of H3K27ac	ENCODE ChIP-Seq (processed)	Number of the activator presence within a TAD or contact domain
Mean location of H3K27me3	ENCODE ChIP-Seq (processed)	Mean location of the repressor within a TAD or contact domain. Query TAD window.
Number of H3K27me3	ENCODE ChIP-Seq (processed)	Number of the repressor presence within a TAD or contact domain. Query TAD window.
Loop category/type	RAD21 from ENCODE ChIP-Seq (processed)	Types: cohesin-associated (CA) and cohesin-independent (CI) loops Query using loop domain window from HICCUPS).
RAD21 distance from loop start and end locations	RAD21 from ENCODE ChIP-Seq (processed)	Mean distance between RAD21 position and loop start or end location. Query using loop domain window from HICCUPS).
Structural effect	Multiple features above	Query loop locations with TAD locations. Classified based on CTCF status and loop types. More details are under Method section
Boundary strength	Summation of insulation scores	High insulation score indicates strong boundaries and vice versa
Chromatin states	Multiple features above	Classified based on CTCF status, loop types, and the number of H3K27ac and H3K27me3 presence. More details are under Method section