공학석사 학위논문

Deep learning-based survival prediction using DNA methylation-derived 3D genomic information

DNA 메틸화 데이터로부터 추출한 3차원 유전체 정보를 활용한 딥 러닝 모델 기반 생존 예측

2023 년 8 월

서울대학교 대학원

협동과정 인공지능전공

양 지 원

Deep learning-based survival prediction using DNA methylation-derived 3D genomic information

DNA 메틸화 데이터로부터 추출한
3차원 유전체 정보를 활용한
딥 러닝 모델 기반 생존 예측

지도교수 김 선

이 논문을 공학석사 학위논문으로 제출함

2023 년 6 월

서울대학교 대학원

협동과정 인공지능전공

양 지 원

양지원의 공학석사 학위논문을 인준함

2023 년 6 월

위 원 장	황대희
부위원장	김선
위 원	임상수

Abstract

Deep learning-based survival prediction using DNA methylation-derived 3D genomic information

Jeewon Yang
Interdisciplinary Program in Artificial Intelligence
College of Engineering
Seoul National University

The development of cancer is strongly linked to the three-dimensional (3D) genome structure. However, the valuable information related to the 3D genome states has not been effectively used in clinical applications, to the best of my knowledge. The main reason for this is the expensive production of Hi-C data, the manifest source of 3D genome information. Therefore, there is a requirement for a new measurement that can be derived from 3D genome-related data, making it more readily available for the 3D genome information to be clinically used.

In this study, I present a novel approach for extracting 3D genome-aware epigenetic features, the epigenetic features that are reflective of the three-dimensional (3D) genome structure, from DNA methylation data. Additionally, I conducted a deep learning-based survival analysis utilizing these features. To generate the 3D genome-aware epigenetic features, the 3D genome structures were reconstructed using the 450K DNA methylation data at an individual level. The results demonstrate that utilizing these features significantly improves the accuracy of risk prediction for seven cancer types. This suggests that the 3D genome information embedded in the 3D genome-aware epigenetic features is highly valuable for predicting the survival, or cancer prognosis.

Furthermore, an in-depth biological analysis revealed that altered DNA methylation levels in risk-high group as defined by the deep learning model are associated with the aberrant activation of genes involved in various cancer-related pathways. Overall, the usage of 3D genome-aware epigenetic features as survival predictors demonstrates their significant clinical importance in seven types of cancer, in addition to their biological significance. All source codes are available on the GitHub repository (https://github.com/jwyang21/3D-genome-risk-prediction).

Keywords: Deep learning, Bioinformatics, DNA methylation, Cancer prognosis prediction, 3D genome, Epigenetics, Survival analysis

Student Number: 2021-26775

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

The 3D genome structure plays a crucial role in determining cell fate and establishing cell identity. It provides spatial constraints that regulate the activity of transcription factors (Stadhouders *et al.*, 2019). This implies that a misconfiguration of the 3D genome structure can activate abnormal transcriptional programs, potentially leading to the malignant transformation of cells and the development of cancer (Gröschel *et al.*, 2014; Umlauf and Mourad, 2019). Moreover, the 3D genome organization is known to drive oncogenic structural variations (Rheinbay *et al.*, 2020), including driver fusion events in which proto-oncogenes hijack enhancers of other genes to promote their expression (Dubois *et al.*, 2022). Collectively, it can be inferred that the 3D genome structure plays a significant role in multiple facets of cancer biology, and any disturbance to its organization can have harmful effects to the cell fate decision. However, as far as I know, there has been no investigation that employs quantified information about the 3D genome in predicting cancer prognosis.

This can be attributed to two primary factors: the limited availability of data capturing the 3D genome conformation in cancer samples and the absence of quantitative measures to assess perturbations in the 3D genome. The most preferred data representation for capturing the 3D genome landscape is contact matrices or contact maps, which are obtained from the high-throughput chromosome conformation capture (Hi-C) experiments (Van Berkum *et al.*, 2010). These matrices/maps offer direct information about the interactions between various genomic regions. In detail, contact maps provide information about the frequency of contact between any two genomic loci, which can help estimate the strength of their interaction. Nevertheless, there are significant constraints when it comes to generating contact matrices, due to the high expense and intricate processes involved in Hi-C experiments (Yardımcı *et al.*, 2019; Zhang *et al.*, 2019a). These difficulties ultimately lead to a restricted supply of 3D genome contact maps. As a result, an alternative dataset that is more abundant and encompasses 3D genomic information is required for the development of prognostic score.

1.2 Motivation

Surprisingly, it has been found that 3D genome configurations can also be reconstructed using DNA methylation data alone (Fortin and Hansen, 2015). This is because the interaction between distant genomic regions results in covarying DNA methylation levels of CpG sites within those regions. As a result, the correlative patterns observed in DNA methylation profiles could reflect the organization of the 3D genome. In particular, the distribution of A/B compartments, which are distinguished from each other in terms of chromatin conformation on a large scale spanning multiple megabases (Liu *et al.*, 2021; Di Stefano and Cavalli, 2022; Magaña-Acosta and Valadez-Graham, 2020), was reconstructed from the DNA methylation data (Fortin and Hansen, 2015). For this purpose, data generated through the DNA methylation microarray techniques, including Illumina Infinium HumanMethylation450 (450K) BeadChip array, were utilized (Bibikova *et al.*, 2011). Unlike Hi-C contact matrices, DNA methylation profiles are available for each patient in numerous cancer types, providing us with more extensive and detailed information.

From Fortin and Hansen (2015), it is observed that the 3D genome information reconstructed from the DNA methylation data (i.e., the PC1 vector representing the A/B compartment distribution) can reproduce that from the Hi-C data to a reliable extent. The authors of (Fortin and Hansen, 2015) attribute the successful replication of 3D genomic information to the long-range correlations embedded in the 450K DNA methylation data. The authors explain that there is a high correlation between DNA methylation levels from two loci belonging to the same compartment compared to the two loci from different compartments, and leveraging these long-range correlations enabled extracting the 3D genome structure from the DNA methylation data.

While the potential of DNA methylation data in inferring the 3D genome has been recognized, the ability to accurately speculate the individual 3D genome state has not yet been attainable. Although the Hi-C data are obtainable at an individual level, the low availability of the Hi-C data itself limits clinical utilization of the Hi-C data. Inspired from this limitation, I have devised a technique for extracting individual 3D genome information from DNA methylation data, henceforth referred to as 3D genome-aware epigenetic features, and used these features to predict the risk of failures for different survival events. Specifically, the differential structure of DNA methylation that approximates the correlative pattern was captured. Using the observation that open sea CpG positions located far from the CpG islands demonstrated the highest predictive capability in 3D genome reconstruction (Fortin and Hansen, 2015), the same probes were employed for my analysis. I employed a deep learning model for survival prediction, utilizing (Katzman *et al.*, 2018). The model used the 3D genome-aware epigenetic features as input to predict risks. Subsequently, a comprehensive survival analysis across multiple cancer types was performed, followed by the interpretation of biological implications. This research introduces a novel method for predicting risks by utilizing the 3D genome-aware epigenetic features. The pan-cancer analysis uncovered notable differences in survival patterns between the risk-high and risk-low groups, indicating the potential of predicted risks as a significant prognostic predictor. The functional annotation of genes located in differentially methylated regions (DMRs) revealed the engagement of DMR genes in numerous cancer-related pathways. Notably, the retinoic acid (RA) signaling pathway, essential for the developmental process (Ozgun *et al.*, 2021), emerged as one of the distinctive pathways. Analysis on the chromatin states of DMRs revealed that the majority of DMRs exhibited inactive or moderately activated states, suggesting a relationship between the altered DNA methylation levels and abnormal gene activation. In conclusion, the epigenetic features reflecting the 3D genome hold significant predictive information for cancer prognosis and carry biological importance.

Chapter 2

Task design and Approach

2.1 Underlying concepts

The entire methodology comprises two main parts: feature engineering and survival prediction using a deep learning model. Initially, features encompassing the 3D genome information are extracted from DNA methylation data, which are subsequently utilized as input for the pan-cancer survival prediction. This section provides a concise overview, and a detailed description of each key concept is presented in the Methods section.

1. The ability to infer individual 3D genome enables quantification of the 3D genome state for each cancer patient, enhancing the clinical applicability of inferred 3D genome structure.

2. The inferred 3D genome structures are expressed as vectors, specifically the first principal components (PC1s), enabling the assessment of dissimilarities between two distinct individual 3D genome states.

3. To construct representations of normal or stem cells' 3D genome struc-

tures, or stem/normal references, the 3D genome states of multiple samples are averaged.

4. The degree of dissimilarity between each individual's 3D genome state and the normal/stem reference is determined by calculating the distances. These distances, hereafter called stem/normal distances, enable the quantification of stem closeness of each sample.

5. Three risk prediction scenarios were examined: utilizing 3D genome-aware, 3D genome-unaware, or no epigenetic features in a deep learning-based risk prediction model.

2.2 Task definition

My research suggests that the 3D genome organization can be represented as PC1 vectors from the DNA methylation levels of open sea CpG positions. These vectors play a crucial role in enhancing the accuracy of risk prediction for cancer patients. In fact, the representation of differential DNA methylation states as PC1s is widely utilized to illustrate chromatin conformation (Lieberman-Aiden *et al.*, 2009; Wang *et al.*, 2016). These PC1 vectors stand for the inferred 3D genome states, with each entry indicating the A/B compartment of the corresponding genomic bins. By utilizing these vectors, it becomes possible to quantify dissimilarities between distinct individual 3D genome structures and extract 3D genome-aware features.

Fig. 2.1 provides an overview of the entire pipeline. In contrast to the approach in (Fortin and Hansen, 2015), which infers a single 3D genome state specific to a particular tissue type using multiple samples (Fig. 2.2), my method rebuilds the 3D genome structure for each individual separately. The dissimilarities between individual 3D genome states are quantified by measuring the distance between PC1s.

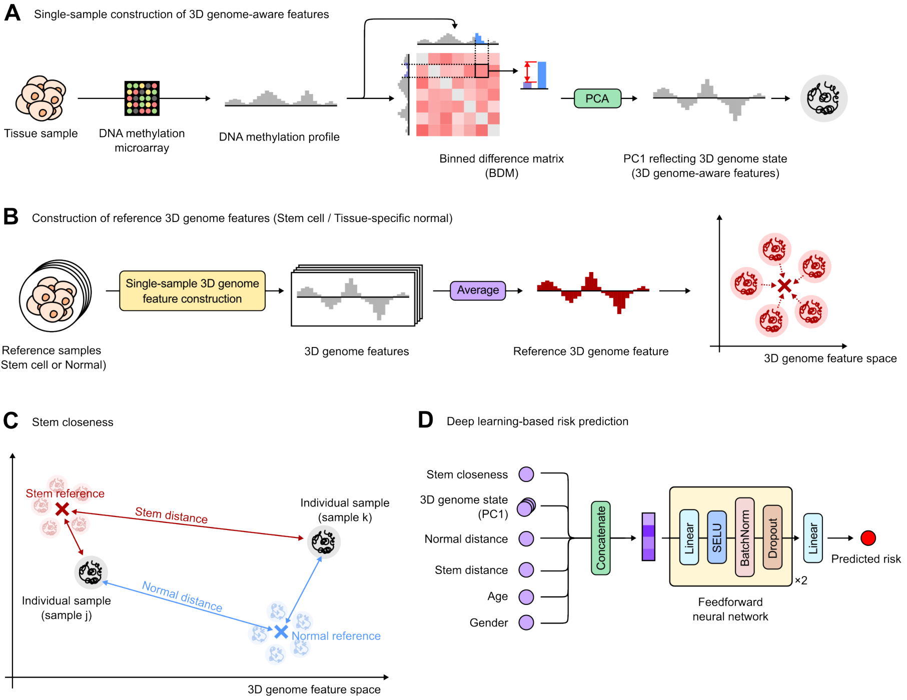

A
Single-sample construction of 3D genome-aware features
Tissue sample
DNA methylation microarray
DNA methylation profile
Binned difference matrix (BDM)
PCA
PC1 reflecting 3D genome state (3D genome-aware features)
B
Construction of reference 3D genome features (Stem cell / Tissue-specific normal)
Reference samples (Stem cell or Normal)
Single-sample 3D genome feature construction
3D genome features
Average
Reference 3D genome feature
3D genome feature space
C
Stem closeness
Stem reference
Stem distance
Individual sample (sample k)
Individual sample (sample j)
Normal distance
Normal reference
3D genome feature space
D
Deep learning-based risk prediction
Stem closeness
3D genome state (PC1)
Normal distance
Stem distance
Age
Gender
Concatenate
Linear
SELU
BatchNorm
Dropout
×2
Linear
Predicted risk
Feedforward neural network

Figure 2.1: Overview of the whole pipeline.

(A) Construction of the binned difference matrix (BDM) using the 450K DNA methylation data. (B) Construction of stem and normal reference of 3D genome states. (C) Computation of normal and stem distances of each sample. (D) Usage of 3D genome-aware epigenetic features, age and gender for a deep learning-based risk prediction.

By utilizing normal/stem references, stem closeness of each individual is assessed. These 3D genome-aware epigenetic features are concatenated along with age and gender, and employed as input for the risk prediction by a feed-forward neural network. Prior to concatenation, each input feature was properly normalized to the range of [0, 1] to prevent the problems which might arise from the inconsistent scales among different features. Subsequently, a pan-cancer survival analysis is conducted, followed by functional analysis of

DMRs to investigate the biological significance of the predicted risks. For a more quantitative explanation of the extraction of 3D genome-aware epigenetic features, please refer to Fig. 2.3.

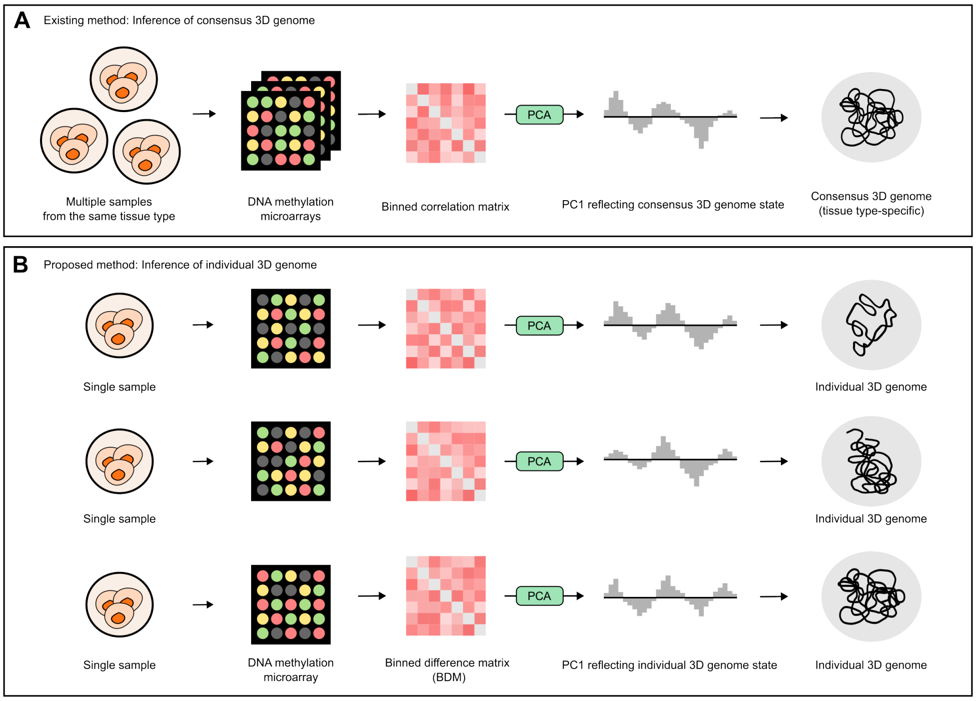
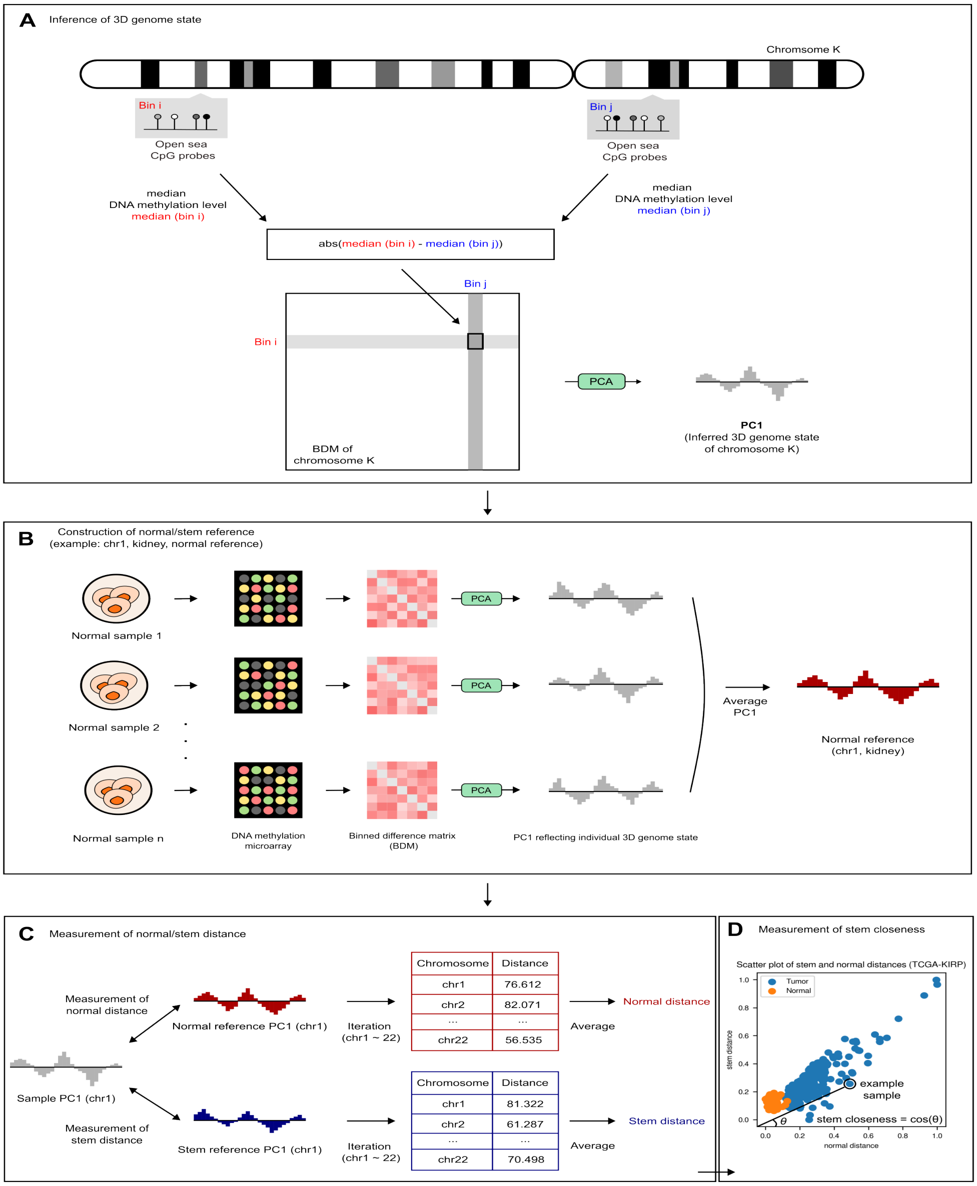


Figure 2.2: The comparison between existing method and the proposed methods for inferring 3D genome structure from the DNA methylation data.

(A) An existing method. DNA methylation data from multiple samples of the same tissue type are utilized to infer a single consensus 3D genome structure. (B) The method proposed in this study. This method focuses on single-sample-based inference of the 3D genome structure using DNA methylation data.



Chromosome	Distance
chr1	76.612
chr2	82.071
...	...
chr22	56.535

Chromosome	Distance
chr1	81.322
chr2	61.287
...	...
chr22	70.498

Figure 2.3: A quantitative explanation of the entire pipeline involved in extracting 3D genome aware epigenetic features from DNA methylation data.

(A) Inference of the 3D genome structure from each autosome of a single sample. (B) Construction of normal/stem references. (C) Measurement of normal/stem distances. (D) Measurement of stem closeness.

Collectively, the deep learning model takes as input the concatenation of 3D genome-aware epigenetic features and features related to survival (age and gender), and outputs the predicted risk for each survival event per individual. A deep learning model is necessitated for this task because of its ability to learn the underlying pattern from the large-dimensional input feature. The feature dimension of fully concatenated input feature exceeds 2700, which makes learning meaningful pattern out of this data a very complicated task. Although there are many other alternative methods for risk prediction which is not deep learning-based, it is considered that the ability of deep learning model to learn the underlying pattern accurately and rapidly is needed for this task. The more detailed explanation on how the deep learning model was used for training and evaluation, including the specifications and the loss function, is provided in 3.8.

Chapter 3

Materials and Methods

3.1 Deriving 3D genome-aware epigenetic features in an individual-level

The 3D genome organization consists of two distinct classes of large genomic compartments with contrasting characteristics: euchromatic A and heterochromatic B compartments, which form the top-level hierarchy of 3D genome structure (Rowley and Corces, 2018). The identification of these compartments can be achieved by analyzing high-throughput measurements of genomic contacts, as inter-compartment contacts are considerably rarer than intra-compartment contacts. This correlative structure of the 3D genome is observed as a prominent plaid pattern in the Hi-C contact frequency matrix or the PC1 values derived from the normalized Hi-C matrix. The Hi-C PC1 values, which capture the largest variability in the Hi-C matrix, indicate the compartmentalization states of different genomic bins (Du *et al.*, 2021; Golloshi *et al.*, 2022; Schmitt *et al.*, 2016).

Inspired by the previous observations of co-varying DNA methylation levels across genomic regions (Fortin and Hansen, 2015), I hypothesized that a matrix representing the absolute difference in DNA methylation levels between two arbitrary genomic bins could also capture the correlative structure of the 3D genome. In other words, I anticipated that regions exhibiting spatial association would display smaller differences in methylation levels. To reflect this concept, I designed a matrix, henceforth called the binned difference matrix (BDM), which quantifies the absolute differences in methylation levels between genomic regions. The detailed procedures for constructing BDM is provided in section 3.2.

3.2 Construction of BDM

This section provides a detailed explanation of the process involved in constructing BDM. BDM aims to capture the differential structure of DNA methylation s across different regions of the genome. To achieve this, it is important to establish a representative value that stands for the overall DNA methylation level for each region, enabling the measurement of differences between these values. To begin, each autosome is divided into bins of size 1Mb. Within each genomic bin, the median value is calculated by considering all the DNA methylation levels of open sea CpG positions located in that bin (Fig. 2.3). This is based on another fundamental principle of BDM, which focuses exclusively on using distant or open sea CpG probes. These probes have demonstrated superior predictive power compared to CpG probes of other positions (such as islands, shelves, and shores) in terms of inferring the 3D genome from DNA methylation levels (Fortin and Hansen, 2015). Subsequently, the absolute difference between these median values is computed for all possible pairs of bins. These resulting values serve as entries for the BDM of the corresponding chromosome. The row and column indices of the BDM are consistent with the bin indices from which the difference value originated. This entire process is repeated for all autosomes. The specific steps involved are depicted in Fig. 2.3. For convenience, the construction of BDM for an arbitrary chromosome, K, is illustrated. Here, K can take any integer value between 1 and 22.

3.3 Investigating the characteristics of BDM

In order to incorporate BDMs into the development of a prognostic score, I posited that certain criteria must be met by the BDMs: (1) containing an ample amount of 3D genome information, (2) exhibiting clear distinctions between tumor and normal groups, and (3) displaying tissue-type specificity. Several experiments were conducted to validate whether the BDMs fulfilled these criteria. To assess whether the 3D genome information is included in the BDMs, I calculated the Pearson Correlation Coefficient (PCC) between the BDM PC1s and the Hi-C PC1s. Additionally, I visualized the BDMs of both tumor and normal groups as heatmaps to examine the differences between two groups. Furthermore, I evaluated all-pairwise PCC values between the averaged BDM PC1s to ensure that the PCC values within homogeneous pairs, consisting of BDM PC1s from the same TCGA cohort, were higher than those from all the other cases.

3.4 Devising a prognostic score from the BDM PC1s

The primary assumption is that the BDM PC1 values reflect the individual 3D genome structure, enabling the measurement of differences between distinct 3D genome states as distances between vectors. Building upon this concept, I evaluated the similarity of each cell to stem cells or normal cells (stem- and normal-likeness, respectively) by comparing inferred 3D genome structure of each cell to the stem/normal reference (Fig. 2.1 and Fig. 2.3). The resulting distances between each BDM PC1 value and the stem/normal reference, referred to as stem/normal distance, represent the dissimilarity of each sample to the states of stem cells or normal cells (Fig. 2.1 and Fig. 2.3). I investigated two distance metrics: the Euclidean distance and the inverse of cosine similarity. For the cosine similarity metric, I added a pseudo count (10^{-15}) to the similarity measure before taking the inverse to avoid division by zero. The biological interpretation of each distance metric suggests that a more stem-like cell would have a higher normal distance and a smaller stem distance. In line with this concept, I formulated a prognostic score, which quantifies the stem closeness of a cell, based on this idea (Eq. 3.1). In Eq. 3.1, d_n and d_s denote normal and stem distance, respectively.

$$\frac{d_n}{\sqrt{(d_n)^2 + (d_s)^2}} \tag{3.1}$$

The more comprehensive explanation of the procedures for extracting the 3D genome-aware epigenetic features from BDMs, along with the underlying rationale, are provided in the section 3.5-3.6.

3.5 Extracting 3D genome-aware epigenetic features from BDM

Once the construction of BDM is complete, the next step involves constructing normal and stem references based on the first principal component (PC1) of the BDMs (Fig. 2.3). Assuming there are n samples from the same tissue type (e.g., kidney), the BDM of chromosome 1 (chr1) can be constructed for each sample. The resulting n PC1 values are averaged to obtain a normal reference for chr1 in kidney. Similarly, a stem reference is computed by replacing the normal samples with stem cells. Afterwards, the normal/stem distances, the distances between normal/stem references and individual samples, are calculated by measuring the distance between the PC1 vectors. Specifically, the distance from a sample's PC1 to the corresponding normal/stem reference is measured for each autosome. This results in a total of 44 values, with half of them obtained using normal references and the other half using stem references. The 22 values computed using the normal references are averaged to yield a single scalar value, or the normal distance. The same process is followed for the stem distance, which involves measuring distances between a sample's PC1 and the stem reference. Finally, the stem closeness of each sample is evaluated based on the stem/normal distances. The normal and stem references are plotted in a two-dimensional Cartesian space, with a straight line connecting the origin and a dot representing each sample. The angle (θ) formed between this line and the x axis is measured, and the cosine of θ ($\cos\theta$) is used as a measure of stem closeness. The process of measuring stem closeness is based on several observations and hypotheses. It has been observed that there is a significant positive correlation between normal and stem distances across all cohorts used in the experiment. The hypothesis is that if the stem and normal distances consistently move in the same direction, then a smaller increase in the stem distance followed by a unit increase (e.g., +0.1) in the normal distance would indicate a higher similarity to stem cells (i.e., stem closeness) for a given sample compared to other samples. The usage of $\cos\theta$ is based on this reasoning, taking into account certain background concepts. Firstly, in the Cartesian 2D space where the x and y axes represent the normal and stem distances, respectively, all points representing the samples are located in the first quadrant since the distances are non-negative. Secondly, the angle (measured in degrees) between the x axis and any arbitrary point situated in the first quadrant falls within the range of . Lastly, within this range of angles, $\cos\theta$ consistently decreases as *theta* increases. Consequently, samples located on a steeper line in the plot would have smaller $\cos\theta$ values compared to those on a line with a gentler slope. This concept was tested across various cancer types, and the results were found to be consistent with the initial assumption.

In detail, the normal samples exhibited smaller $\cos\theta$ values than the tumor samples, indicating their proximity to the y axis. The most distinct example of this pattern, illustrated in Fig. 2.3, is the case of TCGA-KIRP. In summary, the measurement of stem closeness is based on the understanding that the relationship between normal and stem distances follows a consistent pattern, and the use of $\cos\theta$ helps quantify this closeness based on the angles formed in the Cartesian 2D space representing the distances. The experimental results across different cancer types align with the expectations derived from these assumptions.

3.6 Figuring out the optimal stem closeness of each cohort

To ascertain the ideal measure of stem closeness for each cancer type, a comprehensive exploration of various parameter combinations was conducted. The analysis involved considering all possible combinations and carefully evaluating their performance. Through this rigorous process, the most effective combination of parameters was identified for each cohort. The specific details of the chosen parameter combinations for each cohort are provided in Table 3.1. Each subsection provides an explanation of the parameters utilized in the experiment, and the methodology employed to determine the optimal combination of parameters for each cohort, respectively.

3.6.1 Parameters

- Distance metric: When calculating the distance between two PC1 vectors, there are two possible options: Euclidean distance and cosine similarity. If cosine similarity is selected, a pseudo count of 10^{-15} is added to enhance the similarity calculation. The resulting similarity value is then inverted to obtain the distance value.

- Matrix type: Interestingly, during preliminary experiments, a fascinating observation was made: the PC1s derived from the inverse exponential of BDM (IEBDM) exhibited remarkably high correlation with the PC1s derived from BDM. Building on this finding, the IEBDM PC1s were also used as an available option in the analysis. In the case of using IEBDM, the regular BDM was substituted with IEBDM in every step of the pipeline depicted in Fig. 2.3.

- Averaging method: To evaluate the stem/normal distances, the average of 22 distances is computed between the BDM PC1s of each individual sample and the corresponding reference PC1s. Two methods were employed for this averaging process: simple averaging and weighted averaging. In the case of weighted averaging, the ratio of each autosome length to the sum of all autosome lengths is utilized as the weights.

- Min-max scaling: If the option of min-max scaling is selected, the normal/stem distances are transformed to fit within the range of [0, 1]. To achieve this, the maximum and minimum values of normal/normal distances for each cohort are recorded. Then, each distance value is scaled using Eq. 3.2. In Eq. 3.2, x_i represents the normal/stem distance of the i-th sample, $min(\mathbf{x})$ is the minimum value of normal/stem distances within the current cohort, and $max(\mathbf{x})$ is the maximum value of normal/stem distances within the current cohort. This scaling procedure takes place between the steps illustrated in Fig. 2.3.

$$\frac{x_i - min(\mathbf{x})}{max(\mathbf{x}) - min(\mathbf{x})} \tag{3.2}$$

- Normalization: When the option of scaling normal/stem distances into the range of [0, 1] is chosen, it is accomplished by applying Eq. 3.3. In Eq. 3.3, x_i represents the normal/stem distance of the i-th sample, and

$max(\mathbf{x})$ signifies the maximum value among the normal/stem distances within the current cohort.

$$\frac{x_i}{max(\mathbf{x})} \tag{3.3}$$

- Standardization: When the option of standardizing PC1 vectors is selected, it is done by applying Eq. 3.4 prior to computing the distance. In Eq. 3.4, $\mathbf{y}$ represents each PC1 vector, y_j represents the j-th entry of $\mathbf{y}$, $mean(\mathbf{y})$ denotes the average value of all entries in the PC1 vector $\mathbf{y}$, and $std(\mathbf{y})$ represents the standard deviation of all entries in $\mathbf{y}$.

$$\frac{y_j - mean(\mathbf{y})}{std(\mathbf{y})} \tag{3.4}$$

- Number of chromosomes (num_{chrom}): From the work of Fortin and Hansen (2015), it was shown that utilizing smaller chromosomes resulted in a decrease in the accuracy of reproducing the 3D genome structure. Additionally, it was observed that using the entire set of chromosomes did not always yield better results compared to using only a portion of genomic bins. Taking these findings into account, the number of autosomes used in the analysis was established as a parameter. In detail, PC1 vectors from chromosome 1 up to chromosome n was employed, where n is an integer ranging from 1 to 22. Fig. 2.3 illustrates the scenario where n is set to 22.

3.6.2 Selecting single optimal score per cohort

Following the log-rank tests using stem closeness scores with various parameter combinations, the scores were initially grouped based on the number of survival events in which each score acted as a significant predictor (m). Since the log-rank test was performed for a total of four survival events (Overall survival; OS, Disease-specific survival; DSS, Disease-free interval; DFI, and Progression-free interval; PFI), the value of m ranges from 0 to 4. The stem closeness scores belonging to the group with $m = 4$ were examined first. If there were no scores present in the current group of interest, m was reduced by 1. From the current group of interest, the sum of p-values (sum_p) was computed based on the results of the log-rank test, where the stem closeness was identified as a significant predictor. Once sum_p of all stem closeness scores, was calculated for all stem closeness scores, any scores that predicted a better prognosis for the high score group compared to the low score group (which contradicts the desired outcome of the score) were excluded. Finally, the remaining scores for each cohort were ranked in ascending order based on sum_p, and the score with the smallest sum_p as selected as the final score. All the reported results of the log-rank tests in this manuscript were based on the stem closeness scores selected through these procedures. The same scores were also employed for the Cox regression analysis, following the log-rank test.

3.7 Hi-C data processing

The 4DN Hi-C processing pipeline (Reiff *et al.*, 2022) was utilized to process Hi-C data. Raw Hi-C sequencing data in the form of fastq files for cancer (Heidari *et al.*, 2014), normal (Schmitt *et al.*, 2016), and stem cell lines (Freire-Pritchett *et al.*, 2017; Zhang *et al.*, 2019b) were downloaded from the Sequence Read Archive (SRA) using sra-tools (v2.10.1) and parallel-fastq-dump. Specifically, the following cell lines were downloaded: hepatocellular carcinoma cell line (SRS2627396), colon cancer cell line (SRS3816279), breast cancer cell line (SRS3505364), esophageal adenocarcinoma cell line (SRS3505365), lung (SRS1704412 and SRS1704413) pancreas (SRS1704415, SRS1704416, SRS1704417, and SRS1704418) and human embryonic stem cells (SRS1688434 and SRS3533281). The sequencing reads were then mapped to the hg19 reference genome using bwa (v0.7.17) (Li and Durbin, 2009). The resulting aligned files (bam) were

Table 3.1: **Finalized combinations of parameters for computing stem closeness.**

Cohort	Distance metric	Matrix type	Averaging method	Minmax scaling	Normalization	Standardization	num_{chrom}
BLCA	Cosine similarity	BDM	Simple average	Used	Not used	Used	8
BRCA	Cosine similarity	BDM	Simple average	Used	Not used	Not used	7
CHOL	Euclidean distance	BDM	Simple average	Not used	Used	Used	19
COAD	Euclidean distance	BDM	Simple average	Used	Not used	Used	20
KIRC	Euclidean distance	BDM	Weighted average	Not used	Used	Not used	3
KIRP	Euclidean distance	BDM	Weighted average	Used	Not used	Not used	1
LIHC	Cosine similarity	IEBDM	Simple average	Used	Not used	Used	9
LUAD	Euclidean distance	BDM	Weighted average	Not used	Used	Used	1
LUSC	Euclidean distance	BDM	Simple average	Used	Not used	Used	14
PAAD	Cosine similarity	BDM	Simple average	Used	Not used	Not used	8
PRAD	Euclidean distance	BDM	Simple average	Used	Not used	Used	2
THCA	Euclidean distance	BDM	Weighted average	Used	Not used	Used	2
UCEC	Euclidean distance	BDM	Simple average	Used	Not used	Used	18

converted and processed as files representing Hi-C pair information using pairtools (v0.3.0) (Song *et al.*, 2022). Subsequently, Hi-C interaction frequency matrices were generated using cooler (Abdennur and Mirny, 2020). Finally, A/B compartment analyses were conducted using FAN-C (Kruse *et al.*, 2020).

3.8 Risk prediction using a feedforward neural network and 3D genome-aware epigenetic features

Given that the aforementioned 3D genome-aware epigenetic features contain cancer-related 3D genomic information, it was hypothesized that incorporating these features would lead to superior performance in survival prediction compared to baseline scenarios that do not utilize these features. In detail, two baseline scenarios were examined: (1) using age and gender as survival predictors without any epigenetic features, and (2) using age and gender along with the 3D genome-unaware epigenetic feature (the average DNA methylation level of open sea CpG positions). For risk prediction, a feedforward neural network, as introduced by (Katzman *et al.*, 2018), was employed (Fig. 2.1). The neural network consisted of two hidden layers, each comprising 128 hidden nodes. During training, the average negative log partial likelihood was utilized as the loss function (Eq. 3.5). In Eq. 3.5, $N_{E=1}$ represents the number of patients for whom the event was observed. The log-risk function, denoted as $\hat{f}$, is estimated by the neural network. The indices i and j are patient indices. $R(T_i)$ represents the set of patients who are at risk of failure at time T_i. The parameter λ denotes the L2 regularization coefficient.

$$l(\theta) = -\frac{1}{N_{E=1}} \sum_{i;E_i=1} [\hat{f}(x_i, \theta) - log \sum_{j \in R(T_i)} (exp(\hat{f}(x_i, \theta)))] + \lambda ||\theta||_2^2 \tag{3.5}$$

The activation function used was the scaled exponential linear units (SELU), and the gradient descent algorithm employed was stochastic gradient descent

(SGD) with nesterov momentum (momentum factor: 0.9). To mitigate overfitting, several techniques were used, including early stopping with a patience of 10, L2 regularization with a coefficient of 10, dropout with a probability of 0.4, and batch normalization. Additionally, a time-based learning rate decay approach was employed reduce the learning rate every epoch. The model's performance was assessed using the Concordance Index (C-index) for the four specific survival events: Overall Survival (OS), Disease-Specific Survival (DSS), Disease-Free Interval (DFI), and Progression-Free Interval (PFI). The C-index measures the level of agreement between the predicted and actual survival, with a higher C-index indicating better model performance. A Python package lifelines, version 0.27.3 (Davidson-Pilon, 2019), was used to compute the C-index.

For each cohort, a dataset was created individually for each survival event. Samples not having available survival data (i.e., the survival time and the binary indicator for the survival event) were excluded. The remaining samples were then randomly divided into training, validation, and test datasets in a ratio of 6:2:2.

3.9 Survival analyses based on predicted risk

After predicting the risks, the significance of the estimated risk as a prognostic predictor was examined using both the log-rank test and Cox regression. For the log-rank test, patients of each cancer type were divided into risk-high and risk-low groups, thresholded by the median risk value. In the case of Cox regression, four covariates were used: age, gender, the average DNA methylation level of open sea CpG positions, and the predicted risk.

3.10 Functional analyses

Considering that the risks are predicted using 3D genome-aware epigenetic features, I surmised that the difference between the risk-high and risk-low groups arises from variations in DNA methylation levels at open sea CpG positions. Moreover, since these DNA methylation levels are embedded with the cancer-related 3D genome information, examining the DMRs between the risk-high and risk-low groups could interpret he black-box behavior of the deep learning model by offering biological explanation of the inter-group differences. Based on this rationale, functional annotation was performed on the DMRs defined by the predicted risks, following the procedures described in the subsequent subsections.

3.10.1 Functional annotation on DMR genes

DMRs were identified as genomic regions that exhibit significant hypomethylation in the risk-high group compared to the risk-low group. To gain insights into the biological implications of DMRs, functional annotation was conducted on all genes located within the DMRs using the python package GSEApy (Fang *et al.*, 2022). The functional annotation was based on the gene set 'GO Biological Process 2015' (Ashburner *et al.*, 2000).

3.10.2 Analysis on the chromatin states in DMR

Chromatin states, which provide epigenetic annotations for noncoding genomic regions, have been recognized to possess the 3D genome information (Ernst and Kellis, 2017; Rowley and Corces, 2018). To determine the impact of altered 3D genome structure on the chromatin states, the relative proportion of each chromatin state within the DMRs was analyzed, shedding light on the states that are most affected.

3.11 Data description

Table 3.2: **Description and composition of TCGA dataset.**

Cohort	Description	n_{tumor}	n_{normal}	n_{total}
BLCA	Bladder urothelial carcinoma	413	21	434
BRCA	Breast invasive carcinoma	790	98	888
CHOL	Cholangiocarcinoma	36	9	45
COAD	Colon adenocarcinoma	299	38	337
ESCA	Esophageal carcinoma	186	16	202
HNSC	Head and Neck squamous cell carcinoma	530	50	580
KIRC	Kidney renal clear cell carcinoma	320	160	480
KIRP	Kidney renal papillary cell carcinoma	276	45	321
LIHC	Liver hepatocellular carcinoma	379	50	429
LUAD	Lung adenocarcinoma	460	32	492
LUSC	Lung squamous cell carcinoma	372	43	415
PAAD	Pancreatic adenocarcinoma	185	10	195
PRAD	Prostate adenocarcinoma	499	50	549
THCA	Thyroid carcinoma	515	56	571
UCEC	Uterine corpus endometrial carcinoma	432	46	478

Table 3.3: Composition of stem cell samples.

Cohort	Description	$n_{samples}$
SC	Stem cell	44
EB	Embryoid body	22
DE	Definitive endoderm	11
ECTO	Ectoderm	11
MESO-5	Mesoderm, 5-days	11

Table 3.4: TCGA cohorts matched to the Hi-C data of normal cell lines (Hutter and Zenklusen, 2018; Kim *et al.*, 2021).

TCGA cohort	Hi-C normal cell line	GEO IDs of Hi-C data
PAAD	Pancreas	GSM2322547, GSM2322548, GSM2322549, GSM2322550
LUSC	Lung	GSM2322544, GSM2322545
LUAD	Lung	GSM2322544, GSM2322545

Table 3.5: PCBC stem cells matched to the Hi-C data of stem cell lines (Salomonis *et al.*, 2016; Kim *et al.*, 2021).

Cohort	Hi-C stem cell line	GEO ID of Hi-C data
PCBC	H9 Human Embryonic Stem Cells	GSM2309023
PCBC	Embryonic stem cell	GSM3263085

Table 3.6: TCGA cohorts matched to the Hi-C data of cancer cell lines (Hutter and Zenklusen, 2018; Kim *et al.*, 2021).

TCGA cohort	Hi-C cancer cell line	Hi-C cancer cell line description	GEO IDs of Hi-C data
BRCA	HCC1954 (Breast cancer cell line)	Breast Cancer	GSM3258551
COAD	SW480 (Colon cancer cell line)	Colon Adenocarcinoma	GSM3399745
ESCA	OE33 (Esophageal adenocarcinoma cell line)	Esophageal Adenocarcinoma	GSM3258552
KIRC	G-401 (kidney cancer cell line)	Kidney Renal Clear Cell Carcinoma	GSM2825105, GSM2825106
LIHC	HepG2 (hepatocellular carcinoma cell line)	Liver Hepatocellular carcinoma	GSM2825569, GSM2825570
LUSC	NCI-H460 (lung cancer cell line)	Lung Squamous Cell Carcinoma	GSM2827554, GSM2827555
PAAD	Panc1 (pancreatic carcinoma cell line)	Pancreatic Cancer	GSM2827313, GSM2827314
PRAD	22Rv1 (prostate cancer cell line)	Prostate Adenocarcinoma	GSM3358191, GSM3358192

Chapter 4

Results and Discussion

4.1 Significant characteristics of BDM PC1

4.1.1 BDM PC1s can approximate Hi-C PC1s

First, the validation process was conducted which aimed to determine whether the BDM PC1s could effectively reproduce the Hi-C PC1s. For this purpose, the 450K DNA methylation data from various TCGA cohorts and stem cells obtained from the Progenitor Cell Biology Consortium (PCBC) were utilized (Goldman *et al.*, 2020; Hutter and Zenklusen, 2018; Salomonis *et al.*, 2016). Additionally, the PC1s derived from the raw Hi-C matrices (Kim *et al.*, 2021; Schmitt *et al.*, 2016) were included in the analysis. Table 3.2 and Table 3.3 provide information on the composition of the TCGA and PCBC datasets, respectively. For detailed information on the processing of the raw Hi-C data, please refer to section 3.7.

The PC1 values were averaged from 10 randomly selected samples within each category (normal, tumor, and stem cells) due to the large number of available samples. For a fair comparison, the BDM PC1s and Hi-C PC1s from the same category and tissue type were paired. (Table 3.4-3.6) The performance of the BDM PC1s in reproducing the Hi-C PC1s was evaluated using PCC. The results demonstrated that the BDM PC1s were able to reproduce the Hi-C PC1s to a satisfactory extent, with a PCC of over 0.5 observed in most cases. Notably, the highest performance was observed when reproducing the Hi-C PC1s of cancer cells (Fig. 4.1).

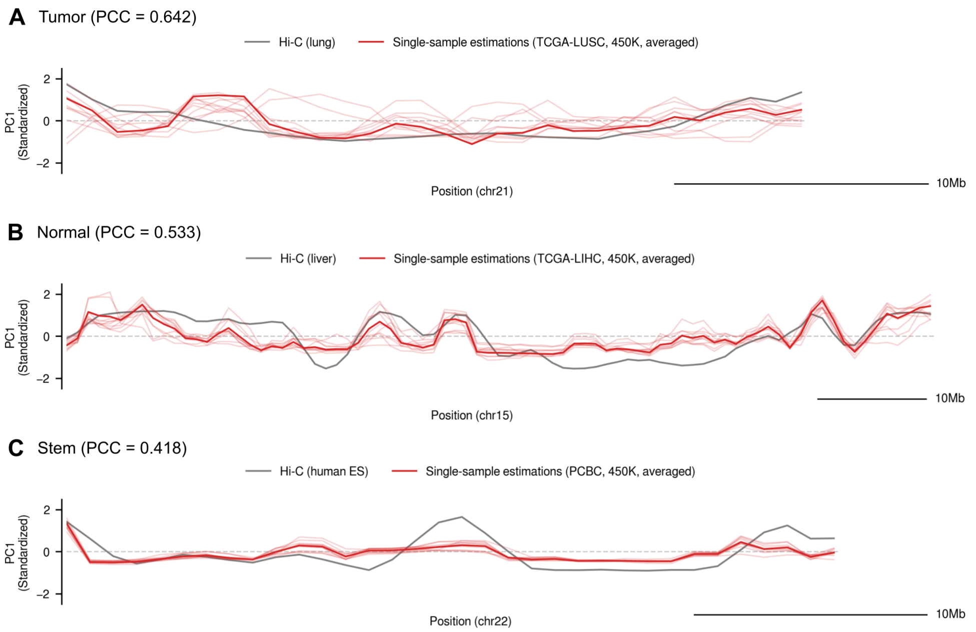


Figure 4.1: Reproduction of Hi-C PC1s from BDM PC1s.

Dark red graphs represent the averaged PC1s from 10 samples, and gray graphs display the Hi-C PC1s. (A) BDM PC1 from tumor samples (TCGA-LUSC, chr21) and Hi-C PC1 from lung squamous cell carcinoma. (B) BDM PC1 from normal samples (TCGA-LIHC, chr15) and Hi-C PC1 from normal lung cells. (C) BDM PC1 from stem cells (PCBC, chr22) and Hi-C PC1s from human embryonic stem cells.

Considering the single-sample nature of my approach, I also conducted a comparison between individual PC1 values. The results showed that the highest PCC among all categories increased to 0.750 (Fig. 4.2) compared to the previous case. This suggests that the 3D genome information captured by the Hi-C PC1 can be better reproduced by utilizing individual BDM PC1 rather than the averaged ones. It is hypothesized that using averaged BDM PC1 values may lower the performance because the well-reproduced individual BDM PC1 values can be diluted when averaged with PC1 values from other samples.

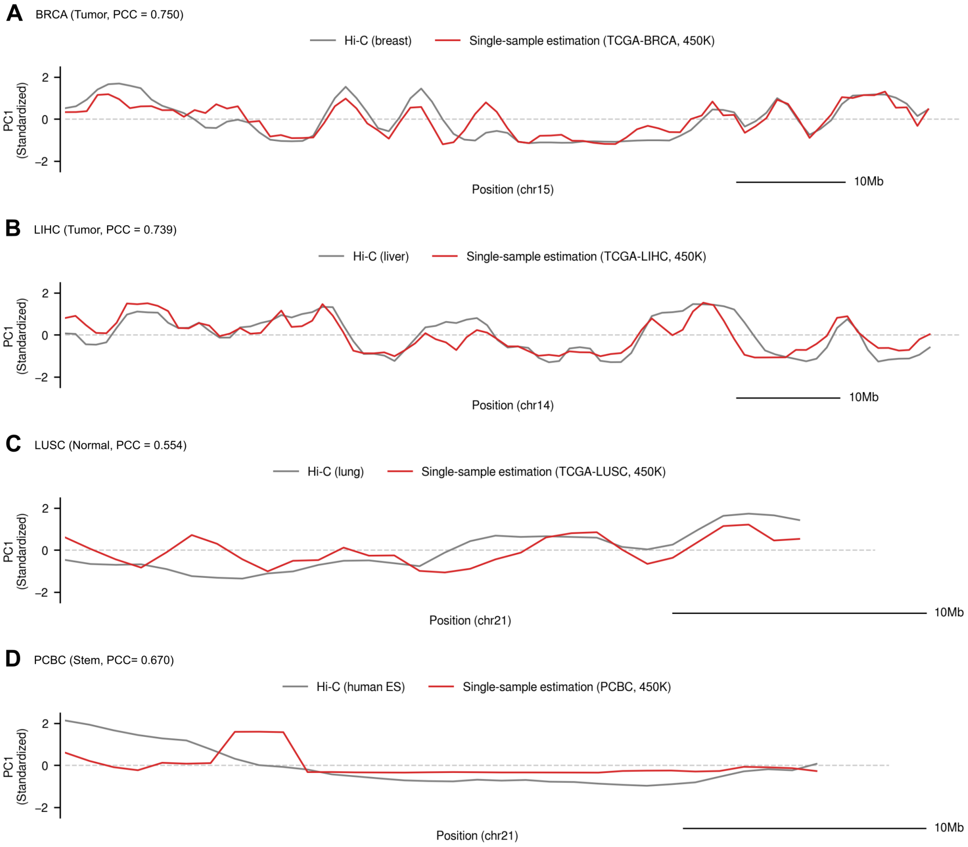


Figure 4.2: A comparison between the individual BDM PC1s and Hi-C PC1s.

Dark red graphs represent the individual PC1s, and gray graphs display the Hi-C PC1s. (A) BRCA tumor samples, breast cancer cells, chr15. (B) LIHC tumor samples, liver hepatocellular carcinoma cells, chr14. (C) LUSC normal samples, normal lung tissue, chr21. (D) PCBC stem cells, H9 human embryonic stem cells, chr21.

4.1.2 BDMs and BDM PC1s capture innate differences between tumor and normal groups

To determine if BDM includes cancer-related 3D genomic data that distinguishes between tumor and normal groups, I compared the BDM heatmaps of these two groups. My analysis revealed a distinct patchy pattern exclusively present in tumor groups across multiple cohorts (Fig. 4.3). Hence, I suggest that BDMs indeed contain information reflecting the inherent dissimilarities between tumor and normal groups. This trend was also observed in the BDM PC1s, as evidenced by the noticeably distinct shapes of the PC1 plots in the two groups across various cohorts (Fig. 4.4).

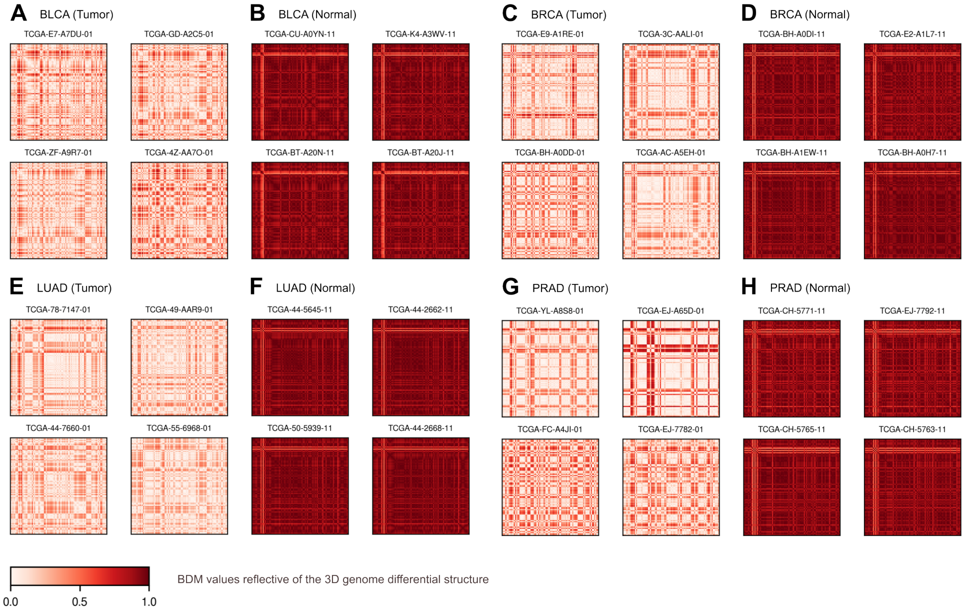


Figure 4.3: A heatmap representation of the binned difference matrices (BDMs) obtained from different TCGA cohorts.

(A) BLCA, tumor samples. (B) BLCA, normal samples. (C) BRCA, tumor samples. (D) BRCA, normal samples. (E) LUAD, tumor samples. (F) LUAD, normal samples. (G) PRAD, tumor samples. (H) PRAD, normal samples.

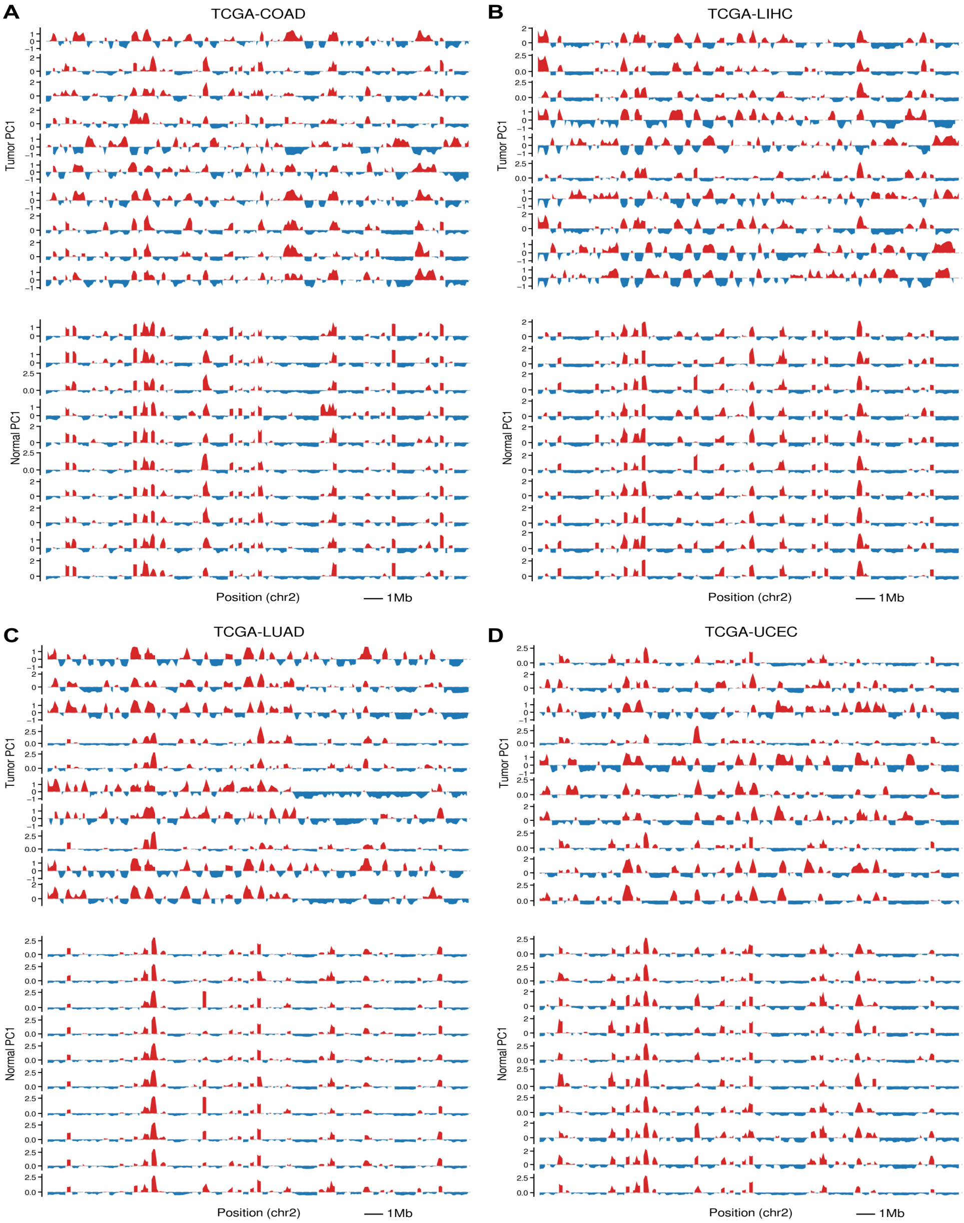


Figure 4.4: The BDM PC1s derived from tumor and normal samples across various TCGA cohorts.

The four cohorts with the prominent disparity between BDm PC1s of the tumor and normal groups are illustrated. (A) COAD, (B) LIHC, (C) LUAD, (D) UCEC.

4.1.3 BDM PC1s are tissue type-specific

Furthermore, to facilitate the pan-cancer clinical application of BDM PC1s, I conceived that these PC1s should exhibit tissue type-specific characteristics. To explore this, I organized the BDM PC1s into distinct pairs: homogeneous pairs comprising PC1s from the same cohort, and heterogeneous pairs containing PC1s from different cohorts. I then assessed the Pearson correlation coefficient (PCC) values for each pair. The analysis revealed that the PCC values for homogeneous pairs were higher compared to those for heterogeneous pairs (Fig. 4.5). Moreover, among the heterogeneous pairs, those consisting of PC1s from cohorts associated with the same tissue type (e.g., KIRP and KIRC) exhibited larger PCC values compared to other pairs. These findings suggest that BDM PC1s contain tissue type-specific information, in addition to capturing the differences between tumor and normal samples.

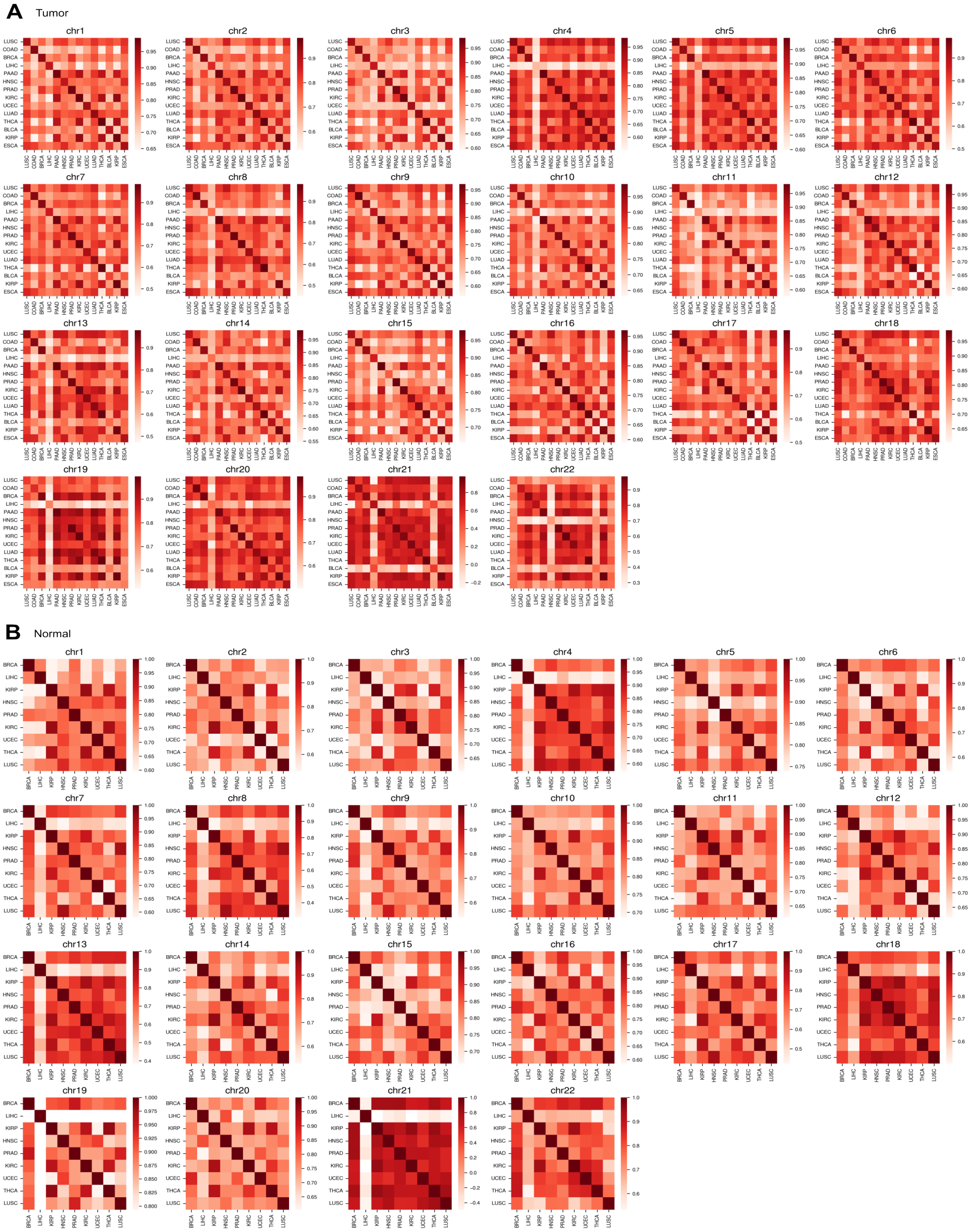


Figure 4.5: The Pearson correlation coefficient (PCC) between averaged BDM PC1s.

(A) Heatmaps displaying the PCC values derived from averaged BDM PC1s of tumor samples. (B) Heatmaps displaying the PCC values derived from averaged BDM PC1s of normal samples.

4.2 Utilizing 3D genome-aware epigenetic features helps survival prediction

Once the characteristics of BDM were examined, a one-dimensional vector was created, comprising the 3D genome-aware epigenetic features (normal/stem distances and references, stem closeness, and BDM PC1s) and the survival-related features (age and gender). This vector was used as an input feature for the deep learning model. To assess the significance of the 3D genome-aware epigenetic features, two baseline scenarios were also explored. The first scenario involved using no epigenetic feature, relying solely on age and gender. The second scenario involved using age, gender, and the 3D genome-unaware epigenetic feature (the average of open sea DNA methylation level) as input. The C-index was used as an evaluation metric.

After the risk prediction, patients from each cancer type were classified into either the risk-high or risk-low group, using the median risk value as the threshold. To determine whether the predicted risk had a significant impact on survival patterns, a log-rank test was performed. I considered the results to be statistically significant if both the validation and test C-index values exceeded 0.65, and if the log-rank test p-value was less than 0.05. As a result, significant findings were observed in seven cohorts, indicating that the predicted risk can serve as an important prognostic indicator (Table 4.1). The outcomes of the log-rank tests are illustrated in Fig. 4.6.

Following the log-rank test, Cox regression analysis was performed. The findings indicated that the predicted risks had greater significance in predicting survival compared to other covariates (Fig. 4.7). These results aligned with the outcomes from the log-rank test, highlighting the detrimental effect of high risk on survival.

Table 4.1: Risk prediction results from the feedforward neural network.

Results from the main scenario, baseline 1 (BS1), and baseline 2 (BS2) are displayed in the top, middle, and bottom section, respectively.

Cohort	Test set C-index	Log-rank p-value	Event
CHOL	0.750	0.029	DSS
KIRC	0.742	0.049	DSS
KIRC	0.705	0.045	OS
KIRP	**0.959**	0.024	DSS
KIRP	0.902	0.004	DFI
KIRP	0.841	**0.001**	OS
KIRP	0.839	0.040	PFI
PAAD	0.688	0.016	PFI
PRAD	0.739	0.026	DFI
THCA	0.742	0.028	OS
PRAD	0.669	0.042	DFI
PRAD	0.654	0.038	DSS
KIRC (BS1)	0.708	**0.001**	OS
KIRP (BS1)	0.741	0.028	DSS
PAAD (BS1)	0.810	0.026	DFI
PRAD (BS1)	0.676	0.011	DFI
THCA (BS1)	**0.987**	0.019	OS
KIRC (BS2)	0.697	**0.004**	OS
THCA (BS2)	**0.929**	0.024	OS
THCA (BS2)	0.884	0.019	OS

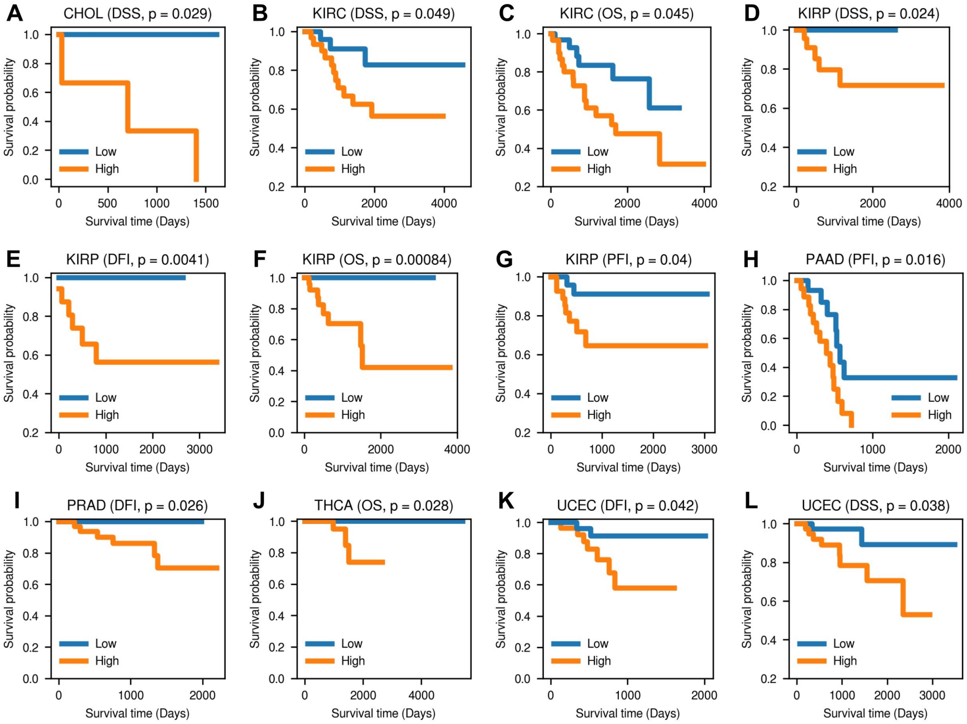


Figure 4.6: Results of log-rank tests based on the risks predicted by the feedforward neural network.

The name of the survival event and the corresponding log-rank test p-value are indicated within parentheses at the top center of each subplot. (A) CHOL (DSS), (B) KIRC (DSS), (C) KIRC (OS), (D) KIRP (DSS), (E) KIRP (DFI), (F) KIRP (OS), (G) KIRP (PFI), (H) PAAD (PFI), (I) PRAD (DFI), (J) THCA (OS), (K) UCEC (DFI), (L) UCEC (DSS).

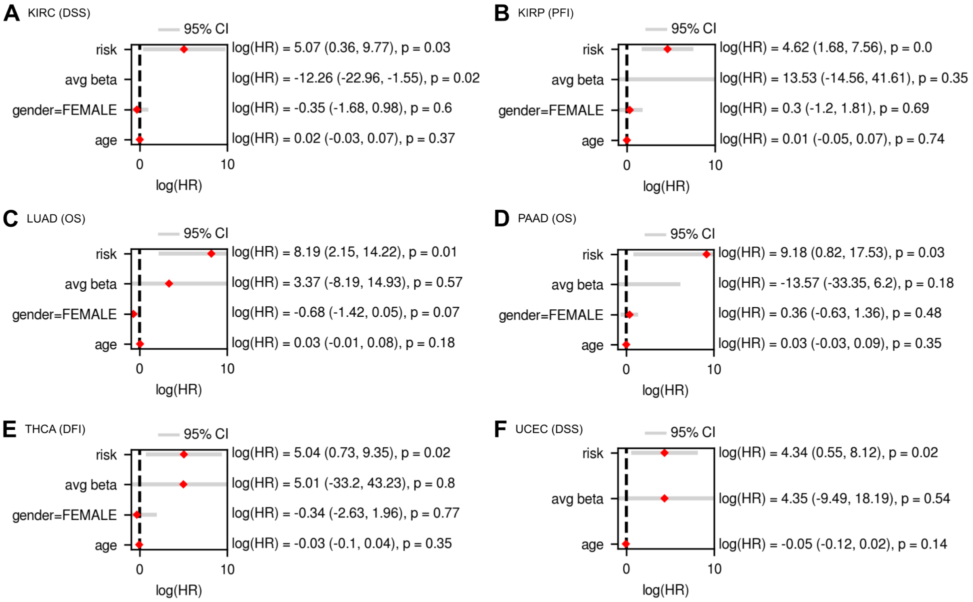


Figure 4.7: The results of Cox regression analysis.

On the right side of each subplot, the logarithm of the hazard ratio for each input covariate is provided, along with the corresponding 95% confidence interval in parentheses, and the p-value. The subplots correspond to different cancer types and survival events. (A) KIRC (DSS), (B) KIRP (PFI), (C) LUAD (OS), (D) PAAD (OS), (E) THCA (DFI), (F) UCEC (DSS).

The outcomes of two baseline cases revealed a decrease in the performance of survival prediction when the 3D genome-aware epigenetic features were excluded from the input feature (Table 4.1; BS1 and BS2). Furthermore, both baseline scenarios exhibited poor performance in predicting events other than OS. Since age is a highly influential predictor of OS, the absence of 3D genome-reflective features may have resulted in the model placing greater emphasis on age, leading to satisfactory performance solely in risk prediction of OS.

4.2.1 The model shows robust performance on external datasets

To validate the model's robust performance on a dataset from different platform, the model was evaluated using the GSE103659 dataset (Edgar *et al.*, 2002). Since GSE103659 comprised patients with glioblastoma (GBM), a comparison was made between the model's performance on this dataset and its performance on TCGA-GBM. The results indicated that the predicted risks served as significant predictors of survival for both datasets (Table 4.2 and Fig. 4.8). It should be noted that GSE103659 had certain limitations compared to TCGA-GBM, such as the absence of gender information and a smaller number of survival events with available data. Considering these limitations, the findings suggest that although using TCGA-GBM yielded more significant results compared to using GSE103659 (Table 4.2), the performance gap to that extent is deemed acceptable. Consistently, the results of Cox regression analysis also confirmed that the predicted risk is a significant predictor of survival in both TCGA-GBM and GSE103659 datasets (Fig. 4.9).

Table 4.2: Risk prediction performance from TCGA-GBM and GSE103659 datasets.

The best performance, indicated by bold text, is determined based on the highest c-index and the lowest p-value.

Cohort	Test set C-index	Log-rank p-value	Event
GBM	**0.743**	0.008	DSS
GBM	0.733	**0.001**	OS
GSE103659	0.721	**0.001**	OS

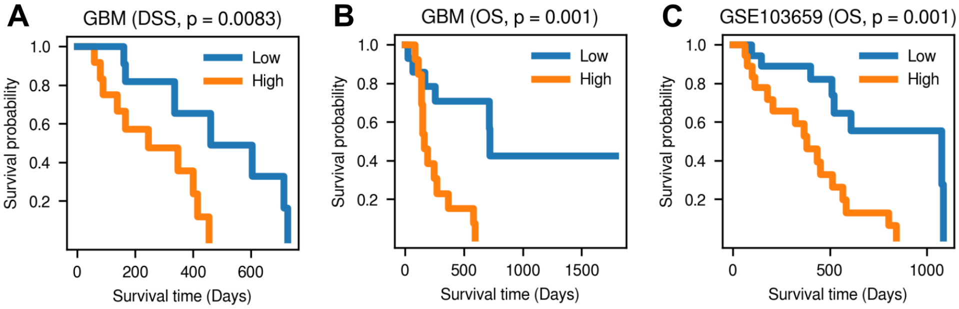


Figure 4.8: Significant log-rank test results obtained from TCGA-GBM and GSE103659 datasets, utilizing the risk predicted from 3D genome-aware epigenetic features.

The top center of each subplot displays the name of the cohort and survival event, followed by the corresponding log-rank test p-value. (A) GBM (DSS), (B) GBM (OS), (C) GSE103659 (OS).

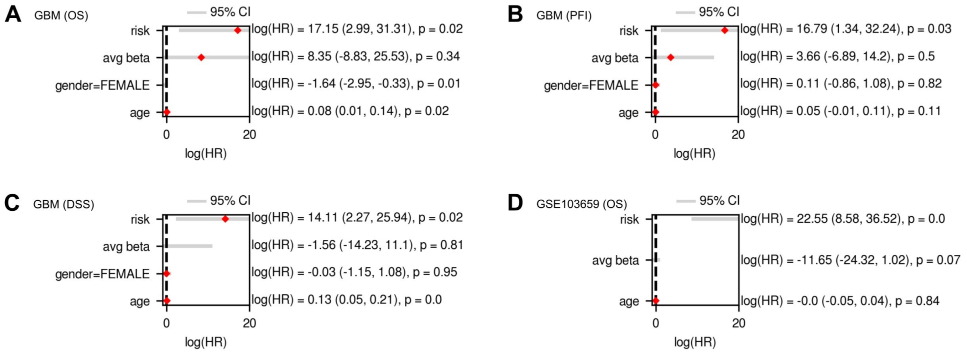


Figure 4.9: Cox regression results from TCGA-GBM and GSE103659 datasets.

On the right side of each subplot, the logarithm of hazard ratio is presented for each input covariate, along with the corresponding 95% confidence interval in parentheses and the p-value. (A) GBM (OS), (B) GBM (PFI), (C) GBM (DSS), (D) GSE103659 (OS).

4.3 Functional annotation on genes in DMR

The functional annotations on genes in DMR, defined by the predicted risks, revealed a significant presence of genes related to the RA signaling pathway. It is widely known that RA binds to its receptor and induces structural changes into euchromatin, thereby promoting the transcription of target genes and playing a crucial role during the developmental process (Ozgun *et al.*, 2021; Ablain and de Thé, 2014). Moreover, RA has cell-type specific effects on cell fate decisions, such as differentiation, apoptosis, or stemness (Mezquita and Mezquita, 2019). Hence, it is plausible that the open sea CpG positions have epigenetic control over the key regulators of development and cell fate, influencing the stemness of cells, and this information is captured by the 3D genome-aware features. Another interesting finding is the presence of genes associated with gas transport in DMRs. This observation could be linked to hypoxia, a condition characterized by low oxygen levels that often occurs in cancer (Eales *et al.*, 2016; Bhandari *et al.*, 2019). Notably, all the DMR genes associated with gas transport were found to encode subunits of hemoglobin, which is responsible for oxygen transport. One of these genes, *HBB*, has been reported to exhibit abnormal expression in various types of cancer (Zheng *et al.*, 2017; Kang *et al.*, 2022). This suggests that genes frequently altered in cancer, even if not directly involved in developmental processes, can be regulated by CpG probes located in open sea positions. Furthermore, enrichment analysis revealed terms related to cell adhesion, such as homophilic cell adhesion via plasma membrane adhesion molecules and cell-cell adhesion via plasma membrane adhesion molecules. Cell adhesion plays a crucial role in cancer progression, as abnormalities in cell adhesion molecules enable tumor cells to better interact with other cells. Increased interactions between cancer cells and endothelium, for example, promote faster metastasis and worsen prognosis

(Läubli and Borsig, 2019; Bendas *et al.*, 2012). Another enriched GO term was related to the mitotic cell cycle, which is closely associated with the biological characteristics of cancer cells. Cancer arises from defects in the cell cycle, with aberrations occurring in cell cycle checkpoints or genes regulating the cell cycle, such as *p53* and *BRCA1* genes. Consequently, cells undergo uncontrolled growth (Williams and Stoeber, 2012; Visconti *et al.*, 2016; Zhang *et al.*, 2020; Oh *et al.*, 2018), leading to cancer development. Lastly, a significantly enriched GO term was related to the gamma-aminobutyric acid (GABA) signaling pathway. GABA is involved in the development of various cell types and acts as an important modulator across different cancer types. Elevated GABA levels significantly enhance the invasive capacity of cancer cells, indicating its contributory role in metastasis. GABA receptors, along with GABA itself, also regulate cell proliferation. Additionally, the gene expression of GABA receptors has been linked to cancer prognosis and tumorigenesis (Zhang *et al.*, 2013; Li *et al.*, 2012; Kanbara *et al.*, 2018; Azuma *et al.*, 2003). Overall, these results demonstrate the enrichment of validated cancer-related pathways in DMR genes. The significant differences in open sea DNA methylation levels of DMR genes between high-risk and low-risk groups suggest the involvement of altered DNA methylation levels of open sea CpG probes in multiple cancer hallmarks. Fig. 4.10 and Table 4.3 provide a comprehensive presentation of the results.

4.4 Inactive chromatin states dominate in DMRs

The chromatin state data (Ernst and Kellis, 2012; Kundaje *et al.*, 2015) was used to examine the distribution of different chromatin states within the DMRs, to identify which states are most affected by the cancer-related 3D genome perturbations. Among all the cohorts listed in Table 4.1, TCGA-PAAD, which had available chromatin state data, was utilized for this analysis.

Table 4.3: Significantly enriched terms found from DMR genes.

Gene ontology (GO) term	Cohorts
Regulation of retinoic acid (RA) receptor signaling pathway	KIRP, PRAD, THCA
Negative regulation of retinoic acid receptor signaling pathway	KIRP, PRAD, THCA
Cell-cell adhesion	KIRC, UCEC
Cell-cell adhesion via plasma-membrane adhesion molecules	KIRC, UCEC
Gas transport	KIRP, UCEC
Homophilic cell adhesion via plasma membrane adhesion molecules	KIRC, UCEC
Calcium-dependent cell-cell adhesion via plasma membrane cell adhesion molecules	KIRC, UCEC
Mitotic cell cycle	KIRC
Gamma-aminobutyric acid (GABA) signaling pathway	UCEC

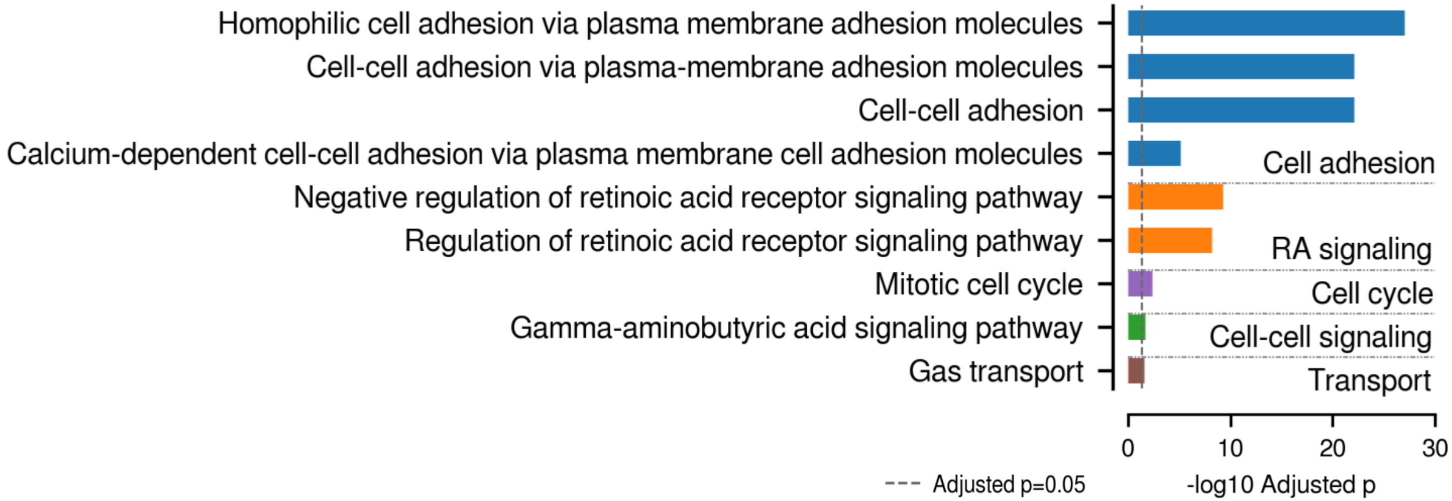


Figure 4.10: $-\log_{10}$**(adjusted p-value) of each Gene Ontology (GO) term enriched by the functional annotation using DMR genes.**

The GO terms are categorized and separated by a horizontal gray line. The black vertical line represents $-\log_{10} 0.05$.

The results revealed that the Quiescent/Low state accounted for the largest proportion of the DMRs, followed by Weak transcription and Heterochromatin states (Fig. 4.11). Both Quiescent/Low and Heterochromatin are inactive states in normal cells, and Weak transcription is a mildly activated state. Therefore, it is postulated that the abnormal hypomethylation of open sea CpG positions, occurring alongside cancer progression, could exert aberrant influences on the DMR genes.

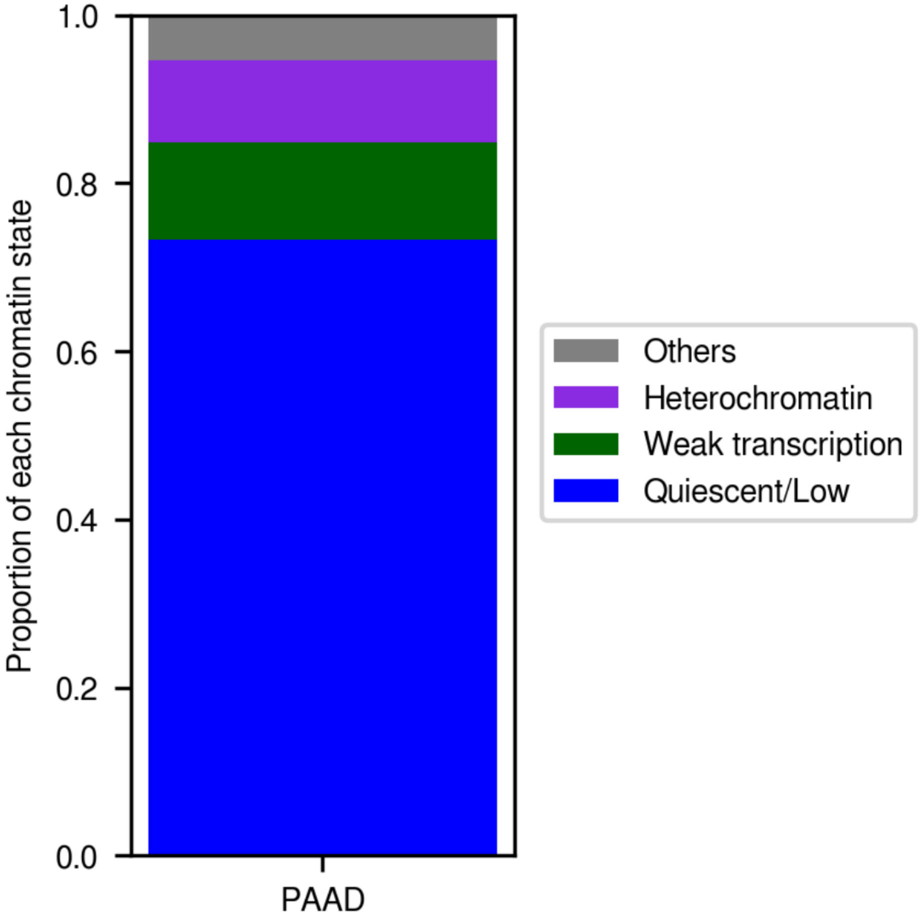


Figure 4.11: The proportion of the chromatin states in DMR.

The x and y axis represent the TCGA cohort and the proportions of different chromatin states, respectively. Excluding the three most dominant states (Quiescent/Low, Heterochromatin, and Weak transcription), all other states are labeled as 'Others'.

4.5 Limitation

The findings of this study have provided valuable insights into the usage of 3D genome-informed epigenetic features for survival prediction. However, certain limitations persist. Firstly, the performance of survival prediction diminishes when narrowing down the scope of survival analysis from cancer as a whole to specific cancer subtypes, as evidenced in section 4.2.2. Therefore, it is imperative to develop more sophisticated approaches to effectively apply this method at the level of cancer subtypes.

In addition to the informative nature of 3D genome-informed epigenetic characteristics, the small number of patients could have contributed to the significant results observed in the log-rank tests (Fig. 4.6). For example, upon examining the dataset used for the log-rank test results in Fig. 4.6A, it was found that there were only four patients in the risk-low group and three patients in the risk-high group. This was incurred by the experimental setup, where patients from each cohort were randomly divided into five folds, and only one fold was utilized as a test set for conducting the log-rank test. Consequently, it is necessary to address this limitation by either increasing the number of patients in the test set by reducing the number of folds or acquiring additional data.

Chapter 5

Conclusion

While the close relationship between the 3D genome structure and the development of cancer has been observed (Rheinbay *et al.*, 2020), a prognostic metric that makes use of the 3D genome information has not yet been developed. This is primarily due to the high cost of generating Hi-C data, which is a manifest source of the 3D genome information, resulting in the limited availability of Hi-C data (Yardımcı *et al.*, 2019). Inspired by the recent discoveries regarding the potential of DNA methylation data to reconstruct the 3D genome information (Fortin and Hansen, 2015), the 3D genome-aware epigenetic features were extracted from 450K DNA methylation data. These features were then used to predict the risk of failure for different survival events by a feedforward neural network. The predicted risk was found to be a significant predictor of survival across various cancer types. An important finding was that excluding the 3D genome-aware features from the input data led to the decreased performance of the model. This suggests that utilizing the 3D genome-aware features facilitates a knowledge-guided risk prediction, resulting in more precise prognostic predictions for cancer. Furthermore, the functional analyses revealed that genes in DMR, defined by the predicted risk values, are involved in a variety of cancer-related pathways, including cell adhesion, the RA signaling pathway, and the mitotic cell cycle. Additionally, a comprehensive analysis of the chromatin states within the DMRs indicated a dominance of inactive or mildly activated states in DMRs. Based on these findings, it is posited that the alterations in DNA methylation levels in the risk-high group are associated with disrupted cancer-related pathways and the abnormal activation of genes. After careful consideration, I suggest that the 3D genome landscape derived from the 450K DNA methylation data, which potentially reflects the aberrantly activated cancer-related genes and pathways, facilitates a more accurate prediction of cancer prognosis.

Bibliography

Abdennur, N. and Mirny, L. A. (2020). Cooler: scalable storage for hi-c data and other genomically labeled arrays. *Bioinformatics*, **36**(1), 311–316.

Ablain, J. and de Thé, H. (2014). Retinoic acid signaling in cancer: the parable of acute promyelocytic leukemia. *International Journal of Cancer*, **135**(10), 2262–2272.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., *et al.* (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, **25**(1), 25–29.

Azuma, H., Inamoto, T., Sakamoto, T., Kiyama, S., Ubai, T., Shinohara, Y., Maemura, K., Tsuji, M., Segawa, N., Masuda, H., *et al.* (2003). γ-aminobutyric acid as a promoting factor of cancer metastasis; induction of matrix metalloproteinase production is potentially its underlying mechanism. *Cancer research*, **63**(23), 8090–8096.

Bendas, G., Borsig, L., *et al.* (2012). Cancer cell adhesion and metastasis: selectins, integrins, and the inhibitory potential of heparins. *International journal of cell biology*, **2012**.

Bhandari, V., Hoey, C., Liu, L. Y., Lalonde, E., Ray, J., Livingstone, J., Lesurf, R., Shiah, Y.-J., Vujcic, T., Huang, X., *et al.* (2019). Molecular landmarks of tumor hypoxia across cancer types. *Nature genetics*, **51**(2), 308–318.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., Delano, D., Zhang, L., Schroth, G. P., Gunderson, K. L., *et al.* (2011). High density dna methylation array with single cpg site resolution. *Genomics*, **98**(4), 288–295.

Davidson-Pilon, C. (2019). lifelines: survival analysis in python. *Journal of Open Source Software*, **4**(40), 1317.

Di Stefano, M. and Cavalli, G. (2022). Integrative studies of 3d genome organization and chromatin structure. *Current Opinion in Structural Biology*, **77**, 102493.

Du, Q., Smith, G. C., Luu, P. L., Ferguson, J. M., Armstrong, N. J., Caldon, C. E., Campbell, E. M., Nair, S. S., Zotenko, E., Gould, C. M., *et al.* (2021). Dna methylation is required to maintain both dna replication timing precision and 3d genome organization integrity. *Cell Reports*, **36**(12), 109722.

Dubois, F., Sidiropoulos, N., Weischenfeldt, J., and Beroukhim, R. (2022). Structural variations in cancer and the 3d genome. *Nature Reviews Cancer*, **22**(9), 533–546.

Eales, K. L., Hollinshead, K. E., and Tennant, D. A. (2016). Hypoxia and metabolic adaptation of cancer cells. *Oncogenesis*, **5**(1), e190–e190.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, **30**(1), 207–210.

Ernst, J. and Kellis, M. (2012). Chromhmm: automating chromatin-state discovery and characterization. *Nature methods*, **9**(3), 215–216.

Ernst, J. and Kellis, M. (2017). Chromatin-state discovery and genome annotation with chromhmm. *Nature protocols*, **12**(12), 2478–2492.

Fang, Z., Liu, X., and Peltz, G. (2022). Gseapy: a comprehensive package for performing gene set enrichment analysis in python. *Bioinformatics*, **39**(1), btac757.

Fortin, J.-P. and Hansen, K. D. (2015). Reconstructing a/b compartments as revealed by hi-c using long-range correlations in epigenetic data. *Genome biology*, **16**(1), 1–23.

Freire-Pritchett, P., Schoenfelder, S., Várnai, C., Wingett, S. W., Cairns, J., Collier, A. J., García-Vílchez, R., Furlan-Magaril, M., Osborne, C. S., Fraser, P., *et al.* (2017). Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem cells. *Elife*, **6**, e21926.

Goldman, M. J., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., Banerjee, A., Luo, Y., Rogers, D., Brooks, A. N., *et al.* (2020). Visualizing and interpreting cancer genomics data via the xena platform. *Nature biotechnology*, **38**(6), 675–678.

Golloshi, R., Playter, C., Freeman, T. F., Das, P., Raines, T. I., Garretson, J. H., Thurston, D., and McCord, R. P. (2022). Constricted migration is associated with stable 3d genome structure differences in cancer cells. *EMBO reports*, **23**(10), e52149.

Gröschel, S., Sanders, M. A., Hoogenboezem, R., de Wit, E., Bouwman, B. A., Erpelinck, C., van der Velden, V. H., Havermans, M., Avellino, R., van

Lom, K., *et al.* (2014). A single oncogenic enhancer rearrangement causes concomitant evi1 and gata2 deregulation in leukemia. *Cell*, **157**(2), 369–381.

Heidari, N., Phanstiel, D. H., He, C., Grubert, F., Jahanbani, F., Kasowski, M., Zhang, M. Q., and Snyder, M. P. (2014). Genome-wide map of regulatory interactions in the human genome. *Genome research*, **24**(12), 1905–1917.

Hutter, C. and Zenklusen, J. C. (2018). The cancer genome atlas: creating lasting value beyond its data. *Cell*, **173**(2), 283–285.

Kanbara, K., Otsuki, Y., Watanabe, M., Yokoe, S., Mori, Y., Asahi, M., and Neo, M. (2018). Gaba b receptor regulates proliferation in the high-grade chondrosarcoma cell line oums-27 via apoptotic pathways. *BMC cancer*, **18**, 1–13.

Kang, N., Qiu, W.-J., Wang, B., Tang, D.-f., and Shen, X.-Y. (2022). Role of hemoglobin alpha and hemoglobin beta in non-small-cell lung cancer based on bioinformatics analysis. *Molecular Carcinogenesis*, **61**(6), 587–602.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. (2018). Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, **18**(1), 1–12.

Kim, K., Jang, I., Kim, M., Choi, J., Kim, M.-S., Lee, B., and Jung, I. (2021). 3div update for 2021: a comprehensive resource of 3d genome and 3d cancer genome. *Nucleic Acids Research*, **49**(D1), D38–D46.

Kruse, K., Hug, C. B., and Vaquerizas, J. M. (2020). Fan-c: a feature-rich framework for the analysis and visualisation of chromosome conformation capture data. *Genome biology*, **21**(1), 1–19.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., *et al.* (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, **518**(7539), 317–330.

Läubli, H. and Borsig, L. (2019). Altered cell adhesion and glycosylation promote cancer immune suppression and metastasis. *Frontiers in immunology*, **10**, 2120.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*, **25**(14), 1754–1760.

Li, Y.-H., Liu, Y., Li, Y.-D., Liu, Y.-H., Li, F., Ju, Q., Xie, P.-L., and Li, G.-C. (2012). Gaba stimulates human hepatocellular carcinoma growth through overexpressed gabaa receptor theta subunit. *World Journal of Gastroenterology: WJG*, **18**(21), 2704.

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., *et al.* (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, **326**(5950), 289–293.

Liu, Y., Nanni, L., Sungalee, S., Zufferey, M., Tavernari, D., Mina, M., Ceri, S., Oricchio, E., and Ciriello, G. (2021). Systematic inference and comparison of multi-scale chromatin sub-compartments connects spatial organization to cell phenotypes. *Nature communications*, **12**(1), 2439.

Magaña-Acosta, M. and Valadez-Graham, V. (2020). Chromatin remodelers in the 3d nuclear compartment. *Frontiers in Genetics*, **11**, 600615.

Mezquita, B. and Mezquita, C. (2019). Two opposing faces of retinoic acid:

induction of stemness or induction of differentiation depending on cell-type. *Biomolecules*, **9**(10), 567.

Oh, M., McBride, A., Yun, S., Bhattacharjee, S., Slack, M., Martin, J. R., Jeter, J., and Abraham, I. (2018). Brca1 and brca2 gene mutations and colorectal cancer risk: systematic review and meta-analysis. *JNCI: Journal of the National Cancer Institute*, **110**(11), 1178–1189.

Ozgun, G., Senturk, S., and Erkek-Ozhan, S. (2021). Retinoic acid signaling and bladder cancer: Epigenetic deregulation, therapy and beyond. *International Journal of Cancer*, **148**(10), 2364–2374.

Reiff, S. B., Schroeder, A. J., Kırlı, K., Cosolo, A., Bakker, C., Lee, S., Veit, A. D., Balashov, A. K., Vitzthum, C., Ronchetti, W., *et al.* (2022). The 4d nucleome data portal as a resource for searching and visualizing curated nucleomics data. *Nature communications*, **13**(1), 1–11.

Rheinbay, E., Nielsen, M. M., Abascal, F., Wala, J. A., Shapira, O., Tiao, G., Hornshøj, H., Hess, J. M., Juul, R. I., Lin, Z., *et al.* (2020). Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature*, **578**(7793), 102–111.

Rowley, M. J. and Corces, V. G. (2018). Organizational principles of 3d genome architecture. *Nature Reviews Genetics*, **19**(12), 789–800.

Salomonis, N., Dexheimer, P. J., Omberg, L., Schroll, R., Bush, S., Huo, J., Schriml, L., Sui, S. H., Keddache, M., Mayhew, C., *et al.* (2016). Integrated genomic analysis of diverse induced pluripotent stem cells from the progenitor cell biology consortium. *Stem cell reports*, **7**(1), 110–125.

Schmitt, A. D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C. L., Li, Y., Lin, S., Lin, Y., Barr, C. L., *et al.* (2016). A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell reports*, **17**(8), 2042–2059.

Song, F., Xu, J., Dixon, J., and Yue, F. (2022). Analysis of hi-c data for discovery of structural variations in cancer. In *Hi-C Data Analysis*, pages 143–161. Springer.

Stadhouders, R., Filion, G. J., and Graf, T. (2019). Transcription factors and 3d genome conformation in cell-fate decisions. *Nature*, **569**(7756), 345–354.

Umlauf, D. and Mourad, R. (2019). The 3d genome: From fundamental principles to disease and cancer. In *Seminars in Cell & Developmental Biology*, volume 90, pages 128–137. Elsevier.

Van Berkum, N. L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L. A., Dekker, J., and Lander, E. S. (2010). Hi-c: a method to study the three-dimensional architecture of genomes. *JoVE (Journal of Visualized Experiments)*, (39), e1869.

Visconti, R., Della Monica, R., and Grieco, D. (2016). Cell cycle checkpoint in cancer: a therapeutically targetable double-edged sword. *Journal of Experimental & Clinical Cancer Research*, **35**(1), 1–8.

Wang, S., Su, J.-H., Beliveau, B. J., Bintu, B., Moffitt, J. R., Wu, C.-t., and Zhuang, X. (2016). Spatial organization of chromatin domains and compartments in single chromosomes. *Science*, **353**(6299), 598–602.

Williams, G. H. and Stoeber, K. (2012). The cell cycle and cancer. *The Journal of pathology*, **226**(2), 352–364.

Yardımcı, G. G., Ozadam, H., Sauria, M. E., Ursu, O., Yan, K.-K., Yang, T., Chakraborty, A., Kaul, A., Lajoie, B. R., Song, F., *et al.* (2019). Measuring the reproducibility and quality of hi-c data. *Genome biology*, **20**(1), 1–19.

Zhang, C., Liu, J., Xu, D., Zhang, T., Hu, W., and Feng, Z. (2020). Gain-of-function mutant p53 in cancer progression and therapy. *Journal of molecular cell biology*, **12**(9), 674–687.

Zhang, S., Chasman, D., Knaack, S., and Roy, S. (2019a). In silico prediction of high-resolution hi-c interaction matrices. *Nature communications*, **10**(1), 1–18.

Zhang, X., Zhang, R., Zheng, Y., Shen, J., Xiao, D., Li, J., Shi, X., Huang, L., Tang, H., Liu, J., *et al.* (2013). Expression of gamma-aminobutyric acid receptors on neoplastic growth and prediction of prognosis in non-small cell lung cancer. *Journal of translational medicine*, **11**(1), 1–10.

Zhang, Y., Li, T., Preissl, S., Amaral, M. L., Grinstein, J. D., Farah, E. N., Destici, E., Qiu, Y., Hu, R., Lee, A. Y., *et al.* (2019b). Transcriptionally active herv-h retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nature genetics*, **51**(9), 1380–1388.

Zheng, Y., Miyamoto, D. T., Wittner, B. S., Sullivan, J. P., Aceto, N., Jordan, N. V., Yu, M., Karabacak, N. M., Comaills, V., Morris, R., *et al.* (2017). Expression of β-globin by cancer cells promotes cell survival during blood-borne dissemination. *Nature communications*, **8**(1), 1–12.

국문초록

암의 발생은 3차원 유전체 구조와 밀접하게 관련 있다. 하지만, 3차원 유전체 구조에 대한 정보는 지금까지 임상적으로 활용되고 있지 않다. 이에 대한 주요한 이유는 3차원 유전체 정보를 가장 직관적으로 제공하는 Hi-C (High-throughput Chromosome Conformation Capture; 고 처리량 염색체 형태 캡처) 데이터의 생산 비용이 매우 높기 때문이다. 따라서, 3차원 유전체 정보를 사용한 새로운 임상적인 척도를 개발한다면, 해당 정보의 임상적 활용 가능성을 높일 수 있다.

본 연구에서는 DNA 메틸화 데이터로부터 3차원 유전체 정보가 내재되어 있는 후성유전적 특징 벡터들을 추출하고, 이를 딥 러닝 기반 생존분석에 활용하는 새로운 방법을 제시한다. 3차원 유전체 정보가 내재되어 있는 후성유전적 특징 벡터들을 추출하기 위해, 개개인의 450K DNA 메틸화 데이터로부터 재구축한 3차원 유전체 구조를 활용한다. 실험 결과, 해당 특징 벡터들을 활용한 경우들이 그렇지 않은 경우들에 비해 다양한 암종에서 생존 예측의 정확도가 더 높았다. 이는 후성유전적 특징 벡터들에 내재되어 있는 3차원 구조에 대한 정보가 암 환자들의 생존 및 예후 예측에 있어서 중요한 예측인자로 작용할 수 있음을 시사한다. 또한 생물학적 분석을 통해, 딥 러닝 모델에 의해 고위험군으로 분류된 환자들에게서 관찰된 DNA 메틸화 수준의 변화가 다양한 암 관련된 패스웨이들의 비정상적인 활성화와 관련 있음이 밝혀졌다. 이를 통해 3차원 정보가 내재되어 있는 후성유전적 특징 벡터들이 임상적으로 중요할 뿐만 아니라 생물학적으로도 의미가 있음을 알 수 있다. 실험에 사용된 코드는 https://github.com/jwyang21/3D-genome-risk-prediction 에서 확인 가능하다.

주요어: 딥 러닝, 생물정보학, DNA 메틸화, 암 예후 예측, 3차원 유전체, 후성유전학, 생존 분석

학번: 2021-26775