Finding Unique Patterns in Transcriptome and Epigenome of Cancer Cells

By-

Arpit Mathur

MT20328

Under the supervision of

Dr. Vibhor Kumar

Indraprastha Institute of Information Technology Delhi

Dec 2022

©Indraprastha Institute of Information Technology (IIITD), New Delhi 2022

Finding Unique Pattern in Transcriptome and Epigenome of Cancer Cells

By-

Arpit Mathur

MT20328

Submitted

in partial fulfillment of the requirements for the degree of Master of Technology

To

Indraprastha Institute of Information Technology Delhi Dec 2022

Certificate

This is to certify that the thesis titled "Finding Unique Pattern in Transcriptome and Epigenome of Cancer Cells" being submitted by Arpit Mathur to the Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology, is an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

Dr. Vibhor Kumar

Dec 2022

Associate Professor

Department of Computational Biology

Indraprastha Institute of Information Technology Delhi, New Delhi 110 020

Acknowledgment

Known words become inadequate to express my gratitude for my mentor **Dr. Vibhor Kumar**, Associate Professor in the Department of Computational Biology IIITD, who initiated me into the realm of research and supervised me with finite patience. Without whose invaluable suggestion and unstinted cooperation, the present work would not have been possible. He made sure my scientific enthusiasm remained blooming, and I thank him for making me part of his projects, lab, and scientific work.

From being a below-average student with lots of backlogs during my Undergraduate years to a transformed student who can now think, visualize and apply tools to understand mysteries of science, I will always be in debt to Indraprastha Institute of Technology Delhi. IIITD gave me wings through excellent research-oriented teaching of faculty members. They liberated me every second they taught or had a conversation with me. Research culture of IIITD is beyond comprehension. IIITD research facility, in terms of infrastructure and technical capabilities, is beyond words to appreciate. I hope I will be able to reflect back and, in the future, give away and contribute to the institute in some capacity.

I want to thank and mention Dr. Neetesh Pandey (PhD) especially**.** He mentored me from the start and was always ready to hear my queries and worries. Special thanks go to all members of 'The ***RegGen Lab***' of the Computational Biology Department at IIITD. My thesis work results from collaboration and teamwork involving healthy discussions and regular feedback. We learned from each other and improved. I would also like to thank the Technical Support facility and IT help desk IIITD for ensuring my work does not halt.

A Tree without strong roots will fall someday. My father, **Dr (Prof) Navin Mathur,** made sure that his support and guidance had no hiccups so my roots remain strong and vibrant. Without his frequent positive interventions, I would not have been able to pursue and complete the degree at IIITD. Thanks, Dad.

ABSTRACT

Background: Studying Chromatin Architecture is paramount to an understanding of cells or nuclei in disease and normal state. With recent advances in genomic technologies and computational power, new domains like Topologically associated domains (TAD) have been discovered. Studying TAD in the context of cancer cells gives insights into how chromatin folding relates to the survival of the patient. Exploiting chromatin interactions from the lens of enhancer-gene interactions is of cardinal value since identifying specific chromatin interactions (enhancer-gene pairs) in disease state cells which are etiology pairs for the disease, and using genomic editing technologies to knockdown these pairs, could be a potential precise and accurate model to beat disease cells, especially cancer cells. Our study is divided into two parts; in the first part, we build a method to understand TAD biology and its implication in estimating patient survival. In the second part of our study, we modified a previously proposed method scEChiA to detect enhancer-gene pairs interactions in cancer-specific cells using RNA-seq profiles. We further validates the predicted interactions with 4D genome[2] and Activity by Contact (ABC) databases

Results: We identified TAD ***chr1_171750000_172350000*** in Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC) cancer type, which according to our curated algorithms and pipelines is found to be the most survival TAD with high survival score. We explored this TAD biology of how genes in this TAD interplay with each other creating a network that ultimately defines this TAD property. From the scEChiA R package, ***we identified several enhancer-***

enhancer, enhancer-gene, and gene-gene interaction pairs in chromosome 11 of Diffuse large B cell lymphoma (DLBCL) cancer type, which was benchmarked with 4D genome and Activity by contact chromatin interaction databases.

Conclusion: Identification of specific TAD, which is most surviving in cancer cell lines, and understanding its underlying biology gives a new definition of TAD property and function. Chromatin interaction results from scEChiA along pipelined developed algorithms give enhancer-gene, enhancer-enhancer, and gene-gene interactions, which could be a database for potential target whose knockdown could be a potential cure to cancer cells.

LIST OF FIGURES

CONTENTS

INTRODUCTION

Topologically Associated domains (TADs) were discovered in 2012 using chromosome conformation techniques like Hi-C[4,5]. These are the regions where DNA sequences within it interact with each other more often outside TAD. They occur across all species, including prokaryotes, and their understanding is of paramount importance in understanding 3D genome architecture, the functionality of the genome, and lastly, how genome interaction unfolds across species across cell lines in a novel way. TAD formation happens when cohesin rings bind to DNA sliding over CTCF molecules whose binding sequences point away from the loop. As soon as each of the rings reaches inward directed CTCF sequence loop formation stops[6]. It is also called or understood as the DNA loop extrusion model[7]. Human cells have a median TAD size of around 1.15 mb[8].

In-vitro Studies have shown that TAD is dynamic in nature, i.e., it shows transient behavior[9]. An important observation that makes studying TAD more important is that TAD boundaries remain conserved across cell types in species. TAD conservation property extends further in some cases across species as well[10]. This Topological Conservation property of TAD helps in identifying disrupted chromatin loops better, along with making way for therapeutics and CRISPR-based approaches on abnormal TAD boundaries. Various studies are now focusing on drug targeting TAD instead of the set of targeting DNA sequences.

Disruption of TAD boundaries can do many perturbations. It can affect gene expression directly or indirectly via changing enhancer-promoter interaction or enhancer-gene

interaction[11]. These disruptions are known to cause cancers or developmental disorders[12]. Hence understanding how genes inside a TAD inter-connect with each other and what it means for the fate of the TAD and the ultimate fate of the cell is of unique interest.

Recent Studies[13] have shown how TAD property is determined by genes present in it. The study describes how the number of genes present in a TAD relates to causing a disease, how the size of a TAD impacts its own survival outcome and how genes are distributed over TADs. These studies do motivate us to more deeply understand the Mechanism behind the interplaying of genes among themselves which define TAD property, and also shed light on TAD evolution. Hence, we have tried to understand TAD biology with respect to cancer by building a pipeline of algorithms.

Another important part of understanding the human genome, particularly TAD properties, is to understand the Mechanism of Chromatin Interactions. There is a profound amount of data and studies that have concluded how studying chromatin interactions can give novel drug targets. Chromatin Interactions through various techniques like Multiplex FISH imaging have identified numerous previously unknown TADs and Sub-TADs[14].

Studying Chromatin Structure gives a new dimension to understanding genome architecture. New Techniques like Chromatin interaction analysis with paired-end tag sequencing (ChIA-PET)[15] and Graph-based approaches[16] have further identified genome-wide long-range chromatin interactions giving further insights into the functioning and properties of TAD. Studying long-range chromatin structures in view of TAD-TAD

interactions gives novel insights into the structural feature of the genome at high resolution and small bin size.

Recent Algorithms, like Armatus[17], have revealed more TAD-like structures, which authors called as Alternative Domains that have not been previously identified as TAD by methods given by Dixon et al. from chromatin conformation data.

Understanding the Genetics of Diseases involves studying chromatin interactions in the purview of enhancer-gene, enhancer-enhancer, and gene-gene interactions. Understanding these three categories of interactions gives a broad and comprehensive approach to finding the genetic pathophysiology of diseases and finding novel biomarkers and therapeutic targets.

A review paper by Daniel J. Gaffney[18] beautifully explains how earlier methods of enhancer-gene interaction predictions did have a limitation in that they consider these interactions within only a range of distance, and they fail to predict long-range interactions that have a high significance. The author explains how Capture-C experiments are only able to detect interactions within a range of 20kb[19], which is not the ground truth.

Expression Quantitative trait loci (eQTL) is a recent method to determine genomic loci that explain the variation of mRNA expression levels[20]. There are cis or local-eQTL, which determines genomic loci located very near to the gene, whereas trans or distal-eQTL locates long-range genomic loci interactions. eQTL method has shown that

genomic interactions can happen up to 100kb distance. Although there are limitations in the eQTL approach as well, sometimes there is a coincidental overlap of GWAS and eQTL results[21]; hence exact causality of predicted genomic loci to query mRNA is questionable.

A recent experimental technique by Charles P et al. [3]. known as CRISPRi-Flow Fish, in a single experiment, maps all enhancers and their target genes without any limiting factor of the distance of gene with enhancer. They named their model as Activity by Contact (ABC) model. Their method is by far the most accurate since it uses an exact readout of enhancer function, which is varying levels of gene expression. Although the ABC model does have some limitations, ABC model accuracy goes down when extended to multiple cell types. But even then, it is the far-most accurate model to give enhancer-gene interactions.

Hence there is a need to build more models to predict enhancer-gene interactions along with enhancer-enhancer and gene-gene interactions, giving a more, broad understanding of the regulatory network concerning enhancer/s and gene/s. To achieve this, we have used the R package scEChiA[1] to predict enhancer-gene interactions.

Chapter 1

Exploiting TADs activity for predicting cancer survival

Brief Introduction:

Recent advances in genomic technologies have revelated Topologically associated domains (TAD) as one of the important genomic structures whose role and association with various Activities of cells is widely being explored and studied. Chapter -1 focuses on how TAD survival outcome could be predicted based on genes (present in that TAD) activity and how TAD's genes interplay with each other to ultimately define TAD property and TAD prognostic score. These studies give more insights into how TAD activity could be used for predicting cancer survival and could become a potential base for genomic editing technologies to treat cancer patients more precisely.

Background and Related Work:

Topologically associated domain (TAD) was discovered in 2012[4]. Since then, numerous studies have been done on understanding genome architecture and chromatin folding in disease and normal state in the context of TAD.

The study by Lifei Li et al. [22] beautifully showed how copy number variation in genes present inside the TAD defines TAD prognostic score and how it can be used to predict patient survival and how TAD prognostic can be used as a biomarker in cancer patients. In our work, ***we computed TAD prognostic scores for different cancer types on the basis of Gene Set Variation Analysis scores of genes present in the TAD.***

Studies have shown how disruption of TAD boundaries causes disease state and effect gene expression by changing enhancer-promoter interactions[11]. The study by Muro et al. [13] in 2019 explained the importance of genes presents inside TAD in defining TAD

property. The study also shed light on how the number of genes in TAD relates to disease state, TAD size relation to its survival, and why the distribution of genes in TAD is way like it is. However, it did not mention what the 'Mechanism' by which these genes present in the TAD interplay with each other to define TAD property. ***This lack of defining Mechanism was our core motivation to propose a mechanism on why and how genes present in TAD define TAD property.***

Theory and Methodology Used: Complete pipeline and protocols followed are shown in a comprehensive manner in Figure A for Chapter 1 work

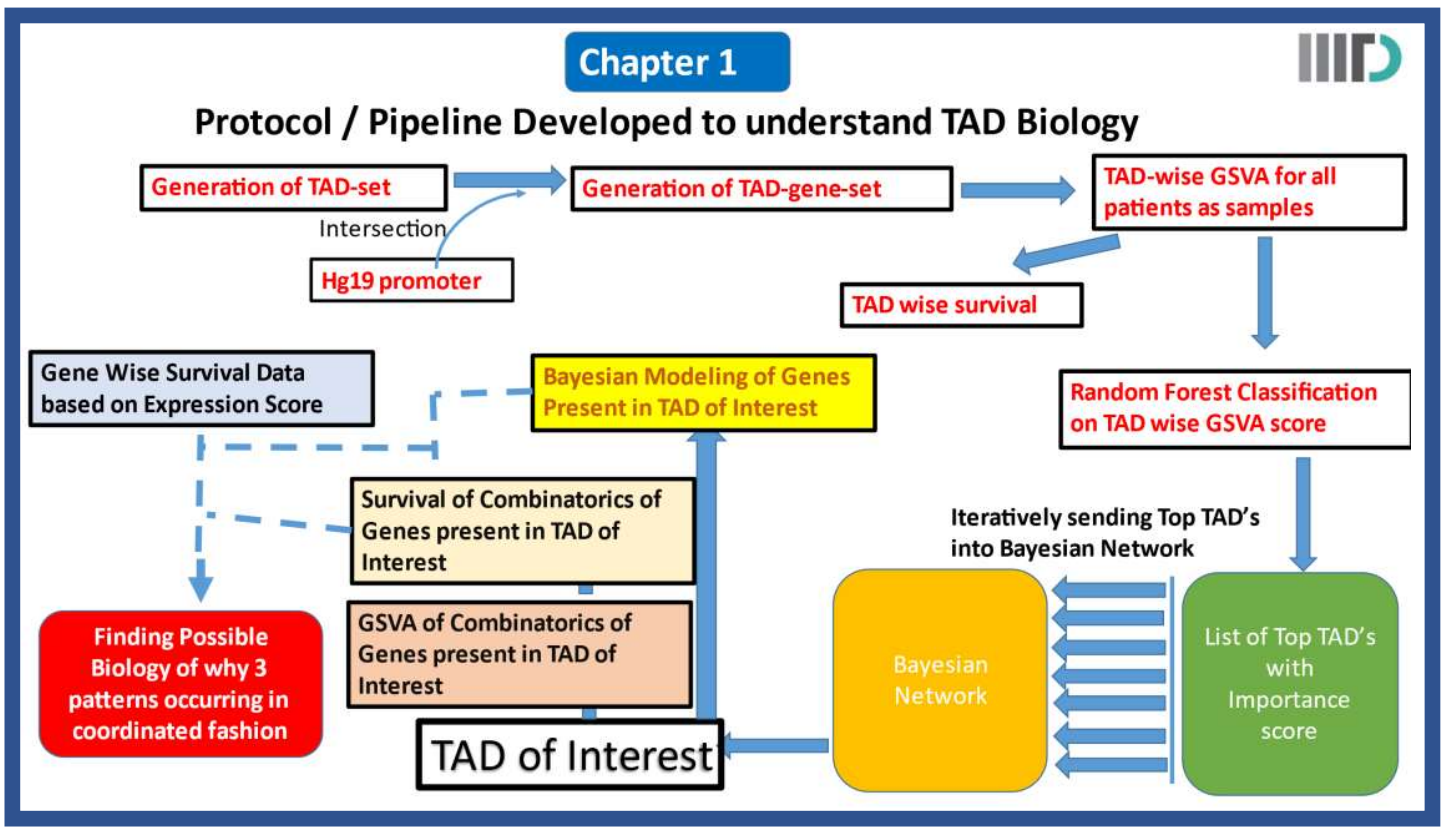

Figure A

1. **Preparing and pre-processing the Data**

As discussed above, TAD boundaries are conserved across cell type[23]. Using this principle, we tried to create a common TAD list for all cancers so that our analysis across TAD would be a PAN-CANCER analysis rather than single cancer oriented.

To create a common TAD list, we first took regions of Topologically Associated domains from the University of Miami TADKB database[24] server for different bin sizes (mainly 10kb to 50kb). Then we performed HiC on four cell lines taken from oral cancer cell lines curated by the Genome Institute of Singapore. 3 cell lines were from primary oral cancer tumors[25–27], while 1 cell line was of metastatic oral cancer[28]. HiC gave us chromatin interactions, and we merged it with the TADKB database to curate a novel common TAD list that can represent every cancer cell genome architecture.

To create a TAD file with genes present in it, we intersected TAD boundaries with the hg19 promoter via bedtools[29] to give us genes present in every TAD. Thus, we created the TAD-gene-set file, which has been used in the following sections. The bin sizes of our TAD-gene-set file were in the range of 0.2 to 70.25 MB. We used this TAD-gene-set file in each TCGA cancer analysis since it remains conserved.

NOTE: I have taken TCGA CESC cell line which is Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma, throughout the project. If anytime the cancer abbreviation is not mentioned, consider it as a CESC cell line.

2. Gene Set Variation Analysis (GSVA)

Gene-set variation analysis developed by Sonja et al. in 2013[30] is a widely used, powerful, and robust technique that estimates unsupervised variation of pathway activity over a sample population. It outperforms other enrichment methods when the Data is highly heterogeneous. We used the GSVA-R package

to calculate the TAD GSVA score for each cancer type. For each TAD, we run GSVA for genes present in the TAD, giving input of gene expression data over different samples (patients). This gives us the GSVA score of each TAD over all samples (patients) for each cancer.

GSVA scores for each TAD over all patients give an impression of how much a particular gene set (of that TAD) is enriched in a particular sample. GSVA computes cumulative density function (CDF) for each gene, then computes the rank of genes (lying in our gene set) via Klimigrov random walk statistics. GSVA score of TAD is interdependent on other TADs as well since we have used the CDF function. GSVA score is actually the Difference between H1 (highest Difference between K-S statistics of random walk for gene lying in gene set and gene, not lying-in gene-set) and H2 (lowest Difference between K-S statistics of random walk for gene lying in gene set and gene, not lying-in gene-set). Positive GSVA scores imply that genes are positively enriched in a given gene set, while negative GSVA score implies just the opposite. For our case, a Positive GSVA score over a patient (sample) implies that genes lying in that TAD on the given patient has high enrichment.

One fundamental challenge in Genomics is to differentiate between correlation and causation. When we apply GSVA, output GSVA scores not only represent correlation scores but also give an impression of causality. We have used this property of GSVA in Bayesian Modeling among genes of specified TAD in our course of study.

For our study, we computed GSVA scores for 20 cancer types and then used these scores to study the relation of TAD among themselves via the random forest and Bayesian Modeling. GSVA scores were also used as a biomarker to calculate the survival -p-value of TAD over all patients in particular cancer.

3. Survival analysis

Survival graphs and their score (p-value) are an important estimate to understand how a particular biomarker activity given over a number of samples (patients) contributes to their survival outcome. Survival p value < 0.05 is generally considered as a cut-off to certify our null hypothesis.

Survival score calculation do require the classification of samples into categories (1/0). This classification cut-off is of paramount importance as it will decide the survival score. This cut-off is the ultimate benchmarking upon which the whole study outcome is inferred and presented. Deciding this cut-off needs proper evaluation and requires a proper understanding of Statistics and the nature of data. In our case, it is oncological-genomic data.

4. Statistical Approaches to handle Survival analysis

In most studies median is set as the cut off which dichotomizes the biomarker, guaranteeing an equal sample size for both groups.[31] There are widespread studies on how cut-off point variation results in a shift of the ROC curve[32]. Altman et al. in 1994 famously have written the dangers of using optimal-cut-off in survival analysis[33] and how it gives rise to data-dredging bias[31]. Finally, in March 1992, Maximally Ranked Statistics by Lausen[34] was proposed, which finds out the maxima of standard statistics of all possible cut points. Monte-Carlo simulations then give conditional p value[31]. Minimizing type-1 error by Altman DG (1994)[33] and α-level adjustment by Miller R (1982)[35] and Mazumdar M (2000), [36] were

some more improvements done to make sure the right cut-off is decided. Lastly, Francisco Tus Tumi mentions that deciding cut-off points will always give some bias[31].

Prof Douglas G Altman (professor at Wolfson college at Oxford) is regarded as the most trustworthy and final voice in medical statistical analysis, particularly in the field of survival analysis on medical/biomedical data. In 1994 did propose the dangers of using optimal cut off[33], but he did give some liberty to use various cut-offs, provided that strong biological explanations are provided to justify these cut-offs, and the optimal cut-off term should be termed minimum-p-value term.

But Prof Altman came down heavily in the year 2006 when he published an article in British Medical Journal[37] where he bluntly said that categorizing biomarkers (in our case, it is GSVA) will give less statistical power, more probability of true positives to becoming false positives and under-estimation of the extent of variation in outcome between groups. Therefore he advocated for the use of Regression in place of cut-off-based survival analysis.

Taking note of Prof Altman's above work, we have used Regression as the basis for survival analysis.

5. Random Forest Classification

Firstly, for each cancer type input matrix was designed, which contains a patient-wise GSVA score for all TAD's including a risk column. This risk column is made on the principle that the higher the number of days to death, the less risk is, and vice-versa. Here we calculate the median of days to death in each cancer type and then denote high risk when days to death are less than the median or low

risk when days to death are more than the median value. This is how the input matrix is designed for the random forest to work upon as input.

The random forest does the classification of features to sort which features are the most appropriate predictors for the response or target variable. Here features are our TAD, and the target variable is the risk column (0 or 1) in factor format.

Here an important question arises of what should be the number of trees we should use in order to correctly classify TAD which targets risk in high value. Since no literature was available on the number of trees used with respect to the number of columns to get the best classifying prediction accuracy, some illustrations showed the relation of trees used and the rows in dataset [38–40]. The study by Oshiro et al. [41] shows how doubling the number of trees does not impact significant performance gain, and it gives rough estimates of using trees in the range of 64-128. But it also showed that as the number of attributes (features) increase, the higher number of trees up to 4096 might be an ideal estimate to be needed. Hence we did not get a true specified range of trees we should use to get to the most optimal classification solution.

In order to fix this problem, we made an iterating algorithm where each time we fixed the number of trees, ran the random classifier forest, and then calculated its error rate. We then select the optimal number of trees to be used based on the condition where the error rate is at its minimum. 'err. Rate' is one of the output parameters in a list of Random forests. It contains a column named Out-of-bag (OOB) error parameter, which is simply the prediction error out of the random forest for each tree used. Simply calculating the median of all OBB errors for all number of trees used gives us the error rate for one iteration.

For iteration number of trees varied from 50 to 2500. Here the maximum number of trees was chosen to be 2500 after many experiments. We observed that the OBB error rate does get to a minimum point between 50 to 2500 trees, and beyond 2500, the OBB error rate does not get to its global minimum further. For cancer wise optimal number of trees also varies. After Running a random forest, we then made a sorted matrix where each TAD (feature) has its Mean-Decrease-Gini-Score. It determines how closely the TAD is related to the risk factor. The higher the Mean-Decrease-Gini-Score higher the association and vice-versa. We stored this matrix and sent it to the Bayesian Modeling R package – 'bn-learn.'

6. Bayesian Modeling

Bayesian Modeling gives probabilistic associations between variables via Directed acyclic graphs. It gives dependencies among variables given a condition has happened. Bayesian networks are an easy-to-understand methodology both mathematically and visually and also consume less computational memory compared with exhaustive probability tables. Many algorithms are present in Bayesian Modeling. There are constraint-based structure algorithms, score-based structure learning algorithms, hybrid structure learning algorithms, local discovery algorithms, and Bayesian network classifiers.

Analysis and Studies by Marco Scutari[42] and Cowell[43] showed that constraint-based and score-based algorithms select identical discrete Bayesian networks. Further study by Marco Scutari[42] showed that a score-based algorithm produces a high-likelihood network giving realistic models and propagating evidence. Also, it showed that score-based algorithms are faster than hybrid and constrain based algorithms.

Hence, we used score-based algorithms in our study. We used Tabu search in place of a hill-climbing algorithm for the simple reason that hill climbing does get stuck in local optima while tabu search maintains a tabu list which holds objects that are taboo for current iteration use for avoiding getting stuck[44,45].

We used Random Forest for the classification of TAD, which is most likely to be affecting the Risk node; Random Forest classification being a supervised bagging algorithm, gives us all TADs with their classification or Mean-Decrease-Gini-Score. Using these lists of TADs with classification scores, we send these TADs into Bayesian networking in an iterating manner where in each iteration, we decide number of TADs to be given as input to the Bayesian network. The whole idea of sending different amounts of TADs into the Bayesian model is to find the most frequently overlapping TAD, which comes from a parent node to a risk node. Using this approach, we can identify TAD of Interest which comes as a frequent parent node to risk node considering the dynamic nature of Bayesian Modeling.

We used nine iterations to submit random forest results to Bayesian Modeling. In each iteration, a specific top TAD out of the random forest was chosen and sent to the bn-learn R package. Specific top TAD numbers for successive iterations were from 20 to 200. We chose the endpoint as 200 because beyond the top 200 TAD, there is almost zero probability that any TAD beyond 200 ranks would be featured in the list of TAD of Interest (parent node to risk node) out of Bayesian analysis.

Here for each iteration, we not only calculated Bayesian network parameters for the complete dataset, but we also calculated AIC (Akaike Information Criterion), which is derived from frequentist probability, and BIC (Bayesian Information Criterion), which is derived from Bayesian probability. Both scores are desired to be low for a good model to fit into the Bayesian network. However, we analyzed

BIC scores since they are more consistent and tolerant and penalize free parameters more strongly than AIC.

We also used one approach to observe whether the TOP parent node does get affected by the number of patients (sample size). To achieve this, we divided our input matrix into two equal halves, then ran the same tabu search Bayesian network on both halves and found which is the top node to risk factor each time. We did this for all nine iterations.

7. GSVA combinatorics and Survival

Once we identify the top prognostic tad for a particular cancer type, we are now interested in how genes present in that TAD work as a team/unit to define TAD property or make that TAD prognostic or not prognostic. So to identify the novel relationship among genes within a TAD, we performed combinatorics GSVA.

The idea is simple. We make all possible combinations of a set of all genes in specified TAD and run GSVA on those combinations. We can call them experimental virtual TAD. After GSVA scores are calculated on them, we compute survival on all those virtual TADs. Survival scores give us a novel understanding of how these genes are interplaying with each other. And we can easily find a correlation among different combinations of genes via their survival p-value.

As discussed above, GSVA scores do give an impression of causality in addition to correlation, so GSVA combinatorics and further survival analysis based on GSVA scores give us a good footprint of both correlation and causality of survival of genes among themselves.

This novel technique sheds light on a completely new gene-regulatory network that defines particular TAD properties which we believe not has been touched upon before by any work by a Research group or individual.

8. Bayesian Modeling of TAD genes

As discussed above, as there is no available data or hypothesis which relates the interplay of genes as a unit to define a TAD property, benchmarking our results has become a major challenge.

To certify our results of GSVA combinatorics, we perform unsupervised Bayesian Modeling via dynamic Tabu search methodology in R package bn-learn[46] to give us a network. We have then benchmarked our GSVA combinatorics gene regulatory network against the Bayesian output of the genes of the same TAD.

Results

1. For each cancer type TAD, a wise GSVA scores file was generated. A sample output via the R Studio window is shown in Fig 1. Columns here are patients, and rows determine TADs. In contrast, each cell value is the GSVA score.

	TCGA.2W.A8YY.01A.11R.A37O.07	TCGA.4J.AA1J.01A.21R.A38B.07
chr10_100025000_102750000	0.02077664	0.38428865
chr10_100100000_101125000	0.02609180	0.64737727
chr10_100220000_101140000	-0.76504323	0.81682787
chr10_101190000_101380000	-0.74350084	0.72017976
chr10_101190000_101480000	-0.68343785	0.73005479
chr10_101375000_102675000	0.10493172	0.27621032

Figure – 1

2. Survival analysis was performed on these GSVA scores and for each TAD survival-p-value. Fig 2 shows a table showing the Survival value (via Regression) for each TAD for CESC cancer type.

TAD	p-value
chr7_2825000_6875000	1.43E-06
chr8_6270000_7270000	4.17E-06
chr7_5900000_6850000	4.80E-06
chr7_6000000_6875000	5.09E-06
chr7_5860000_6900000	5.26E-06
chr3_165010000_167380000	8.43E-06
chr7_6550000_6970000	8.56E-06
chr9_32580000_33010000	1.11E-05
chr7_6530000_6860000	1.22E-05
chr7_2825000_5900000	1.94E-05
chr6_139690000_141920000	1.97E-05
chr7_2750000_2900000	2.06E-05
chr8_2300000_7075000	2.73E-05
chr7_5200000_5775000	2.79E-05
chr7_5820000_6150000	3.94E-05
chr1_146500000_148570000	3.99E-05
chr1_146400000_149650000	6.72E-05
chr1_146400000_149725000	6.72E-05
chr8_6690000_7250000	6.98E-05
chr1_171750000_172350000	7.95E-05

Figure 2

3. Random forest input file for the CESC cancer type is shown in Figure 3

	days_to_death	risk	TAD_chr10_100025000_102750000	TAD_chr10_100100000_101125000
TCGA-C5-A0TN-01	348	1	-0.610388027	-0.559280423
TCGA-C5-A1BE-01	2094	0	0.333246196	0.371126061
TCGA-C5-A1BF-01	570	1	0.322512789	-0.337301005
TCGA-C5-A1BM-01	2520	0	-0.544520009	-0.330243411
TCGA-C5-A1BN-01	166	1	0.234574171	0.027301039
TCGA-C5-A1BQ-01	604	1	-0.666347864	-0.605866942
TCGA-C5-A1M5-01	2052	0	0.385485204	0.452835375
TCGA-C5-A1M6-01	955	0	-0.331853606	-0.013578323
TCGA-C5-A1M9-01	1065	0	0.43804424	0.189130332
TCGA-C5-A1MH-01	1186	0	-0.307214335	0.006763374
TCGA-C5-A1MI-01	1083	0	-0.49890431	-0.318487011
TCGA-C5-A1MJ-01	14	1	-0.153298384	0.285728766
TCGA-C5-A1MK-01	74	1	-0.325744141	-0.542886144
TCGA-C5-A1ML-01	636	0	-0.541599545	-0.495068878
TCGA-C5-A1MN-01	1245	0	-0.007877562	-0.513055752
TCGA-C5-A2LZ-01	3046	0	0.422896192	0.44412476
TCGA-C5-A2M2-01	1011	0	0.251605557	-0.067643489
TCGA-C5-A3HF-01	543	1	0.469200082	0.784334613

Figure 3

4. Figure 4 shows the error rate for the different numbers of trees used as input for CESC cancer type. Here we can see that when we use 1050 trees in the random forest for CESC cancer, we get the least error rate. Hence for further analysis, we have used 1050 trees.

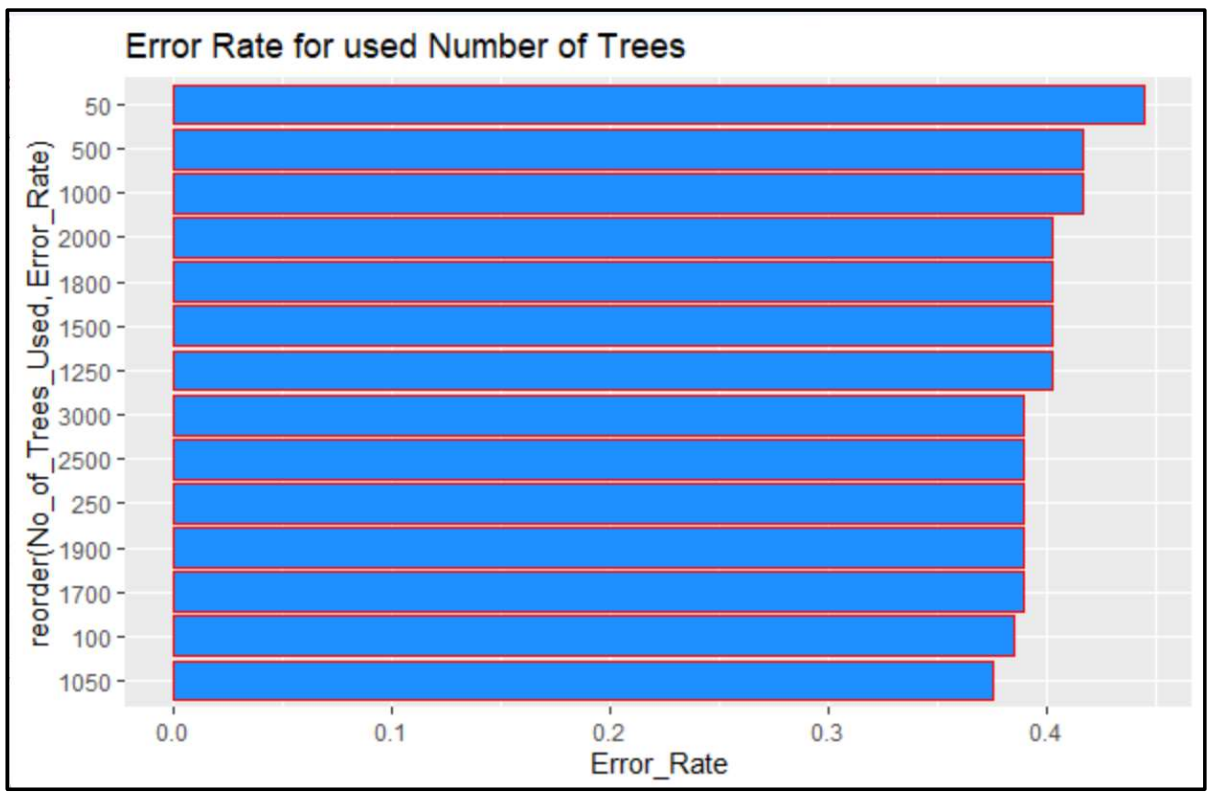

Figure 4

5. Random forest output showing each TAD importance score which gives its correlation with risk factor (built on the basis of days to death value) shown in Figure-5

TAD	MeanDecreaseGini
TAD_chr3_118840000_119430000	0.13357175
TAD_chr1_171750000_172350000	0.13038994
TAD_chr9_34390000_34610000	0.11675401
TAD_chr7_2825000_6875000	0.10916424
TAD_chr18_5190000_6520000	0.10091579
TAD_chr14_96020000_96970000	0.09414148
TAD_chr14_65475000_66525000	0.08861155
TAD_chr1_87550000_89300000	0.08783845
TAD_chr8_14290000_14750000	0.08684187
TAD_chr14_65460000_66570000	0.08625348
TAD_chr18_5060000_6530000	0.08599453

Figure 5

6. A comparative Table is shown in Figure 6, which shows how many times particular TADs appear as the parent node of the Risk factor in CESC cancer type. Bayesian analysis was run in 9 iterations by taking the different number of top TADs from the output of the Random Forest.

TAD	Freq
chr1_171750000_172350000	3
chr7_2825000_6875000	2
chr14_42070000_45360000	1
chr3_118600000_119375000	1
chr3_186910000_187440000	1

Figure 6

7. The comparative matrix of CESC cancer type made out of the Bayesian network shown in Figure 7 shows the number of TADs as a parent node to risk, the top parent node, and their AIC and BIC score. It also shows TOP TAD in the first half of the input matrix and likewise in the second half.

top-tads taken	no of parents node detected	top-parent-node-name	strenth-to-risk-node	aic-score	bic-score	top-tad-in-first-half-bnlearn	top-tad-in-second-half-bnlearn
20	2	TAD_chr8_25020000_259700...	-4.283559	-199.4968	-283.7334	TAD_chr3_118840000_1...	TAD_chr11_20650000_22180000
30	2	TAD_chr18_47010000_47980...	-1.881234	-148.2430	-303.0563	TAD_chr14_96020000_9...	TAD_chr14_96020000_96970000
50	3	TAD_chr5_94890000_951900...	-1.374074	-140.0049	-369.9482	TAD_chr1_171750000_1...	TAD_chr8_13625000_15600000
75	3	TAD_chr5_28980000_312900...	-4.181335	-122.2261	-409.0860	TAD_chr7_5860000_690...	TAD_chr15_85000000_86375000
100	2	TAD_chr14_42070000_45360...	-7.568089	-573.6792	-917.4558	TAD_chr7_2825000_687...	TAD_chr8_13625000_15600000
120	0	NA	NA	-1119.67...	-1508.98...	TAD_chr9_34390000_34...	TAD_chr8_13625000_15600000
150	0	NA	NA	-2023.40...	-2481.01...	NA	TAD_chr8_13625000_15600000
175	0	NA	NA	-2629.24...	-3143.77...	NA	TAD_chr8_13625000_15600000
200	0	NA	NA	-3562.41...	-4133.85...	NA	NA

Figure 7

8. When TADs from Figure 6, which appear the greatest number of times as parent node to risk, was merged with the database of survival-value for the same CESC cancer, we got results as shown in Figure 8

TAD	freq	p_value
chr1_171750000_172350000	3	7.95241216245418e-05
chr7_2825000_6875000	2	1.43074233762058e-06
chr14_42070000_45360000	1	0.0510488574252965
chr3_118600000_119375000	1	0.000160954011812913
chr3_186910000_187440000	1	9.11872113308478e-05

Figure 8

9. Since as we have observed that TAD chr1_171750000_172350000 comes the greatest number of times as parent node and also Top parent node to risk node when we take 100 or 200 Top TAD from the random forest. Figure 9 shows a Bayesian network representing connections of TAD chr1_171750000_172350000 with other TADs. It can be seen that our TAD chr1_171750000_172350000 is acting as the parent node to the risk node. Figure10 shows the survival graph of TAD chr1_171750000_172350000

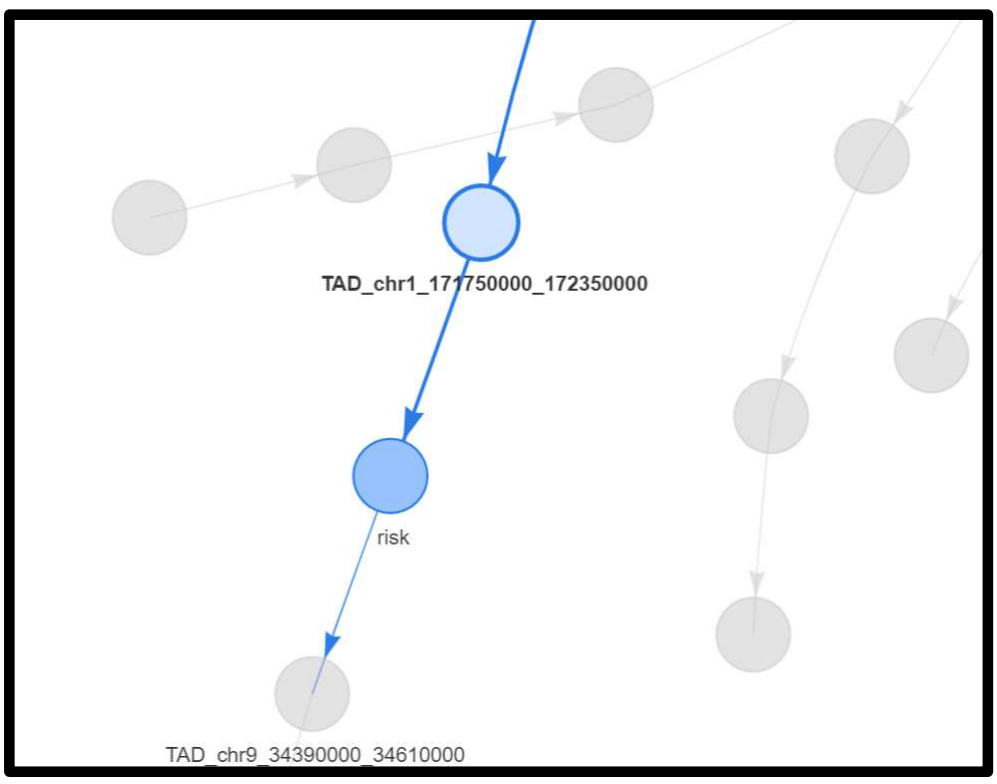

Figure 9

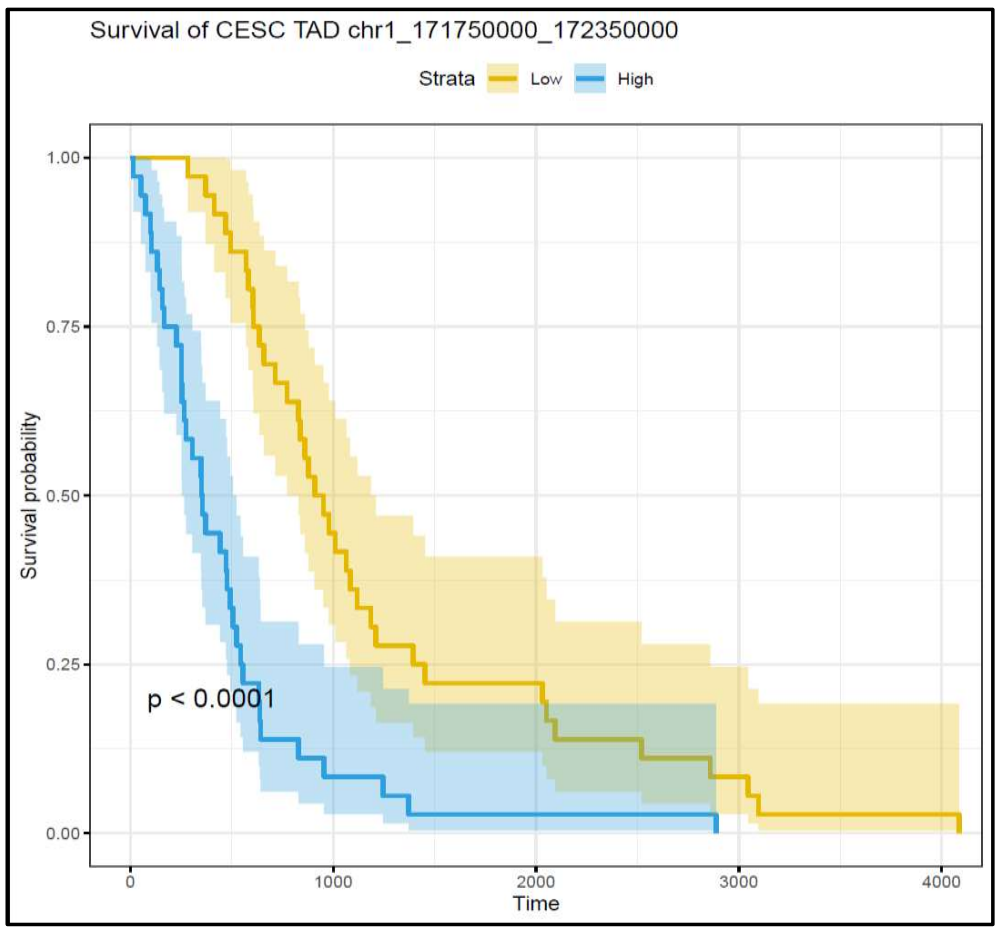

Figure 10

10. Combinatorics of the genes present TAD chr1_171750000_172350000 in CESC cancer type were made. Here we removed pseudo-genes and non-coding RNA from combinatorics. After the removal of pseudo genes and non-coding RNA, we had five genes in our TAD. Making its combinatorics, we got 32 combinations matrix as shown in Figure 11

Exp-TAD-1_METTL13	METTL13		
Exp-TAD-1_DNM3	DNM3		
Exp-TAD-1_MIR214	MIR214		
Exp-TAD-1_DNM3OS	DNM3OS		
Exp-TAD-1_MIR199A2	MIR199A2		
Exp-TAD-2_METTL13-DNM3	METTL13	DNM3	
Exp-TAD-2_METTL13-MIR214	METTL13	MIR214	
Exp-TAD-2_METTL13-DNM3OS	METTL13	DNM3OS	
Exp-TAD-2_METTL13-MIR199A2	METTL13	MIR199A2	
Exp-TAD-2_DNM3-MIR214	DNM3	MIR214	
Exp-TAD-2_DNM3-DNM3OS	DNM3	DNM3OS	
Exp-TAD-2_DNM3-MIR199A2	DNM3	MIR199A2	
Exp-TAD-2_MIR214-DNM3OS	MIR214	DNM3OS	
Exp-TAD-2_MIR214-MIR199A2	MIR214	MIR199A2	
Exp-TAD-2_DNM3OS-MIR199A2	DNM3OS	MIR199A2	
Exp-TAD-3_METTL13-DNM3-MIR214	METTL13	DNM3	MIR214
Exp-TAD-3_METTL13-DNM3-DNM3OS	METTL13	DNM3	DNM3OS
Exp-TAD-3_METTL13-DNM3-MIR199A2	METTL13	DNM3	MIR199A2
Exp-TAD-3_METTL13-MIR214-DNM3OS	METTL13	MIR214	DNM3OS

Figure 11

11. GSVA of each combinatoric TAD chr1_171750000_172350000 was calculated as shown in Figure 12

	TCGA.2W.A8YY.01A.11R.A37O.07	TCGA.4J.AA1J.01A.21R.A38B.07	TCGA.BI.A0VR.01A.11R.A10U.07
chr1_171750000_172350000	0.084859264	-0.042390296	-0.662703937
Exp-TAD-1_METTL13	0.654127616	-0.673385544	-0.091368169
Exp-TAD-1_DNM3	0.451534215	-0.217956948	0.38494458
Exp-TAD-1_MIR214	-0.171395558	0.285231309	-0.944365986
Exp-TAD-1_DNM3OS	-0.630333376	0.526340566	-0.692557881
Exp-TAD-1_MIR199A2	-0.170882013	0.285744854	-0.943852441
Exp-TAD-2_METTL13-DNM3	0.552854575	-0.445690319	0.193184205
Exp-TAD-2_METTL13-MIR214	0.446544125	-0.37578228	-0.856181787
Exp-TAD-2_METTL13-DNM3OS	0.030488953	-0.19601975	-0.576034568
Exp-TAD-2_METTL13-MIR199A2	0.447037391	-0.37540619	-0.855624472
Exp-TAD-2_DNM3-MIR214	0.176435172	0.03363862	-0.654749285
Exp-TAD-2_DNM3-DNM3OS	-0.212912492	0.233542006	-0.335248671
Exp-TAD-2_DNM3-MIR199A2	0.177033294	0.033895404	-0.654123791
Exp-TAD-2_MIR214-DNM3OS	-0.416565687	0.405803304	-0.818496961
Exp-TAD-2_MIR214-MIR199A2	-0.17114611	0.2855003	-0.944149619
Exp-TAD-2_DNM3OS-MIR199A2	-0.417069692	0.406060087	-0.818240178
Exp-TAD-3_METTL13-DNM3-MIR214	0.418662125	-0.245669531	-0.609079865
Exp-TAD-3_METTL13-DNM3-DNM3OS	0.184510239	-0.148331927	-0.285018383

Figure 12

12. Using these GSVA scores, a Bayesian network model of all five functional genes in our TAD chr1_171750000_172350000 was calculated, as shown in Figure 13

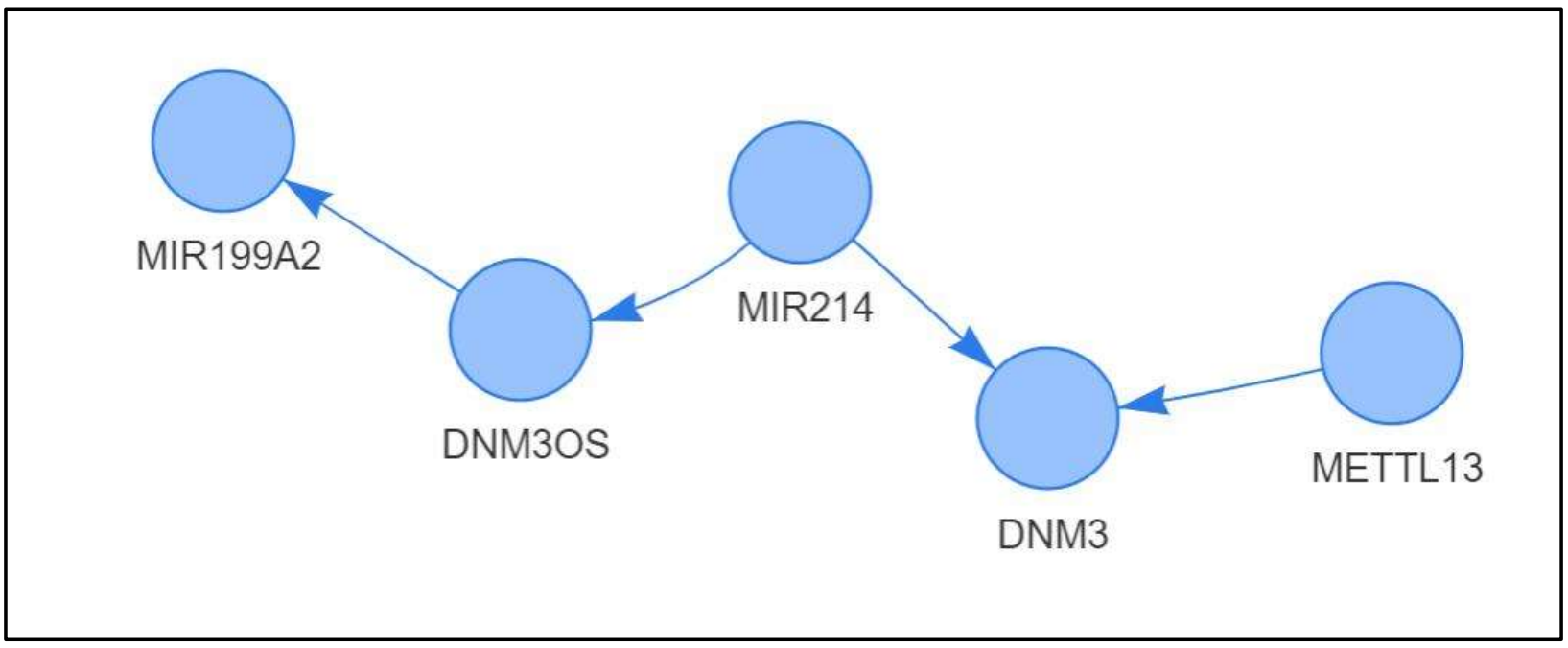

Figure 13

13. Based on Combinatorics GSVA of our TAD chr1_171750000_172350000, the survival p-value was also calculated. Figure 14 shows the survival -the value of TADs in ascending order (sorted). In Figure 14, it is mentioned for each row as 'Experimental (Exp)-TAD- (number of genes in a cluster) **names of genes in cluster**. The cluster of genes varies from 1 to 5 since we have five functional genes in total in our TAD. In total, there are 32 experimental TADs.

TAD	p-value
Exp-TAD-4_METTL13-DNM3-MIR214-MIR199A2	2.44E-05
Exp-TAD-4_METTL13-MIR214-DNM3OS-MIR199A2	4.39E-05
Exp-TAD-2_METTL13-MIR199A2	4.60E-05
Exp-TAD-2_METTL13-MIR214	4.62E-05
Exp-TAD-3_METTL13-MIR214-MIR199A2	4.92E-05
Exp-TAD-5_METTL13-DNM3-MIR214-DNM3OS-MIR199A2	7.95E-05
Exp-TAD-3_METTL13-DNM3-MIR199A2	0.000165723
Exp-TAD-3_METTL13-DNM3-MIR214	0.000165871
Exp-TAD-3_METTL13-DNM3OS-MIR199A2	0.000447931
Exp-TAD-3_METTL13-MIR214-DNM3OS	0.000449892
Exp-TAD-3_DNM3-MIR214-MIR199A2	0.001074155
Exp-TAD-1_MIR214	0.001219824
Exp-TAD-2_MIR214-MIR199A2	0.001219824
Exp-TAD-1_MIR199A2	0.001219825
Exp-TAD-4_METTL13-DNM3-DNM3OS-MIR199A2	0.00263217
Exp-TAD-4_METTL13-DNM3-MIR214-DNM3OS	0.002635888
Exp-TAD-3_MIR214-DNM3OS-MIR199A2	0.003023336
Exp-TAD-4_DNM3-MIR214-DNM3OS-MIR199A2	0.003053381
Exp-TAD-2_DNM3-MIR199A2	0.00411795
Exp-TAD-2_DNM3-MIR214	0.004121122
Exp-TAD-2_DNM3OS-MIR199A2	0.015645562
Exp-TAD-2_MIR214-DNM3OS	0.015668145
Exp-TAD-1_METTL13	0.022350786
Exp-TAD-3_DNM3-MIR214-DNM3OS	0.048938629
Exp-TAD-3_DNM3-DNM3OS-MIR199A2	0.048949964
Exp-TAD-2_METTL13-DNM3	0.071744452
Exp-TAD-2_METTL13-DNM3OS	0.106601449
Exp-TAD-3_METTL13-DNM3-DNM3OS	0.274651274
Exp-TAD-1_DNM3	0.600942388
Exp-TAD-2_DNM3-DNM3OS	0.841355654
Exp-TAD-1_DNM3OS	0.848654453

Figure 14

14. Further for analysis, we have also calculated gene-wise Survival for CESC cancer type. Figure 15 shows the survival of only five functional genes of our TAD chr1_171750000_172350000.

Genes	p-value
MIR214	0.001219824
MIR199A2	0.001219825
METTL13	0.022350786
DNM3	0.600942388
DNM3OS	0.848654453

Figure 15

Observations and Inferences:

1. From Figure 2, Survival analysis of CESC cancer type showed that TAD chr8_27170000_27460000 shows the best survival p-value (lowest) **But Supervised Random Forest analysis (target to risk node) output from Figure 6 shows that for CESC cancer type, TAD chr3_118840000_119430000 comes at the top on the basis of Mean Decrease Gini Score.**

    a. Since Random Forest does helps in finding importance of TAD in survival modeling, it gives a better idea of the TAD which affects risk factors; we can conclude that TAD highlighted by random forest **chr3_118840000_119430000** which also has a significant association to survival p-value; hence it is a better choice of further study.

2. TAD's selected by random forest were used in Bayesian analysis are sent in iteration to Bayesian network, we observed from Figure 7 that each time parent node is different for the different number of input TAD given by Random Forest.

When we crosschecked parent ***nodes' survival p-value for all iterations, we interestingly found that all have significant survival p-values. Hence it signifies that survival-p-value is an important parameter to deciding TAD prognostic nature, although not the only criterion.***

3. From Figure 7 itself, we can see that when we divide the matrix in half and pass it into the Bayesian network, there is a change in the TOP parent node to the risk node. This does ***signify that Bayesian Modeling also depends on the sample size taken***. Although we also observed that while the TOP parent node may change when we input half matrix, some parent nodes to a risk factor (when the first half and second half are input to the Bayesian network) remains common.

4. Figure 7 also shows AIC and BIC scores, but we neglect these parameters for iterations up to the top 200 TADs for the simple reason that up to 200 TAD, there is a significant change in the parent node list. We can surely consider AIC and BIC scores when we take TADs beyond the Top 200 from Random Forest. Then ***AIC and BIC scores (the lower, the better) would give us an idea of which model fits best.***

5. From Figure 8, we clearly get to observe that making a frequency table for noting down which TAD repeatedly comes as a parent node to risk node gives us a better approach to identifying our TAD of Interest since Bayesian networks are dynamic in nature for each iteration. ***We observe our TAD of Interest as TAD chr1_171750000_172350000, which has a significantly low p-value also.***

6. The Bayesian network in figure 13 shows the ***clear linkage of our TAD of Interest chr1_171750000_172350000 to the risk node,*** and also, it's the only parent node that targets the risk node.

7. To understand our TAD of Interest chr1_171750000_172350000 biology, GSVA Combinatorics and subsequent survival analysis were done as shown in Figure 12 and Figure 14. Then we did ***Bayesian Modeling of genes present in TAD of Interest chr1_171750000_172350000 as shown in Figure 13. It answers a fundamental question about TAD functioning and its properties defined by genes present in it.*** If we closely look at the Survival Outcomes of the combinatorics, as in Figure 14, we observe here that ***when all TAD genes GSVA scores are taken into account, they give the best survival score of all of the combinatorics.*** This nature can be explained by again looking at the Bayesian Modeling of genes (Figure 13) and Gene wise survival in CESC cancer based on expression data (Figure 15) with Combinatorics Survival in a comparative manner.

TAD	p-value
Exp-TAD-4_METTL13-DNM3-MIR214-MIR199A2	2.44E-05
Exp-TAD-4_METTL13-MIR214-DNM3OS-MIR199A2	4.39E-05
Exp-TAD-2_METTL13-MIR199A2	4.60E-05
Exp-TAD-2_METTL13-MIR214	4.62E-05
Exp-TAD-3_METTL13-MIR214-MIR199A2	4.92E-05
Exp-TAD-5_METTL13-DNM3-MIR214-DNM3OS-MIR199A2	7.95E-05
Exp-TAD-3_METTL13-DNM3-MIR199A2	0.0001657
Exp-TAD-3_METTL13-DNM3-MIR214	0.0001659
Exp-TAD-3_METTL13-DNM3OS-MIR199A2	0.0004479
Exp-TAD-3_METTL13-MIR214-DNM3OS	0.0004499
Exp-TAD-3_DNM3-MIR214-MIR199A2	0.0010742
Exp-TAD-1_MIR214	0.0012198
Exp-TAD-2_MIR214-MIR199A2	0.0012198
Exp-TAD-1_MIR199A2	0.0012198
Exp-TAD-4_METTL13-DNM3-DNM3OS-MIR199A2	0.0026322
Exp-TAD-4_METTL13-DNM3-MIR214-DNM3OS	0.0026359
Exp-TAD-3_MIR214-DNM3OS-MIR199A2	0.0030233
Exp-TAD-4_DNM3-MIR214-DNM3OS-MIR199A2	0.0030534
Exp-TAD-2_DNM3-MIR199A2	0.0041179
Exp-TAD-2_DNM3-MIR214	0.0041211
Exp-TAD-2_DNM3OS-MIR199A2	0.0156456
Exp-TAD-2_MIR214-DNM3OS	0.0156681
Exp-TAD-1_METTL13	0.0223508
Exp-TAD-3_DNM3-MIR214-DNM3OS	0.0489386
Exp-TAD-3_DNM3-DNM3OS-MIR199A2	0.04895
Exp-TAD-2_METTL13-DNM3	0.0717445
Exp-TAD-2_METTL13-DNM3OS	0.1066014
Exp-TAD-3_METTL13-DNM3-DNM3OS	0.2746513
Exp-TAD-1_DNM3	0.6009424
Exp-TAD-2_DNM3-DNM3OS	0.8413557
Exp-TAD-1_DNM3OS	0.8486545

Figure 14

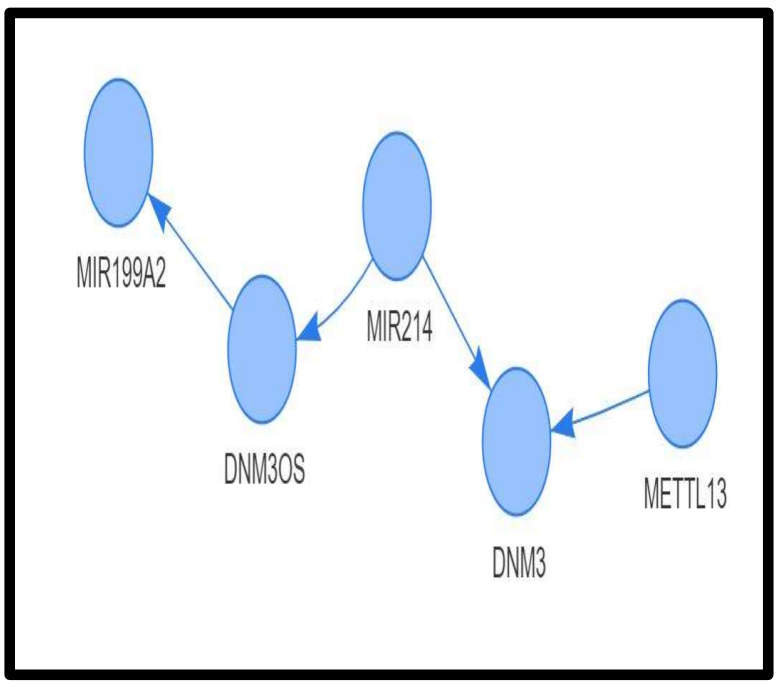

Figure 13

Genes	p-value
MIR214	0.001219824
MIR199A2	0.001219825
METTL13	0.022350786
DNM3	0.600942388
DNM3OS	0.848654453

Figure 15

Here we observe that TAD ***chr1_171750000_172350000*** itself (or the five combination TAD) has a low p-value compared to most of the combinatorics but has a high p-value compared to some high survival TAD.

Note: Here, we define survival score as the score which gives the best survival of the gene or the lowest survival p-value and vice-versa.

We observe that Exp-TAD, which has one gene only, has the same trend of the p-value as shown in Figure 15, where gene-wise survival is shown.

Now we also observe that gene METTL13, in combination with low survival score genes like DNM3OS and DNM3, becomes low survival score than only the METTL13 gene, implicating that ***when there is a combination of genes, the overall survival score depends on two factors, one factor is how good survival score is of individual gene, and the other factor is of how well-combined survival score.*** A combination of both these factors determines the overall experimental TAD survival score. We can confirm this conclusion by looking at other genes like MIR214 and how survival score diminishes when MIR214 combines with DNM3 and DNM3OS genes.

Another beautiful observation is that when MI214 and MIR199A2 genes are in combination, it performs less than individual MIR214 gene expression, although both genes have a high survival score. This is explained by looking at the Bayesian network in figure 13, whereby it is shown clearly how MIR199A2 also activates the DNM3OS gene (which is a very low survival score gene); hence when M1214 and MIR199A2 are in the same TAD, their survival score is low because of hidden linkage between MIR199A2 and DNM3OS. This conclusion shows how ***TAD's overall survival score also depends on the third factor as well which is how genes interplay with each other.*** Any linkage to a low survival score gene would reduce TAD's overall survival score, whereas linkage to high survival genes would upgrade TAD's overall survival score.

Hence overall TAD survival score depends on three factors, as depicted below.

We can thus conclude that survival score of TAD is combination of two factors, first factor being Combination of genes survival score present in the TAD and second being survival scores due to interplay of gene network present in TAD.

Hence by looking at the Bayesian network along with gene survival score when expressed independently, our survival results of combinatorics become crosschecked and validated.

The ***Difference in survival p-value of different combinatorics is due to the survival impact of independent genes when expressed independently, along with how these genes interact with one another.***

Hence even if genes present in combinations do not have the best survival score when expressed independently, they can rank the TAD better in survival score if they impact high survival-scoring genes. Hence even if a weak surviving gene is present in the TAD, it will still make a good amount of Difference in the overall Survival of the TAD if it is a parent node of high survival genes like MIR199A2, MIR214 or METTL13.

Hence ***Survival of TAD depends on every individual gene on account of how is their survival score when expressed independently plus how every individual gene is impacting other genes, impacting other high survival genes would increase the overall TAD survival score, whereas if they impact other low survival genes, TAD overall survival score is bound to decrease.*** Hence these patterns are observed.

Hence these results give the fundamental impression of how genes interplay within a TAD and how a TAD property is defined by the union of genes and not only by the nature of individual genes.

8. Further we searched for every gene, literature on how it impacts Cervical Cancer so that we could validate our study.

a. The study by Xian et al. [47] clearly showed how **MIR199A2, which is a member of the microRNA-199a family, reduces cell proliferation and Invasion of Cervical Cancer cells**[47]**.** They showed how the expression of MIR199A2 is low in CESC cancer cell lines, and high expression of MIR199A2 inhibits and promotes AKT/mTOR pathway by targeting B7-H3 *hence suppressing the* growth of cervical cancer tumor cells. ***This validates our results and hypothesis as to why experimental TADS with MIR199A2 showed a high survival score.***

b. The study by Song et al. [48]. shows how **MIR214 is a known tumor-suppressive microRNA in cervical cancer**. A study showed how SPINT1-AS1, a non-coding RNA, and MIR214 counter-attack each other, which decides the fate of cervical cancer cells. **MIR214 power over SPINT1-AS1 suppresses cervical cancer cell growth.** ***This validates our results and hypothesis as to why TAD with MIR214 shows a high survival score.***

c. The study by Li et al. [49] showed how **METTL13 overexpression increases cisplatin sensitivity of cervical cancer cells** and further suppresses the growth of cervical cancer cells. They explained how METTL13 diminishes the functioning of the Receptor for advanced glycation and products (RAGE) and thereby **protects cervical cells from becoming cancerous**. ***This validates our results of why METTL13 expression in TAD leads to high Survival in CESC cell lines.***

d. The same study by Song et al. [48], which **showed MIR214 role in cancer suppression, also explains how DNM3OS is a transcript of MIR214** and how **SP1NT1-AS1 blocks cleavage of DNM3OS which in turn represses MIR214 bio-genesis** activating cervical tumor growth**. *Hence it somehow validates why our Bayesian network shows the impact of MIR214 on DNM3OS.***

e. The study by Jing Fa[50] shows how DNM3 plays an important role in cervical cancer tumor repression. **Low expression of DNM3 is directly associated with a high grade of cervical cancer.** The authors showed how **over-expression of DNM3 in cervical cancer patients inhibited the invasion and growth of cervical cancer cells**. This also somehow explains why our Bayesian network does show activating of DNM3 by METTL13 and MIR214. Activating by **METTL13 and MIR214** (both of which suppress CESC cancer cells) also **hampers cervical cancer growth by further activating DNM3**. Hence ***this study does support our Bayesian Model network of genes.***

Figure 16 shows how all Genes in TAD chr1_171750000_172350000 interplay with each other and how they suppress cervical tumor growth by targeting various pathways. It also shows combinatorics Survival Score.

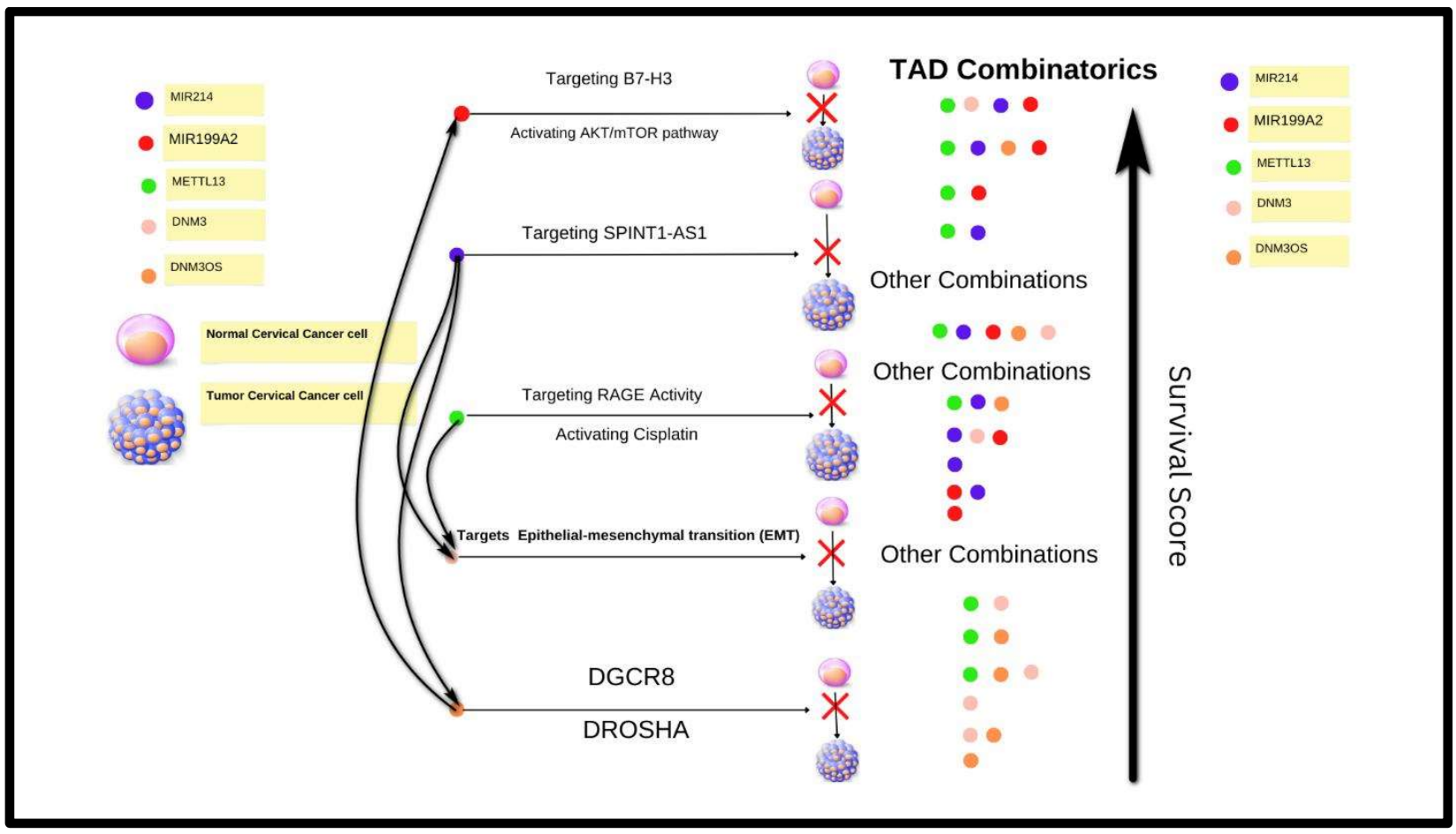
MIR214
MIR199A2
METTL13
DNM3
DNM3OS
Normal Cervical Cancer cell
Tumor Cervical Cancer cell
Targeting B7-H3
Activating AKT/mTOR pathway
Targeting SPINT1-AS1
Targeting RAGE Activity
Activating Cisplatin
Targets Epithelial-mesenchymal transition (EMT)
DGCR8
DROSHA
TAD Combinatorics
Other Combinations
Other Combinations
Other Combinations
MIR214
MIR199A2
METTL13
DNM3
DNM3OS
Survival Score

Figure 16

Chapter 2

Utilizing chromatin domain architecture and highlighting relevant enhancers and their target

Brief Introduction:

Chapter 2 focuses on how the prediction of chromatin interactions is being made by using only RNA-seq data of mRNA and enhancers in a cancer-specific manner. Chromatin interactions were predicted from R package scEChiA[1] developed by Dr. Vibhor Kumar's lab using RNA seq data of enhancers and genes. These predictions were computed based on various algorithms and designed Methodology. These predictions were benchmarked against the 4D genome database[2] and Activity by Contact[3] chromatin interaction database. These benchmarked results were analyzed for different categories giving insights into how good and reliable scEChiA interactions predictions are and how these interaction databases could be used as the base for genomic editing technologies to treat cancer patients.

Background and Related Work:

Chromatin interactions play a vital role in determining the expression level of genes and, ultimately, cell fate. Studying these interactions is of paramount importance, especially in cases of diseased (cancer) cells, since it gives insights into the genomic etiology of the disease.

Most of the studies predicting chromatin interactions are built from chromosome conformation techniques like 3C,4C,5C, and HiC, but recent advances in genomic technologies like Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) have provided more precise chromatin interactions.

Most of the chromatin interaction studies in the context of enhancer-gene interactions are on the basis of the physical distance between erna and mrna. The assumption that enhancers only target genes that are in close proximity is not ground truth anymore. Hence these studies do not represent genome-wide erna-mRNA interactions.

But a recent technique by Fulco et al. [3]. is known as CRISPRi-FlowFISH, which maps enhancers directly through perturbation of gene function. This technique predicts enhancer-gene interactions on the basis of two factors, one being the approximate contact frequency between enhancer and promoter and the second factor being the enhancer activity. Combining these factors, the authors proposed Activity by Contact Score for all interactions. In our work, we have used the scEChiA R package, which predicts chromatin interactions using RNA-seq data, and benchmarked our output interactions with the 4D genome and Activity by contact database to make technical and biological conclusions.

Theory and Methodology Used:

An overview of the entire Methodology followed has been described in Figure B.

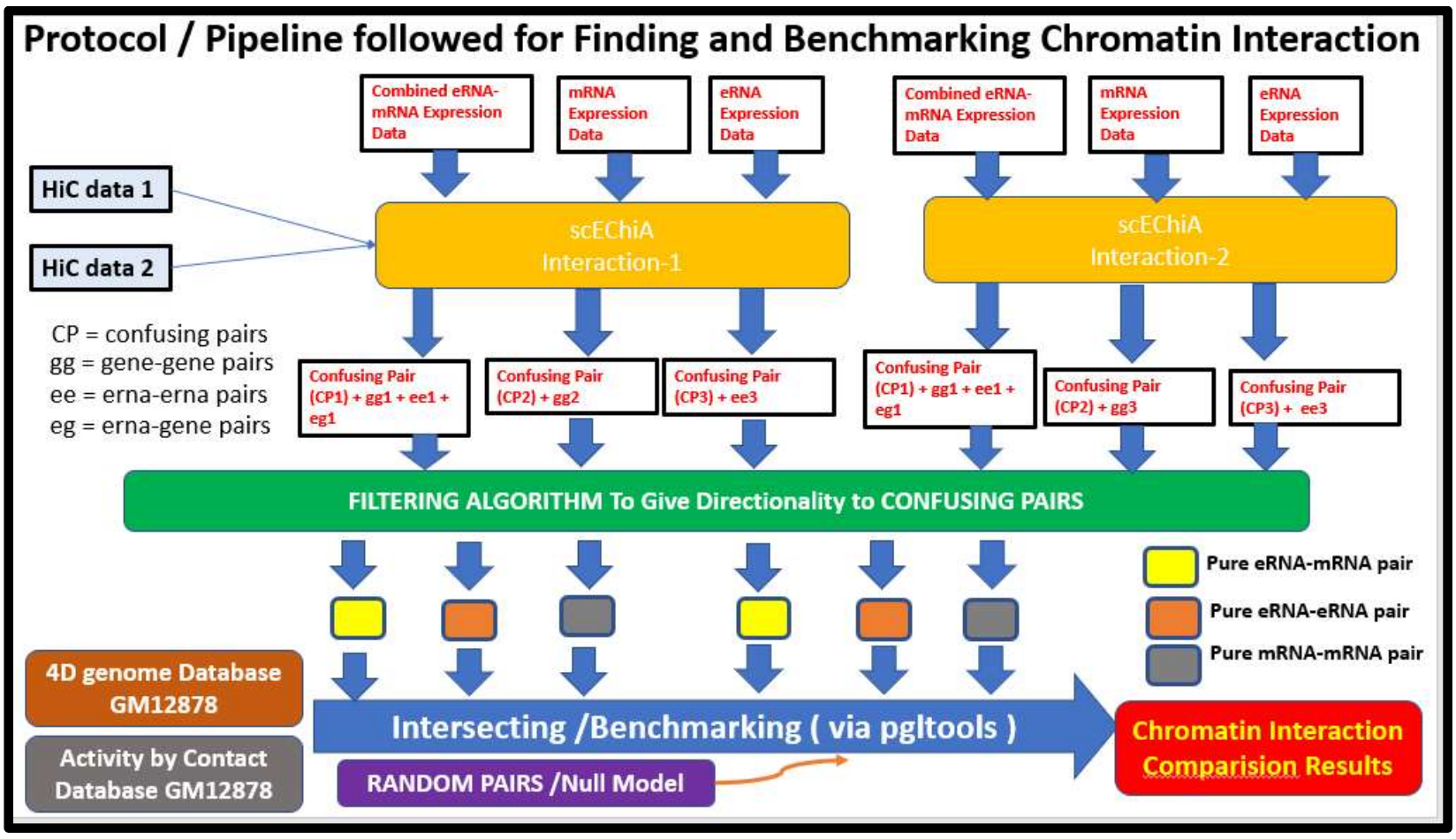

Figure B

1. Working of scEChiA R package:

We are going to use scEChiA for our study to find chromatin interactions; hence it is important to understand the basics of how scEChiA works.

The scEChiA R package primarily computes chromatin interactions based on its two defined functions, "**Interaction_Prediction_1**" and "**Interaction_Prediction_2,**" whose functioning has been described in detail in the subsequent section

For both functions, there are some common input variables:

- Both take expression data sets where the first three columns depict chromosome number, chromosome start, and chromosome end

position. After the third column, expression data for each sample is present.

- Both functions require a user-defined bin size upon which it decides the range of base-pair for each set of chromatin parts that are interacting.
- Both functions require user-defined start and end sample points upon which it calculates interactions through a series of algorithms.

Difference between two functions:

- Interaction_1 takes into account HiC profiles as background information to decide on the penalty term (described in subsequent sections). For computing the average of HiC files, the rhomatAvg function is used, which computes the average of two different HiC files. Results of rhomatAvg are used in Interaction_1 as an input parameter which represents background information as a HiC matrix, and it helps decide the penalty term Rho(p) for the function.
- Interaction_2 function takes a parameter 'Rho constant' which is defined by the user, and it acts like a constant penalty for all predicted interactions.

Mathematical algorithms upon which scEChiA predictions are based:

- **Merging of read-count peaks**: It firstly divides genomic data into user-defined bin sizes and then merges peaks from the read-count matrix

provided, after which log transformation of the newly formed count matrix is taken.

- **Gaussian Graphical Model**: In 2008, Friedman et al. [51] described a technique called a graphical lasso, which estimates regularized covariance matrix and its inverse (which is used for partial correlation calculation). The main Advantage of using a graphical lasso is that for the low association of variables, it reduces the partial correlation between peak pairs by introducing a penalty term rho.
- **Penalty term rho (p):** It forms the very core structure of mathematics that is used in scEChiA. Penalty terms are in the form of a matrix, where each row of the penalty matrix is the penalty for each predicted set of chromatin interactions. Penalty term usage in two functions (interaction_1 and interaction_2) is different. For the Interaction_1 function penalty term is a matrix, where it is computed from a formula as shown in Figure 17 below.

$$p_{ij} = \frac{\delta}{h_{ij}+\varepsilon}$$

Figure 17 (Credits: Pandey et al. [1])

Here h_{ij} represents the average enrichment level of chromatin interaction between genomic bins approximated by using background HiC data.

ε Represents a false count, which halts the penalty terms increase when there is no chromatin interaction found. δ is a variable parameter that decides the trade-off degree between the number of predicted interactions and accuracy.

- **Matrix Factorization method:** After the penalty term has been applied, we use the matrix factorization method, which gives a better estimation of co-

occurrence. It gives a better estimation of how strongly are our predicted interactions correlated.

2. Preparing and pre-processing the Data:

We collected gene expression data across all patients (samples) across all cancers from the TCGA portal. This was available in GEO accession no GSM1536837 Andrea Bild's Lab, University of Utah. We then collected enhancer (eRNA) quantification data from the MD-Anderson Bioinformatics portal for all TCGA patients for all cancers. Both expression data of mRNA and erna also contained information on their location (chromosome no, start position, and end position)

We then merged both genes (mRNA) expression data with eRNA expression data (along with their positions) for all patients (column-wise) and **built mrna & erna concatenated matrix** (patient column-wise) for each cancer type.

Here, since for eRNA only the start position was provided and not the end position, hence we kept the enhancer-end position as (eRNA start + 1600 base pairs) since eRNA can lie range from -800 to +800 bp in length.

Here we took TCGA- DLBC (Lymphoid Neoplasm Diffuse Large B-cell Lymphoma) cancer type in our study. Hence, wherever there is no mention of the cancer type, consider it TCGA-DLBC.

Please note that we took **DLBC chromosome number 11** for our analysis scEChiA performs or computes chromatin interaction for one chromosome (Intra-chromosome only).

3. Concatenated matrix of mRNA-erna to scEChiA R package:

Once concatenated matrix was made, we took it as input for various functions in the scEChiA R package. scEChiA calculates chromatin interactions by two methods,

a. One Method function is defined as Interaction_2, which takes a constant penalty rho(p) for every predicted interaction.
b. Another method function is defined as Interaction_1, which takes background information of two different HiC data files to estimate the penalty for each set of chromatin interaction sets.

4. Input to Interaction-2 function:

Once we have prepared the data, we input our data into interaction-2 for different bin sizes. Here we chose bin sizes like 2kb,10kb,25kb, and 30kb. For each bin size, our constant Rho (p) or penalty score was kept constant at 0.1.

- Since we were more interested in finding enhancer-gene interactions along with enhancer-enhancer and gene-gene interactions, as our Data is a concatenated matrix of mRNA and enhancers, we ran scEChiA Interaction_2 function first for the complete data set (mRNA + enhancers expression data) then we ran it for only mRNA expression data and then we ran it only for enhancer expression data.
- It was done so that we get true pairs of gene-gene interactions, true pairs of enhancer-enhancer interactions, and true pairs of enhancer-gene

interactions. We also did this deliberately because of the fact that when the complete matrix is run, we get some **CONFUSING PAIRS** (pairs which do not give a clear impression that they are true eRNA-eRNA pair or true mRNA-mRNA pair or true eRNA-mRNA pair).

- Hence to determine where these CONFUSING PAIRS actually belong (either to eRNA-eRNA or to eRNA-mRNA or to mRNA-mRNA category), we further developed a **FILTERING ALGORITHM** that can filter these confusing pairs. Some confusing pairs (< 5 %) remain in confusing pairs; hence we remove these pairs for benchmarking in further analysis.

5. Input to Interaction_1 function:

Same as mentioned above in point 4, we took similar bin sizes classes of 2kb,10kb,25kb, and 30kb for interaction_1 input. Here we took two different HiC profiles. One HiC profile we used was **IMR90** which is fibroblast from lung tissue. We took IMR90 from **4D Nucleome Data portal**[52] (4DN portal), Accession no 4DNFIH7TH4MF from study and experiments by Rao et al. [53]. Another HiC profile data was of k562, which are lymphoblast cells obtained from bone marrow. Accession, no of k562 HiC from the 4DN portal, is 4DNFITUOMFUQ from study and experiments by Rao et al. [53]

- As explained in point 4, we performed a similar procedure in the interaction_1 function also by predicting interactions for the complete dataset (erna + mRNA), only the eRNA data set and only the mRNA data set.

- For getting chromatin interactions from HiC data for different resolutions, we used the straw R package from Aidenlab[54]. It gives a data frame with HiC chromatin interaction data.
- A similar **FILTERING ALGORITHM** was applied to sort or give directionality to **CONFUSING PAIRS**.

6. Sorting out scEChiA prediction Results:

Once we got predicted interactions from both functions (Interaction_1 and Interaction_2) for different bin sizes, we benchmarked our results with available chromatin interaction data from 4D genome webserver[2] and Activity by Contact GM12878 Chromatin Interaction database[55]. 4D genome web server contains chromatin interactions for different cell types. It also contains chromatin interactions obtained from various techniques like 3C, 4C, 5C, Capture-C, HiC, ChIA-PET, and IM-PET. We filtered out chromatin interactions for the GM12878 cell type, which is for B-lymphocyte in Blood tissue. It was chosen because our chromatin interactions were done from the DLBC cell line, which is Lymphoid Neoplasm Diffuse Large B-cell Lymphoma is also from blood tissue. Activity by contact (ABC) database by Joseph Nasser et al. [55] gives chromatin predictions sorted out by three scores – Activity, HiC contact, and ABC score. We filtered out those interactions which had ABC score > 0. There we over 1 million ABC interactions.

- Benchmarking or comparing chromatin interactions was done using the python library 'pgltools' by Greenwald et al. [56]. Pgltools compare two chromatin interaction datasets and find the intersection of the two.
- Before intersecting results from GPL tools, chromatin interaction output from scEChiA was sorted out by three algorithms. The first algorithm was

used to make sure that the start position was less than the end location in both loci of chromatin interaction. The second algorithm was used to make sure that locus A comes before locus B in our data set. The third and last algorithm was used to make sure that the end position of locus A was less than the end position of locus B. These three algorithms were used on both data sets (data set out of scEChiA and data set from the 4D genome).

- After this, both the data set (data set out of scEChiA and data set from the 4D genome) were sorted according to the protocol mentioned by pgltools python library command- 'pygl.sort(data)'.

7. Filtering Algorithm:

The basic Methodology followed in the Filtering algorithm is shown in Figure C

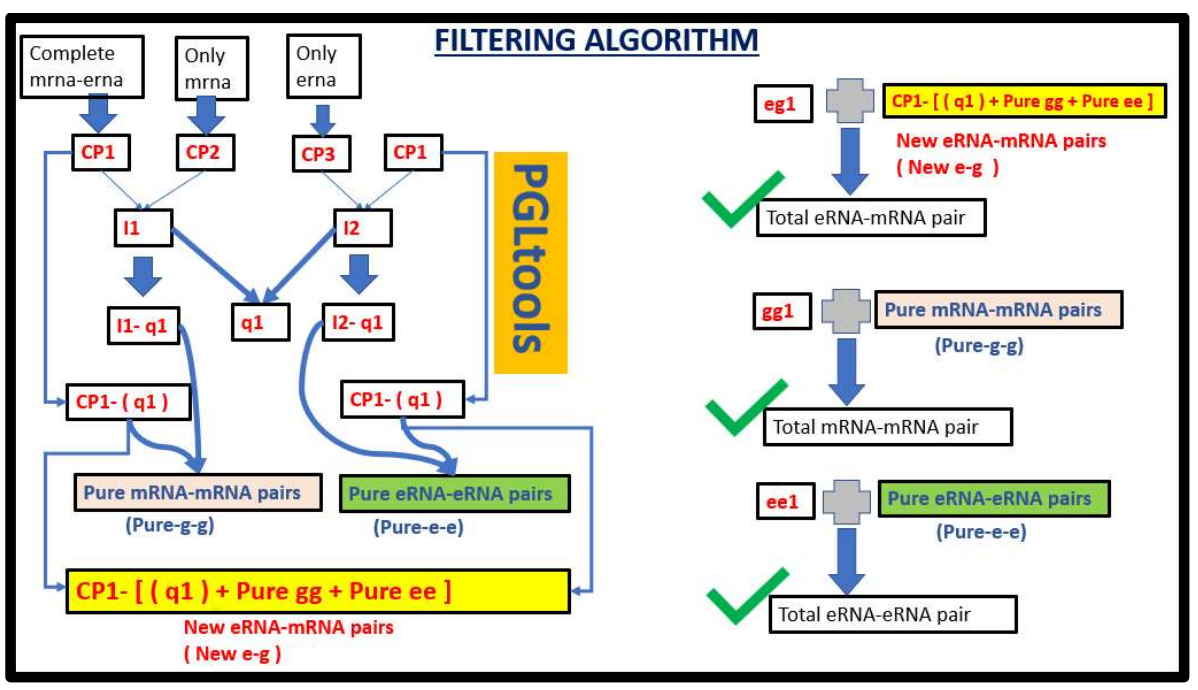

Figure C

- As explained in points 4 and point 5, scEChiA output does give confusing pairs when input is complete (mRNA-erna expression data). Confusing pairs arise irrespective of the fact that it was the interaction_1 function

used or the Interaction_2 function used. To know exactly whether the confusing pair is an erna-erna or mRNA-mRNA, or erna-mRNA interaction, a FILTERING algorithm was applied.

- The filtering algorithm is based on very simple observation. The observation is that when only the mrna input matrix is passed in scEChiA, it is bound to give only mRNA-mRNA interactions, and when only the erna input matrix is passed, it is bound to give only erna-erna interactions. Hence by this concept, every Confusing pair arising from only the mRNA input matrix (CP2) and Confusing pair arising from only the erna input matrix (CP3) are actually in ground truth only mRNA-mRNA pairs (for CP2) and erna-erna pairs (for CP3). Basically, it means that CP2 and CP3 representing confusing pairs are False positives.
- So if we intersect CP1 with CP2 (via pgltools), the output will give expected gg (mRNA-mRNA) pairs (l1), and when we intersect CP1 with CP3, the output will give expected ee (erna-erna) pairs (l2). Hence from CP1, we extracted some expected gg and ee pairs. Hence CP1 pairs got some directionality.
- Now it is possible that gg pairs (l1) and ee pairs (l2) have some intersection again (we call it q1 here as shown in Figure C). These intersections are confirmed confusing pairs which can't be labeled as pure gg or pure ee pairs. Hence we remove them from both expected gg pairs (l1) and expected ee pairs (l2). Hence further intersection of l1 and l2 does not give any result. **q1 thereby represents our algorithm error rate.**
- An important step is to remove these q1 interactions from CP1 data as well (CP1-q1) because this q1 will not get directionality further.

- These (CP1-q1) pairs, when intersected with (I1-q1), will give New Pure gg pairs (Pure-g-g), and the intersection of (CP1-q1) with (I2-q1) will give New Pure ee pairs (Pure-e-e).
- Now we have got new pure gg and ee pairs from CP1. Also, we have found out that those pairs (q1) which will never be able to get directionality. Hence the simple observation that CP1 was derived from Complete erna-mRNA data set implies that pairs not coming in q1 or in new gg pairs or in new ee pairs are actually erna-mRNA interactions (New e-g). This observation gives us new erna-mRNA pairs.
- Hence the majority of our CP1 confusing pairs have now been given directionality except q1 pairs, which still remain confused.
- Finally, we merge eg1 with New e-g pairs to get overall erna-mRNA pairs. Merging gg1 with Pure g-g gave us overall mRNA-mRNA pairs, and merging ee1 with Pure e-e gave us overall erna-erna pairs. These pairs are ready to be analyzed and benchmarked.
- For four bin sizes (30kb,25kb,10kb,2kb) and for each category (erna-erna, mRNA-mRNA, erna-mRNA), we got overall 12 such chromatin interaction pairs for chromatin interaction function_2. For Chromatin interaction_function_1, since the strawR package was only able to process HiC files at 10kb and 5kb resolution hence, we had six total files.

8. Random Interaction Pairs (Null-Model) Generation:

- Before arriving at any final conclusion, it is always advisable to build and benchmark random pairs (or null models) as well, as it represents a better idea of whether our predictions are true positives or not.
- For every 12 files (as explained in point 7 above) in chromatin interactions, we created a random interactions dataset of the same dimension as the concerned file so as to benchmark it. Random data set was created by using the R commands 'expand.grid()' and 'sample()' functions. Hence for the chromatin interaction_2 function, we got 24 files to be benchmarked, and for the chromatin interaction_1 function, we got 12 files to be benchmarked.

9. Benchmarking of scEChiA chromatin interaction results

- Finally, the output chromatin interaction data set out of scEChiA was intersected with the 4D genome chromatin interaction data set and with Activity by Contact GM12878 chromatin interaction database.
- Benchmarking like intersecting of chromatin interactions was done by pgltools command' pygl. intersect2d()'.
- Benchmark results were plotted in a bar plot in a comparative manner so as to understand our results and hypothesis in a better way.

RESULTS:

1. Firstly, chromatin interactions from scEChiA function 2 for different bin sizes along with random pairs of different bin sizes of the same dimensions as of interaction_2 function output for each category (erna-erna, erna-mRNA, and mRNA-mRNA) were intersected with the database of 4D genome GM12878. Its result is shown in Figure 18.

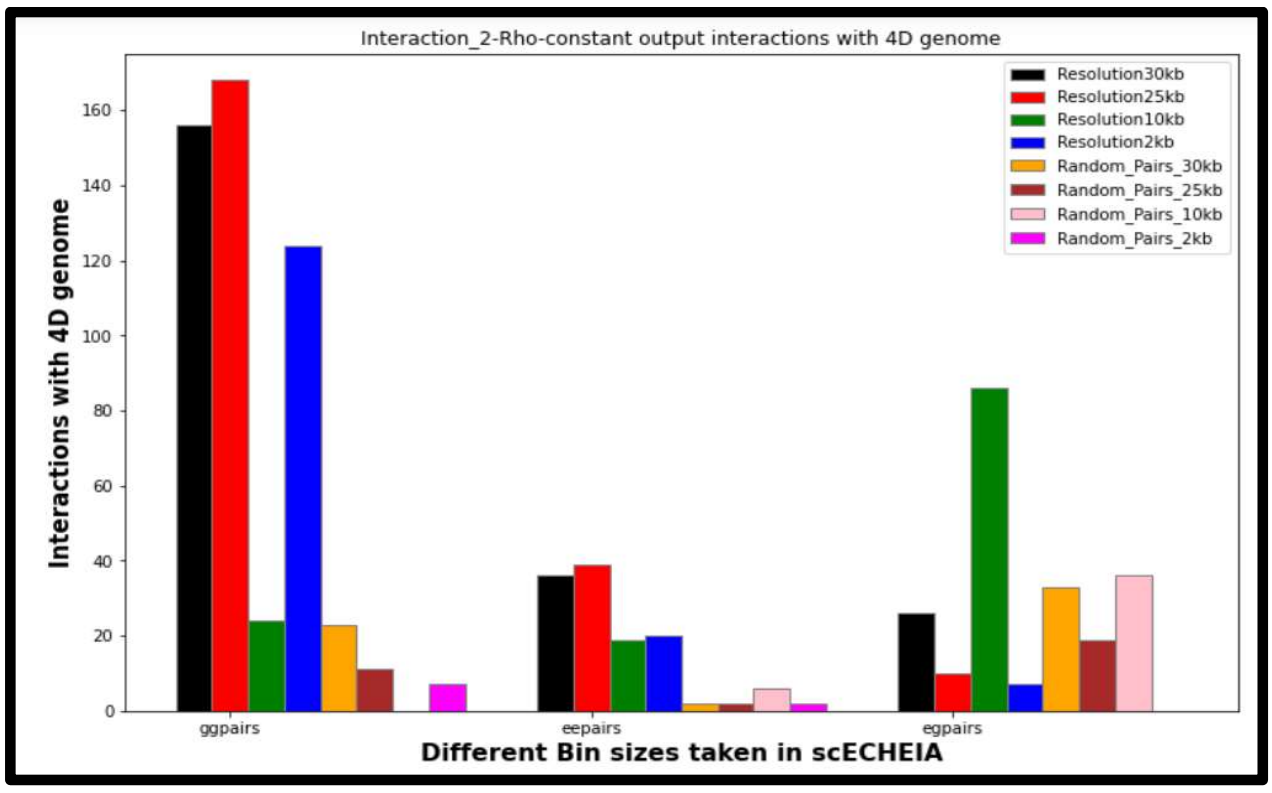

Figure 18

2. Similarly, as done in the previous step for interaction_2 output results, a similar Methodology was applied to interaction_1 output results. Figure 19 shows comparative interaction results.

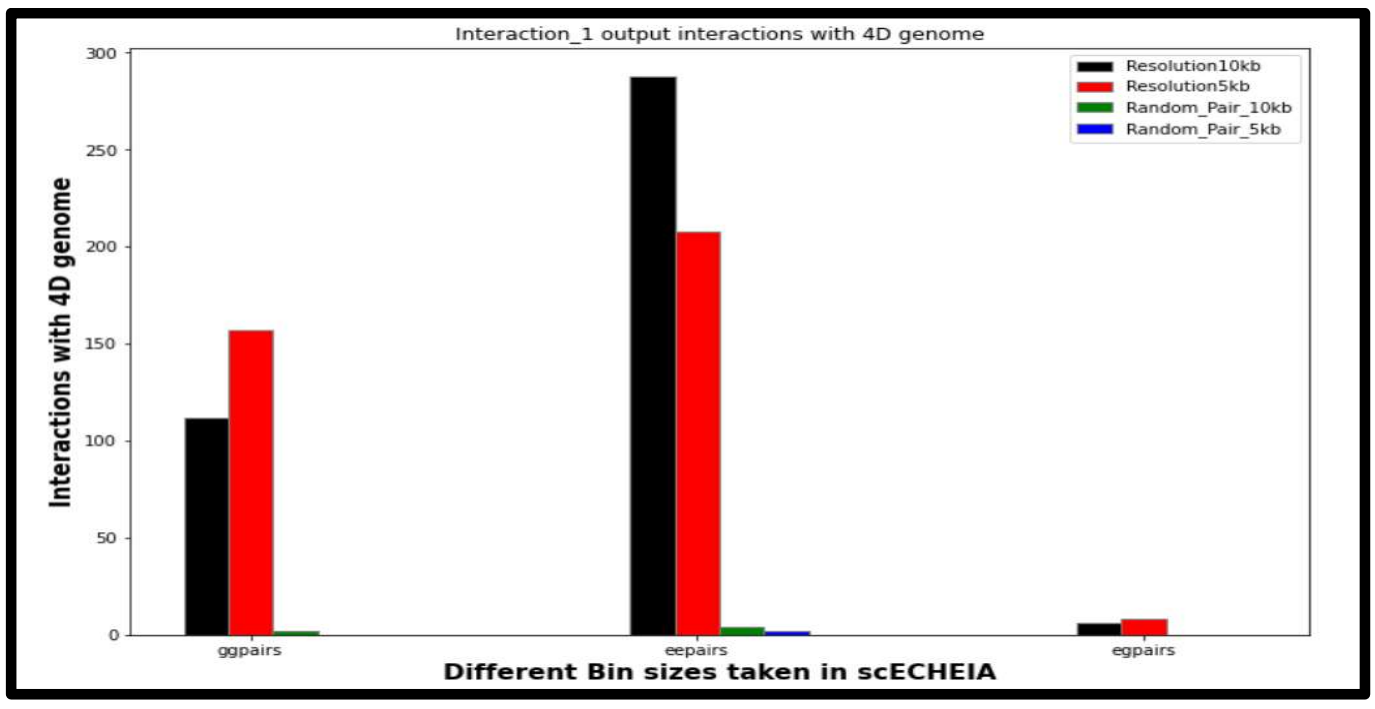

Figure 19

3. Repeating the previously followed Methodology, we did it for interaction_2 output results, but this time we changed the benchmarking base to Activity by

contact predictions for GM12878. Figure 20 shows the intersected results for the same.

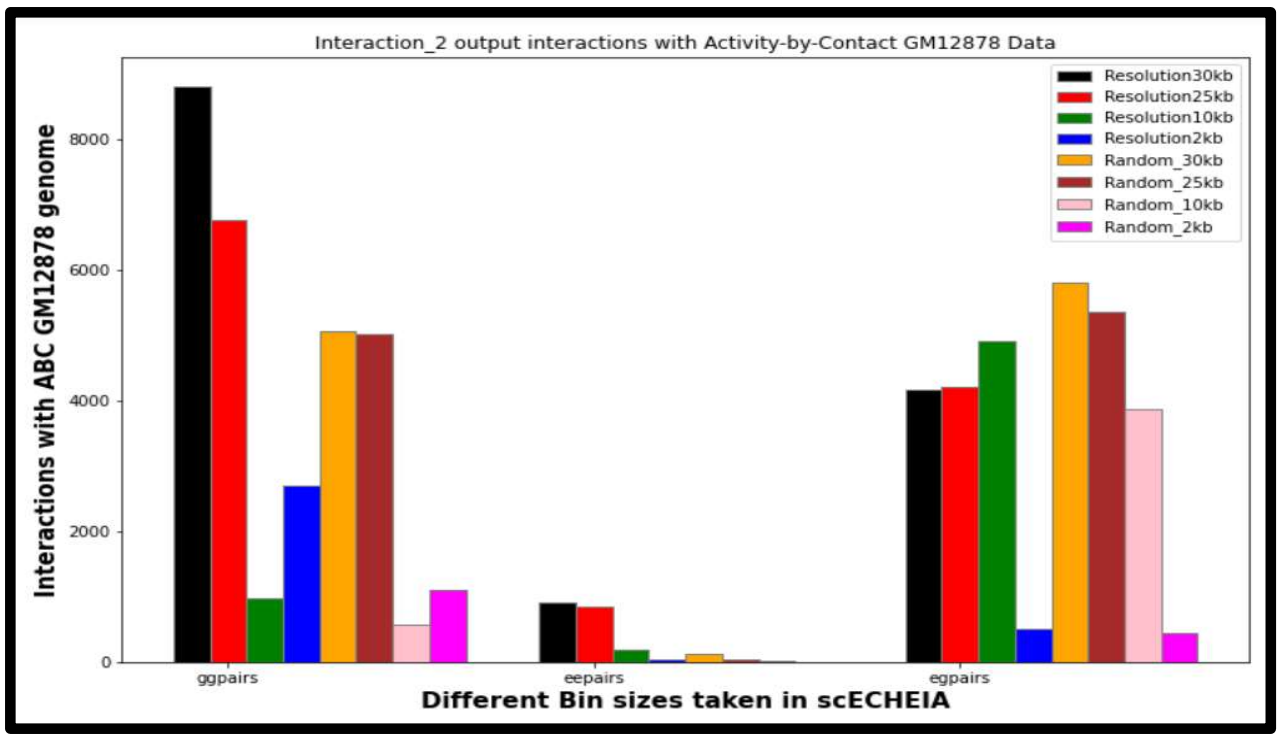

Figure 20

4. Benchmarking of Interaction_1 results was also done with the base as Activity by contact GM12878 data. Figure 21 shows the results.

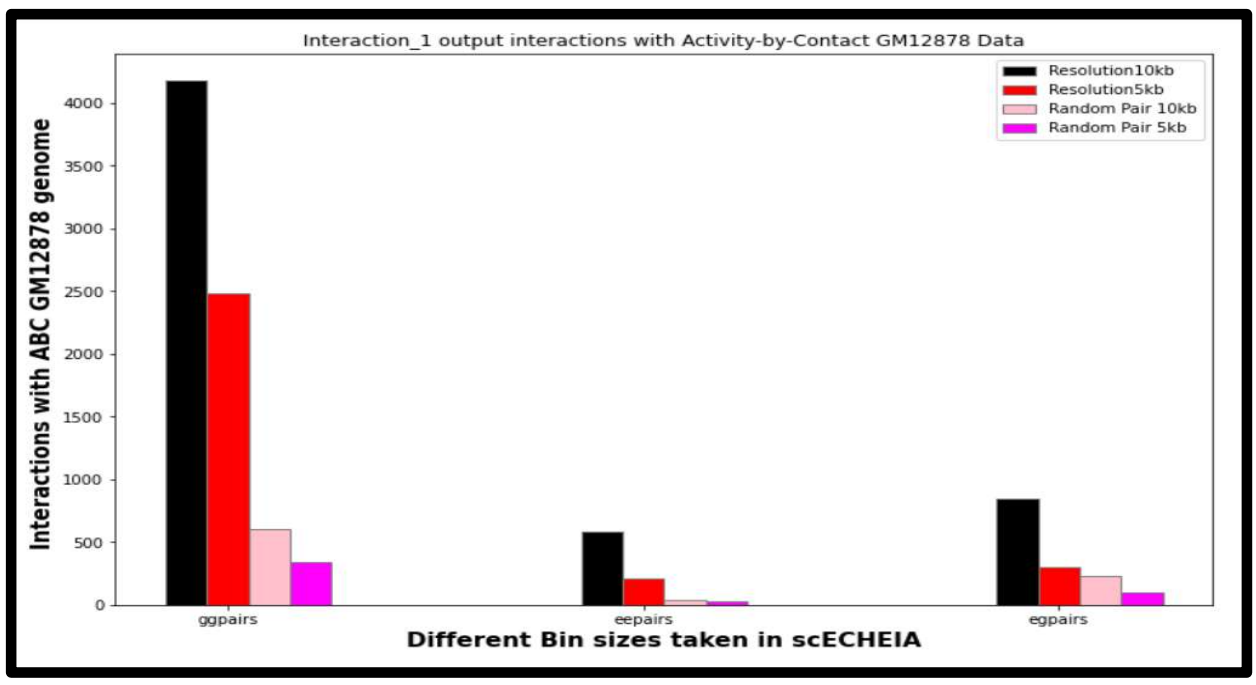

Figure 21

Observation and Inferences:

1. Clearly, when the Activity by Contact database is used, the number of intersected interactions increases exponentially when compared with the 4D genome database.
2. If we see for interaction_2 function, then for gg pairs average increase in intersection comes out to be around 40 times. While for ee pairs, it is nearly 14 times, and for eg pairs, it is near to 181 times.
3. If we see for interaction_1 function, then for gg pairs, there was a 26 times increase; for ee pairs, it was a 1.5 times increase, and for eg pairs, the increase was 88 times.
4. This abnormal exponential increase can be explained by the fact that Activity by contact had over 1 million predicted interactions, while the 4D genome had only about 5K interactions.
5. For Interaction_2 function
    - From Figure 18 and Figure 20, we can observe that for interaction_2 function output, for gg and ee pairs (almost), scEChiA seems to be more optimal for a 25kb bin size. While, eg, the pair's most optimal result is at 10kb resolution.
    - Fig 18 and Fig 20 also show that random pairs intersections for gg and ee pairs give satisfactory results as random pairs for particular bin sizes are always less than their scEChiA output data having the same bin size. This shows that there are fewer false positives in gg pairs results of the interaction_2 function.
    - For eg, in pairs, the interaction_2 function does give false positives, as we can observe in figure 18 and figure 20 that for eg, in pairs at low resolution (30kb,25kb), random (null) models outperform actual scEChiA output data of the same bin sizes. But at high resolution (10kb,2kb), there are fewer false positives as null models do not outperform actual results.

6. For Interaction_1_function

- From Fig 19 and Fig 21, Random pairs (or null models) do not outperform actual scEChiA output intersections with the same bin sizes. This shows that there is a very high showcase that our results in the interaction_1 function for all bin sizes have the least false positives and can be trusted. This is also true for all three categories like of gg or ee or, eg, pairs.
- Also, if we compare interaction_1 result with interaction_2 results across bin sizes and categories, there is one common observation that when the interaction_1 function is used, there is a significant difference between actual intersections and random model intersections compared to Difference seen in interaction_2 function results. This gives a very strong impression that interaction_1 result are more trustworthy since false positives are low.

7. The accuracy of our filtering algorithm drops a little bit in the Interaction_1 function compared to the interaction_2 function for all bin sizes. This might be expected due to the fact that interaction_1 does take into account the HiC data of two different cell types while calculating the penalty for each row of prediction interactions, while the interaction_2 penalty term is constant for each row (given by the user). Hence there are more chances that predicted chromatin interactions would be of confusing pairs that arise in interaction_1 that are found in the q1 category (shown in Figure C- filtering algorithm), which are basically related to both mRNA-mRNA and erna-erna category. This is because of the fact that background information of different cell types is also used to infer chromatin interaction of entirely different cell type

Conclusion & Future Possibility

Chapter 1

Exploiting TADs activity for predicting cancer survival

Conclusion:

By forming a novel TAD-gene set and via its GSVA scores, **we are able to get TAD prognostic score** (survival p-value). Hence, we were able to **identify top prognostic TAD** for 20 TCGA cancers. Comparing those TAD prognostic scores in all cancers **gives us PAN-CANCER analysis of prognostic TAD,** which may be helpful in novel drug design that can target multiple cancer by attacking common top prognostic TAD. Through passing Top prognostic TADs into our curated pipeline of various algorithms of Random Forest and Bayesian Analysis, we were **able to find the most statistically significant TAD** that is most prognostic considering the Bayesian network score along with its survival-p-value. For ***TCGA CESC cancer, we identified chr1_171750000_172350000 as our TAD of Interest.***

We then computed our TAD of Interest genes combinatorics, which gave 32 combinations, and we computed GSVA and subsequent survival analysis on this 32 combinatorics. Further Bayesian networking of our TAD of Interest genes was computed on the basis of their GSVA score. **Comparing 32 combinatorics survival scores with gene's individual and independent survival scores along with Bayesian networking of TAD of Interest genes, we were able to explain the observed gene's Bayesian networking pattern** by cross-validating it with combinatorics survival score coupled with the gene's independent survival score.

TAD of Interest Genes **METTL13, MIR214, DNM3, DNM3OS, and MIR199A2** in relation to cervical cancer **were found to be highly significant** through various published studies.

That **validated our Combinatorics survival score trend** as to why when all genes appear together, the TAD becomes a highly surviving TAD with the lowest survival p-value. We finally compiled these results in **Figure 16 to visualize how TAD_*chr1_171750000_172350000* is the most surviving TAD in cervical cancer**, with mentioning each gene co-networking along with their individual pathway to combat cervical tumor cells, help TAD of Interest become the best surviving TAD.

This showed how genes lying within a TAD have a strong correlation network among themselves, and these interactions, along with their individual behavior, play an important role in determining TAD functional property, specifically its prognostic score.

We also would like to mention that our results, observations, and conclusions contradict with findings of Helen S long et al. [57] ***that genes lying in the same TAD have less functional correlation among themselves.***

Future Possibility of our work:

Our work gives a pipeline to help researchers identify the most significant prognostic TAD for each different cancer. Further, it **explains TAD genes networking** among themselves and coupled with their individual pathways, **it sheds light on TAD biology** and **explains its survival pattern concerning cancer type**. This **can help researchers develop drug targets for the TADs** based on their survival score and TAD biology as an output from our work.

Chapter 2

Utilizing chromatin domain architecture and highlighting relevant enhancers and their target

Conclusion:

As mentioned above, earlier techniques to determine chromatin interaction do have a limitation in that they detect these interactions within a distance of 20kb. ***Our results from scEChiA predict chromatin interactions well beyond the limit of 20kb giving a genome-wide view of chromatin interactions.*** There were a good number of intersected interactions with the 4D genome database, while there was a ***significantly high number of chromatin interactions with Activity by Contact GM12878 interaction*** database, suggesting that our scEChiA model gives interactions on the same line as what Activity by contact model predicts. ***ScEChiA interaction_1 result which used HiC as background data to decide dynamic penalty terms, give more trustworthy or less false positive results than the interaction_2 function***, which uses constant penalty. The assumption of enhancer size can dramatically change the output structure of scEChiA and, subsequently, the number of intersected chromatin interactions. We took the assumption that enhancers are up to 1600 base pairs (bp) in length. Deciding on the length of enhancers or taking the different lengths of enhancers in the same scEChiA input matrix can impact the number and type of chromatin interactions.

Enhancer-gene interaction predictions from interaction-2 results for low resolution (30kb and 25kb) are not reliable, but for high resolution (10kb, 2kb), these enhancer-gene interactions give low false positives, so they are reliable. Consideration of HiC profiles used for background information in interaction-1 function also changes (although not very significantly) the nature and number of predicted chromatin interactions. Further usage of different combinations of HiC profiles also impact the

accuracy of the filtering algorithm. Varying Parameter **δ** (as explained in the theory and methodology section of chapter 2), which decides a number of predictions out of scEChiA at the cost of accuracy, can give more accurate or less false positives results from both interaction_1 and interaction_2 functions.

Future work & Possibility:

- ***We have proposed a novel method by which, only utilizing RNA-seq data, we can identify chromatin interactions in different categories like enhancer-enhancer or gene-gene or enhancer-gene interactions***. The predictions could be used directly as the base for CRISPR-based approaches to interested knockout pair of enhancers or genes for understanding genomics of cancer/diseases and ultimately arise as a genomic technology to treat patients (especially cancer patients)
- ***Combining our results and prognostic TAD from Chpater-1 together with scEChiA results, we can develop and understand chromatin interaction in greater depth, giving more insight into how chromatin interactions relate to diseases/cancer and in the context of TAD***, making CRISPR-based approaches more accurate to target specific interactions/TAD in cancer/cell-line specific manner.

Bibliography

1. Pandey, N., Omkar Chandra, Mishra, S. & Kumar, V. Improving Chromatin-Interaction Prediction Using Single-Cell Open-Chromatin Profiles and Making Insight Into the Cis-Regulatory Landscape of the Human Brain. *Front. Genet.* **12**, (2021).
2. Teng, L., He, B., Wang, J. & Tan, K. 4DGenome: a comprehensive database of chromatin interactions. *Bioinforma. Oxf. Engl.* **31**, 2560–2564 (2015).
3. Fulco, C. P. *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
4. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
5. Ibrahim, D. M. Three-dimensional chromatin in disease: What holds us together and what drives us apart? *Curr. Opin. Cell Biol.* 9 (2020).
6. Organization of Chromosomal DNA by SMC Complexes | Annual Review of Genetics. https://www.annualreviews.org/doi/10.1146/annurev-genet-112618-043633.
7. Dixon, J. R., Gorkin, D. U. & Ren, B. Chromatin Domains: the Unit of Chromosome Organization. *Mol. Cell* **62**, 668–680 (2016).
8. McArthur, E. & Capra, J. A. Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched for heritability. *Am. J. Hum. Genet.* **108**, 269–283 (2021).
9. Fudenberg, G. *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep.* **15**, 2038–2049 (2016).
10. The Three-Dimensional Organization of Mammalian Genomes | Annual Review of Cell and Developmental Biology. https://www.annualreviews.org/doi/10.1146/annurev-cellbio-100616-060531.

11. Jost, D., Vaillant, C. & Meister, P. Coupling 1D modifications and 3D nuclear organization: data, models and function. *Curr. Opin. Cell Biol.* **44**, 20–27 (2017).

12. Lupiáñez, D. G., Spielmann, M. & Mundlos, S. Breaking TADs: How Alterations of Chromatin Domains Result in Disease. *Trends Genet.* **32**, 225–237 (2016).

13. Muro, E. M., Ibn-Salem, J. & Andrade-Navarro, M. A. The distributions of protein coding genes within chromatin domains in relation to human disease. *Epigenetics Chromatin* **12**, 72 (2019).

14. Bintu, B. *et al.* Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* **362**, eaau1783 (2018).

15. Li, G. *et al.* ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biol.* **11**, R22 (2010).

16. Jodkowska, K. *et al.* Three-dimensional connectivity and chromatin environment mediate the activation efficiency of mammalian DNA replication origins. 644971 Preprint at https://doi.org/10.1101/644971 (2019).

17. Filippova, D., Patro, R., Duggal, G. & Kingsford, C. Identification of alternative topological domains in chromatin. *Algorithms Mol. Biol.* **9**, 14 (2014).

18. Gaffney, D. J. Mapping and predicting gene–enhancer interactions. *Nat. Genet.* **51**, 1662–1663 (2019).

19. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47**, 598–606 (2015).

20. Expression quantitative trait loci. *Wikipedia* (2022).

21. Nica, A. C. & Dermitzakis, E. T. Expression quantitative trait loci: present and future. *Philos. Trans. R. Soc. B Biol. Sci.* **368**, 20120362 (2013).

22. Li, L., Barth, N. K. H., Pilarsky, C. & Taher, L. The presence of copy number variants in specific topologically associating domains has prognostic value in many cancer types. 777573 Preprint at https://doi.org/10.1101/777573 (2019).
23. Eres, I. E. & Gilad, Y. A TAD Skeptic: Is 3D Genome Topology Conserved? *Trends Genet.* **37**, 216–223 (2021).
24. Liu, T. *et al.* TADKB: Family classification and a knowledge base of topologically associating domains. *BMC Genomics* **20**, 217 (2019).
25. BioSamples < EMBL-EBI. https://www.ebi.ac.uk/biosamples/samples/SAMN09741268.
26. BioSamples < EMBL-EBI. https://www.ebi.ac.uk/biosamples/samples/SAMN09740487.
27. BioSamples < EMBL-EBI. https://www.ebi.ac.uk/biosamples/samples/SAMN09740836.
28. GEO Accession viewer. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3313056.
29. bedtools: a powerful toolset for genome arithmetic — bedtools 2.30.0 documentation. https://bedtools.readthedocs.io/en/latest/index.html.
30. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics* **14**, 7 (2013).
31. Tustumi, F. Choosing the most appropriate cut-point for continuous variables. *Rev. Colégio Bras. Cir.* **49**, e20223346 (2022).
32. Youden, W. J. Index for rating diagnostic tests. *Cancer* **3**, 32–35 (1950).
33. Altman, D. G., Lausen, B., Sauerbrei, W. & Schumacher, M. Dangers of using 'optimal' cutpoints in the evaluation of prognostic factors. *J. Natl. Cancer Inst.* **86**, 829–835 (1994).
34. Lausen, B. & Schumacher, M. Maximally Selected Rank Statistics. *Biometrics* **48**, 73 (1992).
35. Miller, R. & Siegmund, D. Maximally Selected Chi Square Statistics. *Biometrics* **38**, 1011–1016 (1982).

36. Mazumdar, M. & Glassman, J. R. Categorizing a prognostic variable: review of methods, code for easy implementation and applications to decision-making about cancer treatments. *Stat. Med.* **19**, 113–132 (2000).
37. Altman, D. G. & Royston, P. The cost of dichotomising continuous variables. *BMJ* **332**, 1080 (2006).
38. pplonski. Answer to 'Does the optimal number of trees in a random forest depend on the number of predictors?' *Cross Validated* https://stats.stackexchange.com/a/474933 (2020).
39. Płoński, P. How many trees in the Random Forest? *MLJAR* https://mljar.com/blog/how-many-trees-in-random-forest/ (2020).
40. Ellis, C. Number of trees in random forests. *Crunching the Data* https://crunchingthedata.com/number-of-trees-in-random-forests/ (2022).
41. Oshiro, T. M., Perez, P. S. & Baranauskas, J. A. How Many Trees in a Random Forest? 15.
42. Scutari, M., Graafland, C. E. & Gutiérrez, J. M. Who Learns Better Bayesian Network Structures - Constraint-Based, Score-based or Hybrid Algorithms? 20.
43. Cowell, R. G. Conditions under which conditional independence and scoring methods lead to identical selection of Bayesian network models. in *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence* 91–97 (Morgan Kaufmann Publishers Inc., 2001).
44. Chapter 10. Local Search. https://docs.optaplanner.org/7.0.0.Beta1/optaplanner-docs/html/ch10.html.
45. Team. *OptaPlanner* https://www.optaplanner.org/community/team.html.
46. Scutari, M. Learning Bayesian Networks with the bnlearn R Package. Preprint at https://doi.org/10.48550/arXiv.0908.3817 (2010).
47. Yang, X., Feng, K.-X., Li, H., Wang, L. & Xia, H. MicroRNA-199a Inhibits Cell Proliferation, Migration, and Invasion and Activates AKT/mTOR Signaling Pathway by Targeting B7-H3 in Cervical Cancer. *Technol. Cancer Res. Treat.* **19**, 1533033820942245 (2020).

48. Song, H., Liu, Y., Liang, H., Jin, X. & Liu, L. SPINT1-AS1 Drives Cervical Cancer Progression via Repressing miR-214 Biogenesis. *Front. Cell Dev. Biol.* **9**, 691140 (2021).

49. Li, R. *et al.* METTL3 increases cisplatin chemosensitivity of cervical cancer cells via downregulation of the activity of RAGE. *Mol. Ther. Oncolytics* **22**, 245–255 (2021).

50. Fa, J. Dynamin 3 overexpression suppresses the proliferation, migration and invasion of cervical cancer cells. *Oncol. Lett.* **22**, 1–8 (2021).

51. Friedman, J., Hastie, T. & Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostat. Oxf. Engl.* **9**, 432–441 (2008).

52. the 4D Nucleome Network *et al.* The 4D nucleome project. *Nature* **549**, 219–226 (2017).

53. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).

54. Straw: rapidly stream data from .hic files. (2022).

55. OSF | Activity-by-Contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. https://osf.io/uhnb4/?view_only=.

56. Greenwald, W. W. *et al.* Pgltools: a genomic arithmetic tool suite for manipulation of Hi-C peak and other chromatin interaction data. *BMC Bioinformatics* **18**, 207 (2017).

57. Long, H. S. *et al.* Making sense of the linear genome, gene function and TADs. *Epigenetics Chromatin* **15**, 4 (2022).

Data Analysis and Software Used

For TAD related analysis, survival graphs and Bayesian network modeling, R was used with various libraries.

For Chromatin Interaction part of work, both R and Python (via Anaconda Service) was used to use various tools, libraries and generate graphs.

For compiling and citing authors, Zotero Software was used.

For checking plagiarism, Grammarly IIITD account provided by IIITD library was used.

Code and Data Availability

All the datasets used have been mentioned in the thesis methodology itself. Data-set and all used R plus Python scripts have been submitted to Dr. Vibhor Kumar's lab, as it is a part of a project that is in the pipeline to be sent to a Journal for Publication.

Any Data or Code/Scripts access request can be made to Dr. Vibhor Kumar on his email id vibhor@iiitd.ac.in

References

1. Eng K. H., Schiller E., Morrel K. On representing the prognostic value of continuous gene expression biomarkers with the restricted mean survival curve. Oncotarget. 2015; 6: 36308-36318. Retrieved from https://www.oncotarget.com/article/6121/text/

2. Brownlee J. Probabilistic Model Selection with AIC, BIC, and MDL. MachineLearningMastery.com. Published October 29, 2019. Accessed December 3, 2022. https://machinelearningmastery.com/probabilistic-model-selection-measures/

3. R Addict Blog. Accessed December 3, 2022. http://r-addict.com/2016/11/21/Optimal-Cutpoint-maxstat.html

4. Difference Between AIC and BIC | Difference Between. Accessed December 3, 2022. http://www.differencebetween.net/miscellaneous/difference-between-aic-and-bic/

5. Geiger D, Heckerman D. Learning Gaussian networks. In: *Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence*. UAI'94. Morgan Kaufmann Publishers Inc.; 1994:235-243.