# Signatures and Utility of Epigenetic Memory at Enhancers

By

**Rajat Talukdar**

**Under the supervision of Dr. Vibhor Kumar**

**Submitted in partial fulfilment of the requirements for the degree of Master of Technology in Computational Biology**

**Center for Computational Biology, Indraprastha Institute of Information Technology - Delhi May 2023**

# Certificate

This is to certify that the thesis titled *"Signatures and Utility of Epigenetic Memory at Enhancers,"* being submitted by **Rajat Talukdar** to the Indraprastha Institute of Information Technology Delhi for the award of Master of Technology in Computational Biology, 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.

May 2023

Dr. Vibhor Kumar

Department of Computational Biology
Indraprastha Institute of Information Technology Delhi
New Delhi 110 020

# Acknowledgments

I would like to express my sincere gratitude to my advisor, **Dr. Vibhor Kumar**, for his invaluable guidance, support, and encouragement throughout this journey. His expertise, insights, and feedbacks have been instrumental in shaping the direction of my research and helping me overcome the various challenges that I encountered.

Further, I would also like to thank the faculty members of the Department of Computational Biology, IIITD for their inspiring lectures, discussions, and feedback that have greatly enriched my knowledge and understanding of the subject matter.

Finally, I would like to extend my heartfelt thanks to my family and friends for their unwavering love, support, and motivation throughout my academic pursuits.

# Abstract

Epigenetic memory is a vital cellular process. It regulates the inheritance of certain efficient traits of normal cells and the traits attained lately by the cells affected by diseases like Cancer from parents to daughter cells. Understanding the epigenome profile and how the molecular basis of epigenetic memory governed by histone modifications and other epigenetic markers are erased and re-established during cellular processes such as embryogenesis and cell differentiation in the stem cells, somatic cells as well as disease cells will have a significant impact in a deeper understanding of cellular development and diseases such as Cancer. Here we try to comprehend the influence of distinct epigenetic marks, such as histone modifications and chromatin accessibility, in defining the chromatin state for having active or poised enhancers, which further influences cell type and physiological condition. In our study, we hypothesize that these epigenetic marks present in active enhancers in the past may not remain bound; however, there might be several residuals left, which influences the physiological condition of the Cell in the present. A profound understanding of how these epigenetic marks in enhancers affect the chromatin state would play a crucial role in advancing the prognostics and diagnostics of disease states and help the advancement of targeted therapeutics for diseases.

# Table of Contents

# List of Figures

# List of Tables

# Introduction

Epigenetics studies heritable alterations in cellular phenotype or gene expression that occur without any changes to the underpinning DNA pattern. Before the discovery of DNA-based genetic inheritance, Conrad Waddington invented the phrase “epigenetics”, which is derived from the Greek word “epigenesis” to connect genetics to the phenotype determinants necessary for the embryo’s development from an undifferentiated condition [1]. Epigenetic events are caused by covalent alteration of DNA and histones, the proteins protecting DNA. Therefore, these alterations control particular gene expression patterns and, in certain instances, may be duplicated and transferred to daughter cells [2].

Epigenetic memory is the persistence of epigenetic alterations that can affect gene expression and cellular function across numerous cell divisions and occasionally across generations [3]. In a way, epigenetic memory describes the traits of germ cells and usually developed cells. It creates hereditary traits in embryonic cells, their offspring, and cells defining a disease condition [2]. Epigenetic memory is crucial for both health and illness development and for organism environmental adaptation. For instance, some environmental variables, like diet, stress, and toxin exposure, can cause epigenetic changes that last for generations and impact offspring's well-being and disease propensity. Epigenetic memory has also been linked to forming cell identity and lineage commitment and preserving stem cell pluripotency and differentiation.

Overall, research on epigenetic memory has significant ramifications for comprehending the intricate interactions between heredity and environment and for creating novel disease prevention and treatment strategies.

# Chapter 1: Epigenetic Memory in Developmental Lineages

## 1.1 Epigenetics determine cell functions.

The epigenome comprises several epigenetic marks, such as an assortment of DNA methylation patterns particular to specific genes, unique combinations of transcription factors, non-coding RNAs, chromatin remodeling factors, polycomb group proteins, Histone post-translational modifications, and other epigenetic memory factors. A gene's expression status is assessed by the unique epigenome specifying the (epi)genomic code linked to that gene. Each gene may have an epigenomic code specific to a particular state of a disease or a cell type. As a result, similar genomes can accumulate various groups of epigenomic code to build the distinctive epigenome, which specifies the general cellular features [4].

In contrast to DNA, histone proteins are subject to several changes, such as acetylation, methylation, ubiquitylation, and phosphorylation. Depending on the type of alteration and the particular amino acid changed, these mutations can result in either gene silence or activation [5]. The nucleosome core particle, chromatin's fundamental structural component, comprises an octamer of histones forming two peripheral heterodimers of H2A and H2B histones, and 146 base pairs of DNA surround a core tetramer of H3 and H4 histones. Covalent modification of DNA by DNA methyl transferases (DNMTs), primarily in the presence of CpG dinucleotides, to produce 5- methylcytosine is associated with transcriptional silencing [6].

While DNA undergoes epigenetic changes due to cytosine methylation, histones with lysine and arginine residues are acetylated and methylated. In contrast, lysines are ubiquitylated and sumoylated, threonines and serines are phosphorylated, and the tails of glutamic acid are poly-ADP-ribosylated [7].

Furthermore, residues of arginines are mono and di-methylated, while lysine residues are mono, di, and tri-methylated, mounting to the intricateness of histone modifications.

Methylation of K4 of H3, i.e., H3K4me2 and H3K4me3, as well as H3K79 and H3K36 methylations and Histone acetylation, are linked to an increase in transcriptional activity, with H3K4 tri-methylation (H3K4me3) designating promoter. At the same time, H3K9me2 and H3K27me3 in the euchromatic area signify a suppressed state, and trimethylations of H3K9 or H4K20 generate heterochromatin [8]. Additionally,5- hydroxymethylcytosine (5-hmC), an oxidized version of 5-mC, has attracted much interest as an epigenetic mark that controls chromatin changes and gene transcriptionin embryonic phases, cellular differentiation, and various malignancies [9]. Overall, it has been discovered that these histone marks are related to the chromatin's and cells' functional condition.

## 1.2 Epigenetic landscape related to stem cell characteristics.

The transcription master regulators Nanog, Oct4, and Sox2 comprise most of the complex network that gives rise to pluripotency and stem cell identity. This network is also believed to possess a feedback system connected to some networks of cofactor-protein interactions [10]. In contrast to adult stem cells like mesenchymal stem cells, which can differentiate into adipocytes, osteocytes, etc., embryonic stem cells demonstrate the ability to differentiate into every cell type of the body, while mesenchymal stem cells exhibit the restricted potential to differentiate into only the parental organ's cell types, these stem cells would display different epigenetic signatures as a result of DMRs, histone changes, and interactions with transcription factors. By controlling the post-translational modification of histones, Protein structures called the Polycomb and Trithorax groups either prevent or promote transcription, respectively [11] [12]. Evidence for the existence of epigenetic cellular memory may include bivalency caused by Polycomb group proteins mediated silencing associated with H3K27me3 and Trithorax group protein-dependent activation indicated by the presence of H3K4me3 in the same loci and co-occupation by

regulatory transcription factors such as Sox2 and Oct4 Additionally to the environment and concurrent transcription factors; the various DNA methylation patterns may also influence this process.[13] Based on the stochastic extrinsic and intrinsic signals, they may continue to be prepared for either repression or induction. These modifications and associations may serve as the epigenetic gatekeepers for controlling genes associated with development in pluripotency and pluripotency. Numerous chromatin remodeling and pluripotency- related transcription factors, including Oct and Sox2, the Polycomb group or Trithorax proteins, and others, are inherited from the mother and are most likely linked to specific DNA sequences, which could enable the sequence of events to begin and advance instantly [14].

*Table 1.1: Major histone marks and their association with transcription or other roles.*

| *Histone mark* | Position |
|---|---|
| macroH2A | Compact chromatin, X chromosome (Repressed) |
| H2A.X | Double-strand DNA breaks |
| H2A.Z | Promoters/transcription start sites (Active) |
| H2BK5Ac | Promoters |
| H2BK120Ac | CpG island promoters |
| cenH3 (CENP- A) | Centromere |
| H3.3 | Promoters (Active) |
| H3K4Ac | Enhancers |
| H3K4Me1 | Promoters and enhancers |
| H3K4Me2 | Transcription start sites, CpG islands, promoters, and enhancers |
| H3K4Me3 | CpG islands, promoters, and enhancers (Active) |
| H3K9Ac | Coding regions (Active) |
| H3K9Me3 | Promoters and enhancers, heterochromatin (Repressed) |
| H3K14Ac | CpG islands, promoters, and enhancers (Active) |
| H3K27Ac | Coding regions (Active) |
| H3K27Me3 | Coding regions, heterochromatin (Repressed) |
| H3K36Me3 | Coding regions (Active) |
| H3K79Me1 | Coding regions (Active) |
| H3K79Me2 | Coding regions (Active) |
| H4K16Ac | Euchromatin (Active) |
| H4K20Me3 | Heterochromatin (Repressed) |

Reprogramming the encoded epigenetic memories during the development and differentiation process has opened up the possibility of putting one's cells in place of the damaged cells. Recent progress in this field could make it possible to convert or dedifferentiate developed somatic cells into pluripotent cells [15]. Similarly, it is simple to imagine how the dedifferentiation process brought on by the loss of epigenetic memories could produce the infamous cancer stem cells that can act as a pool for the disease's relapse.

In this study, we analyze the epigenetic profile involving various epigenetic marks such as histone modification marks, transcription factor, polycomb group proteins, chromatin modifiers, etc., in ESCs and K562 cells to detect potential epigenetic markers which can promote epigenetic memory during stem cell differentiation. Analyzing the epigenome of germ cells will reveal details about the biology of cells since combinatorial chromatin alterations can uncover cis-regulatory areas [16]. We also examined the epigenome of human embryonic stem cell enhancers and tracked their state in other progenitor cells, such as neuronal progenitor cells, mesenchymal cells, BMP4 mesendoderm cells, and BMP4 trophoblast cells, all derived from human embryonic stem cells. The hypothesis for this study is that enhancers that were active in the past may not remain bound by the transcription factor, but there might be multiple residual marks left. The levels of these enduring residual epigenetic memories are influenced by the cells from which they were derived, and the experimental techniques used to produce pluripotency. Determining how to deal with these enduring epigenetic memories will be a significant obstacle for researchers as they work to create stem cell-based therapy.

## 1.3 Chip-seq Analysis with Dfilter

Dfilter is a generalized signal identification program that uses linear filters to optimize ROC-AUC for next-generation massively parallel sequencing data analysis. Because of this, it is the perfect tool for spotting peaks in the tag profile of many sequencing methods, including ChIP-seq, DNase-seq, ATAC-seq, etc. Additionally, GC bias correction and read-count estimation on a specific collection of peaks can be done using DFilter.

DFilter has been developed to find enriched states and regulatory regions using the tag count information produced by next-gen sequencing. It has been produced utilizing a generalization technique to analyze data from many assays. Our research used Dfilter to analyze chip-seq data and create a spatial heat map of different epigenetic markers in stem cells and their lineages to investigate the chromatin state in active enhancers. Four types of input files are supported by the DFilter version. A bam or sam file, bedgraph, and bed format for raw tags files is also an option. Combining tag profiles from many assays to uncover regulatory elements is a unique feature of DFilter that makes it a more versatile tool. This feature of DFilter is useful when two or more libraries represent the same regulatory element, such as many histone acetylations at enhancers or several tests to signify open-chromatin. Spatial heat maps can help visualize ChIP-seq data, allowing you to see signal intensity distribution across a genome or a genomic region. Outline of the steps to generate a Spatial heat map using Dfilter are as follows:

1. **Peak Files** were generated by peak calling using the chip-seq data, a n d the input should contain filenames of a sample and control, containing tags from high throughput sequencing. The files can be in bam, bedgraph, or bed formats. Additional option '-pe' for paired-end bam or sam files should be added; otherwise, the tags will be treated as single-end reads in peak calling.
   *run_dfilter.sh -d=CHIPfiles.bed, -c=INPUTfiles.bed -o=OUTPUTPEAKfile.bed f=bed -ks=100 -lpval=6 -nonzero –wig*
   - -d=CHIPfiles.bed": specifies the input file for the ChIP data.
   - -c=INPUTfiles.bed": specifies the input file for the control data.
   - -o=OUTPUTPEAKfile.bed": specifies the output file for the filtered peaks.
   - -f=bed": specifies the input file format as BED.
   - -ks=100": specifies the kernel size for the smoothing function.
   - -lpval=6": specifies the log10 p-value cutoff for peak detection.
   - -nonzero": specifies that only nonzero signals should be considered.
   - -wig": specifies that a WIG file of the smoothed data should also be output.

2. **Normcore** was generated using the peak files in the above step. The normcore in ChIP-seq data analysis typically refers to the normalization of read counts or peak intensities across different samples or experimental conditions. Normalization is crucial because it helps to account for differences in sequencing depthor observed variability that can affect the interpretation of ChIP-seq data.
   *run_normscore.sh File_with_PeakFileNames* File_with_ChIP-FileNames*File_with_Control-Filenames(or NA)*genome* fileformat∗same/NAdivide/NAtagcount−window− size/NAzero −mean = auto/filterSize*
   - *File with PeakFileNames (compulsory) is the file name containing (list)names of files having peaks.*
   - *File with PeakFileNames (compulsory) is the file name containing (list)names of files having peaks.*
   - *File with Control-Filenames (optional) is the file name containing (list)of tag-files for control.*
   - *Genome (compulsory) the genome and its version such as hg19, hg18, mm9, etc.*
   - *Format (compulsory) of tag files bed/bam/sam.*
   - *Divide/NA (optional with NA) divide option is passed when ChipSeq- data hasto be divided by the tag density of control.*
   - *tag-count-window-size/NA (optional with NA)if the user needs tag count ina larger or smaller window than the default (1kb)*

*3.* **Clusters** were generated using K-means clustering by taking the union of peak files and normalized peak scores. K-means clustering can be used to group these peaks into clusters based on their signal values. The technique operates by repeatedly allocating each peak to the closest cluster centroid (i.e., the cluster's center), then updating the cluster centroids depending on the mean signal levels of the peaks in each cluster. The process is repeated until the cluster assignments stop changing or the required number of iterations has been reached. r*un difcluster.sh unionPeak-file* peak score* number of classes* Fold*

*threshold/NA{pca} run*

- *unionPeak-file is the file name containing the list of peaks in a 3-columned format (output from normscore function).*
- *Peak score is the normalized tag counts (from different samples or chip) at peaks provided to the program (output from normscore function).*
- *number of clases a integer value greater than 1(like 2 3 or 10) number of clusters wanted.*
- *Fold_threshold, or NA is a numeric value (like 1.54 or 5) to cluster only those peaks which show fold change above the provided threshold; NA can also be provided if it is not wanted.*
- *Pca if the principal component analysis has to be performed on the selected peaks; if this option is used, then the output will be saved.*

4. **A spatial Heatmap** was generated using the normalized data from the above steps. It helps to visualize the signal intensity across the genome or a genomic region of interest.

   *run_plotmany.sh kmeanClusters.bed ChipfileNames ControlfileNames hg19 bam*

   - kmeanClusters.bed: This BED file generated from the above step contains the results of a k-means clustering analysis on ChIP-seq data. It has four columns: chromosome, start position, end position, and cluster-ID.
   - ChipfileNames: This is a list of ChIP-seq BAM files containing aligned reads for the protein of interest.
   - ControlfileNames: This is a list of BAM files that contain the aligned reads for the control samples.
   - hg19: This is the genome assembly version used to align the reads.

- Bam: This indicates that the input files are BAM files, a compressed binary format for storing DNA sequencing data.

Interpreting the spatial heat map results may involve identifying regions enriched for the protein of interest, comparing the signal intensity across different samples or conditions, or correlating the ChIP-seq signal with other genomic featuressuch as gene expression, DNA methylation, and histone modifications.

## 1.4 Data Collection

ChIP-seq data can be used to study various protein-DNA inter- actions, including transcription factor binding, histone modifications, and chromatin accessibility. By understanding these interactions, researchers can gain insights into how gene expression is regulated and how changes in gene expression may contribute to diseases like cancer. It's important to note that ChIP-seq data generation can be a technically challenging process, and some factors can affect the quality and reproducibility of the data, including antibody specificity, chromatin fragmentation, and sequencing depth. As a result, careful experimental design, quality control, and validation of the results are crucial for obtaining meaningful insights from ChIP-seq experiments. ChIP-seq signal data of various epigenetic marks for different developmental lineages were downloaded from several sources.

- **H1-ESC** (https://genome.ucsc.edu/cgi-bin/hgFileUi?g=wgEncodeBroadHistone.) They are derived from the UCSC genome browser, Histone Modifications by ChIP-seq from ENCODE/Broad Institute.
- **K562** (https://genome.ucsc.edu/cgi-bin/hgFileUi?g=wgEncodeBroadHistone.) They are derived from the UCSC genome browser, Histone Modifications by ChIP-seq from ENCODE/Broad Institute.

- **NPC derived from H1-ESC** (https://www.ncbi.nlm.nih.gov/geo/query/ acc.cgi). Derived from Gene Expression Omnibus by NCBI. The dataset canbe accessed by their GSE accession number (i.e., GSM675542, GSM753429, GSM818039, GSM818056, GSM908957, GSM956010) and freely downloaded by command line utilities like wget. The dataset includes chip-seq signals of various histone modifications and DNA methylation from bisulfite-seq data.

- **Mesenchymal cells derived from H1** (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi), derived from Gene Expression Omnibus by NCBI. The dataset can be accessed by their GSE accession number (i.e., GSM753437, GSM767344, GSM767352, GSM818041) and freely downloaded by command line utilities like wget. The dataset includes chip-seq signals of various histone modifications and DNA methylation from bisulfite-seq data.

- **BMP4 Mesendoderm Cells derived from H1** (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi). Derived from Gene Expression Omnibus by NCBI.The dataset can be accessed by their GSE accession number (i.e., GSM752968, GSM752978, GSM752982, GSM807401) and freely downloaded by command line utilities like wget. The dataset includes chip-seq signals of histone modifications and DNA methylation from bisulfite-seq data.

- **BMP4 Trophoblast Cells derived from H1** (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi). Derived from Gene Expression Omnibus by NCBI. The dataset can be accessed by their GSE accession number (i.e., GSM753436, GSM753439, GSM818054) and freely downloaded by command line utilities like wget. The dataset includes chip-seq signals of histone modifications and DNA methylation from bisulfite-seq data.

# 1.5 RESULTS

## 1.5.1 Chromatin states of K562 cell line.

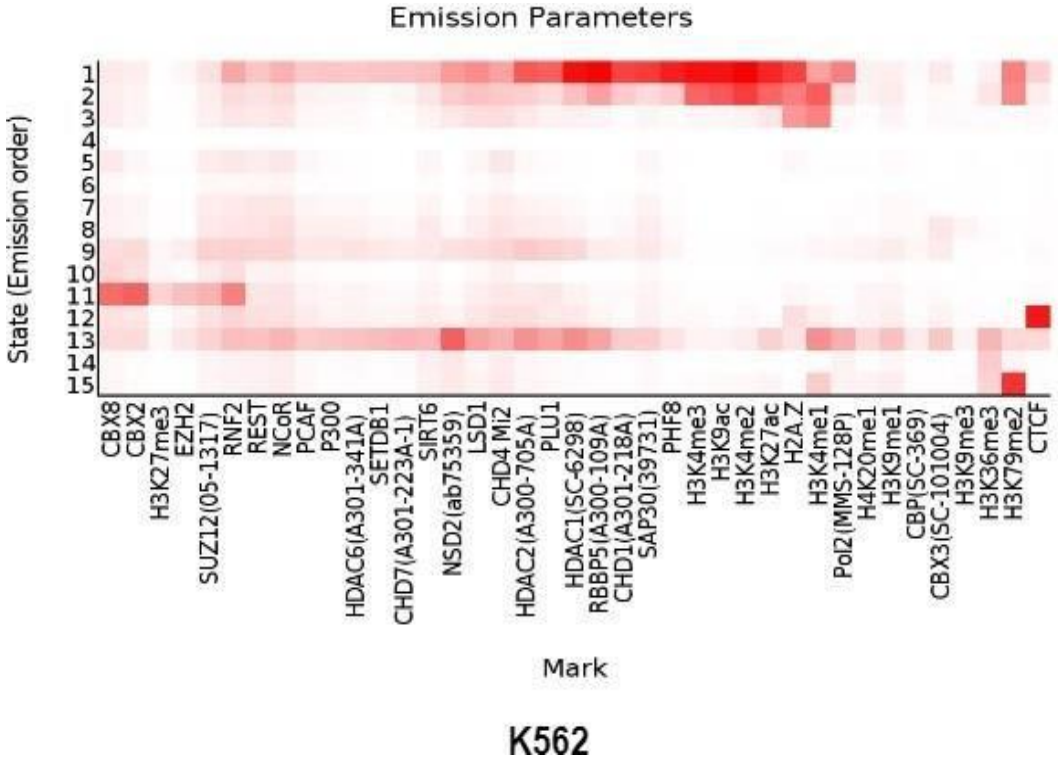



*Figure 1.1: Chromatin state of K562 cell line having different histone modifications and epigenetic marks.*

### 1.5.2 Chromatin states of H1-ESC cell line.

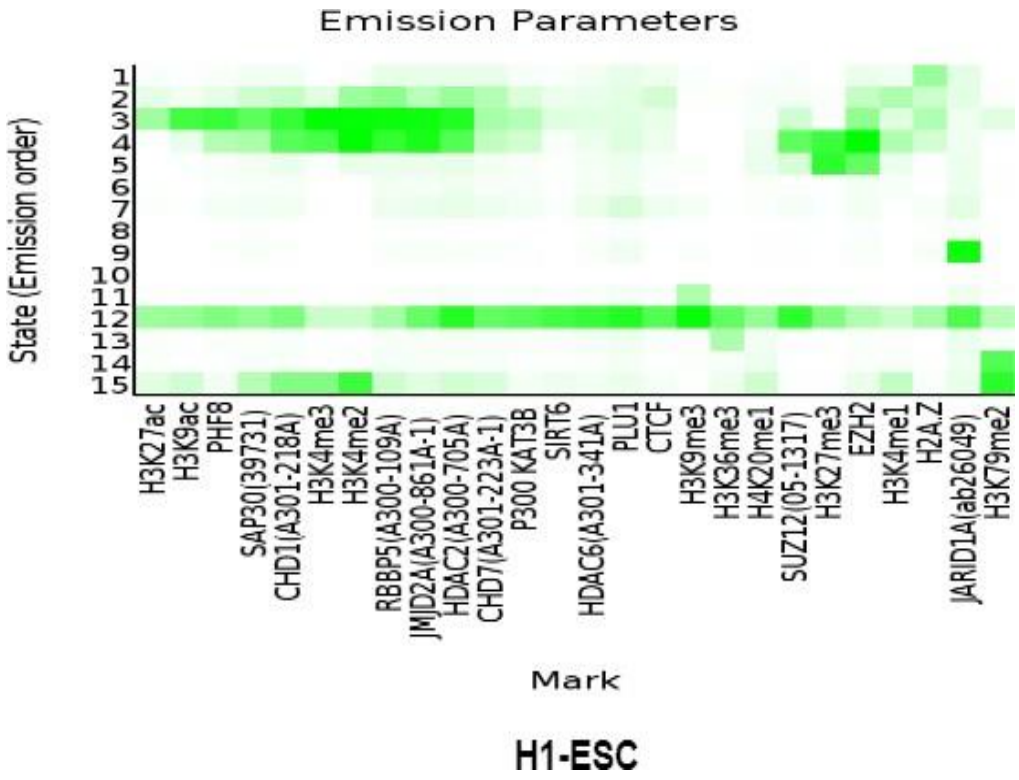



*Figure 1.2: Chromatin state of H1-ESC cell line having different histone modifications and epigenetic marks.*

### 1.5.3 Tracking the chromatin state of Neuronal Progenitor cells derived from H1 ESC for human H1ESC Enhancers.

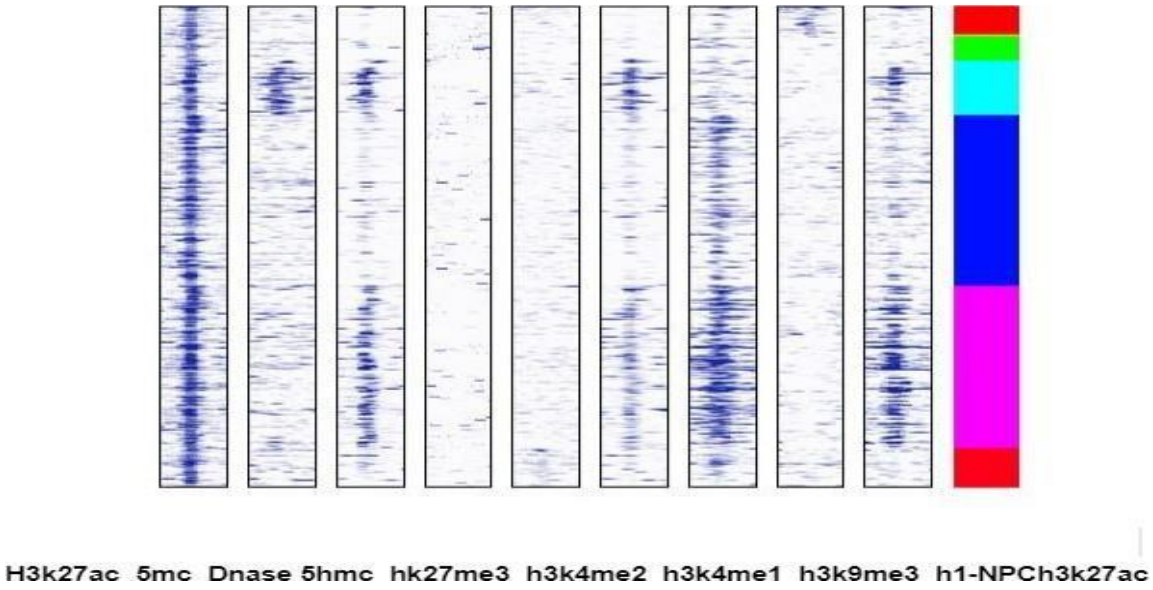



*Figure 1.3: Chromatin states of neural progenitor cells derived from H1-ESC having different histone modifications can be potential memory states during the differentiation of neural progenitor cells from embryonic stem cells.*

### 1.5.4 Tracking the chromatin state of Mesenchymal cells derived from H1-ESC for human H1ESC Enhancers.

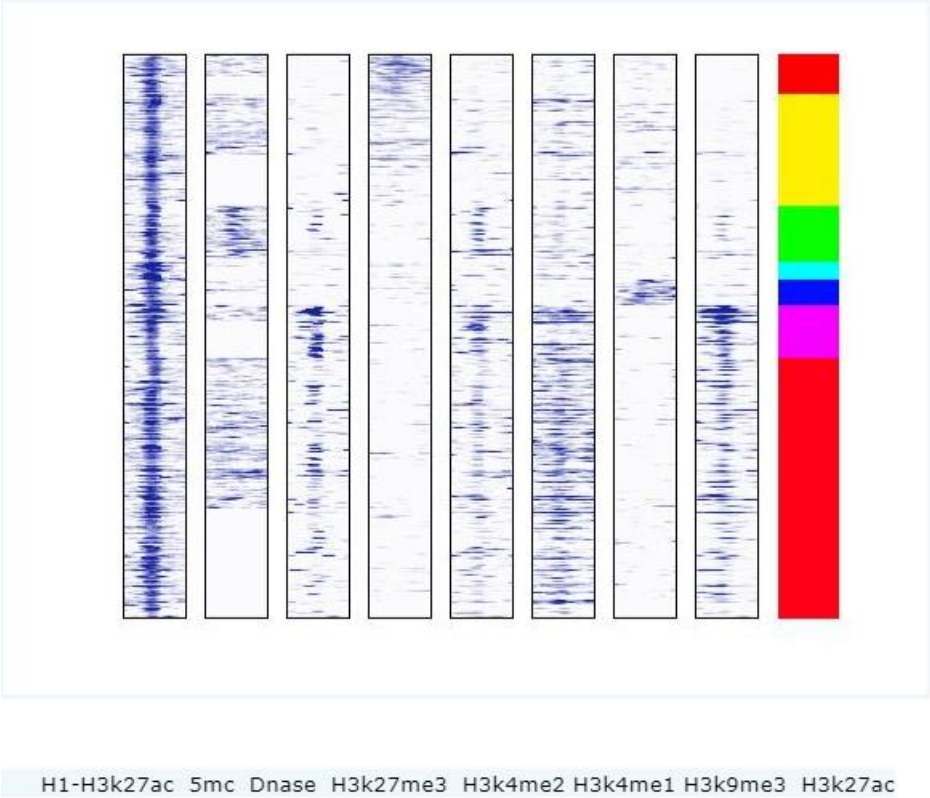



*Figure 1.4: Chromatin states of Mesenchymal cells derived from H1-ESC having different histone modifications can be potential memory states during the differentiation of neural progenitor cells from embryonic stem cells.*

### 1.5.5 Tracking the chromatin state of BMP4 Mesendoderm cells derived from H1-ESC for human H1ESC Enhancers.

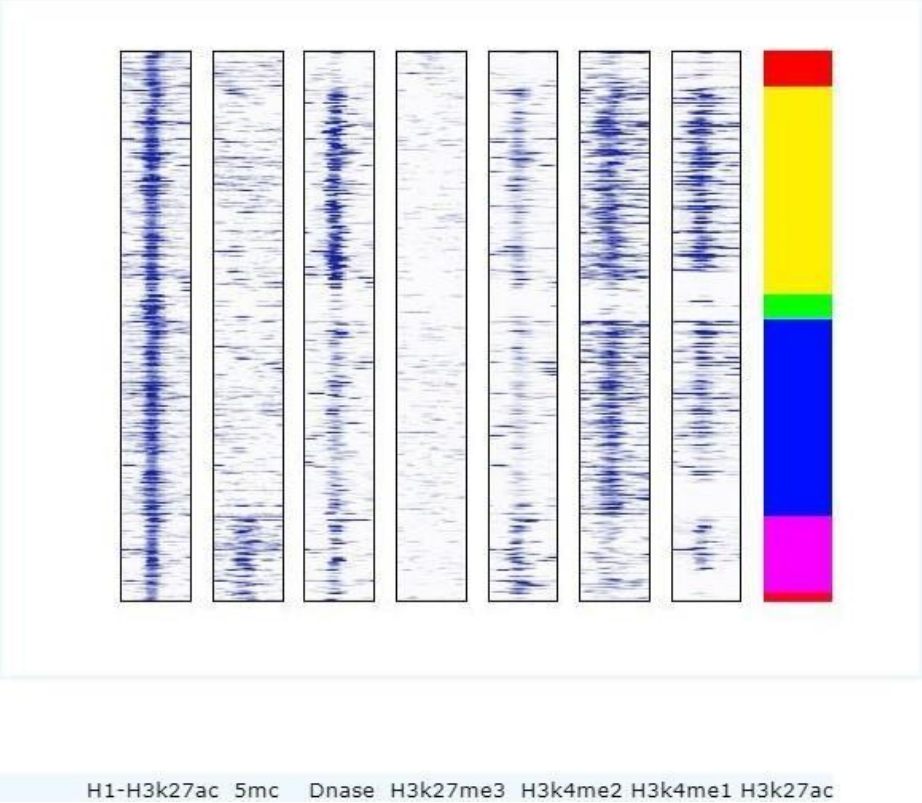



*Figure 1.5: Chromatin states of Mesendoderm cells derived from H1-ESC with different histone modifications can be potential memory states during the differentiation of Mesendoderm cells from embryonic stem cells.*

### 1.5.6 Tracking the chromatin state of BMP4 Trophoblast cells derived from H1-ESC for human H1-ESC enhancers.

### 1.1.1

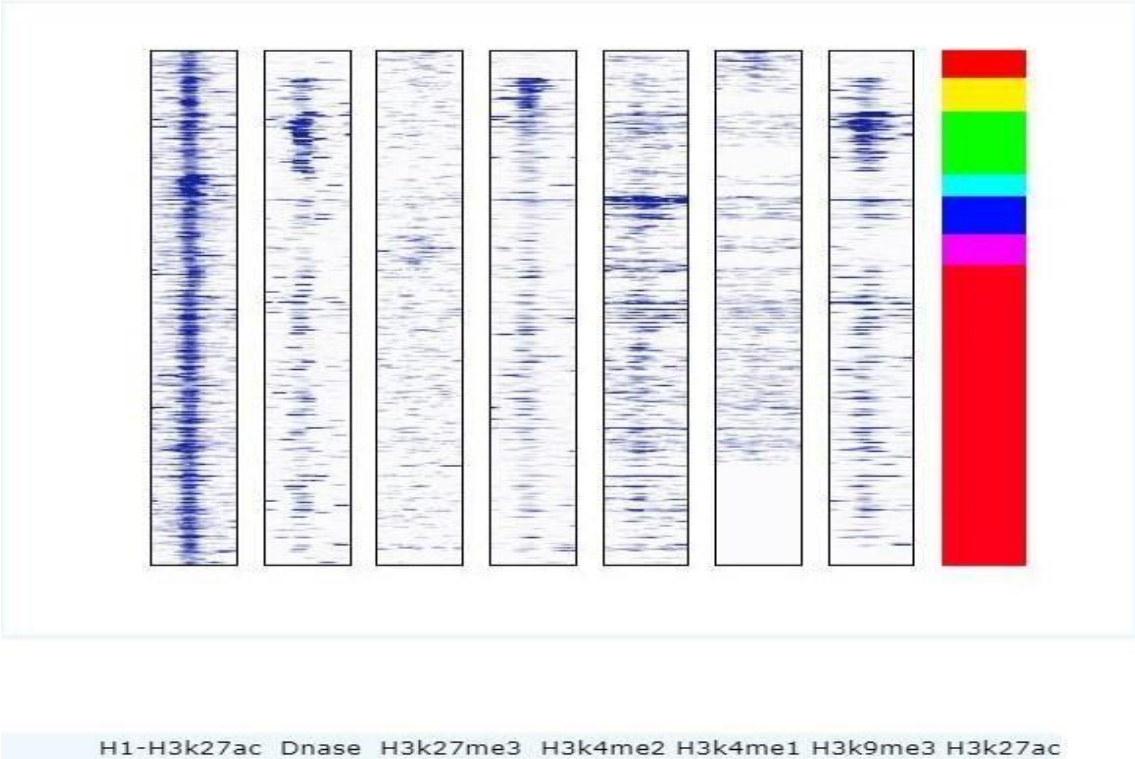



*Figure 1.6: Chromatin states of Trophoblast cells derived from H1-ESC having different histone modifications can be potential memory states during the differentiation of Trophoblast cells from embryonic stem cells.*

# Chapter 2 : Epigenetic Memory in Disease Progression

## 2.1 Cancer as an Epigenetic Disease

Cancer is a chronic, curable disease threatening human life and has emerged as a significant worldwide health issue. The growth and spread of cancer, formally considered a genetic disorder, is now recognized to entail anomalies of the epigenome in addition to genetic alterations. In mammals, tissue-specific gene expression patterns demand epigenetic processes for adequate growth and sustenance. Disrupting epigenetic structure can lead to gene activity changes and cancer cells' biological transformation. One of the hallmarks of cancer is the global changes in the epigenetic landscape [17]. Recent developments in cancer epigenetics have revealed significant reprogramming of the epigenetic machinery in all disease aspects, including histone modification and DNA methylation. Cancer cells lack the specific methylation and chromatin state architecture that controls the normal cellular homeostasis of gene expression patterns. Dense hypermethylation of the CpG islands linked to gene regulatory areas occurs concurrently with a worldwide genomic hypomethylation in the transformed cell's genome. The chromosomal instability, activation of endogenous parasite sequences, loss of imprinting, unauthorized expression, aneuploidy, and mutations resulting from these abrupt changes may also contribute to the transcriptional suppression of tumor suppressor genes [18]. Understanding the role of epigenetics in cancer is vital for developing new diagnostic and therapeutic approaches. For example, drugs that target epigenetic modifications have shown promise in clinical trials for certain types of cancer. Oncogene activation or tumor suppressor gene (TSG) repression are thought to be contributing factors to the development of cancer, these genes are always persistent with epigenetic alterations. The status of the switch that controls the turning “open” and “off” states of gene expression is DNA methylation. The most well-known epigenetic modification pathway in cancer cells is the hypermethylation of CGI promoters, which has been strongly linked to numerous cancer types. Other than abnormal DNA methylation, malignancies that follow the

CSC model, having bidirectional conversions are crucial that take advantage of unbalanced histone modification. The development of embryonic stem cells (ESCs) is where the bivalent histone marks, activating H3K4me3 and repressive H3K27me3, are initially addressed [51]. Oncofetal genes can be partially deregulated in cancer cells by various forms of Cancer, which partially recapitulate this bivalency. Other than abnormal DNA methylation, malignancies that follow the CSC model, in which bidirectional interconversions are crucial, take advantage of unbalanced histone modification. The development of embryonic stem cells (ESCs) is where the two bivalent histone marks, the repressive H3K27me3, and the activating H3K4me3 mark, are initially pointed. [51]. Oncofetal genes can be partially deregulated in cancer cells by various forms of Cancer, which partially recapitulate this bivalency [9].

Environmental factors, which are diet, toxins exposure, and stress, can also impact the epigenetic marks on the DNA, potentially increasing the risk of developing cancer. Understanding the role of epigenetic changes in cancer development and progression is an essential area of research that may aid in developing novel solutions for cancer diagnosis and treatment [17]. The changeable nature of epigenetic aberrations has paved the way for the prospective field of epigenetic therapy, which has already advanced with the FDA's recent approval of three epigenetic medications for the treatment of cancer.

## 2.2 Epigenetic Landscape in Cancer

In cancer, epigenetic changes are often observed and can add to the progression of the disease. These changes can occur at various stages of cancer development, including initiation, promotion, and metastasis. A critical aspect of the epigenetic landscape in Cancer is DNA methylation. Methylation is a chemical modification of DNA that can silence genes by preventing their expression. In Cancer, hypermethylation of specific tumor suppressor genes can lead to their inactivation, allowing cancer cells to grow and divide uncontrollably. Another significant epigenetic change in Cancer is alterations in histone modifications. These changes can affect the

structure of chromatin, the proteins, and the DNA complex in the nucleus of a cell and ultimately impact gene expression. For example, alterations in histone acetylation or methylation may lead to the repression or activation of genes involved in Cancer development. Overall, understanding the epigenetic landscape in Cancer is an essential area of research, as it can aid in a deeper understanding of the underlying mechanisms of the disease and potentially lead to the discovery of novel treatments.

### 2.2.1 DNA Methylation

The most researched epigenetic mechanism is DNA methylation regulation, which can alter gene expression without changing genetic information and takes part in several biological processes, including genomic stability, regulation of transcription, embryogenesis, and progression of Cancer.[20] Human genomic DNA frequently undergoes DNA methylation alterations, which covalently link CpG dinucleotides' cytosine groups and modify by adding methyl groups to their fifth carbon to create 5-methyl-cytosine. The human genome's CpG dinucleotide distribution is uneven, and promoter regions are where they are most prevalent [21]. CpG islands are regions with a more significant percentage of CpG found in the promoters of more than 60% of genes. Genes can become inactive through hypermethylation of a CpG island. However, CpG islands of transcriptionally active DNA sequences are generally unmethylated. For instance, a tiny subset of CpG islands is methylation in a tissue-specific manner to limit gene expression, even though most CpG islands in developing and differentiated tissues remain methylated. DNA hypermethylation can occasionally result in aberrant gene activation [22].

DNA methyl-transferases (DNMTs) are required for the catalysis of DNA methylation. There are five different varieties of DNMTs, which are DNMT1, DNMT3a, DNMT3b, DNMT2, and DNMT3L. The first three types are thought to have methyl- transferase activity. De-novo methylation is used to methylate unmethylated DNA double-strands by Dnmt3 and Dnmt3b, while DNMT1 primarily controls maintenance methylation (i.e., detecting the modified DNA strand and methylating the corresponding strand followed by that) [23]. The epigenetic imbalance

caused by abnormal DNA methylation is a crucial element in the development of tumors. Compared to somatic cells, cancer cells have higher levels of methylation in the promoter regions of several cancer suppressor genes, including MGMT, CdH1, E- cadherin, and BRCA1. Conversely, cancer cells have lower levels of methylation throughout their entire genome, linked to higher levels of proto-oncogene expression such as PAX2, ABCB1, and cyclinD2 [22].

### 2.2.2 Histone Modification

Histone modification is linked to replication, transcription, and repair of DNA through the interaction of Histone and DNA and Histone and Histone interaction. It affects the chromatin structure in contrast to DNA methylation, which occurs on gene sequences [24]. The free N-terminal allows nucleosome histones to be subject to many epigenetic controls. Histone modification can be classified as acetylation, methylation phosphorylation, adenylation, ubiquitination, and ADP ribosylation based on the many ways of action. The regulatory mechanisms of acetylation and methylation are well- known [22]. Acetylation of histones The most well-studied histone alteration, histone acetylation, iscrucial for chromatin structure, Cancer, and gene regulation. It is widely acknowledged that high acetylation stimulates gene expression while low acetylation suppresses it [25]. Histone acetyltransferase (HAT) and HDAC control the frequent modification of histone acetylation on lysine sites of histone types H3 and H4. Acetylation of histones is crucial for the growth of Cancer. Tumor suppressor genes and proto-oncogenes can interact with HDAC and HAT, interfering with how these genes are regulatedthroughout tumor cell growth, metastasis, and apoptosis [26].

The methylation of histones is also another critical in epigenetic regulation. Histone de- methylase and histone methyl invertase catalyze the reversible modification of the N- terminal arginine and lysine residues of histones H3 and H4. The specific biological activity of histone methylation, which controls the activation and silencing of gene transcription, depends on various sections of lysine or arginine amino acid residue sites, as well as the type of methylation [27]. For instance,

histone H3 is trimethylated at lysine K4 and K9 (H3K4me3&H3K9me3), which activates transcription. In contrast, histone H3 is methylated at lysine 9 (H3K9me), suppressing transcription. Cancer is directly linked to dysregulation in histone methylation. For instance, LSD1 is considered a potential target for treating acute leukemia, and it removes the methyl group from H3K4 and H3K9 sites as an active lysine demethylase. MLL1can also cause H3K4 methylation, which results in acute lymphoblastic leukemia. [22]

## 2.3 Epigenetic Memory in Cancer

H3K4me1 is a critical mark in promoting epigenetic memory. It is found in the genome frequently, is associated with enhancers, and has been enriched at enhancer regions throughout the genome. This modification is distinct from another known modification, such as H3K4me3m, associated with active gene promoters. This modification also aids in the recruitment of specific proteins and transcription factors to the enhancer site, where they can interact with the DNA and nearby promoters to increase or decrease gene expression. This way, H3K4me1 regulates gene expression and is a vital component in the epigenetic landscape that controls gene expression [28]. Open chromatin refers to the relaxed state of chromatin structure that allows access to DNA by transcription factors, DNA repair enzymes, and other proteins. Open chromatin is essentia l for controlling gene expression, as it enables the binding of transcription factors to specific regions of DNA and facilitates the initiation of transcription. DNAse enzymes are endonucleases that can cut the DNA backbone, and their activity is often used to identify open chromatin regions in the genome. DNAse hypersensitive sites (DHSs) are DNA regions more susceptible to cleavage by DNAse enzymes due to their open chromatin structure.

Enhancers are DNA sequences essential in regulating gene expression by recruiting proteins and transcription factors to initiate gene transcription. Enhancer priming typically involves pioneer TF binding and H3K4me1 pre-marking, which can occur before and aid in activating the following enhancers. It has been seen enhancers activated in differentiated macrophages lose TF

binding and H3K27ac instantly while retaining H3K4me1 for much longer. H3K4me1 persistence was suggested to aid in the induction of faster and stronger enhancers upon restimulation. Enhancers, in part, maintain H3K4me1, accumulate fewer heterochromatin marks, and remain accessible and sensitive to transcriptional activators [29]. Overall, the presence of H3K4me1 in enhancers is a crucial indicator of their regulatory activity, and it provides insights into gene regulation mechanisms in various cellular processes and diseases.

In this study, we are trying to discover whether H3K4me1and some transcription factors bind enhancers during different stages and progression of Cancer, but not DNAse binding, which is a lack of DNAse hypersensitive sites i.e., our hypothesis is during development and disease progression, H3K4me1 is persistent even once enhancers become decommissioned, lose their responsiveness to transcriptional activators and facilitate their eventual reactivation.

## 2.4 Chip-Seq Analysis with ChromHMM

ChromHMM is a tool used to learn and characterize chromatin states. ChromHMM may combine numerous datasets on chromatin, such as ChIP-seq data on diverse histone modifications, to identify the most common spatial and combinatorial markings in patterns. Under a multivariate Hidden Markov Model, ChromHMM explicitly models each chromatin mark's existence or absence. The genome of different cells is systematically annotated using the model that is produced. ChromHMM facilitates the biological characterization of each state by utilizing. A large amount of functional and annotation datasets to automatically calculate enrichment of states. ChromHMM is apowerful tool for analyzing chromatin state data because it uses a statistical approachto model the co-occurrence of multiple chromatin marks across the genome. It can identify different combinations of histone modifications that define specific chromatinstates, such as active promoters, enhancers, repressed regions, and transcriptionally active or inactive regions.

ChromHMM can also integrate other data types, such as DNA accessibility or transcription factor binding, to improve the accuracy of its predictions. The out-put of ChromHMM analysis is a set

of chromatin state annotations for the genome, which can be viewed in a genome browser or used for downstream analyses, such as gene expression profiling or functional enrichment analysis. One of the strengths of ChromHMM is that it can handle large-scale chromatin profiling data from multiple cell types or conditions, enabling researchers to compare chromatin states across different samples or to identify cell-type-specific regulatory elements. This is particularly useful for studying developmental processes or disease states where changes inchromatin structure and gene expression patterns are expected. ChromHMM can handle missing or incomplete data shared in large-scale chromatin profiling experiments. ChromHMM uses an expectation-maximization algorithm to estimate each genomic region's most likely chromatin state, even when data is missing for some chromatin marks. ChromHMM has been used in many applications, such as identifying disease-associated variants affecting chromatin states, predicting enhancer-promoter interactions, and characterizing cell-type-specific gene regulation. It is a widely adopted tool in the epigenomics research community, and its output is highly reproducible across different datasets and analysis pipelines. Our study used ChromHMM to analyze the chromatin state of varying cancer stages and their cell types. The following steps were followed:

- *Data preprocessing:* The first step is to preprocess the raw chromatin profiling data and generate files in the appropriate format for ChromHMM input. This typically involves aligning the sequencing reads to the reference genome, calling peaks for each chromatin mark, and converting the peak files to a binary for- mat indicating whether each genomic position is marked or unmarked for each chromatin mark.

- *Training the model:* Next, the ChromHMM model is trained using the preprocessed data. This involves specifying the number of chromatin states to be identified, selecting the chromatin marks to be used, and running the ChromHMM training algorithm to learn the probabilities of each chromatin state based on theobserved chromatin profiles.

- *State annotation:* Once the ChromHMM model has been trained, it can annotate the chromatin states for the entire genome. This involves running the ChromHMM annotation algorithm on the preprocessed data to assign each genomic position to one of the identified chromatin states.

- *Visualization and interpretation:* The final step is to visualize and interpret the results. This involves generating heatmaps of the chromatin states, identifying genomic regions enriched for particular states, and exploring the functional implications of the specified states using Gene ontology.

ChromHMM is a powerful tool for analyzing chromatin states; it requires careful attention to data preprocessing and model selection to ensure accurate results. Several options and parameters can be adjusted to optimize the analysis for a particular dataset. Our study used Pancreatic Cancer metastatic and primary cell chip seq data and Breast cancer cells, both wild-type and mutant.

## 2.5 Data Collection and Preprocessing

### 2.5.1 Data Collection

ChromHMM requires genome-wide chromatin data, such as ChIP-seq data for histone modifications or DNase-seq data for DNA accessibility. We obtain these data sets for our study's cell type for Cancer.ChromHMM also requires a control dataset (often referred to as input or background) to normalize the chromatin data. This is typicallya sample of the same cell type or tissue processed similarly but without the antibody used for ChIP-seq. The chromatin and input data should be in BAM or BED format, containing the reads mapped to the reference genome at specific genomic locations.

Chip-sq primary and metastatic pancreatic cancer and breast cancer wild-type and mutant data were obtained from Gene expression omnibus. To analyze ChIP-Seq data, SRA files were downloaded from NCBI Gene Expression Omnibus. GEO is a public data repository for gene expression data, and related metadata, generated from high- throughput molecular biology experiments. Datasets can be accessed using their GSE accession number and freely downloaded by command line utilities like wget.

- **Primary Pancreatic cancer cell Chip-seq Data** (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE99311). Derived from Gene Expression Omnibus by NCBI. Chip-Ses data of the three pancreatic primary cancer cells T3, T6 and T23 having histone marks of H3K4me1, H3K27ac, and DNAse open chromatin be accessed by their GSE accession number and freely downloaded by command line utilities like wget.

- **Metastatic Pancreatic Cancer Cell Chip-seq Data**.Derived from Gene Expression Omnibus by NCBI. Chip-Ses data of the three pancreatic metastatic cancer cells M1L, M10P, and M3P having histone marks of H3K4me1, H3K27ac and DNAse open chromatin be accessed by their GSE accession number and freely downloaded by command line utilities like wget. (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE99311).

- **Wild-type Breast Cancer Cell Chip-Seq Data** (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE159886). Derived from Gene Expression Omnibus by NCBI. Chip-Seq data of the three Wild-type Breast cancer cells Sample 3, Sample 4, and Sample 6, having histone marks of H3K4me1, H3K27ac, and DNAse open chromatin, be accessed by their GSE accession number and freely downloaded by command line utilities like wget.

- **Mutant Breast Cancer Cell Chip-Seq Data**. (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE159886). Derived from Gene Expression Omnibus by NCBI.Chip-Seq data of the three mutant Breast cancer cells Sample 9, Sample 12, and Sample 13 having histone marks of H3K4me1, H3K27ac, and DNAse open chromatin

be accessed by their GSE accession number and freely downloaded by command line utilities like wget.

### 2.5.2 Data Preprocessing

The preprocessing of ChIP-seq data involves several steps to ensure the quality of the data before downstream analysis.

- *Quality control:* Check the quality of the raw sequencing data using a tool such as FastQC.
- *Alignment:* Align the trimmed reads to a reference genome using a read alignments tool such as Bowtie2, BWA, or HISAT2. The alignment file is typically stored in SAM/BAM format.
- *PCR duplicates removal:* Removing PCR duplicates to avoid biases caused by over-amplification of some areas of the genome. Software such as Picard or samtools can be used for this step.
- *Peak calling:* Identifies genome regions that are enriched for the protein of interest using peak calling software such as MACS2 or SICER. This step will generate a list of genomic regions with significantly more reads mapped to them than expected by chance.

These preprocessing steps are critical for ensuring the accuracy and reliability of the downstream analysis of ChIP-seq data. Tools that were used for preprocessing steps are as follows:

- *Bigwig2bed:* For downstream analysis of chip-seq data, we would be required to use either BAM alignment or BED format. Bigwig2bed is used to convert wig data or bigwig data to bed format. The wig2bed tool typically works by taking a wiggle file as input and outputting a bed file with the same data by Parsing the input wiggle file and extracting

the relevant information, such as chromosome, start position, end position, and score or coverage data.

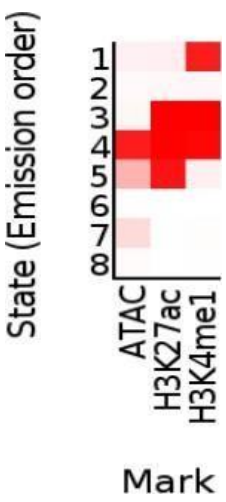
- *SRA Toolkit:* For fetching raw chip-seq data SRA toolkit was used. The SRA Toolkit is a software package developed by the National Center for Biotechnology Information (NCBI) to facilitate retrieving and analyzing data from the Sequence Read Archive (SRA). The SRA Toolkit has command-line tools for downloading, converting, and analyzing SRA data. These tools include fastq-dump, which can be used to convert SRA files into standard FASTQ format, and SRA Blast, which can be used to perform BLAST searches against SRA data.

- *STAT Aligner:* For aligning raw reads, STAR aligner was used. STAR aligner is designed to align high-throughput sequencing accurately reads to a reference genome. STAR aligner uses a two-pass mapping algorithm that generates a genome index and then aligns the RNA sequencing reads to the index. The first pass generates a genome index that contains information on all possible splice junctions within the reference genome. In the second pass, the STAR aligner uses this index to map the reads to the genome, and it can detect novel and known splice junctions with high accuracy.

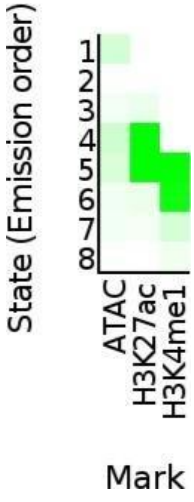
- *Bedtools:* Bedtools is a software suite developed for working with genomic interval data. It provides tools for performing v a r i o u s operations on genomic intervals, including set operations such as union, intersection, difference, filtering, and annotation. The software is named after the “BED” file format, a simple tab- delimited text format used to represent genomic intervals.

## 2.6 RESULTS

### 2.6.1 Tracing the chromatin states of Pancreatic Cancer cell

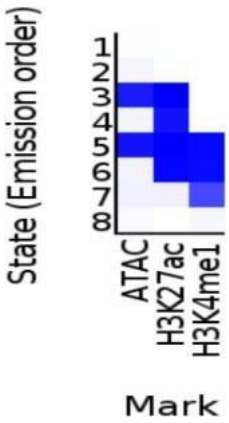



(a) Chromatin states of Primary Pancreatic Cancer cell T3

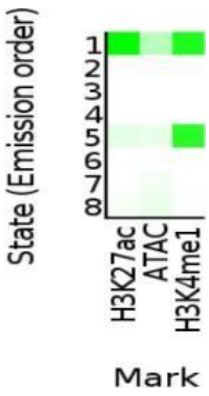



(b) Chromatin states of Metastatic Pancreatic Cancer cell M1L

*Figure 2.1: Chromatin states of Primary (T3) and Metastatic (M1L) Pancreatic Cancer Cell having open chromatin with H3K4me1 is a possible site of memory.*

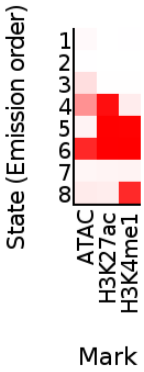



(a) Chromatin states of Primary Pancreatic Cancer cell T23

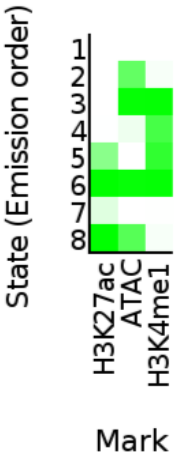



(b) Chromatin states of Metastatic Pancreatic Cancer cell M10P.

*Figure 2.2: Chromatin states of Primary (T23) and Metastatic (M10P) Pancreatic Cancer Cell having open chromatin with H3K4me1 is a possible site of memory.*



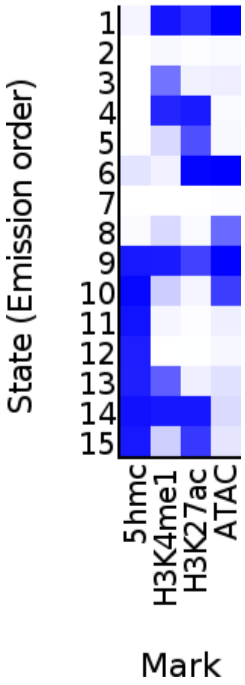

(a) Chromatin states of Primary Pancreatic Cancer cell T6



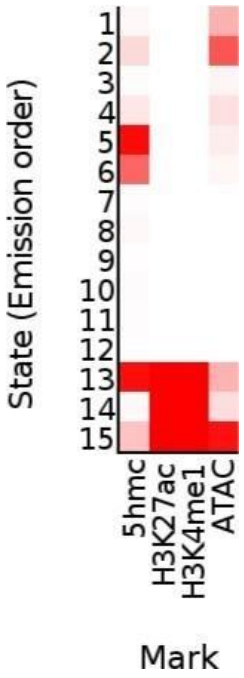

(b) Chromatin states of Metastatic Pancreatic Cancer cell M3P.

*Figure 2.3: Chromatin states of Primary (T6) and Primary (M3P) Pancreatic Cancer Cell having open chromatin with H3K4me1 is a possible site of memory.*

### 2.6.2 Tracing the chromatin states of Breast Cancer cell

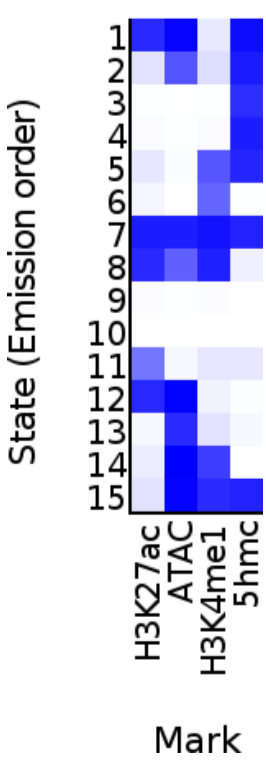



(a) Chromatin states of Wild-type Breast Cancer cell Sample 3

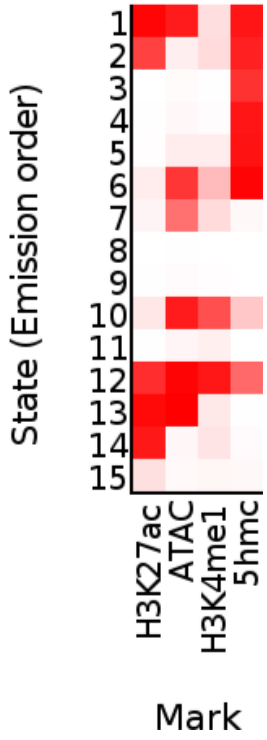



(b) Chromatin states of Mutant Breast Cancer cell Sample 13

*Figure 2.4: Chromatin states of Wild-type (Sample 3) and Mutant (Sample 13) Breast Cancer Cell having open chromatin with H3K4me1 is a possible site of memory.*

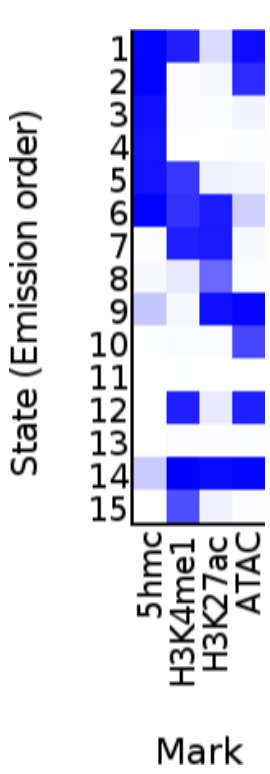



(a) Chromatin states of Wild-type Breast Cancer cell Sample 5

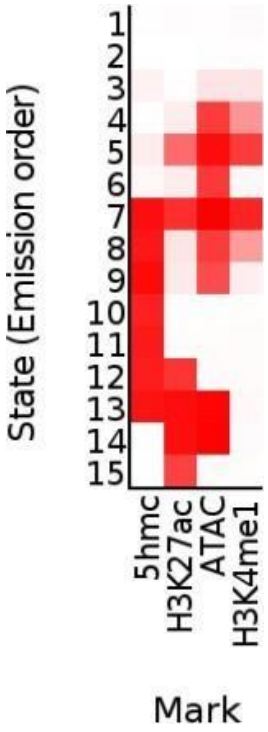



(b) Chromatin states of Mutant Breast Cancer cell Sample 12

*Figure 2.5: Chromatin states of Wild-type (Sample 5) and Mutant (Sample 12) Breast Cancer Cell having, having open chromatin with H3K4me1 is a possible site of memory.*




(a) Chromatin states of Wild-type Breast Cancer cell Sample 6




(b) Chromatin states of Mutant Breast Cancer cell Sample 9

*Figure 2.6: Chromatin states of Wild-type (Sample 6) and Mutant (Sample 9) Breast Cancer Cell having, having open chromatin with H3K4me1 is a possible site of memory.*

# Chapter 3: Post-hoc Analysis for the insights about cancer development

Precision medicine is an idea and a practice that follows a deliberate and organized approach to treating illnesses like cancer [30]. The precision medicine project is heavily focused on treating cancer, but improvements in targeted, efficient therapies could also help treat a wide range of other chronic conditions. Precision oncology strives to match each cancer patient with the most precise and effective treatment based on the patient's genetic profile. Given that every cancer patient has a unique genetic profile and that the profile may change over time, more people may benefit if therapeutic alternatives can be tailored to the individual in cancer treatment [31].

Precision therapy, also known as personalized medicine, is an approach to cancer treatment that involves tailoring treatment plans to an individual's unique genetic makeup, lifestyle, and other personal factors. Precision therapy aims to provide more effective and targeted treatment while minimizing side effects. One type of precision therapy used in cancer treatment is targeted therapy, which entails using medications that specifically target specific proteins or genes associated with the development and propagation of cancer. These drugs are often designed to block the activity of specific overactive or mutated molecules in cancer cells, leaving normal cells relatively unaffected. Another type of precision therapy is immunotherapy, which utilizes the immune system of the body to combat cancer cells. Cancer cells need specific proteins to avoid the immune system, and immunotherapy medications can either boost the immune system to attack such proteins or inhibit them. Overall, precision therapy has shown promise in improving outcomes for certain types of cancer, particularly those that are difficult to treat with traditional chemotherapy and radiation therapy.

## 3.1 Epigenetic Modifiers in Precision Therapy of Cancer

Immune treatments have transformed the way that Cancer is treated in recent years. To improve patient outcomes, it is essential to increase sensitivity to immune therapies because most patients resist these treatments on a main or secondary level. Epigenetic modifiers might be helpful as therapeutic agents since, according to some recent lines of evidence, they have inherent immunomodulatory abilities [32]. Epigenetic modifiers are a class of drugs that can alter the structure and its associated proteins without changing the under- lying DNA sequence. Epigenetic changes can regulate gene expression, and alterations in epigenetic marks have been implicated in many diseases, including Cancer. Several epigenetic modifiers include DNA methylation inhibitors, histone deacetylase inhibitors, histone methyl-transferase inhibitors, histone demethylase inhibitors, and bromodomain and extra-terminal (BET) inhibitors.

Histone deacetylase inhibitors (HDACis) and DNA methyl transferase inhibitors (DNMTis), which have been authorized for treatment in several types of hematologic malignancies, werethe first epigenetic modifiers to be created. DNMTis, which are cytidine analogs, prevent DNMT from functioning and make them degrade when integrated into DNA. The re-expression of abnormally repressed proteins, such as cancer-associated antigens, tumor suppressor genes, and parts of the antigen presentation apparatus, is encouraged by the subsequent loss of DNA methylation [33].

Histone deacetylase inhibitors, such as vorinostat and panobinostat, block the enzyme activity of removal of acetyl groups from histones, which changes the chromatin structure, thereby causing a change in gene expression. These drugs are authorized to treat cutaneous T-cell lymphoma and multiple myeloma. Histone methyl- transferase and histone demethylase inhibitors, such as tazemetostat and GSKJ4, target enzymes that add or remove methyl groups from histone proteins, leading to changes in chromatin structure and gene expression. Bromodomain and extra-terminal (BET) inhibitors, such as JQ1 and OTX015, target proteins called bromodomain and extra-

terminal (BET) proteins, which help to regulate gene expression. These drugs are being investigated to treat various cancers, including leukemia, lymphoma, and solid tumors.

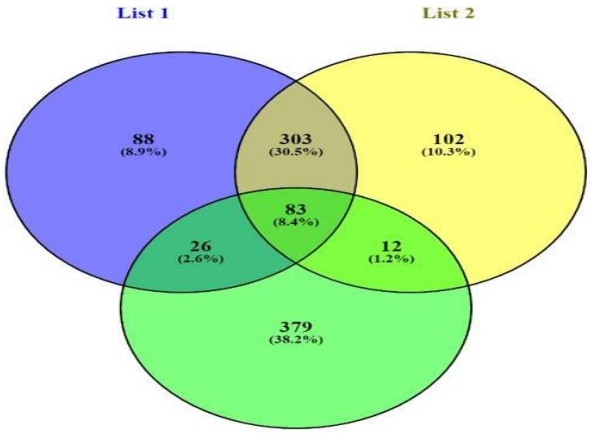
## 3.2 Targeting Neighbouring Enhancers for Cancer Hallmark Genes for Precision Therapy

Over the past few decades, researchers have identified several “hallmark” genes and Master Regulator that play critical roles in the development and progression of cancer. One of the aims of this study is to find out such master regulators in different cancer profiles from the neighboring memory enhancers which have the memory mark (H3K4me1 without DNA hypersensitivity) and target them for personalized therapy for cancer. This study primarily focused on two hallmarks of Cancer for analysis: evasion of apoptosis and immune responsiveness.

Apoptosis is a programmed cell death mechanism that plays a critical role in maintaining tissue homeostasis and preventing the growth of damaged or abnormal cells. Cancer cells, however, have developed several mechanisms to evade apoptosis, considered one of the hallmarks of cancer. Cancer cells can evade apoptosis through several means, such as mutations in genes that regulate apoptosis, upregulation of anti- apoptotic proteins, and downregulation of pro-apoptotic proteins. By evading apoptosis, cancer cells can survive and continue to grow and divide uncontrollably. Immune responsiveness is another hallmark of cancer, as cancer cells can evade the immune system and avoid destruction by immune cells. Cancer cells can do this through several mechanisms, including downregulating the expression of antigens on their surface, producing immunosuppressive molecules, and promoting the formation of an immunosuppressive microenvironment. Together, evasion of apoptosis and immune responsiveness enable cancer cells to grow and divide uncontrollably, invade surrounding tissues, and metastasize to distant sites in the body. Understanding these hallmarks of cancer is critical for developing new cancer therapies and identifying new targets for cancer treatment.

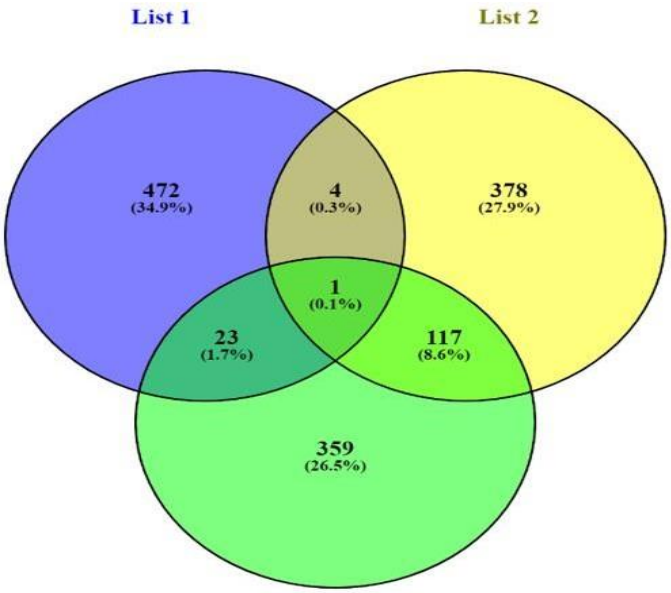
Gene ontology, a standardized system used to describe genes and their products and to annotate and analyze gene function, was performed in different Cancer Profiles in neighboring memory

enhancers to find out the standard biological process among these profiles for further analysis. The genes common to these pathways could be a target in memory enhancers.




| COMMON BIOLOGICAL PROCESSES IN MEMORY STATE OF PRIMARY PANCREATIC CANCER CELLS | | |
|---|---|---|
| dendritic spine maintenance | sensory perception of chemical stimulus | focal adhesion assembly |
| T cell mediated cytotoxicity | negative regulation of muscle contraction | pantothenate metabolic process |
| susceptibility to natural killer cell mediated cytotoxicity | diet induced thermogenesis | sterol metabolic process |
| regulation of cytokine secretion | terpenoid catabolic process | positive regulation of macrophage migration |
| necrotic cell death | positive regulation of monocyte chemotaxis | regulation of tumor necrosis factor secretion |
| necroptotic process | detection of chemical stimulus involved in sensory perception of smell | regulation of protein secretion |
| central nervous system myelin formation | positive regulation of secretion by cell | nuclear-transcribed mRNA catabolic process |
| programmed necrotic cell death | double-strand break repair via classical nonhomologous end joining | programmed cell death involved in cell development |
| positive regulation of fibroblast proliferation | protein lipidation | adherens junction organization |
| negative regulation of T-helper 17 cell lineage commitment | negative regulation of transcription involved in G1/S transition of mitotic cell cycle | positive regulation of secretion |
| positive regulation of mononuclear cell migration | oligosaccharide biosynthetic process | positive regulation of microtubule polymerization |
| plasma membrane repair | cellular response to nitrogen dioxide | RNA catabolic process |
| endosome to melanosome transport | T-helper 2 cell activation | regulation of metanephric glomerular mesangial cell proliferation |
| cellular response to nitrosative stress | positive regulation of CD4-positive, alpha-beta T cell costimulation | ventral spinal cord development |
| regulation of high voltage-gated calcium channel activity | chemokine (C-C motif) ligand 11 production | positive regulation of leukocyte mediated cytotoxicity |
| pyrimidine nucleoside triphosphate biosynthetic process | positive regulation of interleukin-4-dependent isotype switching to IgE isotypes | positive regulation of mRNA metabolic process |
| cellular response to glucose starvation | CD4-positive, alpha-beta T cell costimulation | sensory perception of smell |
| positive regulation of protein secretion | memory T cell activation | positive regulation of T cell costimulation |
| spinal cord association neuron differentiation | galactosylceramide biosynthetic process | embryonic hemopoiesis |
| cell differentiation in spinal cord | negative regulation of T-helper 1 cell differentiation | sensory perception |
| bradykinin catabolic process | asparaginyl-tRNA aminoacylation | positive regulation of respiratory burst |
| positive regulation of isotype switching to IgE isotypes | detection of chemical stimulus | ventral spinal cord interneuron differentiation |
| CMP salvage | detection of chemical stimulus involved in sensory perception | viral genome replication |
| detection of stimulus involved in sensory perception | cardiac cell fate specification | positive regulation of cell killing |
| lysine biosynthetic process via aminoadipic acid | defense response to nematode | mRNA catabolic process |
| detection of stimulus | positive regulation of cytokine secretion | micturition |
| regulation of anion channel activity | positive regulation of cellular extravasation | lipoprotein biosynthetic process |
| positive regulation of DNA replication | positive regulation of macrophage chemotaxis | |

*Figure 3.1: Ven diagram showing standard gene function between three primary pancreatic cancer cell patients followed by the list of common biological functions among these cells.*

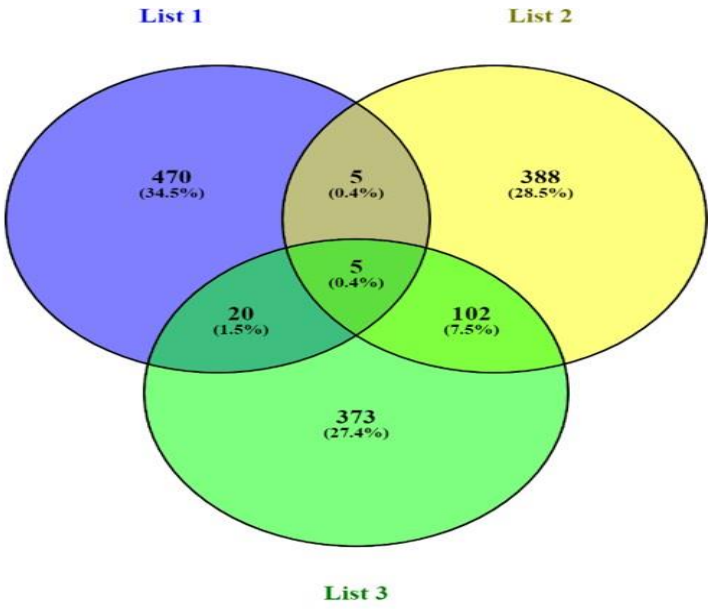



Results:

```
1 common element in "List 1", "List 2" and "List 3":
glycosylceramide biosynthetic process
```

| COMMON BIOLOGICAL PROCESSES IN MEMORY STATE OF METASTATIC PANCREATIC CANCER CELLS |
| --- |
| Glycosylceramide biosynthetic process |

*Figure 3.2: Ven diagram showing standard gene function between three metastatic pancreatic cancer cell patients, followed by the list of common biological functions among these cells.*

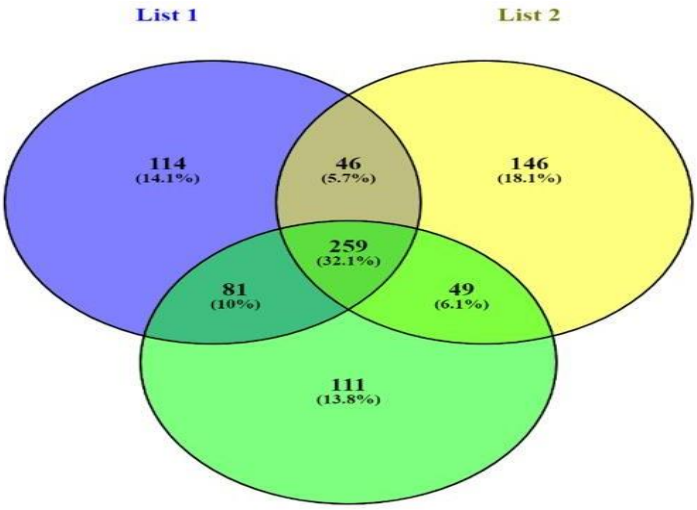



Results:

```
5 common elements in "List 1", "List 2" and "List 3":
regulation of leukocyte activation
regulation of cell activation
regulation of lymphocyte activation
negative regulation of immune system process
negative regulation of protein import into nucleus, tra
```

| COMMON BIOLOGICAL PROCESSES IN MEMORY STATE OF MUTANT BREAST CANCER CELLS |
|---|
| regulation of leukocyte activation |
| regulation of cell activation |
| regulation of lymphocyte activation |
| negative regulation of immune system process |
| negative regulation of protein import into nucleus, translocation |

*Figure 3.3: Ven diagram showing standard gene function between three wild-type breast cancer cell patients, followed by the list of common biological functions among these cells.*




Results:

```
259 common elements in "List 1", "List 2" and "List
leukocyte activation
secretion by cell
positive regulation of immune system process
regulation of immune response
immune effector process
exocytosis
```

| COMMON BIOLOGICAL PROCESSES IN MEMORY STATE OF WILD-TYPE BREAST CANCER CELLS | |
|---|---|
| leukocyte activation | immune response-activating signal transduction |
| secretion by cell | immune response-regulating cell surface receptor signaling pathway |
| positive regulation of immune system process | immune response-activating cell surface receptor signaling pathway |
| regulation of immune response | phagocytosis |
| immune effector process | response to external biotic stimulus |
| exocytosis | endocytosis |
| regulation of defense response | hemostasis |
| response to other organism | regulation of innate immune response |
| regulated exocytosis | inflammatory response |
| interspecies interaction between organisms | response to biotic stimulus |
| symbiosis, encompassing mutualism through parasitism | coagulation |
| cell activation involved in immune response | blood coagulation |
| regulation of cytokine production | immune system development |
| leukocyte activation involved in immune response | cellular macromolecule catabolic process |
| multi-organism cellular process | hematopoietic or lymphoid organ development |
| viral process | macromolecule catabolic process |
| leukocyte mediated immunity | Fc receptor signaling pathway |
| myeloid leukocyte activation | apoptotic process |
| positive regulation of immune response | regulation of type I interferon production |
| myeloid cell activation involved in immune response | regulation of hemopoiesis |
| myeloid leukocyte mediated immunity | regulation of gene expression, epigenetic |
| leukocyte degranulation | viral life cycle |
| granulocyte activation | posttranscriptional regulation of gene expression |
| neutrophil activation | regulation of immune effector process |
| neutrophil mediated immunity | regulation of intrinsic apoptotic signaling pathway |
| neutrophil activation involved in immune response | negative regulation of type I interferon production |
| immune response-regulating signaling pathway | Fc-gamma receptor signaling pathway |
| neutrophil degranulation | regulation of catabolic process |
| innate immune response | Fc-gamma receptor signaling pathway involved in phagocytosis |
| activation of immune response | autophagy |
| proteolysis involved in cellular protein catabolic process | positive regulation of mRNA catabolic process |
| ribonucleoprotein complex biogenesis | response to insulin |
| ncRNA processing | peptidyl-serine phosphorylation |
| positive regulation of mRNA metabolic process | ubiquitin-dependent protein catabolic process |
| mRNA metabolic process | regulation of proteolysis |
| cellular nitrogen compound catabolic process | positive regulation of innate immune response |
| nucleus organization | regulation of leukocyte degranulation |
| unsaturated fatty acid metabolic process | negative regulation of cytokine production |
| T cell receptor signaling pathway | regulation of viral process |
| positive regulation of programmed cell death | multi-organism metabolic process |

| COMMON BIOLOGICAL PROCESSES IN MEMORY STATE OF WILD-TYPE BREAST CANCER CELLS contd.. | |
|---|---|
| regulation of leukocyte activation | cytokine-mediated signaling pathway |
| regulation of cell activation | positive regulation of cellular catabolic process |
| wound healing | endosomal transport |
| platelet activation | apoptotic signaling pathway |
| regulation of lymphocyte activation | response to type I interferon |
| Fc receptor mediated stimulatory signaling pathway | negative regulation of protein modification process |
| mitochondrion organization | regulation of leukocyte cell-cell adhesion |
| regulation of cellular amide metabolic process | innate immune response activating cell surface receptor signaling pathway |
| myeloid cell differentiation | protein catabolic process |
| hemopoiesis | negative regulation of transferase activity |
| response to wounding | regulation of mitochondrion organization |
| regulation of cell cycle phase transition | negative regulation of immune system process |
| regulation of translation | chromatin organization |
| covalent chromatin modification | histone modification |
| regulation of mitotic cell cycle phase transition | response to cytokine |
| intracellular protein transport | ncRNA metabolic process |
| response to bacterium | activation of innate immune response |
| lymphocyte activation | positive regulation of hydrolase activity |
| regulation of T cell activation | positive regulation of mitochondrion organization |
| defense response to other organism | negative regulation of intracellular signal transduction |
| regulation of cellular catabolic process | actin cytoskeleton organization |
| cellular response to DNA damage stimulus | negative regulation of intrinsic apoptotic signaling pathway |
| interaction with host | mitotic cell cycle |
| cellular component disassembly | regulation of defense response to virus |
| positive regulation of organelle organization | regulation of response to biotic stimulus |
| regulation of autophagy | stimulatory C-type lectin receptor signaling pathway |
| mitotic cell cycle process | leukocyte differentiation |
| mitochondrial transport | RNA processing |
| antigen receptor-mediated signaling pathway | protein phosphorylation |
| single-organism intracellular transport | regulation of symbiosis, encompassing mutualism through parasitism |
| nucleobase-containing compound catabolic process | response to lipopolysaccharide |
| regulation of mitochondrial membrane permeability | modification-dependent macromolecule catabolic process |
| positive regulation of release of cytochrome c from mitochor | positive regulation of catabolic process |
| modification by symbiont of host morphology or physiology | organic cyclic compound catabolic process |
| nuclear-transcribed mRNA catabolic process | positive regulation of cell death |
| negative regulation of cell cycle phase transition | Ras protein signal transduction |
| mitochondrial membrane organization | gene silencing |
| RNA catabolic process | modification-dependent protein catabolic process |
| regulation of transferase activity | cellular protein catabolic process |
| mRNA catabolic process | protein autophosphorylation |
| regulation of leukocyte mediated immunity | leukocyte migration |

*Figure 3.4: Ven diagram showing standard gene function between three mutant breast cancer cell patients, followed by the list of common biological functions among these cells.*

| Primary Paancretic Cancer Cell T3 | |
|---|---|
| Regulation of endothelial cell apoptotic process | AKR1C3,ANGPT1,AND6,BRAF,CD40LG,FASLG,FGA,FGB,FGG,FOXO3,GAS6,GATA3,ICAM1,IL13 |
| T cell proliferation involved in immune response | CD86,LILRB1,TNFSF18 |

| Primary Paancretic Cancer Cell T6 | |
|---|---|
| Regulation of endothelial cell apoptotic process | AKR1C3,ANGPT1,AND6,BRAF,CD40LG,FASLG,FGA,FGB,FGG,FOXO3,GAS6,GATA3,ICAM1,IL13 |
| T cell activation nvolved in immune response | APBB1IP,BCL3,CD1C,CD86,CLEC4D,CLEC4E,EIF2AK4,EOMES,F2RL1,FCER1G,FOXP1,GATA3,GPR183,HMGB1,ICAM1 |

| Primary Paancretic Cancer Cell T6 | |
|---|---|
| Regulation of intrinsic apoptotic signaling pathway in response to DNA da | C11orf82,CD44,CXCL12,FBXO18,KDM1A,MUC1,PGAP2,RPS3,SIRT1,USP47 |
| T cell proliferation nvolved in immune response | TNFSF18 |

| Metastatic Pancreatic Cancer Cell M1L | |
|---|---|
| Regulation of intrinsic apoptotic signaling pathway | ARHGEF2,CXCL12,IVNS1ABP,KDM1A,MCL1,MUC1,PARK7,PINK1,PTGS2,TXNDC12 |
| T cell activation involved in immune response | APBB1IP,CD1C,FCER1G,GATA3,LY9,RC3H1,RORC,SEMA4A,SLAMF6,TNFSF18 |

| Metastatic Pancreatic Cancer Cell M3P | |
|---|---|
| Epithelial cell apoptotic process | ARF6,DNMT3A,GSN,JAG2,KRT8,MAP3K5,PIK3CG,RB1,SIX3,STK3,TGFBR2,TNF |
| B cell proliferation involved in immune response | CD180,TLR4 |

| Metastatic Pancreatic Cancer Cell M10P | |
|---|---|
| Regulation of apoptotic cell clearance | C4A,C4B,CCL2,HMGB1 |
| T cell activation involved in immune response | APBB1IP,BCL3,CD1C,CD86,CLEC4D,CLEC4E,EOMES,FCER1G,FOXP1,GATA3,GPR183,HMGB1,IFNA1,IFNA14,IFNA17, |

*Figure 3.5: Genes for apoptotic signaling and immunoregulation found through ontology in neighboring memory enhancers of Primary and Metastatic Pancreatic Cancer.*

| Wild-type Breast Cancer Cell Sample3 | |
|---|---|
| Regulation of endothelial cell apoptotic process | AKR1C3,ANGPT1,ANO6,BRAF,CD40LG,FASLG,FGA,FGB,FGG,FOXO3,GAS6,GATA3,ICAM1,IL13 |
| T cell proliferation involved in immune response | CD86,LILRB1,TNFSF18 |

| Wild-type Breast Cancer Cell Sample5 | |
|---|---|
| Positive regulation of apoptotic process | ABL1,ABR,ACER2,ADAM8,ADAMTSL4,ADCY10,ADIPOQ,ADM,ADORA2A,AES,AGT,AGTR2,AIFM1,AIFM2,AKAP13 |
| Positive regulation of immune response | ABI1,ABL1,ACTB,ACTG1,ACTR2,ACTR3,ADA,ADAM8,ADORA2B,AIM2,ANXA1,AP1G1,ARPC1A,ARPC1B,ARPC2,ARPC3,ARPC4 |

| Wild-type Breast Cancer Cell Sample6 | |
|---|---|
| Positive regulation of apoptotic process | ABL1,ABR,ACER2,ACVR1C,ADAM8,ADAMTSL4,ADCY10,ADIPOQ,ADM,ADORA2A,AES,AGT,AGTR2,AIFM1,AIFM2,AKAP13,AKR1C3,AKT1 |
| Positive regulation of immune response | ABI1,ABL1,ACTB,ACTG1,ACTR2,ACTR3,ADA,ADAM8,ADORA2B,AIM2,ANXA1,AP1G1,APOB,ARPC1A,ARPC1B |

| Mutant Breast Cancer Cell Sample9 | |
|---|---|
| Regulation of endothelial cell apoptotic process | AKR1C3,ANGPT1,ANO6,BRAF,CD40LG,FASLG,FGA,FGB,FGG,FOXO3,GAS6,GATA3,ICAM1,IL13,IL4,KDR,KRIT1,NFE2L2,PDCD4,RGCC |
| T cell activation involved in immune response | APBB1IP,BCL3,CD1C,CD86,CLEC4D,CLEC4E,EIF2AK4,EOMES,FCER1G,FOXP1,GATA3,GPR183,HMGB1,ICAM1,IFNA1 |

| Mutant Breast Cancer Cell Sample13 | |
|---|---|
| Regulation of apoptotic signaling pathway | ACAA2,ACKR3,ACVR1,ADORA2A,AGT,AKT1,ANKRD2,ARHGEF2,ARRB2,AVP,BAD,BAG5,BAK1,BBC3 |
| positive regulation of immune response | ABL1,ACTB,ACTG1,ACTR2,ACTR3,ADA,ADAM8,ADORA2B,AIM2,ANXA1,ARPC1A,ARPC1B,ARPC2,ARPC3,ARPC5,BAG6,BAIAP2,BCAR1 |

| Muatant Breast Cancer Cell Sample12 | |
|---|---|
| Regulation of apoptotic signaling pathway | ACAA2,ACKR3,ACVR1,ADORA2A,AKT1,ANKRD2,APAF1,ARHGEF2,ARRB2,ATPIF1,AVP,BAD,BAG5,,CRADD,CREB3,CREB3L1 |
| positive regulation of immune response | ABI1,ABL1,ACTB,ACTG1,ACTR2,ACTR3,ADA,ADAM8,ADORA2B,AIM2,ANXA1,AP1G1,ARPC1A,ARPC1B,ARPC2,ARPC3,ARPC4,ARPC5 |

*Figure 3.6: Genes for apoptotic signaling and immunoregulation found through ontology in neighboring memory enhancers of Wild-type and Mutant Breast Cancer memory enhancers.*

## 3.3 Motif Analysis using HOMER

Homer comes with tools for analyzing sequencing data, including tools for quality checking, read mapping to a reference genome, differentially expressed genes identification, enriched motif identification, and transcription factor binding site identification. Examining regulatory elements in genomics applications (DNA only, no protein) was the motivation behind developing HOMER's unique motif discovery method. It uses two groups of sequences and a differential motif discovery algorithm to find the regulatory elements that are especially enriched in one set compared to the other. It combines ZOOPS scoring (zero or one occurrence per sequence) with hypergeometric enrichment computations (or binomial) to calculate motif enrichment.

Additionally, HOMER tries to take sequenced bias in the dataset into consideration. Although it was created with ChIP-Seq and promoter analysis in mind, it can be used to solve just about any nucleic acid pattern-finding issue. To identify the regulatory elements that are particularly abundant in one set of sequences when compared to the other, two groups of sequences are used, along with a differential motif identification technique. It combines binomial, hypergeometric enrichment calculations with a ZOOPS score (zero or one occurrence per sequence) to determine motif enrichment. HOMER makes a further effort to account for the dataset's sequencing bias. Although ChIP-Seq and promoter analysis were the two primary purposes for its creation, they may be applied to almost any nucleic acid pattern discovery problem.

In this study, Homer was used to analyzing the motifs of the transcription factors, which is bound to the memory enhancers having apoptotic and immunoregulatory genes in different cancer profile of subjects. These transcription factors could act as a master regulator promoting epigenetic memory, which can be targeted in patients for personalized therapy for cancer.

The following steps were used to analyze Motifs using HOMER:

a. *Preparing Input Data:* It needs to be prepared in the appropriate format before analyzing data with Homer. This involves converting your data into a suitable file format or

organizing it in a specific way.

b. *Running Motif Discovery:* The first step in motif analysis with Homer is to use its motif discovery tool to identify potential motifs in data. This tool uses various algorithms to identify patterns in data that are statistically significant.

c. *Motif Scanning:* Once potential motifs have been identified with the motif discovery tool, Homer's motif scanning tool is used to search your genomic sequences for these motifs. This can help identify potential transcription factor binding sites and other critical regulatory elements.

d. *Motif comparison:* Homer's motif comparison tool is used to compare the motifs that have been identified with known motifs in databases such as JASPAR or TRANSFAC. This helps to identify potential transcription factors that may be binding to the motifs.

e. *Motif annotation:* Homer's motif annotation tool annotates the motifs with additional information, such as gene ontology terms or pathway information. This helps in a better understanding of the functional significance of the motifs that have been identified.

| Rank | Motif | Name |
|---|---|---|
| 1 |  | HEB(bHLH)/mES-Heb-ChIP-Seq(GSE53233)/Homer |
| 2 |  | EBF1(EBF)/Near-E2A-ChIP-Seq(GSE21512)/Homer |
| 3 |  | Ascl1(bHLH)/NeuralTubes-Ascl1-ChIP-Seq(GSE55840)/Homer |
| 4 |  | GLIS3(Zf)/Thyroid-Glis3.GFP-ChIP-Seq(GSE103297)/Homer |
| 5 |  | E2A(bHLH)/proBcell-E2A-ChIP-Seq(GSE21978)/Homer |
| 6 |  | ZNF692(Zf)/HEK293-ZNF692.GFP-ChIP-Seq(GSE58341)/Homer |
| 7 |  | Ascl2(bHLH)/ESC-Ascl2-ChIP-Seq(GSE97712)/Homer |
| 8 |  | Prop1(Homeobox)/GHFT1-PROP1.biotin-ChIP-Seq(GSE77302)/Homer |
| 9 |  | THRb(NR)/HepG2-THRb.Flag-ChIP-Seq(Encode)/Homer |
| 10 |  | Myf5(bHLH)/GM-Myf5-ChIP-Seq(GSE24852)/Homer |
| 11 |  | c-Myc(bHLH)/mES-cMyc-ChIP-Seq(GSE11431)/Homer |

| Rank | Motif | Name |
|---|---|---|
| 1 |  | THRa(NR)/C17.2-THRa-ChIP-Seq(GSE38347)/Homer |
| 2 |  | NF1(CTF)/LNCAP-NF1-ChIP-Seq(Unpublished)/Homer |
| 3 |  | c-Myc(bHLH)/mES-cMyc-ChIP-Seq(GSE11431)/Homer |
| 4 |  | BMAL1(bHLH)/Liver-Bmal1-ChIP-Seq(GSE39860)/Homer |
| 5 |  | Usf2(bHLH)/C2C12-Usf2-ChIP-Seq(GSE36030)/Homer |
| 6 |  | DLX1(Homeobox)/BasalGanglia-Dlx1-ChIP-seq(GSE124936)/Homer |
| 7 |  | Tlx?(NR)/NPC-H3K4me1-ChIP-Seq(GSE16256)/Homer |
| 8 |  | Lhx3(Homeobox)/Neuron-Lhx3-ChIP-Seq(GSE31456)/Homer |
| 9 |  | Pdx1(Homeobox)/Islet-Pdx1-ChIP-Seq(SRA008281)/Homer |
| 10 |  | Pknox1(Homeobox)/ES-Prep1-ChIP-Seq(GSE63282)/Homer |

| Rank | Motif | Name |
|---|---|---|
| 1 |  | RBPJ:Ebox(?,bHLH)/Panc1-Rbpj1-ChIP-Seq(GSE47459)/Homer |
| 2 |  | IRF:BATF(IRF:bZIP)/pDC-Irf8-ChIP-Seq(GSE66899)/Homer |
| 3 |  | Reverb(NR),DR2/RAW-Reverba.biotin-ChIP-Seq(GSE45914)/Homer |
| 4 |  | ZNF675(Zf)/HEK293-ZNF675.GFP-ChIP-Seq(GSE58341)/Homer |
| 5 |  | E2F3(E2F)/MEF-E2F3-ChIP-Seq(GSE71376)/Homer |
| 6 |  | ZNF416(Zf)/HEK293-ZNF416.GFP-ChIP-Seq(GSE58341)/Homer |
| 7 |  | GATA3(Zf),DR4/iTreg-Gata3-ChIP-Seq(GSE20898)/Homer |
| 8 |  | TCFL2(HMG)/K562-TCF7L2-ChIP-Seq(GSE29196)/Homer |
| 9 |  | HIF-1a(bHLH)/MCF7-HIF1a-ChIP-Seq(GSE28352)/Homer |
| 10 |  | DUX(Homeobox)/C2C12-Dux-ChIP-Seq(GSE87279)/Homer |
| 11 |  | Tcf3(HMG)/mES-Tcf3-ChIP-Seq(GSE11724)/Homer |

*Figure 3.7: Motifs found in neighboring memory enhancers of primary pancreatic cancer cells T3, T6 & T23 respectively.*

| Rank | Motif | Name |
|---|---|---|
| 1 |  | Lhx2(Homeobox)/HFSC-Lhx2-ChIP-Seq(GSE48068)/Homer |
| 2 |  | HOXA1(Homeobox)/mES-Hoxa1-ChIP-Seq(SRP084292)/Homer |
| 3 |  | AP-2gamma(AP2)/MCF7-TFAP2C-ChIP-Seq(GSE21234)/Homer |
| 4 |  | NeuroD1(bHLH)/Islet-NeuroD1-ChIP-Seq(GSE30298)/Homer |
| 5 |  | Oct6(POU,Homeobox)/NPC-Pou3f1-ChIP-Seq(GSE35496)/Homer |
| 6 |  | Pdx1(Homeobox)/Islet-Pdx1-ChIP-Seq(SRA008281)/Homer |
| 7 |  | Tbx21(T-box)/GM12878-TBX21-ChIP-Seq(Encode)/Homer |
| 8 |  | Olig2(bHLH)/Neuron-Olig2-ChIP-Seq(GSE30882)/Homer |
| 9 |  | AP-2alpha(AP2)/Hela-AP2alpha-ChIP-Seq(GSE31477)/Homer |
| 10 |  | Phox2b(Homeobox)/CLBGA-PHOX2B-ChIP-Seq(GSE90683)/Homer |

| Rank | Motif | Name |
|---|---|---|
| 1 |  | BHLHA15(bHLH)/NIH3T3-BHLHB8.HA-ChIP-Seq(GSE119782)/Homer |
| 2 |  | Twist2(bHLH)/Myoblast-Twist2.Ty1-ChIP-Seq(GSE127998)/Homer |
| 3 |  | Ascl1(bHLH)/NeuralTubes-Ascl1-ChIP-Seq(GSE55840)/Homer |
| 4 |  | MyoD(bHLH)/Myotube-MyoD-ChIP-Seq(GSE21614)/Homer |
| 5 |  | Tcf12(bHLH)/GM12878-Tcf12-ChIP-Seq(GSE32465)/Homer |
| 6 |  | TCF4(bHLH)/SHSY5Y-TCF4-ChIP-Seq(GSE96915)/Homer |
| 7 |  | Ascl2(bHLH)/ESC-Ascl2-ChIP-Seq(GSE97712)/Homer |
| 8 |  | Tbx20(T-box)/Heart-Tbx20-ChIP-Seq(GSE29636)/Homer |
| 9 |  | Atoh1(bHLH)/Cerebellum-Atoh1-ChIP-Seq(GSE22111)/Homer |
| 10 |  | GRE(NR),IR3/A549-GR-ChIP-Seq(GSE32465)/Homer |

| Rank | Motif | Name |
|---|---|---|
| 1 |  | PU.1-IRF(ETS:IRF)/Bcell-PU.1-ChIP-Seq(GSE21512)/Homer |
| 2 |  | Hoxd13(Homeobox)/ChickenMSG-Hoxd13.Flag-ChIP-Seq(GSE86088)/Homer |
| 3 |  | Pit1+1bp(Homeobox)/GCrat-Pit1-ChIP-Seq(GSE58009)/Homer |
| 4 |  | DLX5(Homeobox)/BasalGanglia-Dlx5-ChIP-seq(GSE124936)/Homer |
| 5 |  | IIoxd12(IIomeobox)/ChickenMSG-Hoxd12.Flag-ChIP-Seq(GSE86088)/Homer |
| 6 |  | BHLHA15(bHLH)/NIH3T3-BHLHB8.HA-ChIP-Seq(GSE119782)/Homer |
| 7 |  | Tbr1(T-box)/Cortex-Tbr1-ChIP-Seq(GSE71384)/Homer |
| 8 |  | DLX2(Homeobox)/BasalGanglia-Dlx2-ChIP-seq(GSE124936)/Homer |
| 9 |  | Hoxd11(Homeobox)/ChickenMSG-Hoxd11.Flag-ChIP-Seq(GSE86088)/Homer |
| 10 |  | ELF3(ETS)/PDAC-ELF3-ChIP-Seq(GSE64557)/Homer |

*Figure 3.8: Motifs found in neighboring memory enhancers of metastatic pancreatic cancer cells M1L,M3P& M10P respectively.*

| Rank | Motif | Name |
|---|---|---|
| 1 |  | GLIS3(Zf)/Thyroid-Glis3.GFP-ChIP-Seq(GSE103297)/Homer |
| 2 |  | EBF1(EBF)/Near-E2A-ChIP-Seq(GSE21512)/Homer |
| 3 |  | Ascl1(bHLH)/NeuralTubes-Ascl1-ChIP-Seq(GSE55840)/Homer |
| 4 |  | CTCF(Zf)/CD4+-CTCF-ChIP-Seq(Barski_et_al.)/Homer |
| 5 |  | ZFX(Zf)/mES-Zfx-ChIP-Seq(GSE11431)/Homer |
| 6 |  | LRF(Zf)/Erythroblasts-ZBTB7A-ChIP-Seq(GSE74977)/Homer |
| 7 |  | BORIS(Zf)/K562-CTCFL-ChIP-Seq(GSE32465)/Homer |
| 8 |  | EBF2(EBF)/BrownAdipose-EBF2-ChIP-Seq(GSE97114)/Homer |
| 9 |  | ZNF711(Zf)/SHSY5Y-ZNF711-ChIP-Seq(GSE20673)/Homer |
| 10 |  | Ascl2(bHLH)/ESC-Ascl2-ChIP-Seq(GSE97712)/Homer |

| Rank | Motif | Name |
|---|---|---|
| 1 |  | E2A(bHLH),near_PU.1/Bcell-PU.1-ChIP-Seq(GSE21512)/Homer |
| 2 |  | EBF1(EBF)/Near-E2A-ChIP-Seq(GSE21512)/Homer |
| 3 |  | EBF2(EBF)/BrownAdipose-EBF2-ChIP-Seq(GSE97114)/Homer |
| 4 |  | PU.1:IRF8(ETS:IRF)/pDC-Irf8-ChIP-Seq(GSE66899)/Homer |
| 5 |  | Snail1(Zf)/LS174T-SNAIL1.HA-ChIP-Seq(GSE127183)/Homer |
| 6 |  | HEB(bHLH)/mES-Heb-ChIP-Seq(GSE53233)/Homer |
| 7 |  | Slug(Zf)/Mesoderm-Snai2-ChIP-Seq(GSE61475)/Homer |
| 8 |  | E2A(bHLH)/proBcell-E2A-ChIP-Seq(GSE21978)/Homer |
| 9 |  | Twist(bHLH)/HMLE-TWIST1-ChIP-Seq(Chang_et_al)/Homer |
| 10 |  | IRF8(IRF)/BMDM-IRF8-ChIP-Seq(GSE77884)/Homer |

| Rank | Motif | Name |
|---|---|---|
| 1 |  | KLF14(Zf)/HEK293-KLF14.GFP-ChIP-Seq(GSE58341)/Homer |
| 2 |  | Maz(Zf)/HepG2-Maz-ChIP-Seq(GSE31477)/Homer |
| 3 |  | PU.1-IRF(ETS:IRF)/Bcell-PU.1-ChIP-Seq(GSE21512)/Homer |
| 4 |  | Elf4(ETS)/BMDM-Elf4-ChIP-Seq(GSE88699)/Homer |
| 5 |  | PU.1(ETS)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer |
| 6 |  | Zfp281(Zf)/ES-Zfp281-ChIP-Seq(GSE81042)/Homer |
| 7 |  | ZNF467(Zf)/HEK293-ZNF467.GFP-ChIP-Seq(GSE58341)/Homer |
| 8 |  | Egr1(Zf)/K562-Egr1-ChIP-Seq(GSE32465)/Homer |
| 9 |  | Etv2(ETS)/ES-ER71-ChIP-Seq(GSE59402)/Homer |
| 10 |  | EHF(ETS)/LoVo-EHF-ChIP-Seq(GSE49402)/Homer |

*Figure 3.9: Motifs found in neighboring memory enhancers of Wild-type breast cancer cells Sample 3, Sample 5 & Sample 6 respectively.*

| Rank | Motif | Name |
|---|---|---|
| 1 |  | ZEB1(Zf)/PDAC-ZEB1-ChIP-Seq(GSE64557)/Homer |
| 2 |  | E2A(bHLH),near_PU.1/Bcell-PU.1-ChIP-Seq(GSE21512)/Homer |
| 3 |  | ZEB2(Zf)/SNU398-ZEB2-ChIP-Seq(GSE103048)/Homer |
| 4 |  | HIC1(Zf)/Treg-ZBTB29-ChIP-Seq(GSE99889)/Homer |
| 5 |  | Pax7(Paired,Homeobox),long/Myoblast-Pax7-ChIP-Seq(GSE25064)/Homer |
| 6 |  | Hoxc9(Homeobox)/Ainv15-Hoxc9-ChIP-Seq(GSE21812)/Homer |
| 7 |  | Rfx2(HTH)/LoVo-RFX2-ChIP-Seq(GSE49402)/Homer |
| 8 |  | SCRT1(Zf)/HEK293-SCRT1.eGFP-ChIP-Seq(Encode)/Homer |
| 9 |  | PBX2(Homeobox)/K562-PBX2-ChIP-Seq(Encode)/Homer |
| 10 |  | RFX(HTH)/K562-RFX3-ChIP-Seq(SRA012198)/Homer |

| Rank | Motif | Name |
|---|---|---|
| 1 |  | KLF14(Zf)/HEK293-KLF14.GFP-ChIP-Seq(GSE58341)/Homer |
| 2 |  | Maz(Zf)/HepG2-Maz-ChIP-Seq(GSE31477)/Homer |
| 3 |  | ZNF467(Zf)/HEK293-ZNF467.GFP-ChIP-Seq(GSE58341)/Homer |
| 4 |  | RXR(NR),DR1/3T3L1-RXR-ChIP-Seq(GSE13511)/Homer |
| 5 |  | Egr1(Zf)/K562-Egr1-ChIP-Seq(GSE32465)/Homer |
| 6 |  | PPARE(NR),DR1/3T3L1-Pparg-ChIP-Seq(GSE13511)/Homer |
| 7 |  | Sp5(Zf)/mES-Sp5.Flag-ChIP-Seq(GSE72989)/Homer |
| 8 |  | ETS1(ETS)/Jurkat-ETS1-ChIP-Seq(GSE17954)/Homer |
| 9 |  | PU.1(ETS)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer |
| 10 |  | WT1(Zf)/Kidney-WT1-ChIP-Seq(GSE90016)/Homer |

| Rank | Motif | Name |
|---|---|---|
| 1 |  | KLF14(Zf)/HEK293-KLF14.GFP-ChIP-Seq(GSE58341)/Homer |
| 2 |  | BATF(bZIP)/Th17-BATF-ChIP-Seq(GSE39756)/Homer |
| 3 |  | Fra1(bZIP)/BT549-Fra1-ChIP-Seq(GSE46166)/Homer |
| 4 |  | Maz(Zf)/HepG2-Maz-ChIP-Seq(GSE31477)/Homer |
| 5 |  | JunB(bZIP)/DendriticCells-Junb-ChIP-Seq(GSE36099)/Homer |
| 6 |  | Fos(bZIP)/TSC-Fos-ChIP-Seq(GSE110950)/Homer |
| 7 |  | Atf3(bZIP)/GBM-ATF3-ChIP-Seq(GSE33912)/Homer |
| 8 |  | Fosl2(bZIP)/3T3L1-Fosl2-ChIP-Seq(GSE56872)/Homer |
| 9 |  | Fra2(bZIP)/Striatum-Fra2-ChIP-Seq(GSE43429)/Homer |
| 10 |  | AP-1(bZIP)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer |

*Figure 3.10: Motifs found in neighboring memory enhancers of Mutant breast cancer cells Sample 9,Sample 12 & Sample 13 respectively.*

# Chapter 4: Conclusion & Future Scope

## 4.1 Conclusion

Epigenetic memory refers to the ability of cells to remember and maintain gene expression patterns established during development or in response to environmental cues. Epigenetic modifications, such as DNA methylation and histone modifications, can influence gene expression by regulating the accessibility of DNA to the transcription machinery. These modifications can be passed on to daughter cells during cell division, allowing them to inherit the same gene expression patterns as the parent cell.

Epigenetic memory is crucial in the context of development, as it allows cells to differentiate into specific cell types and maintain their identity throughout the organism's lifespan. During early embryonic development, cells undergo a series of epigenetic modifications that bring about specific gene expression patterns required for forming different tissues and organs. Once established, these gene expression patterns can be maintained through cell division and passed to daughter cells. Epigenetic memory is also thought to play a role in disease development. Aberrant epigenetic modifications can result in gene expression alterations related to various diseases. Our study shows that H3K4me1 is critical in promoting epigenetic memory in cancer progression. This mark helps to recruit specific proteins and transcription factors to the enhancer site, where they can interact with the DNA and nearby promoters to increase or decrease gene expression. This way, H3K4me1 regulates gene expression and is an essential component of the epigenetic landscape that controls gene expression in diseases; cancer profiles of different subjects were used to study the importance of H3K4me1 in the cancer progression, sites having H3K4me1 but no DNA hypersensitivity is tagged as sites of epigenetic memory, These sites were further studied for the hallmarks of Cancer namely immunoregulation promotion of apoptosis. Master regulators or transcription factors were found from these memory sites, which could be targeted with specific

epigenetic modifiers such as HDAC is or DNMTs for personalized precision therapy in Cancer patients.

Overall, epigenetic memory is an important mechanism that allows cells to maintain stable gene expression patterns over time and respond to changing environmental cues. Understanding how epigenetic memory is established and maintained is crucial for understanding normal development and disease pathogenesis. Our study suggests that targeting epigenetic memory sites could be potential therapeutics for cancer treatment.

## 4.2 Future Scope

Analyzing more data with different profiles of Cancer and other diseases to find the significance of the memory site (H3K4me1 without open-chromatin profile), and through which we might discover a new state. This may provide some data for ML etc. Study other residues that could act as potential epigenetic marks, such as a few small RNAs, few types of enzymes - like Histone modifying enzyme chromatin structure- based proximity, etc.

# Bibliography


[1] C. H. Waddington *et al.*, "An introduction to modern genetics." *An introduction to modern genetics.*, 1939.

[2] C. Martin and Y. Zhang, "Mechanisms of epigenetic inheritance," *Current opinion in cell biology*, vol. 19, no. 3, pp. 266–272, 2007.

[3] S. Henikoff and J. M. Greally, "Epigenetics, cellular memory, and gene regulation," *Current Biology*, vol. 26, no. 14, pp. R644–R648, 2016.

[4] S. Thiagalingam, "Epigenetic memory in development and disease: Unraveling the mechanism," *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, vol. 1873, no. 2, p. 188349, 2020.

[5] A. T. Annunziato, "Split decision: What happens to nucleosomes during DNA replication?*," *Journal of Biological Chemistry*, vol. 280, no. 13, pp. 12 065–12 068, 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021925819605272

[6] H. L. Kaufman, S. Wadler, and K. Antman, *Molecular targeting in oncology*. Springer Science & Business Media, 2007.

[7] T. Kouzarides, "Chromatin modifications and their function," *Cell*, vol. 128, no. 4, pp. 693–705, 2007.

[8] A. Santenard and M.-E. Torres-Padilla, "Epigenetic reprogramming in mammalian reproduction: contribution from histone variants," *Epigenetics*, vol. 4, no. 2, pp. 80–84, 2009.

[9] M. A. Hahn, P. E. Szab'o, and G. P. Pfeifer, "5-hydroxymethylcytosine: a stable or transient dna modification?" *Genomics*, vol. 104, no. 5, pp. 314–323, 2014

[10] J. Wang, S. Rao, J. Chu, X. Shen, D. N. Levasseur, T. W. Theunissen, and S.H. Orkin, "A protein interaction network for pluripotency of embryonic stem cells," *Nature*, vol. 444, no. 7117, pp. 364–368, 2006.

[11] B. Schuettengruber, H.-M. Bourbon, L. Di Croce, and G. Cavalli, "Genome

regulation by polycomb and trithorax: 70 years and counting," *Cell*, vol. 171, no. 1, pp. 34–57, 2017.

[12] L. A. Boyer, K. Plath, J. Zeitlinger, T. Brambrink, L. A. Medeiros, T. I. Lee, S. S. Levine, M. Wernig, A. Tajonar, M. K. Ray *et al.*, "Polycomb complexes repress developmental regulators in murine embryonic stem cells," *nature*, vol. 441, no. 7091, pp. 349–353, 2006.

[13] B. E. Bernstein, T. S. Mikkelsen, X. Xie, M. Kamal, D. J. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig, K. Plath *et al.*, "A bivalent chromatin structure marks keydevelopmental genes in embryonic stem cells," *Cell*, vol. 125, no. 2, pp. 315– 326, 2006.

[14] M. A. Surani, K. Hayashi, and P. Hajkova, "Genetic and epigenetic regulators of pluripotency," *Cell*, vol. 128, no. 4, pp. 747–762, 2007.

[15] J. M. Slack, "Metaplasia and transdifferentiation: from pure biology to the clinic," *Nature reviews Molecular cell biology*, vol. 8, no. 5, pp. 369–378, 2007.

[16] N. D. Heintzman, G. C. Hon, R. D. Hawkins, P. Kheradpour, A. Stark, L. F. Harp, Z. Ye, L. K. Lee, R. K. Stuart, C. W. Ching *et al.*, "Histone modifications at human enhancers reflect global cell-type-specific gene expression," *Nature*, vol.459, no. 7243, pp. 108–112, 2009.

[17] S. Sharma, T. K. Kelly, and P. A. Jones, "Epigenetics in cancer," *Carcinogenesis*,vol. 31, no. 1, pp. 27–36, 2010.

[18] M. Esteller and J. G. Herman, "Cancer as an epigenetic disease: DNA methylation and chromatin alterations in human tumours," *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, vol. 196, no. 1, pp. 1–7, 2002.

[19] Y. Lu, Y.-T. Chan, H.-Y. Tan, S. Li, N. Wang, and Y. Feng, "Epigenetic regulation in human cancer: the potential role of epi-drug in cancer therapy," *Molecular cancer*, vol. 19, pp. 1–16, 2020.

[20] I. Huh, X. Wu, T. Park, and S. V. Yi, "Detecting differential DNA methylation from sequencing of bisulfite converted DNA of diverse species," *Briefings in bioinformatics*, vol. 20, no. 1, pp. 33–46, 2019.

[21] D. Liu, X. Meng, D. Wu, Z. Qiu, and H. Luo, “A natural isoquinoline alkaloid with antitumor activity: studies of the biological activities of berberine,” *Frontiers in pharmacology*, vol. 10, p. 9, 2019.

[22] W. Jiang, T. Xia, C. Liu, J. Li, W. Zhang, and C. Sun, "Remodeling the epi- the genetic landscape of cancer—application potential of flavonoids in the prevention and treatment of cancer,” *Frontiers in Oncology*, p. 2331, 2021.

[23] H. Zhang, Z. Lang, and J.-K. Zhu, “Dynamics and function of DNA methylation in plants,” *nature reviews Molecular cell biology*, vol. 19, no. 8, pp. 489–506, 2018.

[24] M. Esteller, “Cancer epigenomics: DNA methylomes and histone-modification maps,” *Nature reviews genetics*, vol. 8, no. 4, pp. 286–298, 2007.

[25] J. Wysocka, T. A. Milne, and C. D. Allis, “Taking lsd1 to a new high,” *Cell*, vol. 122, no. 5, pp. 654–658, 2005.

[26] K. J. Falkenberg and R. W. Johnstone, “Histone deacetylases and their inhibitors in cancer, neurological diseases, and immune disorders," *Nature Reviews Drug discovery*, vol. 13, no. 9, pp. 673–691, 2014.

[27] D.-G. Fu, “Epigenetic alterations in gastric cancer,” *Molecular medicine reports*, vol. 12, no. 3, pp. 3223–3230, 2015.

[28] A. Local, H. Huang, C. P. Albuquerque, N. Singh, A. Y. Lee, W. Wang, C. Wang, J. E. Hsia, A. K. Shiau, K. Ge *et al.*, “Identification of h3k4me1-associated proteinsat mammalian enhancers,” *Nature Genetics*, vol. 50, no. 1, pp. 73–82, 2018.

[29] T. Bleckwehl, G. Crispatzu, K. Schaaf, P. Respuela, M. Bartusel, L. Benson, S. J. Clark, K. M. Dorighi, A. Barral, M. Laugsch *et al.*, “Enhancer-associated h3k4 methylation safeguards in vitro germline competence,” *Nature Communications*, vol. 12, no. 1, p. 5771, 2021.

[30] F. S. Collins and H. Varmus, “A new initiative on precision medicine,” *New England journal of Medicine*, vol. 372, no. 9, pp. 793–795, 2015.

[31] S. H. Shin, A. M. Bode, and Z. Dong, "Precision medicine: the foundation of future cancer therapeutics,” *Npj precision oncology*, vol. 1, no. 1, p. 12, 2017.

[32] S. Aspeslagh, D. Morel, J.-C. Soria, and S. Postel-Vinay, “Epigenetic modifiers as

new immunomodulatory therapies in solid tumors," *Annals of Oncology*, vol. 29, no. 4, pp. 812–824, 2018.

[33] K. B. Chiappinelli, C. A. Zahnow, N. Ahuja, and S. B. Baylin, "Combining epigenetic and immunotherapy to combat cancer combining epigenetic and immunotherapy to combat cancer," *Cancer Research*, vol. 76, no. 7, pp. 1683– 1689, 2016.