Dissertation
submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

presented by

Qi Wang, M.Sc.
Born in Liaoning, China
Oral-examination: 8 Oct. 2018

Integrative methods for epigenetic profiling in cancer and development

Referees:	Prof. Dr. Roland Eils
	Prof. Dr. Benedikt Brors

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Qi Wang
August 2018

Place, Date

Signature

Abstract

DNA mutation, epigenetic alteration, and gene expression are three major molecular components that distinguish cancer from normal cells. Although it is widely accepted that epigenetic modifications can greatly affect the expression of the target genes, because of the complex combinations of epigenetic marks, together with the interactions between multiple non-coding regulatory elements, measuring the epigenetic effects on gene expression is not an easy task. Nevertheless, it is estimated that epigenetic modifications have a greater effect than DNA mutations on tumorigenesis. In addition, epigenetic alterations are the initiating factor in some chromosome abnormalities and aberrant gene expression, making the study of epigenetic alterations a central aspect in understanding the underlying mechanisms in cancer and cell development.

The aim of this thesis is to conduct qualitative and quantitative analysis on differential epigenetic modifications. To this end, a variety of existing approaches were applied in the ChIP-Seq analyses of six histone marks on glioblastoma data from four distinct subtypes. The results depict a comprehensive landscape of active and poised regulatory elements specific to glioblastoma subtypes, which describe the different aspects of tumor progression. However, the descriptive model of multiple histone marks (ChromHMM and peak calls) was also shown to be prone to various biases and artifacts. Moreover, some models also neglect the quantitative information of ChIP-Seq data, making it inadequate in addressing the magnitude of epigenetic modifications in gene expression levels. Therefore, in the second part of my work, I designed an integrated, network based approach, in which I integrated two levels of epigenetic information: the signal intensities of each epigenetic mark, and the relationships between promoters and distal regulatory elements known as enhancers. Applying this approach to a variety of test cases, it predicts a number of candidate genes with significant epigenetic alterations, and comprehensive benchmarking validated these findings in cancer and stem cell developments.

In summary, as increasing amounts of epigenetic data become available, the computational approaches employed in this study would be highly relevant in both comparative and integrative analysis on the epigenetic landscape. The discovery of novel epigenetic targets in cancers, not only unfolds the fundamental mechanisms in tumorigenesis and development, but also serves as an emerging resource for molecular diagnosis and treatment.

Zusammenfassung

DNA-Mutationen, epigenetische Veränderungen und Genexpression sind drei wichtige molekulare Eigenschaften, die Krebszellen von normalen Zellen unterscheiden. Es ist allgemein anerkannt, dass epigenetische Veränderungen die Expression der Zielgene stark beeinflussen können. Aufgrund der komplexen Kombinationen von epigenetischen Markierungen und der Wechselwirkungen zwischen mehreren nicht-kodierenden regulatorischen Elementen bleibt die Bestimmung der epigenetischen Effekte auf Genexpression eine Herausforderung. Es wird angenommen, dass epigenetische Veränderungen eine größere Auswirkung als DNA-Mutationen auf die Tumorgenese haben. Darüber hinaus liegen bei Chromosomenanomalien und anomaler Genexpression oft epigenetische Alterationen zugrunde, was die Untersuchung von epigenetischen Mechanismen zu einer zentralen Frage für das Versändnis der Krebs- und Zellentwicklung macht. Ziel dieser Studie ist es, qualitative und quantitative Untersuchungen zu differentiellen epigenetischen Modifikationen durchzuführen. Zu diesem Zweck wurde eine Vielzahl von existierenden Ansätzen in den ChIP-Seq-Analysen von sechs Histonmarkierungen von Glioblastomdaten aus vier verschiedenen Subtypen angewendet. Die Ergebnisse zeigen eine umfassende Landschaft aktiver und ruhender regulatorischer Elemente, die spezifisch für bestimmte Glioblastom-Subtypen sind. Die Modelle für Histonmarkierungen (ChromHMM- und Peak-Calls) erwiesen sich jedoch in dieser Studie ebenfalls als anfällig für Verzerrungen und Artefakte. Darüber hinaus vernachlässigt das ChromHMM-Modell auch die quantitative Information von ChIP-Seq-Daten und macht es somit ungeeignet, das Ausmaß epigenetischer Modifikationen in den Genexpressionsniveaus zu berücksichtigen. Aus diesem Grund habe ich im zweiten Teil meiner Arbeit ein generisches Modell für integrative Untersuchungen erstellt. Mit diesem Modell können zwei Ebene epigenetischer Informationen, nämlich die die Signalintensitäten jeder Histonmarkierung und die Zusammenhänge von Enhancern und Promotoren berücksichtigt werden. Mein Ansatz sagt eine Reihe von Kandidaten mit signifikanten epigenetischen Veränderungen voraus, und in einem umfassenden Benchmarking mit einer Vielzahl von epigenetischen Datensätzen konnten diese Ergebnisse in Krebs- und Stammzellentwicklungen bioinformatisch validiert werden. Die Entdeckung neuartiger epigenetischer Targets bei Krebserkrankungen beleuchtet nicht nur die grundlegenden Mechanismen der Tumorgenese und -entwicklung, sondern dient auch als Quelle für die molekulare Diagnose und Behandlung.

Contents

List of Abbreviations

4C	Circularized chromosome conformation capture
ANOVA	Analysis of variance
AUC	Area under ROC curve
BAM	Binary alignment map
BED	Browser extensible data
BIC	Bayesian information criterion
bp	base pair
CGI	CpG islands
ChIA-PET	Chromatin interaction analysis with paired-end tag sequencing
ChIP	Chromatin immunoprecipitation
ChIP-Seq	Chromatin immunoprecipitation sequencing
CLL	Chronic lymphocytic leukemia
CRC	Colorectal cancer
DIPG	Diffuse intrinsic pontine glioma
DNA	Deoxyribonucleic acid
ECDF	Empirical cumulative distribution function
EMT	Epithelial-mesenchymal transition
ESC	Embryonic stem cell
GBM	Glioblastoma multiforme
GO	Gene ontology
GRCh37/hg19	Genome reference consortium human reference 37
GRCh38/hg38	Genome reference consortium human reference 38
GSC	Glioma stem cells
H3K27ac	histone H3 acetylation at lysine 27
H3K27me3	histone H3 trimethylation at lysine 27
H3K36me3	histone H3 trimethylation at lysine 36
H3K4me1	histone H3 monomethylation at lysine 4
H3K4me3	histone H3 trimethylation at lysine 4
H3K9me3	histone H3 trimethylation at lysine 9
HKG	Housekeeping gene
HMM	Hidden markov model
HPRD	Human protein reference database
kb	kilobases
LGG	Low level glioma
Mb	megabases
MSC	Mesenchymal stem cell
NPC	Neural progenitor cell
OG	Oncogene

P-E	Promoter-enhancer
PTC	Papillary thyroid cancer
RNAPII	RNA polymerase II
ROC	Receiver operating characteristic
SAM	Sequence alignment map
SUMO	Small ubiquitin-like modifier
TAD	Topological associated domain
TCGA	The cancer genome atlas
TF	Transcription factor
TSC	Trophoblast stem cell
TSG	Tumor suppressor gene
TSS	Transcription start site
WGBS	whole-genome bisulfite sequencing

Chapter 1

Introduction

1.1 General concepts in epigenetics regulation

Eukaryotes have a much more complex chromosomal structure than prokaryotes, which offers more layers of transcriptional regulation. In particular, the non-coding genome size in eukaryotes is orders of magnitude larger that in prokaryotes, suggesting that the complexities of the transcriptional regulation, rather than the number of genes, relate to the organismal complexity. Epigenetics is one of the mechanisms which allows a cell to alter transcription without changing the DNA sequences, and is usually reversible [1]. This is beneficial by enabling the cell to quickly adapt to the needs of development and rapid changes of the environment. But abnormal epigenetic modifications may also cause persistent activating of cell cycle control genes or deactivating of DNA repair genes and result in permanent growth of the cell, and in the end lead to cancer. Here I will explain from the 4W (what, where, when, who) aspects of epigenetic regulation which compose the fundamentals of this study.

1.1.1 What are the components in epigenetic regulation?

In this section, I will describe the main types of epigenetic modifications, leaving the description of the molecular actors of these changes to section (1.1.4).

DNA methylation

DNA methylation transforms cytosine into 5-methylcytosine in CpG dinucleotides. This transformation occurs through the action of the DNA-methyltransferase enzymes, namely Dnmt1, Dnmt3a and Dnmt3b. The first enzyme is responsible for DNA methylation maintenance in replication, while the latter two as responsible for *de novo* methylation of unmethylated DNA.

Many genes are transcriptionally repressed by CpG methylation at their regulatory domains [2]. Global hypomethylation is a common feature in carcinogenesis [3], e.g. colorectal [4] and ovarian [5] cancer. On the other hand, transcription of many tumor suppressor genes are inhibited by hypermethylation [6], affecting many pathways such as apoptosis (DAPK), cell cycle (p16), cell adherence (CDH1, CDH13), DNA repair (hMLH1, MGMT), and detoxification (GSTP1) etc. [7]. As an example, ERα gene was progressively hypermethylated while global DNA became hypomethylated during colorectal cancer development [8].

DNA hypermethylation in the gene bodies, on the contrary, is often associated with high transcription [9, 10], and is most frequently observed in housekeeping genes [11]. The gene body methylation can recruit spliceosomal proteins [12] and is probably related to regulation of alternative splicing [12–14].

Histone modification

The core histone proteins H2A, H2B, H3 and H4 each form dimers and combine into an octameric structure together with the linker H1 protein. This structure is wrapped by a $\sim$ 147 bp sequence of DNA. This unit is called *nucleosome*, which constitutes a basic component in chromatin.

Covalent modifications on the residues of the histone proteins affect the accessibility and activity of the chromatin DNA, by modifying the chemical and physical properties of the protein-DNA interaction. Histone methylation and acetylation were found more than 40 years ago [15]. Later, modifications occurring at the lysine, serine, threonine and arginine residues of histone proteins, and a variety of modifications including methylation, acetylation, phosphorylation, citrullination, ubiquitination, SUMOylation and ADP-ribosylation were discovered (Fig. 1.1).

Histone proteins are very conserved across mammals, and their mutations are associated with several cancers, such as diffuse intrinsic pontine glioma (DIPG) [16] and chondroblastoma [17]. Dysregulation of histone modifying enzymes can lead to oncogenesis, making them potential targets for cancer drugs.

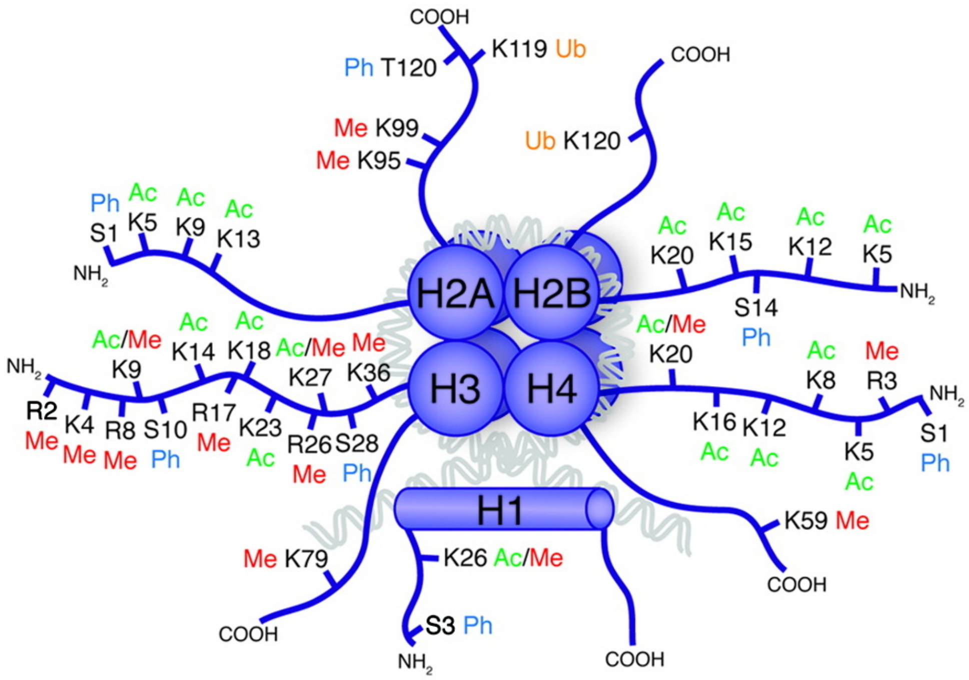


Figure 1.1: Histone modifications and organization of nucleosome. Methylation (Me), acetylation (Ac), ubiquitination (Ub), and phosphorylation (Ph) mostly occur at the N-terminal tail of the histones. Adapted from Epigenetics. 2012;7(8):823-40 [18].

Chromosomal conformation changes

Chromosome has a highly organized three-dimensional architecture, arranging distal regulatory elements in spatial vicinity of their target gene promoters. Chromosome condensation changes have been well characterized throughout the various stages of cell cycle. Recently, mounting evidences suggest that conformation changes are also accompanied with cancer progression, cause formation of aberrant chromosomal loops [19], and lead to homologous recombination [20] and probably gene fusion, loss of imprinting [21], or activation of oncogene transcription [22]. A few mediators, such as ERG overexpression, are known to induce global chromatin conformation reorganization in prostate cancer [23] and breast cancer [24] cell lines. The rearrangement of chromosomal conformation results spatial proximity of ERα binding loci, and promote cancer proliferation [24]. Therefore, profound knowledge of chromatin conformation changes is needed in further understanding cancer pathology.

Topologically associating domain (TAD) are defined based on chromosome conformation, in which genomic loci within TAD have remarkable higher contact probabilities. Cohesin and CTCF co-localize at TAD boundaries that insulate chromatin contacts from outside of the TAD [25].

1.1.2 Where do epigenetic modifications take place?

Promoter

The gene promoter is a genomic region at which the transcription of the gene is initiated. The eukaryotic core promoter contains a RNA polymerase binding site located at $\sim$ 34bp upstream of the transcription start site (TSS) [26, 27], whereas specific transcription factor binding sites are located $\sim$ 250 bp upstream of the TSS. Often, the promoter definition is extended to up to ±2kb around the TSS that may contain additional regulatory sequences [28]. In this study, I will generally adopt this extended definition.

It is estimated that 60%-70% of human gene promoters harbor CpG islands (CGI) [29, 30]. The CpGs in these CGIs are predominantly non-methylated [31]. However, in certain conditions (for example certain cancer such as glioblastoma), these CGIs can be hypermethylated, which results in chromatin compaction and leads to transcriptional repression [2]. In addition to repression due to DNA methylation, these CGIs can also be repressed via polycomb-mediated repression [31]. This alternative mechanism of repression leads to the trimethylation of lysine 27 on histone H3 via the action of the enzyme EZH2, resulting in a closed chromatin conformation.

Actively transcribed promoters show histone H3 lysine 4 trimethylation (H3K4me3), which is recognized by the plant homeodomain (PHD) finger of the TFIID subunit TAF3 [32], wherein H3K9ac and H3K14ac enhanced TFIID interaction [33]. H3K4me3 also recruits the nucleosome remodeling factor (NURF) [34] and pre-mRNA splicing protein CHD1 [35] to facilitate transcription elongation and splicing. Recently, the broadness of H3K4me3 was also found to play a role in the transcription activity of promoters [36].

Histone H3 acetylation at lysine 27 (H3K27ac) is enriched at the promoters of transcriptionally active genes [37]. As opposed to H3K27ac, H3 trimethylation at lysine 27 (H3K27me3) at the promoters inhibit transcription. Co-occurrences of H3K4me3 and H3K27me3, two modifications with opposing effects, are often observed in cultured embryonic stem cells [38–40] and are termed as "bivalent domain". The bivalent configuration reduces transcription activity, yet allows timely activation upon differentiation signals [39].

Enhancer

Enhancers are short *cis*-elements that interact with promoter through chromosomal loops to increase gene transcription [41, 42]. Most of the enhancers are located within ±1 Mb of the transcription start site (TSS) of their target genes [43], including intergenic but also intronic regions [44]. Enhancers can be bound by specific transcription factors which either activate or repress the binding of general transcription factors (GTFs) and RNA polymerase II (RNAPII) on promoters [45, 46]. The specific combination, ordering and spacing of the binding sites are believed to play an important role for the precise activity of these regulatory elements. The DNA sequences of enhancer, as opposed to promoters, show poor conservation across species [47] and tissues [48].

Enhancer activity in time and across conditions and tissues is regulated through the epigenetic modifications, including H3K27ac [49] and H3K4me1 [50, 51]. In particular, cell type specificity of enhancers is unveiled from the unique H3K4me1 patterns in different mammalian cells [51]. Other markers, such as the presence of DNase I hypersensitive sites [52] or binding of the transcriptional co-activator p300 [53] are used to identify enhancers. Similar to promoters, enhancers marked with H3K27ac are termed active [54], and enhancers with H3K27me3 are bivalent [55]. Recent evidence suggests that many enhancers are led to the transcription of non-coding RNAs (eRNA) [56], which can be used as another mark of enhancer activity. Finally, numerous recent studies have shown that enhancers are often clustered together in broader domains to form ”super-enhancer”, resulting in much higher activities of RNAPII binding and eRNA transcription than individual enhancers [57, 58]. In embryonic stem cells, super-enhancers are regulated by a small number of genes so called ”master regulator”. Super-enhancers are also found in other cell types, where they play a major role the control of cell identity and regulation of cell type-specific genes [59].

Insulator

Insulators are another type of *cis*-regulatory elements, which are rich in repeated sequences such as CCCTC. Insulators are often found to be bound by the transcriptional repressor CTCF. Together with cohesin, they form a complex at the boundaries of topologically-associated domain (TAD). Alterations in their sequences and aberrant expression or dysfunction of CTCF may drive cancer, by altering the TAD structure and enabling the aberrant regulation of oncogenes by distant enhancers. However, other elements are also needed for TAD formation, such as TFIIIC [60]. Depending on the type of these additional proteins, the complex can have three regulatory activities: (1) isolating the active chromatin from the repressive chromatin [61], in which USF1 complex form a barrier between the euchromatin and heterochromatin [62]. (2) blocking enhancer from interacting with promoter [63]. (3) promoting enhancer-promoter interactions, which CTCF forms TAF3/CTCF/cohesin complex at the core promoter region [64].

1.1.3 When are epigenetic modifications altered?

Environmental exposures

Several environmental exposures, such as carcinogens, infections, nutrition can affect epigenome modifiers, resulting in epigenetic dysregulation in a variety of cells, and ultimately increasing the risk of carcinogenesis [65, 66].

Environmental contaminants may contain both genotoxic and non-genotoxic agents, such as pirinixic acid (WY-14643) [67], trichloroacetic acid and dichloroacetic acid [68], which cause global DNA hypomethylation. Metals, for instance, nickel (Ni), arsenic (As), lead (Pd), chromium (Cr), cadmium (Cd), have been reported to affect both DNA methylation (Ni [69,70], As [71], Pd [72,73], Cr [74, 75], Cd [76, 77]) and histone modification (Ni [78–82], As [83–87], Pb [73], Cr [88,89]). In addition, benzene exposure plays a role in both acute myeloid leukemia (AML) [90] and chronic myeloid leukemia (CML) [91] via both hypomethylation and hypermethylation [92].

Bacteria, such as *Helicobacter pylori*, can affect the human epigenome in two ways: either by inducing global hypomethylation [93, 94], or promoting mutagenesis with mutagen like N-methyl-N-nitrosourea [95]. Viruses, such as Epstein–Barr virus (EBV), mediate via tumor supressor BIM [96] and PRDM1 [97] silencing through hypermethylation in EBV-positive Burkitt's lymphoma. Similarly, hepatitis B viruses (HBV) has been found to induce hypermethylation of several genes (RASSF1A, GSTP1, CHRNA3, and DOK1) in hepatocellular carcinoma (HCC) [98]. In addition, human papillomavirus (HPV) infection is associated with promoter hypermethylation of a variety of tumor suppressor genes, such as p16, CDH1, RARβ, MGMT, DAPK, DCC, GALR1, and GALR2 [99–101] in head and neck squamous cell carcinoma (HNSCC). Mechanisms behind these aberrant DNA methylation patterns may be attributed to overexpression of DNA methyltransferase (DNMT) during viral infection [102].

Nutrition has also been shown to impact the epigenome. Maternal diet can impact the baby's epigenetic status in *in utero*, particularly high-fat diet increases obesity risk in the offspring [103]. A study has shown how maternal uptake of folic acid altered the epigenome in an agouti mice model, causing a shift in the coat color of offspring [104] (Fig. 1.2). In another example, maternal diet with genistein, a phytoestrogen from soy, protected the offsprings of agouti mice from inherited obesity through epigenetic alterations [105].

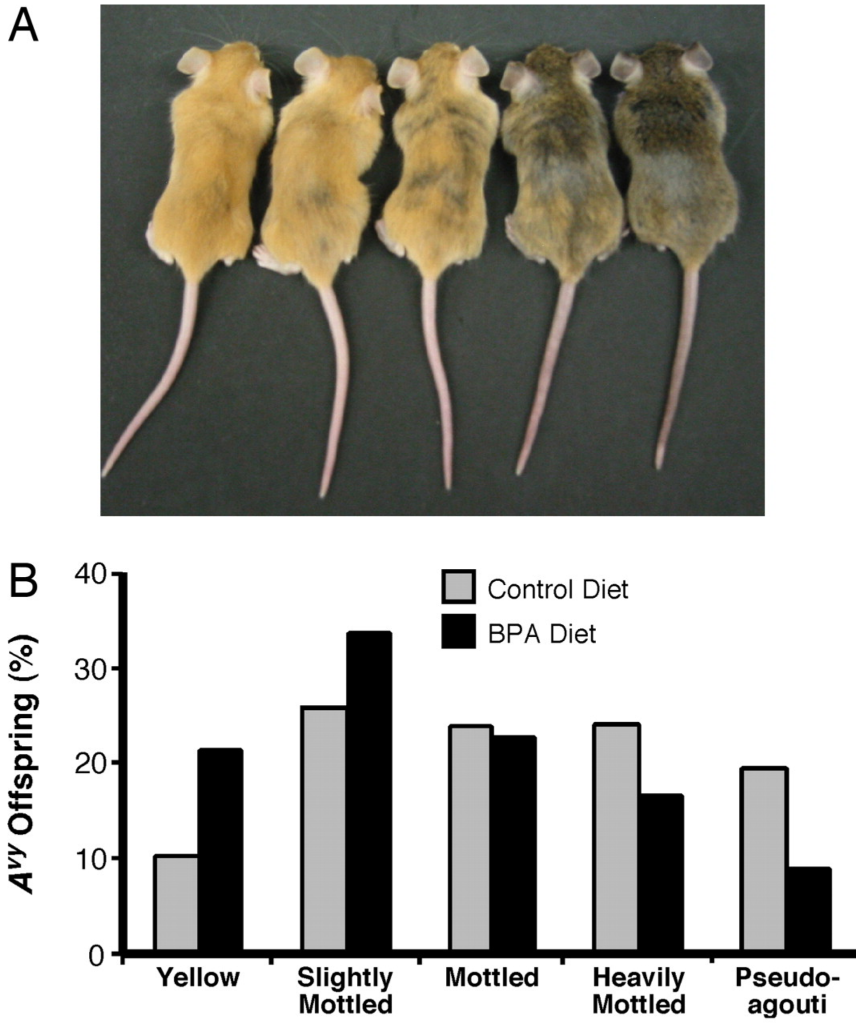


Figure 1.2: Maternal diet with different doses of methyl donors partially restored DNA methylation related to agouti gene, causing a shift in offspring coat colors towards normal. Taken from Proc Natl Acad Sci USA. 2007;104(32):13056-61 [104].

Development

Epigenetic reprogramming takes place in the early stages of mammalian embryo development. Still, a substantial number of epigenetic loci is not altered [106]. For example, hypomethylation of a transposable element upstream of the gene A in agouti mice is kept in offsprings, resulting in inheritance of ectopic expression of the Agouti protein [107]. Also, DNA methylation is almost completely erased during two phases in mammalian embryo development, namely gametogenesis (in oocyte or sperm) and early embryogenesis [108]. Then, genes except housekeeping and tissue-specific genes [109] are re-methylated. Specifically in females, X-chromosome inactivation occurs in a random X chromosome through the mediation of X-inactive specific transcript (Xist) non-coding gene. Promoter CGIs are reported to be heavily methylated on the inactive X chromosome in somatic cells [110, 111], whereas the other non-CGI CpGs are mostly unmethylated on the inactive X chromosome [112]. Heterochromatin is established on the inactivated X chromosome, shutting down the expression of most of the genes [113].

Epigenetic modifications also play an important role in menstruation. The epigenetic alterations in endometrium during the menstrual cycle are mainly driven by the expression changes of epigenetic modulators [114], and the expression of such genes are affected by ovarian hormones estrogen and progesterone [115]. Histone acetylation regulates vascular endothelial growth factor pathway (VEGF) [116] during angiogenesis and facilitates embryo implantation [117](Fig. 1.3).

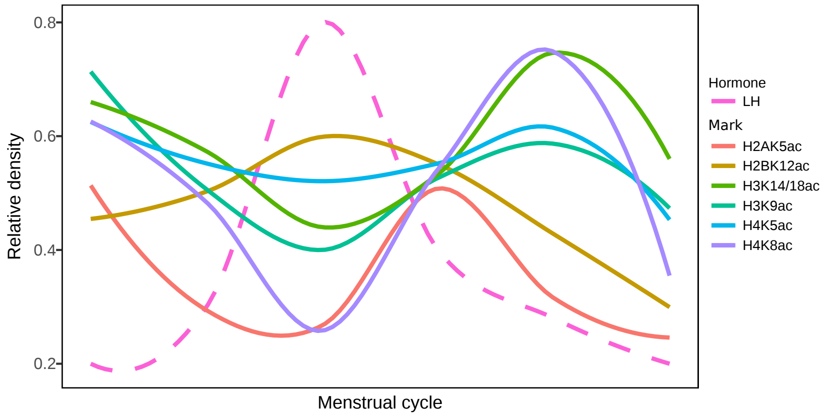


Figure 1.3: Global acetylation levels of endometrial progenitor cells determined by western blotting during menstrual cycle, data taken from [117]

.

Global loss of DNA methylation in non-CpG islands is observed with increasing age [6], while CpG islands tend to become hypermethylated. Hypermethylation is known to affect genes involved in slow wound healing [118] and loss of teeth and hair [119]. Methylome changes are also associated with aging-related diseases, such as Parkinson's disease [120], Alzheimer's disease [121], Huntington's disease [122].

Diseases

Aberrant epigenetic modifications contribute to several effects in carcinogenesis, such as chromosomal instability, reactivation of transposable elements, and loss of imprinting. It is estimated that promoter hypermethylation plays a more dominate role than sequence mutations in gene silencing [123]. Global hypomethylation is frequently observed during cancer progression, which may cause overexpression of oncogenes [124,125]. In particular, hypomethylation at enhancers are much more frequent than at promoters in colorectal cancer [124].

Hypomethylation of transposable elements, such as small interspersed nuclear elements (SINE), and long interspersed nuclear elements (LINE) is frequent in cancer. Given that LINE-1 comprises $\sim$ 17% of the human genome [126], LINE-1 hypomethylation has been reported in a variety of cancers (bladder cancer [127], colorectal cancer [128, 129], gastric cancer [130], breast cancer [131], multiple myeloma [132], hepatocellular carcinoma [133], urothelial carcinoma [134]), and leads to the activation of oncogenes [127,128].

Moreover, dysregulation of transposable elements is responsible for chromosome instability [135, 136], and contributes to chromosomal translocations in acute myeloid leukemia (AML) [137], T-cell acute lymphoblastic leukemia (T-ALL) [138], chronic myelogenous leukemia (CML) [139], breast cancer [140,141], gastric cancer [142] and ovarian cancer [143].

Imprinting disorders are found in $\sim$ 30% of colorectal cancer (CRC) patients [144]. In some cases, the insulin-like growth factor II gene (IGF2) is hypomethylated, causing dual expression in both alleles. Although IGF2 overexpression alone is not sufficient for tumorigenesis, it stimulates the growth of many carcinomas [145], and leads to poor prognosis.

Centromeres are mostly hypermethylated in normal cells, but in cancer they are found to be often demethylated [3], promoting mitotic recombination, and eventually possibly leading to aneuploidy [146,147].

1.1.4 Who are involved in epigenome regulation?

DNA methyltransferases

DNA methyltransferases (DNMTs) are involved in two roles: maintenance of methylation (DNMT1) and de novo methylation of CpG sites (DNMT3a and

DNMT3b) [148, 149]. DNMT overexpression is frequent in a variety of cancers, such as AML [150, 151], CML [150], breast cancer [152, 153], colorectal cancer [154, 155], hepatocellular carcinomas [156, 157], pancreas cancer [158], prostate cancer [159], and esophageal squamous cell carcinoma [160], and causes global changes in the methylome. DNMT inhibitors prevent aberrant methylation, can reverse hypomethylation of oncogenes and hypermethylation of tumor suppressor genes [157, 161], making tumors more sensitive to chemotherapeutic treatment [162]. Cancer can evade immunoresponse by silencing the expression of cancer-testis antigens, such as NY-ESO-1. Treating with DNMT inhibitors can induce the expression of these antigens, allowing them to be recognized by T cells [163].

Histone acetyltransferases

Histone modifications are performed by two types of modifying enzymes known as epigenetic ”writers”, such as kinases, ubiquitin ligases, histone methyltransferase (HMT) and histone acetyltransferase (HAT) and ”erasers”, such as phosphatases, deubiquitinases, histone deacetylase (HDAC) and histone demethylases (KDM) which remove these modifications. Alteration of histone-modifying genes are frequently observed in cancers [164]. For examples, HDACs regulate various cancer hallmarks, including cell differentiation, cell cycle and proliferation, migration and metastasis, angiogenesis and apotosis [165]. HDAC overexpression is frequent in a variety of cancers [166, 167]. Hence, HDACs are considered as potential drug targets for cancer treatment and HDAC inhibitors are currently under clinical trial for a variety of cancers, such as glioblastoma [168–170].

In particular, enhancer of zeste homolog 2 (EZH2) is a histone methyltransferase which adds methyl group to lysine 27 on histone 3. EZH2 overexpression is frequent in breast cancers [171], wherein it adds H3K27me3 to tumor suppressor genes, causes transcriptionally repression. Therefore, EZH2 is a therapeutic target and EZH2 inhibitors are currently under developing for clinical trials [172].

Transcription factors

Transcription factors (TF) recognize specific DNA sequences [173], alone or with other proteins in a complex. They can promote or block the recruitment of RNA polymerase to specific genes [174, 175]. The DNA methylation level of the binding sequence can affect TF binding, and lead to increase or decrease in transcriptional activity. For example, Yin *et al.* found that 34% are enhanced by CpG methylation of their recognition sequences, while the others might be inhibited or not affected [176]. The enhanced ones are mostly belong to transcription factors of the extended homeodomain family.

Histone modifications and chromatin accessibility impact the binding activity as well. Transcription factor binding requires the relaxation of chromatin, as it rarely binds to dense [177] or repressed [178] chromatin. Exceptions to this are so called pioneer factors such as FoxA1, which are capable of binding condensed chromatin, leading to its rearrangement and recruitment of further TFs.

In embryonic stem cells (ESC), gene expression that establishes and maintains ESC state is controlled by a few master transcription factors [179–181]. Whyte *et al.* found that the master transcription factors bind large clusters of enhancers, termed as ”super-enhancers” [182]. The ESC master transcription factors Oct4, Sox2 and Nanog were found to regulate each other through binding to super-enhancers and form the a so-called ”core regulatory circuitry” [183, 184]. As such, super-enhancers are densely occupied by these the master regulators during ESC development [182].

1.2 Bioinformatic approaches to the epigenetic study

1.2.1 ChIP-Seq analysis

Chromatin immunoprecipitation sequencing (ChIP-Seq) is widely used in resolving the genomic positions of histone modifications, transcription factors (TFs) and other non-histone proteins. In ChIP protocols, chromatin is sheared to $\sim 150 - 500$ bp fragments, and a specific antibody is used to bind the protein of interests. After purification, the bound DNA is sequenced for analyzing the binding locations. Control samples (referred as ”input”) which are not subject to immunoprecipitation, are often used in estimating the background noises and correcting the biases from GC content [185, 186] and copy number variation [187], etc. As sequencing costs decrease and more histone antibodies are available, many different histone marks are profiled in a variety of cell types or tissues [188].

Depending on the type of histone modifications (or ”histone marks”), the binding regions either display sharp peaks or broad domains, whereas TFs typically display sharp peaks [189]. The sharp peaks can be identified without a sequenced control [190], whereas the identification of broad domain typically requires control especially in case of low enrichment levels.

Enriched regions can be identified using bioinformatic tools called peak callers. Peak caller usually determines the significance of the signal levels in enriched regions, and also estimates the false discovery rate (FDR) relative to the control signal. MACS [191] and SICER [192] are two widely used peak callers. MACS focuses on the local enrichment, and has initially been developed to detect sharp peaks, while SICER joins spatial clusters of signals into broad domain by setting gap size and window size, making it ideal for calling broad peaks like H3K27me3 and H3K9me3. In its first version, MACS used to truncate broad peaks into many separate small peaks. In the next version MACS2, broad peak calling was added, which merges nearby highly enriched regions into a broad region using a looser cutoff.

While peak callers allow one to discriminate the binding events or the regions enriched for specific histone marks from background noise, in many cases, one also needs to compare ChIP-Seq intensities in order to detect differential binding. Many tools have been developed to perform comparative ChIP-Seq analysis between two conditions, such as ChIPComp [193], ChIPDiff [194], ChIPnorm [195], csaw [189], DBChIP [196], DiffBind [197], MAnorm [198], RSEG [199], each specialized in particular scenarios (a decision tree model for tools selection was proposed by Steinhauser *et al.* [200]) in terms of the availability of replicates or ChIP-Seq inputs. As an application example, using DiffBind in quantitative investigation of estrogen receptor-α (ERα) binding intensities in primary and metastasis breast cancer patients, Ross-Innes *et al.* found that differential ER-bindings were associated with the prognosis of breast cancer [201].

1.2.2 WGB-Seq analysis

Whole-genome bisulfite sequencing (WGB-Seq or WGBS) is widely used in analyzing DNA methylation. In WGB-Seq, the DNA is treated with sodium bisulfite. Following this treatment, unmethylated cytosines are deaminated to uracils and converted to thymidines during sequencing, while methylated cytosines are still read as cytosines. Using the sequencing coverage of methylated and unmethylated reads, methylation levels are measured either as beta-values (1.1) or M-values (1.2), where a constant offset α is added to the denominator in cases where the sequencing coverage (Cov_{meth} and Cov_{unmeth}) is low. The beta-value ranges from 0 to 1 (unmethylated to fully methylated), and is more intuitive, whereas the M-value usually has a much broader range, but is more statistically valid in various tests [202].

$$Beta = \frac{Cov_{meth}}{Cov_{meth} + Cov_{unmeth} + \alpha} \tag{1.1}$$

$$M = log_2(\frac{Cov_{meth} + \alpha}{Cov_{unmeth} + \alpha}) \tag{1.2}$$

The beta-values and M-values are easily inter-convertible with formula (1.3) and (1.4). M-values are widely used in many tools which detect differentially methylated regions (DMR), such as MethylAction [203], RnBeads [204], while other tools may directly infer DMRs using sequencing coverage [205]. I explicitly use M-values in determining the methylation thresholds in section (2.3.2).

$$Beta = \frac{2^M}{2^M + 1} \tag{1.3}$$

$$M = log_2(\frac{Beta}{1 - Beta}) \tag{1.4}$$

1.2.3 Chromatin accessibility

Regions of open chromatin are associated with nucleosome-free regions exhibiting enrichment of DNase I hypersensitive sites (DHS) [206,207]. DNase-Seq [208], assay for transposase accessible chromatin using sequencing (ATAC-Seq) [209] and formaldehyde-assisted isolation of regulatory elements and sequencing (FAIRE-Seq) [210] are often used as low-cost alternatives to factor-specific ChIP-Seq for general purpose chromatin accessibility studies [211].

FAIRE-Seq does not need antibodies as it is based on formaldehyde crosslinking. As nucleosomes depleted chromatin are very inefficiently crosslinked to protein [212], their locations are captured by FAIRE-Seq for sequencing. In DNase-Seq, sequences bound by regulatory proteins are protected from DNase I digestion, and are further sequenced for the identification of open chromatin regions and hence potential binding regions across the genome. DNase-Seq has high sensitivity at promoters [213], yet it has been limited to the requirement for the presence of DHSs [214]. In ATAC-Seq, Tn5 transposase carrying sequencing primers is used to cleave genomic DNA at nucleosome-free regions. Deep sequencing of the purified regions provides open chromatin locations in the genome.

Chromatin accessibility is closely related to enhancer activity, and has been used in enhancer inference in a number of studies [215–217]. Especially ATAC-Seq is increasingly used to map open and potentially active enhancer regions in a cost effective way. Given that ATAC-Seq can be performed on a very low number of cells (even down to 500 cells), it has the potential to detect enhancers active only in rare subpopulations of cells and the resolution to precisely identify the active enhancer region.

1.2.4 Chromatin states

Chromatin state represents a descriptive classification of genomic regions that is based on specific combinations of chromatin-associated proteins and histone modifications. A state describes the probability (between 0 and 1) of the presence of a particular histone mark. For example, a promoter state may have an emission probability of H3K4me3 close to 1 and H3K36me3 close to 0. Given m histone marks, the theoretical number of possible combinations of these marks will be $m!$ (factorial m). The number quickly becomes too large to analyze when $m > 4$, yet some of the patterns are very infrequent or produced from technical artifacts. There are several methods to generalize these patterns. One of which is based on hidden Markov models (HMM) and can be defined as (1.5).

$$\lambda = (A, B, \pi) \tag{1.5}$$

where A is a matrix of state transition probabilities, B is a vector of state emission probabilities and π is a vector of initial state distributions. A and π are initialized from Bernoulli or gaussian random variables, and refined using the Baum-Welch algorithm. Transition probabilities represent the frequencies of co-occurrence of each possible combination of the neighboring states. The model is used to infer the chromatin segmentation from a sequence of hidden states, using the Viterbi or the posterior decoding algorithm.

There are several tools implementing HMMs in chromatin states recognition, such as ChromHMM [218], and EpicSeg [219]. ChromHMM allows one to define chromatin states through the presence of histone marks, and was widely used to annotate the epigenome in the ENCODE and Roadmap projects. The epigenetic signals are assigned to non-overlapping bins of 200 bp (which roughly mimic the nucleosome DNA sizes). ChromHMM has been applied on 111 Roadmap primary cell lines and 16 ENCODE cell lines with six histone marks [220].

The interpretation of chromatin states is done according to prior biological knowledge. For example, the combination of H3K4me1 with H3K27ac is known to mark active enhancers, hence the corresponding state will carry this name. As the exact number of chromatin states is unknown, a ChromHMM model with a large number of states can be pruned so that the state which has the least distance to the nearest is removed. Besides histone marks, ChromHMM was also used in combination with ATAC-Seq and WGB-Seq [221]. ChromHMM has been used in characterizing cancer subtypes [222,223], and associated with other cancer-specific regions, such as hypomethylated regions (HMR) [224], and differential gene expression [223].

1.2.5 High-throughput sequencing based techniques to study 3D chromatin organization

Since chromatin conformation capture (3C) was invented by Dekker *et al.* [225], various chromatin conformation capture techniques have been developed, including circular chromosome conformation capture (4C [226], or "one-vs-all"), chromosome conformation capture carbon copy (5C [227] or "many-vs-many"), chromatin interaction analysis with paired-end tag sequencing (ChIA-PET) [228] and Hi-C [229] ("all-vs-all"), which all allow detection of long-range DNA interactions. In these methods, DNA-protein complexes are crosslinked with formaldehyde. After ligation of the interacting chromatin, the DNA is digested with restriction enzymes. The resulting DNA fragments are then sequenced and mapped to the genome, allowing identification of the genomic locations of the distal interacting chromatins. The Hi-C resolution has been increased from initially 100 kb [230] to 40 kb [231] and, more recently, further down to 1kb [232]. Due to multiple steps in the protocol and the low-yield of ligation products [233], Hi-C requires large amounts of starting material, which makes it not applicable to small amount of cells, for example from cancer biopsies.

In additional to Hi-C, ChIA-PET includes an immunoprecipitation step to enrich for chromatin that is bound by a specific protein, e.g. transcription factors, insulator proteins (CTCF) or the elements of the basal transcription machinery (RNA-PolII, etc.). Capture-C [234] includes an additional pull-down of the biotinylated fragments with magnetic beads, allowing to capture fragments which interact with e.g. promoter regions.

Overall, the "C" techniques can be used to generate contact probability profiles and validate chromatin interactions, which make it essential in predicting promoter-enhancer interactions.

1.2.6 Predicting promoter-enhancer interactions

Predicting chromatin interactions, especially promoter-enhancer (P-E) interactions is an alternative approach when Hi-C datasets are not available or impossible to obtain due to limited amount of samples. The most basic prediction method is to simply select the nearest promoter of the enhancer, but the accuracy is merely about 40% [235, 236]. The accurany can be improved by restricting the predicted interactions within the TAD domains [231, 237], or requiring conserved sequence patterns at the binding sites [238, 239].

Other approaches select several candidate promoters within a certain distance from the enhancer, and predict the interaction probability based on activities from other assays, such as DNase I hypersensitivity [240, 241].

Recently, several supervised methods were developed to identify cell-type specific P-E interactions. IM-PET [236], RIPPLE [242], and JEME [243] were implemented based on a random forest (RF) classifer. They use epigenetic modifications (DNase, H3K27ac, H3K27me3 and H3K4me1), promoter-enhancer distances, binding motifs and enhancer RNAs (eRNAs) as features and trained with known interactions from ChIA-PET data. The authors claimed that they achieved high prediction accuracy (70%-90% of the AUPR) on various cell lines [242, 243].

1.3 Goals and structure of this thesis

In this thesis, I will describe two approaches in using epigenetic datasets in a integrative manner in order to achieve a better understanding of regulatory mechanisms.

In the first part (Chapter 2), I will a describe several bioinformatic approaches to analyze the epigenomic profile in a specific cancer type, namely glioblastoma multiforme (GBM). In particular, I will describe possible artifacts in the identification of genomic regions enriched for specific histone marks, and benchmark two peak calling algorithms. In addition, I will describe the chromatin states obtained in the different subtypes of GBM, and discuss the differences observed. Following the observation of the subtype differences, I will show how subtype classification can be obtained from histone marks, and compare these classifications with alternate ones.

In the second part (Chapter 3), I will present a novel method which (i) integrates multiple epigenetic marks into one single score, and (ii) takes into account the contribution from various regulatory elements, namely promoters and distal enhancers. This method uses a graph theoretical framework, and can be applied to any differential analysis of the epigenetic profiles between two conditions. Using random walk methods on the graph of enhancer-promoter interactions, single genes can be ranked according to the amount of epigenetic alterations in their regulome. I will show a comprehensive benchmarking of this method using datasets from various cancer types and developmental processes. This method can also take into account relations between genes (such as gene-gene interactions) in the form of an embedding network, from which gene modules can be extracted, and compared to specific pathways.

In the last Chapter (Chapter 4), I will discuss the results and provide some outlook.

Chapter 2

Glioblastoma epigenetics

2.1 Introduction

Glioblastoma multiforme (GBM) is a deadly and frequent brain tumor in adults [244]. In the WHO classification, it is classified as grade IV, and the tumor cells are undifferentiated or anaplastic. Grade II and III gliomas are termed lower-grade gliomas (LGG), and can in certain cases evolve to secondary GBMs. The median survival time of patients who were diagnosed as GBM in the United States is less than one year [245]. The cell of origin of GBMs is still controversial, and various studies have studied to what extend central nervous system (CNS) cells can leading to tumor initiation, with mixed results. However, it is believed that neural stem cells (NSC) play a central role in the GBM initiation.

Comparing to the normal brain cells, the tumor cells harbor a significant number of mutations, both in DNA sequences as well as epigenetic modifications [246, 247]. Using gene expression datasets, GBMs were classified into four basic groups according to gene expression patterns, which are proneural, neural, classical and mesenchymal [246], and characterized by either frequent mutations or high expression of signature genes (Table 2.1). The survival time for each subgroups after aggressive treatment is highest in neural, followed by classical, mesenchymal, and proneural which has the lowest survival time [246]. The gene expression classification was found to be problematic afterwards because the proneural group contains two very distinct subgroups (termed "IDH" and "RTK I") which can be distinguished by investigating their DNA methylation pattern.

GBMs with IDH mutations gain the ability to produce 2-hydroxyglutarate (2-HG), which affects the function of enzymes that are dependent on α-ketoglutarate [248], including DNA methyltransferase [249] and histone lysine demethylases [250, 251]. In the latest GBM classifier, the neural subtype is no long present due to high normal brain cells contamination [252]. By subtyping with DNA methylation from 450K methylation array, Noushmehr *et al.* discovered that the IDH1 mutation status defines two distinct DNA methylation patterns in 272 glioblastoma tumors from TCGA, and defined two major subgroups called Glioma CpG island methylator phenotype (G-CIMP) positive (CIMP+) and negative (CIMP-) [253]. Turcan *et al.* further confirmed that mutated IDH1 has the capability to remodel the methylome [254]. Besides IDH1, Sturm *et al.* also proposed another two subgroups within the CIMP- defined by the mutation status of H3F3A gene [247].

Table 2.1: Glioblastoma classification and signature genes *

	Mutation	Overexpression
Classical	EGFR	EGFR
Proneural	TP53, IDH1, PDGFRA	PDGFRA
Mesenchymal	NF1, PTEN, TP53	
Neural		neuronal genes

* Signature genes refer to the mutated and differential expressed genes from Verhaak *et al.*

In this study, I analyzed six histone modification datasets of GBMs from the DKFZ-HIPO project (Table 2.2). The 60 GBMs in this study were classified into IDH (proneural), MES (mesenchymal), RTK I (classical) and RTK II (proneural) using a previously published 450K methylation array classifier [247], and the classification is in coincidence with WGBS and RNA-Seq subtyping. Six histone modifications (H3K4me1, H3K4me3, H3K27ac, H3K36me3, H3K9me3, H3K27me3) were analyzed using ChIP-Seq.

2.2 Regulatory regions in GBMs

2.2.1 Identification of enriched regions of epigenetic marks

ChIP-Seq reads were mapped to the reference genome GRCh37/hg19 using the Bowtie aligner. Uniquely mapped reads and individual input controls were used for the peak calling with MACS2 [191] using default settings. Broad peaks are generated both using SICER [255] and MACS2 with "-broad" option. The broad peaks were 1.5-3.2 fold longer than the narrow

Table 2.2: GBM datasets in this study

Subtype	WGB-Seq	RNA-Seq	H3K4me1	H3K4me3	H3K9me3	H3K27me3	H3K27ac	H3K36me3
IDH	15	15	6	6	6	6	6	5
MES	15	15	4	4	4	4	4	4
RTK I	15	15	5	5	5	5	5	5
RTK II	15	15	5	5	5	5	5	5

peaks, and 23%-42% less in number (except H3K9me3). Visual inspection in IGV [256] showed that MACS2 did not detect many enriched domains for broad histone marks (H3K9me3, H3K27me3), and ended up in underestimating both in number and length of the peaks (Fig. 2.1).

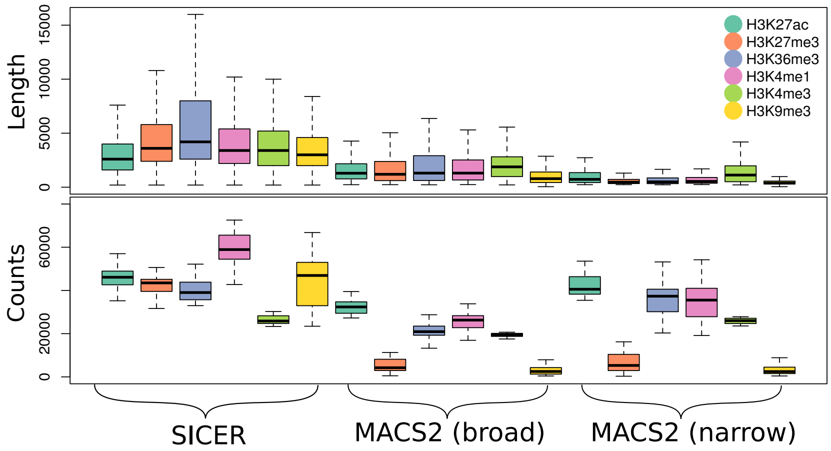


Figure 2.1: Comparison of length and number of peaks across histone marks for two peak callers. SICER outperformed MACS2 in calling broad peaks.

The histone modifications show distinct patterns related to the subtypes. To identify subtype specific enhancers and their associated pathways, I analyzed the peak calls of four histone marks (H3K4me1, H3K4me3, H3K27ac, H3K27me3) in each subtype. The co-enrichment of distinct histone marks at promoters and enhancers are classified into five functional categories, and further filtered according to subtype specificity (Fig. 2.2). In order to identify GBM specific alterations in promoter and enhancer regions, I also compared with peak calls of normal brain samples from the Roadmap project [220]. In total, there are 13 normal brain samples from six different brain tissues (Angular gyrus, Anterior caudate, Cingulate gyrus, Hippocampus middle, Mid frontal lobe, and Substantia nigra), and each has the same set of six histone marks as this study.

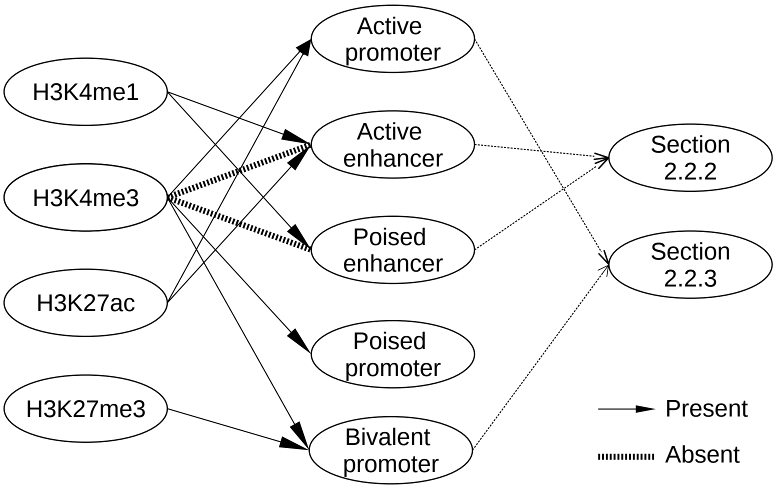


Figure 2.2: Schematic diagram of promoter and enhancer analysis.

2.2.2 Active and poised enhancers

I classified enhancers into two categories as active and poised ones. Active enhancers are marked by both H3K4me1 and H3K27ac [235,257], and poised enhancers do not have the H3K27ac mark [258]. Poised promoters are considered as the outcome of either pre-marking or persisting for extended time after loss of activation [51]. I used the H3K4me1 broad peaks identified from MACS2 to define enhancer regions. Similar to the methods in other studies [259, 260], subtype specific promoters and enhancers were identified by requiring their existence in at least two samples from the same subgroup, and only the regions with no H3K4me3 surrounding ±2 kb of the TSS were considered as enhancers, otherwise they were considered as promoters. Active enhancers and promoters are more frequent than the poised ones in both numbers and genomic length in all four GBM subtypes (Fig. 2.3).

Signals of active enhancers were compared using an analysis of variance (ANOVA) approach across the four subtypes, whereas H3K27ac enrichments of the active enhancers were calculated with (2.1).

$$Score_{H3K27ac} = \log_2\left(\frac{\frac{Count_{H3K27ac}}{Libsize_{H3K27ac}} \times Libsize_{min} + \alpha}{\frac{Count_{Input}}{Libsize_{Input}} \times Libsize_{min} + \alpha}\right) \qquad (2.1)$$

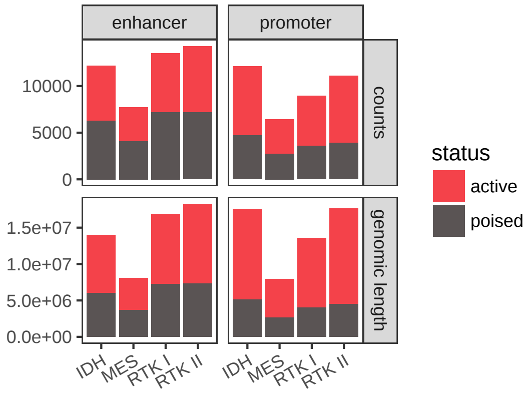


Figure 2.3: Number and length of promoters and enhancers in each category.

where $Count_{H3K27ac}$ and $Count_{Input}$ are the total number of reads mapped to the enhancer in H3K27ac and control datasets, respectively. $Libsize_{H3K27ac}$ and $Libsize_{Input}$ represent the total library sizes for H3K27ac and control, respectively. $Libsize_{min}$ is calculated as $min(Libsize_{H3K27ac}, Libsize_{Input})$, and a constant number α was added to stabilize enrichments when read counts are low. Subtype active enhancers were selected with criteria (FDR <0.1, and log fold change >1), which resulted in 343 IDH, 54 MES, 153 RTK I, 625 RTK II specific active enhancers. I performed functional enrichment analysis of the neighbouring genes of subtype active enhancers using GREAT [261]. The enhancers are assigned to genes based on the "basal plus extension" rule, in which a regulatory domain of ±1Mb from the basal domain (5 kb upstream and 1 kb downstream from the TSS) was searched. GREAT used the entire genome as the background and identified a number of pathways (Fig 2.5) as enriched with active enhancer regulation in each subtype. I selected a few GBM relevant gene ontology (GO) terms as shown in Fig 2.5, wherein the MES subtype has too few active enhancers and show no enrichment in any of the databases.

From the GO enrichment of subtype specific active enhancers, the IDH subtype has more enhancers neighboring with the genes of SREBP signaling pathway, which is associated with poor survival in GBMs [262]. IDH also has a significant amount of G-protein coupled glutamate receptor related enhancers, which have important roles in tumorigenesis [263]. On the other hand, RTK II has several very significantly enriched enhancers associated with well studied pathways in glioblastoma, such as ERBB signaling pathway [264], and fibroblast growth factor receptor signaling pathway [265]. In addition, enhancers regulating neurotrophin signaling pathway are also enriched in RTK II, which are connected with glioma invasion [266].

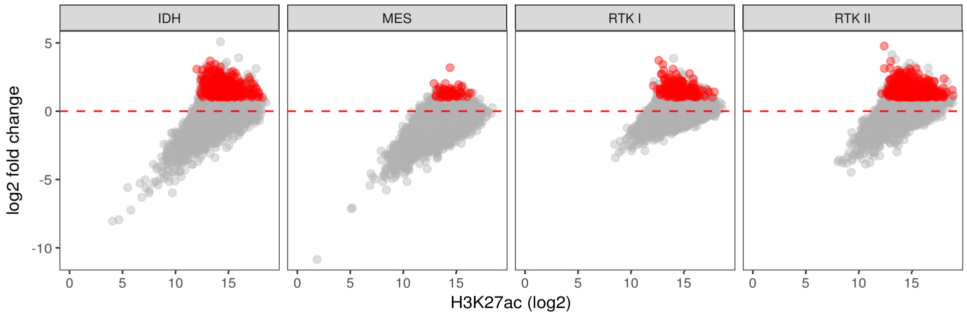


Figure 2.4: H3K27ac signal in each subtype versus log fold changes of H3K27ac by comparing each subtype to the other three subtypes. Subtype specific active enhancers (FDR <0.1, and log fold change >1) are colored in red.

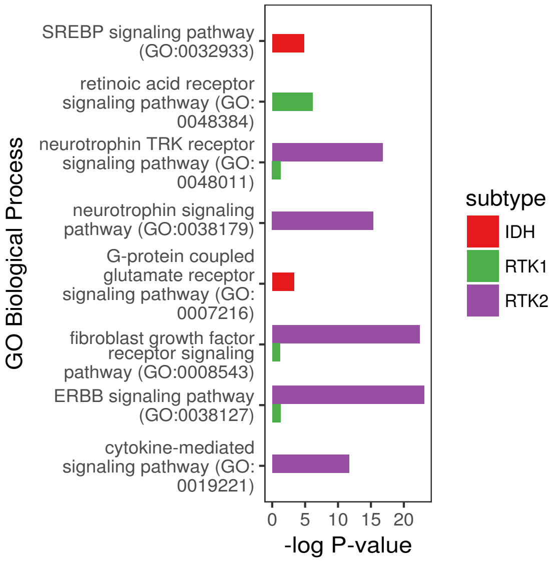
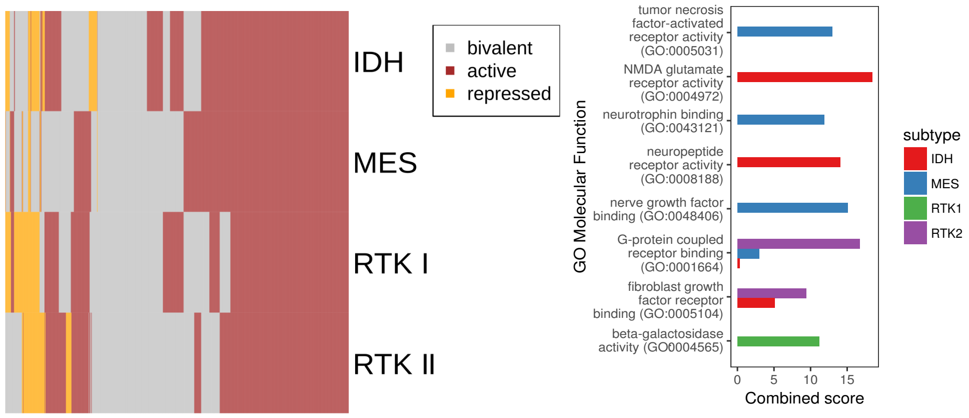


Figure 2.5: GO Biological Process of subtype specific active enhancers. Selected ontologies from top ten enriched terms are presented, and only bars with significant p-values are shown in each subtype.

2.2.3 Loss of promoter bivalency in GBMs

I also examined the genomic regions within ±2 kb of the TSS that are marked with both H3K27me3 and H3K4me3 in normal brain, and defined as so-called ”bivalent promoters”. I analyzed the promoters which lose bivalency in GBMs. Among 2812 bivalent promoters from the normal brain tissues (defined by the fact that at least 5 out of 8 samples have both H3K4me3 and H3K27me3), 51%-57% become active in GBMs (at least three samples in IDH, RTK I, RTK II, or at least two samples in MES lose H3K27me3), while most of the remaining ones stay bivalent, and only a small proportion become repressed by losing H3K4me3 mark (Fig. 2.6 a). GO enrichment of activated promoters in GBMs show enrichment of glioma related terms (Fig 2.6 b), in which glutamate receptor activity in IDH and fibroblast growth factor receptor binding in RTK I are also significantly enriched, suggesting these two categories of genes undergo intensive epigenetic regulation with both promoters and enhancers in GBMs. The MES subtype has a significant number of genes related to neurotrophin binding which become activated, and these genes are known to be involved in epithelial-mesenchymal transition (EMT) [266], which is a typical process in the MES subtype. The common bivalent promoters in both normal brains and GBMs are enriched in developmental genes such as the Hox genes, which are also often observed in ESCs in previous studies [267, 268].



(a) (b)

Figure 2.6: The status and functional enrichment of normal brain bivalent promoters in GBMs. (a) Each cell in the figure represents a bivalent promoter in normal brain, which can be either activated (loss of H3K27me3) or repressed (loss of H3K4me3), or still bivalent in one of the GBM subtypes. (b) GO Molecular Function enrichment of subtype activated bivalent promoters. Terms are measured with combined scores from EnrichR in each subtype.

An investigation on individual promoters reveals that a number of GBM related genes have been activated by losing the H3K27me3 marks. For example, RUNX1 is involved in migration, invasion, and angiogenesis of GBM cells [269]. A general loss of H3K27me3 marks at RUNX1 promoters has been observed in all GBMs as opposed to the normal brain, and led to higher gene expression. The same is true for MYC and MYCN, which are oncogenes highly expressed in GBMs [270, 271]. A few genes are found with exclusive loss of H3K27me3 in RTK I subtype, including ASCL1 [272,273], DLL3 [274], NKX2.2 [275], OLIG2 [276, 277], and SOX11 [278], and lead to specific gene expression in RTK I (unpublished data from other studies). H3K27me3 signals in the house keeping genes are also compared, in which both GBMs and normal brains have low levels of H3K27me3 (Fig. 2.7).

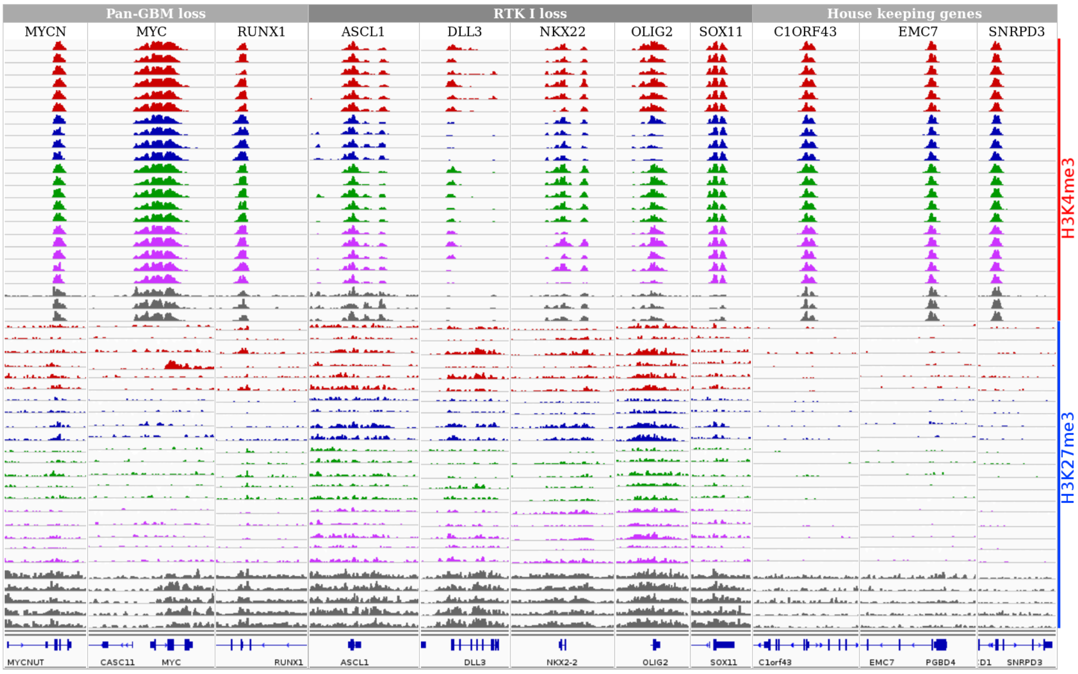


Figure 2.7: IGV screenshots of a few genes losing H3K27me3 marks in promoters, led to higher gene expression in all GBMs or RTK I. Colors of the H3K27me3 and H3K4me3 track represent IDH in red, MES in blue, RTK I in green, and RTK II in purple.

2.2.4 Epigenetic subtyping in GBM

To get a general view of subtype specific histone modification patterns in GBMs, I performed differential binding analysis with DiffBind [197]. GBM subtyping using histone marks is not as common as subtyping using DNA methylation or RNA expression. Not only because the costs of ChIP-Seq are not attractive comparing to methylation array, but also because the normalization of histone marks is more difficult than DNA methylation, since the intensities of histone modification at genome loci can range from 0 to an indefinite value, whereas DNA methylation only ranges from 0 to 1. However, histone mark subtyping for this project is of practical use, since it allows me to validate the accuracy of peak calls by comparing to the known subtype classifications. Here I show that for some of the histone marks subtyping is feasible. With proper peak caller, one can also achieve very good accuracy in GBM classification (Fig. S2).

A number of heatmaps (Fig. S2) using only one histone mark were generated using DiffBind, which presented an initial clustering of the samples from the cross-correlations of merged peak set. Every peak caller and every histone mark individually, is inadequate to produce a satisfying classification of the GBM subtypes (Fig. 2.8). SICER has overall best accuracy for subtyping. It is noticeable that the classification with H3K4me3 peaks from MACS2 and H3K27me3 peaks from SICER both achieved the best performance, taking as a reference of the classification based on DNA methylation patterns. From visual inspection of a large number of loci in the genome browser, I assume that the H3K4me3 peaks from MACS2 are accurate and H3K4me3 is indeed a good classifier which has been also shown in breast cancer subtyping [279]. From visual inspection again, H3K27me3 peaks from SICER are more accurate than those from MACS2. As H3K27me3 modifications are often found mutually exclusive with DNA methylation on a majority of genomic locations [280], classification using H3K27me3 is expected to achieve similar accuracy as the DNA methylation classifier.

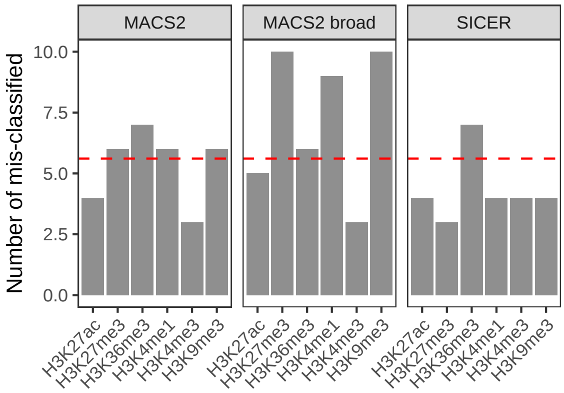


Figure 2.8: Number of mis-classified samples across histone marks and peak callers. The average number of mis-classified samples is 5.6 (indicated with red dashed line), whereas SICER succeeded in lowering the number for every epigenetic mark expect H3K36me3.

Making use of all histone marks seems a more feasible solution [281]. But instead of pooling all peak sets together, I built the concatenation of the feature matrices of six histone marks computed by DiffBind, which produced an augmentation of feature matrix (2.2).

$$Overall_{(k\times(m+\cdots+n))} = \begin{bmatrix} H3K27ac_{(k\times m)} & \cdots & H3K9me3_{(k\times n)} \end{bmatrix} \tag{2.2}$$

where k indicates the number of samples, and $m, \cdots, n$ are the number of merged peaks for each mark. Clustering with the *overall* matrix gives the most similar classification with the subtype definitions (Fig. 2.9).

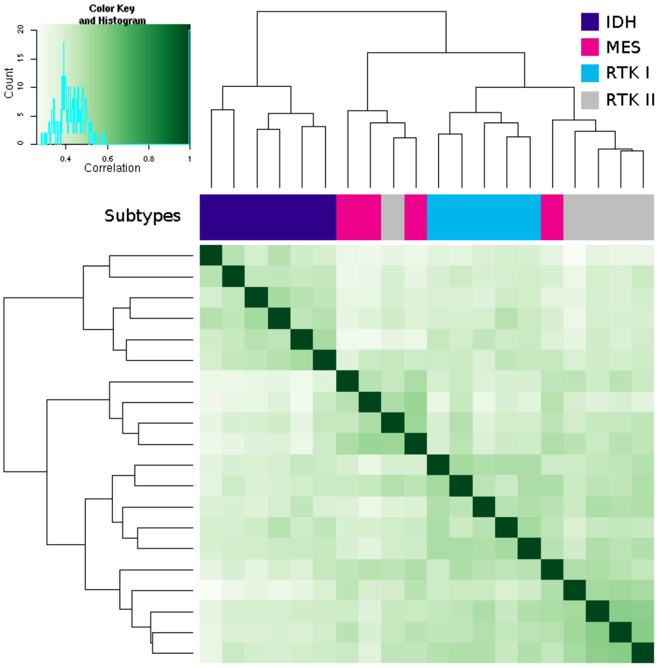


Figure 2.9: Correlation heatmap using all bound sites from all histone marks

2.3 Chromatin states

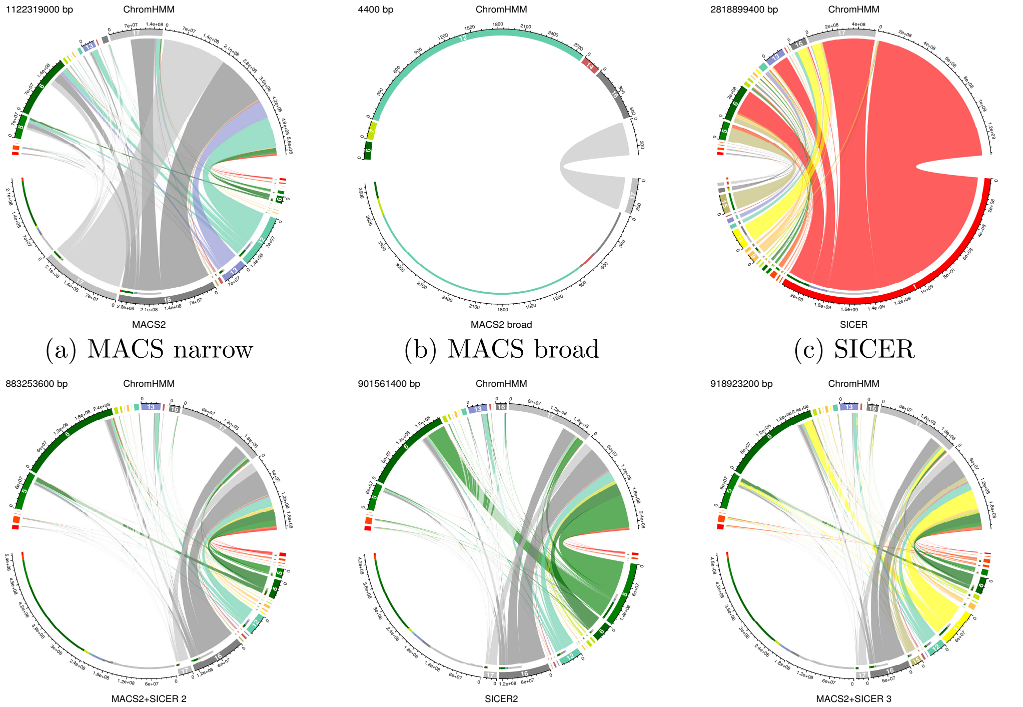
In this part, I performed analysis on the general combination of six histone marks, which is referred to as chromatin states. The analysis focuses on a small number of most prevalent combinations of histone marks, to interpret the validity of the model and transitions between different subtypes.

2.3.1 Roadmap 18-states model

Here I took the emission probabilities and transition probabilities from the 18-states model of the Roadmap project [220], and applied this model to each GBM sample to infer the epigenome states. Two approaches are employed in transforming the count data into the binary values. Either from a Poisson background distribution built in ChromHMM, which is estimated from aligned reads with a sample-specific threshold, or from the presence or absence of peaks from the peak callers. After examining the ChromHMM segmentations from both approaches, I found that the latter one did not work well on certain types of chromatin states. The reason is that the border of either short or long peaks of some histone marks identified by MACS2 or SICER are not exact, making the combination biased towards a few types of chromatin states (Fig. 2.10). From these results, the ChromHMM model from MACS2 narrow peaks (with looser cutoffs, otherwise the H3K9me3 and H3K27me3 peaks were largely absent) (Fig. 2.10 a), MACS2 broad peaks (Fig. 2.10 b), and SICER peaks (Fig. 2.10 c) over-represented ZNF repeats, weak polycomb repressed, and promoters/enhancers, respectively, while ZNF repeats and weak polycomb repressed were largely missing in the ChromHMM model from MACS2 broad peaks and SICER peaks, respectively. Selective combinations of MACS2 broad peaks and SICER peaks still over/under-estimated weak transcription (Fig. 2.10 d,e,f), polycomb repressed (Fig. 2.10 d,e,f), and enhancer (Fig. 2.10 f) states.

In general, the built-in binarization command "binarizeBam" is more accurate in making binarized inputs, owing to that it is capable of setting different cutoffs specific to the Poisson background distribution of different marks, as opposed to using the same cutoff for all datasets. The final model confirmed the findings from section (2.2.3) that many gene promoters indeed lose bivalency (Fig. 2.6 and 2.11).

Since ChromHMM only uses the histone marks, I wanted to analyze the relationship of chromatin states with DNA methylation. Therefore, I took the beta-values of DNA methylation from corresponding samples, and plotted subgroup-wide distributions of DNA methylation in each chromatin states. From (Fig. 2.12), the active promoters show the lowest methylation levels as expected, and the gene bodies, which correspond to the transcription states, are highly methylated. Enhancers have an overall relatively high level of DNA methylation, probably due to a number of enhancer located within gene bodies (10%-20% across samples, primarily intronic). The RTK I subtype has the overall lowest DNA methylation, which can also be observed from the genome-wide patterns of methylation. The polycomb repressed chromatin (PRC) regions show a relatively low methylation as well, which



(a) MACS narrow

(b) MACS broad

(c) SICER

(d) 4 MACS broad (K27ac, K36me3, K4me1, K4me3), 2 SICER (K27me3, K9me3)

(e) 3 MACS narrow (K27ac, K4me1, K4me3), 3 SICER (K27me3, K36me3, K9me3)

(f) 3 MACS broad (K27ac, K36me3, K4me3), 3 SICER (K27me3, K4me1, K9me3)

Figure 2.10: Various combinations of MACS and SICER peak calls in chromatin binarization (bottom half circle), compared with the binarization from the built-in function of ChromHMM (top half circle). The ChromHMM segmentations are made with the 18-states model from the Roadmap project, and the genomic length from all combinations and the built-in binarized input are compared against each other. Chromatin states and their transitions are colored the same as Fig. 2.12.

is also observed in prostate cancer [282], owing to the mutually exclusive occupancy of PRC and DNA methylation at CpG sites. In addition, bivalent promoters in the normal brain samples are very lowly methylated comparing to the tumor samples. This phenomenon is also frequently observed in other cancers, such as colorectal cancer [283, 284], prostate cancer [282, 285], lymphomas [286] and cancer cell lines [287]. As a result, hypermethylation in bivalent promoters makes the expression of these genes even lower.

State transitions between subtypes and normal brain samples show that the promoter and transcription states are relative consistent across samples, and switching usually takes place between different variants of the same cat-

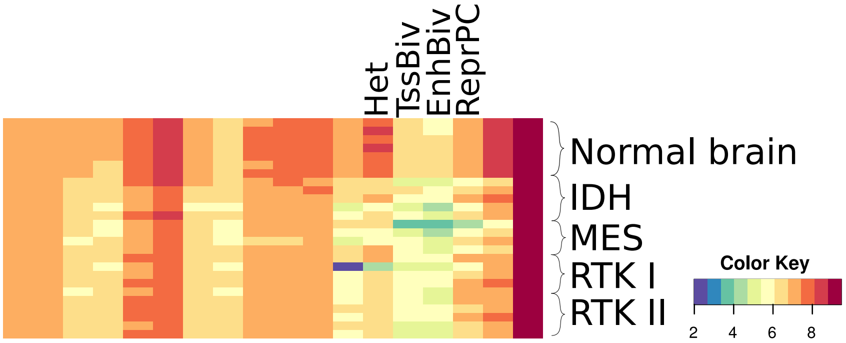


Figure 2.11: Bivalent regions as proportions in GBMs and normal brain. GBMs have an overall trend of losing bivalent and repressed chromatin states.

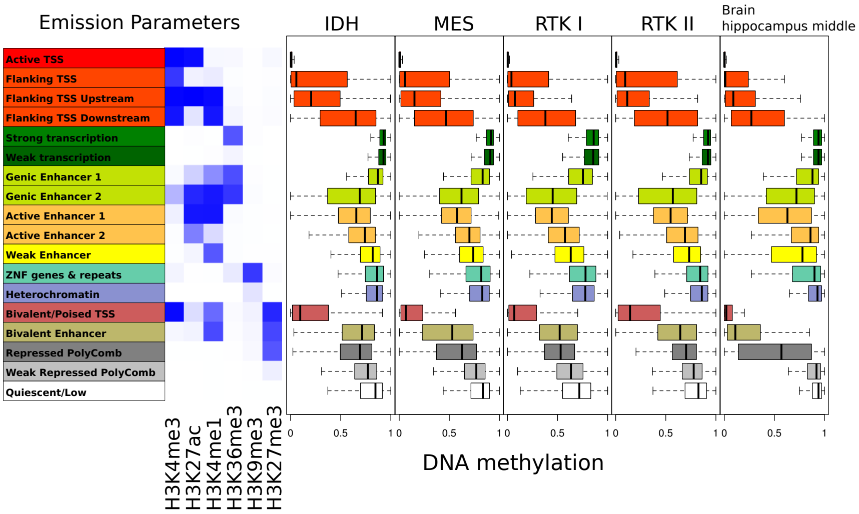
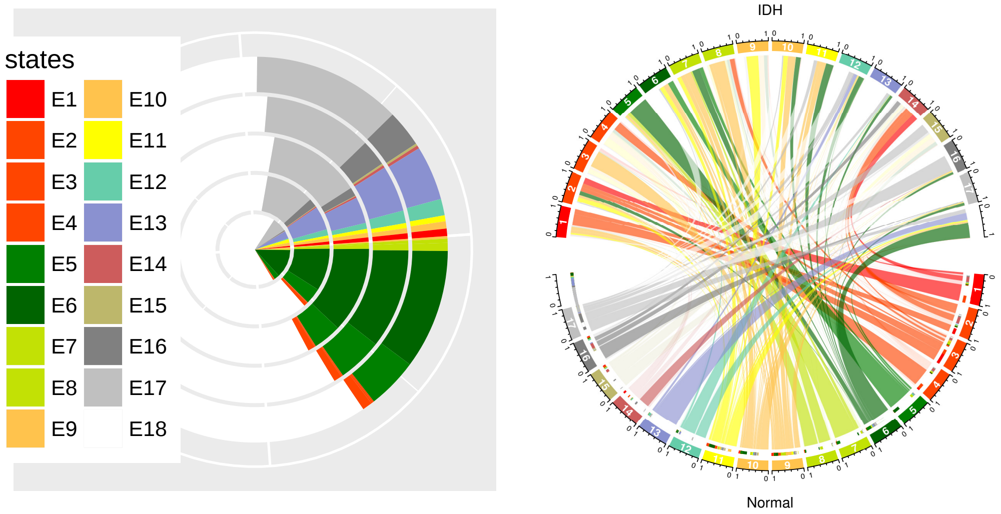


Figure 2.12: DNA methylation distributions of chromatin segmentation of the 18-states model

egories, e.g. from weak transcription to strong transcription, or from active promoter to bivalent promoter (Fig. S3). The other states are more variable, and transitions are more frequent between different categories, e.g. from enhancers to heterochromatin. There is no noticeable excessive transition of any state between subtypes, or between GBMs and normal brains, yet the genomic regions of enhancers still show strong subtype specificity (Fig. 2.13).

2.3.2 Customized ChromHMM model

Apart from segmenting the GBM epigenomes into a fixed combinations of chromatin states, I also learned several novel specific models from the char-



(a) (b)

Figure 2.13: Chromatin states comparison. (a) Genomic length of chromatin states, from the inner circle to outer circle are: normal brain, IDH, MES, RTK I, RTK II. (b) States transition shown in percentage between five IDHs and seven normal brain samples. Promoter and transcription states are more consistent as transitions mainly occur between the active/poised or strong/weak variants.

acteristics of epigenetic marks in each subtype. Learning a new model from binarized GBM inputs gives a different number of states. In these models, GBMs show a lack of bivalent domains, especially bivalent enhancers, which is possibly a sign of dysregulation (Fig. 2.14). This also confirms the observation made previously of a general loss of bivalency in GBMs compared to normal brain tissues.

Besides modeling with the pre-defined ChromHMM model from Roadmap and the same histone marks (Fig. 2.14), I profiled my own model using six histone modifications, and DNA methylation. In order to binarize the CpG sites into two mutual exclusive categories, I took the M-values and fitted a Poisson model. Plot of genome-wide CpG sites for each GBM subtype (Fig. 2.15) shows that the M-values follow a clear Poisson distribution, while the beta-values display a bimodal distribution. Therefore, the M-value representation is more homoscedastic and statistically valid in methylation analysis. M-values of every 200bp bins are calculated from the average methylation level for each sample. By fitting the M-values to sample-wise Poisson models and defining DNA methylation below the respective threshold as unmethylated (Fig. 2.16), I got 34%-38% of the 200bp bins across samples as un-

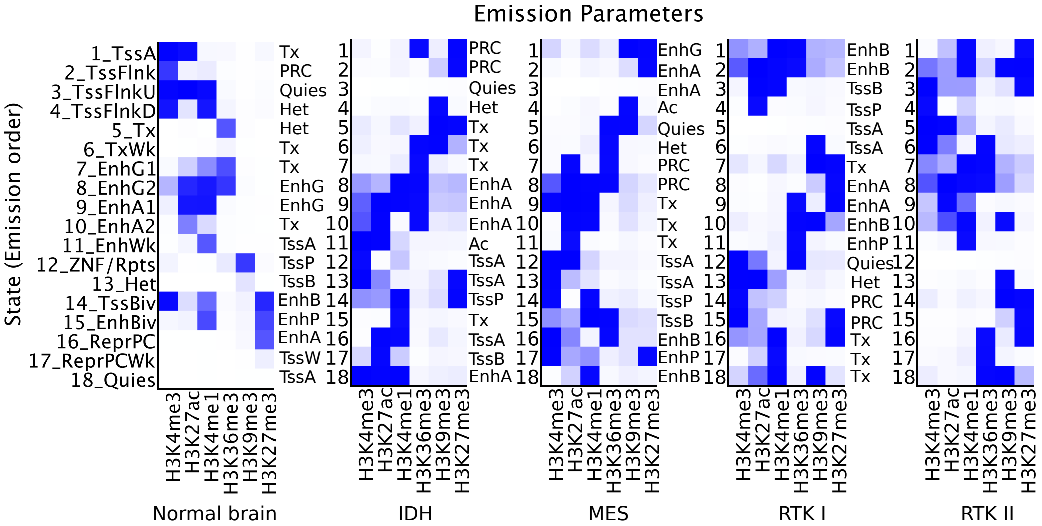
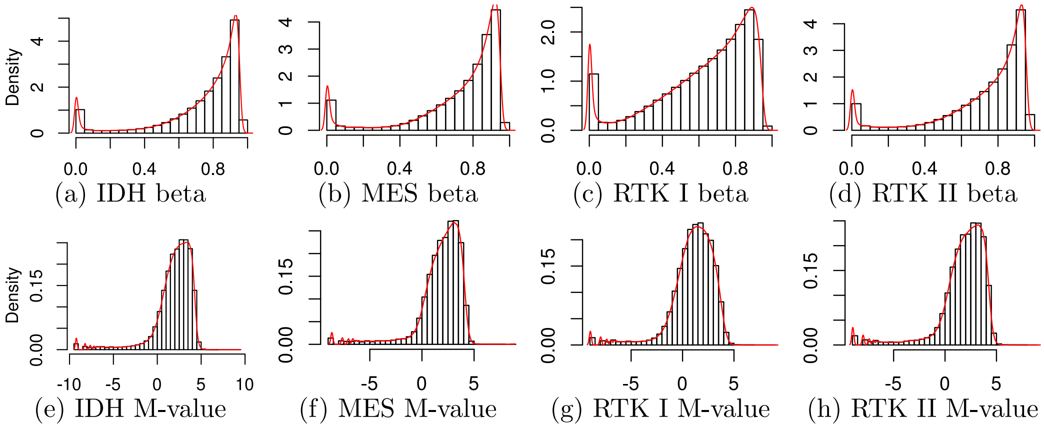


Figure 2.14: Comparing the 18-states model learned from GBM binary inputs.

methylated.



(a) IDH beta (b) MES beta (c) RTK I beta (d) RTK II beta

(e) IDH M-value (f) MES M-value (g) RTK I M-value (h) RTK II M-value

Figure 2.15: Density plot of beta-values and M-values in each GBM subtype.

Since there is no obvious way of deciding how many chromatin states I should use in training the model, I trained a series of HMMs with states ranging from 15 to 30 for 200 iterations with the default initialization method, and estimated the optimum number of states using Bayesian information criterion (BIC), which is defined as (2.3).

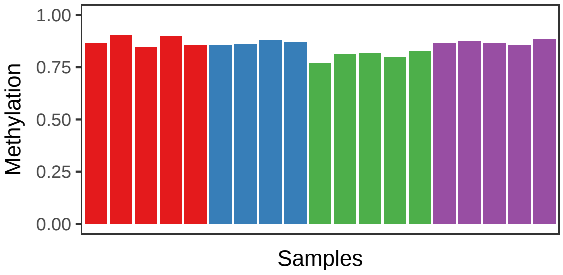


Figure 2.16: DNA methylation thresholds across samples. Red, blue, green, purple represents IDH, MES, RTK I, and RTK II, respectively. Methylation level above thresholds are considered as methylated.

$$BIC = ln(n) \times (m^2 + k \times m - 1) - 2ln(\hat{L}) \tag{2.3}$$

where k indicates the number of parameters of the underlying distribution of the observation process. As the Poisson distribution has only one parameter, hence $k = 1$. m indicates the number of states, and n indicates the total length of all the observation sequences that the HMM was trained with, which is the number of bins used to train the model. $\hat{L}$ is the maximized value of the likelihood function of the model, which was computed by ChromHMM. Generally, the BIC becomes lower as the number of states increases.

I started from the model with 30 states, which has the lowest BIC score among 16 models, and iteratively removed states from this model with the "StatePruning" command in ChromHMM. The emission probabilities of the removed states are redistributed to its transitioned states uniformly. Afterwards, initial parameters were estimated from the resulting set of models for learning another model with reduced number of states. Using the initial parameters from the pruned model, I relearned a model with reduced number of states until it did not contain duplicated states and infrequent states ($\leq$ 0.05% of genome coverage), and still contains the imperative states including active promoters/enhancers, bivalent promoters, transcription, heterochromatin, and polycomb repressed regions. The final model contains 23 states, in which 53% of the genome is in a high methylation states. Promoters in this model fall into two distinct groups, either active and lowly-methylated, or poised and high-methylated. Enhancers can be either highly-methylated or lowly-methylated, regardless of their active status (Fig. 2.17).

In summary, epigenetic profiling using my own models leads to more refined states, especially when DNA methylation is included. Either the technical bias in different histone marks causes heterogeneity in chromatin

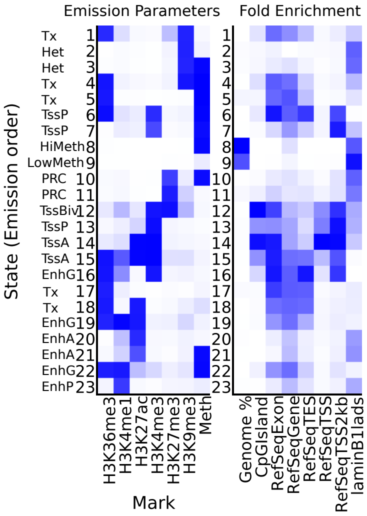


Figure 2.17: A ChromHMM integrating DNA methylation as chromatin states

states modeling, or the GBMs really have a disorganized composition of chromatin states. Either way makes the comparison of chromatin profile across samples more complicated. Therefore, using the same chromatin state model, which is the Roadmap 18-states model, seems to be a good solution in comparative chromatin states analysis.

2.4 Subtype specific differential histone modification patterns

As I learned from the differential expression profiles of GBMs, differential histone modification sites (DHMSs) located in the vicinity of genes contribute to a number of subtype specific gene expressions (Fig. 2.7). Most of the differential binding site detecting tools perform comparison between only two biological conditions [189, 201, 288], and they are not suitable in discriminating subtype specific differential binding for multiple conditions. While one can treat multiple conditions as one condition and compare with the rest, this approach is questionable when there is high heterogeneity in the conditions which are grouped to form the control.

In the GBM study, I assume the presence of 16 differential binding patterns, which are (1) four cases for which bindings are only present in one subtype (Fig. 2.18b 2-5), (2) six cases for which bindings are present in either two subtypes (Fig. 2.18b 6-11), (3) four cases for which bindings are present in either three subtypes (Fig. 2.18b 12-15), (4) two cases for which are not differential binding, either bound on all subtypes, or bound on none of the subtypes (Fig. 2.18b 1,16). ANOVA does not allow one to discover differential patterns specific to more than one subtype. However, such patterns are believed to be present as for example IDH and RTK I belong to a common subgroup (at least from the point of view of gene expression), namely the proneural group.

Hence I developed a new method in calling subtype specific differential binding events, which uses deep neural network to classify epigenome signal into all theoretical binding patterns. The method is efficient in identifying differential binding patterns specific to more than one subtype, and will be useful in more complex binding patterns recognition, e.g. specific binding involving normal brain tissues and low grade glioma in this study.

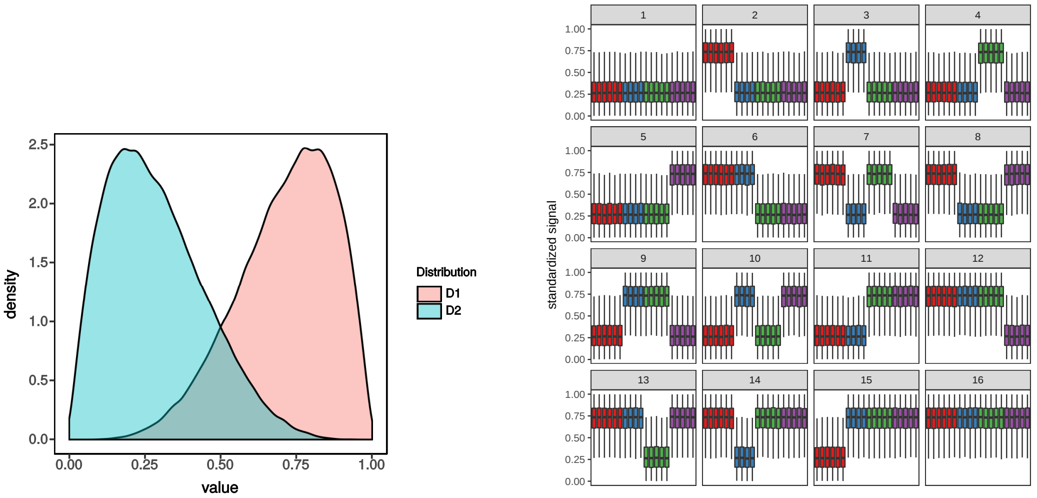
2.4.1 Simulated data

In order to train the neural networks, I built simulated datasets. Synthetic data has been widely used in both physics [289] and biology, e.g., differential binding [189] and expression patterns recognition [290]. Due to the lack of observations of some patterns in real data, synthetic data can be used for training the differential binding model. In this scenario, the "bound" and "unbound" level of H3K27ac signals can be represented from two skewed distributions, with each ranges from 0 (lowest binding signal) to 1 (highest binding signal), and 10,000 cases are sampled from each of the distributions (Fig. 2.18 a).

Since every subtype only has two binding states, I generated 2^4 sets of synthetic data as much as the number of patterns I wish to discover. The datasets are labeled as "$1, 2, 3, \cdots, 14, 15, 16$" for training with a supervised learning approach (Fig. 2.18 b).

2.4.2 Differential modification patterns from neural network classifer

Neural network (NN) is an algorithm which can simulate any mathematical function. Here I use a feedforward neural network to classify the H3K27ac binding into 16 categories. The input to the NN is a 20-dimensional vector (one dimension for each sample) representing each case in the synthetic data. I used a three-layer model (Fig. 2.19). In the fully-connected (dense)



(a) Densities of two sample distributions. D1 and D2 are two beta distributions from different parameters, with D1 representing enrichment in epigenetic signals, and D2 representing depletion in epigenetic signals.

(b) Distributions and labels of synthetic data for all combinations from the four subtypes. Every sample in each subtype is sampled from the same distribution, wherein red, blue, green, purple represent IDH, MES, RTK I, and RTK II, respectively.

Figure 2.18: Synthetic data distributions.

layer, each layer's input can be represented using (2.4) from the inputs of it's previous layer.

$$a_j^i = \sigma(\sum_k (w_{jk}^i \cdot a_k^{i-1}) + b_j^i) \tag{2.4}$$

where σ is the activation function, and in the i^{th} layer, w_{jk}^i denotes the weight from the k^{th} neuron in the previous layer to the j^{th} neuron, b_j^i represents the bias of the j^{th} neuron, and a_j^i is the activation value of the j^{th} neuron.

In the dense layer, the network applies a rectified linear activation function (ReLU) (2.5). I also added a dropout layer between every two dense layers to prevent overfitting. The final layer consists of 16 neurons, which equals the number of patterns I want to predict.

$$\sigma(x) = \{ \begin{array}{ccc} 0 & \text{if} & i < 0 \\ x & \text{if} & i \geq 0 \end{array} \tag{2.5}$$

To train the model, I chose a function called cross-entropy cost function to adjust weights and biases after every iteration (2.6).

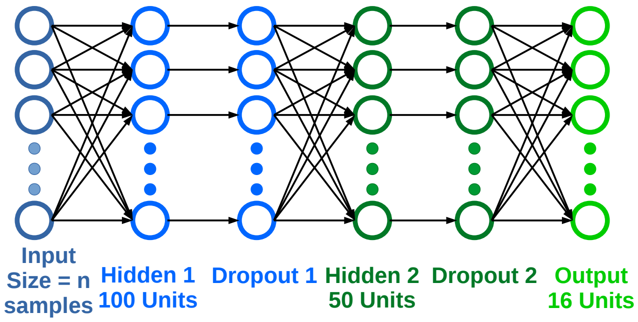


Figure 2.19: Model architecture of deep neural network

$$C_{CE}(W, B, S^r, E^r) = -\sum_j [E^r_j \ln a^L_j + (1 - E^r_j) \ln (1 - a^L_j)] \tag{2.6}$$

where W is the neural network's weights, B is the neural network's biases, S^r is the input of a single training sample, and E^r is the desired output of that training sample. To minimize the loss of cost function, I chose one stochastic gradient descent function frequently used called "RMSProp" (Root Mean Square Propagation) (2.7).

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{E[g^2]_t + \epsilon}} g_t \tag{2.7}$$

where η is the learning rate, $E[g^2]$ is the RMSprop running average of the past squared gradients, ϵ is a fudge factor for preventing divide-by-zero, and g_t is the gradient.

After 100 epoch of training, the network reached 95% accuracy on the generated labels. Comparing the network approach to classical k-means [291], it partitions the data into much more unbalanced groups (Fig. 2.20). Looking at the clusters produced from k-means with 16 centers, all the cluster centers are well spread, and the size of groups are balanced (Fig. 2.21 a,c). However, the patterns from k-means does not really reflect the differences in subtype specificity, as the clusters mainly differ in overall binding levels but not necessarily in binding shapes (Fig. 2.21 a,c). On the other hand, although some patterns in the NN do not have the same level of signals (typically the all "on" and all "off" binding patterns, Fig 2.21 b,d), the patterns are much more subtype specific. Most importantly, some patterns discovered by NN are not present in the k-means clusters, and these patterns are still of biological importance as will be discussed later.

In the subtype specific active promoters, I identified several genes which are in concordance with biological processes of GBM subtypes (Fig. 2.22).

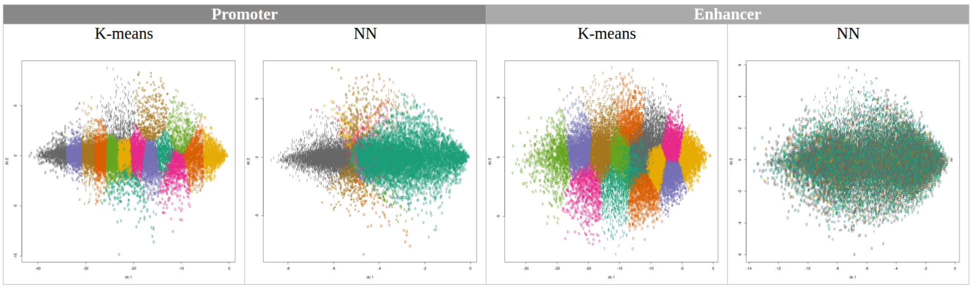
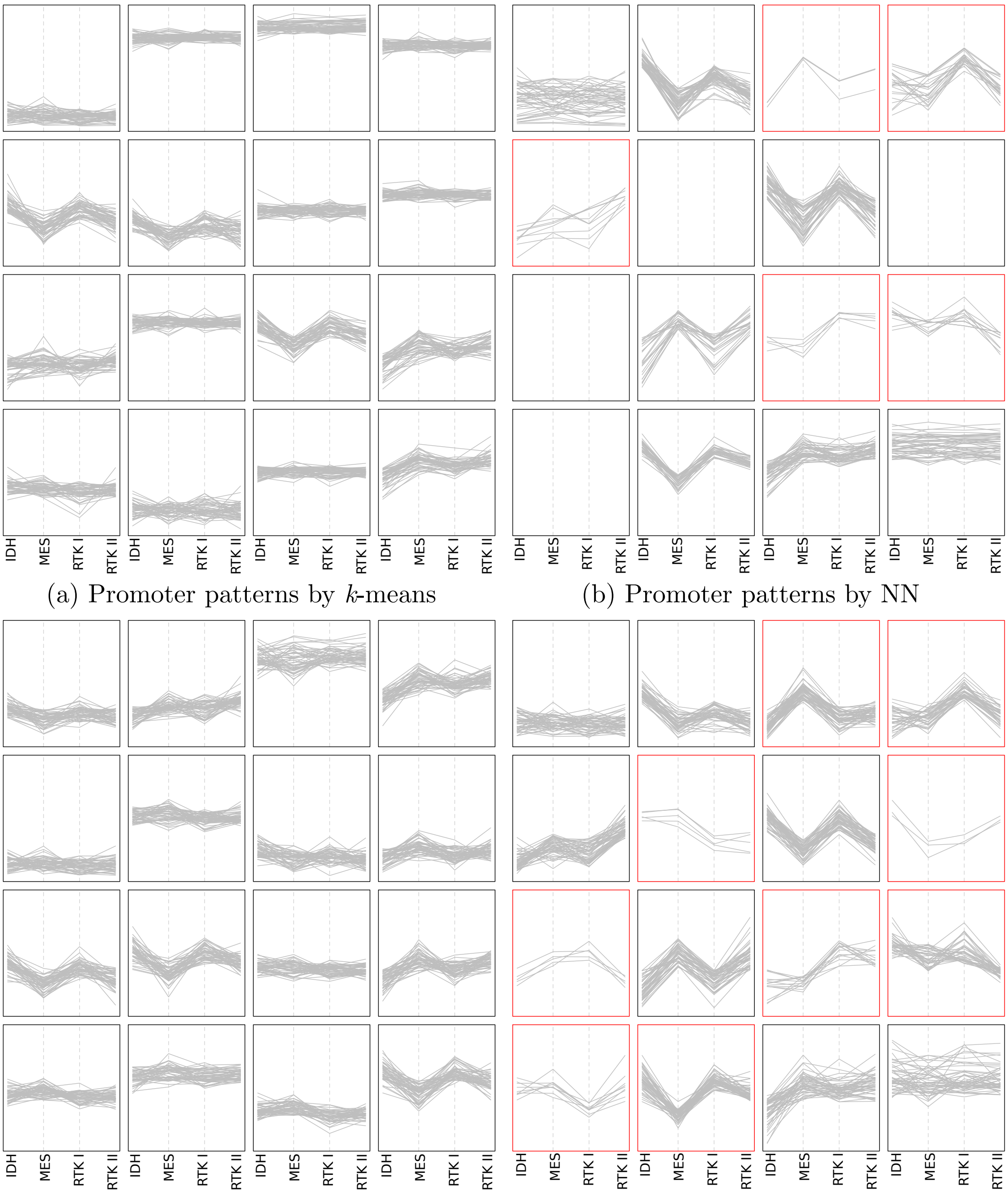


Figure 2.20: Differential binding sites projected to discriminant coordinates. Each predicted category is filled with a distinct color. The sixteen patterns are not linearly separatable in these views.

For example, mesenchymal specific active gene CXCL14 [292] and SGMS2 [293] have been reported to play an important role in epithelial-to-mesenchymal transition (EMT), which is associated the tumor differentiation in the MES subtype [294]. BCAT1, which is only expressed in IDH wild-type tumors, is exclusively inactivated in IDH subtype, which is reported for aberrant BCAT1 promoter methylation as well as promoter deacetylation [295]. KCND2, a potassium voltage-gated channel gene [296] in glioblastoma patients correlated with poor survival, is highly expressed in RTK I [297]. Inactivation of OLIG2 in glioma stem cells (GSCs) results in mesenchymal phenotypes [297], and OLIG2 is lowly expressed in MES in accordance of its lack of H3K27ac. Furthermore, loss of OLIG2 function results in mesenchymal transformation in proneural GBM subtype [298]. The ASCL1 is associated with high H3K27ac levels in all GBM subtypes except MES, and is also reported lowly expressed in mesenchymal and normal brain [273].

In summary, the deep neural network classifier presented accurate pattern recognition in the simulated data with conceived patterns, while these patterns have been proven useful in associating genes with the subtype specific pathways. As any pattern can be generalized in synthetic dataset for supervised learning, the deep learning approach might be proven useful in many other applications in the GBM study, such as recognition of specific epigenetic patterns in transcription factor binding sites, or co-occurrence patterns of multiple epigenetic marks.



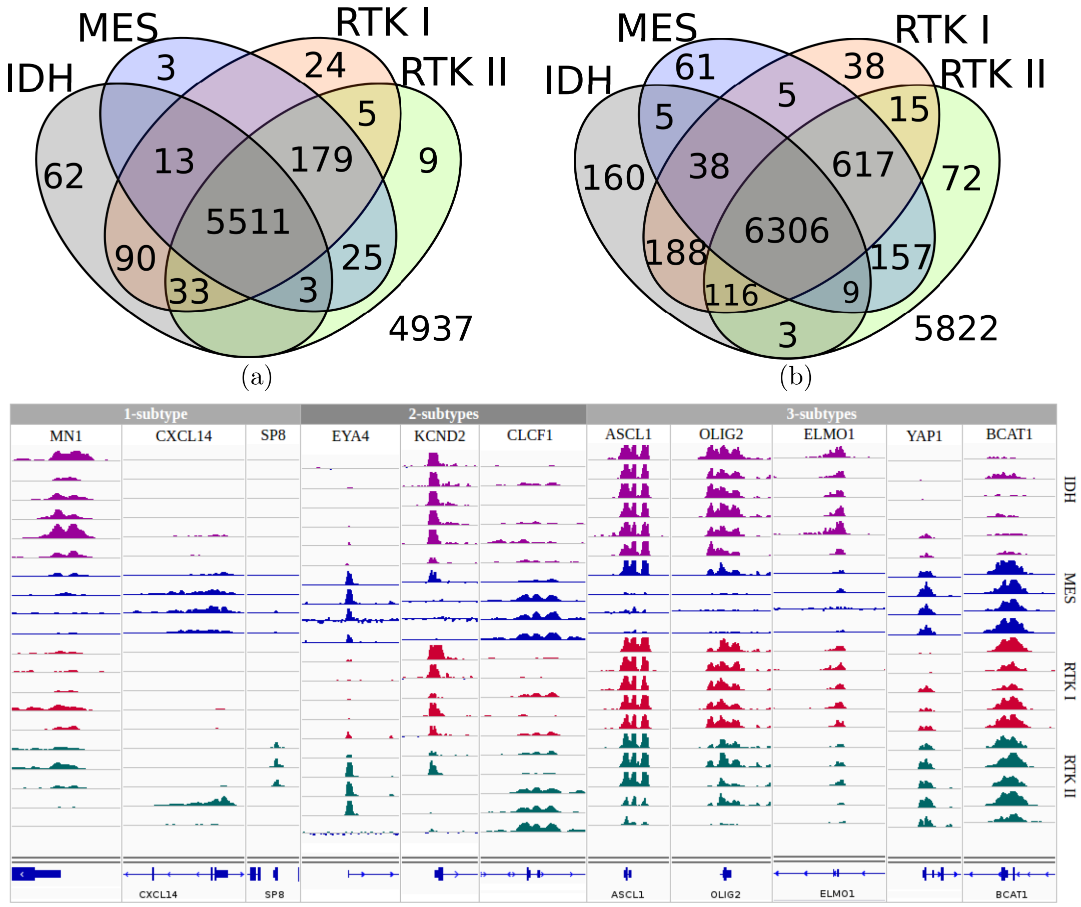
(a) Promoter patterns by *k*-means

(b) Promoter patterns by NN

(c) Enhancer patterns by *k*-means

(d) Enhancer patterns by NN

Figure 2.21: Differential binding patterns discovered from two approaches. Although *k*-means produced more balanced clusters, the patterns of interest were not all present (in red frames). On the contrary, the clusters in NN are unbalanced, but have clearer subtype specific patterns.



(a)

(b)

(c)

Figure 2.22: Differential H3K27ac modification specific to combinations of subtypes. (a) Number of differential H3K27ac promoters specific to the combination of subtypes. (b) Number of differential H3K27ac enhancers specific to the combination of subtypes. (c) A few examples with differential H3K27ac levels from IGV screenshots.

Chapter 3

Integrative analysis of differential epigenetic alterations

3.1 Introduction

Integrating multi-omics data is challenging due to high variability and noises across different data types, yet it is essential in cancer research since cancer usually harbors all type of alterations, either in DNA sequences or epigenetic modifications. Although tools such as ChromHMM have proven their usefulness in studying the combinatorial patterns of multiple epigenetic marks, it is limited to binary measurement such as presence or absence of the peaks. In case where enrichment peaks are present in all biological conditions, the intensity of epigenetic modification level may dramatically affect the binding probabilities. Therefore a quantitative comparison between epigenetic data is necessary in understanding the impact of a binding event. However there are very limited ways in doing this. Although meta-analysis [299] and time course analysis [300] approaches are common in RNA-Seq studies, they a generally not applicable in studying multiple epigenetic marks. In this chapter, I present my approach in integrating multiple epigenetic datasets, which leads to more insightful results in understanding the underlying biological relationship of epigenetic alterations.

In this approach, I will deliberately treat gene expression independently of epigenetic modifications. One of the justifications for this choice comes from the concept of epigenetic priming, an event in which epigenetic modifications initiate before gene expression. Hence, epigenetic alterations and gene expression appear to be decoupled from this point of view. The notion of epigenetic priming is important in cancer studies because it might allow one to identify potential oncogenic processes through epigenetic alterations before they become detectable from gene expression or at protein level. To investigate the concept of epigenetic priming, I considered datasets containing different time points during developmental progression. For example, I studied four stages in neural progenitor cells (NPC) development into neuroepithelial (NE, day 12), early radial glial (ERG, day 12), mid radial glial (MRG, day 35) and late radial glial (LRG, day 80), in which Ziller *et al.* observed that gain of H3K4me1 and loss of DNA methylation appeared in the early stages of the differentiation from ESC to NPC [301]. By comparing the epigenetic patterns in promoter regions of NE stage to gene expression patterns of all five stages, the result shows that the correlations of epigenetic marks with expression at later stages, as opposed to the NE stage, reaches highest level (Fig. 3.1), indicating the expression level is more related to the epigenetic modifications in earlier stages. Comparing to histone modifications, DNA methylation have longer term effect, which confer later expression in development [302]. Epigenetic priming is the reason why my approach focuses mostly on epigenetic alterations rather that gene expression. The delay in gene expression in this case is understandable since time is needed for mRNA accumulation after the epigenetic regulations take place.

To address such concepts in differential epigenetic analyses, I developed a new method termed ”cancer regulatory landscapes” (*crl*) which integrates the quantitative information from multiple epigenetic marks, on genome-wide non-coding regulatory elements, allowing one to discover significantly epigenetically altered genomic regions and pathways. In the benchmarking, I proved that the genes of interests found using this method are highly relevant to the cancer and developmental test cases.

3.2 Data sources

3.2.1 Epigenetic datasets

As illustrated in Fig. 3.1, both histone modifications and DNA methylation can exhibit epigenetic priming. Besides epigenetic modifications in promoters, I also included genome-wide epigenetic alterations in non-coding regulatory elements, which covers both alternative promoters and enhancers. With WGBS data I am able to evaluate the effects of lowly methylated regions (LMRs) inside enhancers, which have been shown to contribute to its activity [303]. Genome-wide epigenome cohorts are publicly available from many consortia, such as NIH Roadmap Epigenomics [304], ENCODE

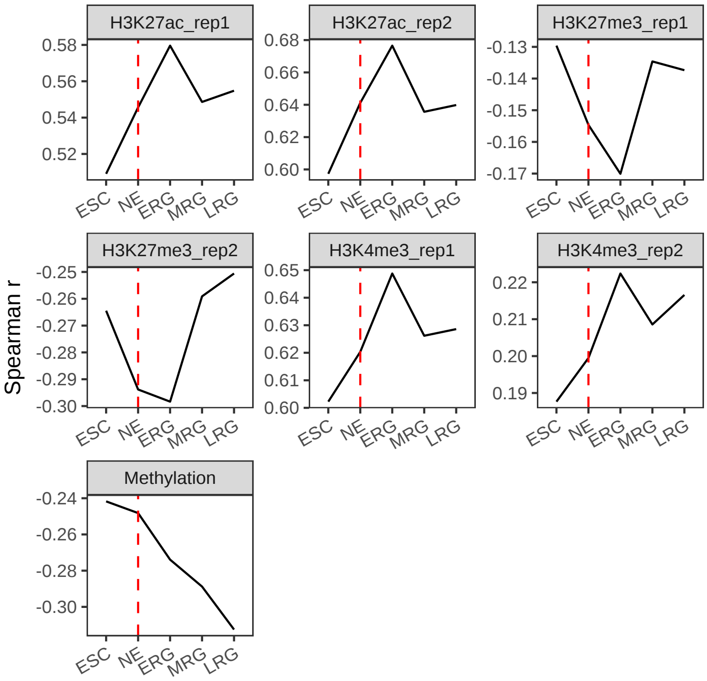


Figure 3.1: Epigenetic priming in neural progenitor development stages. Each histone mark has two replicates, and Spearman correlation between epigenetic patterns of the NE stage and gene expression of all stages are shown. Histone modifications present short-term effects, while DNA methylation present long-term effects.

[188], Blueprint [305], and the International Human Epigenome Consortium (IHEC) [306]. These resources allow me to investigate the epigenetic relationships between embryonic stem cells and differentiated cells, or between tumor and normal tissues (Table 3.1).

Due to the public policies of some data providers, their data are generally provided for visualization purpose. Mostly only Wig and BigWig format files [307] are accessible, rather than raw sequences (Fastq) or alignment files (BAM/BED). Therefore statistical methods specific to raw counts [308, 309] are not applicable in my study. The peak calls were done by these data providers, and these peak regions are used in the validation of cell-type specific enhancers in later sections.

For restricting the epigenetic comparisons to the genomic loci of interests (promoters and enhancers), I downloaded genomic coordinates of promoters from the eukaryotic promoter database (EPD) [310], and enhancers from the GeneHancer database [311]. There are $\sim$ 285,000 enhancers in GeneHancer

Table 3.1: Test cases for phenotypic studies.

Tests	Controls	Num.*	Data**	Accession codes/src
Neural Progenitor Cells (NPC)	Embryonic stem cells	8	61	GSE16256
Neuroepithelial (NE)	Embryonic stem cells	5	10	GSE62193
Early radial glial (ERG)	Embryonic stem cells	5	10	GSE62193
Mid radial glial (MRG)	Embryonic stem cells	5	10	GSE62193
Mesenchymal stem cells (MSC)	Embryonic stem cells	8	51	GSE16256
Trophoblast stem cells (TSC)	Embryonic stem cells	8	64	GSE16256
Chronic lymphocytic leukemia (CLL)	B cells from healthy cases	4	121	CEEHRC
Lower grade glioma (LGG)	Hippocampus middle, Inferior temporal lobe, Mid frontal lobe	7	112	CEEHRC, GSE17312
Colorectal cancer (CRC)	Sigmoid colon from healthy cases	7	154	CEEHRC
Papillary thyroid cancer (PTC)	Thyroid from healthy cases	7	54	CEEHRC

* Number of epigenetic marks.

** Number of epigenetic datasets.

database, incorporated from four different sources: the Encyclopedia of DNA Elements (ENCODE) [312], the Ensembl regulatory build [313], the VISTA Enhancer Browser [314], and the functional annotation of the mammalian genome (FANTOM) project [235]. This database contains enhancers for a large number of cell-types and cell-lines, and tissues. Both datasets were converted from GRCh38 to GRCh37 using the LiftOver tool [315].

The promoter coordinates were extended to ±1000 base pairs around the original coordinates. A BigWig file consists of a number of blocks, each containing a declaration of a fixed or variable genomic region. The numerical signals from BigWigs for region $i \in [m, n]$ were calculated as $S_i = \sum_m^n s_i$, where the promoter/enhancer ranges from m to n, and s_i is the signal for each genomic window in the region. The enhancers from GeneHancer database are collected from all tissues. To avoid unspecific enhancers, I require enhancers to overlap with H3K4me1 peaks of at least two samples in the tissue I am studying. Fig. 3.2 shows the number of common and specific enhancers among four cancer types.

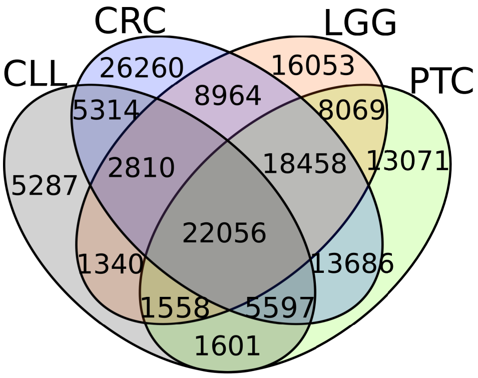


Figure 3.2: Number of common and specific cancer enhancers among four cancer test cases.

3.2.2 Data processing procedure

The epigenetic data between samples have large heterogeneity and must be normalized before statistical comparisons can be made. The promoters and enhancers are of different genomic length, so I divided the intensities by their length before the normalization. The data are heavily right skewed, and I used Box-Cox transformation (3.1) to transform the data into normal distributions.

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log y, & \text{if } \lambda = 0 \end{cases} \tag{3.1}$$

By finding the most likely λ that minimizes the variation, a universal λ of ~ 0.182 was proposed to apply to all data. After power-transformation, I checked the global variability across and within biological groups using *quantro* [316]. It estimated the variabilities were caused by technical variation (e.g. batch effects), and a global normalization is applicable.

I used quantile normalization which was firstly implemented in microarray analysis by Bolstad *et al.* [317], and quickly adopted to a variety of data types such as RNA-Seq analysis [255, 318–320], DNA methylation [321], ChIP-Seq [322, 323]. This method makes the quantiles of each distribution equal. After normalization, except a number of genomic regions showing no signal of epigenetic modifications, the rest of the data follows normal distribution (Fig. 3.3), which is the prerequisite for using differential principal component analysis in the next step.

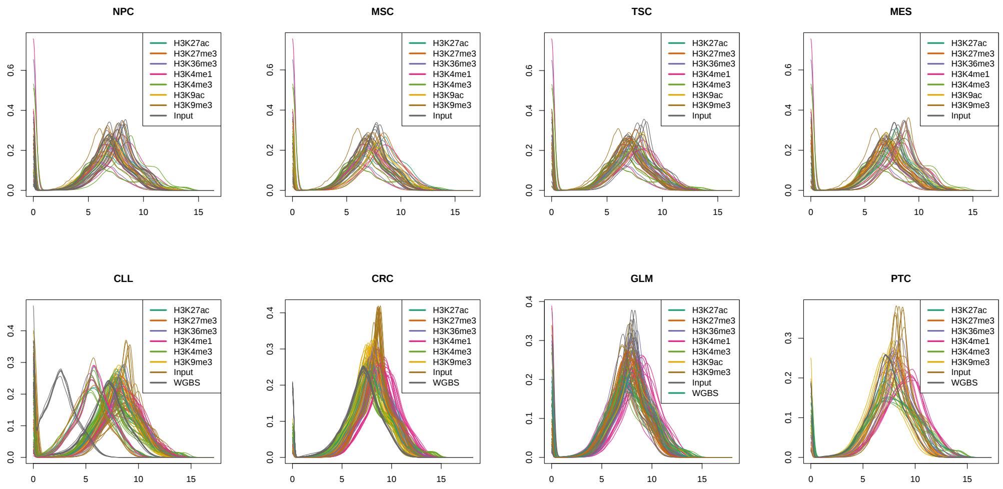
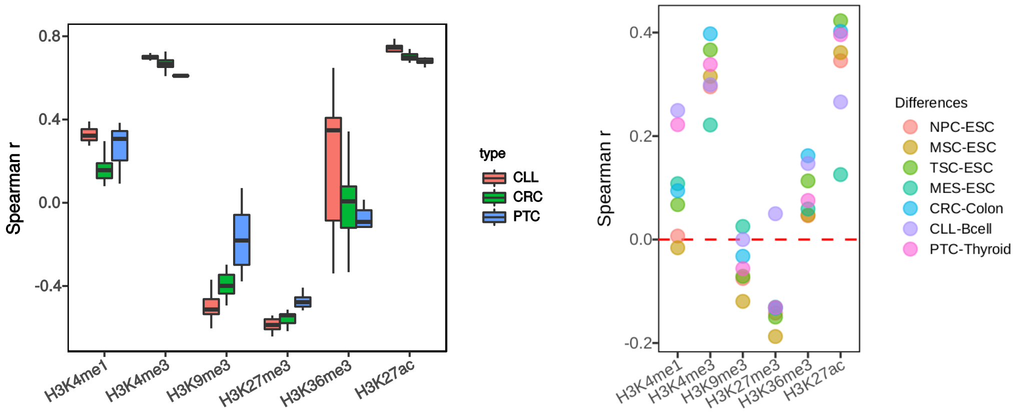


Figure 3.3: Data distributions of epigenetic marks after Box-Cox transformation.

Correlation structures of data

I firstly compared the aforementioned processed data with gene expression to verify that the normalization maintains the correlation structure. The result shows that for each sample, the gene expression positively correlates with the epigenetic signal in the promoter regions of H3K4me3 and H3K27ac, and negatively correlates with the epigenetic signal of H3K9me3 and H3K27me3 (Fig. 3.4. a).

Although some epigenetic marks correlate very well with the gene expression of the corresponding dataset, I wanted to perform a differential analysis between conditions. Therefore, I tried to resolve the relationships of expression differences and the differences of epigenetic modifications between two biological conditions, e.g. tumor cells and normal cells. Again, I normalized the data using quantile normalization, making the two groups have same standard deviations. Afterwards I took the average differences of epigenetic modification levels and gene expression levels. The correlations of differential histone modification with expression in the promoter regions are considerably lower comparing to the previous test, but still correlate positively with activation marks and negatively with repression marks (Fig. 3.4. b). The weaker correlations may be attributed to the fact that I have not included the epigenetic modification outside of the promoter.



(a) Correlation of histone mark with RNA expression

(b) Correlation of histone mark differences between two groups with gene expression differences

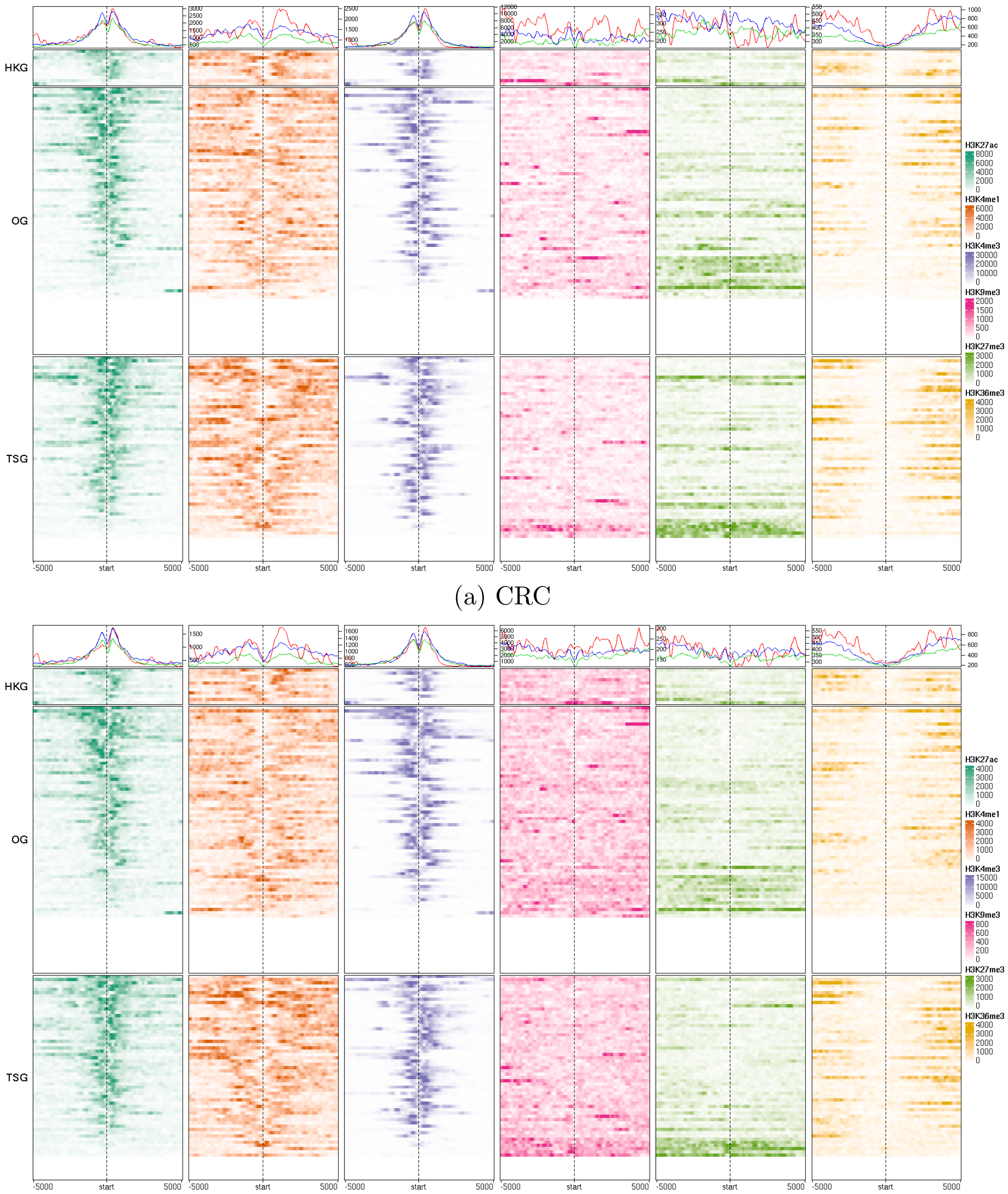
Figure 3.4: Correlation of epigenetic marks with gene expression

Modifications at oncogenes and tumor suppressor genes

With respect to the role in cancer development, it has been known that hyperacetylation of oncogenes (OG) results in an increase of their gene expression, whereas hypoacetylation of tumor suppressor genes (TSG) reduce their expression levels [324]. OG and TSG are both retrieved from a compiled list by Walker *et al.* [325] (Supplementary table S1). Here I used 11 housekeeping genes (HKG) with constant expression level from RNA-Seq profiles as control [326]. The epigenetic modification levels around the TSS of HKG, OG and TSG are not significantly altered comparing to each other (Fig. 3.5 and Fig. S1), suggesting that differential epigenetic modifications mainly occur at distal regulatory regions.

3.2.3 Multivariate data analysis

In order to represent the overall differences from multiple epigenetic modifications, a single measure is needed to represent the variances between different datasets. After subtracting each epigenetic mark with the control, the Pearson's correlations between the average values of each epigenetic mark at promoters and enhancers across the samples indicate that there are strong positive correlations between the repressive marks (H3K27me and H3K9me3) and



(a) CRC

(b) Colon

Figure 3.5: Histone mark signals around OG, TSG, HKG in CLL and normal B cells, for the other test cases, see supplementary fig. S1

active marks (H3K4me1, H3K4me3 and H3K27ac), as well as negative correlation between active marks and repressive marks (Fig. 3.6). Co-occurrence of epigenetic marks is common for multiple activation marks [33]. In addition, DNA methylation and repressive marks are also negatively correlated, as they are often replaced with each other during gene silencing [327, 328].

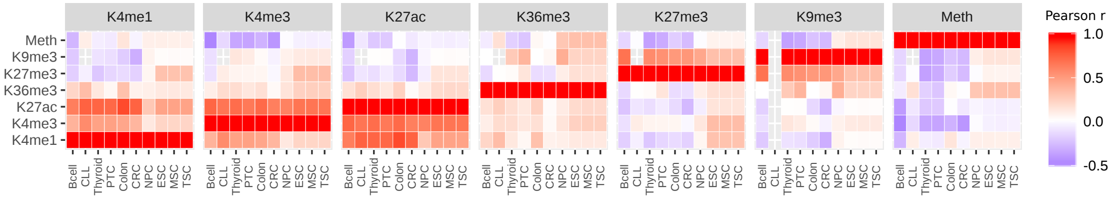
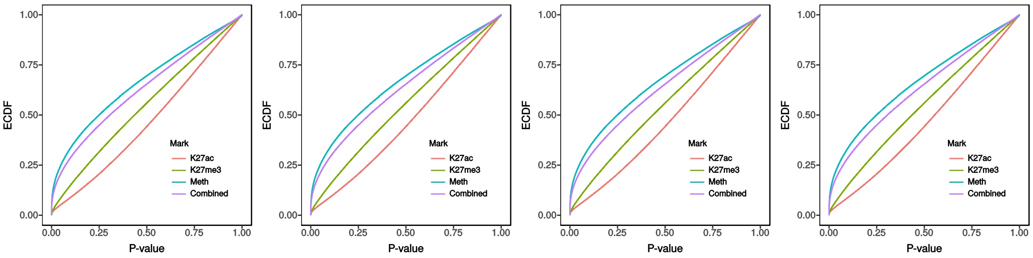


Figure 3.6: Epigenetic marks in promoter and enhancer regions show strong correlations with each other

To statistically test the differences between two biological conditions with each epigenetic mark, I used the epigenetic datasets of embryonic stem cells (ESC) and compared with their differentiated forms (NPC, MSC, TSC, MES). The datasets are available in BAM formats. The p-values of the differential epigenetic signals in both promoter and enhancer regions were computed with ChIPComp for ChIP-Seq, and BiSeq for WGBS. The combined p-values using Fisher's method (3.2) are very close to the smallest p-value in the every comparison (Fig. 3.7). In this situation, combining p-values from multiple hypothesis testing is not applicable as it may lead to severe inflation of false positive rates when applied to highly correlated datasets [329]. Indeed, Fisher's method makes the assumption of independence of the tests, which is not fulfilled here.

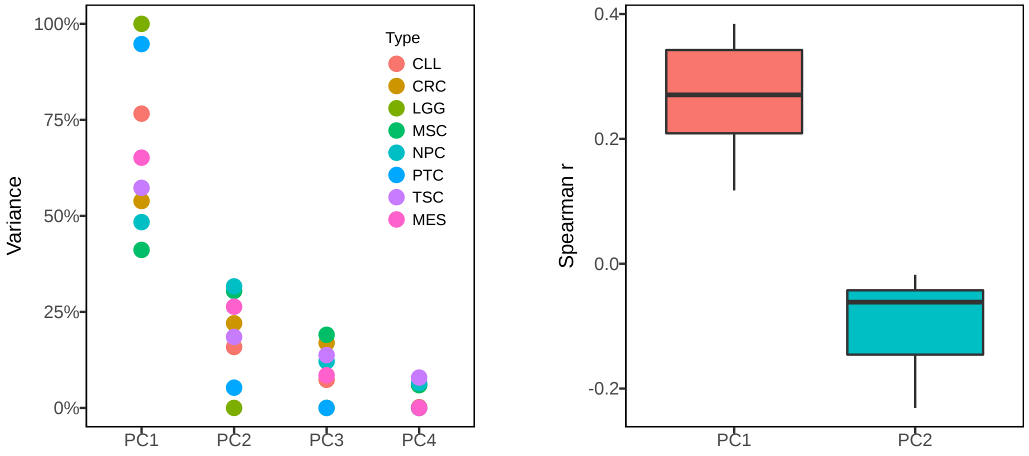
$$X_{2k}^2 \sim -2\sum_{i=1}^{k} ln(p_i) \tag{3.2}$$



(a) MSC (b) NPC (c) TSC (d) MES

Figure 3.7: Combined p-values from histone modifications and DNA methylation

Alternatively, I used a method based on principal component analysis (PCA) applied to differential signals to represent the overall epigenomic differences. Differential principal component analysis (dPCA) is one of the methods built on singular value decomposition (SVD) that compares differential epigenetic signals across multiple histone marks and replicates between two biological groups [330]. It takes the arithmetic means of each epigenetic datasets across replicates, and summarizes the observed differences between the two groups into a matrix D with genomic loci as rows and datasets as columns. The primary difference between the dPCA and conventional principal component analysis (PCA) is that it analyses the underlying true differences by decomposing D into two matrices: $D = \Delta + E$, where Δ is the underlying true differences which I am interested in, and E is the random sampling noise. E is calculated as $E = \sigma^2\Omega$, where Ω represents the diagonal matrix of eigenvalues, and σ^2 is estimated from a normal distribution over all loci. Since the number of genomic loci is typically much larger than the number of epigenetic datasets, SVD can be used to decompose the matrix $\Delta = B \times V'$, where V' is a transposed diagonal matrix, and $B = (\beta_1, \beta_2, \cdots, \beta_n)$ in which β_j characterizes the variation in Δ contributed by pattern v_j.



(a)

(b)

Figure 3.8: Characteristics of dPCs. (a) Variances explained by each dPCs. (b) Correlation of dPCs with gene expression differences in all test cases, in which dPC1 is positively correlated with gene expression, and dPC2 is slightly negatively correlated with gene expression.

In these test cases, like the correlation of histone mark differences with gene expression differences (Fig. 3.4 b), the dPCs still have correlation with the gene expression differences (Fig. 3.8 b). Among the dPCs, dPC1 explained $\sim 40\% - 100\%$ variances (Fig. 3.8 a), and usually these variances

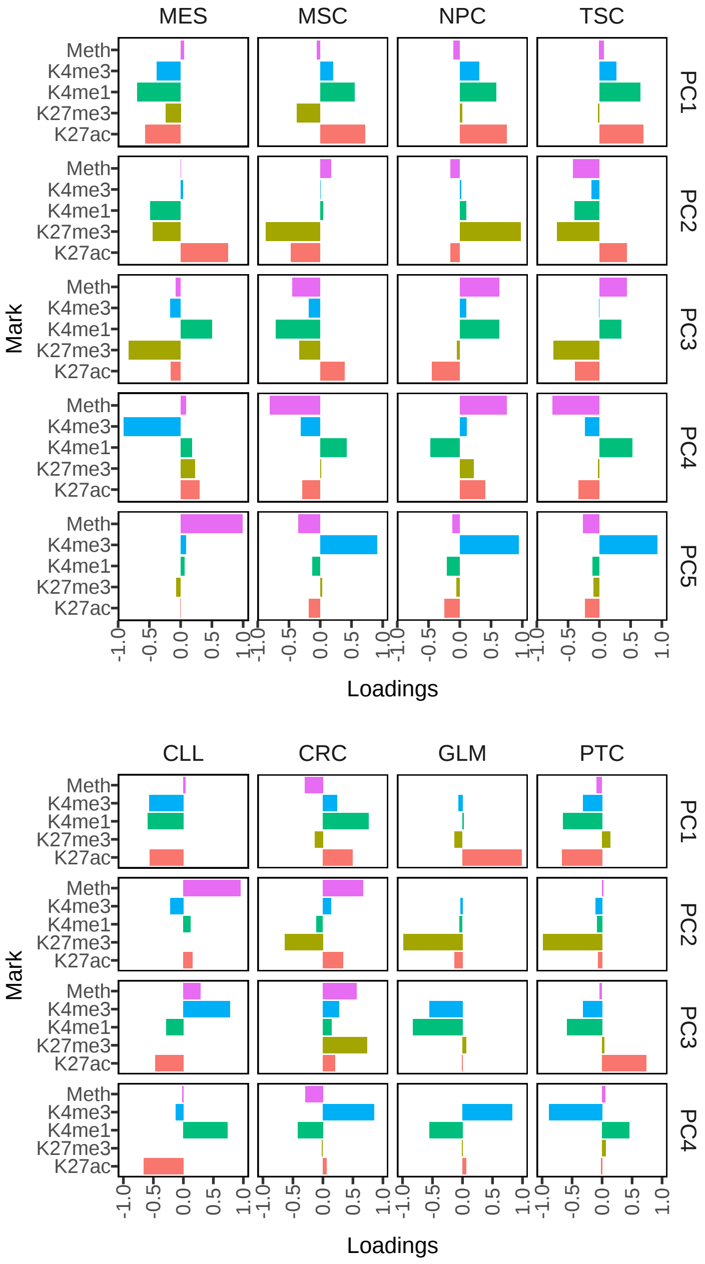


Figure 3.9: Histone mark contributions to each PC

are mainly contributed by one or two epigenetic marks. Coinciding with the results from Ji *et al.*, dPC1 appear to be mainly driven by active epigenetic marks (Fig. 3.9).

Plotting the computed dPC1 values in both promoter and enhancer regions in the context of the three of the activation marks (H3K27ac, H3K4me1 ,H3K4me3) against each biological condition, indicates that dPC1 alone is able to represent all three marks (Fig. S4 a, b). Other dPCs, taking dPC2 for example, are not representative for the above three marks (Fig. S4 c, d).

3.3 Network representation of promoter-enhancer relationships

3.3.1 Enhancer-promoter interactions

Both *in vivo* [331] and *in vitro* [50], one enhancer can regulate multiple promoters, and one promoters can be under the control of multiple enhancers, too. Therefore, enhancer-promoter interactions can be presented as a bipartite graph, in which both enhancers and promoters are represented as vertices, and directed edges link enhancers to their target promoters. In this oriented graph, the vertices are weighted according to the magnitude of epigenetic alterations at enhancers and promoters (as measured by the dPCs), and the edges are weighted according to the probability of such promoter-enhancer interactions, as will be described in the next section.

3.3.2 Estimating interaction frequencies between enhancers and promoters

Enhancers have been found to interact with promoters on the same chromosome (*cis*) or different chromosomes (*trans*). It is estimated that most often, the enhancers are within a limited distance from the target promoter of the same chromosome, a fact which is supported either by polymer physics [332, 333] or looking at the interaction densities from both experimental data (CD34 and GM12878 [334])(17 blood cells [335]) and *in silico* predictions [236,243]. Taking the promoter-enhancer (P-E) interactions from these observations and predictions, most of the interactions occur within ±1 Mb of the TSS (Fig. 3.10).

Chromatin interaction data in cancer are generally not available to us. Moreover, besides the dysregulation of histone modifications and DNA methylation, the chromosome loops in cancer cells are believed to be altered [336–338] comparing to their normal counterparts. I want to develop a method which can be applied to many different biological contexts, for which, in general, no experimental interaction data is available. Therefore, I chose to implement a universal model for interaction probabilities, fitted on a large set of experimental datasets. The interaction probabilities can be modeled

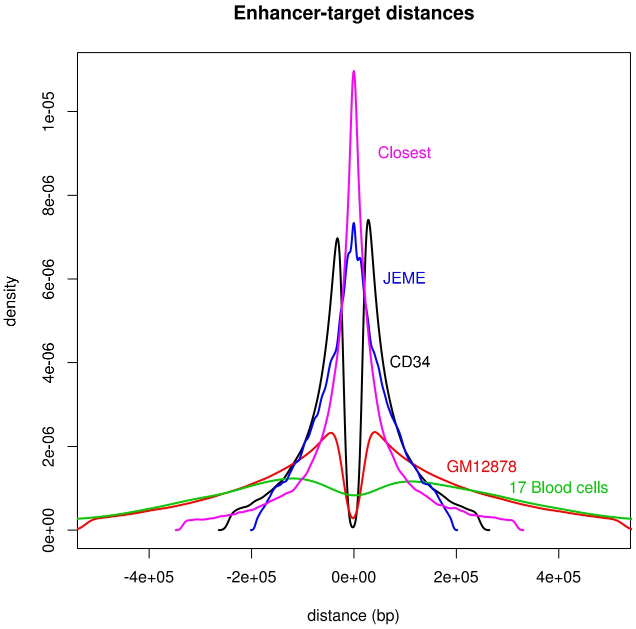


Figure 3.10: Interaction density from capture Hi-C and imputation data. Most of the inferred chromatin interactions occur within ±1Mb range from the target.

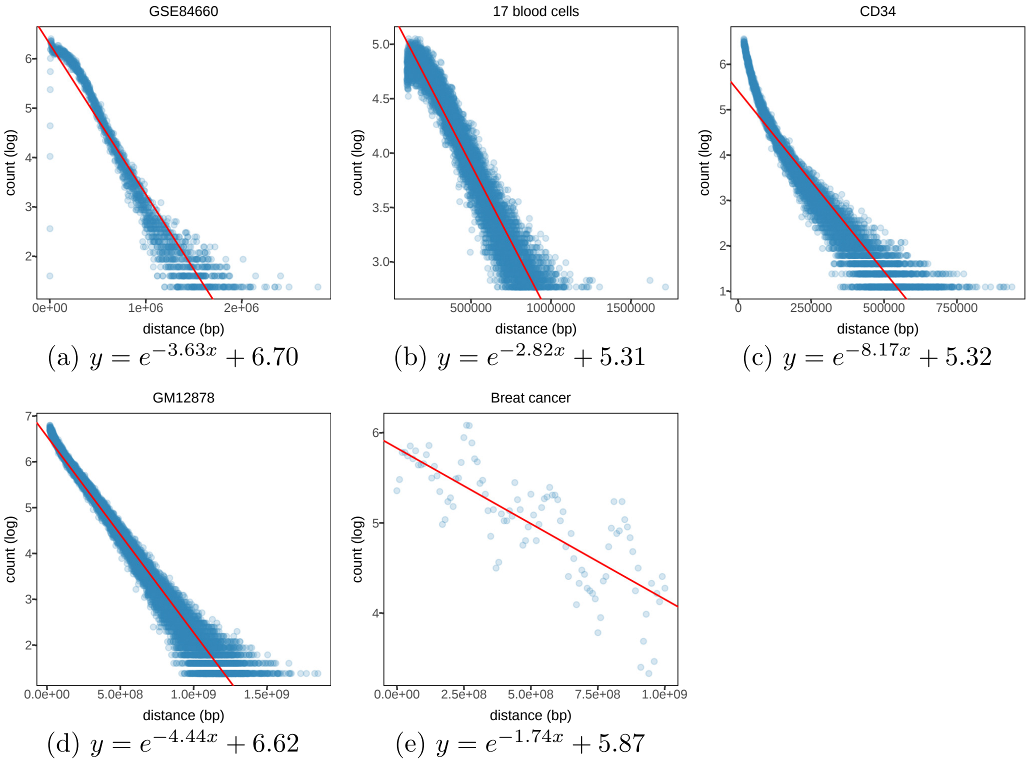
using an exponential function based on P-E distances [236]. For each genomic region window $i \in (m, m + 200bp], j \in (n, n + 200bp]$ for a given promoter position m and enhancer position n, I counted the number of interactions f_{ij} for each interaction (i, j). Afterwards, the probability y was fitted to an exponential function (3.3).

$$y \sim exp(f_{ij}, d_{ij}) \tag{3.3}$$

Although physical interactions are generally organized along the whole chromosome and even trans-chromosomes, functional interactions are more likely to be limited within topologically associating domains (TADs) [237]. Therefore, I also restricted the promoter and its interacting enhancers to stay within the same topologically associating domains (TAD). Although TADs are generally believed to be tissue-specific [339], a test using TADs from five human cell lines (embryonic stem cell, mesendoderm cell, mesenchymal stem cell, neural progenitor cell, trophoblast-like cell) provided by Schmitt *et al.* [339], has shown that 57% of the promoters have the same contacting profile in at least 80% of the cell lines.

The likelihoods of enhancer-promoter interaction can be mapped to enhancer-promoter distances with a power-law decay function [340]. I estimated consistency of contact frequency profiles from several publicly available capture Hi-C datasets, including GM12878 (E-MTAB-2323 [334]), 17 blood cells (EGAS00001001911 [335]), breast cancer (PRJEB23968 [341]), stem cells (GSE84660 [342]), and colorectal cancer (EGAS00001001085 [343]). During processing the capture Hi-C data, I required ≥ 10 reads mapped to the other end of the fragment to infer a reliable interaction.

Discrete binning is used to estimate the parameters in the distance-decay function. Afterwards, an interpolation method, as implemented by Lajoie *et al.* [340] is used. By fitting the average number of interations falling into each bin against the distances to the promoter to an exponential function, I obtained exponent values ranged from -8.17 to -1.74 (Fig. 3.11). As exponent coefficients ranging from -1 to -30 display similar performances in later benchmarking (Fig. 3.14 b), I choose -20 in the following tests.



(a) $y = e^{-3.63x} + 6.70$ (b) $y = e^{-2.82x} + 5.31$ (c) $y = e^{-8.17x} + 5.32$

(d) $y = e^{-4.44x} + 6.62$ (e) $y = e^{-1.74x} + 5.87$

Figure 3.11: Probability density functions of several capture Hi-C contact frequencies. y is the probability of the contacts, while x represents the distance (in Mb) of the interacting region to the promoter.

3.3.3 Ranking genes using personalized PageRank

I defined a meta-gene as the union of all promoters of the gene and its targeting enhancers. I adopted PageRank to summarize the weights of promoters and connected enhancers into a unique meta-gene score. PageRank is originally designed in the valuation the importance of web pages [344]. It was also adopted in bioinformatics in ranking the impact of nodes in metabolic network [345] or gene ontology network [346]. PageRank yields an importance score computed through a random walk process, in which a walker starts from a random vertex, and walks to another connected vertex randomly. This process can be repeated many times. In the end a rank is used to represent the frequency of visit of each vertex, which is calculated as (3.4).

$$PR(u) = c \sum_{v \in B(u)} \frac{PR(v)}{N_v} \tag{3.4}$$

where u represents a vertex, and $B(u)$ are the incoming vertices linked to u, $PR(u)$ and $PR(v)$ represent ranks of vertices u and v, respectively. N_v denotes the number of outgoing links from node v.

Depending on the network structure, in a scenario where there are only incoming links to a vertex, but no outgoing links from that vertex, the walker will stop at the vertex and the process terminates. To solve this problem, a reset parameter α is added to allow the walker to restart at any other random vertex, therefore the final rank becomes (3.5).

$$PR(u) = (1 - \alpha) + \alpha \sum_{v \in B(u)} \frac{PR(v)}{N_v} \tag{3.5}$$

In practice α is usually set to 0.85, which means the walker has 85% probability to follow a outgoing link from current vertex, and 15% probability to hop to a random vertex.

In another scenario specific to my application, the walker has a preference for some vertices or links over the other ones, and therefore weights of vertices and edges are introduced. In a "personalized" PageRank, the rank is also dependent on the weights of the incoming and outgoing links (3.6).

$$PR(u) = (1 - \alpha) + \alpha \sum_{v \in B(u)} PR(v) W^{in}_{(v,u)} W^{out}_{(v,u)} \tag{3.6}$$

where $W^{in}_{(v,u)}$ and $W^{out}_{(v,u)}$ are calculated based on the number and weight of incoming links and outgoing links of vertex v and u.

I used personalized PageRank implemented in *igraph* [190] to uncover important epigenetic alterations for a gene by taking into account of the regulatory contribution of enhancers. Upon setting the weights of vertices in the random walk, the promoters or enhancers with more significant epigenetic alterations have higher weights, which is represented by the dPCs previously computed. The dPCs are sorted in decreasing order regardless of their direction of alteration, as several studies suggest that both up and down regulation of histone acetylation can contribute to gene activation [347], as well as hypomethylation and hypermethylation both affect transcription factor binding, depending on the preferences of transcription factors [176].

I also set edge weights in accordance to the probabilities of chromatin contacts to ensure that only highly confident enhancers are contributing. Since the enhancer-promoter network is a directed graph, all the enhancer scores will eventually be attributed to their interacting promoter. This ensure that, even in a case where the promoter shows little epigenetic alterations, the corresponding meta-gene might has a high score due to the contribution of enhancers. In a next section, I will specifically discuss such cases. In the end, PageRank returns a vector with the rank scores for all meta-genes, in which all the genes are ordered according to the cumulative score of their promoters and associated enhancers.

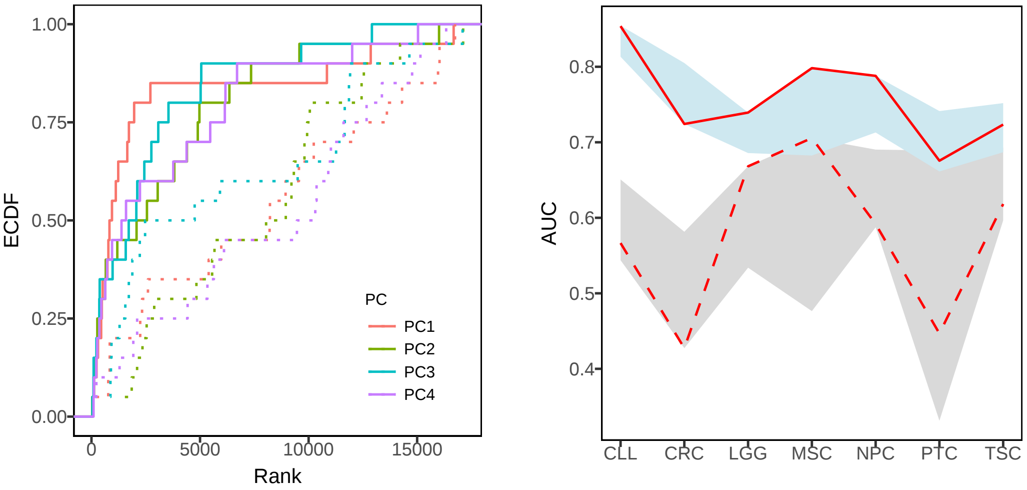
3.3.4 Benchmarks

In order to validate the outlined procedure, I designed several benchmarking strategies which I will discuss in the next sections.

Benchmarking using rank lists

The PageRank algorithm sorts the genes into descending order which is in accordance to the significance of alterations from both promoter and enhancers (abbreviated as ”PromEnh” rank list). First, I found that the PromEnh rank list is more relevant to the biological conditions than the rank list derived from the dPC1 order of only promoters (abbreviated as ”PromOnly” rank list). In order to show how much improvements I have using the PromEnh rank list, I compiled a list of 14 $\sim$ 36 marker genes for each biological test case (Supplementary table S2 and S3), either selected from comprehensive literature reviews [348–352], or cancer signature databases, including COSMIC [353], Intogen [354], MalaCards [355], etc.

Afterwards, the receiver-operating characteristics (ROC) of cumulative fraction of markers genes through the ranked list of genes is determined for assessing the sensitivity and specificity of the ranking. The rank list can be generated from other dPCs as well. Taking the area under curve (AUC) in the CLL case as an example, as 70%-80% of the CLL specific marker genes are enriched in the top 20% genes of the PromEnh rank list, while for the PromOnly rank list this value is below 60% (Fig. 3.12 a). This conclusion is also true for other test cases as well as most of the other dPCs (Fig. 3.12 b).



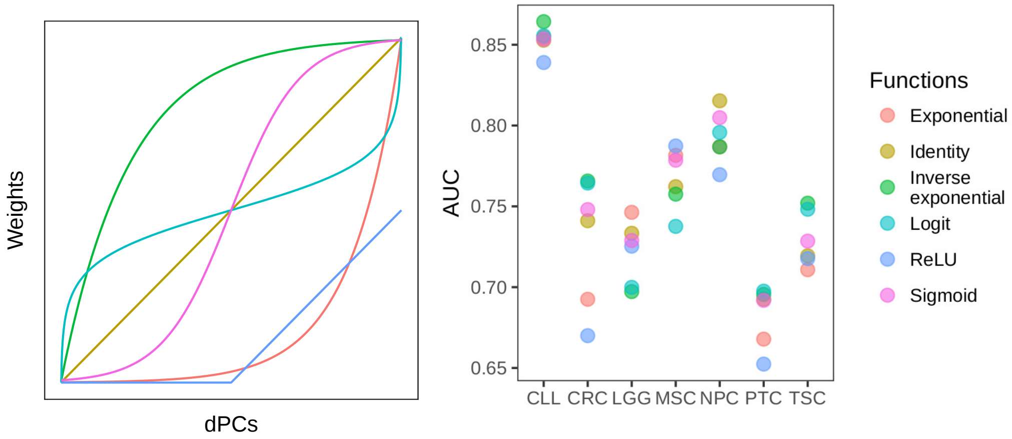
(a)

(b)

Figure 3.12: ROC of CLL and AUC of all test cases. (a) CLL ROC of PC1-4 using PromEnh rank list (solid lines) and PromOnly rank list (dashed lines). (b) AUC of dPC1 are plotted as red lines (solid: PromEnh, dashed: PromOnly), and AUCs from other dPCs are plotted as shaded bands between the minimum and maximum scores (lightblue: PromEnh, grey: PromOnly).

Benchmarking with transformed vertex weights

So far it is still questionable whether dPC scores are directly related to the importance of the vertices. Indeed, it could be that, below a certain threshold, small differences in epigenetic signals (and hence small dPC scores) have no impact on the state of the gene, and should have a zero contribution to the overall score. Hence, I tested a number of transformations of the raw dPC scores using functions frequently used in data transformation in artificial neural networks (Table 3.2) to introduce non-linearity into the vertices weights (Fig. 3.13 a). Benchmarking with dPC1 using different transformation functions ended up with similar AUCs (Fig. 3.13 b). Therefore, I directly used dPCs as the vertex weights to avoid introducing any undesired side effect.



(a) Function shapes

(b) AUCs

Figure 3.13: Transforming functions and AUCs of each function

Table 3.2: Weights transforming functions

Function	Equation
Sigmoid	$f(x) = \frac{1}{1+e^{-x}}$
Logit	$f(x) = ln(\frac{x}{1-x})$
Exponential	$f(x) = e^x$
Inverse exponential	$f(x) = ln(x)$
Identity	$f(x) = x$
Rectified linear unit (ReLU)	$f(x) = \{ \begin{array}{lll} 0 & \text{if} & i < \frac{n}{2} \\ x & \text{if} & i \geq \frac{n}{2} \end{array}$

Benchmarking with biological evidences

The PromEnh rank lists not only show better coincidence with selected marker genes comparing to the PromOnly rank lists, but also show better tissue specificity and are enriched with oncogenes in cancer samples. To show this, tissue specific genes were retrieved from ARCHS4_Tissues [356] which is provided with the EnrichR tool [357]. Adjusted p-values from EnrichR indicate tissue specific enrichment in corresponding test cases. The most significant term from the enrichment of the top 1000 genes in each PromEnh rank list shows a remarkable enrichment of corresponding tissues, while the most significant terms from the top 1000 genes in PromOnly rank list are not relevant to the corresponding tissues (Table 3.3). Given the fact that super-enhancers near cell type specific genes often accumulate disease associated non-coding variants [58, 358], this phenomenon is not a coincidence as it is applicable to all the cancer test cases, implying enhancers play a dominant role in cell type specific regulations.

Table 3.3: Most enriched tissues corresponding to the PromEnh and PromOnly rank list of test cases

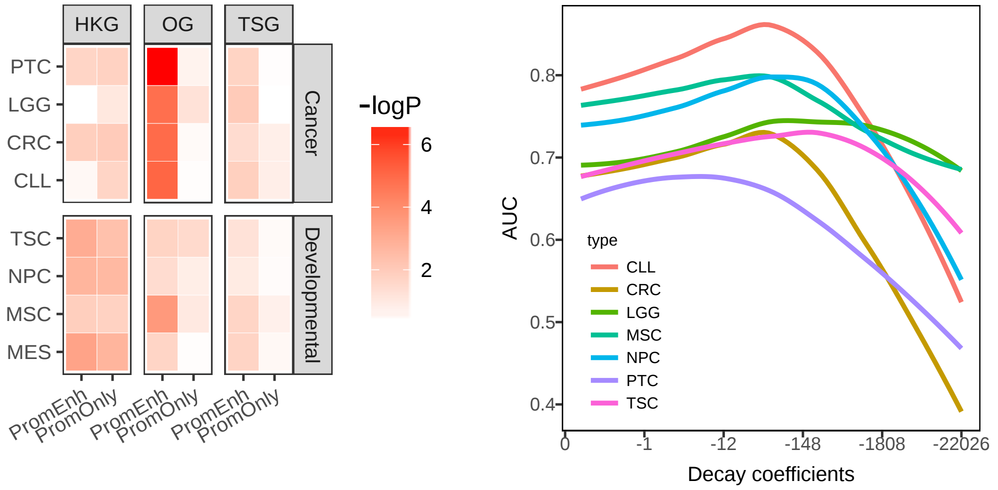
Test case	PromEnh tissue	Adj. P*	PromOnly tissue	Adj. P*
CLL	CD19+ B cells	6.1e-10	Breast (bulk)	2.3e-23
CRC	Small intestine (bulk)	1.5e-16	Spinal cord	2.5e-108
LGG	Prefrontal cortex	3.9e-24	Testis (bulk)	0.07
PTC	Thyroid (bulk)	2.1e-15	Brain (bulk)	7.0e-33
NPC	Spinal cord	3.2e-27	Renal cortex	1.6e-4
MSC	Astrocyte	5.7e-37	Fibroblast	3.9e-24
TSC	Fibroblast	3.9e-24	Lung (bulk)	6.8e-12

* Adjusted p-value (Benjamini-Hochberg method)

In addition, I found that oncogenes are also highly ranked in the PromEnh list. I performed Wilcoxon-Mann-Whitney tests on the positions of oncogenes (OG), tumor suppressor genes (TSG), and housekeeping genes (HKG) in the PromEnh rank list against a uniformly distributed rank list of the same length. Comparing to TSG, HKG and random gene sets, OG show a higher ranking which might be interpreted as a tendency to be under strong epigenetic regulation in cancer cells. This phenomenon is not observed in the test cases with normal cells (NPC, MSC, TSC, MES) (Fig. 3.14 a). On the contrary, the test cases related to developmental processes show that HKG are significantly ranked higher when epigenetic regulation is taken into account. However, all p-values are insignificant when performed using PromOnly rank list, suggesting that the ranks of OGs and HKGs are mainly explained by enhancer contribution. Looking at individual genes, BRAF, KRAS are associated with stronger enhancer regulation. On the other hand, TP53 is often associated with weaker enhancer regulation (Fig. 3.15).

Benchmarking with distance based functions

As the promoter-enhancer contact frequencies discussed above, the distances from the enhancer to the TSS of the promoter determine the likelihood of promoter-enhancer interaction, and the effect can be modeled with power-law decay. I tested a series of exponent values in the decay function from which the edge weights were calculated as (3.7).



(a) Wilcoxon tests of OG,TSG,HKG ranks comparing to uniformly distributed rank. OGs in cancer are significantly ranked higher in PromEnh rank list, whereas they are not significant in developmental test cases

(b) AUCs from a range of exponential decay coefficients applied to promoter-enhancer distances. From 0 (no distance decay) to -22026 (edge weight reduced by 99% when P-E distance is over 200bp)

Figure 3.14: Rank list benchmarking.

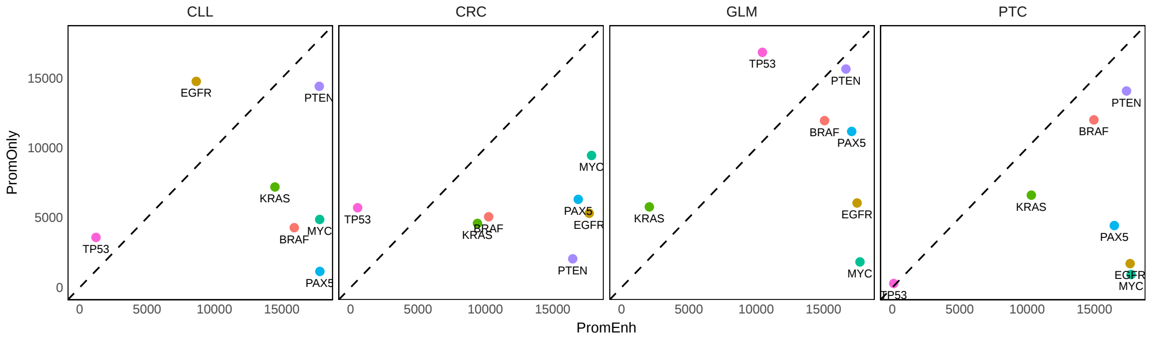


Figure 3.15: Ranking positions of well known genes in PromEnh and PromOnly list. Genes on the bottom-left corner are ranked low in both lists, while genes on the top-right corner are ranked high in both lists. Genes on the top-left corner are only ranked high in PromOnly list, and genes on bottom-right corner are only ranked high in PromEnh list

$$w_i = exp(m \times d_i + \beta) \tag{3.7}$$

where m is the coefficient I am estimating, and d_i indicates the distance (in Mb) between the enhancer to the promoter, and β is the intercept from the fitted function. Therefore, setting $m = 0$ implies that all enhancers have the same weights regardless of their distances to the promoter. Also calculated from (3.7), when $m \approx -3912$, the weight of an enhancer with a distance of 1kb to the promoter drops by 98% comparing to an enhancer at the same genomic location of the promoter, which essentially means that the contributions of enhancers in the PageRank are negligible in this scenario, and the AUCs are very close to ranking promoters only. Therefore, a coefficient $m \in (-10, -30)$ seems suitable in maximizing the AUC and matching with the experimental estimation (Fig. 3.14 b), without exaggerating the contribution from enhancers.

Robustness of the gene ranking under random perturbations

To validate that the contributions from enhancers are not an artifact, I used degree-preserving random perturbations, which rewires the endpoints of the edges with a 50% probability randomly to another vertex in a graph. The randomization can be realized using a rewired promoter-enhancer network in PageRank. During the test with 100 permutations of different rewired network structures, I used the same marker genes as in the benchmarking, and the AUCs with randomly assigned enhancers dropped 10%-20% for most of the test cases (Fig. 3.16). Nevertheless, the AUCs from rank list including random enhancers still outperformed the rank list using promoter only. I assume that the reason is because the rewiring did not change the number of enhancers that a gene might have. Therefore, the ranking of cancer and development marker genes will benefit from enhancers no matter which enhancers were linked to them, but the ranks of the other genes are disordered and hence the AUCs are lower.

3.4 Network representation of gene relationships

3.4.1 Network construction

Considering the top ranked genes could be artifacts, the relatively high ranked genes with known biological functions are more appealing to me, even though they may not be the highest in the rank list. The most common way to perform functional annotation of a rank list of genes is by doing a gene-set enrichment analysis (GSEA). But applying GSEA on high-throughput epigenetic data is still debatable, as severe biases towards genes related to

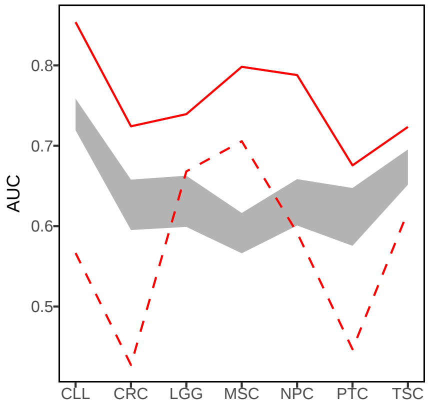


Figure 3.16: AUCs of PromEnh rank list from randomized promoter-enhancer interactions. The grey band regions indicate the quantile ranges from benchmarking each with 100 different rewired promoter-enhancer networks, whereas the red lines show the AUCs with the original promoter-enhancer interactions from PromEnh (solid line) and PromOnly (dashed line) rank lists.

development and differentiation have been mistakenly reported in a wide range of DNA methylation studies [359]. Therefore, I performed a network analysis over the rank lists. The rank lists are examined in a context of biological networks such as protein-protein interaction networks, and significantly altered genes might appear clustered together in accordance with their biological functions. The context can be known signaling pathways, co-expression, or protein-protein interactions etc. Here I chose Human Protein Reference Database (HPRD) as the reference network [360]. HPRD contains manually curated scientific information of most human proteins related to their biological functions, including protein-protein interactions, post-translational modifications, enzyme-substrate relationships and disease associations.

I used *igraph* to find communities from HPRD via short random walks. Multiple edges and self-loops are removed. I limited the genes of interests to a percentage of the top ranked genes, and the edge weights are computed as the average rank of the two connected genes. The edge directions are ignored in HPRD in random walk. Using the highly ranked genes as "seeds", the random walk clustering returns several dense subnetworks. The genes in the subnetworks are non-overlapping. This method recovered more cancer-related genes due to the fact they have more interacting partners than non-cancer genes [361–363]. In a test I took out the top fifteen largest subnetworks, generated with thresholds of the top 10%-50% ranked genes.

The genes enriched in the subnetworks exhibit higher oncogene frequencies comparing to the frequencies in the PromEnh rank list. As the frequencies of oncogenes dropped along with the PromEnh rank list, the oncogene frequencies stayed stable in the extracted subnetworks, implying the network clustering succeeded in selecting the biological meaningful genes in cancer (Fig. 3.17).

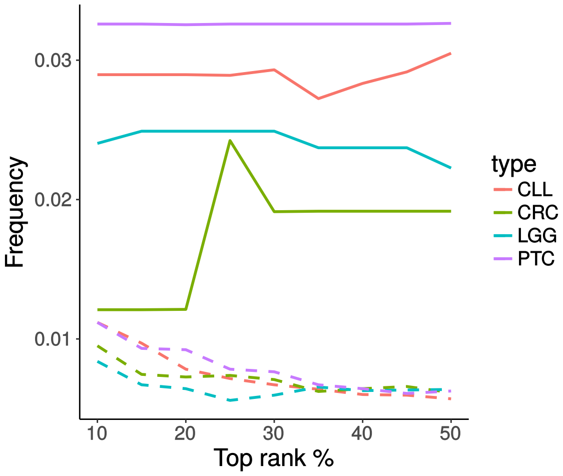


Figure 3.17: Frequencies of oncogenes in the network clustering. The solid lines represent frequencies of oncogenes in the enriched subnetworks built from top 10%-50% genes of PromEnh rank list, while the dashed lines indicate the frequencies of oncogenes in the top 10%-50% of the PromEnh rank list.

Identifying modules and pathways in diseases

Functional analysis of the subnetworks obtained as described previously uncovers a few interesting pathways for my test cases by analyzing the genes in these subnetworks with EnrichR. Many of them are general biological processes and signaling pathways, yet some of them show specific functions related to the cancer types. In table 3.4 I listed a few of the subnetworks ranked in descending order of the sum of vertex weights.

In the center of the subnetworks reside the hub regulators, which are linked by many genes with epigenetic alterations. For example, SMAD2 and SMAD3 expression have been shown to be regulated by histone modifications of their promoters [383]. In the following part, I will highlight two cancer specific pathways.

Table 3.4: Top ranked specific functions from network clustering

Case	Biological functions	Hub regulators*	Adj. P**
CLL	Toll-like receptor signaling pathway	TLR1, TLR4 [364]	1.0e-7
CRC	HIF-1 signaling pathway [365, 366]	EPAS1 [367]	7.6e-4
LGG	TGF-β signaling pathway [368]	SMADs	1.8e-6
LGG	Notch signaling pathway [350]	NOTCH2 [369]	7.5e-14
PTC	PI3K-Akt signaling pathway [370]	FGFs [371, 372]	4.0e-9
MSC	Osteoblast signaling pathway	PTH [373]	1.0e-3
NPC	Hedgehog signaling pathway [374]	ZIC3 [375]	7.6e-9
NPC	Axon guidance [376]	Ephrins [377, 378]	2.8e-18
TSC	MAPK cascade [379, 380]	MAPK1 [381]	6.5e-16
TSC	EPO receptor signaling pathway [382]	PTPRC	0.015

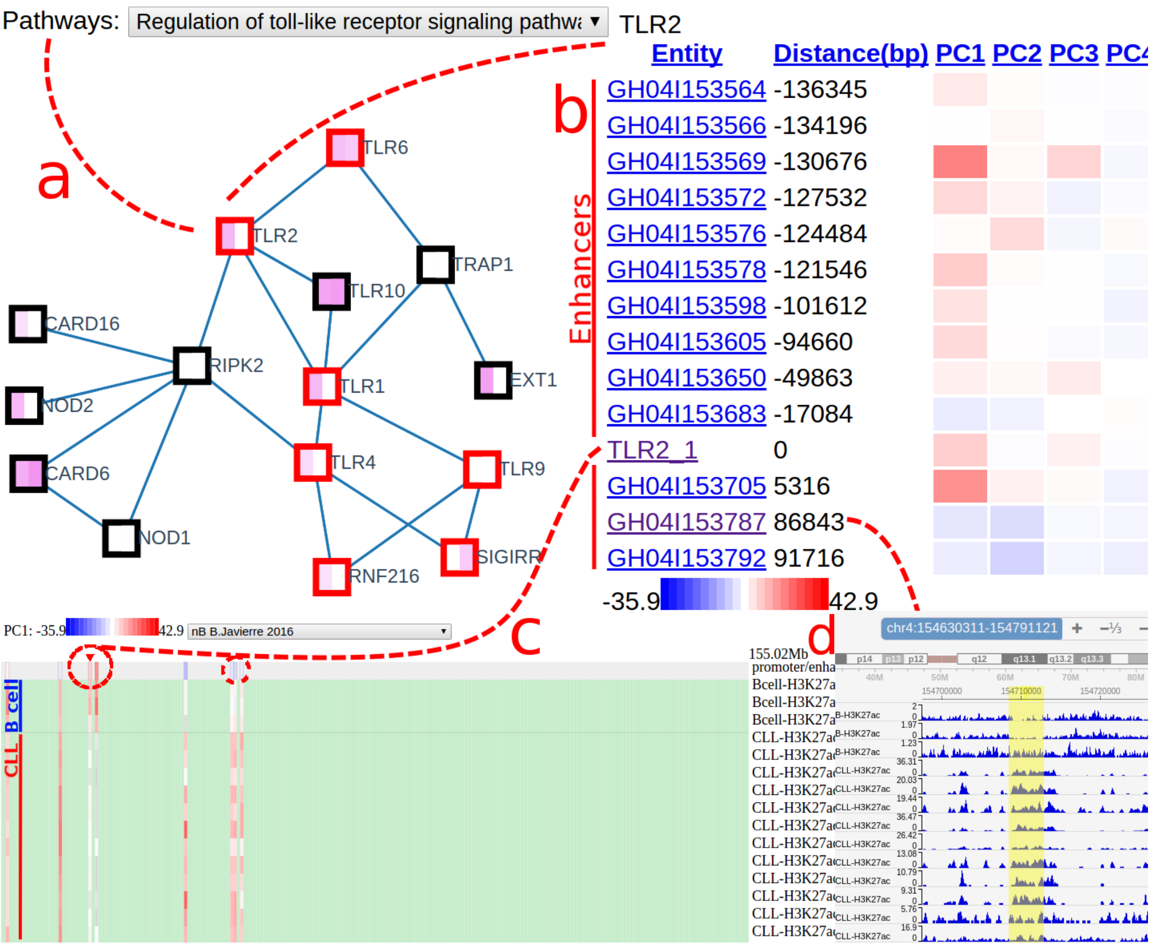
* Hub regulator indicates a gene surrounded by several significantly altered genes.
** Adjusted P-value (Benjamini-Hochberg method)

Chronic lymphocytic leukemia

Toll-like receptors (TLR) are iconic markers in both normal and malignant B-cells. They mediate innate immune response via pattern recognition of antigens. TLR4 and TLR9 gene expression were lower in CLL than in healthy individuals [364, 384], while TLR2 was highly expressed in both CLL [384] and acute myeloid leukemia (AML) [385]. Accordingly, I observed both hyperacetylation and hypoacetylation of the TLR genes and their neighboring enhancers in my test case (Fig. 3.18), which may lead to their differential expression in the end. The toll-like receptor signaling pathway is only recovered from network analysis, and GSEA did not identify this pathway, which indicates network analysis is powerful in enriching pathways with few members.

Colorectal cancer

Colorectal cancer (CRC) is a malignant cancer affecting colon or rectum, and accounts for $\sim 9\%$ of all cancer deaths [386]. Network clustering of the epigenetic alterations of promoters and enhancers suggests that HIF-1 pathway is the top candidate. HIF-2α overexpression is frequent in multiple cancers, and is associated with poor prognosis [365, 366]. I further identified EGLN3 and HIF-2α (EPAS1), whose scores are mainly contributed from



Entity	Distance(bp)	PC1	PC2	PC3	PC4
GH04I153564	-136345				
GH04I153566	-134196				
GH04I153569	-130676				
GH04I153572	-127532				
GH04I153576	-124484				
GH04I153578	-121546				
GH04I153598	-101612				
GH04I153605	-94660				
GH04I153650	-49863				
GH04I153683	-17084				
TLR2_1	0				
GH04I153705	5316				
GH04I153787	86843				
GH04I153792	91716				

Figure 3.18: Toll-like receptor signaling pathway is associated with strong epigenetic alterations in CLL. Here I attached a screenshot from *crl* program to illustrate the analysis procedure: a. Network browser showing enriched pathways, corresponding genes are highlighted in red on selection from the drop list; b. Clicking on a gene of interest will direct to a web page showing neighboring enhancers, as well as their dPCs in colored ranges; c. By clicking the promoter ID of the gene, a heatmap-like epigenome browser appears, showing the intensities from each epigenetic mark between tests and controls; d. Clicking the enhancer IDs will allow users to read additional information from GeneCards [384], as well as their predicted targets; Clicking a track in the epigenome browser will redirect users to WashU Epigenome Browser [260], showing a more detailed view of epigenetic signals for that point of genomic region.

epigenetic alterations of nearby enhancers (GH14I033835 and GH02I046346, respectively). These genes are also differentially expressed between CRC and normal cells. Baba *et al.* also confirmed EPAS1 overexpression in a cohort of 731 colorectal cancers [387], and Yoshimura *et al.* proved EPAS1 is associated with high grade colorectal cancer in 88 patients [367]. The enriched network not only revealed EGLN3 and EPAS1, but also included HIF-1α, MYC and MAX as potential regulatory factors.

3.4.2 Discovery of novel cancer related enhancers

Besides studying the pathways enriched for epigenetic alterations, I also analyzed the highly altered enhancers associated with certain genes of interests. PAX5 and MYC are two candidate genes discovered from their consistent high ranking in PromEnh list from all the test cases (Fig. 3.15). As their ranking in the PromOnly are not high, it suggests PAX5 and MYC are extensively regulated by enhancers. PAX5 is a key TF involved in B-cell development, and its promoters have no significant epigenetic alterations in CLL. However, taking the nearby enhancers into consideration, this gene is associated with several hyperacetylated and hypomethylated enhancers, one of which located 330 kilobases (kb) upstream of the PAX5 TSS has been found as extensively mutated in CLL [388] (Fig. 3.19). Deletion of this enhancer resulted in a 40% reduction in the expression of PAX5 and chromatin interaction of this enhancer and PAX5 have been proven from chromosome conformation capture sequencing (4C-Seq) analysis [388].

By analyzing the significantly altered oncogenes across cancer cohorts, I identified a large enhancer region, also referred to as super-enhancer, regulating MYC [58]. The enhancers are located around 20 $\sim$ 200kb downstream of MYC, overlapping with the long non-coding (lncRNA) PVT1. PVT1 has been found to positively influence MYC expression [389–391], and is also considered as an oncogene. Cho *et al.* found that PVT1 competes with MYC for access to this super enhancer, and disruption of the PVT1 promoter leads to greater enhancer activity of MYC as well as an increase in expression [392]. The enrichments in H3K27ac and H3K4me1 signals of this enhancer have been observed in all cancer test cases (Fig. 3.20). Also, *cis*-interactions were detected using an orthogonal conformation capture technique [393]. The Hi-C profiles from the most similar available tissues or cell lines of each cancer type (K562, Huvec, GM12878 [232], Hippocampus [339] were used in upholding the chromatin interactions in CLL, CRC, PTC, LGG, respectively) are visualized using Hi-C data Browser [394] and placed above the genomic view. Another enhancer in the vicinity of MYC, known as the "Blood ENhancer Cluster" (BENC) [395], has also been recognized from increased H3K27ac levels in CLL (Fig. 3.20 c) and LGG (Fig. 3.20 g).

3.4.3 Development of an R package for integrative epigenetics analysis

I have developed an R package named "*crl* " incorporating the differential analysis steps described in this chapter. *Crl* features a variety of tools for visualizing network clusters and epigenetic alterations, which is crucial for both hypothesis generation and detection of potential artifacts. Comparing with other network presentation tools, such as Cytoscape [396], the display of networks in *crl* is fully adapted to highlight differential epigenetic alterations from both the PromEnh and PromOnly rank list (Fig. 3.18 a), allowing one to inspect principal components of epigenetic marks for the gene of interests, and further navigate to a snapshot of the genomic regions of epigenetic tracks from each biological replicates, while existing tools, such as the WashU Epigenome Browser [260], load very slowly when visualizing a large amount of epigenetic tracks (Fig. 3.18 c).

An alpha version of this tools is available in GitHub for inspecting the enriched pathways for the cancer and development test cases. It can be accessed from `http://qwang-big.github.io/crl-web/` .

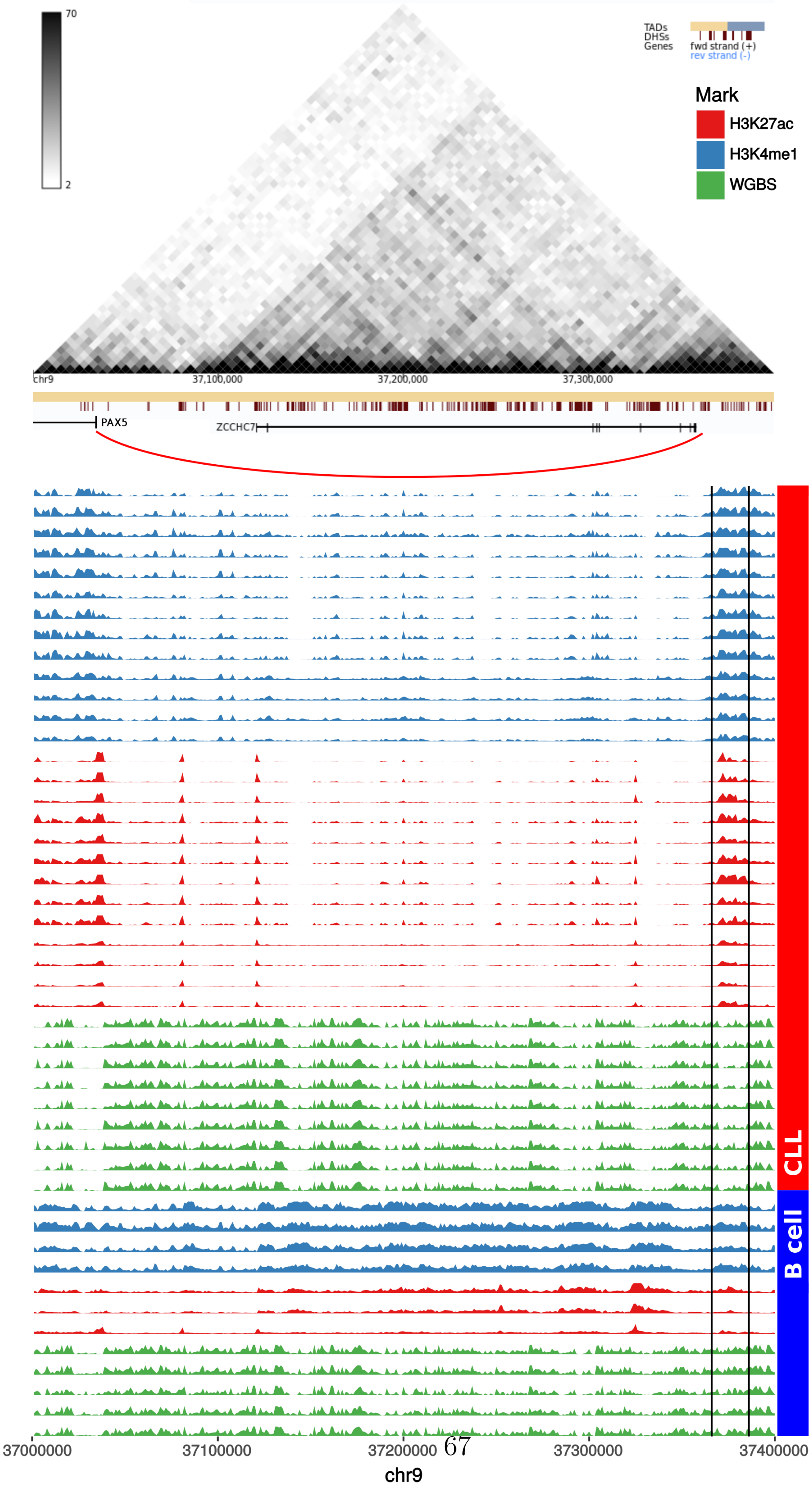
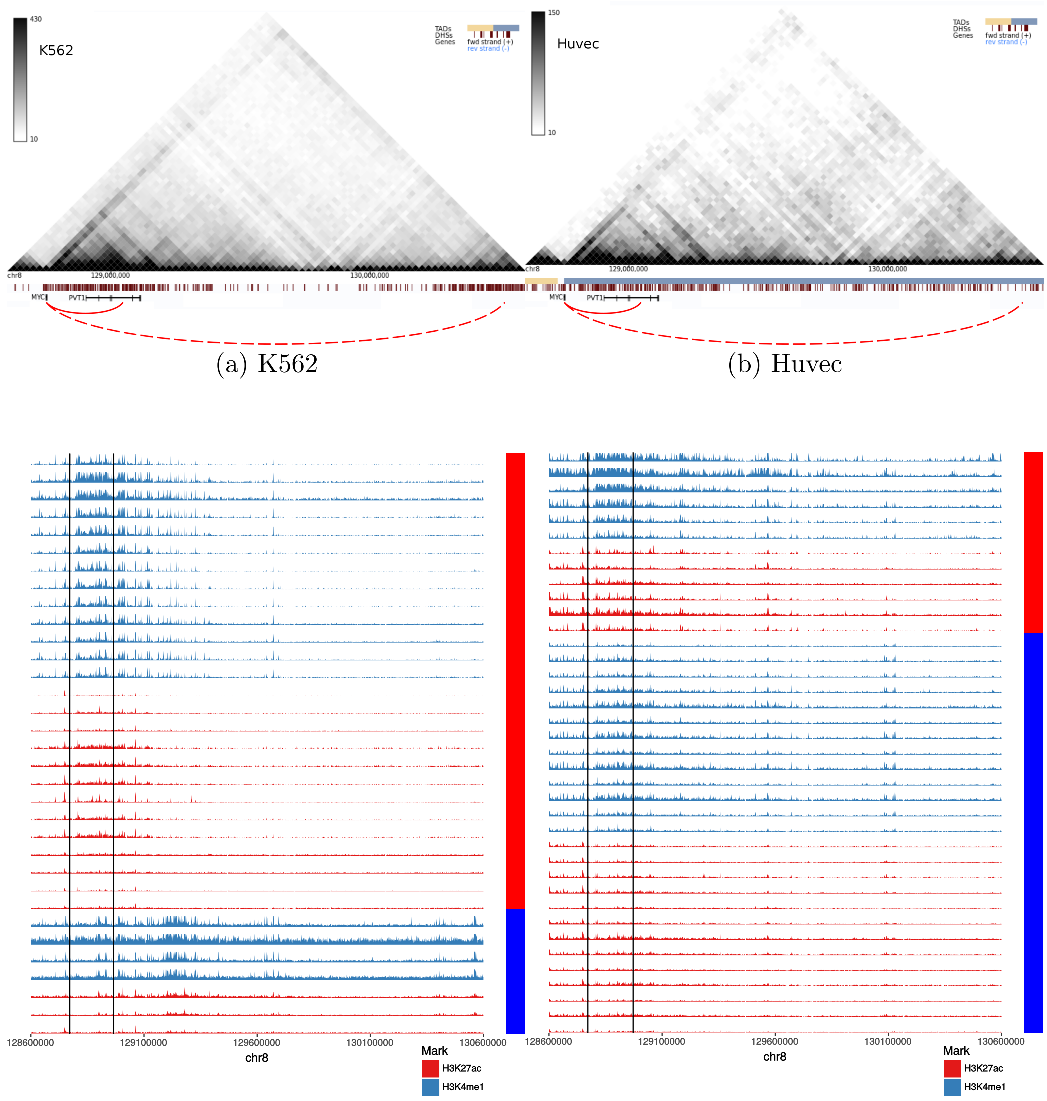


Figure 3.19: A known PAX5 enhancer in CLL exhibits hyperacetylation and hypomethylation.



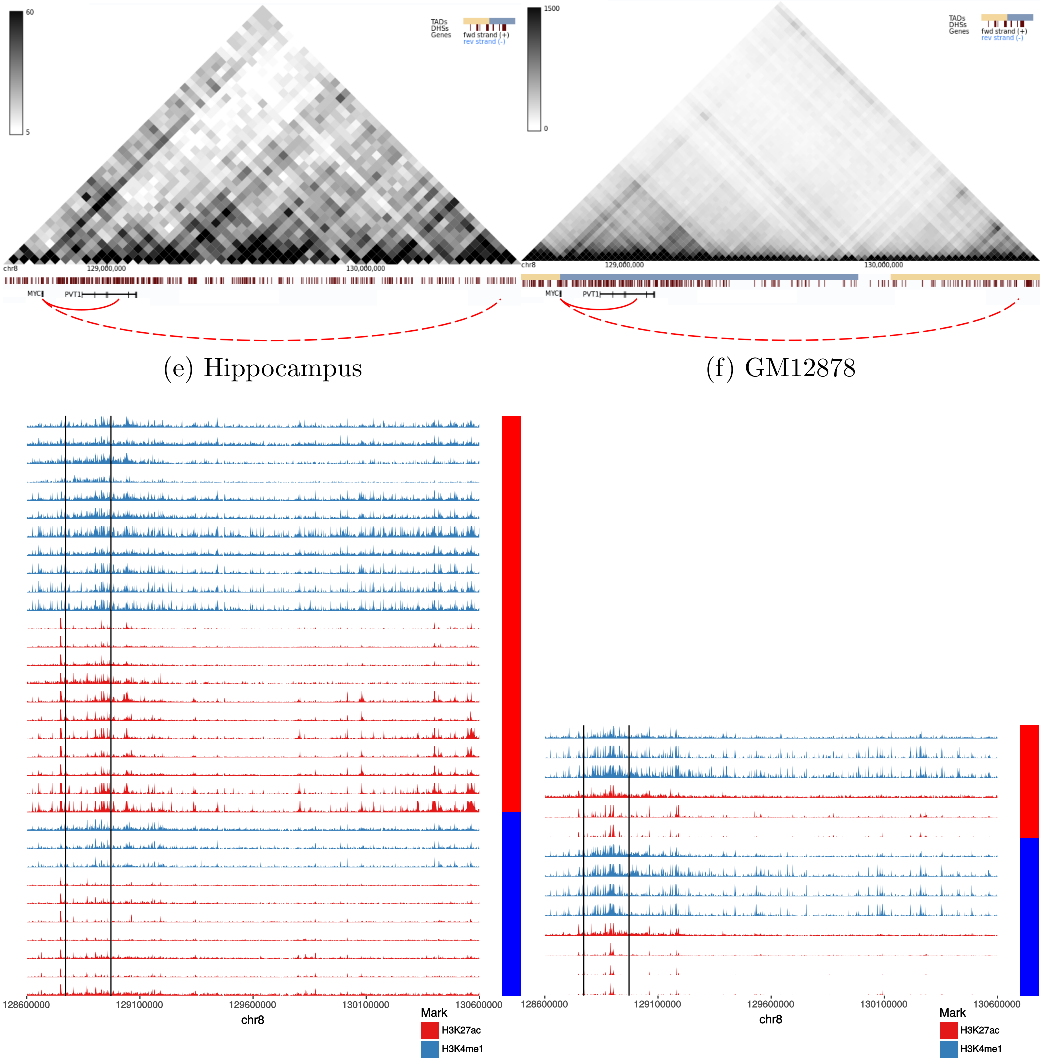
(a) K562

(b) Huvec

(c) CLL

(d) CRC

Figure 3.20: Continued on next page.



(e) Hippocampus

(f) GM12878

(g) LGG

(h) PTC

Figure 3.20: A super enhancer near MYC exhibits hyperacetylation in cancers. The proposed super enhancer regions are bordered with black solid lines on the genome tracks, wherein the cancer samples are marked with red strips, and normal samples are marked with blue strips. Solid red lines indicate the chromatin interaction between MYC and this enhancer in Hi-C, and dashed lines indicate the interactions between MYC and another superenhancer "BENC".

Chapter 4

Discussion

In this thesis I presented several approaches to incorporate and combine multiple epigenetic data types, from multiple regulatory loci with complex relationships, to describe the underlying regulatory mechanisms in cell development and cancer. Generally, these approaches perform either quantitative or qualitative analysis of epigenetic modifications. Qualitative analyses are efficient in solving simple problems, such as definition and classification of regulatory regions, but become incompetence when complex relationships and multiple dependencies are involved. Quantitative analyses, on the other hand, are powerful when it comes to intensities and probabilistic levels of data integration, yet comprehensive benchmarking is essential in defining and validating the modeling assumptions. Both approaches are covered in this thesis, and their applicable scenarios and efficiencies are discussed in the applications of a variety of cancer and development test cases.

4.1 Descriptive analysis of the epigenetic modifications

In first main topic, I focused on the descriptive analysis of the epigenetic modifications. I analyzed enriched regions (peaks) from the ChIP-Seq of histone modifications in a particular cancer type, namely glioblastoma multiforme (GBM). As the presence of the peaks is the only feature in descriptive analysis, the accuracy of peak calling is crucial in making conclusions. Therefore, various settings with several peak callers have been tested, and they presented similar outputs at sharp peak regions (H3K4me1, H3K4me3, H3K27ac), but vary largely in the broad peaks (H3K9me3, H3K27me3, H3K36me3). Balancing the trade-off between sensitivity and specificity, I tend to use the same peak caller for all analyses, as the GBM classification with SICER led to a remarkable agreement with other subtyping approaches, such as 450k methylation array and WGBS. I also inferred regulatory elements directly from the positions and combinations of these peaks, and I distinguished active/poised promoters and enhancers. The common and specific regulatory elements in GBMs and normal brain tissues reflect the distinct possible mechanisms in tumorigenesis and progression by means of both pathway analyses (Fig. 2.6) and subtype specific signature genes (Fig. 2.7)

4.1.1 Chromatin states

In order to comprehensively depict all the combinations of epigenetic modifications for GBMs, I employed the widely used tool ChromHMM. There are several options in building ChromHMM, in both the binarization step and modeling step. Firstly, I tested binarized signals from peak callers, and from the built-in Poisson model of ChromHMM. Binarization using peak callers has been previously applied on building a 38 states model including 6 histone marks and 13 transcription factors by Predeus *et al.* [397], wherein they used SICER peaks with parameters for calling narrow peaks (200bp window size, 200bp gap size). However, owing to the broadness of epigenetic marks, they still need to adjust the gap size (600bp) for H3K27me3 and H3K36me3. Also, as I investigated the benchmarking of ChromHMM in GBMs, the binarized signals from peak calls with the same parameters failed to address the actual boundaries of either active or repressed regions, causing overestimation or underestimation of several types of chromatin states (Fig. 2.10). On the other hand, ChromHMM applies independent cut-off for each mark using a Poisson distribution, which makes it more suitable to use in the binarization step. In the modeling step, regarding the number of states and marks, I either directly used the Roadmap 18-states model, or learned a new HMM from the binarized data. Although I have tested a variety of numbers of states and epigenetic marks (even including DNA methylation), the model varies largely across different samples and subtypes. In addition, the models contain a large redundancy in some chromatin states, either caused by sequencing biases in ChIP-Seq data, or the underlying heterogeneity of GBM epigenomes. In the end, I used the Roadmap 18-states model for the downstream analyses, to make the result comparable to previous Roadmap analyses.

4.1.2 Subtype specific epigenetic patterns

In the end of the GBM study, I discussed the potential use of deep neural network in subtype specific epigenetic patterns classification. Deep learning is one of the most popular modern machine learning approaches. Comparing to other traditional classification methods, deep learning greatly reduces the need for feature engineering, which is one of the most time-consuming parts of machine learning practice. Specifically, comparing to the *k*-means method I used in classifying subtype specific epigenetic modification patterns, it avoids choosing the optimum number of clusters, yet it guarantees discovery of the patterns used in training. The disadvantage is that some patterns are rare, while neural networks require a large amount of data to train. This shortcoming can be compensated by making use of the synthetic data. From the test cases with one of the histone mark, the approach allowed me to identify the desired patterns of subtype active promoters, and a literature mining suggests they are highly relevant to the specific phenotypes, which makes it feasible to apply this approach with even more epigenetic marks and complex patterns.

4.2 Quantitative analysis of the epigenetic modifications

The second main topic of the thesis is on the quantitative analysis of the epigenetic modifications. To achieve this I analyzed the epigenetic contrasts between two biological conditions. Several evidences suggested that the magnitude of epigenetic differences correlates with gene expression levels. Nonetheless, epigenetic comparison is not equivalent to differential gene expression analysis. As illustrated in the ”epigenetic priming” section (3.1), differential gene expression is the possibly delayed outcome of differential epigenetic modifications, which highlights the importance of differential epigenetic analysis. In order to analyze the underlying variances of multiple epigenetic modifications, the epigenetic signal intensities of eight human cells (four cancer and four normal cells) and five stem cells (ESC and four ESC-derived cells) are generated based on the sequencing densities of pre-defined promoter and enhancer regions. As I was using the annotated enhancers from all the tissues [311], enhancers were further filtered with H3K4me1 to eliminate the unspecific enhancer regions.

4.2.1 Data validity and reliability

In the beginning, I started from analyzing the data quality by testing multiple correlations between epigenetic modification and global gene expression level, gene expression differences of two groups, and correlations between multiple epigenetic marks. These verified my assumptions that epigenetic intensities are important in epigenetic regulation. Epigenetic modification are generally categorized into two types, which are active and repressive. The epigenetic marks of the same type are highly positively correlated and negatively correlated with the other type. Therefore, a straightforward combination of multiple hypothesis testing of multiple epigenetic marks is not feasible as high rate of false positives may be introduced due to the redundancy between marks [329]. However, the highly correlated structure of epigenetic marks is preferable in a dimensionality reduction approach, making it possible to represent these marks with a few principal components using dPCA [330].

4.2.2 Dimensionality reduction

Many challenges arise in dimensional reduction step. In the first place, data normalization is challenging due to the different background noises between epigenetic marks. So far I use normalization methods developed in RNA-Seq analyses, which does not make use of the ChIP-Seq control. In my current approach, the signal levels are estimated from uniform background noise distribution. However many studies have shown that the influences from chromatin accessibility [398], GC content [185,186], copy number variation [187], may cause biases in the noise model. Furthermore, dPCA explicitly use normal distribution to estimate signal-to-noise ratio, which requires stabilization via variance transformations.

Also, for the follow-up analyses of promoters and enhancers ranking with multiple epigenetic marks, linear approximations are used in nearly all levels of data integration. Specifically, as I want to factorize the high-dimensional epigenetic data, the dPCA linearly maps the data points to a low-dimensional latent space. Because I only used PC1 for analyzing the differences from multiple epigenetic marks, although the low-dimension representations of some of the datasets were satisfying (Fig. 3.8 a), there is not always a linear function that can explain the relationship between genomic datasets [242]. Therefore, non-linear dimensional reduction techniques, such as t-stochastic neighbor embedding (t-SNE) and autoencoder may be expected to have greater potential in complex models.

4.2.3 Incorporating non-coding elements

There is a great necessity in incorporating distal non-coding regulatory elements into differential epigenetic comparisons. For example, although H3K4me1 is a sign of activate chromatin, the correlation of H3K4me1 inside promoters with gene expression is fairly weak (Fig. 3.4). In fact H3K4me1 signals extend to ±5kb region of the TSS, and positively associated with gene expression [399]. Therefore, a method is required in order to capture distal epigenetic modifications in the functional non-coding regions, primarily enhancer elements. As the region around promoter may contain many enhancers, and in turn one enhancer can distally interact with many promoters [400]. The relationships are not sufficiently captured in one-to-one or many-to-one presentation. Accordingly, graph data structure is an ideal solution for depicting such relationships, as it also enables mapping enhancer-promoter regulations into directed links. Although there are numerous algorithms in picking up useful information from a graph, PageRank seems the most appropriate in solving such problem. At the time of writing, PageRank has been applied in prioritizing candidate genes from protein-protein interaction [345, 401] and microarray [346] data, but was never used in interpreting biological contribution from non-coding elements and epigenetic data.

As the magnitude of epigenetic alterations at promoters and enhancers can be interpreted as vertex weight, and enhancer-promoter interaction confidence can be interpreted as edge weight, a "personalized" PageRank is more suitable in modeling such complex regulatory relationships than conventional PageRank. Nonetheless, since edge weights are solely estimated from P-E distances, it is still arguable that false positive rate of chromatin interactions is uniformly distributed along the genomic region or across samples [402]. Although a universal distance decay function for generating enhancer-promoter probabilities is practicable in my approach, one can expect employing more specific contact profiles (from Hi-C or ChIA-PET) would result in more realistic interaction probabilities.

4.2.4 Benchmarking

To test whether the epigenetically significantly altered candidates from previous analyses make biological sense, I performed a number of tests on the gene ordering from PageRank, which measures the overall alteration from both enhancer and promoter epigenetic modifications. The benchmarking was done using four cancer and three developmental test cases, each compared with corresponding normal tissues or embryonic stem cells. Benchmarking on different principal components, different signature genes sets, a variety of edge weights and node weights generating functions, as well as the permutations of promoter-enhancer network structures, all these analyses supported my hypothesis, that the enhancers are the most important elements in cell differentiation and cancer progression. Besides finding novel signature genes specific to epigenetic regulation in tumorigenesis, the enhancers identified from this approach would also be good candidates for functional studies of enhancers in carcinogenesis.

4.3 Integrative analysis tools

Encapsulating the above quantitative analysis into an R package provides great usability for biologists to test the epigenetic cohorts in their studies. It is more desirable if one can examine the epigenetic alteration within the context of reference database such as HPRD or STRING [403]. The web applications of *crl* allows to trace back these layers of information which are hidden during dimension reduction. With these test cases, I highlighted a few pathways and enhancers of interests, which coincide with previous descriptions of PAX5 and MYC enhancers. Despite the fact that high MYC expression is common in cancers, coding mutation of MYC is not prevalent in cancers. MYC expression is considered to be precisely controlled through epigenetic mechanisms. These tools also generated many enhancer/oncogene/pathway candidates of oncological interests for the biologists to investigate.

4.4 Outlook

In conclusion, the approaches presented in the thesis provide novel solutions to study the genome-wide epigenetic cohort studies in development and cancer, as well as hypothesizing to the discovery of epigenetic hotspots for experiment. Epigenetic mechanisms can be proposed from comprehensive analyses of regulatory elements specific to cancer subtypes, integration of multiple epigenetic datasets, inference from biological networks, etc., which highlight various abnormalities in cancer progression. Given that epigenome alterations are reversible upon treatment with epigenetic drugs, it is advantageous over gene editing in cancer treatment. Therefore, computational aided precise oncology will shed a new light on both cancer diagnosis and epigenetic therapies.

Bibliography

[1] Willbanks A, Leary M, Greenshields M, et al. The evolution of epigenetics: From prokaryotes to humans and its biological consequences. Genetics and Epigenetics. 2016;1(8):25–36.

[2] Jones PL, Veenstra GJC, Wade PA, et al. Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. Nature Genetics. 1998;19(2):187–191.

[3] Ehrlich M. DNA hypomethylation in cancer cells. Epigenomics. 2009;1(2):239–259.

[4] Feinberg AP, Vogelstein B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. Nature. 1983;301(5895):89–92.

[5] Jones PA, Baylin SB. The fundamental role of epigenetic events in cancer; 2002.

[6] Gonzalo S. Epigenetic alterations in aging. Journal of Applied Physiology. 2010;109(2):586–597.

[7] Esteller M. CpG island hypermethylation and tumor suppressor genes: A booming present, a brighter future; 2002.

[8] Shen R, Tao L, Xu Y, et al. Reversibility of aberrant global DNA and estrogen receptor-alpha gene methylation distinguishes colorectal precancer from cancer. International journal of clinical and experimental pathology. 2009;2(1):21–33.

[9] Zemach A, McDaniel IE, Silva P, et al. Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science (New York, NY). 2010;328(5980):916–9.

[10] Feng S, Cokus SJ, Zhang X, et al. Conservation and divergence of methylation patterning in plants and animals. Proceedings of the National Academy of Sciences of the United States of America. 2010;107(19):8689–94.

[11] Kim MY, Zilberman D. DNA methylation as a system of plant genomic immunity; 2014.

[12] Lev Maor G, Yearim A, Ast G. The alternative role of DNA methylation in splicing regulation; 2015.

[13] Maunakea AK, Nagarajan RP, Bilenky M, et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature. 2010;466(7303):253–257.

[14] Wang X, Hu L, Wang X, et al.. DNA Methylation Affects Gene Alternative Splicing in Plants: An Example from Rice; 2016.

[15] ALLFREY VG, FAULKNER R, MIRSKY AE. Acetylation and Methylation of Histones and Their Possible Role in the Regulation of Rna Synthesis. Proceedings of the National Academy of Sciences of the United States of America. 1964;51(1938):786–94.

[16] Wu G, Broniscer A, McEachron TA, et al. Somatic histone H3 alterations in pediatric diffuse intrinsic pontine gliomas and non-brainstem glioblastomas. Nature genetics. 2012;44(3):251–3.

[17] Behjati S, Tarpey PS, Presneau N, et al. Distinct H3F3A and H3F3B driver mutations define chondroblastoma and giant cell tumor of bone. Nature Genetics.

2013;45(12):1479–1482.

[18] Tollervey JR, Lunyak VV. Epigenetics: judge, jury and executioner of stem cell fate. Epigenetics. 2012;7(8):823–40.

[19] Klocker H, Eder IE, Comuzzi B, et al. Androgen Receptor Function in Prostate Cancer Progression. In: Prostate Cancer. Springer, Boston, MA; 2007. p. 87–105. Available from: `http://dx.doi.org/10.1007/978-1-59745-224-3{_}6`.

[20] Hu J, Zhang Y, Zhao L, et al. Chromosomal Loop Domains Direct the Recombination of Antigen Receptor Genes. Cell. 2015;163(4):947–959.

[21] Vu TH, Nguyen AH, Hoffman AR. Loss of IGF2 imprinting is associated with abrogation of long-range intrachromosomal interactions in human cancer cells. Human Molecular Genetics. 2010;.

[22] Hnisz D, Weintrau AS, Day DS, et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science. 2016;351(6280):1454–1458.

[23] Rickman DS, Soong TD, Moss B, et al. Oncogene-mediated alterations in chromatin conformation. Proceedings of the National Academy of Sciences. 2012;109(23):9083–9088.

[24] Mourad R, Hsu PY, Juan L, et al. Estrogen Induces Global Reorganization of Chromatin Structure in Human Breast Cancer Cells. PLoS ONE. 2014;.

[25] Schwarzer W, Abdennur N, Goloborodko A, et al. Two independent modes of chromatin organization revealed by cohesin removal. Nature. 2017;551(7678):51–56.

[26] Chalkley GE, Verrijzer CP. DNA binding site selection by RNA polymerase II TAFs: a TAF(II)250-TAF(II)150 complex recognizes the initiator. The EMBO journal. 1999;18(17):4835–45.

[27] Smale ST. Transcription initiation from TATA-less promoters within eukaryotic protein-coding genes; 1997.

[28] Koudritsky M, Domany E. Positional distribution of human transcription factor binding sites. Nucleic Acids Research. 2008;.

[29] Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proceedings of the National Academy of Sciences. 2006;103(5):1412–1417.

[30] Mirabella AC, Foster BM, Bartke T. Chromatin deregulation in disease; 2016.

[31] Deaton AM, Bird A. CpG islands and the regulation of transcription. Genes & Development. 2011;25(10):1010–1022.

[32] Vermeulen M, Mulder KW, Denissov S, et al. Selective Anchoring of TFIID to Nucleosomes by Trimethylation of Histone H3 Lysine 4. Cell. 2007;131(1):58–69.

[33] Karmodiya K, Krebs AR, Oulad-Abdelghani M, et al. H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells. BMC genomics. 2012;13:424.

[34] Wysocka J, Swigut T, Xiao H, et al. A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. Nature. 2006;442(7098):86–90.

[35] Sims RJ, Millhouse S, Chen CF, et al. Recognition of Trimethylated Histone H3 Lysine 4 Facilitates the Recruitment of Transcription Postinitiation Factors and Pre-mRNA Splicing. Molecular Cell. 2007;28(4):665–676.

[36] Chen K, Chen Z, Wu D, et al. Broad H3K4me3 is associated with increased transcription elongation and enhancer activity at tumor-suppressor genes. Nature Genetics. 2015;47(10):1149–1157.

[37] Wang Z, Zang C, Rosenfeld JA, et al. Combinatorial patterns of histone acetylations

and methylations in the human genome. Nature Genetics. 2008;40(7):897–903.

[38] Azuara V, Perry P, Sauer S, et al. Chromatin signatures of pluripotent cell lines. Nature Cell Biology. 2006;8(5):532–538.

[39] Bernstein BE, Mikkelsen TS, Xie X, et al. A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. Cell. 2006;125(2):315–326.

[40] Mikkelsen TS, Ku M, Jaffe DB, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature. 2007;448(7153):553–560.

[41] Blackwood EM, Kadonaga JT. Going the distance: A current view of enhancer action; 1998.

[42] Pennacchio LA, Bickmore W, Dean A, et al.. Enhancers: Five essential questions; 2013.

[43] Maston GA, Evans SK, Green MR. Transcriptional Regulatory Elements in the Human Genome. Annual Review of Genomics and Human Genetics. 2006;7(1):29–59.

[44] Kowalczyk MS, Hughes JR, Garrick D, et al. Intragenic Enhancers Act as Alternative Promoters. Molecular Cell. 2012;.

[45] Fuda NJ, Ardehali MB, Lis JT. Defining mechanisms that regulate RNA polymerase II transcription in vivo; 2009.

[46] Koch F, Fenouil R, Gut M, et al. Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. Nature Structural and Molecular Biology. 2011;18(8):956–963.

[47] Schmidt D, Wilson MD, Ballester B, et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science (New York, NY). 2010;328(5981):1036–40.

[48] May D, Blow MJ, Kaplan T, et al. Large-scale discovery of enhancers from human heart tissue. Nature Genetics. 2012;.

[49] Creyghton MP, Cheng AW, Welstead GG, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences. 2010;107(50):21931–21936.

[50] Heintzman ND, Stuart RK, Hon G, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature Genetics. 2007;.

[51] Calo E, Wysocka J. Modification of Enhancer Chromatin: What, How, and Why?; 2013.

[52] Dorschner MO, Hawrylycz M, Humbert R, et al. High-throughput localization of functional elements by quantitative chromatin profiling. Nature Methods. 2004;.

[53] Visel A, Blow MJ, Li Z, et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature. 2009;.

[54] Stasevich TJ, Hayashi-Takanaka Y, Sato Y, et al. Regulation of RNA polymerase II activation by histone acetylation in single living cells. Nature. 2014;516(7530):272–275.

[55] Charlet J, Duymich CE, Lay FD, et al. Bivalent Regions of Cytosine Methylation and H3K27 Acetylation Suggest an Active Role for DNA Methylation at Enhancers. Molecular Cell. 2016;62(3):422–431.

[56] Li W, Notani D, Rosenfeld MG. Enhancers as non-coding RNA transcription units: Recent insights and future perspectives; 2016.

[57] Hah N, Benner C, Chong LW, et al. Inflammation-sensitive super enhancers form domains of coordinately regulated enhancer RNAs. Proceedings of the National Academy of Sciences. 2015;.

[58] Hnisz D, Abraham BJ, Lee TI, et al. Super-enhancers in the control of cell identity and disease. Cell. 2013;155(4):934–47.

[59] Whyte WA, Orlando DA, Hnisz D, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell. 2013;153(2):307–319.

[60] Ong CT, Corces VG. CTCF: An architectural protein bridging genome topology and function; 2014.

[61] Kim TH, Abdullaev ZK, Smith AD, et al. Analysis of the Vertebrate Insulator Protein CTCF-Binding Sites in the Human Genome. Cell. 2007;.

[62] Huang S, Li X, Yusufzai TM, et al. USF1 Recruits Histone Modification Complexes and Is Critical for Maintenance of a Chromatin Barrier. Molecular and Cellular Biology. 2007;.

[63] Wood AM, Van Bortle K, Ramos E, et al. Regulation of Chromatin Organization and Inducible Gene Expression by a Drosophila Insulator. Molecular Cell. 2011;.

[64] Liu Z, Scannell DR, Eisen MB, et al. Control of embryonic stem cell lineage commitment by core promoter factor, TAF3. Cell. 2011;146(5):720–31.

[65] Casey SC, Vaccari M, Al-Mulla F, et al.. The effect of environmental chemicals on the tumor microenvironment; 2015.

[66] Chai H, Brown RE. Field effect in cancer-an update. Annals of clinical and laboratory science. 2009;39(4):331–7.

[67] Ge R, Wang W, Kramer PM, et al. Wy-14,643-induced hypomethylation of the c-myc gene in mouse liver. Toxicological Sciences. 2001;62(1).

[68] Tao L, Yang S, Xie M, et al. Effect of trichloroethylene and its metabolites, dichloroacetic acid and trichloroacetic acid, on the methylation and expression of c-Jun and c-Myc protooncogenes in mouse liver: prevention by methionine. Toxicological sciences : an official journal of the Society of Toxicology. 2000;54(2):399–407.

[69] Sen P, Costa M. Induction of Chromosomal Damage in Chinese Hamster Ovary Cells by Soluble and Particulate Nickel Compounds: Preferential Fragmentation of the Heterochromatic Long Arm of the X-Chromosome by Carcinogenic Crystalline NiS Particles. Cancer Research. 1985;45(5):2320–2325.

[70] Lee YW, Klein CB, Kargacin B, et al. Carcinogenic nickel silences gene expression by chromatin condensation and DNA methylation: a new model for epigenetic carcinogens. Molecular and Cellular Biology. 1995;15(5):2547–2557.

[71] Zhao CQ, Young MR, Diwan Ba, et al. Association of arsenic-induced malignant transformation with DNA hypomethylation and aberrant gene expression. Proceedings of the National Academy of Sciences of the United States of America. 1997;94(20):10907–10912.

[72] Senut MC, Cingolani P, Sen A, et al. Epigenetics of early-life lead exposure and effects on brain development. Epigenomics. 2012;4(6):665–674.

[73] Devóz PP, Gomes WR, De Araújo ML, et al. Lead (Pb) exposure induces disturbances in epigenetic status in workers exposed to this metal. Journal of Toxicology and Environmental Health - Part A: Current Issues. 2017;80(19-21):1098–1105.

[74] Kondo K, Takahashi Y, Hirose Y, et al. The reduced expression and aberrant methylation of p16INK4a in chromate workers with lung cancer. Lung Cancer. 2006;53(3):295–302.

[75] Takahashi Y, Kondo K, Hirose T, et al. Microsatellite instability and protein expression of the DNA mismatch repair gene, hMLH1, of lung cancer in chromate-exposed workers. Molecular Carcinogenesis. 2005;42(3):150–158.

[76] Takiguchi M, Achanzar WE, Qu W, et al. Effects of cadmium on DNA-(Cytosine-5) methyltransferase activity and DNA methylation status during cadmium-induced

cellular transformation. Experimental Cell Research. 2003;286(2):355–365.

[77] Huang D, Zhang Y, Qi Y, et al. Global DNA hypomethylation, rather than reactive oxygen species (ROS), a potential facilitator of cadmium-stimulated K562 cell proliferation. Toxicology Letters. 2008;179(1):43–47.

[78] Karaczyn AA, Golebiowski F, Kasprzak KS. Truncation, deamidation, and oxidation of histone H2B in cells cultured with nickel(II). Chemical Research in Toxicology. 2005;18(12):1934–1942.

[79] Broday L, Peng W, Kuo MH, et al. Nickel compounds are novel inhibitors of histone H4 acetylation. Cancer Research. 2000;60(2):238–241.

[80] Chen H, Ke Q, Kluz T, et al. Nickel Ions Increase Histone H3 Lysine 9 Dimethylation and Induce Transgene Silencing. Molecular and Cellular Biology. 2006;26(10):3728–3737.

[81] Ke Q, Davidson T, Chen H, et al. Alterations of histone modifications and transgene silencing by nickel chloride. Carcinogenesis. 2006;27(7):1481–1488.

[82] Golebiowski F, Kasprzak KS. Inhibition of core histones acetylation by carcinogenic nickel(II). Molecular and Cellular Biochemistry. 2005;279(1-2):133–139.

[83] Zhou X, Sun H, Ellen TP, et al. Arsenite alters global histone H3 methylation. Carcinogenesis. 2008;29(9):1831–1836.

[84] Zhong CX, Mass MJ. Both hypomethylation and hypermethylation of DNA associated with arsenite exposure in cultures of human cells identified by methylation-sensitive arbitrarily-primed PCR. Toxicology Letters. 2001;122(3):223–234.

[85] Chai CY, Huang YC, Hung WC, et al. Arsenic salts induced autophagic cell death and hypermethylation of DAPK promoter in SV-40 immortalized human uroepithelial cells. Toxicology Letters. 2007;173(1):48–56.

[86] Mass MJ, Wang L. Arsenic alters cytosine methylation patterns of the promoter of the tumor suppressor gene p53 in human lung cells: A model for a mechanism of carcinogenesis. Mutation Research - Reviews in Mutation Research. 1997;386(3):263–277.

[87] Chanda S, Dasgupta UB, GuhaMazumder D, et al. DNA hypermethylation of promoter of gene p53 and p16 in arsenic-exposed people with and without malignancy. Toxicological Sciences. 2006;89(2):431–437.

[88] Schnekenburger M, Talaska G, Puga A. Chromium Cross-Links Histone Deacetylase 1-DNA Methyltransferase 1 Complexes to Chromatin, Inhibiting Histone-Remodeling Marks Critical for Transcriptional Activation. Molecular and Cellular Biology. 2007;27(20):7089–7101.

[89] Sun H, Zhou X, Chen H, et al. Modulation of histone methylation and MLH1 gene silencing by hexavalent chromium. Toxicology and Applied Pharmacology. 2009;237(3):258–266.

[90] Natelson EA. Benzene-induced acute myeloid leukemia: a clinician's perspective. American journal of hematology. 2007;82(9):826–30.

[91] McMichael AJ, Spirtas R, Kupper LL, et al. Solvent exposure and leukemia among rubber workers: an epidemiologic study. Journal of occupational medicine : official publication of the Industrial Medical Association. 1975;17(4):234–9.

[92] Bollati V, Baccarelli A, Hou L, et al. Changes in DNA methylation patterns in subjects exposed to low-dose benzene. Cancer Research. 2007;67(3):876–880.

[93] Xie Y, Zhou JJ, Zhao Y, et al. H. pylori modifies methylation of global genomic DNA and the gastrin gene promoter in gastric mucosal cells and gastric cancer cells. Microbial Pathogenesis. 2017;108:129–136.

[94] Chan AOO, Lam SK, Wong BCY, et al. Promoter methylation of E-cadherin gene

in gastric mucosa associated with Helicobacter pylori infection and in gastric cancer. Gut. 2003;52(4):502–6.

[95] Sugiyama A, Maruta F, Ikeno T, et al. Helicobacter pylori infection enhances N-methyl-N-nitrosourea-induced stomach carcinogenesis in the Mongolian gerbil. Cancer Research. 1998;58(10):2067–2069.

[96] Paschos K, Smith P, Anderton E, et al. Epstein-Barr virus latency in B cells leads to epigenetic repression and CpG methylation of the tumour suppressor gene Bim. PLoS Pathogens. 2009;5(6).

[97] Zhang T, Ma J, Nie K, et al. Hypermethylation of the tumor suppressor gene PRDM1/Blimp-1 supports a pathogenetic role in EBV-positive Burkitt lymphom. Blood Cancer Journal. 2014;4(11).

[98] Lambert MP, Paliwal A, Vaissière T, et al. Aberrant DNA methylation distinguishes hepatocellular carcinoma associated with HBV and HCV infection and alcohol intake. Journal of Hepatology. 2011;54(4):705–715.

[99] Sanchez-Cespedes M, Esteller M, Wu L, et al. Gene promoter hypermethylation in tumors and serum of head and neck cancer patients. Cancer research. 2000;60(4):892–895.

[100] Kato K, Hara A, Kuno T, et al. Aberrant promoter hypermethylation of p16 and MGMT genes in oral squamous cell carcinomas and the surrounding normal mucosa. Journal of Cancer Research and Clinical Oncology. 2006;132(11):735–743.

[101] Shaw RJ, Liloglou T, Rogers SN, et al. Promoter methylation of P16, RARβ, E-cadherin, cyclin A1 and cytoglobin in oral cancer: Quantitative evaluation using pyrosequencing. British Journal of Cancer. 2006;94(4):561–568.

[102] Hattori N, Ushijima T. Epigenetic impact of infection on carcinogenesis: mechanisms and applications. Genome medicine. 2016;8(1):10.

[103] Keleher MR, Zaidi R, Shah S, et al. Maternal high-fat diet associated with altered gene expression, DNA methylation, and obesity risk in mouse offspring. PLoS ONE. 2018;13(2).

[104] Dolinoy DC, Huang D, Jirtle RL. Maternal nutrient supplementation counteracts bisphenol A-induced DNA hypomethylation in early development. Proceedings of the National Academy of Sciences of the United States of America. 2007;104(32):13056–61.

[105] Dolinoy DC, Weidman JR, Waterland RA, et al. Maternal genistein alters coat color and protects Avy mouse offspring from obesity by modifying the fetal epigenome. Environmental Health Perspectives. 2006;114(4):567–572.

[106] Gkountela S, Zhang KX, Shafiq TA, et al. DNA demethylation dynamics in the human prenatal germline. Cell. 2015;161(6):1425–1436.

[107] Morgan HD, Sutherland HGE, Martin DIK, et al. Epigenetic inheritance at the agouti locus in the mouse. Nature Genetics. 1999;23(3):314–318.

[108] Smith ZD, Chan MM, Humm KC, et al. DNA methylation dynamics of the human preimplantation embryo. Nature. 2014;.

[109] Cedar H, Bergman Y. Programming of DNA Methylation Patterns. Annual Review of Biochemistry. 2012;81(1):97–117.

[110] Weber RG, Hoischen a, Ehrler M, et al. Frequent loss of chromosome 9, homozygous CDKN2A/p14(ARF)/CDKN2B deletion and low TSC1 mRNA expression in pleomorphic xanthoastrocytomas. Oncogene. 2007;26(7):1088–97.

[111] Meissner A, Mikkelsen TS, Gu H, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature. 2008;454(7205):766–770.

[112] Yasukochi Y, Maruyama O, Mahajan MC, et al. X chromosome-wide analyses of ge-

nomic DNA methylation states and gene expression in male and female neutrophils. Proceedings of the National Academy of Sciences. 2010;107(8):3704–3709.

[113] Marahrens Y. X-inactivation by chromosomal pairing events; 1999.

[114] Krusche CA, Vloet AJ, Classen-Linke I, et al. Class I histone deacetylase expression in the human cyclic endometrium and endometrial adenocarcinomas. Human Reproduction. 2007;22(11):2956–2966.

[115] Guo JZ, Gorski J. Estrogen effects on modifications of chromatin proteins in the rat uterus. Journal of Steroid Biochemistry. 1989;32(1 PART 1):13–20.

[116] Heryanto B, Lipson KE, Rogers PAW. Effect of angiogenesis inhibitors on oestrogen-mediated endometrial endothelial cell proliferation in the ovariectomized mouse; 2003.

[117] Munro SK, Farquhar CM, Mitchell MD, et al.. Epigenetic regulation of endometrium during the menstrual cycle; 2010.

[118] Langton AK, Herrick SE, Headon DJ. An extended epidermal response heals cutaneous wounds in the absence of a hair follicle stem cell contribution. Journal of Investigative Dermatology. 2008;128(5):1311–1318.

[119] Yan M, Zhang Z, Brady JR, et al. Identification of a novel death domain-containing adaptor molecule for ectodysplasin-A receptor that is mutated in crinkled mice. Current Biology. 2002;12(5):409–413.

[120] Wullner U, Kaut O, DeBoni L, et al. DNA methylation in Parkinson's disease. Journal of neurochemistry. 2016;139 Suppl:108–120.

[121] Irier HA, Jin P. Dynamics of DNA Methylation in Aging and Alzheimer's Disease. DNA and Cell Biology. 2012;31(S1):S–42–S–48.

[122] De Souza RAG, Islam SA, McEwen LM, et al. DNA methylation profiling in human Huntington's disease brain. Human Molecular Genetics. 2016;25(10):2013–2030.

[123] Vogelstein B, Papadopoulos N, Velculescu VE, et al.. Cancer genome landscapes; 2013.

[124] Berman BP, Weisenberger DJ, Aman JF, et al. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear laminag-associated domains. Nature Genetics. 2012;44(1):40–46.

[125] Eden A, Gaudet F, Waghmare A, et al.. Chromosomal instability and tumors promoted by DNA hypomethylation; 2003.

[126] International Human Genome Sequencing Consortium, Human I, Sequencing G. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860–921.

[127] Wolff EM, Byun HM, Han HF, et al. Hypomethylation of a LINE-1 promoter activates an alternate transcript of the MET oncogene in bladders with cancer. PLoS Genetics. 2010;6(4).

[128] Hur K, Cejas P, Feliu J, et al. Hypomethylation of long interspersed nuclear element-1 (LINE-1) leads to activation of protooncogenes in human colorectal cancer metastasis. Gut. 2014;63(4):635–646.

[129] Antelo M, Balaguer F, Shia J, et al. A High Degree of LINE-1 Hypomethylation Is a Unique Feature of Early-Onset Colorectal Cancer. PLoS ONE. 2012;7(9).

[130] Shigaki H, Baba Y, Watanabe M, et al. LINE-1 hypomethylation in gastric cancer, detected by bisulfite pyrosequencing, is associated with poor prognosis. Gastric Cancer. 2013;16(4):480–487.

[131] Van Hoesel AQ, Van De Velde CJH, Kuppen PJK, et al. Hypomethylation of LINE-1 in primary tumor has poor prognosis in young breast cancer patients: A retrospective cohort study. Breast Cancer Research and Treatment. 2012;134(3):1103–1114.

[132] Aoki Y, Nojima M, Suzuki H, et al. Genomic vulnerability to LINE-1 hypomethylation is a potential determinant of the clinicogenetic features of multiple myeloma. Genome Medicine. 2012;4(12).

[133] Zhu C, Utsunomiya T, Ikemoto T, et al. Hypomethylation of Long Interspersed Nuclear Element-1 (LINE-1) is Associated with Poor Prognosis via Activation of c-MET in Hepatocellular Carcinoma. Annals of Surgical Oncology. 2014;21(4):729–735.

[134] Kreimer U, Schulz WA, Koch A, et al. HERV-K and LINE-1 DNA Methylation and Reexpression in Urothelial Carcinoma. Frontiers in Oncology. 2013;3.

[135] Howard G, Eiges R, Gaudet F, et al. Activation and transposition of endogenous retroviral elements in hypomethylation induced tumors in mice. Oncogene. 2008;27(3):404–408.

[136] Sunami E, de Maat M, Vu A, et al. LINE-1 hypomethylation during primary colon cancer progression. PLoS ONE. 2011;6(4).

[137] Schichman SA, Caligiuri MA, Strout MP, et al. ALL-1 Tandem Duplication in Acute Myeloid Leukemia with a Normal Karyotype Involves Homologous Recombination between Alu Elements1. Cancer Research. 1994;54(16):4277–4280.

[138] O'Neil J, Tchinda J, Gutierrez A, et al. Alu elements mediate <i>MYB</i> gene tandem duplication in human T-ALL. The Journal of Experimental Medicine. 2007;204(13):3059–3066.

[139] Jeffs AR, Benjes SM, Smith TL, et al. The BCR gene recombines preferentially with Alu elements in complex BCR-ABL translocations of chronic myeloid leukaemia. Human Molecular Genetics. 1998;7(5):767–776.

[140] Morse B, Rotherg PG, South VJ, et al. Insertional mutagenesis of the myc locus by a LINE-1 sequence in a human breast carcinoma. Nature. 1988;333(6168):87–90.

[141] Teugels E, De Brakeleer S, Goelen G, et al. De novo Alu element insertions targeted to a sequence common to the BRCA1 and BRCA2 genes. Human mutation. 2005;26(3):284.

[142] Oliveira C, Senz J, Kaurah P, et al. Germline CDH1 deletions in hereditary diffuse gastric cancer families. Human Molecular Genetics. 2009;18(9):1545–1555.

[143] Tang Z, Steranka JP, Ma S, et al. Human transposon insertion profiling: Analysis, visualization and identification of somatic LINE-1 insertions in ovarian cancer. Proceedings of the National Academy of Sciences. 2017;114(5):E733–E740.

[144] Cui H, Cruz-Correa M, Giardiello FM, et al. Loss of IGF2 imprinting: a potential marker of colorectal cancer risk. Science (New York, NY). 2003;299(5613):1753–5.

[145] Roglerss CE, Yangs D, Rossettin L, et al. Altered Body Composition and Increased Frequency of Diverse Malignancies in Insulin-like Growth Factor41 Transgenic Mice*. THE JOURNAL OF B˜OLOGICAL CHEMISTRY. 1994;269(19):13779–13784.

[146] Shi W, Dirim F, Wolf E, et al. Methylation Reprogramming and Chromosomal Aneuploidy in In Vivo Fertilized and Cloned Rabbit Preimplantation Embryos1. Biology of Reproduction. 2004;71(1):340–347.

[147] Esteller M. ABERRANT DNA METHYLATION AS A CANCER-INDUCING MECHANISM. Annual Review of Pharmacology and Toxicology. 2005;45(1):629–656.

[148] Clouaire T, Stancheva I. Methyl-CpG binding proteins: Specialized transcriptional repressors or structural components of chromatin?; 2008.

[149] Singh S, Li SSL. Epigenetic effects of environmental chemicals bisphenol A and phthalates; 2012.

[150] Mizuno SI, Chijiwa T, Okamura T, et al. Expression of DNA methyltransferases DNMT1, 3A, and 3B in normal hematopoiesis and in acute and chronic myelogenous leukemia. Blood. 2001;.

[151] Lu R, Wang P, Parton T, et al. Epigenetic Perturbations by Arg882-Mutated DNMT3A Potentiate Aberrant Stem Cell Gene-Expression Program and Acute Leukemia Development. Cancer Cell. 2016;.

[152] Agoston AT, Argani P, Yegnasubramanian S, et al. Increased protein stability causes DNA methyltransferase 1 dysregulation in breast cancer. The Journal of biological chemistry. 2005;.

[153] Butcher DT, Rodenhiser DI. Epigenetic inactivation of BRCA1 is associated with aberrant expression of CTCF and DNA methyltransferase (DNMT3B) in some sporadic breast tumours. European Journal of Cancer. 2007;.

[154] Ibrahim AEK, Arends MJ, Silva AL, et al. Sequential DNA methylation changes are associated with DNMT3B overexpression in colorectal neoplastic progression. Gut. 2011;.

[155] Nosho K, Shima K, Irahara N, et al. DNMT3B expression might contribute to CpG island methylator phenotype in colorectal cancer. Clinical Cancer Research. 2009;.

[156] Saito Y, Kanai Y, Nakagawa T, et al. Increased protein expression of DNA methyltransferase (DNMT) 1 is significantly correlated with the malignant potential and poor prognosis of human hepatocellular carcinomas. International Journal of Cancer. 2003;.

[157] Zhao Z, Wu Q, Cheng J, et al. Depletion of DNMT3A suppressed cell proliferation and restored PTEN in hepatocellular carcinoma cell. Journal of Biomedicine and Biotechnology. 2010;.

[158] Peng DF, Kanai Y, Sawada M, et al. DNA methylation of multiple tumor-related genes in association with overexpression of DNA methyltransferase 1 (DNMT1) during multistage carcinogenesis of the pancreas. Carcinogenesis. 2006;.

[159] Kobayashi Y, Absher DM, Gulzar ZG, et al. DNA methylation profiling reveals novel biomarkers and important roles for DNA methyltransferases in prostate cancer. Genome Research. 2011;.

[160] Zhao SL, Zhu ST, Hao X, et al. Effects of DNA methyltransferase 1 inhibition on esophageal squamous cell carcinoma. Diseases of the esophagus : official journal of the International Society for Diseases of the Esophagus / ISDE. 2011;.

[161] Pathania R, Ramachandran S, Mariappan G, et al. Combined inhibition of DNMT and HDAC blocks the tumorigenicity of cancer stem-like cells and attenuates mammary tumor growth. Cancer Research. 2016;76(11):3224–3235.

[162] Soengas MS, Capodieci P, Polsky D, et al. Inactivation of the apoptosis effector Apaf-1 in malignant melanoma. Nature. 2001;409(6817):207–211.

[163] Wargo JA, Robbins PF, Li Y, et al. Recognition of NY-ESO-1+ tumor cells by engineered lymphocytes is enhanced by improved vector design and epigenetic modulation of tumor antigen expression. Cancer Immunology, Immunotherapy. 2009;58(3):383–394.

[164] Li Y, Seto E. HDACs and HDAC Inhibitors in Cancer Development and Therapy. Cold Spring Harbor perspectives in medicine. 2016;6(10).

[165] Ellis L, Hammers H, Pili R. Targeting tumor angiogenesis with histone deacetylase inhibitors; 2009.

[166] Marquard L, Poulsen CB, Gjerdrum LM, et al. Histone deacetylase 1, 2, 6 and acetylated histone H4 in B- and T-cell lymphomas. Histopathology. 2009;.

[167] Adams H, Fritzsche FR, Dirnhofer S, et al. Class I histone deacetylases 1, 2 and 3 are

highly expressed in classical Hodgkin's lymphoma. Expert opinion on therapeutic targets. 2010;.

[168] Svechnikova I, Almqvist PM, Ekström TJ. HDAC inhibitors effectively induce cell type-specific differentiation in human glioblastoma cell lines of different origin. International Journal of Oncology. 2008;32(4):821–827.

[169] Adamopoulou E, Naumann U. HDAC inhibitors and their potential applications to glioblastoma therapy. OncoImmunology. 2013;2(8).

[170] Wallace IV GC, Haar CP, Vandergrift WA, et al. Multi-targeted DATS prevents tumor progression and promotes apoptosis in ectopic glioblastoma xenografts in SCID mice via HDAC inhibition. Journal of Neuro-Oncology. 2013;114(1):43–50.

[171] Holm K, Grabau D, Lövgren K, et al. Global H3K27 trimethylation and EZH2 abundance in breast tumor subtypes. Molecular Oncology. 2012;6(5):494–506.

[172] Yamagishi M, Uchimaru K. Targeting EZH2 in cancer therapy; 2017.

[173] Tjian R. The binding site on SV40 DNA for a T antigen-related protein. Cell. 1978;.

[174] Roeder RG. The role of general initiation factors in transcription by RNA polymerase II. Trends in biochemical sciences. 1996;21(9):327–35.

[175] Nikolov DB, Burley SK. RNA polymerase II transcription initiation: a structural view. Proceedings of the National Academy of Sciences of the United States of America. 1997;94(1):15–22.

[176] Yin Y, Morgunova E, Jolma A, et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science. 2017;356(6337).

[177] Venkatesh S, Workman JL. Histone exchange, chromatin structure and the regulation of transcription; 2015.

[178] Xie H, Hoffmann HM, Iyer AK, et al. Chromatin status and transcription factor binding to gonadotropin promoters in gonadotrope cell lines. Reproductive biology and endocrinology : RB&E. 2017;15(1):86.

[179] Ng HH, Surani MA. The transcriptional and signalling networks of pluripotency; 2011.

[180] Orkin SH, Hochedlinger K. Chromatin connections to pluripotency and cellular reprogramming; 2011.

[181] Young RA. Control of the embryonic stem cell state; 2011.

[182] Whyte WA, Orlando DA, Hnisz D, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell. 2013;.

[183] Boyer LA, Lee TI, Cole MF, et al. Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells. Cell. 2005;.

[184] Loh YH, Wu Q, Chew JL, et al. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nature Genetics. 2006;.

[185] Dohm JC, Lottaz C, Borodina T, et al. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Research. 2008;.

[186] Kuan PF, Chung D, Pan G, et al. A statistical framework for the analysis of ChIP-Seq data. Journal of the American Statistical Association. 2011;.

[187] Vega VB, Cheung E, Palanisamy N, et al. Inherent signals in sequencing-based Chromatin-ImmunoPrecipitation control libraries. PLoS ONE. 2009;.

[188] ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science (New York, NY). 2004;306(5696):636–40.

[189] Lun ATL, Smyth GK. Csaw: A Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. Nucleic Acids Research. 2015;44(5):1–10.

[190] Rye MB, Sætrom P, Drabløs F. A manually curated ChIP-seq benchmark demonstrates room for improvement in current peak-finder programs. Nucleic Acids Re-

search. 2011;39(4).

[191] Zhang Y, Liu T, Meyer CA, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biology. 2008;9(9).

[192] Zang C, Schones DE, Zeng C, et al. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics. 2009;25(15):1952–1958.

[193] Chen L, Wang C, Qin ZS, et al. A novel statistical method for quantitative comparison of multiple ChIP-seq datasets. Bioinformatics. 2015;.

[194] Xu H, Wei CL, Lin F, et al. An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. Bioinformatics. 2008;.

[195] Nair NU, Das Sahu A, Bucher P, et al. Chipnorm: A statistical method for normalizing and identifying differential regions in histone modification chip-seq libraries. PLoS ONE. 2012;.

[196] Liang K, Keles S. Detecting differential binding of transcription factors with ChIP-seq. Bioinformatics (Oxford, England). 2012;.

[197] Stark R, Brown G. DiffBind: differential binding analysis of ChIP-Seq peak data; 2011. Available from: `http://bioconductor.org/packages/release/bioc/vignettes/DiffBind/inst/doc/DiffBind.pdf`.

[198] Shao Z, Zhang Y, Yuan GC, et al. MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. Genome Biology. 2012;.

[199] Song Q, Smith AD. Identifying dispersed epigenomic domains from ChIP-Seq data. Bioinformatics. 2011;.

[200] Steinhauser S, Kurzawa N, Eils R, et al. A comprehensive comparison of tools for differential ChIP-seq analysis. Briefings in Bioinformatics. 2016; p. bbv110.

[201] Ross-Innes CS, Stark R, Teschendorff AE, et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature. 2012;481(7381):389–393.

[202] Du P, Zhang X, Huang CC, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010;11.

[203] Bhasin JM, Hu B, Ting AH. MethylAction: Detecting differentially methylated regions that distinguish biological subtypes. Nucleic Acids Research. 2016;.

[204] Assenov Y, Müller F, Lutsik P, et al. Comprehensive analysis of DNA methylation data with RnBeads. Nature Methods. 2014;.

[205] Hebestreit K, Dugas M, Klein HU. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. In: Bioinformatics; 2013. p. 1647–53.

[206] Stalder J, Larsen A, Engel JD, et al. Tissue-specific DNA cleavages in the globin chromatin domain introduced by DNAase I. Cell. 1980;20(2):451–460.

[207] Gross D. Nuclease Hypersensitive Sites In Chromatin. Annual Review of Biochemistry. 1988;.

[208] Boyle AP, Davis S, Shulha HP, et al. High-Resolution Mapping and Characterization of Open Chromatin across the Genome. Cell. 2008;132(2):311–322.

[209] Buenrostro JD, Giresi PG, Zaba LC, et al. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature Methods. 2013;.

[210] Giresi PG, Kim J, McDaniell RM, et al. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Research. 2007;.

[211] Hesselberth JR, Chen X, Zhang Z, et al. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nature Methods. 2009;.

[212] Giresi PG, Lieb JD. Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements). Methods (San Diego, Calif). 2009;48(3):233–9.

[213] Kumar V, Muratani M, Rayan NA, et al. Uniform, optimal signal processing of mapped deep-sequencing data. Nature Biotechnology. 2013;31(7):615–622.

[214] Yan H, Tian S, Slager SL, et al. Genome-Wide Epigenetic Studies in Human Disease: A Primer on -Omic Technologies. American journal of epidemiology. 2016;183(2):96–109.

[215] Daugherty AC, Yeo RW, Buenrostro JD, et al. Chromatin accessibility dynamics reveal novel functional enhancers in C. elegans. Genome Research. 2017;27(12):2096–2107.

[216] Quillien A, Abdalla M, Yu J, et al. Robust Identification of Developmentally Active Endothelial Enhancers in Zebrafish Using FANS-Assisted ATAC-Seq. Cell Reports. 2017;20(3):709–720.

[217] Bowman SK. Discovering enhancers by mapping chromatin features in primary tissue; 2015.

[218] Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nature methods. 2012;9(3):215–6.

[219] Mammana A, Chung HR. Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome. Genome Biology. 2015;.

[220] Kundaje A, Meuleman W, Ernst J, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;.

[221] Thurner M, van de Bunt M, Torres JM, et al. Integration of human pancreatic islet genomic data refines regulatory mechanisms at type 2 diabetes susceptibility loci. eLife. 2018;.

[222] Taberlay PC, Statham AL, Kelly TK, et al. Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer. Genome Research. 2014;.

[223] Hall AW, Battenhouse AM, Shivram H, et al. Bivalent chromatin domains in glioblastoma reveal a subtype-specific signature of glioma stem cells. Cancer Research. 2018;.

[224] Lin IH, Chen DT, Chang YF, et al. Hierarchical clustering of breast cancer methylomes revealed differentially methylated and expressed breast cancer genes. PLoS ONE. 2015;.

[225] Dekker J, Rippe K, Dekker M, et al. Capturing chromosome conformation. Science. 2002;.

[226] Simonis M, Klous P, Splinter E, et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nature Genetics. 2006;38(11):1348–1354.

[227] Dostie J, Richmond TA, Arnaout RA, et al. Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. Genome Research. 2006;16(10):1299–1309.

[228] Fullwood MJ, Liu MH, Pan YF, et al. An oestrogen-receptor-α-bound human chromatin interactome. Nature. 2009;462(7269):58–64.

[229] Belton JM, McCord RP, Gibcus JH, et al. Hi-C: A comprehensive technique to capture the conformation of genomes. Methods. 2012;58(3):268–276.

[230] Lieberman-Aiden E, Van Berkum NL, Williams L, et al. Comprehensive mapping

of long-range interactions reveals folding principles of the human genome. Science. 2009;326(5950):289–293.

[231] Dixon JR, Selvaraj S, Yue F, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;.

[232] Rao SSP, Huntley MH, Durand NC, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159(7):1665–1680.

[233] Bourgo RJ, Singhal H, Greene GL. Capture of associated targets on chromatin links long-distance chromatin looping to transcriptional coordination. Nature Communications. 2016;.

[234] Hughes JR, Roberts N, McGowan S, et al. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. Nature genetics. 2014;46(2):205–12.

[235] Andersson R, Gebhard C, Miguel-Escalada I, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;.

[236] He B, Chen C, Teng L, et al. Global view of enhancer-promoter interactome in human cells. Proceedings of the National Academy of Sciences. 2014;111(21):E2191–E2199.

[237] Nora EP, Lajoie BR, Schulz EG, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature. 2012;485(7398):381–385.

[238] Kikuta H, Laplante M, Navratilova P, et al. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. Genome Research. 2007;17(5):545–555.

[239] Larkin DM, Pape G, Donthu R, et al. Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. Genome Research. 2009;19(5):770–777.

[240] Thurman RE, Rynes E, Humbert R, et al. The accessible chromatin landscape of the human genome. Nature. 2012;489(7414):75–82.

[241] Ernst J, Kheradpour P, Mikkelsen TS, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473(7345):43–49.

[242] Roy S, Siahpirani AF, Chasman D, et al. A predictive modeling approach for cell line-specific long-range regulatory interactions. Nucleic Acids Research. 2015;.

[243] Cao Q, Anyansi C, Hu X, et al. Reconstruction of enhancer–target networks in 935 samples of human primary cells, tissues and cell lines. Nature Genetics. 2017;.

[244] Dobes M, Khurana VG, Shadbolt B, et al. Increasing incidence of glioblastoma multiforme and meningioma, and decreasing incidence of Schwannoma (2000-2008): Findings of a multicenter Australian study. Surgical neurology international. 2011;2:176.

[245] Johnson DR, O'Neill BP. Glioblastoma survival in the United States before and during the temozolomide era. Journal of neuro-oncology. 2012;107(2):359–64.

[246] Verhaak RGW, Hoadley KA, Purdom E, et al. Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell. 2010;17(1):98–110.

[247] Sturm D, Witt H, Hovestadt V, et al. Hotspot Mutations in H3F3A and IDH1 Define Distinct Epigenetic and Biological Subgroups of Glioblastoma. Cancer Cell. 2012;22(4):425–437.

[248] Xu W, Yang H, Liu Y, et al. Oncometabolite 2-hydroxyglutarate is a competitive inhibitor of α-ketoglutarate-dependent dioxygenases. Cancer Cell. 2011;.

[249] Yang Z, Jiang B, Wang Y, et al. 2-HG Inhibits Necroptosis by Stimulating DNMT1-Dependent Hypermethylation of the RIP3 Promoter. Cell Reports. 2017;.

[250] Zhang W, Xu J. DNA methyltransferases and their roles in tumorigenesis. Biomarker Research. 2017;.

[251] Lu C, Thompson CB. Metabolic regulation of epigenetics; 2012.

[252] Q W, B H, X H, et al. Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the Microenvironment. Cancer Cell. 2017;32(1):42–56.e6.

[253] Noushmehr H, Weisenberger DJ, Diefes K, et al. Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma. Cancer Cell. 2010;17(5):510–522.

[254] Turcan S, Rohle D, Goenka A, et al. IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. Nature. 2012;483(7390):479–483.

[255] Dillies MA, Rau A, Aubert J, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in Bioinformatics. 2013;14(6):671–683.

[256] Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. Nature biotechnology. 2011;29(1):24–6.

[257] Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: From properties to genome-wide predictions; 2014.

[258] Heinz S, Romanoski CE, Benner C, et al.. The selection and function of cell type-specific enhancers; 2015.

[259] Lin CY, Erkek S, Tong Y, et al. Active medulloblastoma enhancers reveal subgroup-specific cellular origins. Nature. 2016;530(7588):57–62.

[260] Zhou X, Lowdon RF, Li D, et al.. Exploring long-range genome interactions using the WashU Epigenome Browser; 2013.

[261] McLean CY, Bristor D, Hiller M, et al. GREAT improves functional interpretation of cis-regulatory regions. Nature Biotechnology. 2010;28(5):495–501.

[262] Lewis CA, Brault C, Peck B, et al. SREBP maintains lipid biosynthesis and viability of cancer cells under lipid- and oxygen-deprived conditions and defines a gene signature associated with poor survival in glioblastoma multiforme. Oncogene. 2015;.

[263] Cherry AE, Stella N. G protein-coupled receptors as oncogenic signals in glioma: emerging therapeutic avenues. Neuroscience. 2014;278:222–36.

[264] Berezowska S, Schlegel J. Targeting ErbB Receptors in High-Grade Glioma. Current Pharmaceutical Design. 2011;.

[265] Haley EM, Kim Y. The role of basic fibroblast growth factor in glioblastoma multiforme and glioblastoma stem cells and in their in vitro culture; 2014.

[266] Johnston ALM, Lun X, Rahn JJ, et al. The p75 neurotrophin receptor is a central regulator of glioma invasion. PLoS Biology. 2007;.

[267] Bernstein BE, Kamal M, Lindblad-Toh K, et al. Genomic maps and comparative analysis of histone modifications in human and mouse. Cell. 2005;120(2):169–181.

[268] Kim TH, Barrera LO, Zheng M, et al. A high-resolution map of active promoters in the human genome. Nature. 2005;436(7052):876–880.

[269] Sangpairoj K, Vivithanaporn P, Apisawetakan S, et al. RUNX1 Regulates Migration, Invasion, and Angiogenesis via p38 MAPK Pathway in Human Glioblastoma. Cellular and Molecular Neurobiology. 2017;.

[270] Herms JW, Von Loewenich FD, Behnke J, et al. C-MYC oncogene family expression in glioblastoma and survival. Surgical Neurology. 1999;.

[271] Bjerke L, Mackay A, Nandhabalan M, et al. Histone H3.3 mutations drive pediatric glioblastoma through upregulation of MYCN. Cancer Discovery. 2013;.

[272] Rheinbay E, Suvà ML, Gillespie SM, et al. An Aberrant Transcription Factor Network Essential for Wnt Signaling and Stem Cell Maintenance in Glioblastoma. Cell Reports. 2013;.

[273] Park NI, Guilhamon P, Desai K, et al. ASCL1 Reorganizes Chromatin to Direct Neuronal Fate and Suppress Tumorigenicity of Glioblastoma Stem Cells. Cell stem cell. 2017;21(2):209–224.e7.

[274] Saunders LR, Bankovich AJ, Anderson WC, et al. A DLL3-targeted antibody-drug conjugate eradicates high-grade pulmonary neuroendocrine tumor-initiating cells in vivo. Science Translational Medicine. 2015;.

[275] Muraguchi T, Tanaka S, Yamada D, et al. NKX2.2 suppresses self-renewal of glioma-initiating cells. Cancer Research. 2011;.

[276] Trépant AL, Bouchart C, Rorive S, et al. Identification of OLIG2 as the most specific glioblastoma stem cell marker starting from comparative analysis of data from similar DNA chip microarray platforms. Tumor Biology. 2015;.

[277] Leelatian N, Ihrie RA. Head of the Class: OLIG2 and Glioblastoma Phenotype; 2016.

[278] Weigle B, Ebner R, Temme A, et al. Highly specific overexpression of the transcription factor SOX11 in human malignant gliomas. Oncology Reports. 2005;.

[279] Chen X, Hu H, He L, et al. A novel subtype classification and risk of breast cancer by histone modification profiling. Breast Cancer Research and Treatment. 2016;157(2):267–279.

[280] Bogdanović O, Long SW, Van Heeringen SJ, et al. Temporal uncoupling of the DNA methylome and transcriptional repression during embryogenesis. Genome Research. 2011;21(8):1313–1327.

[281] Kelley DZ, Flam EL, Izumchenko E, et al. Integrated analysis of whole-genome ChIP-Seq and RNA-Seq data of primary head and neck tumor samples associates HPV integration sites with open chromatin marks. Cancer Research. 2017;77(23):6538–6550.

[282] Gal-Yam EN, Egger G, Iniguez L, et al. Frequent switching of Polycomb repressive marks and DNA hypermethylation in the PC3 prostate cancer cell line. Proceedings of the National Academy of Sciences of the United States of America. 2008;.

[283] Hinoue T, Weisenberger DJ, Lange CPE, et al. Genome-scale analysis of aberrant DNA methylation in colorectal cancer. Genome Research. 2012;22(2):271–282.

[284] Hahn MA, Li AX, Wu X, et al. Loss of the polycomb mark from bivalent promoters leads to activation of cancer-promoting genes in colorectal tumors. Cancer Research. 2014;74(13):3617–3629.

[285] Ohm JE, McGarvey KM, Yu X, et al. A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. Nature Genetics. 2007;39(2):237–242.

[286] Kretzmer H, Bernhart SH, Wang W, et al. DNA methylome analysis in Burkitt and follicular lymphomas identifies differentially methylated regions linked to somatic mutation and transcriptional control. Nature Genetics. 2015;47(11):1316–1325.

[287] Bernhart SH, Kretzmer H, Holdt LM, et al. Changes of bivalent chromatin coincide with increased expression of developmental genes in cancer. Scientific Reports. 2016;6(1):37393.

[288] Liang K, Keles S. Detecting differential binding of transcription factors with ChIP-seq. Bioinformatics (Oxford, England). 2012;28(1):121–2.

[289] Baldi P, Sadowski P, Whiteson D. Searching for exotic particles in high-energy physics with deep learning. Nature Communications. 2014;5.

[290] Wang HW, Sun HJ, Chang TY, et al. Discovering monotonic stemness marker genes from time-series stem cell microarray data. BMC Genomics. 2015;16.

[291] Hartigan JA, Wong MA. A K-Means Clustering Algorithm. Applied Statistics. 1979;28(1):100–108.

[292] Cao B, Yang Y, Pan Y, et al. Epigenetic silencing of CXCL14 induced colorectal cancer migration and invasion. Discovery medicine. 2013;16(88):137–47.

[293] Meshcheryakova A, Svoboda M, Tahir A, et al. Exploring the role of sphingolipid machinery during the epithelial to mesenchymal transition program using an integrative approach. Oncotarget. 2016;7(16):22295–323.

[294] Thiery JP. Epithelial–mesenchymal transitions in tumour progression. Nature Reviews Cancer. 2002;.

[295] Tönjes M, Barbus S, Park YJ, et al. BCAT1 promotes cell proliferation through amino acid catabolism in gliomas carrying wild-type IDH1. Nature medicine. 2013;19(7):901–908.

[296] Joshi AD, Parsons DW, Velculescu VE, et al. Sodium ion channel mutations in glioblastoma patients correlate with shorter survival. Molecular Cancer. 2011;10.

[297] Puget S, Philippe C, Bax DA, et al. Mesenchymal transition and PDGFRA amplification/mutation are key distinct oncogenic events in pediatric diffuse intrinsic pontine gliomas. PloS one. 2012;7(2):e30313.

[298] Kupp R, Shtayer L, Tien AC, et al. Lineage-Restricted OLIG2-RTK Signaling Governs the Molecular Subtype of Glioma Stem-like Cells. Cell reports. 2016;16(11):2838–2845.

[299] Sudmant PH, Alexis MS, Burge CB. Meta-analysis of RNA-seq expression data across species, tissues and studies. Genome Biology. 2015;16(1).

[300] Sun X, Dalpiaz D, Wu D, et al. Statistical inference for time course RNA-Seq data using a negative binomial mixed-effect model. BMC Bioinformatics. 2016;17(1).

[301] Ziller MJ, Edri R, Yaffe Y, et al. Dissecting neural differentiation regulatory networks through epigenetic footprinting. Nature. 2014;518(7539):355–359.

[302] Reik W. Stability and flexibility of epigenetic gene regulation in mammalian development; 2007.

[303] Stadler MB, Murr R, Burger L, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature. 2011;480(7378):490–495.

[304] Bernstein BE, Stamatoyannopoulos JA, Costello JF, et al.. The NIH roadmap epigenomics mapping consortium; 2010.

[305] Adams D, Altucci L, Antonarakis SE, et al.. BLUEPRINT to decode the epigenetic signature written in blood; 2012.

[306] Stunnenberg HG, Consortium TIHE, Hirst M, et al. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. Cell. 2016;167(5):1145–1149.

[307] Kent WJ, Zweig AS, Barber G, et al. BigWig and BigBed: Enabling browsing of large distributed datasets. Bioinformatics. 2010;26(17):2204–2207.

[308] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics (Oxford, England). 2010;.

[309] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;.

[310] Périer RC, Praz V, Junier T, et al. The eukaryotic promoter database (EPD). Nucleic acids research. 2000;28(1):302–3.

[311] Fishilevich S, Nudel R, Rappaport N, et al. GeneHancer: genome-wide integration

of enhancers and target genes in GeneCards. Database. 2017;2017(1).

[312] ENCODE Project Consortium AIEoDEitH. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74.

[313] Zerbino DR, Wilder SP, Johnson N, et al. The Ensembl Regulatory Build. Genome Biology. 2015;16(1).

[314] Visel A, Minovitsky S, Dubchak I, et al. VISTA Enhancer Browser - A database of tissue-specific human enhancers. Nucleic Acids Research. 2007;35(SUPPL. 1).

[315] Hinrichs AS. The UCSC Genome Browser Database: update 2006. Nucleic Acids Research. 2006;34(90001):D590–D598.

[316] Hicks SC, Irizarry RA. quantro: A data-driven approach to guide the choice of an appropriate normalization method. Genome Biology. 2015;.

[317] Bolstad BM, Irizarry Ra, Astrand M, et al. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics (Oxford, England). 2003;19(2):185–93.

[318] Cloonan N, Forrest ARR, Kolle G, et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nature Methods. 2008;5(7):613–619.

[319] Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology. 2010;11(3):R25.

[320] Bullard JH, Purdom E, Hansen KD, et al. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 2010;11.

[321] Yousefi P, Huen K, Schall RA, et al. Considerations for normalization of DNA methylation data by Illumina 450K BeadChip assay in population studies. Epigenetics. 2013;8(11):1141–1152.

[322] Bilodeau S, Kagey MH, Frampton GM, et al. SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state. Genes and Development. 2009;23(21):2484–2489.

[323] Kasowski M, Grubert F, Heffelfinger C, et al. Variation in transcription factor binding among humans. Science. 2010;328(5975):232–235.

[324] Di Cerbo V, Schneider R. Cancers with wrong HATs: The impact of acetylation. Briefings in Functional Genomics. 2013;12(3):231–243.

[325] Walker EJ, Zhang C, Castelo-Branco P, et al. Monoallelic expression determines oncogenic progression and outcome in benign and malignant brain tumors. Cancer research. 2012;72(3):636–44.

[326] Eisenberg E, Levanon EY. Human housekeeping genes, revisited; 2013.

[327] Hagarman JA, Motley MP, Kristjansdottir K, et al. Coordinate Regulation of DNA Methylation and H3K27me3 in Mouse Embryonic Stem Cells. PLoS ONE. 2013;8(1).

[328] Yang X, Hu B, Hou Y, et al.. Silencing of developmental genes by H3K27me3 and DNA methylation reflects the discrepant plasticity of embryonic and extraembryonic lineages; 2018.

[329] Dai H, Leeder JS, Cui Y. A modified generalized fisher method for combining probabilities from dependent tests. Frontiers in Genetics. 2014;5(FEB).

[330] Ji H, Li X, Wang Qf, et al. Differential principal component analysis of ChIP-seq. Proceedings of the National Academy of Sciences. 2013;110(17):6789–6794.

[331] Visel A, Rubin EM, Pennacchio LA. Genomic views of distant-acting enhancers; 2009.

[332] Fudenberg G, Mirny LA. Higher-order chromatin structure: Bridging physics and biology; 2012.

[333] Pombo A, Nicodemi M. Physical mechanisms behind the large scale features of

chromatin organization. Transcription. 2014;5(2):e28447.

[334] Mifsud B, Tavares-Cadete F, Young AN, et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nature Genetics. 2015;47(6):598–606.

[335] Javierre BM, Sewitz S, Cairns J, et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell. 2016;167(5):1369–1384.e19.

[336] Zeitz MJ, Ay F, Heidmann JD, et al. Genomic Interaction Profiles in Breast Cancer Reveal Altered Chromatin Architecture. PLoS ONE. 2013;8(9).

[337] Elemento O, Rubin MA, Rickman DS. Oncogenic transcription factors as master regulators of chromatin topology: A new role for ERG in prostate cancer; 2012.

[338] Ferraro A. Altered primary chromatin structures and their implications in cancer development; 2016.

[339] Schmitt AD, Hu M, Jung I, et al. A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. Cell Reports. 2016;17(8):2042–2059.

[340] Lajoie BR, Dekker J, Kaplan N. The Hitchhiker's guide to Hi-C analysis: Practical guidelines. Methods. 2015;72(C):65–75.

[341] Baxter JS, Leavy OC, Dryden NH, et al. Capture Hi-C identifies putative target genes at 33 breast cancer risk loci. Nature Communications. 2018;9(1).

[342] Rubin AJ, Barajas BC, Furlan-Magaril M, et al. Lineage-specific dynamic and pre-established enhancer-promoter contacts cooperate in terminal differentiation. Nature genetics. 2017;49(10):1522–1528.

[343] Jäger R, Migliorini G, Henrion M, et al. Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. Nature Communications. 2015;6.

[344] Brin S, Page L. The anatomy of a large scale hypertextual Web search engine. Computer Networks and ISDN Systems. 1998;30(1/7):107–17.

[345] Iván G, Grolmusz V. When the web meets the cell: Using personalized PageRank for analyzing protein interaction networks. Bioinformatics. 2011;27(3):405–407.

[346] Morrison JL, Breitling R, Higham DJ, et al. GeneRank: Using search engine technology for the analysis of microarray experiments. BMC Bioinformatics. 2005;6.

[347] Kurdistani SK, Tavazoie S, Grunstein M. Mapping global histone acetylation patterns to gene expression. Cell. 2004;117(6):721–733.

[348] Lin CS, Xin ZC, Dai J, et al.. Commonly used mesenchymal stem cell markers and tracking labels: Limitations and challenges; 2013.

[349] Luo Y, Cai J, Liu Y, et al. Microarray analysis of selected genes in neural stem and progenitor cells. Journal of Neurochemistry. 2002;83(6):1481–1497.

[350] Kotliarova S, Fine HA. SnapShot: glioblastoma multiforme. Cancer cell. 2012;21(5):710–710.e1.

[351] Ten Hacken E, Guièze R, Wu CJ. SnapShot: Chronic Lymphocytic Leukemia. Cancer cell. 2017;32(5):716–716.e1.

[352] Ciccone M, Ferrajoli A, Keating MJ, et al. SnapShot: chronic lymphocytic leukemia. Cancer cell. 2014;26(5):770–770.e1.

[353] Forbes SA, Beare D, Boutselakis H, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic acids research. 2017;45(D1):D777–D783.

[354] Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, et al. IntOGen-mutations identifies cancer drivers across tumor types. Nature Methods. 2013;10(11):1081–1084.

[355] Rappaport N, Nativ N, Stelzer G, et al. MalaCards: an integrated compendium for diseases and their annotation. Database : the journal of biological databases and

curation. 2013;2013:bat018.

[356] Lachmann A, Torre D, Keenan AB, et al. Massive mining of publicly available RNA-seq data from human and mouse. Nat Commun. 2018;9(1):1366.

[357] Kuleshov MV, Jones MR, Rouillard AD, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic acids research. 2016;44(W1):W90–7.

[358] Khan A, Mathelier A, Zhang X. Super-enhancers are transcriptionally more active and cell-type-specific than stretch enhancers. bioRxiv. 2018;.

[359] Geeleher P, Hartnett L, Egan LJ, et al. Gene-set analysis is severely biased when applied to genome-wide methylation data. Bioinformatics (Oxford, England). 2013;29(15):1851–7.

[360] Keshava Prasad TS, Goel R, Kandasamy K, et al. Human Protein Reference Database–2009 update. Nucleic acids research. 2009;37(Database issue):D767–72.

[361] Lin J, Gan CM, Zhang X, et al. A multidimensional analysis of genes mutated in breast and colorectal cancers. Genome Research. 2007;17(9):1304–1318.

[362] Jonsson PF, Bates PA. Global topological features of cancer proteins in the human interactome. Bioinformatics. 2006;22(18):2291–2297.

[363] Davoli T, Xu AW, Mengwasser KE, et al. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. Cell. 2013;155(4):948–62.

[364] Arvaniti E, Ntoufa S, Papakonstantinou N, et al. Toll-like receptor signaling pathway in chronic lymphocytic leukemia: Distinct gene expression profiles of potential pathogenic significance in specific subsets of patients. Haematologica. 2011;96(11):1644–1652.

[365] Keith B, Johnson RS, Simon MC. HIF1α and HIF2α: sibling rivalry in hypoxic tumour growth and progression. Nature reviews Cancer. 2011;12(1):9–22.

[366] Qing G, Simon MC. Hypoxia inducible factor-2α: a critical mediator of aggressive tumor phenotypes; 2009.

[367] Yoshimura H, Dhar DK, Kohno H, et al. Prognostic Impact of Hypoxia-Inducible Factors 1α and 2α in Colorectal Cancer Patients. Clinical Cancer Research. 2004;10(24):8554–8560.

[368] Joseph JV, Balasubramaniyan V, Walenkamp A, et al. TGF-β as a therapeutic target in high grade gliomas - promises and challenges. Biochemical pharmacology. 2013;85(4):478–85.

[369] Tchorz JS, Tome M, Cloëtta D, et al. Constitutive Notch2 signaling in neural stem cells promotes tumorigenic features and astroglial lineage entry. Cell death & disease. 2012;3:e325.

[370] Nozhat Z, Hedayati M. PI3K/AKT Pathway and Its Mediators in Thyroid Carcinomas. Molecular diagnosis & therapy. 2016;20(1):13–26.

[371] Eggo MC, Hopkins JM, Franklyn JA, et al. Expression of fibroblast growth factors in thyroid cancer. The Journal of clinical endocrinology and metabolism. 1995;80(3):1006–11.

[372] St Bernard R, Zheng L, Liu W, et al. Fibroblast growth factor receptors as molecular targets in thyroid carcinoma. Endocrinology. 2005;146(3):1145–53.

[373] Yu B, Zhao X, Yang C, et al. PTH Induces Differentiation of Mesenchymal Stem Cells by Enhancing BMP Signaling. J Bone Miner Res J Bone Miner Res. 2012;27(9):2001–2014.

[374] Wu SM, Choo ABH, Yap MGS, et al. Role of Sonic hedgehog signaling and the expression of its components in human embryonic stem cells. Stem cell research.

2010;4(1):38–49.

[375] Kumar A, Declercq J, Eggermont K, et al. Zic3 induces conversion of human fibroblasts to stable neural progenitor-like cells. Journal of molecular cell biology. 2012;4(4):252–5.

[376] Hinck L. The versatile roles of "axon guidance" cues in tissue morphogenesis; 2004.

[377] Qiu R, Wang X, Davy A, et al. Regulation of neural progenitor cell state by ephrin-B. The Journal of cell biology. 2008;181(6):973–83.

[378] Jiao Jw, Feldheim DA, Chen DF. Ephrins as negative regulators of adult neurogenesis in diverse regions of the central nervous system. Proceedings of the National Academy of Sciences. 2008;105(25):8778–8783.

[379] Adams RH, Porras A, Alonso G, et al. Essential role of p38α MAP kinase in placental but not embryonic cardiovascular development. Molecular Cell. 2000;6(1):109–116.

[380] Vaillancourt C, Lanoix D, Le Bellego F, et al. Involvement of MAPK signalling in human villous trophoblast differentiation. Mini reviews in medicinal chemistry. 2009;9(8):962–73.

[381] Daoud G, Amyot M, Rassart E, et al. ERK1/2 and p38 regulate trophoblasts differentiation in human term placenta. The Journal of physiology. 2005;566(Pt 2):409–23.

[382] Fairchild Benyo D, Conrad KP. Expression of the Erythropoietin Receptor by Trophoblast Cellsin the Human Placenta. Biol Reprod. 1999;60(4):861–870.

[383] Tang YN, Ding WQ, Guo XJ, et al. Epigenetic regulation of Smad2 and Smad3 by profilin-2 promotes lung cancer growth and metastasis. Nature Communications. 2015;6(May 2014):8230.

[384] Rybka J, Butrym A, Wróbel T, et al. The Expression of Toll-Like Receptors in Patients with B-Cell Chronic Lymphocytic Leukemia. Archivum Immunologiae et Therapiae Experimentalis. 2016;64:147–150.

[385] Rybka J, Butrym A, Wróbel T, et al. The expression of Toll-like receptors in patients with acute myeloid leukemia treated with induction chemotherapy. Leukemia research. 2015;39(3):318–22.

[386] World Cancer Research Fund and American Institute of Cancer Research (WCRF and AICR). Food, nutrition, physical activity, and the prevention of cancer: a global perspective. World Cancer Research Fund International. 2007; p. 517.

[387] Baba Y, Nosho K, Shima K, et al. HIF1A overexpression is associated with poor prognosis in a cohort of 731 colorectal cancers. American Journal of Pathology. 2010;176(5):2292–2301.

[388] Puente XS, Beà S, Valdés-Mas R, et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. Nature. 2015;526(7574):519–524.

[389] Tseng YY, Moriarity BS, Gong W, et al. PVT1 dependence in cancer with MYC copy-number increase. Nature. 2014;512(1):82–86.

[390] Kim T, Cui R, Jeon YJ, et al. Long-range interaction and correlation between MYC enhancer and oncogenic long noncoding RNA CARLo-5. Proceedings of the National Academy of Sciences. 2014;.

[391] Shi J, Whyte WA, Zepeda-Mendoza CJ, et al. Role of SWI/SNF in acute leukemia maintenance and enhancer-mediated Myc regulation. Genes and Development. 2013;.

[392] Cho SW, Xu J, Sun R, et al. Promoter of lncRNA Gene PVT1 Is a Tumor-Suppressor DNA Boundary Element. Cell. 2018;173(6):1398–1412.e22.

[393] Schwartzman O, Mukamel Z, Oded-Elkayam N, et al. UMI-4C for quantitative and targeted chromosomal contact profiling. Nature Methods. 2016;13(8):685–691.

[394] Wang Y, Zhang B, Zhang L, et al. The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. bioRxiv. 2017; p. 112268.

[395] Bahr C, Von Paleske L, Uslu VV, et al. A Myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies. Nature. 2018;553(7689):515–520.

[396] Shannon P, Markiel A, Ozier O, et al. Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research. 2003;13(Karp 2001):2498–2504.

[397] Predeus AV, Gopalakrishnan S, Huang Y, et al. Targeted Chromatin Profiling Reveals Novel Enhancers in Ig H and Ig L Chain Loci. The Journal of Immunology. 2014;192(3):1064–1070.

[398] Rozowsky J, Euskirchen G, Auerbach RK, et al. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nature Biotechnology. 2009;.

[399] Heintzman ND, Hon GC, Hawkins RD, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature. 2009;459(7243):108–112.

[400] Mohrs M, Blankespoor CM, Wang ZE, et al. Deletion of a coordinate regulator of type 2 cytokine expression in mice. Nature Immunology. 2001;.

[401] Jadamba E, Shin M. A novel approach to significant pathway identification using pathway interaction network from PPI data. Biochip Journal. 2014;8(1):22–27.

[402] Paulsen J, Rødland EA, Holden L, et al. A statistical model of ChIA-PET data for accurate detection of chromatin 3D interactions. Nucleic Acids Research. 2014;42(18).

[403] von Mering C, Huynen M, Jaeggi D, et al.. STRING: A database of predicted functional associations between proteins; 2003.

[404] Islam MO, Kanemura Y, Tajria J, et al. Functional expression of ABCG2 transporter in human neural stem/progenitor cells. Neuroscience Research. 2005;52(1):75–82.

[405] Castro DS, Martynoga B, Parras C, et al. A novel function of the proneural factor Ascl1 in progenitor proliferation identified by genome-wide characterization of its targets. Genes and Development. 2011;25(9):930–945.

[406] Molofsky AV, Pardal R, Iwashita T, et al. Bmi-1 dependence distinguishes neural stem cell self-renewal from progenitor proliferation. Nature. 2003;425(6961):962–967.

[407] Peh GSL, Lang RJ, Pera MF, et al. CD133 Expression by Neural Progenitors Derived from Human Embryonic Stem Cells and Its Use for Their Prospective Isolation. Stem Cells and Development. 2009;18(2):269–282.

[408] Bang SY, Kwon SH, Yi SH, et al. Epigenetic activation of the Foxa2 gene is required for maintaining the potential of neural precursor cells to differentiate into dopaminergic neurons after expansion. Stem Cells Dev. 2015;24(4):520–533.

[409] Kirkeby A, Grealish S, Wolf DA, et al. Generation of Regionally Specified Neural Progenitors and Functional Neurons from Human Embryonic Stem Cells under Defined Conditions. Cell Reports. 2012;1(6):703–714.

[410] Kim DY, Hwang I, Muller FL, et al. Functional regulation of FoxO1 in neural stem cell differentiation. Cell Death and Differentiation. 2015;22(12):2034–2045.

[411] Fathi A, Hatami M, Hajihosseini V, et al. Comprehensive gene expression analysis of human embryonic stem cells during differentiation into neural cells. PLoS ONE. 2011;6(7).

[412] Livesey FJ, Young TL, Cepko CL. An analysis of the gene expression program of mammalian neural progenitor cells. Proceedings of the National Academy of

Sciences of the United States of America. 2004;101(5):1374–1379.

[413] Zhao J, Yao Y, Xu C, et al. Expression of GAP-43 in fibroblast cell lines influences the orientation of cell division. International journal of developmental neuroscience : the official journal of the International Society for Developmental Neuroscience. 2011;29(4):469–74.

[414] Doetsch F, Caillé I, Lim DA, et al. Subventricular Zone Astrocytes Are Neural Stem Cells in the Adult Mammalian Brain. Cell. 1999;97(6):703–716.

[415] Middeldorp, Boer, Sluijs, et al. GFAPdelta in radial glia and subventricular zone progenitors in the developing human cortex. Development (Cambridge, England). 2010;137:313–321.

[416] Maurer MH, Geomor HK, Bürgers HF, et al. Adult neural stem cells express glucose transporters GLUT1 and GLUT3 and regulate GLUT3 expression. FEBS Letters. 2006;580(18):4430–4434.

[417] Kobayashi T, Kageyama R. Hes1 regulates embryonic stem cell differentiation by suppressing Notch signaling. Genes to Cells. 2010;15(7):689–698.

[418] Shimojo H, Ohtsuka T, Kageyama R. Dynamic Expression of Notch Signaling Genes in Neural Stem/Progenitor Cells. Frontiers in Neuroscience. 2011;5.

[419] Kaneko J, Chiba C. Immunohistochemical analysis of Musashi-1 expression during retinal regeneration of adult newt. Neuroscience Letters. 2009;450(3):252–257.

[420] Murdoch B, Roskams AJ. A Novel Embryonic Nestin-Expressing Radial Glia-Like Progenitor Gives Rise to Zonally Restricted Olfactory and Vomeronasal Neurons. Journal of Neuroscience. 2008;28(16):4271–4282.

[421] Steiner B, Zurborg S, Hörster H, et al. Differential 24 h responsiveness of Prox1-expressing precursor cells in adult hippocampal neurogenesis to physical activity, environmental enrichment, and kainic acid-induced seizures. Neuroscience. 2008;154(2):521–529.

[422] Heng YHE, McLeay RC, Harvey TJ, et al. NFIX Regulates Neural Progenitor Cell Differentiation During Hippocampal Morphogenesis. Cerebral cortex (New York, NY : 1991). 2012; p. 1–19.

[423] Yang X, Klein R, Tian X, et al. Notch activation induces apoptosis in neural progenitor cells through a p53-dependent pathway. Developmental Biology. 2004;269(1):81–94.

[424] Cui XY, Hu QD, Tekaya M, et al. NB-3/Notch1 pathway via Deltex1 promotes neural progenitor cell differentiation into oligodendrocytes. Journal of Biological Chemistry. 2004;279(24):25858–25865.

[425] Dominici C, Moreno-Bravo JA, Puiggros SR, et al. Floor-plate-derived netrin-1 is dispensable for commissural axon guidance. Nature. 2017;545(7654):350–354.

[426] Li XJ, Du ZW, Zarnowska ED, et al. Specification of motoneurons from human embryonic stem cells. Nat Biotechnol. 2005;23(2):2121–2152.

[427] Basch ML, Bronner-Fraser M, García-Castro MI. Specification of the neural crest occurs during gastrulation and requires Pax7. Nature. 2006;441(7090):218–222.

[428] Blake JA, Ziman MR. Pax genes: regulators of lineage specification and progenitor cell maintenance. Development. 2014;141(4):737–751.

[429] Pankratz MT, Li XJ, LaVaute TM, et al. Directed Neural Differentiation of Human Embryonic Stem Cells via an Obligated Primitive Anterior Stage. Stem Cells. 2007;25(6):1511–1520.

[430] Vives V, Alonso G, Solal AC, et al. Visualization of S100B-positive neurons and glia in the central nervous system of EGFP transgenic mice. Journal of Comparative Neurology. 2003;457(4):404–419.

[431] Matsumoto S, Banine F, Struve J, et al. Brg1 is required for murine neural stem cell maintenance and gliogenesis. Developmental Biology. 2006;289(2):372–383.

[432] Venere M, Han YG, Bell R, et al. Sox1 marks an activated neural stem/progenitor cell in the hippocampus. Development. 2012;139(21):3938–3949.

[433] Bergsland M, Werme M, Malewicz M, et al. The establishment of neuronal properties is controlled by Sox4 and Sox11. Genes and Development. 2006;20(24):3475–3486.

[434] Graham V, Khudyakov J, Ellis P, et al. SOX2 functions to maintain neural progenitor identity. Neuron. 2003;39(5):749–765.

[435] Ellis P, Fagan BM, Magness ST, et al. SOX2, a persistent marker for multipotential neural stem cells derived from embryonic stem cells, the embryo or the adult. Developmental Neuroscience. 2004;26(2-4):148–165.

[436] Wang TW, Stromberg GP, Whitney JT, et al. Sox3 expression identifies neural progenitors in persistent neonatal and adult mouse forebrain germinative zones. Journal of Comparative Neurology. 2006;497(1):88–100.

[437] Scott CE, Wynn SL, Sesay A, et al. SOX9 induces and maintains neural stem cells. Nature Neuroscience. 2010;13(10):1181–1189.

[438] Uittenbogaard M, Chiaramello a. Expression of the bHLH transcription factor Tcf12 (ME1) gene is linked to the expansion of precursor cell populations during neurogenesis. Brain research Gene expression patterns. 2002;1(2):115–121.

[439] Nakagawa T, Miyazaki T, Miyamoto O, et al. Regional expression of the radial glial marker vimentin at different stages of the kindling process. Epilepsy Research. 2004;61(1-3):141–151.

[440] Arai F, Ohneda O, Miyamoto T, et al. Mesenchymal Stem Cells in Perichondrium Express Activated Leukocyte Cell Adhesion Molecule and Participate in Bone Marrow Formation. The Journal of Experimental Medicine. 2002;195(12):1549–1563.

[441] Frobel J, Hemeda H, Lenz M, et al. Epigenetic rejuvenation of mesenchymal stromal cells derived from induced pluripotent stem cells. Stem Cell Reports. 2014;3(3):414–422.

[442] Obara C, Takizawa K, Tomiyama K, et al. Differentiation and molecular properties of mesenchymal stem cells derived from murine induced pluripotent stem cells derived on gelatin or collagen. Stem Cells International. 2016;2016.

[443] Abdallah BM, Boissy P, Tan Q, et al. dlk1/FA1 regulates the function of human bone marrow mesenchymal stem cells by modulating gene expression of pro-inflammatory cytokines and immune response-related factors. The Journal of biological chemistry. 2007;282(10):7339–51.

[444] Abdallah BM, Jensen CH, Gutierrez G, et al. Regulation of human skeletal stem cells differentiation by Dlk1/Pref-1. Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research. 2004;19(5):841–852.

[445] Kubo H, Shimizu M, Taya Y, et al. Identification of mesenchymal stem cell (MSC)-transcription factors by microarray and knockdown analyses, and signature molecule-marked MSC in bone marrow by immunohistochemistry. Genes to Cells. 2009;14(3):407–424.

[446] Almalki SG, Agrawal DK. Key transcription factors in the differentiation of mesenchymal stem cells; 2016.

[447] Cui LL, Nitzsche F, Pryazhnikov E, et al. Integrin α4 Overexpression on Rat Mesenchymal Stem Cells Enhances Transmigration and Reduces Cerebral Embolism After Intracarotid Injection. Stroke. 2017;48(10):2895–2900.

[448] Ball SG, Shuttleworth A, Kielty CM. Inhibition of platelet-derived growth factor receptor signaling regulates Oct4 and Nanog expression, cell shape, and mesenchymal

stem cell potency. Stem Cells. 2012;30(3):548–560.

[449] Farahani RM, Xaymardan M. Platelet-Derived Growth Factor Receptor Alpha as a Marker of Mesenchymal Stem Cells in Development and Stem Cell Biology. Stem Cells International. 2015;2015:1–8.

[450] Han SM, Han SH, Coh YR, et al. Enhanced proliferation and differentiation of Oct4- And Sox2-overexpressing human adipose tissue mesenchymal stem cells. Experimental and Molecular Medicine. 2014;46(6).

[451] Matic I, Antunovic M, Brkic S, et al. Expression of OCT-4 and SOX-2 in bone marrow-derived human mesenchymal stem cells during osteogenic differentiation. Macedonian Journal of Medical Sciences. 2016;4(1).

[452] Park SB, Seo KW, So AY, et al. SOX2 has a crucial role in the lineage determination and proliferation of mesenchymal stem cells through Dickkopf-1 and c-MYC. Cell Death and Differentiation. 2012;19(3):534–545.

[453] Tiwari N, Tiwari VK, Waldmeier L, et al. Sox4 Is a Master Regulator of Epithelial-Mesenchymal Transition by Controlling Ezh2 Expression and Epigenetic Reprogramming. Cancer Cell. 2013;23(6):768–783.

[454] Bradshaw AD, Sage EH. SPARC, a matricellular protein that functions in cellular differentiation and tissue response to injury; 2001.

[455] Ivaska J, Pallari HM, Nevo J, et al.. Novel functions of vimentin in cell adhesion, migration, and signaling; 2007.

[456] Rhee C, Lee BK, Beck S, et al. Mechanisms of transcription factor-mediated direct reprogramming of mouse embryonic stem cells to trophoblast stem-like cells. Nucleic acids research. 2017;45(17):10103–10114.

[457] Kubaczka C, Senner C, Araúzo-Bravo MJ, et al. Derivation and maintenance of murine trophoblast stem cells under defined conditions. Stem Cell Reports. 2014;2(2):232–242.

[458] Douglas GC, VandeVoort CA, Kumar P, et al.. Trophoblast stem cells: Models for investigating trophectoderm differentiation and placental development; 2009.

[459] Ohinata Y, Tsukiyama T. Establishment of trophoblast stem cells under defined culture conditions in mice. PLoS ONE. 2014;9(9).

[460] Peiffer I, Belhomme D, Barbet R, et al. Simultaneous differentiation of endothelial and trophoblastic cells derived from human embryonic stem cells. Stem cells and development. 2007;16(3):393–402.

[461] Strumpf D. Cdx2 is required for correct cell fate specification and differentiation of trophectoderm in the mouse blastocyst. Development. 2005;132(9):2093–2102.

[462] Chen Y, Wang K, Gong YG, et al. Roles of CDX2 and EOMES in human induced trophoblast progenitor cells. Biochemical and Biophysical Research Communications. 2013;431(2):197–202.

[463] Li Y, Moretto-Zita M, Soncin F, et al. BMP4-directed trophoblast differentiation of human embryonic stem cells is mediated through a ΔNp63+ cytotrophoblast stem cell state. Development (Cambridge, England). 2013;140(19):3965–76.

[464] Schulz LC, Ezashi T, Das P, et al. Human Embryonic Stem Cells as Models for Trophoblast Differentiation. Placenta. 2008;29(SUPPL.):10–16.

[465] Lee CQE, Gardner L, Turco M, et al. What Is Trophoblast? A Combination of Criteria Define Human First-Trimester Trophoblast. Stem Cell Reports. 2016;6(2):257–272.

[466] Kidder BL, Palmer S. Examination of transcriptional networks reveals an important role for TCFAP2C, SMARCA4, and EOMES in trophoblast stem cell maintenance. Genome Research. 2010;20(4):458–472.

[467] Tanaka S, Kunath T, Hadjantonakis AK, et al. Promotion of trophoblast stem cell proliferation by FGF4. Science (New York, NY). 1998;282(5396):2072–5.

[468] Haffner-Krausz R, Gorivodsky M, Chen Y, et al. Expression of Fgfr2 in the early mouse embryo indicates its involvement in preimplantation development. Mechanisms of Development. 1999;85(1-2):167–172.

[469] Selesniemi K, Reedy M, Gultice A, et al. Transforming growth factor-beta induces differentiation of the labyrinthine trophoblast stem cell line SM10. Stem cells and development. 2005;14(6):697–711.

[470] Selesniemi K, Albers RE, Brown TL. Id2 Mediates Differentiation of Labyrinthine Placental Progenitor Cell Line, SM10. Stem Cells and Development. 2016;25(13):959–974.

[471] Liang H, Zhang Q, Lu J, et al. MSX2 Induces Trophoblast Invasion in Human Placenta. PLoS ONE. 2016;11(4).

[472] Yagi R, Kohn MJ, Karavanova I, et al. Transcription factor TEAD4 specifies the trophectoderm lineage at the beginning of mammalian development. Development (Cambridge, England). 2007;134(21):3827–36.

[473] Rubio-Perez C, Tamborero D, Schroeder MP, et al. In Silico Prescription of Anticancer Drugs to Cohorts of 28 Tumor Types Reveals Targeting Opportunities. Cancer Cell. 2015;27(3):382–396.

[474] Wang L, Lawrence MS, Wan Y, et al. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. The New England journal of medicine. 2011;365(26):2497–506.

[475] Guarini A, Marinelli M, Tavolaro S, et al. Atm gene alterations in chronic lymphocytic leukemia patients induce a distinct gene expression profile and predict disease progression. Haematologica. 2012;97(1):47–55.

[476] Landau DA, Carter SL, Stojanov P, et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013;152(4):714–726.

[477] Baliakas P, Hadzidimitriou A, Sutton LA, et al. Recurrent mutations refine prognosis in chronic lymphocytic leukemia. Leukemia. 2015;29(2):329–36.

[478] Alhourani E, Othman MAK, Melo JB, et al. BIRC3 alterations in chronic and B-cell acute lymphocytic leukemia patients. Oncology letters. 2016;11(5):3240–3246.

[479] Landau DA, Tausch E, Taylor-Weiner AN, et al. Mutations driving CLL and their evolution in progression and relapse. Nature. 2015;526(7574):525–530.

[480] Jebaraj BMC, Kienle D, Bühler A, et al. BRAF mutations in chronic lymphocytic leukemia. Leukemia & lymphoma. 2013;54(6):1177–82.

[481] Rodríguez D, Bretones G, Quesada V, et al. Mutations in CHD2 cause defective association with active chromatin in chronic lymphocytic leukemia. Blood. 2015;126(2):195–202.

[482] Quesada V, Conde L, Villamor N, et al. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. Nature Genetics. 2011;44(1):47–52.

[483] Ghobrial IM, Bone ND, Stenson MJ, et al. Expression of the chemokine receptors CXCR4 and CCR7 and disease progression in B-cell chronic lymphocytic leukemia/ small lymphocytic lymphoma. Mayo Clin Proc. 2004;79(3):318–25.

[484] Möhle R, Failenschmid C, Bautz F, et al. Overexpression of the chemokine receptor CXCR4 in B cell chronic lymphocytic leukemia is associated with increased functional response to stromal cell-derived factor-1 (SDF-1). Leukemia. 1999;13(12):1954–9.

[485] Barretina J, Juncà J, Llano A, et al. CXCR4 and SDF-1 expression in B-cell

chronic lymphocytic leukemia and stage of the disease. Annals of Hematology. 2003;82(8):500–505.

[486] Crowther-Swanepoel D, Qureshi M, Dyer MJS, et al. Genetic variation in CXCR4 and risk of chronic lymphocytic leukemia. Blood. 2009;114(23):4843–6.

[487] Ojha J, Secreto CR, Rabe KG, et al.. Identification of recurrent truncated DDX3X mutations in chronic lymphocytic leukaemia; 2015.

[488] Young E, Noerenberg D, Mansouri L, et al. EGR2 mutations define a new clinically aggressive subgroup of chronic lymphocytic leukemia. Leukemia. 2017;31(7):1547–1554.

[489] Jeromin S, Weissmann S, Haferlach C, et al. SF3B1 mutations correlated to cytogenetics and mutations in NOTCH1, FBXW7, MYD88, XPO1 and TP53 in 1160 untreated CLL patients. Leukemia. 2014;28(1):108–117.

[490] Havelange V, Pekarsky Y, Nakamura T, et al. IRF4 mutations in chronic lymphocytic leukemia. Blood. 2011;118(10):2827–2829.

[491] Martínez-Trillos A, Navarro A, Aymerich M, et al. Clinical impact of MYD88 mutations in chronic lymphocytic leukemia. Blood. 2016;127(12):1611–1613.

[492] Rossi D, Rasi S, Fabbri G, et al. Mutations of NOTCH1 are an independent predictor of survival in chronic lymphocytic leukemia. Blood. 2012;119(2):521–529.

[493] Clifford R, Louis T, Robbe P, et al. SAMHD1 is mutated recurrently in chronic lymphocytic leukemia and is involved in response to DNA damage. Blood. 2014;123(7):1021–1031.

[494] Rossi D. SAMHD1: A new gene for CLL; 2014.

[495] Buchner M, Fuchs S, Prinz G, et al. Spleen tyrosine kinase is overexpressed and represents a potential therapeutic target in chronic lymphocytic leukemia. Cancer Research. 2009;69(13):5424–5432.

[496] Baudot AD, Jeandel PY, Mouska X, et al. The tyrosine kinase Syk regulates the survival of chronic lymphocytic leukemia B cells through PKCdelta and proteasome-dependent regulation of Mcl-1 expression. Oncogene. 2009;28(37):3261–3273.

[497] Hoellenriegel J, Coffey GP, Sinha U, et al. Selective, novel spleen tyrosine kinase (Syk) inhibitors suppress chronic lymphocytic leukemia B-cell activation and migration. Leukemia. 2012;26(7):1576–1583.

[498] Wiestner A. ZAP-70 expression identifies a chronic lymphocytic leukemia subtype with unmutated immunoglobulin genes, inferior clinical outcome, and distinct gene expression profile. Blood. 2003;101(>12):4944–4951.

[499] Chen L, Widhopf G, Huynh L, et al. Expression of ZAP-70 is associated with increased B-cell receptor signaling in chronic lymphocytic leukemia. Blood. 2002;100(13):4609–4614.

[500] Sausen M, Leary RJ, Jones S, et al. Integrated genomic analyses identify ARID1A and ARID1B alterations in the childhood cancer neuroblastoma. Nature Genetics. 2013;45(1):12–17.

[501] Wiestler B, Capper D, Holland-Letz T, et al. ATRX loss refines the classification of anaplastic gliomas and identifies a subgroup of IDH mutant astrocytic tumors with better prognosis. Acta Neuropathologica. 2013;126(3):443–451.

[502] Jiao Y, Killela PJ, Reitman ZJ, et al. Frequent ATRX, CIC, FUBP1 and IDH1 mutations refine the classification of malignant gliomas. Oncotarget. 2012;3(7):709–22.

[503] Dahiya S, Emnett RJ, Haydon DH, et al.. BRAF-V600E mutation in pediatric and adult glioblastoma; 2014.

[504] et al McLendon, R, Friedman, A, Bigner, D, Van Meir, EG, Brat, DJ, Mastro-

gianakis, GM, Olson, JJ. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008;455(7216):1061–1068.

[505] Knobbe CB, Reifenberger J, Reifenberger G. Mutation analysis of the Ras pathway genes NRAS, HRAS, KRAS and BRAF in glioblastomas. Acta Neuropathologica. 2004;108(6):467–470.

[506] Toth J, Egervari K, Klekner A, et al. Analysis of EGFR gene amplification, protein over-expression and tyrosine kinase domain mutation in recurrent glioblastoma. Pathology oncology research : POR. 2009;15(2):225–229.

[507] Mukasa A, Wykosky J, Ligon KL, et al. Mutant EGFR is required for maintenance of glioma growth in vivo, and its ablation leads to escape from receptor dependence. Proceedings of the National Academy of Sciences of the United States of America. 2010;107(6):2616–2621.

[508] Baumgarten P, Harter PN, Tönjes M, et al. Loss of FUBP1 expression in gliomas predicts FUBP1 mutation and is associated with oligodendroglial differentiation, IDH1 mutation and 1p/19q loss of heterozygosity. Neuropathology and Applied Neurobiology. 2014;40(2):205–216.

[509] Parsons DW, Jones S, Zhang X, et al. An integrated genomic analysis of human glioblastoma multiforme. Science (New York, NY). 2008;321(5897):1807–1812.

[510] Cohen AL, Holmen SL, Colman H. IDH1 and IDH2 mutations in gliomas.; 2013.

[511] Xu P, Zhang A, Jiang R, et al. The Different Role of Notch1 and Notch2 in Astrocytic Gliomas. PLoS ONE. 2013;8(1).

[512] Gallia GL, Rand V, Siu IM, et al. PIK3CA Gene Mutations in Pediatric and Adult Glioblastoma Multiforme. Molecular Cancer Research. 2006;4(10):709–714.

[513] Kita D, Yonekawa Y, Weller M, et al. PIK3CA alterations in primary (de novo) and secondary glioblastomas. Acta Neuropathologica. 2007;113(3):295–302.

[514] Ozawa T, Brennan CW, Wang L, et al. PDGFRA gene rearrangements are frequent genetic events in PDGFRA-amplified glioblastomas. Genes and Development. 2010;24(19):2205–2218.

[515] Chakravarty D, Pedraza AM, Cotari J, et al. EGFR and PDGFRA co-expression and heterodimerization in glioblastoma tumor sphere lines. Scientific Reports. 2017;7(1):9043.

[516] Quayle SN, Lee JY, Cheung LWT, et al. Somatic Mutations of PIK3R1 Promote Gliomagenesis. PLoS ONE. 2012;7(11).

[517] Wang L, He S, Yuan J, et al. Oncogenic role of SOX9 expression in human malignant glioma. Medical Oncology. 2012;29(5):3484–3490.

[518] Gao J, Zhang JY, Li YH, et al. Decreased expression of SOX9 indicates a better prognosis and inhibits the growth of glioma cells by inducing cell cycle arrest. International Journal of Clinical and Experimental Pathology. 2015;8(9):10130–10138.

[519] Labreche K, Simeonova I, Kamoun A, et al. TCF12 is mutated in anaplastic oligodendroglioma. Nature Communications. 2015;6.

[520] Cancer Genom Atlas. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487(7407):330–337.

[521] Berg M, Danielsen SA, Ahlquist T, et al. DNA sequence profiles of the colorectal cancer critical gene set KRAS-BRAF-PIK3CA-PTEN-TP53 related to age at disease onset. PloS one. 2010;5(11):e13978.

[522] De Roock W, Claes B, Bernasconi D, et al. Effects of KRAS, BRAF, NRAS, and PIK3CA mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: A retrospective consortium analysis. The Lancet Oncology. 2010;11(8):753–762.

[523] Mao C, Wu XY, Yang ZY, et al. Concordant analysis of KRAS, BRAF, PIK3CA mutations, and PTEN expression between primary colorectal cancer and matched metastases. Scientific Reports. 2015;5(1):8065.

[524] Zhang X, Nagahara H, Mimori K, et al. Mutations of epidermal growth factor receptor in colon cancer indicate susceptibility or resistance to gefitinib. Oncology reports. 2008;19(6):1541–4.

[525] Oh BY, Lee RA, Chung SS, et al. Epidermal growth factor receptor mutations in colorectal cancer patients. Journal of the Korean Society of Coloproctology. 2011;27(3):127–32.

[526] Gross ME, Zorbas MA, Danels YJ, et al. Cellular Growth Response to Epidermal Growth Factor in Colon Carcinoma Cells with an Amplified Epidermal Growth Factor Receptor Derived from a Familial Adenomatous Polyposis Patient. Cancer Research. 1991;51(5):1452–1459.

[527] Bos JL, Fearon ER, Hamilton SR, et al. Prevalence of ras gene mutations in human colorectal cancers. Nature. 1987;327(6120):293–7.

[528] Hao Y, Samuels Y, Li Q, et al. Oncogenic PIK3CA mutations reprogram glutamine metabolism in colorectal cancer. Nature Communications. 2016;7:11971.

[529] Roper J, Hung KE. Molecular Mechanisms of Colorectal Carcinogenesis. In: Molecular Pathogenesis of Colorectal Cancer. New York, NY: Springer New York; 2013. p. 25–65.

[530] Ngeow J, Heald B, Rybicki LA, et al. Prevalence of germline PTEN, BMPR1A, SMAD4, STK11, and ENG mutations in patients with moderate-load colorectal polyps. Gastroenterology. 2013;144(7).

[531] Molinari F, Frattini M. Functions and Regulation of the PTEN Gene in Colorectal Cancer. Frontiers in oncology. 2013;3(January):326.

[532] Miyaki M, Iijima T, Konishi M, et al. Higher frequency of Smad4 gene mutation in human colorectal cancer with distant metastasis. Oncogene. 1999;18(20):3098–3103.

[533] Lü B, Fang Y, Xu J, et al. Analysis of SOX9 expression in colorectal cancer. American journal of clinical pathology. 2008;130(6):897–904.

[534] Folsom AR, Pankow JS, Peacock JM, et al. Variation in TCF7L2 and increased risk of colon cancer: the Atherosclerosis Risk in Communities (ARIC) Study. Diabetes care. 2008;31(5):905–9.

[535] Xu Y, Pasche B. TGF-β signaling alterations and susceptibility to colorectal cancer; 2007.

[536] Xing M. Molecular pathogenesis and mechanisms of thyroid cancer; 2013.

[537] Agrawal N, Akbani R, Aksoy BA, et al. Integrated Genomic Characterization of Papillary Thyroid Carcinoma. Cell. 2014;159(3):676–690.

[538] Landa I, Ibrahimpasic T, Boucai L, et al. Genomic and transcriptomic hallmarks of poorly differentiated and anaplastic thyroid cancers. Journal of Clinical Investigation. 2016;126(3):1052–1066.

[539] Kimura ET, Nikiforova MN, Zhu Z, et al. High prevalence of BRAF mutations in thyroid cancer: Genetic evidence for constitutive activation of the RET/PTC-RAS-BRAF signaling pathway in papillary thyroid carcinoma. Cancer Research. 2003;63(7):1454–1457.

[540] Yoo SK, Lee S, Kim SJ, et al. Comprehensive Analysis of the Transcriptional and Mutational Landscape of Follicular and Papillary Thyroid Cancers. PLoS Genetics. 2016;12(8).

[541] Howell GM, Hodak SP, Yip L. RAS mutations in thyroid cancer. The Oncologist. 2013;18:926–932.

[542] Nagy R, Ganapathi S, Comeras I, et al. Frequency of germline PTEN mutations in differentiated thyroid cancer. Thyroid : official journal of the American Thyroid Association. 2011;21(5):505–510.

Appendix

Table S1: Oncogenes, tumor suppressor genes, and housekeeping genes used in the analysis

Oncogenes					
ABL1	ABL2	AKT1	AKT2	ATF1	BCL11A
BCL2	BCL3	BCL6	BCR	BRAF	CARD11
CBLB	CBLC	CCND1	CCND2	CCND3	CDX2
CTNNB1	DDB2	DDIT3	DDX6	DEK	EGFR
ELK4	ERBB2	ETV4	ETV6	EWSR1	FEV
FGFR1	FGFR1OP	FGFR2	FUS	GOLGA5	GOPC
HMGA1	HMGA2	HRAS	IRF4	JUN	KIT
KMT2A	KRAS	LCK	LMO2	MAF	MAFB
MAML2	MDM2	MECOM	MET	MITF	MPL
MYB	MYC	MYCL	MYCN	NCOA4	NFKB2
NRAS	NTRK1	NUP214	PAX8	PDGFB	PIK3CA
PIM1	PLAG1	PPARG	PTPN11	RAF1	REL
RET	ROS1	SMO	SS18	TCL1A	TET2
TFG	TLX1	TPR	USP6		
Tumor suppressor genes					
APC	ARHGEF12	ATM	BCL11B	BLM	BMPR1A
BRCA1	BRCA2	CARS	CBFA2T3	CDH1	CDH11
CDK6	CDKN2C	CEBPA	CHEK2	CREB1	CREBBP
CYLD	DDX5	EXT1	EXT2	FBXW7	FH
FLT3	FOXP1	GPC3	IDH1	IL2	JAK2
MAP2K4	MDM4	MEN1	MLH1	MSH2	NF1
NF2	NOTCH1	NPM1	NR4A3	NUP98	PALB2
PML	PTEN	RB1	RUNX1	SDHB	SDHD
SMARCA4	SMARCB1	SOCS1	STK11	SUFU	SUZ12
SYK	TCF3	TNFAIP3	TP53	TSC1	TSC2
VHL	WRN	WT1			
Housekeeping genes					
C1orf43	CHMP2A	EMC7	GPI	PSMB2	PSMB4
RAB7A	REEP5	SNRPD3	VCP	VPS29	

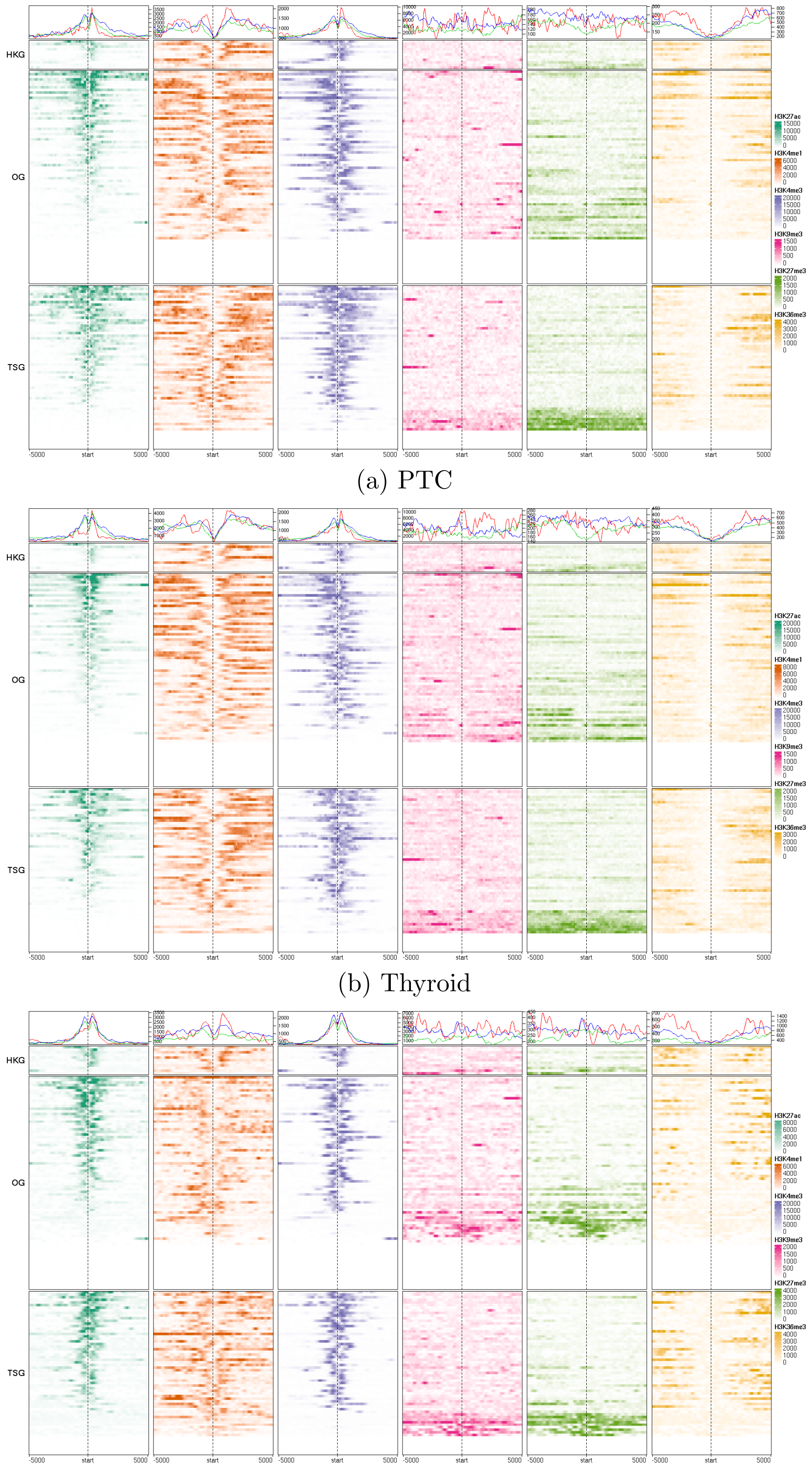
Table S2: Stem cell differentiation marker genes

Neural Progenitor Cells (NPC)			
ABCG2 [404]	ASCL1 [405]	BMI1 [406]	CD133 [407]
CXCR4 [349]	FOXA2 [408,409]	FOXO1 [410]	FZD9 [411]
GAP43 [412,413]	GFAP [414,415]	GLUT1 [416]	HES1 [417,418]
MAP2 [409]	MSI1 [419]	NES [420]	NEUROD1 [421]
NFIX [422]	NOTCH1 [423,424]	NTN1 [425]	OTX2 [426]
PAX3 [427,428]	PAX5 [428]	PAX6 [409,428,429]	PAX7 [428]
PAX8 [428]	S100B [430]	SMARCA4 [431]	SOX1 [429,432]
SOX11 [433]	SOX2 [434,435]	SOX3 [436]	SOX4 [433]
SOX9 [437]	SYP [409]	TCF12 [438]	VIM [439]
Mesenchymal Stem Cells (MSC)			
ALCAM [440]	ANPEP [441]	CD44 [441,442]	CD70 [348]
DLK1 [443,444]	ENG [348,442]	ETV1 [445]	ETV5 [445]
FOXP1 [445]	GATA4 [446]	GATA6 [445]	HMGA2 [445]
ITGA4 [447]	ITGB1 [442]	MYOD1 [446]	NANOG [441,448]
NCAM1 [441]	NT5E [441]	OCT4 [448]	PDGFRA [448,449]
POU5F1 [450,451]	PPARG [446]	RUNX2 [441,446]	SIM2 [445]
SOX11 [445]	SOX2 [450–452]	SOX4 [453]	SOX9 [441,446]
SPARC [441,454]	THY1 [348,442]	VIM [455]	
Trophoblast Stem Cells (TSC)			
ARID3A [456]	BMP4 [457]	CD9 [458]	CDH1 [459]
CDX1 [460]	CDX2 [457,460–462]	CGA [460,463,464]	CGB [458,463,464]
ELF5 [459,465]	EOMES [457,459,463,466]	ESRRB [459]	ETS2 [457]
FGF4 [467]	FGFR2 [457,459,468]	FURIN [457]	GATA2 [460,464]
GATA3 [457]	GCM1 [464]	HAND1 [460]	ID2 [469,470]
IGFBP3 [464]	KRT7 [458,464]	MMP9 [464]	MSX2 [464,471]
SMARCA4 [466]	SOX2 [459,464]	TEAD4 [472]	TFAP2C [457]
TFAP2C [457,459,466]			

Table S3: Cancer marker genes

Chronic Lymphocytic Leukemia (CLL)	
ARID1A (2.41 [388,473])	ATM (9 [474], 4.14 [473,475])
BCOR (1.72 [476])	BIRC3 (2.5 [477], 19.7 [478])
BRAF (3.7 [479], 2.8 [388,480])	CHD2 (5.3 [481], 4.8 [479,482])
CXCR4 (OE [483–486])	DDX3X (1.03 [473], 2.4 [474], 1.72 [487])
EGR2 (3.8 [479,488])	FBXW7 (1.03 [473], 2.5 [474,489])
IRF4 (1.5 [388,490])	MYD88 (2.2 [477], 4 [491], 8 [474], 5.17 [473,489])
PAX5 [388]	NOTCH1 (3.1 [473], 4 [474], 8 [477], 11.3 [488,489,492])
SAMHD1 (11 [493,494])	SF3B1 (11.2 [477], 15 [474], 7.93 [473,474,489])
SYK (OE [495–497])	TP53 (10.4 [477], 15 [474], 7.1 [489], 8.62 [473])
XPO1 (2.76 [473], 3.4 [489])	ZAP70 (OE [498,499])
Lower Grade Glioma (LGG)	
ARID1A (11 [500], 5.92 [473])	ARID1B (11 [500], 2.37 [473])
ATRX (42.6 [501,502])	BRAF (15 [503] , 1.85 [504,505])
CIC (20.12 [502])	EGFR (OE [506], A [507], 23.22 [504], 4.14 [473])
FUBP1 (10.65 [508])	IDH1 (77.51 [509,510])
IDH2 (3.55 [510])	NF1 (5.92 [473,504,509])
NOTCH1 (7.69 [473] , OE [511])	PIK3CA (6.51 [473], 10.03 [504,509,512,513])
PDGFRA [514,515]	PIK3R1 (5.92 [473,504,509,516])
PTEN (4.14 [473], 30.34 [504,509])	RB1 (1.78 [473,504,509])
SOX9 (OE [517,518])	TCF12 (3.55 [519])
TP53 (50.89 [473], 30.61 [504,509])	
Colorectal Cancer (CRC)	
APC (79.04 [520])	BRAF (16 [521], 4.7 [522], 3 [520,523])
EGFR (12-22 [524–526])	KRAS (40 [522], 43 [520,521,523,527])
FBXW7 (10 [520])	PIK3CA (14.5 [522], 15 [520,521,523,528])
SMAD2 (3.4 [520,529])	PTEN (14 [521], 4 [520,522,523,530,531])
SMAD3 (4.3 [520,529])	SMAD4 (8.6 [520,529,532])
SOX9 (3.49 [473], 4 [520,533])	TCF7L2 (9.17 [473,534], 12 [520])
TGFBR2 (3.49 [473], 2 [520,535])	TP53 (59 [520])
Papillary Thyroid Cancer (PTC)	
AKT1 (15 [536])	ALK (10 [536])
ARID1B (1 [537] [538])	BRAF (35.8 [539], 56.52 [473])
CTNNB1 (25 [536])	EGFR (5 [536])
EIF1AX (1.5 [537,540])	HRAS (20-40 [536,541])
KMT2C (1 [537] [538])	KRAS (20-40 [536,541])
NDUFA13 (15 [536])	NRAS (20-40 [536], 8.07 [473,541])
PIK3CA (1–2 [536])	PTEN (4.8 [536,542])
TG (2.7 [537])	TP53 (25 [536])
ZFHX3 (1.7 [537])	

Numbers in the brackets represent the expression (OE stands for overexpression) or mutation (A: amplification, number: percentage of mutation rate) of the gene.

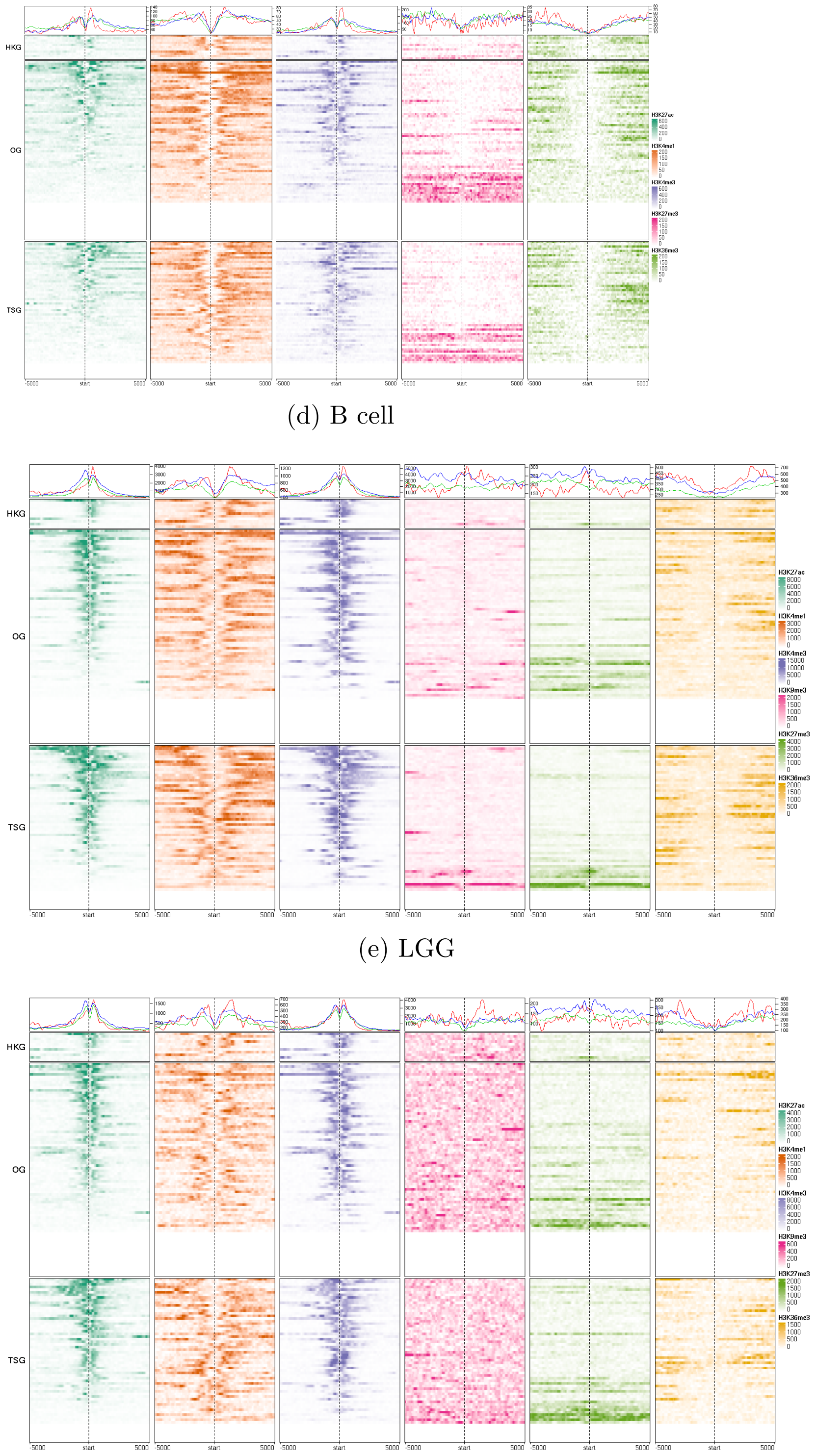


(a) PTC

(b) Thyroid

(c) CLL

Figure S1: Continued on next page.

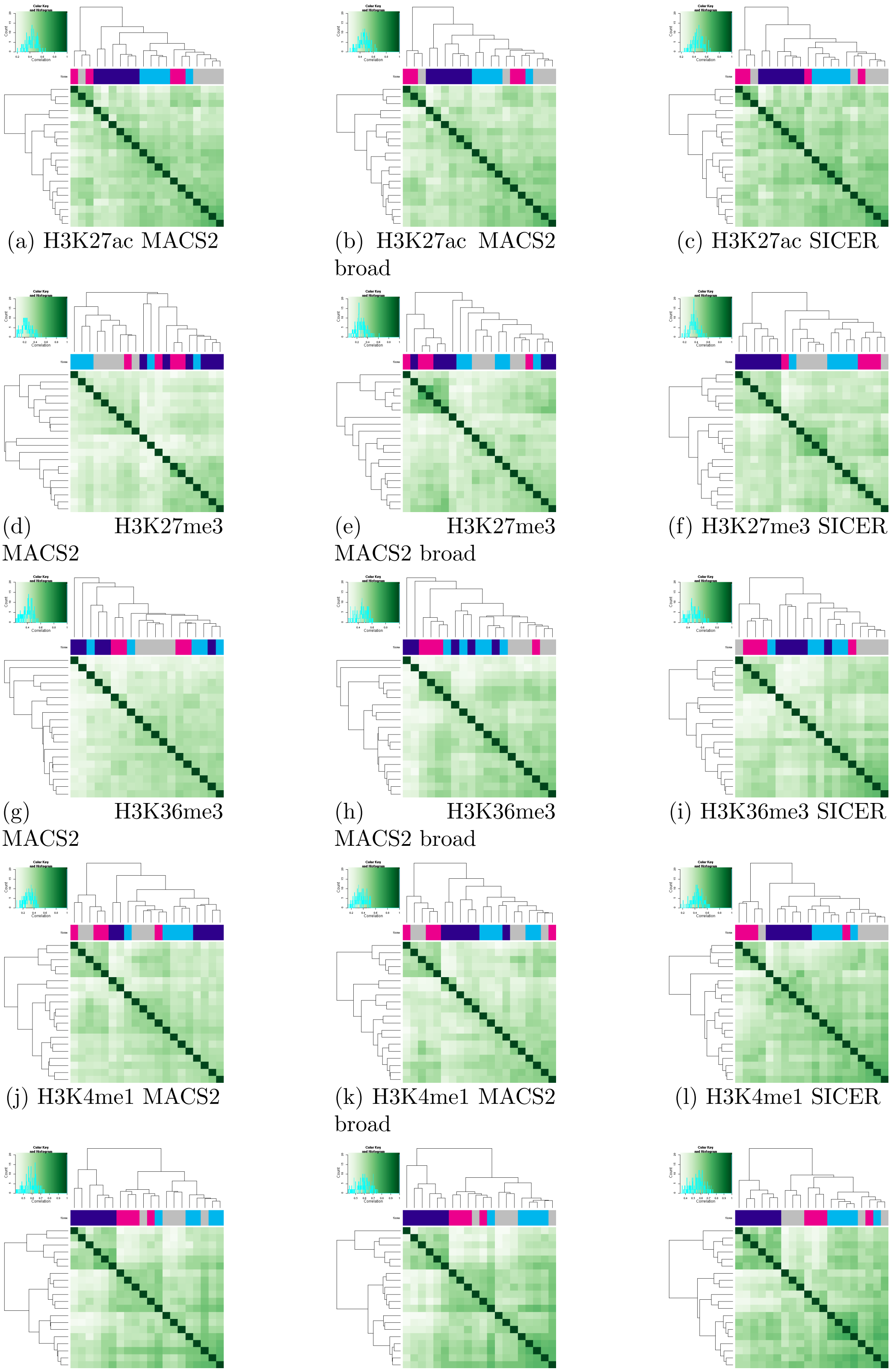


(d) B cell

(e) LGG

(f) Normal brain

Figure S1: Histone mark signals around oncogenes (OG), tumor suppressor genes (TSG), housekeeping genes (HKG)



(a) H3K27ac MACS2

(b) H3K27ac MACS2 broad

(c) H3K27ac SICER

(d) H3K27me3 MACS2

(e) H3K27me3 MACS2 broad

(f) H3K27me3 SICER

(g) H3K36me3 MACS2

(h) H3K36me3 MACS2 broad

(i) H3K36me3 SICER

(j) H3K4me1 MACS2

(k) H3K4me1 MACS2 broad

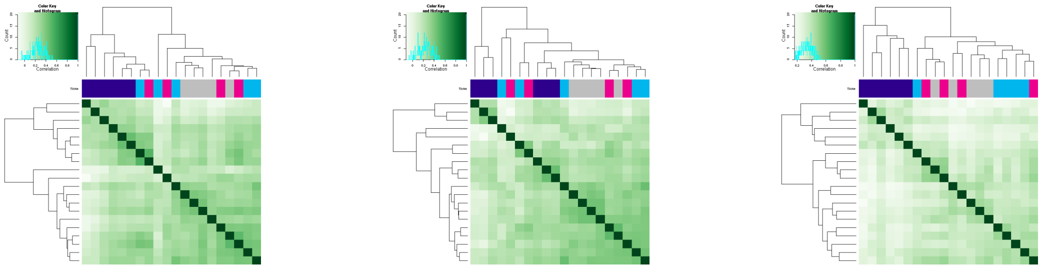
(l) H3K4me1 SICER

(m) H3K4me3 MACS2

(n) H3K4me3 MACS2 broad

(o) H3K4me3 SICER

Figure S2: Continued on next page.

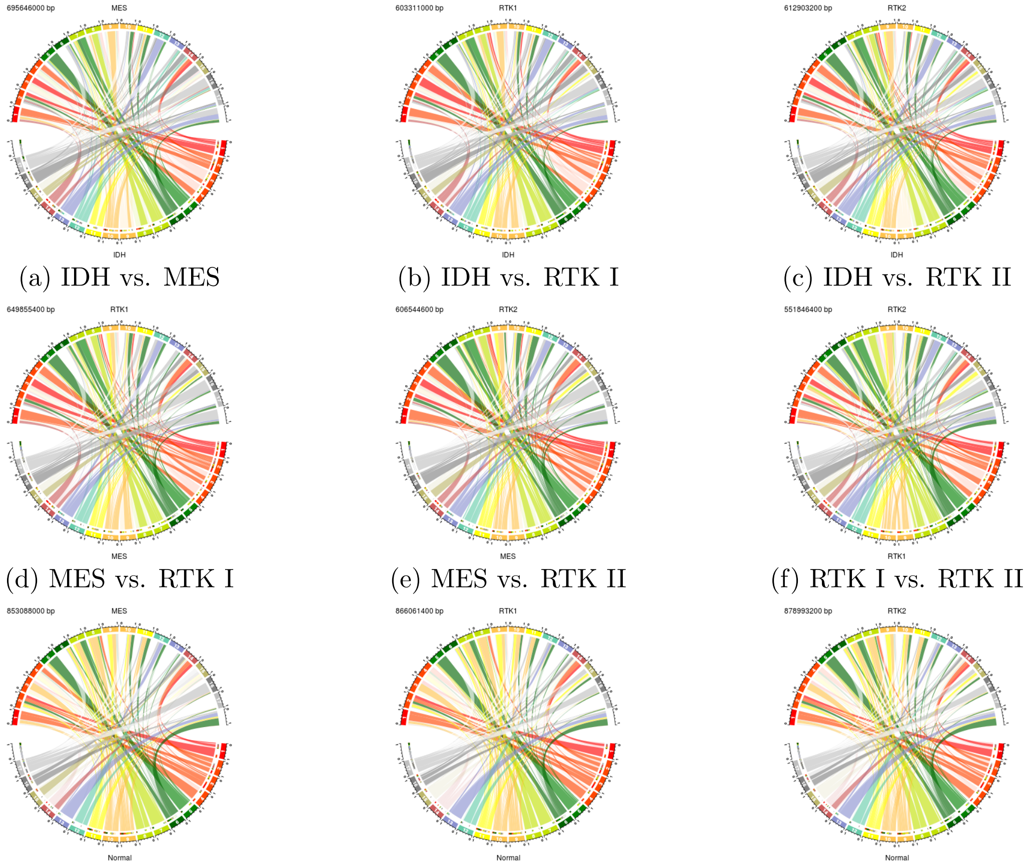


(p) H3K9me3 MACS2

(q) H3K9me3 MACS2 broad

(r) H3K9me3 SICER

Figure S2: Correlation heatmap using all bound sites for each histone mark.



(a) IDH vs. MES

(b) IDH vs. RTK I

(c) IDH vs. RTK II

(d) MES vs. RTK I

(e) MES vs. RTK II

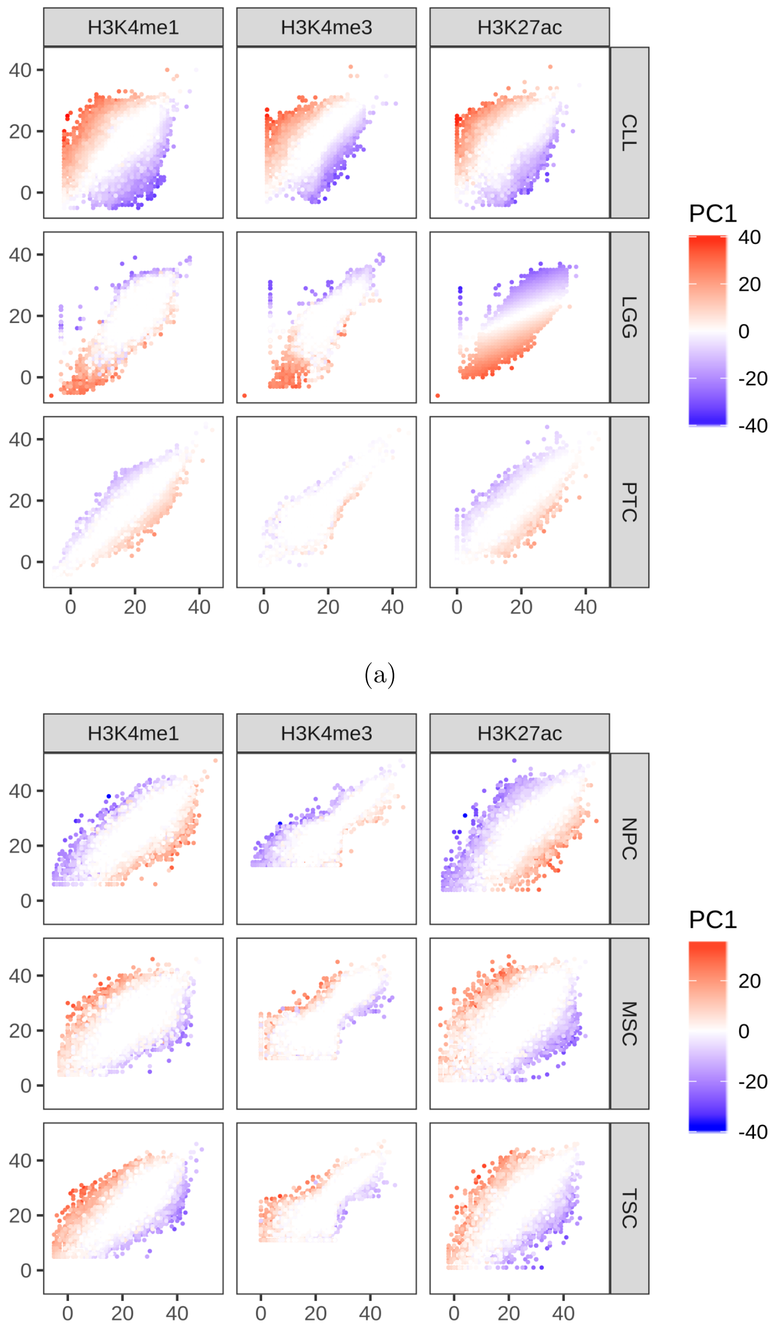
(f) RTK I vs. RTK II

(g) Normal vs. MES

(h) Normal vs. RTK I

(i) Normal vs. RTK II

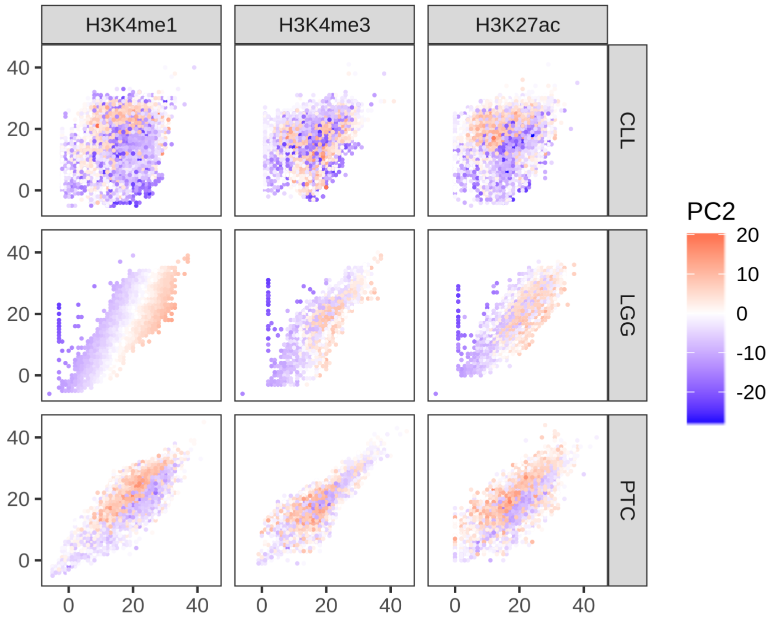
Figure S3: States transitions shown in percentage between each subtypes and normal brain



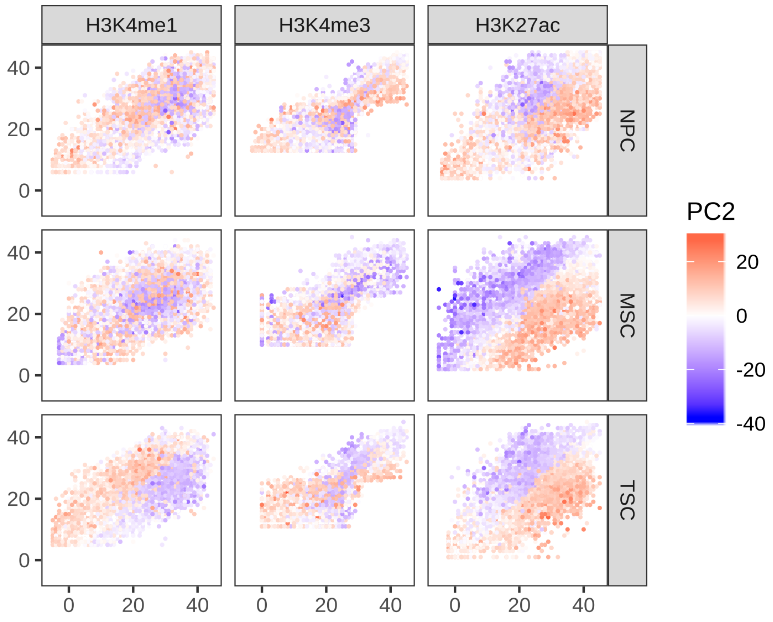
(a)

(b)

Figure S4: Continued on the next page.



(c)



(d)

Figure S4: dPC1 and dPC2 for three histone marks. The mean log intensities for every genomic locus in group one and group two from six test cases are plotted against each other, and the nodes are colored according to their dPC scales.

Acknowledgement

First, I would like to express my sincere gratitude to my supervisor Dr. Carl Herrmann for the continuous support of this project, for his patience, trust and great passion. I am deeply grateful to him for introducing me into the epigenetics world. He has always been there to talk to, provide suggestions, and give encouragement. His guidance helped me in all the time of writing this thesis and my Ph.D study.

I would like to thank the CRG group members and many people that helped through my Ph.D period, as well as the helpful and friendly atmosphere in eilslabs. I would like to thank (in alphabetical order) Ashwini Kumar Sharma, Andrés Felipe Quintero Moreno, Calvin Chan, Jing Yang, Michael Fletcher, Sebastian Steinhauser, Yonghe Wu, Zuguang Gu and many other people for their helpful comments and discussions, and special thanks to my supervisors Prof. Dr. Roland Eils and Prof. Dr. Benedikt Brors and Dr. Anne-Claude Gavin as an external TAC member. I would like to thank Dr. Sevin Turcan and Dr. Jürgen Pahle for serving as members on my oral exam committee and for giving me valuable comments on my thesis.

In addition, I want to thank my parents, my wife and my son who have always been standing by my side to support me. It is their caring and encouragement made me infused my research with love and conscientiousness.

Lastly, I would like to thank DKFZ for supporting three years of my Ph.D with the DKFZ Doctoral Fellowship, and the staff of HIGS as well as the alumni for fulfilling me a wonderful time in Heidelberg.