Inference of epigenetic subnetworks and expression-based analysis using a breast cancer dataset

by

Anqi Jing

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Computer Engineering

Department of Electrical and Computer Engineering

University of Alberta

Abstract

Changes in gene expression have been thought to play a crucial role in various types of cancer. With the advance of high-throughput experimental techniques, many genome-wide studies are underway to analyze underlying mechanisms that may drive the changes in gene expression. It has been observed that the change could arise from altered DNA methylation. However, the knowledge about the degree to which epigenetic changes might cause differences in gene expression in cancer is currently lacking. By considering the change of gene expression as the response of altered DNA methylation, we introduce a novel analytical framework to identify epigenetic subnetworks in which the methylation status of a set of highly correlated genes is predictive of a set of gene expression. By detecting highly correlated modules as representatives of the regulatory scenario underling the gene expression and DNA methylation, the dependency between DNA methylation and gene expression is explored by a Bayesian regression model with the incorporation of *g-prior* followed by a strategy of an optimal predictor subset selection. The subsequent network analysis indicates that the detected epigenetic subnetworks are highly biologically relevant and contain many verified epigenetic causal mechanisms. Moreover, a survival analysis indicates that they might be effective prognostic factors associated with patient survival time.

The alterations in gene expression are often ignored as stochastic noises, specifically those arising from variations in transcriptional regulation or biochemical modifications within cells. To evaluate if such alterations contribute

to cancer progression, we performed an expression-based analysis to detect exclusively expression-altered (EEA) genes, i.e., genes with altered expression not caused by genetic mutations, and we investigated the pattern of their aberrant expression in breast cancer. Based on these investigations, we found that the alterations in EEA genes are instigated by hypoxia-related molecular events, predominantly in two groups of genes that control chromosomal instability (CIN) and remodel tumor microenvironment (TME). We conclude that alterations are not stochastic and that hypoxia induces CIN and TME remodeling to permit further tumor progression.

Preface

The work presented in Chapter 3 was performed through a collaboration with Prof. Franco Vizeacoumar and his team at the University of Saskatchewan. With the initial concepts formed during discussions between Prof. Franco Vizeacoumar and Prof. Jie Han, I carried out the data analysis under their supervision. Prof. Franco Vizeacoumar and his students further performed the follow-up ingenuity pathway analysis and drug analysis (Chapter 3.3.4). Prof. Franco Vizeacoumar and Prof. Jie Han further revised the manuscript and supervised the research. The manuscript "Expression-based analyses indicate a central role for hypoxia in driving tumor plasticity through microenvironment remodeling and chromosomal instability" has been submitted to npj Systems Biology and Applications and is currently under review.

Acknowledgements

First, I would like to thank my supervisor, Prof. Jie Han for his guidance and help during my MSc program. Also I would like to thank Prof. Franco Vizeacoumar for his guidance in my project. The knowledge and attitude toward research I have learned from them will significantly benefit my future career.

Many thanks to all the students in our lab for their support and company: Honglan Jiang, Yidong Liu, Yuanzhuo Qu, Siting Liu, Mohammad Saeed Ansari and Oleg Oleynikov.

I want to express my gratitude to my parents Mr. Lei Jing and Ms. Liping Zhao. I am also very greatly thankful to my brother, Mr. Tian Jing. Their love and support really mean a lot to me.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Human cancer is one of the leading causes of morbidity and death around the world [87]. According to World Health Organization's report, approximately 10 million people are diagnosed with cancer and more than 6 million die of the disease every year [87]. It is well known that cancer is a dynamic disease driven by a series of abnormal genetic changes. These changes can occur at various levels and take many different forms, including the gain or loss of chromosomes, DNA (Deoxyribonucleic acid) point mutations, deletions and insertions, which can alter the function or stability of their protein products, further causing uncontrollable malignancy and damage to neighbour cells. With the development of sequencing technologies, a large number of projects based on cancer genomics takes the advantage of recent technologies to study the abnormalities in genes that may drive the development and growth of cancer, in order to improve the understanding of the biological mechanism of cancer and develop new methods for the diagnosis and treatment of cancer patients.

In the past decades, a large number of studies have been started to investigate and discover the genetic changes that could be associated with the growth and development of different types of cancer [15][11][101][70]. The results of these studies have illustrated the cancer genome landscape of genetic changes and provided us a fundamental understanding of the molecular bias of multiple cancer types [28][42]. For instance, numerous studies have identified the high frequency of mutations in the HER2 gene and suggested that the HER2 gene might act as a prognostic factor in multiple types of cancer, including

lung [64][92], bladder [70] and breast [15][6]. In addition to the study on genetic mutations, a large scale of projects has sought to understand the role of copy number alteration (gain or loss of chromosomes) in cancer progression [9][105]. For example, a study [9] has presented a high-resolution analysis of somatic copy number alterations from thousands of tumor samples across multiple cancer types, which demonstrated a strong tendency for significant somatic copy number alterations in one cancer type to be also found in several others. A recent finding [18] has revealed that tumors harboring a high level of copy number alterations are unlikely to respond to immune checkpoint blockade immunotherapies, an increasingly promising treatment option for many cancers. This finding may help doctors to recommend tailored therapies and predict patient outcome.

The classic view that cancer progression is driven by genetic changes including mutations and chromosomal abnormalities, and later on epigenetic alterations have been considered as crucial in the progression of cancer. Epigenetic alteration refers to the functionally relevant changes to the genome without changing DNA sequences [91], including DNA methylation, histone modification, etc. Such epigenetic alterations have been investigated in numerous studies [49][8], which revealed that they are likely to be responsible for the reduced or increased expression in DNA repair genes and be the cause of genetic instability characteristic of cancers in early cancer progression. The change in DNA methylation leading to an aberrant gene expression has been considered to play a crucial role not only in cellular development and differentiation but also in disease progression [91]. Many studies have been conducted for the identification of aberrant DNA methylation sites in cancer [91][36][94], but there are fewer studies on the degree to which epigenetic changes might cause differences of gene expression in cancer. Thus it motivates us to discover the association between epigenetic changes and altered gene expression and investigate the varying level of DNA methylation that could drive differences in gene expression. We expect that such investigations could shed light on novel biological mechanisms and reveal potential therapeutic targets.

In addition, the pattern of genetic changes including mutations and copy

number alterations has been widely and deeply explored in multiple types of cancer. We noticed that relatively fewer studies regarding aberrant gene expression in cancer were performed compared to the study on genomic changes. Since the changes in gene expression were often ignored as stochastic noise, specifically those arising from variations in transcriptional regulation or biochemical modifications within cells, we performed a computational analysis aiming to address the questions: what are the most significant changes that occur within the transcriptome (i.e., the set of all mRNA) of cancer cells and how do they contribute to tumor development?

This thesis consists of two main sections. In Chapter 2, we introduce a novel analytical framework for the discovery of associations between DNA methylation and gene expression. Aberrant gene expression is considered as the response of DNA methylation predictors and epigenetic subnetworks are identified in which the methylation status of a set of highly correlated genes is predictive of a set of gene expression. The subsequent pathway and network analyses indicate that the detected epigenetic subnetworks are highly correlated with cancer genes and cancer-related pathways. Multiple direct causal epigenetic mechanisms are detected in our study and verified in studies reported in the literature. A survival analysis reveals that the subnetworks might be effective prognostic factors associated with survival time. These results indicate that the detected epigenetic subnetworks could be a starting point to uncover underlying epigenetic mechanisms in breast cancer.

In Chapter 3, we perform a computational analysis based on the aberrant gene expression in breast cancer, starting with the detection of exclusively expression-altered (EEA) genes. Then the pattern of EEA genes across breast cancer samples is investigated to learn how they contribute to cancer progression. We found that the alterations in EEA genes are not stochastic noises, and that those alterations in early steps of cancer progression are instigated by hypoxia-related molecular events, predominantly in two groups of genes that control chromosomal instability (CIN) and remodel tumor microenvironment (TME).

In summary, the work reported in this thesis makes new contributions

to cancer biomarker research. In Chapter 2, the detected epigenetic subnetworks that contain verified epigenetic mechanisms may help uncover underlying mechanisms in breast cancer. In Chapter 3, the analysis on EEA genes highlights a therapeutic potential of targeting CIN and TME events in triple negative breast cancer (TNBC) tumors. These results could help to discover underlying cancer mechanisms and ultimately improve cancer diagnosis and therapy.

Chapter 2

Discovery of epigenetic subnetworks

2.1 Introduction

2.1.1 Motivation

With the advance of high-throughput experimental techniques, a tremendous amount of genomic-wide omics data has been available, which revolutionizes the study of cancer by making it possible to discover potential biomarkers and biological mechanisms at the genome level. The change in the expression level of gene regulation has been considered to play an important role in various types of cancers. To understand the roles of genes involved in cancer, a comparison of gene regulation is typically performed for tumor and control conditions to explore meaningful patterns and relationships in biological data. Differential expression analysis is usually conducted by testing the significance of changes in the expression level of genes between two conditions such as disease and control. It has been successful in discovering biomarkers associated with the cancer phenotype and cancer progression [101][54][80]. For instance, Welsh et al. [101] analyzed the patterns of gene expression between normal and tumor samples of prostate cancer, which revealed important genes acting within biochemical pathways and encoding diagnostic potential molecules. LaTulippe [54] identified highly and significantly differentially expressed genes in multiple functional categories revealing critical cellular activities that contribute to clinical heterogeneity and provide diagnostic and therapeutic targets. Although

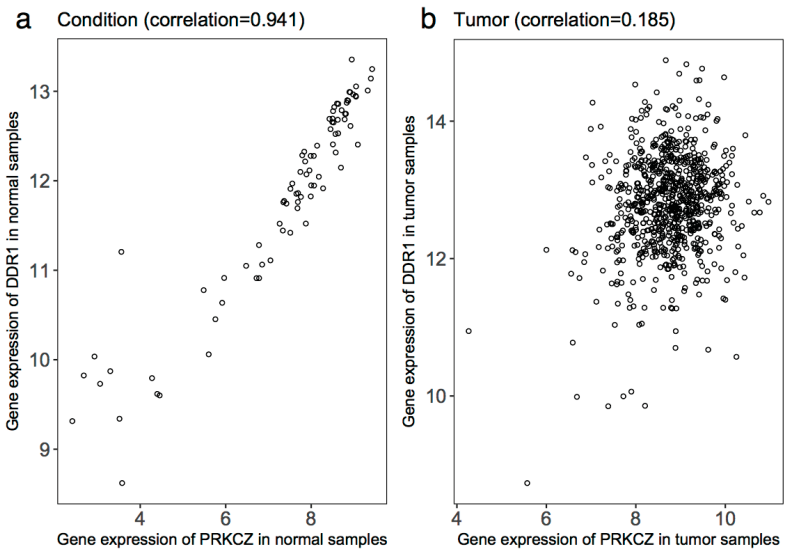


Figure 2.1: An example of differential co-expression between normal and tumor samples in breast cancer. Each dot denotes one sample. Figure (a) presents gene expression levels of DDR1 and PRKCZ in normal samples and (b) presents their gene expression levels in tumor samples. A strong co-expression pattern between gene DDR1 and gene PRKCZ can be observed in normal samples (a) but not in tumor samples (b).

differential expression studies have been successful in discovering cancer genes, gene expression datasets contain more information than that differential expression analysis can extract [24]. Recent investigations have gone beyond differential expression analysis and tried to identify genes involved in a differential coexpression pattern. Differential coexpression refers to the changes in a gene-gene correlation between two conditions (Fig. 2.1) and it can discover dysfunctional regulations that would not be discovered by differential expression analyses [14][57].

Changes in gene expression in tumors typically are driven by two main factors. One is the combination of mutations of transcription factor affinities binding to DNA regulatory sequences, including the copy number variation, point mutation and chromosomal alterations. Aside from that, epigenetic mechanisms are the other main factor that causes the aberration of expression levels of genes. Epigenetics is a relatively new research field and there are fewer studies on epigenetic mechanisms that underlie gene regulation than studies on genetic mutations. DNA methylation is one epigenetic mechanism, which can

alter gene expression by causing the stable silencing or activation of particular genes without changing DNA sequences. It remains throughout cell divisions and can be inherited by daughter cells lasting for multiple generations. It is of great interest to cancer study since it is potentially reversible and could be returned to normal function with appropriate drugs, which makes it an excellent target for anticancer therapies [33].

It has been observed that a large proportion of differential gene expression and differential co-expression relationships could arise from altered DNA methylations [91]. The change in DNA methylation has led to differential patterns of gene expression, considered as crucial in not only cellular development and differentiation but also in disease progression [91]. Numerous studies have been conducted to identify aberrant DNA methylation sites in cancer [91][36][94]. However, we have relatively little knowledge about the degree to which epigenetic changes might explain differences in gene expression levels in cancer. Thus it motivates us to discover associations between altered DNA methylation and aberrant gene expression and investigate the varying level of DNA methylation that could drive differences in gene expression. We expect such investigation could shed light on novel biological mechanisms and reveal potential therapeutic targets.

2.1.2 Related work

The development of high-throughput profiling in biological researches enables the vast amount of public large-scale genomic-wide DNA omics data available for various types of cancer, which in turn provides the opportunity to analyze the epigenetic mechanisms at the whole genome level. A recent database TCGA (The Cancer Genome Atlas) [100] has profiled and collected multidimensional omics data at DNA, RNA, protein and epigenetic levels for hundreds of clinical tumors, making the integrative analysis of epigenetic mechanisms at the whole genome level possible. There is now a growing biological interest in the analysis of methylation profiles to extract DNA methylation patterns at a statistical level. For example, Hinoue et al. [34] identified four DNA methylation-based subgroups of colorectal cancer exhibiting characteristic ge-

netic and clinical features, which provided novel insights regarding the role of subgroup-specific DNA hypermethylation in gene silencing. Varley et al. [95] provided an atlas of DNA methylation across diverse and well-characterized samples and analyzed dynamic DNA methylation patterns in 82 cell lines and tissues, which discovered the role of DNA methylation in gene regulation and disease. Gevaert et al. [26] developed an algorithm *MethylMix* to identify differentially methylated genes and applied it to 12 cancer sites and performed a pan-cancer analysis by combining all cancer sites. In addition, they identified novel methylation-driven subgroups with clinical implications reflecting new similarities across malignantly transformed tissues.

Although the pattern of DNA methylation has been extensively investigated, how gene modules or pathways are deregulated through DNA methylation is far from understood. More specifically, approaches for simultaneously analyzing methylation and gene expression data need to be developed to discover how DNA methylation deregulates gene expression in cancer. West et al. [102] have proposed a method *EpiMod* to address whether differential DNA methylation is associated with a given phenotype of interest in the context of a protein interaction network. It started from constructing a weighted co-methylation network in the context of the human interactome model in which the edge weight represents the association between DNA methylation profiles in two connecting genes, and subsequently applied a local community detection algorithm (spin-glass) to identify differential methylation hotspots around differentially methylated genes by maximizing the sum of weights. They demonstrated the existence of epigenetic modules associated with phenotypes by applying the method to cancer and ageing. However this approach was restricted to the DNA methylation data. Jiao et al. [40] proposed a new approach *FEM* by expanding *EpiMod* by defining the edge weight as the combination of two statistical associations of co-methylation and co-expression. Encoding the two associations into edge weight allowed it to identify epigenetically deregulated modules in which genes showing coordinated differential methylation and differential expression. They identified the previously known deregulated pathway driving endometrial cancer development and an up-streamer of the

well-known progesterone receptor tumour suppressor pathway. It is well acknowledged that the existence of anti-correlations between DNA methylation and gene expression, i.e., the changes in DNA methylation cause the silencing of gene expression. The method *FEM* detected the anti-correlated epigenetically correlated modules, however a recent study [95] found the positive correlation between the two types of data that the increased methylation is associated with the higher level of gene expression. By assuming the existences of both negative and positive correlations, Ma et al. [63] has proposed a multiple network algorithm *EMDN* by constructing differential co-expression and differential co-methylation networks respectively and subsequently identified the common modules appeared at both networks. *EMDN* can recognize both positively and negatively correlated modules. They demonstrated that the identified modules can serve as biomarkers to predict breast subtypes and estimate the survival time of patients. However, Wang et al. [99] pointed out that only a small proportion of the alteration in DNA methylation leads to a corresponding change in gene expression at the same gene, therefore identifying gene modules restricted to the association between DNA methylation and gene expression at the same or adjacent genes may miss important links between the two changes. To overcome this limitation, they have proposed a multivariate regression framework *NsRRR* to identify relationships between any varying level of DNAmethylation and changes in expression of any genes. By considering expression levels of genes as responses of DNA methylation levels, they extracted a group of genes in which the expression level of a subset of genes could be regressed on the DNA methylation level of remaining genes.

Inspired by *NsRRR*, to further understand the relationship between gene expression and DNA methylation, we propose a novel framework to identify epigenetic subnetworks consisting of a set of genes with aberrant gene expression or aberrant DNA methylation level, in which the altered gene expression are deregulated by DNA methylation. Different from *NsRRR* [99] which evaluated the association at the individual gene level, we extract a high level representation of regulatory scenarios underling both gene expression and DNA methylation in the form of gene modules, and then quantify the association

between DNA methylation and gene expression at module level by a regression model. Since module-level analysis could increase the association signal and provide insight into the biological behaviours [97], we expect that regression analysis at module level could provide complementary information to the analysis at single gene level [99] and shed light on the discovery of new epigenetic mechanisms.

2.1.3 Contribution

We propose an analytical framework for the discovery of epigenetic subnetworks in which aberrant gene expression is deregulated by DNA methylation levels. More specifically, we consider aberrant gene expression as the response of DNA methylation predictors. It starts with the discovery of predictor and response modules on a weighted differential network, and subsequently quantify the relationship between DNA methylation predictor modules and gene expression response modules via a Bayesian regression model with the incorporation of known protein-interaction priors. For each response module, the best subset of predictor variables are selected based on Bayesian information criterion (BIC). Statistical significance tests and biological relevance analysis are performed to assess the performance of the model.

The contribution lies in the following points:

(1) A novel method is proposed to detect epigenetic subnetworks by incorporating prior biological knowledge as g-priors [106], a type of priors for the regression coefficients in Bayesian regression model. It detects more significantly correlated epigenetic subnetworks than the alternative model without prior information. It shows that encoding biological network information as g-priors successfully guided the selection of epigenetic subnetworks.

(2) The detected epigenetic subnetworks are more enriched in biological processes and signalling pathways compared with those detected by EMDN, which indicates that evaluating the association between gene expression and DNA methylation at the module level would increase the association signal and shed light on the underlying mechanisms.

(3) The pathway analysis indicates that the detected epigenetic subnetworks

are highly correlated with cancer genes and cancer-related pathways. Moreover, the subnetworks contain multiple direct causal mechanisms which are verified in other scientific papers, which indicates the capability of our method to detect the true epigenetic mechanisms. A survival analysis reveals that the subnetworks might be effective prognostic factors associated with patient survival time. Overall, these results indicate that the detected epigenetic subnetworks could be a starting point to uncover underlying epigenetic mechanisms in breast cancer and reveals potential therapeutic targets.

2.1.4 Outline

Section 2.2 describes preliminary information and background knowledge, including the introduction to dataset, significance test, and models used in our method. In 2.3, we introduce the framework in detail, as well as evaluation metrics. In 2.4 and 2.5, we present one simulation study and one case study, respectively. In 2.6, we make a conclusion of this chapter.

2.2 Background

The development of our method combines several threads of previous researches. In this chapter, we first introduce the preliminary information about the gene expression and DNA methylation data, as well as the approach to differential analysis. Statistical significance tests are used, so we introduce the basis of significance test and discuss several types of significance test. Moreover, we describe one existing algorithm for nonnegative matrix factorization, which is employed for the detection of gene modules in this framework. The development of this method depends on fundamental concepts of Bayesian regression within the context of Bayesian inference, therefore in the last section we discuss the theory of Bayesian regression model.

2.2.1 Introduction to dataset

High-throughput experimental techniques have generated large-scale omics data to help the understanding of gene function and the enhancement of disease treatment. In recent years, a tremendous amount of genomic data for various types of cancers has been collected by the Cancer Genome Atlas (TCGA). It uses different techniques to collect and analyze the cancer data from thousands of patients for 30 different types of cancer, and the techniques include gene expression profiling, copy number variation profiling, SNP genotyping, genome wide DNA methylation profiling, microRNA profiling, and so on [100]. In this chapter, we utilize two types of data, DNA methylation and RNA sequencing, of 786 breast cancer samples from TCGA to study the epigenetically deregulated subnetworks in breast cancer.

Gene expression

RNA sequencing (RNA-seq) utilizes the next-generation sequencing to reveal the presence and quantity of RNA in a biological sample at a given moment [98]. The sample RNA first undergoes fragmentation [67] and subsequently is sequenced by a sequencing machine such as Illumina Genome Analyzer and HiSeq in a massively parallel fashion. After aligning short reads to the genome,

the read coverage depth can be counted to measure the expression level [22]. In this thesis, we collect HiSeq RSEM gene-normalized RNA-seq data from TCGA[100] as gene expression data. In this work, we use Hi-Seq RNA-seq data as gene expression data provided by the database TCGA [100] and quantified at the gene level using RSEM (RNA-Seq by Expectation Maximization), which is a software package for estimating gene and isoform expression levels for RNA-seq data [58]. In total, we collect the expression data for 20531 genes in 786 tumor samples and 84 normal samples.

Data preprocessing After collecting gene expression data, the gene expression matrix $X[x_{ij}]$ is obtained with samples in columns and genes in rows, where x_{ij} represents the gene expression level of gene i in sample j. Then the data preprocessing is performed. First, genes with missing values in more than 30% of samples are removed. The remaining missing values are imputed by using k-nearest neighbours (KNN) averaging method [93]. Supposing the expression value of genen i in sample j (x_{ij}) is missing, KNN would find k (an integer, typically small) nearest genes in sample j with the expression value most similar to gene i in other samples by computing the Euclidean distance. The missing value of x_{ij} is computed by averaging the expression levels of k closest genes [93]. In this work, k is set to 10.

Differential analysis of gene expression data We identify genes that are differentially expressed between conditions of tumor and normal while constructing the differential gene expression network. Two criteria are applied on the selection of differentially expressed genes: fold change and t-test. Fold-change is a biological assessment of changes in gene expression between tumor and normal conditions and the fold-change for gene g is estimated as:

$$fold\text{-}change_g = log_2 \frac{GE_g^{tumor}}{GE_g^{normal}}, \tag{2.1}$$

where GE_g^{tumor} and GE_g^{normal} denote the average expression of gene g in tumor and normal samples, respectively.

The t-test statistic is used in hypothesis testing to determine if the gene is significantly differentially expressed, which is defined as:

$$t_g = \frac{\hat{\beta}_g - \beta_0}{se_g(\hat{\beta}_g - \beta_0)}, \tag{2.2}$$

where $\hat{\beta}_g$ is the contrast estimator - the fold change of gene g between conditions of tumor and normal, and β_0 is set to 0 in our case. The $se_g(\hat{\beta}_g - \beta_0)$ stands for the standard error of differences in average expression values of gene g between tumor and normal samples. However, t-test statistic has a drawback that the vartiance estimates can be skewed by genes having a very low variance [39] which leads to a large t-statistic. Thus, those genes with a very low variance can be falsely detected as differentially expressed. Moreover, t-test has a low statistical power for studies with few samples [68]. Consequently, an alternative strategy limma [85] has been proposed and widely accepted to improve the power and accuracy of variance estimation. It uses Empirical Bayes to derive the moderated t-statistic by borrowing information from the population of other genes to aid with the inference about each individual gene. It defines the variation as:

$$\tilde{se_g^2} = \frac{d_0 se_0^2 + d_g se_g^2}{d_0 + d_g}, \tag{2.3}$$

where se_0^2 is the overall estimate variance for all genes, se_g^2 the deviation variation for gene g, $\frac{d_0}{d_0+d_g}$ the weight coefficient associated with all genes, and $\frac{d_g}{d_0+d_g}$ associated with gene g. The moderated t-statistic is defined as:

$$\tilde{t_g} = \frac{\hat{\beta}_g}{\tilde{se}_g\sqrt{v_g}}, \tag{2.4}$$

where v_g denotes the corresponding diagonal element of the estimated covariance matrix. Thus p-value is obtained from the moderated t-statistic implemented by the package limma [85]. Since lots of tests for different genes are performed simultaneously, the multiple testing problem occurs and results in a large number of false positives. Thus, an adjustment for p-value is performed for addressing the multiple testing problem by Benjamini and Hochberg's correction [7], which introduces an adjusted p-value for each test to reduce the number of false discoveries.

We combine the two assessments fold-change and p-value to detect differentially expressed genes from both biological and statistical points of views. By convention, the threshold for p-value is commonly set to 0.05 or 0.01 [75]. We set a restrictive threshold for the adjusted p-value as 0.01 to select the significantly differentially expressed genes. Genes are considered as differentially expressed genes if p-values are less than or equal to 0.01 and the absolute value of fold-change is larger than or equal to 2.

DNA methylation

Summarize DNA methylation value at the gene level DNA methylation is the process by which the methylation groups are added to DNA molecules [37], which can change the activity of a DNA segment without changing the DNA sequence. Illumina Methylation Assay provides a robust profiling platform which uses the "BeadChip" technology to generate a comprehensive genome-wide profiling of human DNA methylation data [37]. The HumanMethylation 450 BeadChip employs both Infinium I and Infinium II assay technologies to enhance the breadth of coverage. It measures the methylation level over 450k sites per sample at single-nucleotide resolution. Infinium I applies two bead types, one for methylated allele and the other for unmethylated allele. Infinium applies II applies one bead type with a unique type of probe allowing detection of both alleles [21]. Extracted from the assays, a β value is defined as the ratio of intensities between methylated and un-methylated alleles across gene regions with sites in different regions, including 5'UTR, first exon, gene body, TSS200, TSS1500 and 3'UTR, as shown in Fig. 2.2. We use the Illumina Infinium 450k DNA methylation data for breast cancer from TCGA.

In Illumina 450K DNA methylation data, many probes may map to different regions associated with the gene. Jiao et al. [40] proposed a scheme designed for Illumina 450k DNA methylation data to summarize DNA methylation values at a gene level, by assessing which methylation probes are most predictive of the gene expression state. Then it was validated by demonstrating that it can successfully retrieve known genes and gene modules. Ma et

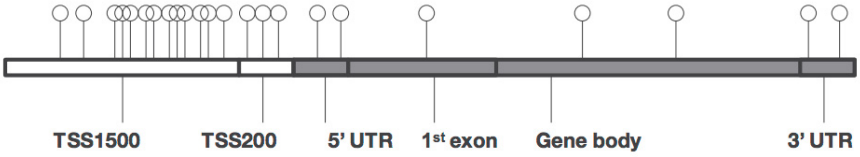


Figure 2.2: Illuminal Infinium 450k provides coverage across gene regions. It measures the methylation level across gene regions with sites in the TS1500, TS200, 5'UTR, first exon, gene body, 3'UTR. (This figure was taken from the datasheet [37].)

al. [63] followed this strategy to assign DNA methylation values to genes in a breast cancer study and successfully identified known gene modules. Therefore we follow this scheme to summarize DNA methylation values at the gene level. Specifically, for a given gene with TSS200 probes, the average β-value of probes mapping to TSS200 is used as the DNA methylation value. If no probes mapped to TSS200, the average β-value of probes mapping to the first exon is considered. If such probes are not available, the average value of probes mapping to TSS1500 is used.

Differential analysis of DNA methylation level Similar to differential analysis of gene expression data, we perform an Empirical Bayes t-test to identify significantly methylated genes, i.e., genes with p-value less than 0.01. Since no significant fold-change between tumor and normal samples in DNA methylation data, we only use p-values to identify differentially expressed genes from the statistical point of view without the assessment of fold-change.

2.2.2 Significance test

Typically, the statistical significance test begins with the statement of null hypothesis that the observed result occurs by chance. A significance test is designed to evaluate the strength of the evidence against the null hypothesis [16]. A result having a statistical significance indicates that it is very unlikely to have occurred given the null hypothesis [16]. It can be measured by the p-value, referring to the likelihood of obtaining the result when the null

hypothesis is true. Therefore a lower p-value indicates that the result has a higher level of significance. A predefined significance level α indicates the probability of the study rejecting the null hypothesis. It is typically defined as 5% or lower [75]. The result is statistically significant if p-value $< \alpha$ [16].

Correlation testing via Fisher transformation

To detect significant differential co-expression/co-methylation relationships between two conditions in gene expression or DNA methylation data, we perform correlation testing via Fisher transformation. We use Pearson correlation coefficient ρ to evaluate the correlation between two genes X and Y in one condition:

$$\rho_{XY} = \frac{cov(X,Y)}{\sigma_x \sigma_y}, \tag{2.5}$$

where $cov(X,Y)$ stands for the covariance between variables X and Y and σ indicates the standard deviation of the respective variable. The Fisher transformation of the correlation coefficient ρ_{XY} is defined as follows:

$$F(\rho_{XY}) = \frac{1}{2} ln(\frac{1+\rho_{XY}}{1-\rho_{XY}}). \tag{2.6}$$

If X and Y have a joint bivariate distribution or the number of samples n is large enough, then $F(\rho_{XY})$ is approximately normally distributed with the mean $\frac{1}{2} ln(\frac{1+\rho_{XY}}{1-\rho_{XY}})$ and standard error $\frac{1}{\sqrt{n-3}}$ [55], where n is the number of samples.

To test whether the correlations ρ_{XY} in two different conditions are the same or different, a test statistic Z is defined as:

$$Z = \frac{F_1(\rho_{XY}) - F_2(\rho_{XY})}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}}, \tag{2.7}$$

where $F_1(\rho_{XY})$ and $F_2(\rho_{XY})$ denote the Fisher transformation of ρ_{XY} in the two conditions, and n_1 and n_2 represent the number of samples in two conditions. It follows the distribution:

$$N(F_1(\rho_{XY}) - F_2(\rho_{XY}), \sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}). \tag{2.8}$$

Under the null hypothesis, i.e., the correlations in the two conditions are the same, the test statistic Z follows $N(0,1)$. Thus the p-value can be calculated with a two-tailed test:

$$p - value = 2 \times (1 - \phi(|Z|)), \tag{2.9}$$

where ϕ is the normal cumulative distribution function.

Hypergeometric test

To evaluate the biological relevance of identified epigenetic subnetworks to the known gene reference sets, a gene set enrichment analysis is performed by a hypergeometric test. Reference sets are obtained from multiple databases and genes in the same reference set are typically involved in the same biological pathway. Let g indicate the number of genes in the reference set, f indicate the gene population size, and d denote the number of genes in the detected subnetwork. The number of genes in the overlap between the subnetwork and the reference set is denoted by n. The variable n follows a hypergeometric distribution under the null hypothesis that the identified epigenetic subnetwork is irrelevant to the reference set. The p-value measuring the significance of enrichment is the tail probability of observing n or more genes in the reference set [78]:

$$p - value = \sum_{k=n}^{min(g,d)} \frac{\binom{g}{k}\binom{f-g}{d-k}}{\binom{f}{d}}. \tag{2.10}$$

Given an epigenetic subnetwork, if the p-value is less than 0.05, we consider it to be significantly enriched in the reference gene set.

2.2.3 Nonnegative matrix factorization

Clustering property

Nonnegative matrix factorization (NMF) is a group of algorithms that can factorize the nonnegative matrix $\mathbf{V}$ into nonnegative matrices $\mathbf{U}$ and $\mathbf{W}$ with lower ranks:

$$\mathbf{V} = \mathbf{UW}. \tag{2.11}$$

The approximation of $\mathbf{V}$ can be achieved by minimizing the Forbenius norm distance between $\mathbf{V}$ and $\mathbf{UW}$:

$$min_{\mathbf{U},\mathbf{W}}||\mathbf{V}-\mathbf{UW}||_F,\ subject\ to\ \mathbf{U}\geq 0,\mathbf{W}\geq 0, \tag{2.12}$$

where $\mathbf{V}\in\mathbb{R}_+^{m\times n}$, $\mathbf{U}\in\mathbb{R}_+^{m\times k}$, $\mathbf{W}\in\mathbb{R}_+^{k\times n}$, $\mathbb{R}_+$ denotes the nonnegative real numbers and $||\cdot||_F$ denotes the Forbenius norm. Typically $k<min\{n,m\}$ and it is assumed to be much smaller than n or m. Assuming that n samples are represented as columns in $\mathbf{V}=[v_1,v_2,...,v_n]$, NMF clusters the columns of input data $\mathbf{V}$ into k clusters. The columns of $\mathbf{U}$ represent the basis of a latent k-dimensional space and columns of $\mathbf{W}$ provide the representation of $v_1,v_2,...,v_n$ in the latent space [48]. The clustering assignment of each sample can be obtained by $\mathbf{W}$, i.e., the largest entry in the corresponding column $\mathbf{W}$ is used to assign membership labels for each data column.

Similarity matrix factorization

The standard NMF could achieve good clustering results when the input data have a linear cluster structure. However, it cannot detect a nonlinear cluster structure. To overcome this issue, Kuang et al. [48] proposed *SymNMF* dealing with the case where data are embedded in a nonlinear relationship. The algorithm takes the similarity matrix $A_{n\times n}$ as input, where the element measures the pairwise similarity between data. The formulation of the nonnegative similarity matrix is represented as:

$$min_{H\geq 0}||A-HH^T||_F, \tag{2.13}$$

where H is a nonnegative matrix with the size $n\times k$ and k is the predefined number of clusters. Similar to the standard NMF, the approximated matrix H captures the cluster structure and the clustering assignment for the i_{th} data is the largest value in the i_{th} row of H.

SymNMF is developed on the basis of the projected Newton algorithm [10]. The projected Newton algorithm aims to find the local minimum of the objective function by using the gradient descent with the inverse of a Hessian matrix at the current point as the search direction. In the gradient descent,

one goes from the current point towards the negative of the gradient of the function, which leads to the fastest decrease of the objective function. The incorporation of the Hessian matrix allows the method to use second-order information when scaling the gradient, but it results in a high computational complexity. *SymNMF* made two improvements on the projected Newton algorithm to reduce computation cost by delaying the update of the scale matrix and setting the scale matrix as a block diagonal matrix. In this work, we aim to find the module structure with a high density in which genes may be involved in the same regulatory pattern. Two similarity matrices for both gene expression and DNA methylation data are constructed to measure the probability that genes belong to the same module. *SymNMF* is applied to the similarity matrices to find the module structure.

2.2.4 Bayesian regression

Bayesian linear regression with prior

Bayesian linear regression is a linear regression approach supplemented by additional information in the form of a prior probability distribution within the context of Bayesian inference. The standard regression model can be considered as the explanation of response y_i, where $i = 1, ..., n$, using a given $k \times 1$ predictor vector x_i:

$$y_i = x_i^T \beta_i + \epsilon_i, \tag{2.14}$$

where β_i is the vector of regression coefficients with size of $k \times 1$, and ϵ_i is the prediction error that is independent, identically normally distributed with $N(0, \sigma^2)$. The parameter β_i can be estimated by minimizing the sum of the squared residuals (SSR). SSR is a measure of model fit quantifying differences between observed and predicted responses:

$$SSR(b) = \sum_{i=1}^{n} (y_i - x_i^T \beta_i)^2 = (y - Xb)^T (y - Xb), \tag{2.15}$$

where X is an $n \times k$ design matrix in which each row is a predictor vector x_i^T, y is the response vector comprised of $[y_1, ..., y_n]^T$, and b is an estimated vector with size $k \times 1$. The global minimum can be found by taking the first order

partial derivative of SSR with respect to b and setting the equation to zero:

$$0 = \frac{dSSR}{db}(\hat{\beta}) = \frac{d}{db}(yy' - b'X'y - y'Xb + b'X'Xb)|_{b=\hat{\beta}} = -2X'y + 2X'X\hat{\beta}; \tag{2.16}$$

$$\hat{\beta} = (X^T X)^{-1} X^T y. \tag{2.17}$$

The estimator $\hat{\beta}$ obtained by minimizing SRR is referred to as ordinary least squares (OLS) estimator.

In contrast to the standard linear regression, the prior probability distributions for the unknown parameters (β, σ^2) are employed in Bayesian regression. The prior probability distirbutions are usually called priors, representing the prior belief about parameters before any evidence is taken into account. In this work, we encode the knowledge of biological relatedness between the response and the predictor as Zellner's g-prior [106] to guide the estimation of regression coefficients.

The g-prior for the regression coefficients β follows a multivariate normal distribution with the prior mean β_0 and the data dependent covariance matrix:

$$\beta|\sigma^2 \sim Normal(\beta^0, g(X^T X)^{-1}\sigma^2), \tag{2.18}$$

where the parameter g is a constant that controls the uncertainty relative to the variance around the prior mean. The g-prior for parameter σ^2 is specified as:

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2} \tag{2.19}$$

Based on Bayes' theorem, the posterior distribution can be parameterized as:

$$p(\beta, \sigma^2|y) \propto p(\beta|\sigma^2, y)p(\sigma^2|y). \tag{2.20}$$

The marginal posterior distributions [106] are given as:

$$p(\beta|\sigma^2, y) \propto N(\frac{g}{g+1}(\beta^{ols} + \frac{\beta_0}{g}), \frac{g}{g+1}\sigma^2(X^T X)^{-1}); \tag{2.21}$$

$$p(\sigma^2|y) \propto IG(\frac{n}{2}, \frac{SSR^2}{2} + \frac{1}{2(g+1)}(\beta^{ols} - \beta_0)X^T X(\beta^{ols} - \beta_0)), \tag{2.22}$$

where IG represents the inverse gamma distribution. From the above equations, the posterior mean for parameter β [106] can be obtained:

$$E(\beta|y) = \frac{g}{g+1}(\frac{\beta_0}{g} + \beta^{ols}). \tag{2.23}$$

Bayesian Information Criterion

Bayesian Information criterion (BIC) is a criterion for model selection [83] and is defined as:

$$BIC = -2ln(\hat{L}) + dln(n), \tag{2.24}$$

where n is the number of samples, d is the number of parameters estimated by the model and $\hat{L}$ is the maximum likelihood value of the model that measures the overall performance of model fit. BIC introduces a penalty for the number of predictors in the model $dln(n)$ to avoid the problem of overfitting. A lower value of BIC indicates a better model fit.

Significance test for regression coefficients

We would like to know that given a regression model, whether or not each predict variable makes contribution to explain the model. Since we assume the prediction error ϵ is normally and independently distributed with a mean of zero and variance of σ^2, t-test on regression coefficients can be performed to check the significance of the regression coefficient. Consider that a k-variable regression model with n observations:

$$Y = \beta X + \epsilon, \tag{2.25}$$

to determine the significance of an individual predictor variable, denoted as x_p, the null and alternative hypotheses are set:

$$H_0 : \beta_p = 0, \ H_1 : \beta_p \neq 0. \tag{2.26}$$

If the null hypothesis H_0 is true, the change of x_p would not give rise to the change of Y and there is no significant correlation between Y and x_p. The test statistic is calculated based on t-test:

$$\begin{aligned} t_p &= \frac{\beta_p^{ols} - \beta_0}{SE(\beta_p^{ols})} \\ &= \frac{\beta_p^{ols}}{SE(\beta_p^{ols})} \sim \tau_{n-k-1}, \end{aligned} \tag{2.27}$$

where β_p^{ols} is the OLS estimator and $SE(\beta_p^{ols})$ is the standard error for x_p. β_0 is the specified value in the null hypothesis, taken to be 0. The standard error

can be obtained by:

$$SE(\beta_p^{ols}) = \sqrt{\frac{SSR(\beta_p^{ols})}{n-k-1}}, \tag{2.28}$$

where $SSR(\beta_p^{ols})$ is the sum of squared error for β_p^{ols}. The t-score t_p has a t-statistic with $n-k-1$ degrees of freedom in the null hypothesis is true. Since it is a two-tailed test, $p-value$ can be calculated as $2 \times (1 - P(t \leq |t_p|))$. If the $p-value$ is less than the predefined threshold, we can say the variable x_p is of significance in the regression model.

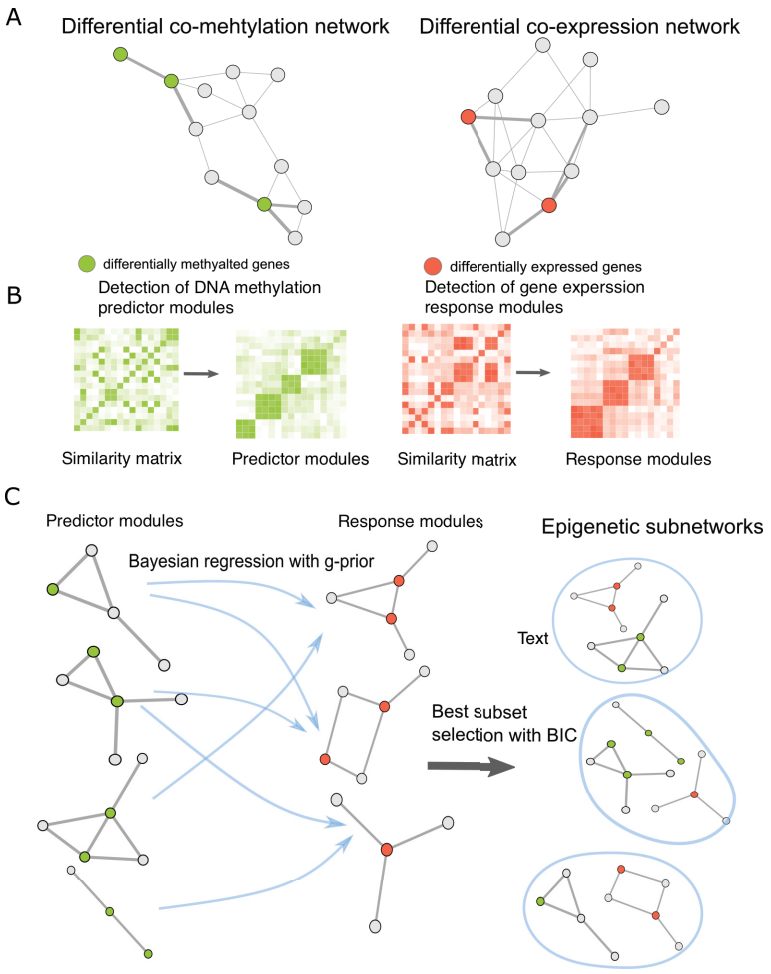


Figure 2.3: Overall framework.

2.3 Method

In this section, the method for the detection of epigenetic subnetworks is introduced in detail.

2.3.1 Overall framework

As shown in Fig. 2.3, the method consists of three main steps. In the first step, differential networks are constructed for gene expression and DNA methylation data in the context of the human interactome network, respectively. Edge weights are assigned according to the differential levels of gene co-expression (co-methylation) between tumor and normal conditions (Fig. 2.3A). Secondly, the similarity matrices are constructed by mapping the edge weights to the val-

ues of matrix elements. The gene expression and DNA methylation modules, in which genes are involved in a regulatory pattern, are discovered by the non-negative matrix factorization to the respective similarity matrix (Fig. 2.3B). Since multiple DNA methylations can drive a systematic change in a expression regulatory pattern, we consider DNA methylation modules as predictors and gene expression as responses. The relationships between DNA methylation predictor modules and gene expression response modules are quantified via a Bayesian regression model with the incorporation prior biological relatedness encoded as g-prior. For each response module, the best subset of predictor variables are selected based on BIC (Fig. 2.3C). Finally, the response modules with its corresponding predictors compose the epigenetic subnetworks, in which the differential expression regulatory pattern results from the varying level of DNA methylation of predictors.

2.3.2 Detection of predictor and response modules

Construction of differential networks

Since the differential expression (DE) has a cascading effect with the emergence of differential co-expression effects (DCE) due to the underlying biological network structures [52], we combine the effects of DE and DCE in the context of a human protein-interaction network [104] to construct the differential gene expression network. The level of DCE between each pair of interacting genes in the protein-interaction network is evaluated. As introduced in section 2.2.2, we use Pearson correlation coefficient ρ_t and ρ_n to evaluate the correlation between two genes X and Y in tumor and normal conditions, respectively. Then Fisher transformations are applied to the Pearson correlation coefficients. Recall that the Fisher transformation is defined as:

$$F(\rho) = \frac{1}{2} ln \frac{1+\rho}{1-\rho}. \tag{2.29}$$

The statistic Z is defined to assess the difference in gene correlation between tumor and normal conditions:

$$Z = \frac{F(\rho_t) - F(\rho_n)}{\sqrt{\frac{1}{n_t-3} + \frac{1}{n_n-3}}}, \tag{2.30}$$

where n_t and n_n denote the number of tumor and normal samples, respectively. The absolute value of Z is used as the edge weight on a pair of interacting genes in the protein interaction network.

The statistical significance on the statistic Z is evaluated:

$$\text{p-value} = 2 \times (1 - \phi(|Z|)). \tag{2.31}$$

After the Benjamini-Hochberg correction, gene pairs with adjusted p-value less than 0.05 are considered to be significantly differentially co-expressed between tumor and normal conditions. To filter out irrelevant genes and reduce false positives, genes are removed that do not show significant DCE with differentially expressed genes in the network.

Analogous, differential methylation data network is constructed in the same way.

Detection of predictor and response modules

The constructed differential networks provide valuable information about the gene regulatory patterns, since a larger value of the edge weight indicates a higher probability that the pair of genes are involved in a gene regulatory module. To detect gene modules, two similarity matrices $A_r[a_{ij}]$ and $A_p[a_{ij}]$ are constructed, based on the differential gene expression DNA methylation network. The matrix element a_{ij} represents the value of the edge weight between the gene i and gene j in the differential network. The discovery of module structures based on the similarity matrix can be formulated as a problem of symmetric nonnegative matrix factorization (NMF). The input similarity matrix $A_r[a_{ij}]$ ($A_p[a_{ij}]$) with size $n \times n$ can be factorized into a low rank matrix H that encodes the latent information embedded in the original similarity matrix, i.e.

$$A \approx H \times H^T, \tag{2.32}$$

where H with size $n \times k$ gives the information on module indicators, i.e., the matrix element h_{ij} in H indicates the confidence of assigning gene i to module j, where $i = 1, 2, ..., n$ and $j = 1, 2, ..., k$. The factorization of matrix A can be

achieved by minimizing the loss function:

$$min_{H \geq 0} \parallel A - H \times H^T \parallel, subject\ to\ H \geq 0. \tag{2.33}$$

We solve this problem using the algorithm *SymNMF* proposed by Da. et al [48]. The output H contains the information on module memberships. Specifically, for each row h_i in H, gene i is assigned into the k_{th} clusters if h_{ik} is the maximum element.

Significance test leading to the optimal selection of predictor and response modules.

The rank K (i.e., the column number of matrix H) determines the number of modules, which is a key parameter that needs to be explored. The choice of K in NMF is often an application-dependent and long-standing problem [108]. In this work, given the weighted differential network where the edge weight indicates the extent of the correlation between two genes, the task is to detect densely connected modules with high modularity. Note that not all detected modules by NMF would have above-average modularity, since it is possible that non-correlated genes would be grouped into a cluster representing isolated associations. Thus the modularity of the detected modules is evaluated by calculating module density [62]:

$$density(M_{ik}) = \frac{\sum_{p \in M_{ik}, q \in M_{ik}} A[a_{pq}]}{|M_{ik}| \times (|M_{ik}| - 1)}, \tag{2.34}$$

where M_{ik} indicates the i_{th} module under the rank k and A is the similarity matrix.

A permutation test is performed to assess the statistical significance of module density by randomly generating modules with the same size as the detected module in the background differential network. This procedure is repeated 1000 times, i.e., for each detected module, 1000 random modules are generated. Under the null hypothesis, the density of random modules is equal to or greater than the observed modules. The significance level (p-value) of the density for the observed module M_{ik} is calculated as:

$$p(M_{ik}) = \frac{\sum_{b=1}^{b=1000} I\{density(M_{ik}^b) \geq density(M_{ik})\}}{1000}, \tag{2.35}$$

where $density(M_{ik}^b)$ indicates the density score of the b_{th} permuted module. If the adjusted p-value is less than 0.05 after the Bonferroni correction, the observed module is considered to be statistically significant.

We expect that with an appropriate value of the rank K, most significant modules showing local regulation patterns would be detected. A wide range value of K is explored and we select the value of K leading to the largest number of detected significant modules as the optimal K. This procedure is performed for both DNA methylation and gene expression data. The optimal values of k_p and k_r are respectively obtained for DNA methylation predictor modules and gene expression response modules, respectively. With the optimal values of k_p and k_r, modules with the adjusted p-value less than 0.05 are selected as DNA methylation predictor modules and gene expression response modules, respectively.

Module quality measures

The module density measures whether or not genes in identified modules are densely connected. The other measure of module quality, separability score [51], is employed to evaluate whether or not a detected module is well separated from other modules in the differential network. The separability score between two modules M_i and M_j is determined by the inter-module adjacency and intra-module adjacency:

$$separability(M_i, M_j) = 1 - \frac{interAdj(M_i, M_j)}{\sqrt{density(M_i) \times density(M_j)}}, \tag{2.36}$$

where $density(M_i)$ and $density(M_j)$ are the intra module densities (defined in Eq. 2.34) for module M_i and M_j, respectively. The inter-module adjacency $interAdj(M_i, M_j)$ is defined as:

$$interAdj(M_i, M_j) = \frac{\sum_{p \in M_i} \sum_{q \in M_j} A[a_{pq}]}{n_i n_j}, \tag{2.37}$$

where A is the similarity matrix and n_i and n_j are the number of genes in M_i and M_j, respectively. The closer the separability score is to 1, the more separated are the M_i and M_j. The permutation test for the observed separability score is performed to obtain the significance level. The separability

and density scores measure the homogeneity and separateness of the detected modules [51]. We use these two measures to validate if the modules are well detected.

2.3.3 Detection of epigenetic subnetworks

In this section, we aim to identify the relationships between predictors and responses by detecting the set of predictors that best explains the variation in expression in the response module. We use the eigengene [50] as the representative of each module in one synthetic profile, since it allows to relate the module to the clinical trait of interest in an easy way and it can also be used as a feature in more complex predictive models including the Bayesian inference model [23]. To select the best subset of predictors for each response, Bayesian linear regression model with an informative g-prior is employed to compute all possible regression models for a response module. The biological relatedness between predictors and responses is encoded as an informative g-prior to guide the search of association between predictors and responses. The best subset of predictors for each response is selected according to the criterion of Bayesian information criterion (BIC).

Module eigengene

We treat each modules as a single unit by constructing the representative eigengene [50]. The eigengene is defined as the first principal component based on singular value decomposition [2]. In detail, let $Y = (y_{il})$ denote the gene expression profile for a response module, where $i = 1, 2, ..., n$ denotes the index of genes and $l = 1, 2, ..., m$ corresponds to the tumor samples. The expression profile for each gene, i.e., each row of Y, is standardized to have the mean 0 and the variance 1. The singular value decomposition of Y is represented as:

$$Y = UDV^T, \tag{2.38}$$

where U is an orthogonal matrix with size $n \times m$ and the columns of U are referred to as the left-singular vectors. V is the orthogonal matrix with size $m \times m$ and the columns of V and D is an $m \times m$ diagonal matrix of the

singular values. The first column of V is referred to as the module eigengene. Similarly, eigengenes of DNA methylation predictor modules are obtained from methylation profiles in the same way.

To evaluate if the module eigengene can represent the module profile well, we calculate the proportion of variances explained by the module eigengene [35] as follows:

$$varExplained(E) = \frac{|d^1|^2}{\sum_j |d^j|^2}, \tag{2.39}$$

where d^1 is the first element in the diagonal matrix D. The large value of $varExplained$ indicates that the module eigengene is properly generated and it can represent the profile well.

Bayesian regression with g-prior

We assume that the response module is associated with a set of predictors via a linear regression model. Given a response module eigengene Y_i and a set of predictor module eigenegenes X_γ, the prediction error

$$\epsilon_{i\gamma} = Y_i - \beta_{i\gamma} \times X_\gamma, \tag{2.40}$$

is assumed to be independent and identically distributed with mean 0 and variance σ^2, where the parameter $\beta_{i\gamma}$ indicates the vector of regression coefficients. Assuming that the response Y_i conditional on X_γ is subject to a multivariate normal distribution:

$$Y_i|X_\gamma, \beta_{i\gamma}, \sigma^2 \sim Normal(\beta_{i\gamma}X_\gamma, \sigma^2_{i\gamma}I), \tag{2.41}$$

where $\sigma^2_{i\gamma}I$ is a variance co-variance matrix that has error $\sigma^2_{i\gamma}$ on the diagonal and zeros for the remaining elements.

We employ Zellener's g-prior [107] to include the prior biological relatedness between responses and predictors. Intuitively, the g-prior controls the uncertainty in the prior belief relative to the variance of the observations around the mean, and the prior distribution of $\beta_{i\gamma}$ conditional on variance $\sigma^2_{i\gamma}$ is formulated as:

$$\beta_{i\gamma}|\sigma^2_{i\gamma} \sim Normal(\beta^0_{i\gamma}, g_{i\gamma}(X^T_\gamma X_\gamma)^{-1}\sigma^2_{i\gamma}), \tag{2.42}$$

where $\beta_{i\gamma}^0$ is the initial guess of mean vector, the term $X_\gamma^T X_\gamma$ is the variance-covariance matrix that provides a prior covariance structure, and $\sigma_{i\gamma}^2$ is data-dependent covariance matrix that can be scaled by a user-defined positive factor $g_{i\gamma}$. The prior information can be integrated into the model by changing the parameter in the prior distribution, thus we can propose a prior guess of the vector of regression coefficients and encode the corresponding prior belief in $g_{i\gamma}$.

As introduced in detail in section 2.2.4, the posterior distribution of $\beta_{i\gamma}$ is given by

$$p(\beta_{i\gamma}|\sigma_{i\gamma}^2, X_\gamma, Y_i) \sim N(\frac{g_{i\gamma}}{g_{i\gamma}+1}(\frac{\beta_{i\gamma}^0}{g_{i\gamma}} + \beta_{i\gamma}^{ols}), \frac{\sigma_{i\gamma}^2 g_{i\gamma}}{g_{i\gamma}+1}(X_\gamma^T X_\gamma)^{-1}), \tag{2.43}$$

where $\beta_{i\gamma}^{ols} = (X_\gamma^T X_\gamma)^{-1} X_\gamma^T Y_i$ is the OLS estimator of $\beta_{i\gamma}$, and the vector of regression coefficients $\beta_{i\gamma}$ can be estimated by the posterior mean and prior $g_{i\gamma}$:

$$\widetilde{\beta_{i\gamma}} = \frac{g_{i\gamma}}{1+g_{i\gamma}} + \frac{1}{1+g_{i\gamma}}\beta_{i\gamma}^{ols}. \tag{2.44}$$

The posterior distribution of $\sigma_{i\gamma}^2$ follows the inverse-gamma distribution and can be estimated as

$$p(\sigma_{i\gamma}^2|X_\gamma, Y_i) \sim IG(\frac{n_\gamma}{2}, \frac{SSR_{i\gamma}}{2} + \frac{(\beta_{i\gamma}^0 - \beta_{i\gamma}^{ols})X^T X \frac{1}{1+g_{i\gamma}}(\beta_{i\gamma}^0 - \beta_{i\gamma}^{ols})}{2}) \tag{2.45}$$

where n_γ is the number of predictors in X_γ and $SSR_{i\gamma}$ is the sum of squares of the residuals of $\beta_{i\gamma}^0$.

Encode prior biological information as g-prior

We evaluate the biological relatedness between each response Y_i and each predictor X_j based on the human protein interaction network [104]. The greater number of interactions between genes in predictor and genes in response, the higher degree of biological relatedness between them. We define the biological relatedness $r_{i\gamma}$ between X_j and Y_i as:

$$r_{ij} = \frac{2E_{ij}}{N_{ij}}, \tag{2.46}$$

where N_{ij} is the total number of genes in predictor X_j and response module Y_i. The parameter E_{ij} denotes the number of interactions between genes in the

predictor and the response. Then a weight parameter μ is added to control the relative influence of the prior biological relatedness in the Bayesian regression model:

$$g_{ij} = \mu r_{ij}. \tag{2.47}$$

When $\mu = 0$, we treat all predictors equally and no prior information is included in the model. The larger the value of g_{ij} is, the more confident we are about that the predictor X_j is associated with response Y_i.

Two factors, the prior coefficient vector $\beta^0_{i\gamma}$ and the scalar $g_{i\gamma}$, need to be set. In practice, we set $\beta^0_{i\gamma}$ to be a vector with all elements having values of zeros, which reflects our prior belief in the very subtle dependence between the predictors and responses. The parameter $g_{i\gamma}$ is originally formulated as a constant to control the confidence in the coefficient $\beta^0_{i\gamma}$. Specifically, a large value of $g_{i\gamma}$ leads the regression coefficients to be centered around $\beta^{ols}_{i\gamma}$. On the other hand, values of $g_{i\gamma}$ with a small value leads to the solution centered around $\beta^0_{i\gamma}$. We extend the formulation of $g_{i\gamma}$ as a scalar vector $\vec{g_{i\gamma}}$ to allow for different levels of the control in the elements in $\beta^0_{i\gamma}$. Each entry in $\vec{g_{i\gamma}}$ corresponds to one predictor, controlling the confidence in the prior belief relative to the variance of the observations around the mean. In this case, the scalar vector $\vec{g_{i\gamma}}$ constructed for response Y_i and the predictor set X_γ is composed of $[g_{ix}]$, where x indicates the index of predictors in the set $X\gamma$.

Best Predictor Subset Selection Based on BIC.

Assuming that k_p predictors are obtained, there are totally 2^{k_p} combinations of predictor variables for each response. We use BIC as the measure to select the best model. Recall that BIC is defined as:

$$BIC = -2ln(\hat{L}) + kln(n), \tag{2.48}$$

where n is the number of observations, k is the number of parameters estimated by the model and $\hat{L}$ is the maximum likelihood value of the model. The expected value of BIC is calculated as:

$$E[BIC_{i\gamma}] = nE[In(\sigma^2_{i\gamma})] + k_\gamma In(n), \tag{2.49}$$

where n is the number of samples and k_γ is the number of predictors in the set γ. The expected value of $\ln(\sigma^2_{i\gamma})$ is calculated as:

$$E[ln(\sigma^2_{i\gamma})] = Digamma(\frac{n}{2}) - In(\frac{SSR_{i\gamma}}{2} + \frac{(\beta^0_{i\gamma} - \beta^{ols}_{i\gamma})G_{i\gamma}X^T_\gamma X_\gamma G_{i\gamma}(\beta^0_{i\gamma} - \beta^{ols}_{i\gamma})}{2}), \tag{2.50}$$

where $G_{i\gamma}$ is a square matrix in which diagonal elements are $\sqrt{\frac{1}{1+g_{i\gamma}}}$ and the remaining elements are all zeros, and $SSR_{i\gamma}$ is the sum of squares of the residuals of the ordinary least squares $\beta^{ols}_{i\gamma}$. Given a response module, the combination of predictors with the smallest expected value of BIC would be selected.

Survival analysis

We hypothesized that the detected modules or subnetworks might be effective prognostic parameters that are associated with the survival time of patients. Thus, a survival analysis was performed. Since the coefficient in a Cox regression model is related to the hazard, i.e., a positive value represents a worse prognosis and a negative value indicates a positive association with survival time [12]. Thus, we devise the prognostic index scores for patients based on the coefficients in the Cox regression model of each module or subnetwork. The prognostic index score for a patient i with a response or predictor module k is defined as:

$$PI_{ki} = \beta^{cox}_k E_{ki}, \tag{2.51}$$

where β^{cox}_k is the Cox regression coefficient for module k and E_{ki} is the value of eigengene of module k for patient i.

For a subnetwork k, the multivariate Cox regression is performed and the prognostic index score for the patient i is defined as:

$$PI_{ki} = \sum_{c \in k} \beta^{cox}_c E_{ci}, \tag{2.52}$$

where β^{cox}_c is the Cox regression coefficient for a module variable c in the subnetwork k and E_{ci} is the value of module eigengene$c$ for the patient i.

Then we divide patients into two groups based on the prognostic index scores: low-risk (the PI score $< 30^{th}$ percentile of the entire PIs) and high-risk

(the PI score $> 70^{th}$ percentile of the entire PIs). Kaplan-Meier estimator is used to generate the survival curves for two groups, followed by the Log-rank test to the significance level on the difference of the survival time in two groups.

2.4 Simulation study

To test if the proposed Bayesian regression model can identify true relationships between predictors and responses, we first applied the method to the simulation datasets. We simulated predefined epigenetic subnetworks consisting of a set of predictors and a response. The result of simulation study is presented to characterize the ability of our method in detecting true epigenetic subnetworks.

2.4.1 Simulation dataset

We generated three sets of studies corresponding to different strengths of association $c_r = (0.3, 0.5, 0.7)$ within response modules. In each study, we generated four predictor modules, and simulated different levels of associations from $\Phi = (0.03, 0.05, 0.1, 0.2, 0.3)$ between predictors and responses to detect the true relationships with respect to different strengths of associations. Since the structure of epigenetic subnetworks was known, the performance of the model can be evaluated by comparing the detected structure to the predefined structure.

First, we simulated four DNA methylation predictors x_1, x_2, x_3, x_4 corresponding to different correlation signals $c_p = (0.3, 0.5, 0.3, 0.5)$, with the same size $n \times p$, where n and p indicate the number of samples and variables, respectively. In practice, we set $n = 200$ and $p = 25$. Let x_i^m denote the methylation level of the variable m in the i_{th} predictor module, which is generated as:

$$x_i^m \sim N(0, \Sigma_i^m), \tag{2.53}$$

where $\Sigma_i^m \sim Inverse-Wishart(60, (1-c_i)I + c_i J)$, c_i is the association signal taken from c_p, $I_{p\times p}$ is the identity matrix and $J_{p\times p}$ is a matrix with all entries as 1.

We generated three response modules y_1, y_2, y_3 with size $n \times q$, corresponding to different levels of correlations, in order to test if different levels of correlation within responses would affect the outcome of our method. In practice, we set $n = 200$ and $q = 30$. Thus, three response modules with correlation

signals $c_r = (0.3, 0.5, 0.7)$ were generated in a similar way as the predictor modules.

We further simulated subnetwork structures by assuming that specific predictors and responses contribute to the non-random associations. Let x_j and y_i denote the profile of the j_{th} module in x and the i_{th} module in y, respectively. The dependency between response y_i and a specific set of predictors is added. Thus the new profile of the i_{th} response y_i^S was obtained as follows:

$$y_i^S = y_i + \sum_{i \in S_i} x_i A + E, \tag{2.54}$$

where S_i indicates the set of predictors having associations with y_i, A is a matrix of size $p \times r$ with elements carrying the association signal from $\Phi = (0.03, 0.05, 0.1, 0.2, 0.3)$, and E is the random noise matrix with size $n \times q$ generated from the independent normal distribution with mean 0 and variance 1. In practice, we set S_1, S_2 and S_3 as $\{x_1\}$, $\{x_2\}$, $\{x_1, x_2\}$, respectively, which means that the response modules y_1 and y_2 are regulated by predictors x_1 and x_2 respectively, while y_3 is regulated by both x_1 and x_2. No predictor is specified for response module y_4.

In addition, we generated a gene-interaction network G to simulate the biological relatedness between responses and predictors by using two parameters: p_c is the probability of the connection between the predictor and response that belong to the same epigenetic subnetwork, and p_{cn} is the probability of the connection between the predictor and response not in a a same epigenetic subnetwork. We set $p_c = 0.1$ and $p_{cn} = 0.05$ such that predictor modules and response modules in the same subnetwork are relatively densely connected, whereas there are fewer links in the rest of epigenetic subnetworks.

2.4.2 Result

Three sets of dataset are simulated corresponding to different levels of correlations within response modules. We applied our method to the three datasets starting with the construction of module eigengenes. Fig. 2.4 shows the simulated methylation profile of two predictor modules across 200 samples with

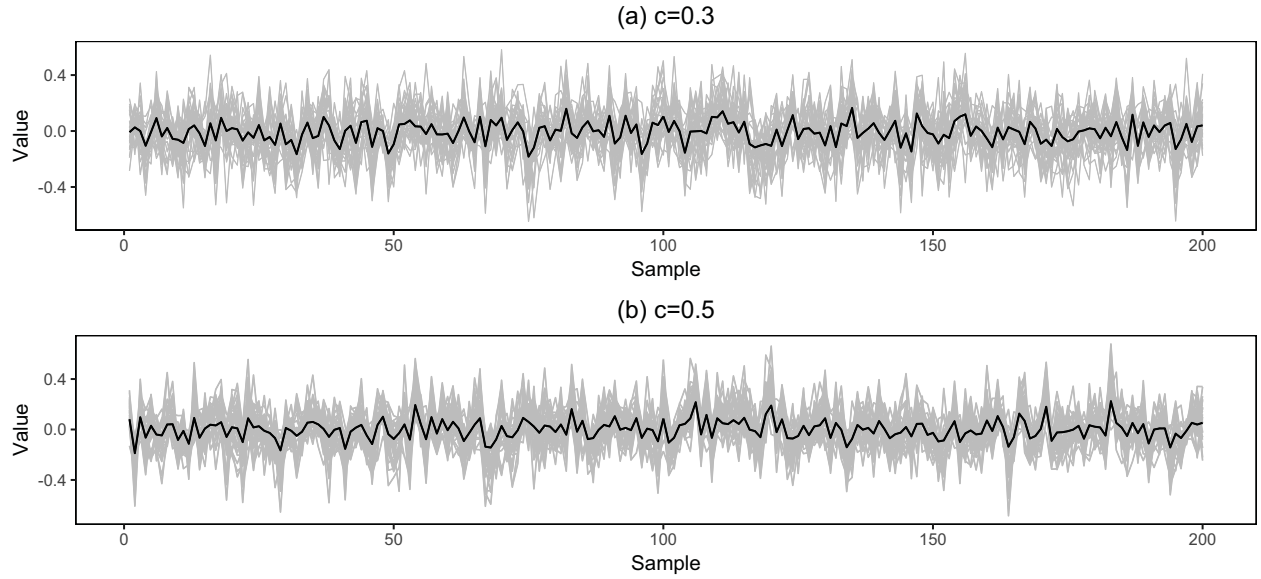


Figure 2.4: Methylation levels of simulated module genes (grey lines) and eigengene (black line) in 200 samples. (a) The correlation signal within the module is 0.3. (b) The correlation signal is set to 0.5.

correlation signal 0.3 and 0.5, respectively. An intuitive illustration of eigengene is shown by the black line in Fig. 2.4. It is highly correlated with the methylation profiles in the module.

For comparison, we applied the standard regression model without incorporation of prior knowledge to the simulated dataset. Results on three datasets y_1, y_2 and y_3 by two methods are shown in Table 2.1a, b and c, respectively.

Table 2.1a shows the result of identification on the first dataset y_1, where 'g-prior' indicates the result of our method with the incorporation of g-prior and 'no prior' indicates the result of the standard regression without prior. A wide range of association strengths between responses and predictors were specified from 0.03 to 0.3. When $a = 0.03$, a very weak association was specified between the response and predictor. In the case where the single predictor is specified to the response, our method can detect almost all true relationships. When stronger associations $a = 0.05, 0.1, 0.2, 0.3$ are specified, our method identified all the true relationships on y_1, y_2 and y_3. However, for the method without the incorporation with g-prior, it resulted in several false positives and cannot identify true relationships as correctly as our method. For example, the false

Table 2.1: Simulation results on three datasets by two methods

(a) Result on y_1

response	true predictor	identified predictors									
		a=0.03		a=0.05		a=0.1		a=0.2		a=0.3	
		g-prior	no prior	g-prior	no prior	g-prior	no prior	g-prior	no prior	g-prior	no prior
y_1^1	x_1	x_1^*	x_1^*	x_1^*	x_1^*	x_1^*	x_1^*	x_1^*	x_1^*	x_1	x_1^*
y_2^1	x_2	x_2^*	x_2^*, x_3^*	x_2^*	x_2^*	x_2^*	x_2^*	x_2^*	x_2^*	x_2^*	x_2^*
y_3^1	x_1, x_2	x_1^*	x_1^*, x_2^*	x_1^*, x_2^*	x_1^*, x_2^*	x_1^*, x_2^*	x_1^*, x_2^*	x_1^*, x_2^*	x_1^*, x_2^*, x_4^*	x_1^*, x_2^*	x_1^*, x_2^*
y_4^1	no predictor	x_3	x_3	x_4	x_4	x_4	x_4	x_3	x_3	x_3	x_3

(b) Result on y_2

response	true predictor	identified predictors									
		a=0.03		a=0.05		a=0.1		a=0.2		a=0.3	
		g-prior	no prior	g-prior	no prior	g-prior	no prior	g-prior	no prior	g-prior	no prior
y_1^2	x_1	x_1^*	x_1^*	x_1^*	x_1^*	x_1^*	x_1^*	x_1^*	x_1^*	x_1^*	x_1^*
y_2^2	x_2	x_2^*	x_2^*	x_2^*	x_2^*	x_2^*	x_1^*, x_2^*	x_2^*	x_2^*	x_2^*	x_1^*, x_2^*
y_3^2	x_1, x_2	x_1^*, x_2^*	x_1^*, x_2^*	x_1^*, x_2^*	x_1^*, x_2^*	x_1^*, x_2^*	x_1^*, x_2^*	x_1^*, x_2^*	x_1^*, x_2^*	x_1^*, x_2^*	x_1^*, x_2^*
y_4^2	no predictor	x_4	x_4	x_4	x_4	x_4	x_4	x_4	x_4	x_4	x_4

(c) Result on y_3

response	true predictor	identified predictors									
		a=0.03		a=0.05		a=0.1		a=0.2		a=0.3	
		g-prior	no prior	g-prior	no prior	g-prior	no prior	g-prior	no prior	g-prior	no prior
y_1^3	x_1	x_1^*	x_1^*	x_1^*	x_1^*	x_1^*	x_1^*	x_1^*	x_1^*	x_1^*	x_1^*
y_2^3	x_2	x_2^*	x_2^*	x_2^*	x_2^*	x_2^*	x_1^*, x_2^*	x_2^*	x_2^*	x_2^*	x_1^*, x_2^*
y_3^3	x_1, x_2	x_1^*, x_2^*	x_1^*, x_2^*	x_1^*, x_2^*	x_1^*, x_2^*	x_2^*, x_4^*	x_1^*, x_2^*	x_1^*, x_2^*	x_1^*, x_2^*, x_4^*	x_1^*, x_2^*	x_1^*, x_2^*
y_4^3	no predictor	x_2	x_2	x_2	x_2	x_2	x_2	x_2	x_2	x_2	x_2

a indicates the association signal

* indicates the regression coefficient is statistically significant.

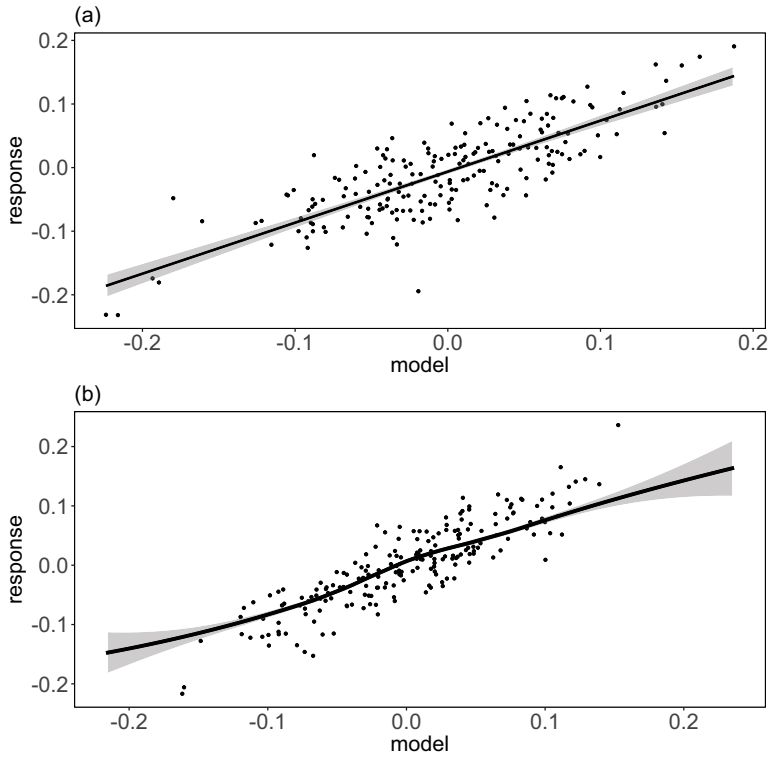


Figure 2.5: Fitting regression models. (a) The model for subnetworks consisting of response y_2^2 and predictor x_2 with association signal 0.1. (b) The model for subnetworks consisting of response y_3^2 and predictor x_1 and x_2 with association signal 0.1.

positive x_2 was detected by the standard model for the response y_2^2 and y_2^3, which was not specified in the relationship. Figure 2.5 shows the two examples of fitted regression models constructed by our method.

2.4.3 Discussion

The simulation analysis demonstrated that our model can identify subnetworks correctly even in the case where a very weak association is specified, while the standard regression model without g-prior resulted in multiple false positives. The g-prior in our method worked as a modifier on the shrinkage incurred on each predictor parameter. A larger value of g-prior corresponds to a smaller shrinkage incurred on the corresponding regression coefficient, making the corresponding variable less likely to be shrunk out of the model. It is worth nothing that it only modified the degree of shrinkage of a predictor, but

not the correlation between responses or the order in which the predictors are selected by the model.

2.5 Case study

In this section, the method was applied to a breast cancer dataset. We describe the results of applying the ideas discussed in previous sections to the task of detection of epigenetic subnetworks. The experiment procedures and results are described and summarized.

2.5.1 Dataset

We collected sample matched level-3 Illumina 450k methylation data and HiSeq RSEM gene-normalized RNA-seq data of breast cancer from TCGA [100]. We followed the strategy used by Jiao et al. [40] to assign the methylation value to a given gene, which was introduced in section 2.2.1 in detail. After data preprocessing, we generated the sample matched gene expression and DNA methylation profiles in 786 invasive ductal carcinoma tumor samples as well as 84 normal samples.

In addition, TCGA provides the corresponding clinical information including the patient status (alive or dead), the survival days (days to last follow-up or days to death). Such information was also collected to perform the survival analysis.

The information of the protein-protein interaction (PPI) was used in the inference procedure. It refers to the physical contact of high specificity between two proteins and it has been studied from multiple perspectives such as molecular dynamics, signal transduction and so on [20]. We downloaded the PPI network from the Protein Interaction Network Analysis (PINA) platform [104], which integrates and annotates the data from six public PPI databases (MINT, IncAct, DIP, BioGRID, HPRD, and MIPS/MPact). The network consists of 166776 edges and 16182 nodes.

2.5.2 Discovery of predictor and response modules

Significance test leading to the optimal selection of rank K

Differential gene expression and DNA methylation networks were constructed by evaluating the differential co-expression and co-methylation in the PPI net-

work. Two respective similarity matrices were generated by mapping the edge weight in the differential networks into the value of matrix elements, where an element indicates the probability that two genes may be involved in a regulatory pattern, i.e., the same module. Next, *SymNMF* was performed on these two similarity matrices to discover predictor and response modules. A wide range of candidate values from 5 to 70 for the number of modules K was explored. We expected that with an appropriate value of K, the most number of modules showing significant high-density would be detected. Given a candidate value for K, density scores were calculated for detected modules. By performing the significance test, the statistical significance of the module density was evaluated. Fig. 2.6 shows the number of the predictor and response modules showing significant density with respect to the parameter K. We observed that with the increase of K, the number of significant predictor and response modules increases to a maximum point followed by a decrease in the number of modules. The maximum number of the significant predictor and response modules were detected when k_p and k_r are set to 50 and 49, respectively. When K exceeded the optimal value, the number of significant modules did not grow with the increase of K any more. It indicated that in the case where K is greater than the optimal value, dissimilar genes were grouped into more non-correlated modules. Finally, 21 significant predictor modules and 39 significant response modules were detected with adjusted p-values less than 0.05.

Module quality measures

Density-based measure As we discussed, we employed the module density to select the significant modules which remain densely connected in the differential networks. Significance levels of density statistics were measured by a permutation test. We showed the result of the permutation test in Fig. 2.7, where we presented the density of observed modules as well as the distribution of the densities of 1000 randomly modules. From Fig. 2.7, we can see that the density scores for detected modules are significantly higher than random scores.

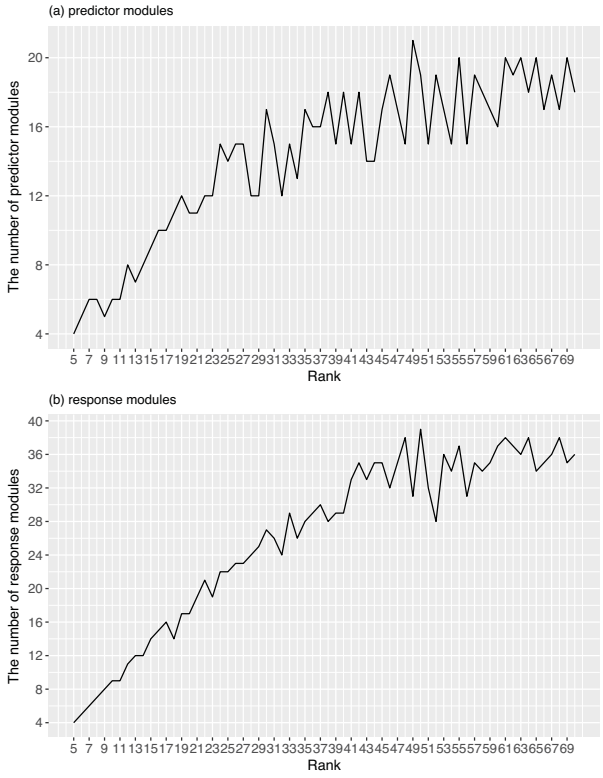


Figure 2.6: The number of predictor modules (a) and response modules (b) showing significant density score with respect to the parameter rank. The x-axis represents the candidate values for parameter K and y-axis represents the number of significant modules.

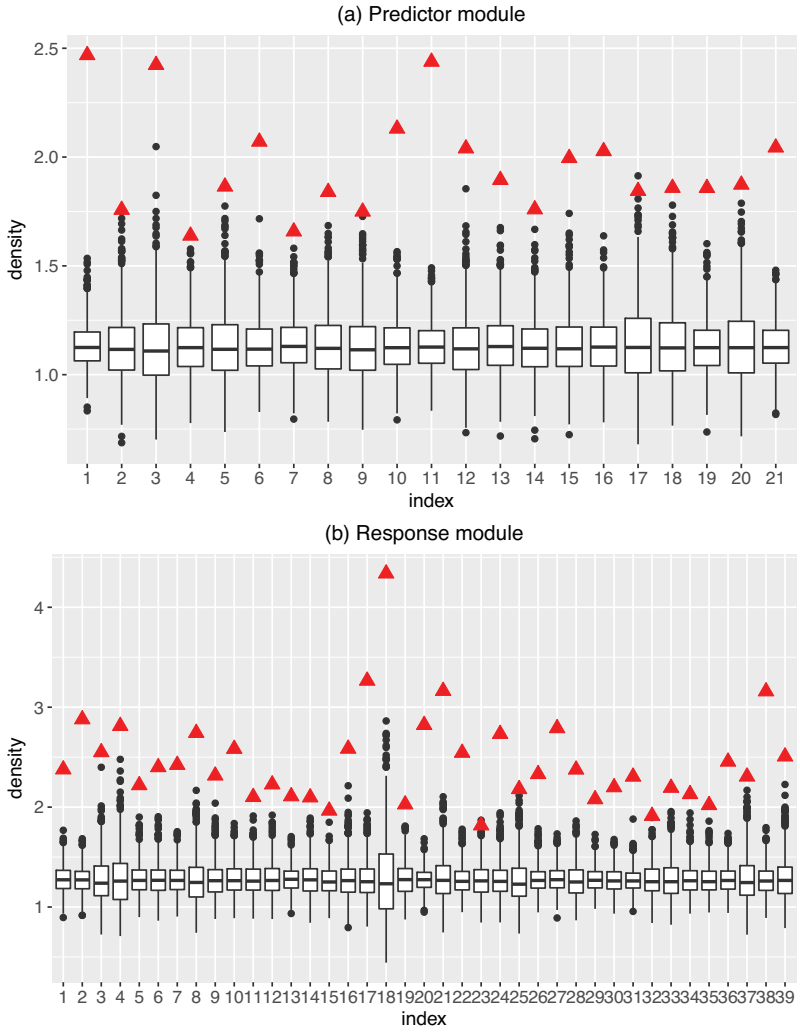


Figure 2.7: Module density scores. (a) The density of identified predictor modules. The red triangles represent observed density scores for predictor modules and boxplots represent the corresponding density scores of 1000 randomly generated modules. (b) The density of identified response modules. The red triangles represent observed density scores for response modules and boxplots represent the corresponding density scores of 1000 randomly generated modules.

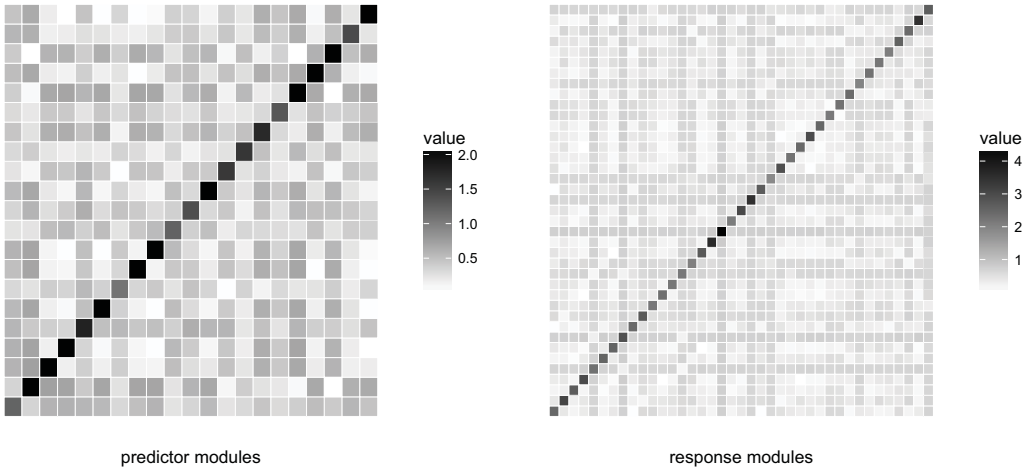


Figure 2.8: Heatmaps of separability and density scores for predictors (left) and responses (right).

Separability-based measure Next we evaluated the separability of identified modules to test if modules remain distinct from others. Separability scores and corresponding p-value were calculated to evaluate the significance levels of the separability for each pair of identified modules. By setting the threshold of p-value as 0.05, we observed that all pairs of predictors and responses are of significant separability. The p-values of separability scores for both predictor and response modules were attached in Appendix A. Two heatmaps (Fig. 2.8) shows the separability and the density scores between each pair of modules, where the off-diagonal blocks represent the separability scores and the diagonal blocks represent module density. Evaluations on the density and separability revealed that the modules are well defined and genes within a module remain densely connected as well as distinct from other modules.

Other measures We calculated $varExplained$, the proportion of the variance explained by module eigengenes, to check if the module profile is well represented by the eigengene. Figure 2.9 shows the boxplots of $varExplained$ for predictor and response modules. The median values of $varExplained$ for predictors and responses were 0.82 and 0.80, respectively, which indicated the eigengene can represent a large proportion of variance of the module profile.

In addition, we evaluated if the detected predictor and response modules

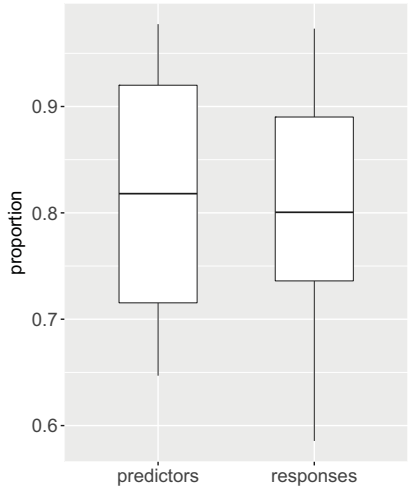


Figure 2.9: The proportion of variance explained by eigengenes

are correlated with patient survival time. We selected the right-censoring tumor samples, i.e., patients with known death time, to measure the correlation between the module eigengene and the survival time of patients. The Pearson correlation coefficients and the corresponding significance levels by the permutation with z-test were calculated. The modules with p-value less than 0.05 are considered to be associated with the patients survival time. We found that 13 out of 39 response modules are significantly correlated with the patient survival time, while no significant correlations between predictors and survival time were found. Fig. 2.10 showed scatterplots between eigengenes and the patient survival time for the 13 response modules.

2.5.3 Discovery of epigenetic subnetworks

Effect of varying weights on prior

The Pearson correlation between the profiles of DNA methylation and gene expression within each detected subnetwork are calculated to evaluate the performance. The p-value on the correlation coefficients after adjustment by a permutation test was obtained for each detected subnetwork. The Fisher's meta analyzed p-value was obtained by combining the set of p-values for all subnetworks into one meta p-value using Fisher's combined probability test to evaluate the overall performance. The weight parameter μ on g-prior was set

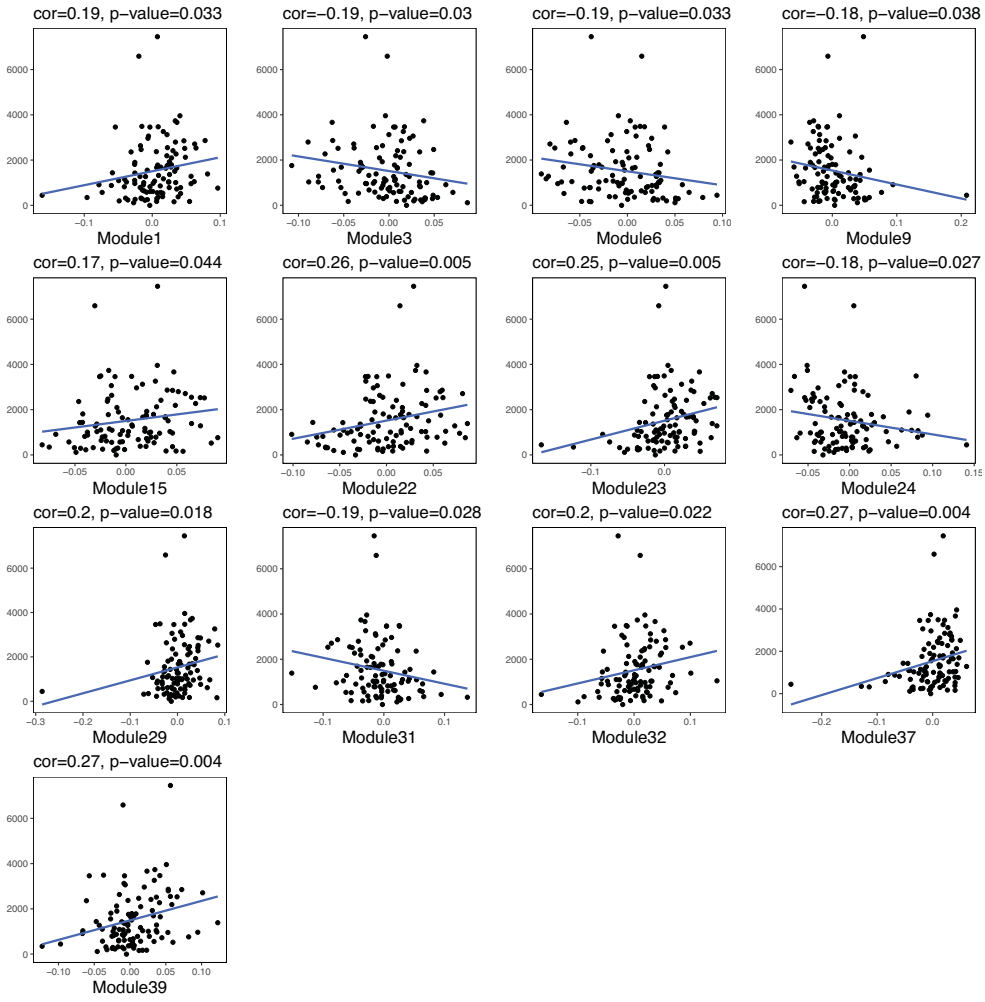


Figure 2.10: Scatterplots between the eigengenes of response modules and the patient survival time. In each figure, a dot represents a patient, and the x-axis and y-axis represent the profile of the module eigengene and the corresponding patient survival days, respectively.

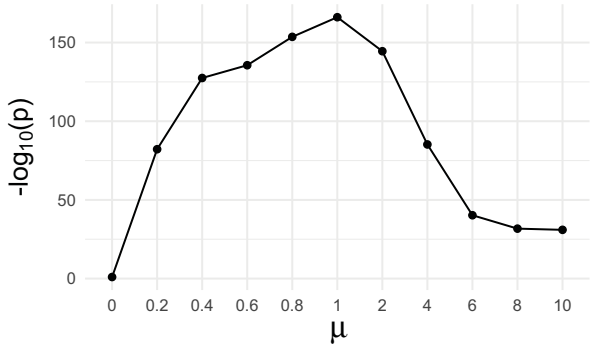


Figure 2.11: Effect of different weights on performance. It shows the negative logarithm of Fisher's meta analyzed p-value with different weight values.

to control the relative influence of prior biological relatedness to the discovery of epigenetic subnetwork. We measured the sensitivity of our method to the weight μ from a wide range of candidates $[0, 0.2, 0.4, 0.6, 0.8, 1, 2, 4, 6, 8, 10]$. Different sets of epigenetic subnetworks were detected and the Fisher's meta analyzed p-values were obtained with respect to different values of the parameter μ.

Fig. 2.11 shows the negative logarithm of Fisher's meta-analyzed p-value for each μ. When $\mu = 0$, no prior information was incorporated. As the value of μ increases, the performance increases to a certain point then followed by a decrease. The best performance was obtained when $\mu = 1$, therefore we selected the value of 1 leading to the most significant correlation within subnetworks as the optimal weight value on the prior. We noticed that in the case where $\mu = 0$, not all detected subnetworks show a significant correlation, which indicated that the incorporation with g-prior contributed to the discovery of significant epigenetic subnetworks.

Table 2.2 shows the detailed regression results. For each response module, the best subset of predictors was selected based on BIC. We assessed the significance level of regression coefficients in each detected subnetworks to evaluate whether the slope of the regression line differs significantly from zero. Table 2.2 shows the detailed regression coefficient and the corresponding significance level of each model. Except for the response module $y18$ and $y21$, all models show a significant relationship between the predictor and the response. Thus,

we removed the two subnetworks and finally 37 epigenetic subnetworks were kept.

We calculated the confidence score (Table 2.2) of each selected predictor x_i for response y_j, which measures the proportion of variance explained by x_i and the confidence in being a true regulation.

2.5.4 Follow up analysis.

Pathway enrichment test and network analysis

To determine the biological functional relevance of the detected epigenetic subnetworks, we performed the pathway enrichment test using reference pathways in the databases KEGG [43], Reactome [17], Biocarta [73], GO [4] and Canonical pathways (CP) downloaded from MSigDB [88]. The subnetwork is considered to be enriched in a reference pathway if a p-value < 0.05 is obtained by Hypergeometric test after correction. First, we examined the functional homogeneity of the detected subnetworks. A set of genes is defined as functional homogeneity if they are enriched in at least one GO category [4]. We found that all detected subnetworks exhibit significanct functional homogeneity since they are all enriched in at least one reference set in GO. Table 2.3 shows the ratio of enriched subnetworks in each database. All detected subnetworks are enriched in at least one reference pathway from Reactome and CP, and 35 out of 39 subnetworks (90%) and 28 out of 39 subnetworks (72%) were enriched in KEGG and Reactome pathways, respectively. In addition, we evaluated the proportion of reference sets enriched for epigenetic subnetworks (Table 2.3) and found that 42.3%, 41.9%, 50.9%, 43.5% and 49.8% of reference sets in GO, KEGG, CP, Reactome and Biocarta were enriched for detected subnetworks, respectively. The results revealed that the detected epigenetic subnetworks are of great biological relevance.

Next we asked if the detected subnetworks were related to cancer, especially the breast cancer. We examined whether the genes in detected epigenetic subnetworks are cancer-related biomarkers. We collected 2027 cancer genes from allOnco database (`http://www.bushmanlab.org/links/genelists`), and 738

Table 2.2: Regression results

Response	Predictor	Coefficient	P-value	Response	Predictor	Coefficient	P-value
y1	x3	−1.04	2.09E−02	y20	x2	−0.69	1.41E−58
	x8	1.00	6.51E−22		x14	−0.27	5.12E−12
y2	x15	−0.63	7.47E−18	y21	x5	−0.04	0.313
	x19	−0.27	1.86E−04	y22	x3	−0.10	3.98E−23
y3	x13	−0.24	2.51E−11		x8	−1.00	1.93E−21
y4	x7	−0.38	1.33E−28	y23	x15	−0.19	1.19E−07
y5	x16	−0.31	5.96E−19	y24	x3	0.32	2.10E−17
y6	x2	−0.08	0.02		x13	−0.41	9.65E−36
y7	x3	−0.41	1.39E−30	y25	x7	−0.27	5.99E−15
	x13	−0.66	2.48E−68	y26	x2	−0.48	1.93E−47
y8	x16	−0.29	6.66E−17	y27	x2	−0.41	1.04E−33
y9	x15	0.76	1.49E−18	y28	x2	−0.32	1.59E−20
	x17	−0.19	5.71E−05	y29	x15	−0.35	4.09E−24
	x19	−0.29	7.53E−05	y30	x3	0.71	5.76E−12
y10	x15	−0.69	3.31E−20		x8	−0.73	1.50E−12
	x19	−0.49	5.42E−11		x15	0.20	3.63E−09
y11	x2	−0.31	6.18E−21	y31	x15	0.83	1.22E−29
	x3	−0.89	3.50E−16		x19	−0.55	1.31E−14
	x8	0.57	4.82E−09	y32	x15	−0.57	7.20E−14
	x11	0.33	2.61E−09		x16	0.16	3.54E−06
y12	x13	0.51	9.24E−40		x19	0.25	7.44E−04
	x18	−0.36	2.26E−21	y33	x2	−0.44	5.19E−17
y13	x15	0.23	2.83E−11		x15	0.46	3.37E−18
y14	x15	−0.69	4.85E−20	y34	x15	−0.60	2.53E−15
	x19	0.52	2.41E−12		x19	0.54	1.33E−12
y15	x3	−0.94	2.70E−19	y35	x13	−0.18	1.51E−06
	x8	0.91	3.78E−18		x15	0.75	6.78E−27
y16	x1	−0.51	3.31E−47		x19	−0.39	1.84E−07
	x3	0.94	1.11E−22	y36	x1	0.33	3.05E−20
	x8	−0.70	6.40E−14		x3	−0.11	3.50E−25
y17	x2	−0.20	2.02E−08		x8	0.88	7.88E−19
y18	x14	0.03	0.36	y37	x15	−0.17	2.57E−06
y19	x15	0.60	4.08E−15	y38	x15	0.93	1.73E−36
	x19	−0.57	5.85E−14		x19	−0.69	1.38E−21
				y39	x16	0.19	4.24E−08

Table 2.3: The result of pathway enrichment tests in detected epigenetic subnetworks.

Reference set	GO	Biocarta	CP	KEGG	Reactome
The ratio of enriched subnetworks	1	0.718	1	0.897	1
The ratio of reference sets enriched for subnetworks	0.423	0.498	0.509	0.419	0.435

breast cancer driver genes from intogen [81] and OncoSearch [56]. On average, 20% of genes in the detected subnetworks were cancer genes and 9% were breast cancer genes. Table 2.4 shows the breast cancer genes in detected epigenetic subnetworks, where the third column 'ratio' represents the ratio between the number of cancer genes and the module size. We found that, except for subnetwork 4, there is at least one breast cancer gene in each detected subnetwork, which reveals that the epigenetic subnetworks are related to breast cancer. In addition, multiple important breast cancer genes were detected in the epigenetic subnetworks, like gene ERBB2 in subnetwork 19, a known proto-oncogene, that encodes HER2, a member of the human epidermal growth factor receptor. Genes TP53BP1 and TP53BP2 were also detected and encode a member of the ASPP (apoptosis-stimulating protein of p53) family of tumor suppressor p53 interacting proteins.

We took the epigenetic subnetwork 16 as an example and performed an extensive analysis for it. The subnetwork 16 contained 20 cancer genes and 9 breast cancer genes (CDK2, PRLR, CDH1, ERBB3, TP53BP2, SRC, MBIP, KDM1A and SERPINE1) and it was enriched in 12 KEGG pathways including two pathways that are specific to the breast cancer: KEGG cell cycle and KEGG P53 signalling pathway. Fig. 2.13 shows the network representation of subnetwork 16, including genes involved in KEGG pathways and genes showing correlations larger than 0.3. Genes acting as predictors were drawn as circles and responses were drawn as squares. Multiple epigenetic mechanisms were detected between predictors and responses. We found that the mechanism between SFN and CDK2 in subnetwork 16 was supported by observations that SFN is a frequently hypermethylated gene [41] [44] emerging as a new

Table 2.4: Breast cancer genes in detected epigenetic subnetworks.

Subnetwork	Breast cancer gene	Ratio
1	GRB7 SRC MED24 NCOA4 BCL2 PIK3CB	6/78
2	IDH1 ESR2 FOXO1 HSD17B12 PTPRJ CDKN2C SRC	7/79
3	CRK EDNRB	2/28
5	FOXM1 CLSPN MELK MMP14	4/41
6	NR3C1 PIK3CA FGF2	3/40
7	VIM ABCB1 GATA3 SFPQ ERBB4 SRC	6/69
8	CDK1 TOP2A	2/27
9	MAPK14 IKBKE SLC9A1 SRC MBIP KDM1A SERPINE1	7/94
10	FASN TWIST1 SRC	3/69
11	CSNK1A1 PTGES3 RUNX1 SRC MBIP KDM1A SERPINE1 AKT2 MAPK1 HMGB1	10/117
12	BMP6 TGM2 SRC	3/71
13	CHD4	1/51
14	CCNB1 RAD51 SRC	3/67
15	BAG3 EGFR SRC	3/70
16	CDK2 PRLR CDH1 ERBB3 TP53BP2 SRC MBIP KDM1A SERPINE1	9/92
17	CAV1	1/34
19	SOS1 PRKCZ SMYD3 ERBB2 SRC	5/68
20	SPTAN1 AHNAK ITGB1 NDRG1 ITSN1 EPAS1 MSN SRC	8/89
22	SMARCA4 MYH14 SRC	3/77
23	SMG1	1/34
24	IGF1R AKT2 PPARG SRC	4/66
25	DNMT3B ETS1	2/28
26	FN1 RBPJ FGFR1 KDM1A TGFBR2 VDR ITGB3	7/49
27	RNF41 SPARC	2/49
28	TFAP2A	1/37
29	FUBP1 IRS1 MAPK3 BTRC PIK3R1 ANK3	6/54
30	RHOA KDM5B MAP3K1 SRC MBIP KDM1A SERPINE1	7/112
31	PRKAR1A FUS MYH9 MKL1 HNRNPM SRC	6/84
32	CSK SMAD2 SOS2 TNPO1 SRC MBIP KDM1A SERPINE1	8/95
33	KDR FHL2 CD9 STAT5B PTN TP73 SRC	7/61
34	GAPDH APEX1 SRC	3/72
35	HDAC1 MUC1 HSPA5 PDIA6 RAB11FIP1 SRC MBIP KDM1A SERPINE1	9/104
36	PAK1 NOTCH1 GSN HSP90AA1 SRC MBIP KDM1A SERPINE1	8/104
38	CTNNB1 HSD17B4 ACO1 CD36 SRC	5/73
39	STUB1	1/29

inhibitor of CDK2 in breast cancer cells [53]. In addition, SFN has an important function in preventing breast tumor cell growth [53] which suggests that SFN may play a therapeutic potential role in cancer prevention by targeting epigenetic machinery. We also observed that CCNA1 has been detected as an epigenetic regulator in Fig. 2.13. Evidence in the literature showed that the differential methylation pattern of CCNA1 was associated with the treatment response in breast cancer and could potentially be a predictive marker to anthracycline/mitomycine sensitivity [46]. Moreover, multiple researches demonstrated that UHRF1 interacting with various proteins in multiple pathways results in the silencing of key tumor suppressor genes in breast cancer [88][84]. In Fig. 2.13, we observed that the methylation pattern of UHRF1 was highly correlated with the expression of multiple breast cancer genes including ERBB3, TP53BP2 and PRLR. In addition, genes like PLCG1 and PTPN6 in subnetwork 16 were also likely to be epigenetic regulators, which was supported by several researches [61][66]. Overall, these findings supported the idea that our method successfully detects epigenetic subnetworks containing verified epigenetic mechanism, and the detected subnetworks could be a starting point to uncover the underlying epigenetic mechanisms.

Survival analysis

We hypothesized that the profiles of gene expression or DNA methylation in detected modules and subnetworks might be effective prognostic parameters associated with survival time. As introduced in Method, we derived the prognostic index score for each patient based on the module profiles. The patients were divided into high-risk and low-risk groups and we performed the log-rank test to validate if the survival times in the two groups are significantly different. First, the survival analysis was performed on predictor and response modules. The results showed that 8 of 39 response modules (Fig. 2.14) can divide patients into two groups in which the survival time of patients of high-risk and low-risk are significantly different. However, no groups in predictor modules showed significantly different survival time. Next the multivariate Cox proportional regression was performed on epigenetic subnetworks and we detected

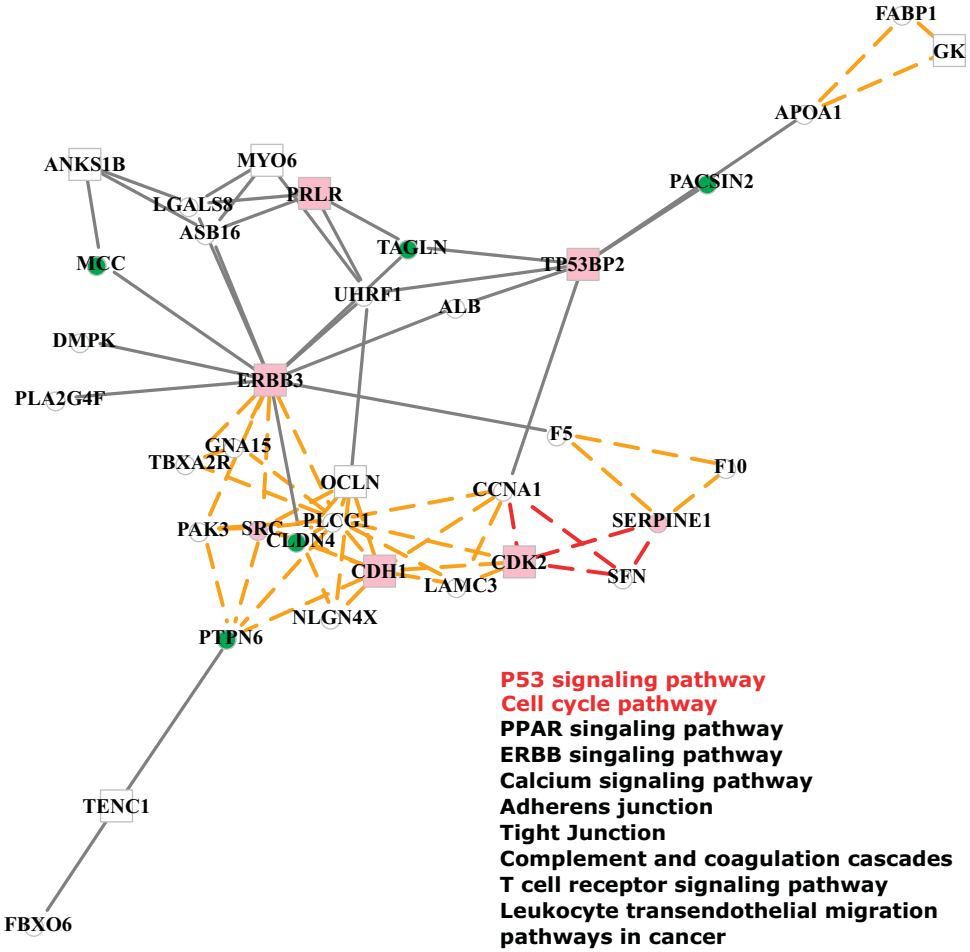


Figure 2.12: Network analysis of the detected subnetwork 16 in breast cancer. Genes that acted as predictors are represented by circles and responses are represented by squares. Pink nodes denote breast cancer driver genes and green nodes denote cancer genes. A grey line indicates that a Pearson correlation coefficient between a predictor and a response is larger than 0.03. Genes enriched in KEGG breast cancer pathways were connected by red dash lines and the yellow dash lines denote other KEGG pathways. The names for KEGG pathways are shown at the right bottom corner, where the red text indicates the breast cancer specific pathways.

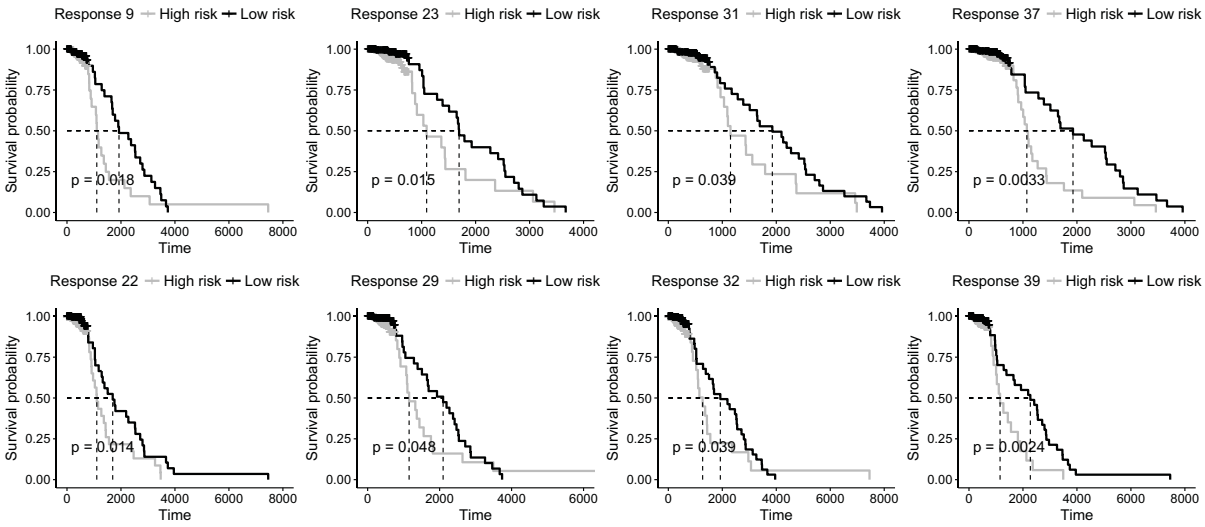


Figure 2.13: Kaplan-Meier survival analysis for patients in response modules.

that 11 of 37 subnetworks (Fig. 2.15) were significantly associated with survival time. In addition to the detected 8 significant response modules, 3 more responses in the subnetworks with the incorporation of DNA methylation predictors (subnetworks 1, 6, 36) showed the significant association with survival time, which indicated that the combinations of DNA methylation predictors and responses in the 3 subnetworks improve the classification of patients. It revealed that predictors and response in these 3 subnetworks jointly impact on the survival time.

Performance comparison

Ma et al. [63] detected 26 epigenetic modules by EMDN using TCGA breast cancer data and calculated the ratio of enriched modules as well as the ratio of enriched reference pathways. The method EMDN was compared with two other methods, EpiMod and FEM. They showed that the results detected by EMDN are more enriched than those achieved by EpiMod [102] and FEM [40]. Since the breast samples used in EMDN, FEM, EpiMod [63] were identical to the data in our paper, we can compare the performance of epigenetic subnetworks detected by EMDN directly. About 40% to 50% of subnetworks detected by EMDN, EpiMod and FEM were enriched in at least one reference set in GO, KEGG, CP, Reactome and Biocarta, which is much lower than the

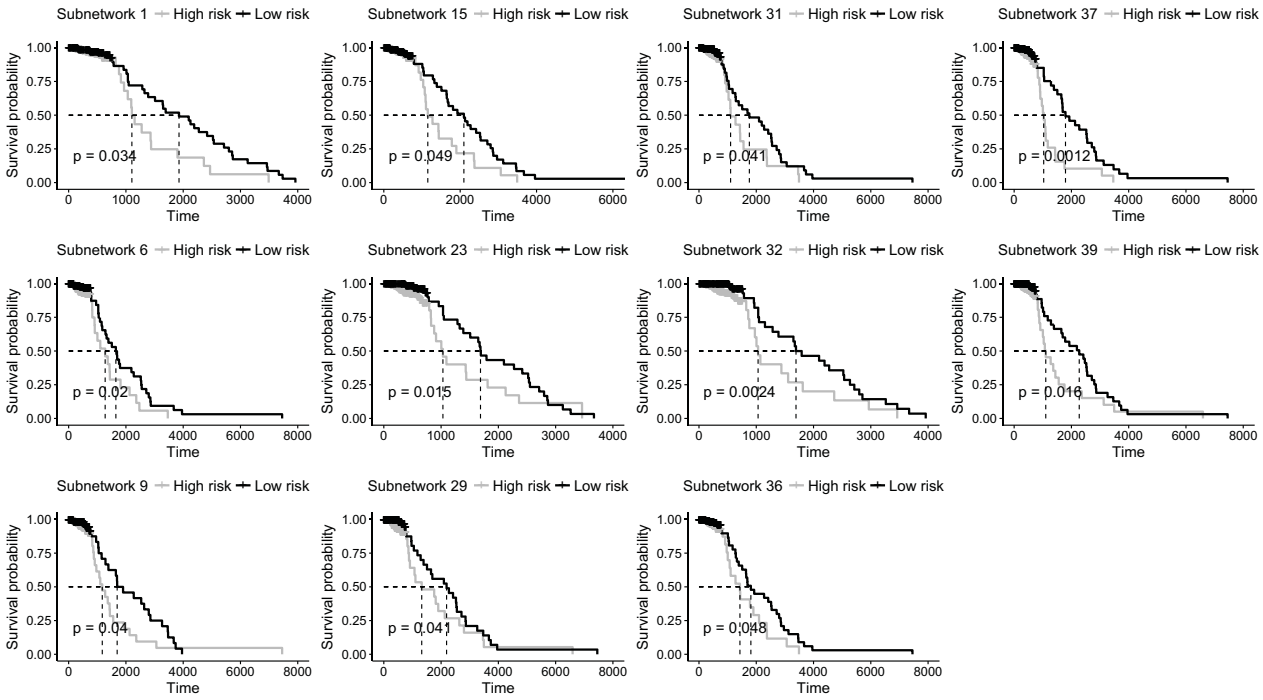


Figure 2.14: Kaplan-Meier survival analysis for patients in epigenetic subnetworks.

ratios achieved by our method. In our method, all of subnetworks detected were enriched in GO, CP and Biocarta, and 89.7% and 79.8% of subnetworks were enriched in KEGG and Reactome, respectively. However, one should note that EMDN did not take protein-interactions into account while EpiMod and FEM employed the PPI network, thus we conclude that incorporation with the biological interaction network may contribute to the discovery of biologically-relevant epigenetic subnetworks. The comparison with EMDN revealed that our framework with incorporation with PPI networks can detect more enriched subnetworks than EMDN, EpiMod and FEM.

2.6 Conclusion

Recent technology developments have enabled simultaneous genomic profiling of biological samples on multiple platforms, resulting in genome-wide DNA methylation and gene expression data. However, a systematic analysis between the two types of data for discovering biologically relevant combinatorial patterns is currently lacking. In this chapter, we present a method to evaluate the association between gene expression and DNA methylation at the module level by Bayesian regression with the incorporation of prior gene interaction knowledge. We first identified gene expression responses and DNA methylation predictors on a weighted differential expression and methylation networks respectively. Through a significance test, modules passing a p-value threshold were considered as predictors or responses. Density-based and separability-based measures in the significance test were used to validate if detected modules are densely connected and well separated from others. The results showed that the detected modules are well defined and that genes within a module show homogeneity and separability. Then we considered an eigengene as the representative of module profiles for a large proportion of variance of module profiles. With the incorporation of prior gene interaction networks as g-prior, we performed Bayesian regression to discover the dependent relationship between predictors and responses, i.e., the best subset of predictors for each response was selected. The application in breast cancer data demonstrated superior performance of our method to detect biologically relevant epigenetic subnetworks.

Overall, Our contributions lie in the following aspects:

(1) We proposed a novel method to detect epigenetic subnetworks by considering a set of highly correlated genes showing the pattern of differential co-expression/methylation instead of considering a single gene as a predictor or response. By comparing with EMDN, EpiMod and FEM which measure the association between gene expression and DNA methylation at the individual gene level, our detected epigenetic subnetworks were much more enriched in biological processes and signalling pathways, which indicates that evalu-

ating the association between gene expression and DNA methylation at the module level would increase the biological association and shed light on the underlying mechanism. Furthermore, our method achieved a larger ratio of enriched subnetworks than that achieved by EMDN. This higher achievement in enrichment ratio is partially due to the construction of significant differential networks with the incorporation of gene interaction information to reduce false positives. The incorporation of the biological interaction networks may contribute to the discovery of enriched epigenetic subnetworks, however it could filter out important cancer genes which were not included in the prior network. Therefore it remained to be a trade-off between filtering out false positives and discovering novel cancer mechanisms, which could be a future research direction for investigation.

(2) By incorporating the prior biological knowledge as g-prior in a Bayesian regression model, it detected more significantly correlated epigenetic subnetworks than the alternative model without g-prior, which showed that encoding biological network information as g-prior effectively guided the selection of epigenetic subnetworks. It is possible to introduce other sources of prior information, such as the derived regulatory interactions in the literature.

(3) The network analysis for the detected epigenetic subnetworks revealed the direct causal mechanisms verified in other scientific papers, which indicated the ability of our method in detecting true epigenetic mechanisms and that the detected epigenetic subnetworks could be a good start to uncover underlying epigenetic mechanisms. Moreover, the survival analysis for detected modules and epigenetic subnetworks indicated that the derived modules might be effective prognostic factors associated with the patients' survival time.

Chapter 3

Analysis of aberrant gene expression in Breast cancer

3.1 Introduction

Alterations that occur within the transcriptome of cancer cells have been observed in multiple types of cancers. Usually, normal cells respond to stress by deploying repair or resistance tools to maintain their genetic integrity and assure survival [38][74]. In contrast, cancer cells typically do not have intact repair tools, which lead to genetic instability. Chromosomal instability (CIN) is a form of genetic instability that causes changes in both the structure and number of chromosomes [5][25][27][29][79][82]. For example, mutations in CIN genes like BRCA1/2 increase the number of deletions up to 50 bps, causing multiple defects within the genome [1]. Progressive accumulation of CIN within a tumor allows development of cell populations with heterogeneous properties. Some of these cells will carry selective survival advantages and will be responsible for further tumor progression [65]. Likewise, overexpression of APOBEC3, a member of the cytidine deaminase gene family, may generate frequent C>T base substitutions also leading to tumor heterogeneity and progression along the malignancy pathway [32]. Understanding the sequence of molecular events essential for tumor progression may not only benefit early detection of malignancies, but may also allow the development of more effective treatment and even prevention strategies. While the role of accumulating genetic mutations in cancer progression has been extensively discussed, it is still

not clear how alterations in gene expression contribute to cancer progression.

Changes in gene expression can be brought about by number of factors, including epigenetic modifications, translation regulation, and differences in mRNA and protein stability [3]. For example, increased activities of growth factor, chemokine and cytokine receptors can set off specific signaling cascades and subsequent changes in gene expression, without any direct involvement of genetic mutations. However, what are the most significant changes that occur within the transcriptome of cancer cells and how they may contribute to tumor development is not clear. Here we use an aggressive malignancy in breast cancer, triple negative breast cancer (TNBC), as a model to explore the role of transcriptomic alterations during early cancer development that are caused not by genomic mutations, but exclusively by differential gene expression. We achieve this by focusing specifically on genes that are heavily up-regulated in the non-amplified regions of the genome. We focused specifically on up-regulated genes because direct inhibition of these molecules may provide viable cancer treatment/prevention options at early stages of tumor development. Remarkably, our analysis of RNA-seq data in 158 TNBC cases revealed that there is indeed a set of expressional changes in two major groups of genes controlled by hypoxia-related factors. These two groups included molecules that regulate CIN and remodel tumor microenvironment (TME). This not only reveals new potential targets for TNBC therapy, but also indicates a critical role for hypoxia in early tumor development.

3.2 Method summary

3.2.1 Dataset

The Cancer Genome Atlas (TCGA) that represents the largest collection of patient samples with information on the mutation status, copy number aberrations (CNA), as well as gene expression patterns at different stages of tumor development. We collected breast cancer samples from TCGA with information on the CNA, gene expression as well as tumor information. According to the tumor stage information, 1078 samples were classified into four stages: from stage I to stage IV. According to the immunohistochemistry markers [71], 158 samples were classified as TNBC tumors in which the ER, PR and HER2 were all negative. With the tumor stage information, we classified TNBC tumors into TNBC-stage I, TNBC-stage II, TNBC-stage III and TNBC-stage IV. In addition, 114 normal samples were collected from TCGA for comparison with tumor sample data.

3.2.2 Identification of differentially expressed genes

We combine the two assessments fold-change and p-value to detect differentially expressed genes from both biological and statistical points of views. Fold-change is a biological assessment of changes in gene expression as represented in Eq. 3.1,

$$fold\text{-}change_i = log_2 \frac{mean(E_i^{tumor})}{mean(E_i^{normal})}, \tag{3.1}$$

where E_i^{tumor} and E_i^{normal} is the mean expression level of gene i in tumor and normal samples, resepectively. Empirical Bayes moderated t-test was applied to assess the statistical significance of differential expression. False discover rate (FDR) is obtained after Benjamini and Hochberg correction. Limma package in R [85] is employed to derive the two assessments of differentially expressed genes. Genes are considered as up-regulated genes if $FDR \leq 0.01$ and $FC \geq 2$. Down-regulated genes are selected if $FDR \leq 0.01$ and $FC \leq -2$. In addition, the frequency-based analysis for each affected gene is performed. By maintaining a two-fold change in the expression level as a minimum re-

quirement for a gene to be considered differentially regulated, the frequency of changes in each differentially expressed gene is calculated as a percentage of patients in whom the gene is up or down-regulated.

3.2.3 Evaluating the concordance between copy number amplification and up-regulated gene expression

As changes in gene expression may arise from the accompanying chromosomal amplifications and deletions and other types of mutations, to account for this, we isolate the differentially expressed genes exclusively from the non-amplified regions of the genome. We evaluate the associations between CNA and up-regulations in gene expression and identify the up-regulations that are driven by CNA. The CNA profile containing the information of the amplification status of each gene in each patient is generated from TCGA. Only the data of patients whose CNA profile and up-regulation status available are considered for this study. To avoid patient heterogeneities, only genes showing amplification over 40% of patients are considered as cancer relevant amplification genes. CNA regions are identified by calculating the percentage of amplification of genes on each chromosome region, and regions with at least 40% of amplification genes are considered as CNA regions. Then we analyze the concordance between CNA and up-regulations by two metrics: fold CNA-associated change and pearson correlation coefficients. For up-regulated gene i at CNA regions, tumors are grouped into two groups S_i^1, S_i^2, where S_i^1 and S_i^2 denote patients with and without gene i getting amplified, respectively. The fold CNA-associated change and empirical bayes t-test are assessed as the same way in section 3.2.2. If gene shows at least 1 positive fold CNA-associated change and FDR smaller than 0.01, it is considered to be associated strongly with CNA. Secondly, pearson correlation coefficient is calculated to quantify the correlation between CNA and gene expression. If a gene with the pearson correlation coefficient larger than 0.3, it is considered as CNA-driven genes as well.

3.2.4 Analysis on the accumulation of aberrant expression

We isolate the differentially regulated genes without the affection of CNA or mutations and name them exclusively expression-altered (EEA) genes. Then we perform progression-based analysis on the up-regulation status of EEA genes to get a general understanding about how those changes accumulate across TNBC samples. Since each sample can provide a snapshot of the accumulation process of the molecular changes [90][31], we consider each tumor sample is recorded at a specific time point during the accumulation of aberrant gene expression. To investigate the aberrant gene expression pattern across TNBC samples, we group samples with similar up-regulation profile into clusters. The binary profile is generated with up-regulation status (0 = no up-regulation, 1 = up-regulation) in rows and patient in columns and the hierarchical clustering with Euclidian distance based on the binary profile is employed to group TNBC samples into clusters.

Assuming that no up-regulatory events appeared in normal status, the sample profile with all zeros is generated representing normal status. The clusters are ordered along a progression path according to the extent of the accumulation of up-regulatory events, which is achieved by neighbour-joining algorithm based on the distance matrix between mean up-regulation status of each cluster. To determine the occurrence of up-regulatory events in each cluster, we generate binary subset vectors to represent the occurrence of each up-regulatory event in each cluster. Assuming four clusters are identified, the subset vectors from (0, 0, 0, 1) to (1, 1, 1, 1) are generated representing fifteen possible subsets of these clusters. For each up-regulatory gene, the cosine similarity between the mean up-regulation profile and each subset is calculated. The gene is assigned to the subset vector with the maximum similarity to its mean up-regulation profile [69].

3.2.5 Ingenuity pathway analysis and hypoxia analysis

Ingenuity pathway analysis (IPA) is performed on the genes from cluster 1 (the cluster with shortest distance to the normal sample) as described in Kramer et al. [47], The gene list is first annotated and the data set underwent various analyses including for core expression to study the interactions. The gene interactions are explored, built and different overlays including pathways, disease and function and molecule activity prediction were applied to obtain the required outputs. Comparison analysis is also performed among the different subpopulations (referred as clusters). Hypoxia analyses are performed using the hypoxia database (http://www.hypoxiadb.com). This database includes 72,000 manually curated entries taken on 3500 proteins extracted from 73 peer-reviewed publications selected from PubMed. As described in Khurana et al., it provides manually curated literature references to support the inclusion of the protein in the database and establish its association with hypoxia [45].

3.2.6 Drug data analyses

The cell lines from the cancerRXgene database are divided into high cluster 1 expression and low cluster 1 expression. This is done by creating a table where the rows were cell lines, the columns are cluster 1 genes and the intersection at each row and column was the expression value of that gene in that cell line. The expression values across all cell lines for each gene are then added together and the mean and standard deviation are calculated. Then each cell line is given a Z score for that gene $Z = (x - \overline{x})/\sigma$, where x is the value, $\overline{x}$ is the mean and σ is the standard deviation. Each cell line then has all of its Z scores summed together to give the total score of c1 gene expression. The 85th percentile and the 15th percentile are then taken to be the cluster 1 high expression and cluster 1 low expression groups respectively. After grouping the cell lines into high and low expression of cluster 1, analysis is ran on the sensitivity of these cell lines to drugs. The drug data is obtained from the cancerrxgene database. For each drug, the IC50 values from the database are taken for each cell line of the high expression cell lines, and each of the low

expression cell lines. The IC50 values for each group are then compared using a Mann-Whitney-U test. Using the percent survival, we generate the graphs and dose data from the cancerrxgene database and fitting a sigmoidal curve to the resulting plot, using the literature IC50 value as an estimator. The sigmoid curve used is of the form.

3.3 Result

3.3.1 Identify differentially expressed genes in breast cancer

To identify genes with aberrant expression patterns, we initially curated all the genes that are differentially regulated. We used the breast cancer-specific tumor samples from TCGA with information on the mutation status, CNA, as well as gene expression patterns at different stages of tumor development. Gene expression in breast tumor samples was compared to the expression of the matching genes in normal samples using fold change and FDR after Empirical Bayes moderated t-test with Benjamini-Hochberg correction. Our initial analyses in overall breast cancer identified 586 genes that were up-regulated and 1446 genes that were down-regulated at multiple stages of cancer progression (Fig. 3.2). We also ran a complementary analysis to identify differentially regulated genes in specifically in TNBC. We found 1127 genes to be up-regulated and 1752 genes down-regulated across multiple stages of TNBC (Fig. 3.1a). The Gene Set Enrichment Analysis (GSEA) indicated that the up-regulated genes in TNBC are enriched for molecules involved in cell cycle regulation and chromatin organization ($p<0.001$) (Fig. 3.1b). Results of our GSEA analysis of genes differentially up-regulated in TNBC tumors correlated well with the previously reported, differentially regulated genes from an independent cohort ($p<0.001$) (Fig. 3.1b) [86], which provides an additional support for the relevance of our observations. While we found a higher abundance of down-regulated genes, compared to the up-regulated genes, no similar significant enrichment was observed within the pool of the down-regulated genes. Similar results were obtained for overall breast cancer (Fig. 3.2b). Taken together, these observations indicated that the application of our approach to the analysis of TCGA data allows identifying subsets of genes differentially regulated in TNBC tumors.

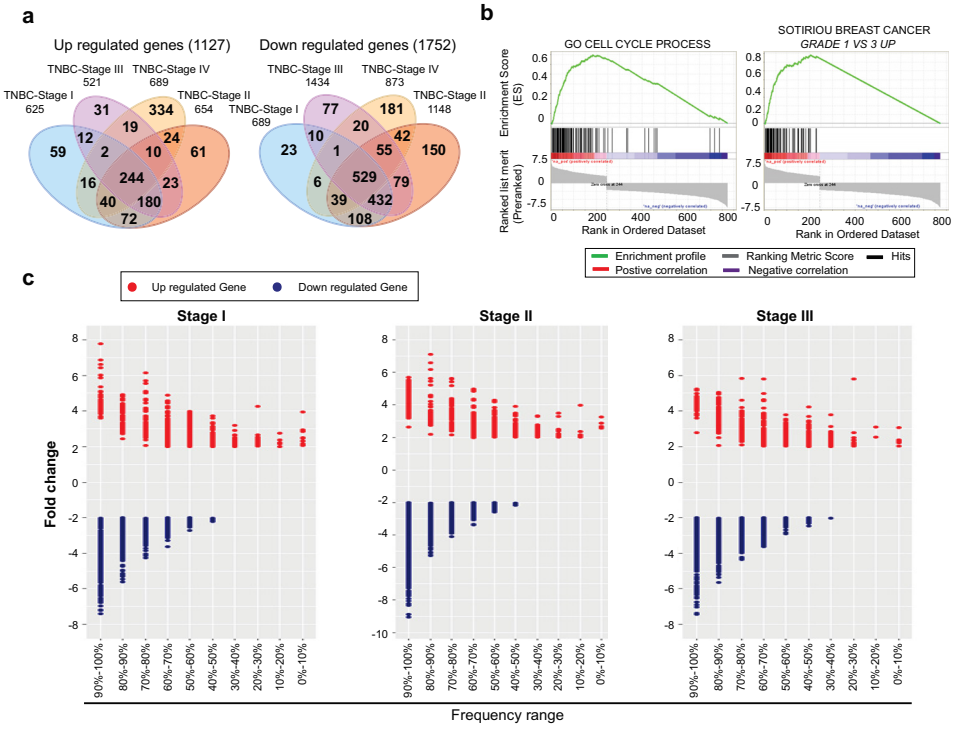


Figure 3.1: Identification of differentially expressed genes in TNBC. (a) Venn diagram of differentially expressed genes in TNBC stage-specific tumors. The number of up and down-regulated genes at each stage of tumor and at the intersection between different stages have been represented. (b) Gene set enrichment analysis for up-regulated genes across all TNBC tumor stages. Gene Set Enrichment Analyses for 244 up-regulated genes (left) across four tumor stages along with previously identified, differentially up-regulated genes (right from Sotiriou et al [86]. (c) Frequency distribution of differential expression in TNBC stage-specific tumors. Dot plot represents the fold change and the frequency range of TNBC stage-specific differentially expressed genes, where the red denotes up-regulated gene and the blue denotes down-regulated gene.

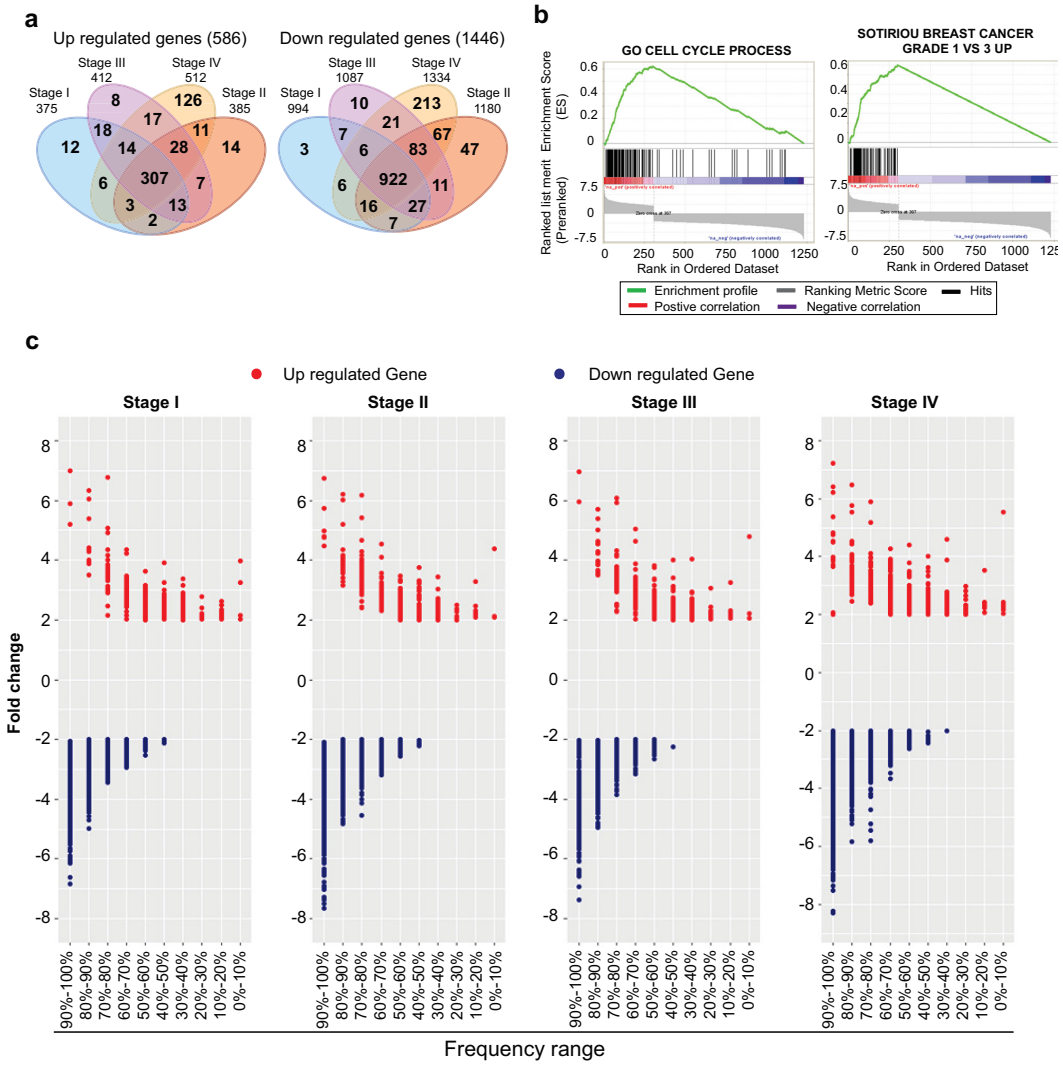


Figure 3.2: Identification of differentially expressed genes in overall breast cancer. (a) Venn diagram of differentially expressed genes in overall tumor samples. (b) Gene set enrichment analysis for up-regulated genes across all tumor stages. It shows the Gene Set Enrichment Analysis for 307 up-regulated genes (left) across four tumor stages along with previously identified, differentially up-regulated genes (right) from Sotiriou et al [86]. (c) Frequency distribution of differential expression in overall patients. This plot presents the fold change and the frequency range of stage-specific differentially expressed genes, where the red denotes up-regulated gene and the blue denotes down-regulated gene.

3.3.2 Not all differentially expressed genes are equally deregulated across the population of breast cancer patients.

While gene expression analysis to identify differentially regulated genes has been a common approach in cancer biology, we attempted to determine, how many of these genes are aberrantly expressed with high frequencies across the population of TNBC patients. We rationalized that common aberrations found in all patients should have arisen earlier in the development of the malignancy, compared to alterations that were found only in a subset of patients. Therefore, we have calculated a frequency of differential expression of each affected gene in TNBC tumors or overall cancer tumors (Fig. 3.1c; Fig. 3.2c). Throughout this analysis, we maintained a two-fold change in the expression level as a minimum requirement for a gene to be considered differentially regulated. The frequency of changes in each differentially expressed gene is calculated as a percentage of patients in whom the gene is up or down-regulated. We found 254 genes were up-regulated and 1197 genes were down-regulated in almost 70% of the TNBC patients. Same analysis was performed for overall breast cancer as well. Unfortunately, there were only two patient samples that were available in TNBC-stage IV in TCGA dataset, which was not sufficient to minimize random effects and carry out significance test. Therefore, we computed our analyses using the larger number of samples involved in the first three stages of TNBC.

Changes in gene expression may not only arise from aberrant expression from an endogenous promoter, but also from accompanying chromosomal amplifications, deletions and other types of mutations. To account for this, we isolated the differentially regulated genes exclusively from the non-amplified/deleted regions of the genome. We identified 77 amplified chromosome regions from the TCGA dataset based on CNA, including several previously reported regions in 1q, 8q, 16p and 20q (Table 3.1) [30], as presented in the circos plot for TNBC (Fig. 3.3a) or overall breast cancer (Fig. 3.4a). We further evaluated the concordance of amplification and gene expression by fold

change with FDR and pearson correlation coefficients. We considered genes likely to be driven by CNA if their pearson correlation coefficient between expression and CNA was greater than 0.3, or they show significant differential CNA-associated expression change (Fig. 3.4b,c). Subsequently, we filtered out from our analysis 20 genes from TNBC patients that were in amplified regions and had strong correlations with chromosome amplification.

We also used somatic mutational analyses of 560 breast cancer whole genome sequencing database available at COSMIC to eliminate any gene that might be differentially expressed because of a mutation [72]. By also excluding 13 genes whose loci information was ambiguous, we finally identified 219 exclusively expression-altered (EEA) genes that elevated their expression in TNBC (Fig. 3.3b) and therefore, may represent good therapeutic targets. Interestingly, we observed multiple distinct patterns of up-regulation with varying frequencies across different cancer stages (Fig. 3.3b). For example, some genes were constitutively up-regulated across all stages (PLK1, UBE2C or KIF4A). Similarly, certain genes were up-regulated mostly at later stages (CCNE1, HMGB3 or NUF2). In contrast to this category, some genes were up-regulated selectively at early stages but were gradually down-regulated through the later stages (MMP1, MMP11 or MMP13). Among the 219 up-regulation events, majority of changes occurred in chromosome 1 and 17 (Fig. 3.4d). Surprisingly, although the expression of some initially up-regulated genes gradually decreased, we did not observe any instance where their expression returned back to normal levels (Fig 3.5).

3.3.3 Progression analysis based on the up-regulation status of EEA genes

Since each sample can provide a snapshot of the accumulation process of the molecular changes [90] [31], we considered each tumor sample is recorded at a specific time point during the accumulation of aberrant gene expression. We expected to gain a general understanding about how the aberrant expression of EEA genes dynamically accumulated by analyzing the pattern of aberrant expression across TNBC patients. First, based on the profile of up-regulation

Table 3.1: Identified amplified chromosome regions.

cytoband	start_position	end_position	cytoband	start_position	end_position
1p11.2	120600000	121500000	20q11.22	32100000	34400000
1q21.1	142600000	147000000	20q11.23	34400000	37600000
1q21.2	147000000	150300000	20q12	37600000	41700000
1q21.3	150300000	155000000	20q13.11	41700000	42100000
1q22	155000000	156500000	20q13.12	42100000	46400000
1q23.1	156500000	159100000	20q13.13	46400000	49800000
1q23.2	159100000	160500000	20q13.2	49800000	55000000
1q23.3	160500000	165500000	20q13.31	55000000	56500000
1q24.1	165500000	167200000	20q13.32	56500000	58400000
1q24.2	167200000	170900000	20q13.33	58400000	63025520
1q24.3	170900000	172900000	8p11.1	43100000	45600000
1q25.1	172900000	176000000	8q11.1	45600000	48100000
1q25.2	176000000	180300000	8q11.21	48100000	52200000
1q25.3	180300000	185800000	8q11.22	52200000	52600000
1q31.1	185800000	190800000	8q11.23	52600000	55500000
1q31.2	190800000	193800000	8q12.1	55500000	61600000
1q31.3	193800000	198700000	8q12.2	61600000	62200000
1q32.1	198700000	207200000	8q12.3	62200000	66000000
1q32.2	207200000	211500000	8q13.1	66000000	68000000
1q32.3	211500000	214500000	8q13.2	68000000	70500000
1q41	214500000	224100000	8q13.3	70500000	73900000
1q42.11	224100000	224600000	8q21.11	73900000	78300000
1q42.12	224600000	227000000	8q21.12	78300000	80100000
1q42.13	227000000	230700000	8q21.13	80100000	84600000
1q42.2	230700000	234700000	8q21.2	84600000	86900000
1q42.3	234700000	236600000	8q21.3	86900000	93300000
1q43	236600000	243700000	8q22.1	93300000	99000000
1q44	243700000	249250621	8q22.2	99000000	101600000
16p13.3	0	7900000	8q22.3	101600000	106200000
16p13.2	7900000	10500000	8q23.1	106200000	110500000
16p13.13	10500000	12600000	8q23.2	110500000	112100000
16p13.12	12600000	14800000	8q23.3	112100000	117700000
16p13.11	14800000	16800000	8q24.11	117700000	119200000
16p12.3	16800000	21200000	8q24.12	119200000	122500000
16p12.2	21200000	24200000	8q24.13	122500000	127300000
16p12.1	24200000	28100000	8q24.21	127300000	131500000
16p11.2	28100000	34600000	8q24.22	131500000	136400000
20q11.21	29400000	32100000	8q24.23	136400000	139900000
			8q24.3	139900000	146364022

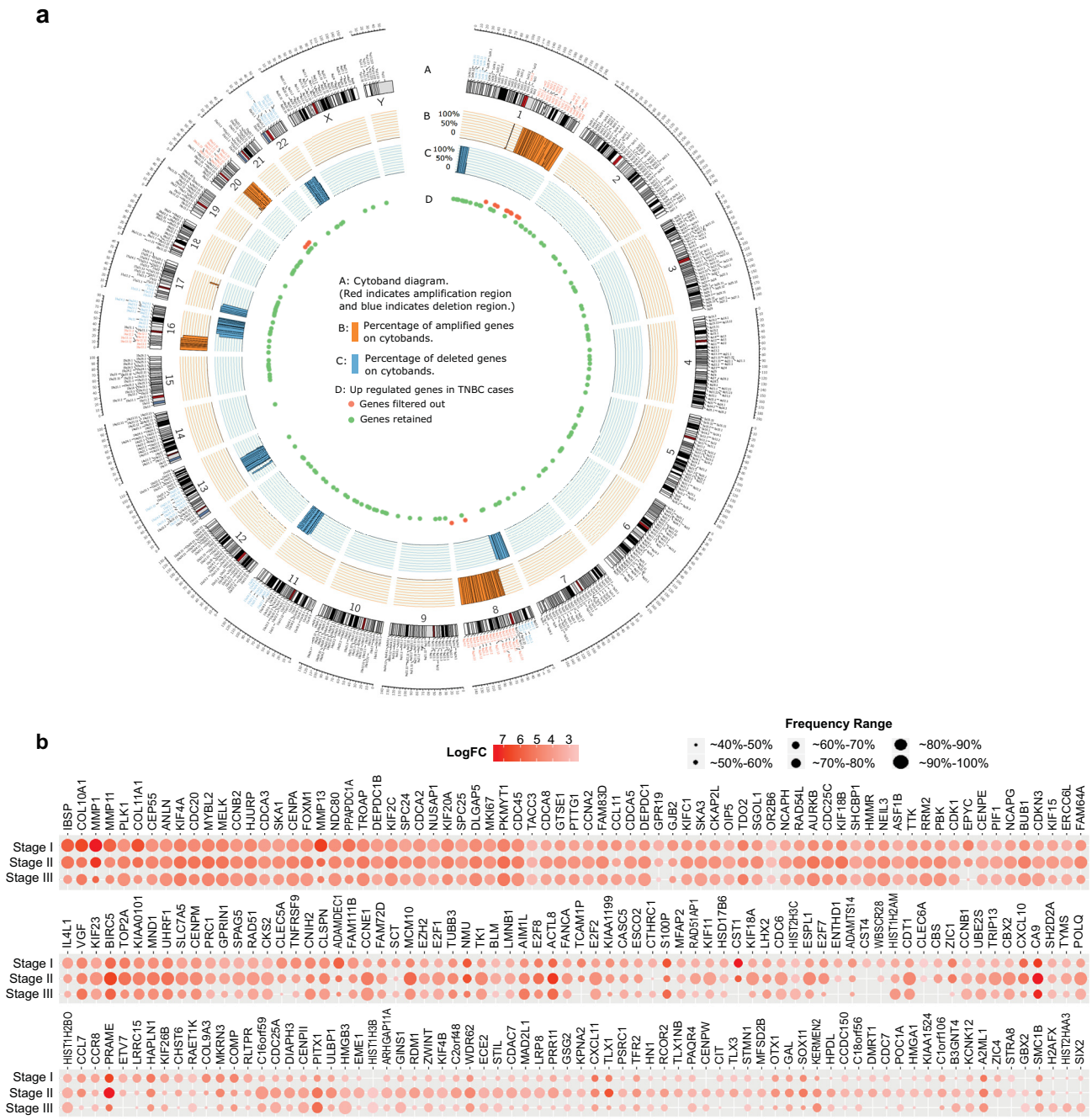


Figure 3.3: Elimination of amplified genes to identify 219 up-regulated events. (a) Amplified chromosome cytobands and up-regulated genes locus. Track A displays the cytoband diagram where the texts in red indicate identified amplified regions. Track B and C display the frequency of genes showing amplification and deletion respectively in at least in 40% of patients in each cytoband. Genes in Fig. 3.1a were mapped to the Track D. (b) Fold change and frequency distribution for genes showing up-regulation in at least 70% of TNBC patients. Nodes in each column represent up-regulated genes with their sizes indicating the frequency of samples and their colors representing the fold change in the specific tumor stage.

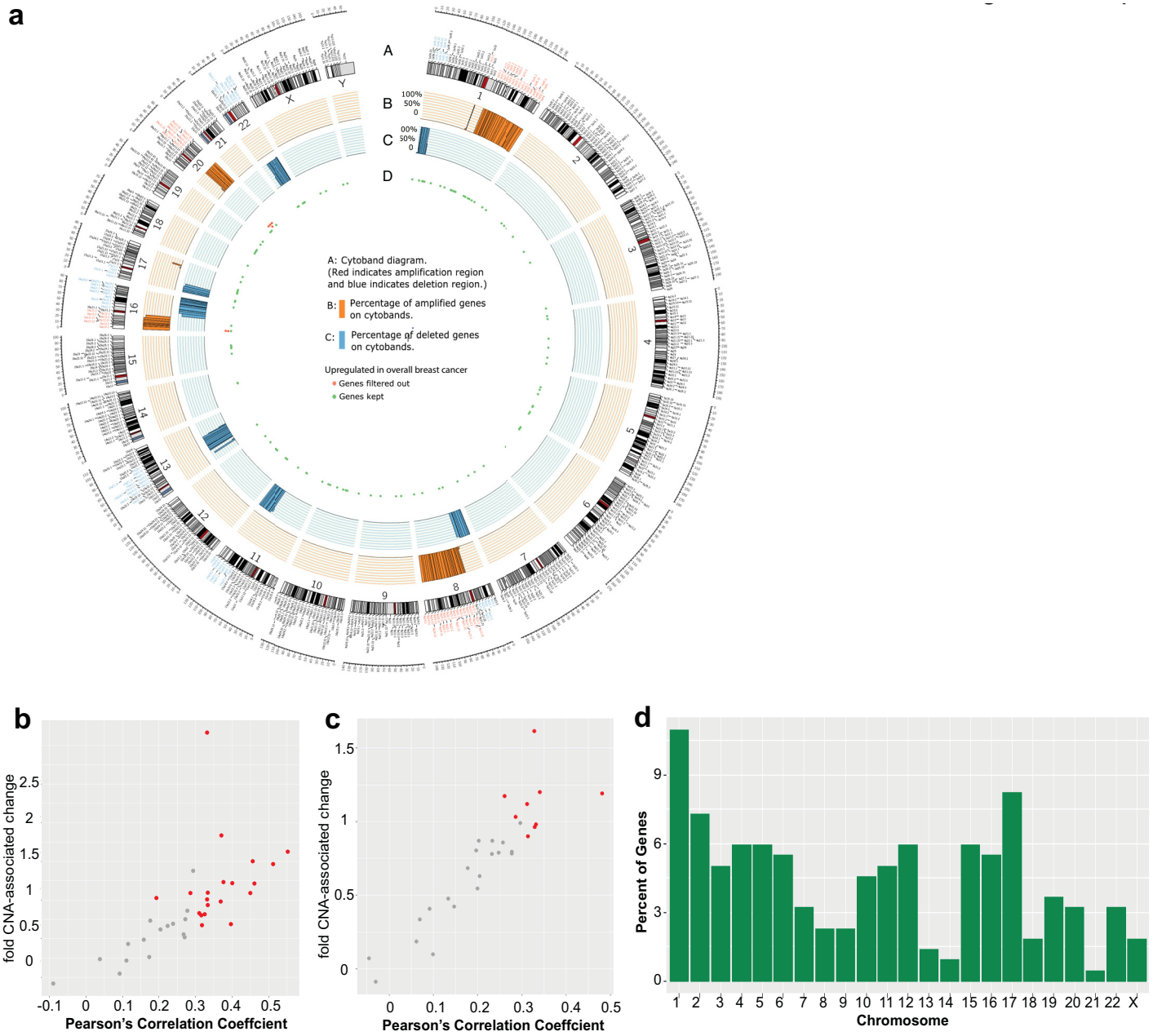


Figure 3.4: Expression pattern of up-regulated genes. (a) Amplified chromosome cytobands and up-regulated genes locus. Track A displays the cytoband diagram where the texts in red indicate identified amplified regions. Track B and C display the frequency of genes showing amplification and deletion respectively at least in 40% of patients in each cytoband. Genes in Fig. 3.2a were mapped to the Track D. (b and c) The evaluation on the concordance between gene expression and amplification. Nodes represent up-regulated genes in overall breast cancer cases in amplified regions, showing pearson correlation coefficient and fold CNA-associated change. Genes in red were considered driven by copy number amplification either in overall breast cancer (b) or in TNBC (c). (d) Distribution of the 219 up-regulated events according to their chromosomal location.

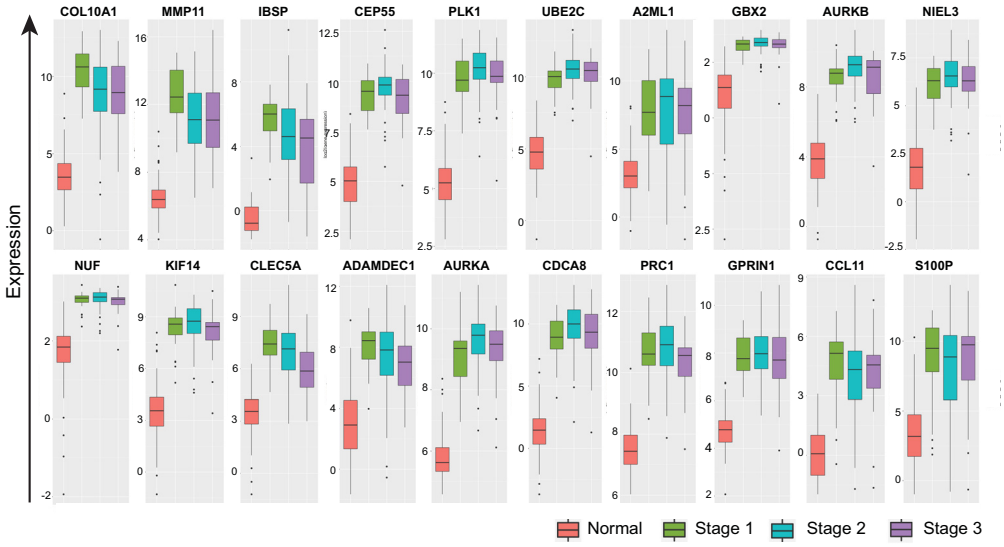


Figure 3.5: Box plots of gene expression at various stages of TNBC tumor. The y-axis represents log2-transformed gene expression and x-axis denotes TNBC stages.

status of EEA genes, we partitioned TNBC samples into groups so that the samples within a group have more similar up-regulation profiles than other samples in different groups. Hierarchical clustering was applied to group tumor samples into clusters. The cluster structure was graphically represented in Fig. 3.6a, which revealed that the most distinguishable cluster C1 is diverged at the highest overhang with the highest dissimilarities from the remaining samples. In addition, several distinguishable branches C2, C3 and C4 were also clustered.

To gain insights into the progression path in the context of accumulation of aberrant expression, we used the vector with 219 elements of all zeros representing no up-regulations of EEA genes in normal status as the root vertex, and constructed a tree-like structure by neighbour-joining to evaluate the extent of the accumulation of up-regulatory events for each cluster(Fig. 3.6b). It showed that cluster C1 is most similar to normal status since it has the shortest distance from root vertex, which suggested up-regulatory events occurred in C1 may act as early events in the early step of the progression path. By measuring the cosine similarity between the up regulation profile and subset vector, we obtained the occurrence of those up-regulatory events in each clus-

ter. We found that a burst of 83 up-regulation events occurred earliest in C1 and 205 up-regulatory events occurred in C2 including 76 events in C1. Also, 211 and 219 up regulation events occurred in C3 and C4, respectively.

We considered the 83 genes occurred in C1 may act as early potential enabling factors within the early tumor progression. To confirm the relevance, we next inquired if these changes in gene expression correlate with the loss of expression of known tumor suppressors. Vogelstein and colleagues identified 70 tumor suppressor genes that when inactivated by intragenic mutations can promote tumorigenesis [103]. We found a strong negative correlation in the expression of the 83 EEA genes and the 74 tumor suppressors (Fig. 3.6c).

3.3.4 TME remodeling and CIN cooperatively drive TNBC

Since the analysis indicated that the 83 up-regulated EEA genes are crucial early events in early tumorigenesis, we next explored the functionalities of these genes. Interestingly, we found a large subset of genes that are known to be involved in remodeling TME, including metalloproteinases (MMP1, MMP11, MMP13, ADAMDEC1, ADAMTS14), chemokine receptors and ligands (CXCL11, CXCL10, CCL11, CCR8), protease inhibitors (CST4, CST1), pH maintenance factors (CAIX), and different collagens (COL9A3, COL10A1). This emphasizes the critical role of extra-cellular matrix and TME remodeling in early tumor progression. Similarly, we also identified several of the EEA genes including, FOXM1, PLK1, BUB1, KIF2C, CDCA2, CDC20, CDKN3, KNL1 to name a few, that are known for their role in CIN and tumor development [96][59][19][77]. This may reflect a selective pressure for additional genetic alterations in early tumors that would allow their further progression. In addition, cluster 1 included genes like DEPDC1B and HMMR that have known roles in both TME remodeling as well as CIN associated functions. Overall, our identification of cluster 1 genes indicates that a burst of expressional changes occurs simultaneously in both CIN and TME remodeling genes very early in tumor development. The literature evidence for the role of cluster 1 genes in TME and CIN was listed in Appendix B.1. If both CIN and TME remodeling ensue simultaneously, we should ask what possible

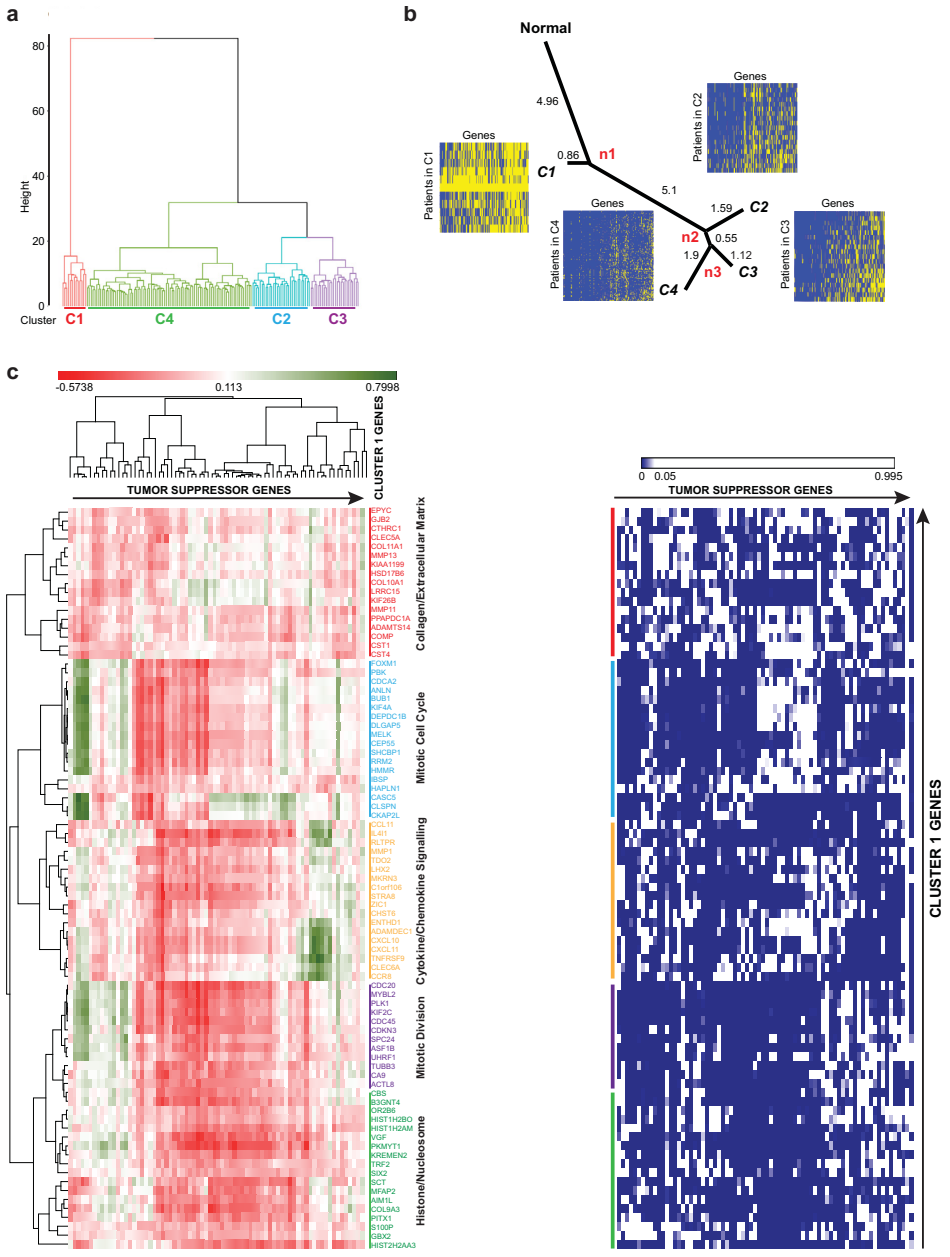


Figure 3.6: Analysis based on the up-regulation profile of EEA genes. (a) Hierarchical clustering. Different colors show TNBC patients clustered into four clusters, represented as Red for Cluster 1 (C1), Purple for Cluster 2 (C2), Blue for Cluster 3 (C3) and Green for Cluster 4 (C4). (b) The accumulation of up-regulatory events along the progression line. The figure shows the progression path based on gene up-regulations from the normal to each cluster. Heat maps with genes in columns and TNBC samples in rows display the up-regulation status (yellow: no up-regulation; blue: up-regulation) for different TNBC clusters. (c) Correlation clustergram of cluster 1 genes compared to known tumor suppressors. Red indicates negative correlation and green indicates positive correlation. The panel on the right represents the significance of the correlation as a heat map. Blue indicates significance (<0.05) and white indicates lack of significance (>0.05).

factors could drive such changes. To address this, we used recently published causal analyses tools [47] available in the Ingenuity Pathway Analysis(IPA). In particular, we performed Upstream Regulator Analysis, and Causal Network Analysis to curate all interactions of cluster 1 genes (Fig. 3.7a,b). Interestingly, a large subset of direct up-stream interactions as well as causal interactions of both the CIN and TME genes (cluster 1), are hypoxia responsive genes [45] (Fig. 3.7a,b; Appendix B.2 and B.3). Invariably, almost 50% of the cluster 1 genes are also associated with poor prognosis (Fig. 3.8a and Fig. 3.9). This strongly suggests that very early in the course of tumor progression gradually increasing hypoxic conditions induce both CIN and TME remodeling to permit survival of cancer cells and their further evolution at later stages of malignancy. Having identified a set of 83 EEA genes that act in early TNBC tumors, we sought to identify drugs that can benefit TNBC treatment and may potentially be also used for cancer prevention. To do this, we selected breast cancer cell lines that overexpress cluster 1 genes and analyzed their sensitivity to drugs using the cancerRXgene database (http://www.cancerrxgene.org). This database provides information on cell line drug sensitivity. The data for 265 drugs and multiple cell lines was examined to identify compounds that are more effective when used selectively with cell lines that highly express cluster 1 genes. We found four drugs, bleomycin, pevonedistat, ponatinib, and WIKI4, that showed a significant decrease in the IC50, for cell lines that highly expressed cluster 1 genes (Fig. 3.8b). Consistent with our identification of several cluster 1 genes being involved in CIN (Fig. 3.7a), our drug analyses indicate that cell lines with high expression of cluster 1 genes are more sensitive to a DNA damaging agent, bleomycin (Fig 3.8b).

3.4 Discussion

Differential gene expression analyses have been traditionally used to examine fluctuations within the transcriptome in a given context for decades. This has been a powerful strategy to identify biomarkers and drug targets. However, tumor genome sequencing has provided new opportunities to re-examine these

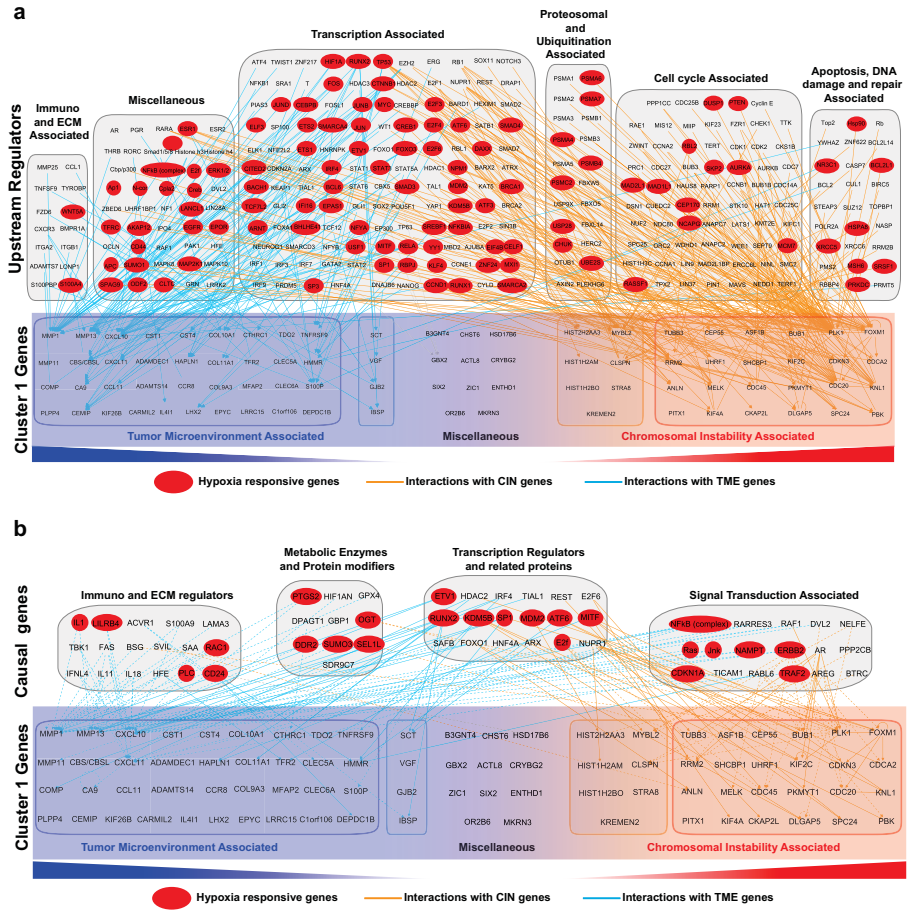


Figure 3.7: IPA analyses showing extensive interaction between hypoxia responsive genes with members of cluster 1 genes. (a) Upstream regulator analysis was performed with IPA for the cluster 1 genes and all the interactions retrieved are presented. Cluster 1 genes are classified into those that are associated with CIN or TME. The upstream genes that are hypoxia responsive, are highlighted in red. (b) Causal network analysis was performed with IPA for the cluster 1 genes and all the interactions retrieved are presented. Cluster 1 genes are classified into those that are associated with CIN or TME. The upstream genes that are hypoxia responsive, are highlighted in red.

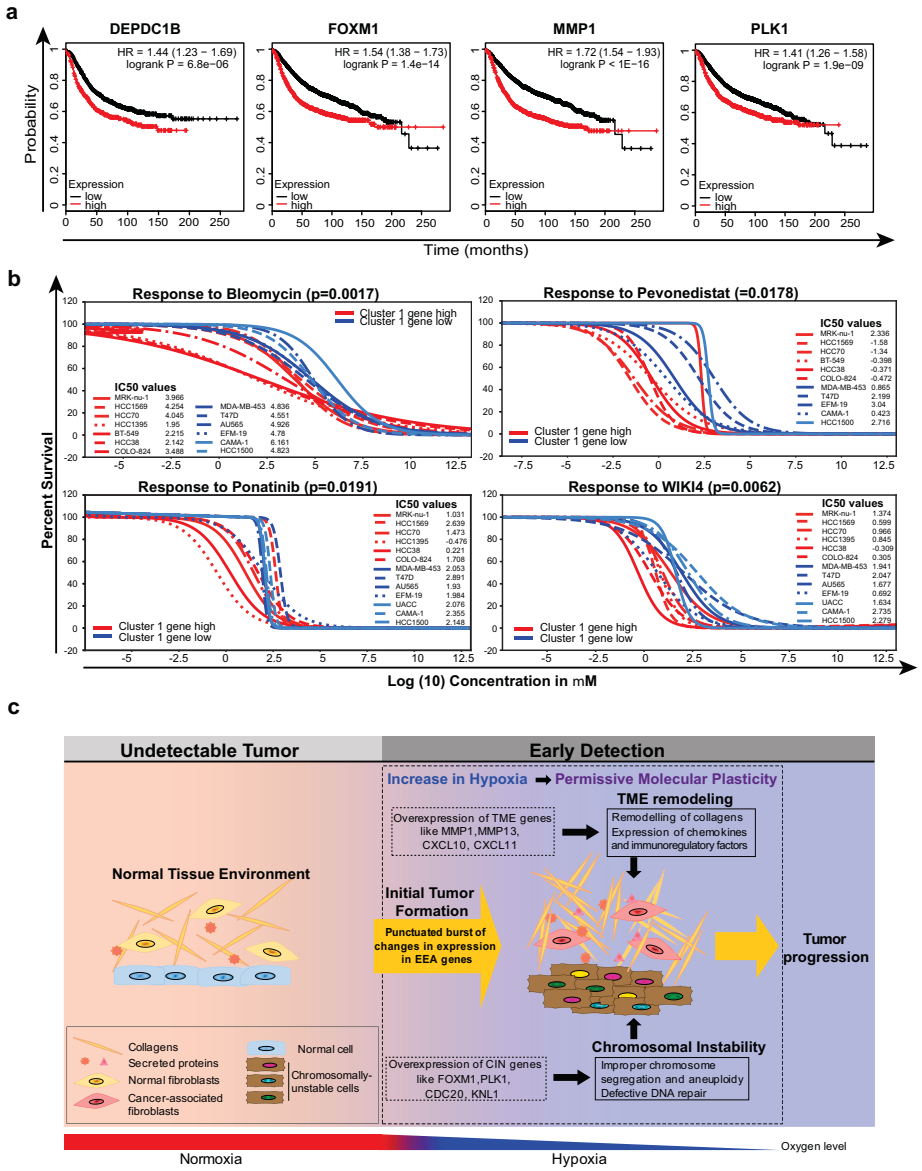


Figure 3.8: Survival plot, Drug response and a model describing the role of cluster 1 genes in tumor evolution. (a) Representative relapse free survival plots of breast cancer patients with low and high expression of cluster 1 genes (b) Dose response curves and IC50 values of drugs targeting cell lines with low and high expression of cluster 1 genes. (c) Schematic model showing the effect of simultaneous burst of CIN and TME-associated genes in response to hypoxia during early cancer initiation.

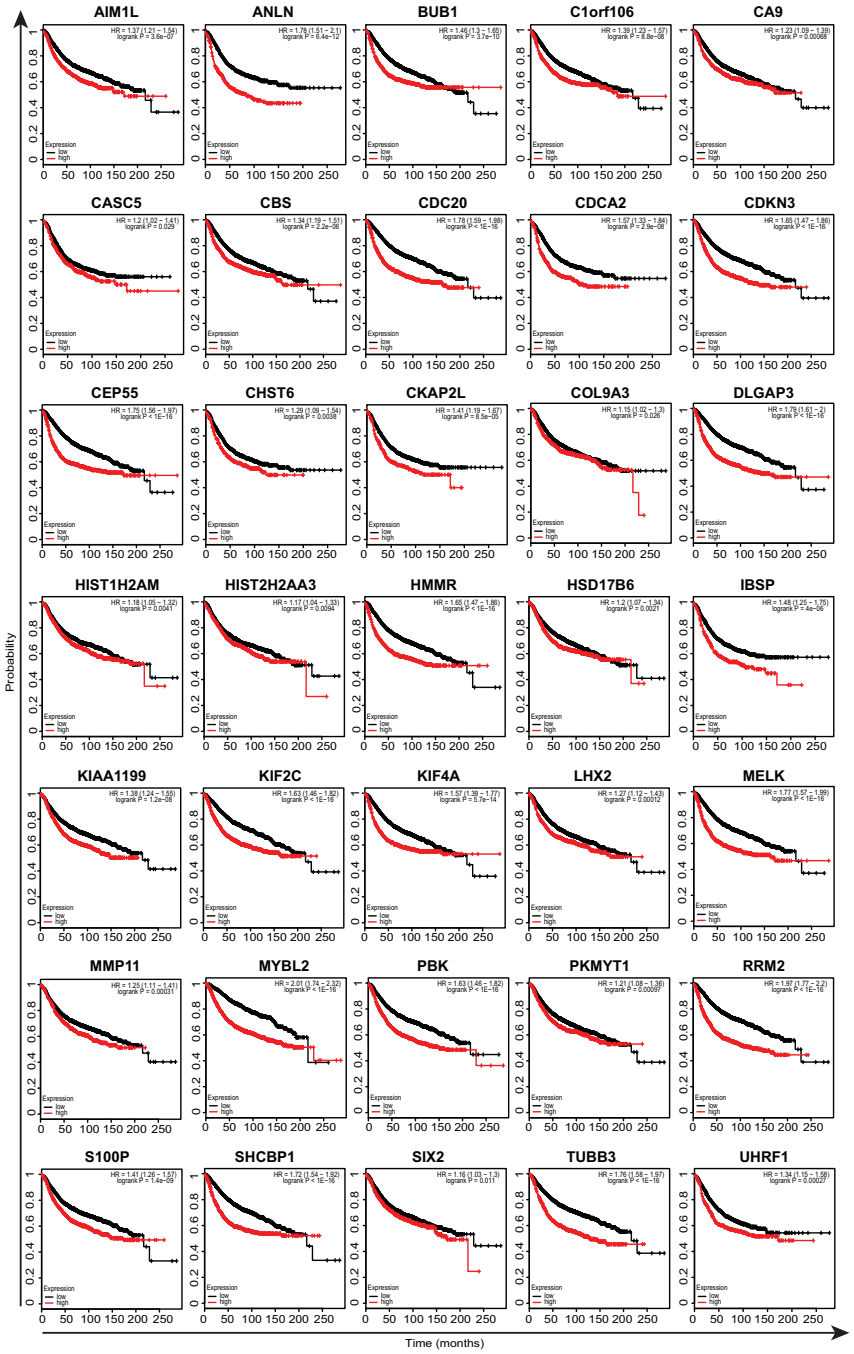


Figure 3.9: Relapse free survival plot in breast cancer patients having low and high expression of cluster 1 genes.

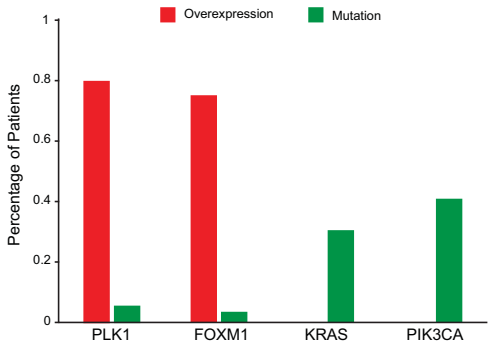


Figure 3.10: Percentage of patients that either overexpress or carry mutations in some of the key cancer genes.

fluctuations in the context of tumor progression. We rationalize that common aberrations detected across all patients arose earlier in the development of the malignancy compared to alterations that were found only in a subset of patients. Based on this, our strategy in this work is to explore the frequency of changes in the expression pattern of genes at different stages of TNBC progression. This is similar to previous studies that explored dynamic changes in mutations or CNA for a given patient at a given stage [89][90][69]. Changes in gene expression, unless constitutively observed, are often ignored as stochastic noise, specifically those that arise from variations in transcriptional regulation or biochemical modifications within cells. Our analyses deliver a number of important observations. First, compared to mutational changes, alterations within the transcriptome are more common and occur at high frequency. For example, the highly significant mutations in genes like PIK3CA or KRAS are observed in 30% of breast cancer patients. In contrast, overexpression of PLK1 or FOXM1 genes is observed in over 90% of patients (Fig. 3.10). Second, more genes are down-regulated compared to up-regulated genes. Third, during cancer progression, the initial singular burst of changes in expression pattern, results in simultaneous accumulation of overexpression of multiple EEA genes. Fourth, early changes in the expression of EEA molecules occur in genes that remodel TME and maintain chromosomal stability. This is most likely because survival within the progressively changing biological landscape during early stages requires cancer cells to both actively adjust to their mi-

croenvironment for their needs and to enhance CIN to facilitate their plasticity and adapt. Indeed, our unbiased genome-wide investigation reveals a strong functional connection between these two mechanisms and a crucial role of their coordinated effort in establishing early tumors. Interestingly, some TME genes, including MMP1, MMP11 and MMP13 proved to be up-regulated at early stages and gradually down-regulated through the later stages, although never achieving their normal levels. This suggests that their activities are essential at all stages of cancer progression, but their higher activity is required in early tumors, where the TME is not adjusted yet to the needs of malignant cells. Fifth, while we know that hypoxic TME can trigger tumor metastasis and invasion at later stages of cancer progression, our causal network analyses suggest that increasing hypoxia may be responsible for the cooperative induction of CIN and TME remodeling much earlier than previously appreciated (Fig. 3.8c). As hypoxic environment is also known to promote the propagation of tumor initiating cells (TICs) [60][13], we suspect that the expressional changes of EEA genes may facilitate this process. This is consistent with our finding that drugs like bleomycin and WIKI4 that efficiently eliminate TIC-enriched cell populations, cause selective lethality to cancer cell lines that overexpress cluster 1 genes (Fig. 3.8b).

Although CIN is nearly ubiquitous in cancer cells, and is considered as an important factor in tumor development, our findings indicate that hypoxic TME of early tumor may function as a trigger of genetic instability. This model is consistent with previous observations, showing that repeated cycles of hypoxia, can down-regulate a number of DNA repair pathways in cancer cells, ultimately leading to genetic instability. In regards to this, the Glazer group has provided one of the first quantitative assessments of how genetic instability can be instigated by TME [76]. Interestingly, several of the core EEA genes that maintain genome stability were experimentally shown to be involved in tumor development [96][59][19][77]. Although some of these examples might be indicative of a direct role for CIN genes in tumorigenesis, in the context of our analyses, we suggest that overexpression of these genes may have enabled cancer cells to acquire properties that allowed them to survive at the early

time and thus, to develop detectable tumors (Fig. 3.8c). In summary, our unbiased comprehensive analyses of the transcriptome directly link the early onset of hypoxia to the collective burst of CIN and TME remodeling factors, which highlights a therapeutic potential of targeting these molecules in TNBC tumors in their earliest detectable stage.

Chapter 4

Conclusion

Recent technology has enabled the multi-platform genomic profiling of biological samples, resulting in genome-wide genetic data in multiple types of cancer. The discovery of cancer biomarkers is of great importance to understanding the biological mechanism of cancer and providing insights into the early diagnosis and efficient treatment of cancer. In this thesis, we have presented a computational method for the discovery of epigenetic mechanisms and a computational analysis based on the aberrant expression in breast cancer, both of which contribute to the discovery of cancer-related biomarkers and mechanisms.

In Chapter 2, we proposed a novel method for the systematic analysis between alterations of DNA methylation and gene expression at the module level using a Bayesian regression model with the incorporation of prior gene-interaction knowledge. We discovered the dependency between DNA methylation predictor and gene expression response, which contained verified epigenetic causal relationships. The results revealed that the detected epigenetic subnetworks are significantly enriched in multiple cancer-related pathways. Hence, they are of great biological relevance and could be a starting point to uncover underlying epigenetic mechanisms.

In Chapter 3, we performed a computational analysis based on aberrant gene expression in breast cancer. We explored the frequency of expressional changes at different stages of TNBC and investigated the pattern of aberrant expression in cancer progression. Our analysis delivered a number of important

observations. Compared to mutational changes, alterations within the transcriptome were more common and occurred at a high frequency. The initial singular burst of changes in expression resulted in simultaneous accumulation of overexpression of multiple EEA genes. Early changes in the expression of EEA molecules occur in genes that remodel TME and maintain chromosomal stability. Also the analysis highlighted a therapeutic potential of targeting these molecules in TNBC tumors.

References

[1] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A.-L. Børresen-Dale, *et al.*, "Signatures of mutational processes in human cancer," *Nature*, vol. 500, no. 7463, p. 415, 2013. 59

[2] O. Alter, P. O. Brown, and D. Botstein, "Singular value decomposition for genome-wide expression data processing and modeling," *Proceedings of the National Academy of Sciences*, vol. 97, no. 18, pp. 10 101–10 106, 2000. 29

[3] T. Arcondeguy, E. Lacazette, S. Millevoi, H. Prats, and C. Touriol, "Vegf-a mrna processing, stability and translation: A paradigm for intricate regulation of gene expression at the post-transcriptional level," *Nucleic acids research*, vol. 41, no. 17, pp. 7997–8010, 2013. 60

[4] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, *et al.*, "Gene ontology: Tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, p. 25, 2000. 49

[5] T. D. Barber, K. McManus, K. W. Yuen, M. Reis, G. Parmigiani, D. Shen, I. Barrett, Y. Nouhi, F. Spencer, S. Markowitz, *et al.*, "Chromatid cohesion defects may underlie chromosome instability in human colorectal cancers," *Proceedings of the National Academy of Sciences*, vol. 105, no. 9, pp. 3443–3448, 2008. 59

[6] K. R. Bauer, M. Brown, R. D. Cress, C. A. Parise, and V. Caggiano, "Descriptive analysis of estrogen receptor (er)-negative, progesterone receptor (pr)-negative, and her2-negative invasive breast cancer, the so-called triple-negative phenotype," *Cancer*, vol. 109, no. 9, pp. 1721–1728, 2007. 2

[7] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300, 1995. 14

[8] C. Bernstein, V. Nfonsam, A. R. Prasad, and H. Bernstein, "Epigenetic field defects in progression to cancer," *World journal of gastrointestinal oncology*, vol. 5, no. 3, p. 43, 2013. 2

[9] R. Beroukhim, C. H. Mermel, D. Porter, G. Wei, S. Raychaudhuri, J. Donovan, J. Barretina, J. S. Boehm, J. Dobson, M. Urashima, *et al.*, "The landscape of somatic copy-number alteration across human cancers," *Nature*, vol. 463, no. 7283, p. 899, 2010. 2

[10] D. P. Bertsekas, "Projected newton methods for optimization problems with simple constraints," *SIAM Journal on control and Optimization*, vol. 20, no. 2, pp. 221–246, 1982. 19

[11] K. L. Bolton, G. Chenevix-Trench, C. Goh, S. Sadetzki, S. J. Ramus, B. Y. Karlan, D. Lambrechts, E. Despierre, D. Barrowdale, L. McGuffog, *et al.*, "Association between brca1 and brca2 mutations and survival in women with invasive epithelial ovarian cancer," *Jama*, vol. 307, no. 4, pp. 382–389, 2012. 1

[12] N. E. Breslow, "Analysis of survival data under the proportional hazards model," *International Statistical Review/Revue Internationale de Statistique*, pp. 45–57, 1975. 33

[13] A. Carnero and M. Lleonart, "The hypoxic microenvironment: A determinant of cancer stem cell evolution," *Inside the Cell*, vol. 1, no. 2, pp. 96–105, 2016. 82

[14] J. K. Choi, U. Yu, O. J. Yoo, and S. Kim, "Differential coexpression analysis using microarray data and its application to human cancer," *Bioinformatics*, vol. 21, no. 24, pp. 4348–4355, 2005. 6

[15] G. Ciriello, M. L. Gatza, A. H. Beck, M. D. Wilkerson, S. K. Rhie, A. Pastore, H. Zhang, M. McLellan, C. Yau, C. Kandoth, *et al.*, "Comprehensive molecular portraits of invasive lobular breast cancer," *Cell*, vol. 163, no. 2, pp. 506–519, 2015. 1, 2

[16] W. contributors, *Null hypothesis — wikipedia, the free encyclopedia*, [Online; accessed 20-March-2018], 2018. [Online]. Available: `\url{https://en.wikipedia.org/w/index.php?title=Null_hypothesis&oldid=830713968}`. 16, 17

[17] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, *et al.*, "The reactome pathway knowledgebase," *Nucleic acids research*, vol. 42, no. D1, pp. D472–D477, 2013. 49

[18] T. Davoli, H. Uno, E. C. Wooten, and S. J. Elledge, "Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy," *Science*, vol. 355, no. 6322, eaaf8399, 2017. 2

[19] G. De Cárcer and M. Malumbres, "A centrosomal route for cancer genome instability," *Nature cell biology*, vol. 16, no. 6, p. 504, 2014. 75, 82

[20] J. De Las Rivas and C. Fontanillo, "Protein–protein interactions essentials: Key concepts to building and analyzing interactome networks," *PLoS computational biology*, vol. 6, no. 6, e1000807, 2010. 41

[21] S. Dedeurwaerder, M. Defrance, E. Calonne, H. Denis, C. Sotiriou, and F. Fuks, "Evaluation of the infinium methylation 450k technology," *Epigenomics*, vol. 3, no. 6, pp. 771–784, 2011. 15

[22] Y. Ding, "Prognostic biomarker detection, machine learning bias correction, and differential coexpression module detection," PhD thesis, University of Pittsburgh, 2014. 13

[23] A. Foroushani, R. Agrahari, R. Docking, L. Chang, G. Duns, M. Hudoba, A. Karsan, and H. Zare, "Large-scale gene network analysis reveals the significance of extracellular matrix pathway and homeobox genes in acute myeloid leukemia: An introduction to the pigengene package and its applications," *BMC medical genomics*, vol. 10, no. 1, p. 16, 2017. 29

[24] A. de la Fuente, "From ?differential expression?to ?differential networking?– identification of dysfunctional regulatory networks in diseases," *Trends in genetics*, vol. 26, no. 7, pp. 326–333, 2010. 6

[25] J. B. Geigl, A. C. Obenauf, T. Schwarzbraun, and M. R. Speicher, "Defining ?chromosomal instability?" *Trends in Genetics*, vol. 24, no. 2, pp. 64–69, 2008. 59

[26] O. Gevaert, R. Tibshirani, and S. K. Plevritis, "Pancancer analysis of dna methylation-driven genes using methylmix," *Genome biology*, vol. 16, no. 1, p. 17, 2015. 8

[27] S. M. Gollin, "Mechanisms leading to chromosomal instability," in *Seminars in cancer biology*, Elsevier, vol. 15, 2005, pp. 33–42. 59

[28] J. Gray and B. Druker, "Genomics: The breast cancer landscape," *Nature*, vol. 486, no. 7403, p. 328, 2012. 1

[29] M Grigorova, J. Staines, H Ozdag, C Caldas, and P. Edwards, "Possible causes of chromosome instability: Comparison of chromosomal abnormalities in cancer cell lines with mutations in brca1, brca2, chk2 and bub1," *Cytogenetic and genome research*, vol. 104, no. 1-4, pp. 333–340, 2004. 59

[30] J. Guenthoer, S. J. Diede, H. Tanaka, X. Chai, L. Hsu, S. J. Tapscott, and P. L. Porter, "Assessment of palindromes as platforms for dna amplification in breast cancer," *Genome research*, vol. 22, no. 2, pp. 232–245, 2012. 69

[31] A. Gupta and Z. Bar-Joseph, "Extracting dynamics from static cancer expression data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 5, no. 2, pp. 172–182, 2008. 63, 70

[32] R. S. Harris, "Molecular mechanism and clinical impact of apobec3b-catalyzed mutagenesis in breast cancer," *Breast Cancer Research*, vol. 17, no. 1, p. 8, 2015. 59

[33] S. Heerboth, K. Lapinska, N. Snyder, M. Leary, S. Rollinson, and S. Sarkar, "Use of epigenetic drugs in disease: An overview," *Genetics & epigenetics*, vol. 6, GEG–S12270, 2014. 7

[34] T. Hinoue, D. J. Weisenberger, C. P. Lange, H. Shen, H.-M. Byun, D. Van Den Berg, S. Malik, F. Pan, H. Noushmehr, C. M. van Dijk, *et al.*, "Genome-scale analysis of aberrant dna methylation in colorectal cancer," *Genome research*, vol. 22, no. 2, pp. 271–282, 2012. 7

[35] S. Horvath and J. Dong, "Geometric interpretation of gene coexpression network analysis," *PLoS computational biology*, vol. 4, no. 8, e1000117, 2008. 30

[36] M. Hu, J. Yao, L. Cai, K. E. Bachman, F. van den Brûle, V. Velculescu, and K. Polyak, "Distinct epigenetic changes in the stromal cells of breast cancers," *Nature genetics*, vol. 37, no. 8, p. 899, 2005. 2, 7

[37] *Humanmethylation450 beadchip*, Illumina, Mar. 2012. 15, 16

[38] P. V. Jallepalli and C. Lengauer, "Chromosome segregation and cancer: Cutting through the mystery," *Nature Reviews Cancer*, vol. 1, no. 2, p. 109, 2001. 59

[39] M. Jeanmougin, A. De Reynies, L. Marisa, C. Paccard, G. Nuel, and M. Guedj, "Should we abandon the t-test in the analysis of gene expression microarray data: A comparison of variance modeling strategies," *PloS one*, vol. 5, no. 9, e12336, 2010. 14

[40] Y. Jiao, M. Widschwendter, and A. E. Teschendorff, "A systems-level integrative framework for genome-wide dna methylation and gene expression data identifies differential gene expression modules under epigenetic control," *Bioinformatics*, vol. 30, no. 16, pp. 2360–2366, 2014. 8, 15, 41, 55

[41] J. Jovanovic, J. A. Rønneberg, J. Tost, and V. Kristensen, "The epigenetics of breast cancer," *Molecular oncology*, vol. 4, no. 3, pp. 242–254, 2010. 51

[42] C. Kandoth, M. D. McLellan, F. Vandin, K. Ye, B. Niu, C. Lu, M. Xie, Q. Zhang, J. F. McMichael, M. A. Wyczalkowski, *et al.*, "Mutational landscape and significance across 12 major cancer types," *Nature*, vol. 502, no. 7471, p. 333, 2013. 1

[43] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "Kegg: New perspectives on genomes, pathways, diseases and drugs," *Nucleic acids research*, vol. 45, no. D1, pp. D353–D361, 2016. 49

[44] S. Kar, D. Sengupta, M. Deb, A. Shilpi, S. Parbin, S. K. Rath, N. Pradhan, M. Rakshit, and S. K. Patra, "Expression profiling of dna methylation-mediated epigenetic gene-silencing factors in breast cancer," *Clinical epigenetics*, vol. 6, no. 1, p. 20, 2014. 51

[45] P. Khurana, R. Sugadev, J. Jain, and S. B. Singh, "Hypoxiadb: A database of hypoxia-regulated proteins," *Database*, vol. 2013, 2013. 64, 77

[46] J. Klajic, F. Busato, H. Edvardsen, N. Touleimat, T. Fleischer, I. Bukholm, A.-L. Børresen-Dale, P. E. Lønning, J. Tost, and V. N. Kristensen, "Dna methylation status of key cell-cycle regulators such as cdkna2/p16 and ccna1 correlates with treatment response to doxorubicin and 5-fluorouracil in locally advanced breast tumors," *Clinical cancer research*, vol. 20, no. 24, pp. 6357–6366, 2014. 53

[47] A. Krämer, J. Green, J. Pollard Jr, and S. Tugendreich, "Causal analysis approaches in ingenuity pathway analysis," *Bioinformatics*, vol. 30, no. 4, pp. 523–530, 2013. 64, 77

[48] D. Kuang, C. Ding, and H. Park, "Symmetric nonnegative matrix factorization for graph clustering," in *Proceedings of the 2012 SIAM international conference on data mining*, SIAM, 2012, pp. 106–117. 19, 27

[49] C. Lahtz and G. P. Pfeifer, "Epigenetic changes of dna repair genes in cancer," *Journal of molecular cell biology*, vol. 3, no. 1, pp. 51–58, 2011. 2

[50] P. Langfelder and S. Horvath, "Wgcna: An r package for weighted correlation network analysis," *BMC bioinformatics*, vol. 9, no. 1, p. 559, 2008. 29

[51] P. Langfelder, R. Luo, M. C. Oldham, and S. Horvath, "Is my network module preserved and reproducible?" *PLoS computational biology*, vol. 7, no. 1, e1001057, 2011. 28, 29

[52] C. A. Lareau, B. C. White, A. L. Oberg, and B. A. McKinney, "Differential co-expression network centrality and machine learning feature selection for identifying susceptibility hubs in networks with scale-free structure," *BioData mining*, vol. 8, no. 1, p. 5, 2015. 25

[53] C. Laronga, H.-Y. Yang, C. Neal, and M.-H. Lee, "Association of the cyclin-dependent kinases and 14-3-3 sigma negatively regulates cell cycle progression," *Journal of Biological Chemistry*, vol. 275, no. 30, pp. 23 106–23 112, 2000. 53

[54] E. LaTulippe, J. Satagopan, A. Smith, H. Scher, P. Scardino, V. Reuter, and W. L. Gerald, "Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease," *Cancer research*, vol. 62, no. 15, pp. 4499–4506, 2002. 5

[55] D. Lawley, "A generalization of fisher's z test," *Biometrika*, vol. 30, no. 1/2, pp. 180–187, 1938. 17

[56] H.-J. Lee, T. C. Dang, H. Lee, and J. C. Park, "Oncosearch: Cancer gene search engine with literature evidence," *Nucleic acids research*, vol. 42, no. W1, W416–W421, 2014. 51

[57] H. K. Lee, A. K. Hsu, J. Sajdak, J. Qin, and P. Pavlidis, "Coexpression analysis of human genes across many microarray data sets," *Genome research*, vol. 14, no. 6, pp. 1085–1094, 2004. 6

[58] B. Li and C. N. Dewey, "Rsem: Accurate transcript quantification from rna-seq data with or without a reference genome," *BMC bioinformatics*, vol. 12, no. 1, p. 323, 2011. 13

[59] M. Li, X. Fang, Z. Wei, J. P. York, and P. Zhang, "Loss of spindle assembly checkpoint–mediated inhibition of cdc20 promotes tumorigenesis in mice," *The Journal of cell biology*, vol. 185, no. 6, pp. 983–994, 2009. 75, 82

[60] Q. Lin and Z. Yun, "Impact of the hypoxic tumor microenvironment on the regulation of cancer stem cell characteristics," *Cancer biology & therapy*, vol. 9, no. 12, pp. 949–956, 2010. 82

[61] C.-Y. Liu, L.-M. Tseng, J.-C. Su, K.-C. Chang, P.-Y. Chu, W.-T. Tai, C.-W. Shiau, and K.-F. Chen, "Novel sorafenib analogues induce apoptosis through shp-1 dependent stat3 inactivation in human breast cancer cells," *Breast Cancer Research*, vol. 15, no. 4, p. 3254, 2013. 53

[62] G. Liu, L. Wong, and H. N. Chua, "Complex discovery from weighted ppi networks," *Bioinformatics*, vol. 25, no. 15, pp. 1891–1897, 2009. 27

[63] X. Ma, Z. Liu, Z. Zhang, X. Huang, and W. Tang, "Multiple network algorithm for epigenetic modules via the integration of genome-wide dna methylation and gene expression data," *BMC bioinformatics*, vol. 18, no. 1, p. 72, 2017. 9, 16, 55

[64] J. Mazieres, S. Peters, B. Lepage, A. B. Cortot, F. Barlesi, M. Beau-Faller, B. Besse, H. Blons, A. Mansuet-Lupo, T. Urban, *et al.*, "Lung cancer that harbors an her2 mutation: Epidemiologic characteristics and therapeutic perspectives," *Journal of clinical oncology*, vol. 31, no. 16, pp. 1997–2003, 2013. 2

[65] N. McGranahan and C. Swanton, "Biological and therapeutic impact of intratumor heterogeneity in cancer evolution," *Cancer cell*, vol. 27, no. 1, pp. 15–26, 2015. 59

[66] R. Medina-Aguilar, C. Pérez-Plasencia, P. Gariglio, L. A. Marchat, A. Flores-Pérez, C. López-Camarillo, and J. G. Mena, "Dna methylation data for identification of epigenetic targets of resveratrol in triple negative breast cancer cells," *Data in brief*, vol. 11, pp. 169–182, 2017. 53

[67] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by rna-seq," *Nature methods*, vol. 5, no. 7, p. 621, 2008. 12

[68] C. Murie, O. Woody, A. Y. Lee, and R. Nadon, "Comparison of small n statistical tests of differential expression applied to microarrays," *BMC bioinformatics*, vol. 10, no. 1, p. 45, 2009. 14

[69] N. Navin, J. Kendall, J. Troge, P. Andrews, L. Rodgers, J. McIndoo, K. Cook, A. Stepansky, D. Levy, D. Esposito, *et al.*, "Tumour evolution inferred by single-cell sequencing," *Nature*, vol. 472, no. 7341, p. 90, 2011. 63, 81

[70] C. G.A. R. Network *et al.*, "Comprehensive molecular characterization of urothelial bladder carcinoma," *Nature*, vol. 507, no. 7492, p. 315, 2014. 1, 2

[71] T. O. Nielsen, F. D. Hsu, K. Jensen, M. Cheang, G. Karaca, Z. Hu, T. Hernandez-Boussard, C. Livasy, D. Cowan, L. Dressler, *et al.*, "Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma," *Clinical cancer research*, vol. 10, no. 16, pp. 5367–5374, 2004. 61

[72] S. Nik-Zainal, H. Davies, J. Staaf, M. Ramakrishna, D. Glodzik, X. Zou, I. Martincorena, L. B. Alexandrov, S. Martin, D. C. Wedge, *et al.*, "Landscape of somatic mutations in 560 breast cancer whole-genome sequences," *Nature*, vol. 534, no. 7605, p. 47, 2016. 70

[73] D. Nishimura, "Biocarta," *Biotech Software & Internet Report: The Computer Software Journal for Scient*, vol. 2, no. 3, pp. 117–120, 2001. 49

[74] M. A. Nowak, N. L. Komarova, A. Sengupta, P. V. Jallepalli, I.-M. Shih, B. Vogelstein, and C. Lengauer, "The role of chromosomal instability in tumor initiation," *Proceedings of the National Academy of Sciences*, vol. 99, no. 25, pp. 16 226–16 231, 2002. 59

[75] R. Nuzzo, "Scientific method: Statistical errors," *Nature News*, vol. 506, no. 7487, p. 150, 2014. 15, 17

[76] T. Y. Reynolds, S. Rockwell, and P. M. Glazer, "Genetic instability induced by the tumor microenvironment," *Cancer research*, vol. 56, no. 24, pp. 5754–5757, 1996. 82

[77] R. M. Ricke, K. B. Jeganathan, and J. M. van Deursen, "Bub1 overexpression induces aneuploidy and tumor formation through aurora b kinase hyperactivation," *The Journal of cell biology*, vol. 193, no. 6, pp. 1049–1064, 2011. 75, 82

[78] I. Rivals, L. Personnaz, L. Taing, and M.-C. Potier, "Enrichment or depletion of a go category within a class of genes: Which test?" *Bioinformatics*, vol. 23, no. 4, pp. 401–407, 2006. 18

[79] A. V. Roschke and E. Rozenblum, "Multi-layered cancer chromosomal instability phenotype," *Frontiers in oncology*, vol. 3, p. 302, 2013. 59

[80] E. Rosivatz, I. Becker, K. Specht, E. Fricke, B. Luber, R. Busch, H. Höfler, and K.-F. Becker, "Differential expression of the epithelial-mesenchymal transition regulators snail, sip1, and twist in gastric cancer," *The American journal of pathology*, vol. 161, no. 5, pp. 1881–1891, 2002. 5

[81] C. Rubio-Perez, D. Tamborero, M. P. Schroeder, A. A. Antolín, J. Deu-Pons, C. Perez-Llamas, J. Mestres, A. Gonzalez-Perez, and N. Lopez-Bigas, "In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities," *Cancer cell*, vol. 27, no. 3, pp. 382–396, 2015. 51

[82] J.-M. Schvartzman, R. Sotillo, and R. Benezra, "Mitotic chromosomal instability and cancer: Mouse modelling of the human disease," *Nature Reviews Cancer*, vol. 10, no. 2, p. 102, 2010. 59

[83] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978. 22

[84] H. Sidhu and N. Capalash, "Uhrf1: The key regulator of epigenetics and molecular target for cancer therapeutics," *Tumor Biology*, vol. 39, no. 2, p. 1 010 428 317 692 205, 2017. 53

[85] G. K. Smyth, "Limma: Linear models for microarray data," in *Bioinformatics and computational biology solutions using R and Bioconductor*, Springer, 2005, pp. 397–420. 14, 61

[86] C. Sotiriou, P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, *et al.*, "Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis," *Journal of the National Cancer Institute*, vol. 98, no. 4, pp. 262–272, 2006. 66–68

[87] B. Stewart, C. P. Wild, *et al.*, "World cancer report 2014," *Health*, 2017. 1

[88] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, *et al.*, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15 545–15 550, 2005. 49, 53

[89] Y. Sun, J. Yao, N. J. Nowak, and S. Goodison, "Cancer progression modeling using static sample data," *Genome biology*, vol. 15, no. 8, p. 440, 2014. 81

[90] Y. Sun, J. Yao, L. Yang, R. Chen, N. J. Nowak, and S. Goodison, "Computational approach for deriving cancer progression roadmaps from static sample data," *Nucleic acids research*, vol. 45, no. 9, e69–e69, 2017. 63, 70, 81

[91] M. M. Suzuki and A. Bird, "Dna methylation landscapes: Provocative insights from epigenomics," *Nature Reviews Genetics*, vol. 9, no. 6, p. 465, 2008. 2, 7

[92] K. Tomizawa, K. Suda, R. Onozato, T. Kosaka, H. Endoh, Y. Sekido, H. Shigematsu, H. Kuwano, Y. Yatabe, and T. Mitsudomi, "Prognostic and predictive implications of her2/erbb2/neu gene mutations in lung cancers," *Lung cancer*, vol. 74, no. 1, pp. 139–144, 2011. 2

[93] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for dna microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001. 13

[94] T. Ushijima, "Detection and interpretation of altered methylation patterns in cancer cells," *Nature Reviews Cancer*, vol. 5, no. 3, p. 223, 2005. 2, 7

[95] K. E. Varley, J. Gertz, K. M. Bowling, S. L. Parker, T. E. Reddy, F. Pauli-Behn, M. K. Cross, B. A. Williams, J. A. Stamatoyannopoulos, G. E. Crawford, *et al.*, "Dynamic dna methylation across diverse human cell lines and tissues," *Genome research*, vol. 23, no. 3, pp. 555–567, 2013. 8, 9

[96] I.-C. Wang, Y.-J. Chen, D. Hughes, V. Petrovic, M. L. Major, H. J. Park, Y. Tan, T. Ackerson, and R. H. Costa, "Forkhead box m1 regulates the transcriptional network of genes essential for mitotic progression and genes encoding the scf (skp2-cks1) ubiquitin ligase," *Molecular and cellular biology*, vol. 25, no. 24, pp. 10 875–10 894, 2005. 75, 82

[97] X. Wang, E. Dalkic, M. Wu, and C. Chan, "Gene module level analysis: Identification to networks and dynamics," *Current opinion in biotechnology*, vol. 19, no. 5, pp. 482–491, 2008. 10

[98] Z. Wang, M. Gerstein, and M. Snyder, "Rna-seq: A revolutionary tool for transcriptomics," *Nature reviews genetics*, vol. 10, no. 1, p. 57, 2009. 12

[99] Z. Wang, E. Curry, and G. Montana, "Network-guided regression for detecting associations between dna methylation and gene expression," *Bioinformatics*, vol. 30, no. 19, pp. 2693–2701, 2014. 9, 10

[100] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, C. G.A. R. Network, *et al.*, "The cancer genome atlas pan-cancer analysis project," *Nature genetics*, vol. 45, no. 10, p. 1113, 2013. 7, 12, 13, 41

[101] J. B. Welsh, L. M. Sapinoso, A. I. Su, S. G. Kern, J. Wang-Rodriguez, C. A. Moskaluk, H. F. Frierson, and G. M. Hampton, "Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer," *Cancer research*, vol. 61, no. 16, pp. 5974–5978, 2001. 1, 5

[102] J. West, S. Beck, X. Wang, and A. E. Teschendorff, "An integrative network algorithm identifies age-associated differential methylation interactome hotspots targeting stem-cell differentiation pathways," *Scientific reports*, vol. 3, p. 1630, 2013. 8, 55

[103] L. D. Wood, D. W. Parsons, S. Jones, J. Lin, T. Sjöblom, R. J. Leary, D. Shen, S. M. Boca, T. Barber, J. Ptak, *et al.*, "The genomic landscapes of human breast and colorectal cancers," *Science*, vol. 318, no. 5853, pp. 1108–1113, 2007. 75

[104] J. Wu, T. Vallenius, K. Ovaska, J. Westermarck, T. P. Mäkelä, and S. Hautaniemi, "Integrated network analysis platform for protein-protein interactions," *Nature methods*, vol. 6, no. 1, p. 75, 2009. 25, 31, 41

[105] T. I. Zack, S. E. Schumacher, S. L. Carter, A. D. Cherniack, G. Saksena, B. Tabak, M. S. Lawrence, C.-Z. Zhang, J. Wala, C. H. Mermel, *et al.*, "Pan-cancer patterns of somatic copy number alteration," *Nature genetics*, vol. 45, no. 10, p. 1134, 2013. 2

[106] A. Zellner, "Bayesian estimation and prediction using asymmetric loss functions," *Journal of the American Statistical Association*, vol. 81, no. 394, pp. 446–451, 1986. 10, 21

[107] ——, "On assessing prior distributions and bayesian regression analysis with g-prior distributions," *Bayesian inference and decision techniques*, 1986. 30

[108] S. Zhang, C.-C. Liu, W. Li, H. Shen, P. W. Laird, and X. J. Zhou, "Discovery of multi-dimensional modules by integrative analysis of cancer genomic data," *Nucleic acids research*, vol. 40, no. 19, pp. 9379–9391, 2012. 27

Appendix A

Significance level of separability scores

Appendix A.1: Significance level of separability scores of predictor modules.
Appendix A.2: Significance level of separability scores of response modules.

A.1

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	0.00E+00																				
2	5.98E-23	0.00E+00																			
3	1.37E-79	8.12E-166	0.00E+00																		
4	5.10E-51	5.05E-97	2.77E-07	0.00E+00																	
5	1.06E-09	5.07E-07	7.50E-06	2.37E-04	0.00E+00																
6	6.10E-50	1.03E-129	1.04E-09	2.96E-03	1.91E-08	0.00E+00															
7	6.35E-09	3.28E-07	7.62E-15	6.48E-11	9.39E-08	2.67E-14	0.00E+00														
8	2.97E-57	1.06E-132	9.19E-08	2.93E-03	2.63E-07	3.14E-06	6.83E-11	0.00E+00													
9	3.05E-66	6.79E-119	3.34E-06	1.44E-02	4.86E-08	9.12E-03	1.86E-18	1.73E-05	0.00E+00												
10	1.80E-04	3.10E-05	5.36E-19	1.40E-17	8.15E-04	7.27E-17	8.27E-04	6.76E-14	6.72E-18	0.00E+00											
11	1.66E-19	3.93E-80	3.80E-19	1.20E-31	2.33E-11	1.26E-24	6.21E-19	6.55E-31	3.61E-21	3.08E-10	0.00E+00										
12	7.34E-24	2.57E-55	2.33E-07	2.64E-02	6.98E-10	4.49E-03	7.18E-08	9.56E-05	4.57E-03	2.61E-11	2.03E-11	0.00E+00									
13	8.35E-05	2.60E-02	7.97E-20	2.21E-13	7.26E-07	1.37E-14	3.14E-02	1.25E-14	3.07E-15	2.98E-02	4.05E-11	6.59E-08	0.00E+00								
14	8.54E-11	1.28E-22	1.96E-08	1.23E-04	1.63E-02	1.93E-06	5.24E-07	7.16E-10	1.04E-04	6.00E-07	4.31E-03	1.27E-04	1.99E-04	0.00E+00							
15	1.61E-36	2.48E-21	1.75E-98	4.20E-73	3.70E-08	9.38E-114	4.42E-03	1.30E-80	9.76E-67	1.05E-07	1.32E-37	5.50E-39	8.08E-13	1.21E-27	0.00E+00						
16	1.62E-23	6.94E-20	4.60E-74	9.39E-48	2.70E-03	4.44E-72	4.81E-10	1.58E-75	3.75E-56	5.97E-08	3.97E-21	4.12E-20	4.42E-05	1.06E-07	1.29E-34	0.00E+00					
17	4.32E-30	2.09E-04	5.90E-149	1.06E-107	1.42E-10	1.42E-148	1.41E-05	6.54E-107	1.89E-107	4.79E-06	3.02E-66	3.94E-53	5.74E-03	3.49E-39	4.07E-10	1.55E-29	0.00E+00				
18	4.35E-42	1.26E-120	4.94E-10	1.21E-04	6.60E-05	5.52E-07	2.36E-06	2.41E-02	1.41E-10	4.35E-12	2.77E-39	3.39E-08	3.26E-17	3.68E-10	1.44E-79	1.48E-65	2.24E-82	0.00E+00			
19	1.72E-28	2.83E-02	7.48E-155	1.11E-72	1.46E-05	4.88E-104	2.26E-17	1.06E-93	5.97E-92	2.70E-06	6.08E-59	1.26E-43	4.50E-02	6.41E-36	2.62E-09	9.13E-49	7.78E-03	1.34E-77	0.00E+00		
20	1.63E-31	4.54E-47	7.75E-06	8.18E-05	2.71E-03	4.56E-09	4.38E-03	4.26E-05	4.45E-08	1.27E-08	5.88E-23	3.31E-05	4.24E-10	9.00E-06	1.28E-32	1.70E-23	7.59E-47	4.56E-04	2.67E-32	0.00E+00	
21	1.05E-47	4.82E-168	6.47E-19	5.56E-03	1.15E-07	2.66E-03	3.29E-12	2.55E-03	4.84E-04	2.03E-13	1.47E-31	1.75E-03	2.49E-14	5.49E-11	9.01E-151	2.43E-108	1.14E-190	5.92E-05	4.01E-165	4.89E-10	0.00E+00

A.2

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	0																			
1	0																			
2	4.93E-92	0.00E+00																		
3	2.31E-07	4.44E-13	0.00E+00																	
4	1.45E-11	1.03E-15	2.78E-03	0.00E+00																
5	1.40E-153	4.27E-147	3.65E-38	1.23E-34	0.00E+00															
6	3.82E-22	8.50E-21	5.16E-03	2.86E-09	8.00E-133	0.00E+00														
7	4.06E-51	8.56E-29	3.90E-09	1.64E-02	4.79E-107	2.73E-34	0.00E+00													
8	1.94E-62	3.88E-60	2.02E-35	2.03E-30	9.53E-10	7.25E-50	1.96E-54	0.00E+00												
9	3.09E-04	8.05E-20	2.15E-03	5.57E-09	2.92E-65	5.65E-10	1.85E-25	7.29E-40	0.00E+00											
10	6.28E-107	3.16E-12	2.20E-27	8.68E-20	1.92E-112	5.61E-77	1.07E-41	1.91E-62	2.91E-57	0.00E+00										
11	6.84E-113	7.46E-65	5.23E-13	2.87E-17	3.44E-88	9.21E-51	4.91E-45	1.28E-49	8.58E-49	7.43E-64	0.00E+00									
12	5.71E-48	1.50E-56	3.65E-06	4.87E-02	6.57E-86	1.33E-34	4.41E-11	1.57E-55	7.28E-26	1.43E-73	1.99E-26	0.00E+00								
13	5.47E-73	1.25E-135	3.28E-21	8.27E-05	1.09E-178	8.88E-87	1.13E-11	1.38E-60	1.24E-46	1.46E-108	1.66E-86	4.20E-22	0.00E+00							
14	4.27E-107	4.30E-93	7.35E-40	3.16E-29	6.87E-09	4.92E-111	2.45E-69	1.30E-08	3.82E-35	1.61E-80	6.09E-75	6.65E-92	4.98E-98	0.00E+00						
15	2.68E-40	1.68E-138	3.92E-13	3.44E-30	9.00E-114	4.39E-17	1.87E-104	1.62E-38	5.54E-38	1.84E-125	1.39E-80	4.52E-78	1.99E-137	1.16E-110	0.00E+00					
16	1.21E-49	2.43E-23	2.16E-08	1.13E-03	6.82E-82	1.56E-25	4.22E-02	1.15E-44	7.38E-24	6.52E-43	1.52E-20	4.46E-11	2.92E-27	2.73E-79	4.77E-81	0.00E+00				
17	1.53E-15	1.71E-28	9.82E-06	8.82E-04	2.51E-84	4.36E-13	1.15E-08	4.92E-44	4.45E-12	5.42E-57	6.50E-59	4.56E-16	3.73E-15	9.36E-56	3.09E-69	4.49E-11	0.00E+00			
18	4.85E-15	2.17E-14	8.20E-11	7.12E-10	7.35E-14	6.88E-13	1.59E-14	1.79E-10	1.35E-13	4.11E-14	1.00E-10	4.38E-15	1.18E-13	3.61E-11	2.45E-10	2.21E-12	5.02E-13	0.00E+00		
19	2.53E-20	2.93E-55	6.53E-03	8.66E-08	5.48E-134	2.10E-16	8.09E-35	4.61E-48	2.77E-06	1.66E-77	8.03E-43	2.15E-15	1.23E-73	1.26E-65	6.56E-34	2.21E-26	2.05E-17	7.23E-11	0.00E+00	
20	1.45E-91	3.18E-133	2.71E-11	2.94E-03	1.62E-212	2.91E-40	7.22E-31	9.00E-70	1.98E-47	1.93E-142	1.68E-33	4.61E-26	1.79E-65	3.07E-139	8.32E-137	2.34E-15	1.31E-19	4.30E-15	3.11E-58	0.00E+00
21	4.44E-63	5.37E-62	1.42E-27	2.36E-21	5.27E-58	5.72E-52	4.02E-63	1.46E-37	2.35E-42	5.48E-62	1.59E-53	5.48E-54	2.38E-83	8.53E-60	1.46E-64	2.81E-44	3.91E-60	2.70E-12	1.42E-53	1.07E-105
22	1.47E-06	3.17E-111	2.23E-17	3.26E-14	3.74E-131	6.80E-49	1.48E-66	6.56E-64	6.89E-11	7.15E-125	3.20E-102	4.80E-52	1.53E-63	1.46E-96	2.72E-74	1.10E-53	2.99E-22	2.00E-15	5.41E-28	2.92E-130
23	1.92E-78	1.12E-91	7.39E-35	1.12E-23	7.67E-54	1.02E-70	1.19E-71	1.18E-33	2.67E-58	1.15E-58	2.27E-67	7.92E-74	1.32E-91	7.45E-45	2.48E-33	6.96E-51	6.33E-84	1.16E-11	1.97E-56	1.11E-128
24	2.63E-75	2.04E-07	2.18E-09	4.08E-07	3.59E-75	3.58E-18	4.54E-11	9.34E-51	2.24E-36	2.26E-20	4.00E-11	4.34E-17	1.68E-44	6.46E-90	4.51E-85	2.33E-07	2.94E-22	6.42E-14	1.37E-28	2.66E-34
25	8.68E-20	8.04E-12	6.16E-07	6.05E-05	1.47E-38	7.97E-14	3.91E-04	3.08E-17	3.70E-09	2.85E-24	1.62E-16	1.09E-06	1.61E-07	4.88E-17	4.84E-44	5.15E-05	4.33E-07	5.43E-11	8.70E-15	4.83E-07
26	3.31E-43	8.40E-116	6.07E-06	5.80E-03	2.28E-162	3.91E-52	5.58E-15	1.60E-56	8.63E-18	3.19E-100	2.38E-39	5.39E-06	2.19E-21	1.22E-120	2.98E-115	2.43E-10	6.60E-23	5.63E-15	1.00E-22	2.89E-45
27	2.02E-83	4.23E-51	2.43E-13	3.45E-04	1.43E-159	1.58E-35	9.33E-09	5.56E-72	1.13E-41	1.40E-65	3.62E-31	4.27E-16	6.99E-42	1.50E-103	6.46E-144	2.28E-07	5.87E-18	2.29E-14	7.83E-46	1.36E-13
28	3.10E-19	6.79E-58	5.47E-07	3.67E-03	2.18E-97	9.31E-32	1.70E-07	7.26E-50	2.15E-07	1.47E-64	9.93E-43	1.14E-04	5.63E-12	7.76E-79	9.25E-90	5.08E-07	7.82E-11	6.29E-15	1.20E-11	1.71E-26
29	6.50E-220	5.07E-192	3.80E-38	5.16E-10	2.95E-168	1.21E-125	4.08E-25	2.77E-74	1.87E-98	1.72E-137	2.44E-46	2.65E-38	1.06E-64	1.40E-135	3.87E-141	4.42E-15	1.10E-67	3.28E-13	6.67E-127	2.11E-78
30	3.59E-203	1.23E-143	4.66E-33	3.28E-10	3.95E-147	1.24E-112	1.04E-15	2.21E-60	5.45E-87	3.49E-91	6.99E-63	1.86E-40	3.22E-54	3.95E-98	3.04E-162	1.84E-11	2.30E-65	8.12E-12	1.24E-98	5.26E-79
31	1.57E-17	3.72E-79	7.29E-07	9.49E-12	4.83E-171	1.04E-15	2.38E-54	3.86E-54	1.36E-15	2.39E-137	7.26E-114	1.28E-42	1.48E-101	9.69E-114	1.19E-31	1.84E-50	3.80E-14	1.77E-12	5.90E-18	4.25E-143
32	2.30E-42	1.16E-110	9.24E-37	1.67E-28	6.73E-86	4.31E-50	1.38E-91	4.08E-42	6.24E-31	2.55E-85	8.68E-76	1.04E-93	4.25E-94	6.19E-85	2.18E-25	7.88E-71	1.64E-55	6.24E-11	1.67E-55	1.60E-147
33	5.00E-23	6.48E-17	2.76E-03	8.61E-06	1.61E-73	2.28E-08	2.14E-23	1.55E-36	2.43E-12	3.37E-31	4.60E-15	3.52E-18	3.61E-67	2.65E-56	3.67E-23	6.36E-20	9.16E-18	3.90E-12	1.33E-08	5.63E-32
34	4.04E-17	1.75E-129	2.71E-09	4.28E-13	3.20E-130	3.28E-28	1.15E-79	1.21E-51	4.43E-19	8.58E-123	6.61E-86	1.60E-30	5.57E-87	2.17E-97	1.32E-20	2.76E-67	5.59E-30	2.29E-12	5.75E-18	4.54E-99
35	3.50E-25	1.36E-75	7.83E-13	1.24E-17	3.41E-118	6.17E-28	1.10E-52	4.51E-56	1.28E-06	1.06E-79	2.39E-70	9.83E-65	1.34E-95	4.07E-92	1.29E-58	1.79E-33	6.85E-29	1.16E-10	3.53E-11	9.10E-142
36	9.77E-151	1.31E-53	1.64E-23	1.90E-08	2.83E-130	7.73E-77	5.96E-16	2.63E-53	9.06E-55	7.80E-45	5.17E-22	4.86E-20	4.98E-71	2.55E-106	1.86E-143	1.22E-07	1.40E-54	5.34E-12	2.33E-70	2.41E-69
37	2.00E-06	9.39E-62	2.63E-11	9.16E-14	1.85E-50	2.75E-18	3.50E-40	6.16E-27	4.74E-11	8.86E-48	4.01E-40	9.76E-37	1.04E-57	1.71E-43	1.68E-04	3.12E-34	6.08E-34	1.38E-09	3.64E-13	3.66E-48
38	1.72E-52	3.15E-15	1.07E-04	4.48E-04	3.79E-117	7.49E-13	3.38E-06	4.74E-60	3.45E-21	2.05E-50	1.14E-34	1.53E-08	2.51E-36	3.29E-105	1.08E-90	1.03E-08	3.30E-08	6.22E-14	5.89E-28	1.01E-13
39	2.12E-79	1.05E-72	2.58E-26	3.87E-21	8.19E-61	8.56E-53	2.02E-51	7.80E-29	2.09E-51	1.82E-50	3.66E-13	3.10E-45	1.45E-73	5.40E-49	2.82E-49	5.32E-35	4.85E-50	8.35E-11	1.39E-57	1.29E-57

21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39
0.00E+00																		
1.27E-96	0.00E+00																	
1.66E-40	8.47E-84	0.00E+00																
4.99E-47	3.27E-91	3.81E-60	0.00E+00															
1.92E-30	1.05E-20	2.35E-35	6.78E-07	0.00E+00														
1.11E-63	1.35E-43	7.96E-75	2.11E-43	1.65E-07	0.00E+00													
5.18E-72	4.30E-108	1.54E-115	8.55E-05	2.52E-05	1.43E-38	0.00E+00												
1.41E-52	5.52E-17	1.66E-65	2.99E-25	1.06E-05	4.17E-05	3.48E-16	0.00E+00											
3.41E-76	1.27E-206	1.17E-76	1.69E-45	7.51E-15	5.68E-48	6.82E-44	2.43E-36	0.00E+00										
1.60E-82	4.85E-189	1.35E-67	1.15E-24	6.60E-06	3.33E-52	9.77E-29	3.01E-38	6.63E-12	0.00E+00									
5.63E-83	5.56E-17	6.02E-73	4.11E-63	1.21E-28	1.22E-52	9.20E-125	6.85E-32	2.26E-258	1.55E-266	0.00E+00								
1.64E-56	1.30E-36	7.47E-17	1.39E-60	1.44E-34	2.27E-94	1.88E-108	6.41E-66	2.07E-123	5.78E-148	1.43E-35	0.00E+00							
1.06E-41	4.21E-50	5.46E-43	5.81E-13	4.39E-19	4.20E-35	1.71E-18	5.84E-28	1.92E-63	1.68E-60	2.81E-19	4.41E-42	0.00E+00						
4.95E-69	4.21E-20	1.28E-62	4.10E-88	1.68E-26	2.37E-42	6.46E-104	3.81E-29	5.45E-155	1.53E-177	8.48E-17	5.58E-42	3.54E-48	0.00E+00					
3.51E-60	6.91E-22	4.67E-58	9.09E-76	1.42E-26	1.96E-55	2.35E-83	5.47E-26	3.88E-155	9.89E-171	1.92E-35	1.06E-35	5.31E-23	2.24E-65	0.00E+00				
4.39E-69	7.56E-125	4.92E-87	5.38E-07	9.50E-08	1.40E-47	3.50E-16	5.70E-23	1.42E-19	1.96E-13	2.96E-136	8.48E-104	3.58E-28	7.78E-114	9.38E-94	0.00E+00			
2.09E-32	6.77E-13	3.78E-18	1.19E-29	2.70E-26	2.46E-27	1.77E-50	4.94E-27	6.16E-50	1.45E-52	1.71E-09	2.59E-06	8.17E-14	1.37E-08	2.79E-13	1.95E-45	0.00E+00		
5.83E-66	2.44E-70	2.06E-81	2.49E-05	2.10E-06	7.40E-15	3.91E-09	1.21E-13	7.23E-54	3.58E-37	1.75E-58	3.76E-110	3.04E-15	8.77E-55	9.28E-63	2.99E-23	1.06E-50	0.00E+00	
2.05E-35	6.75E-78	2.76E-21	5.23E-24	2.98E-23	1.47E-76	5.25E-45	6.02E-47	4.87E-19	6.83E-22	6.42E-73	3.38E-35	3.83E-44	1.62E-53	7.88E-59	2.85E-18	2.06E-34	2.48E-50	0.00E+00

Appendix B

Analyses of cluster 1 genes and its interactions

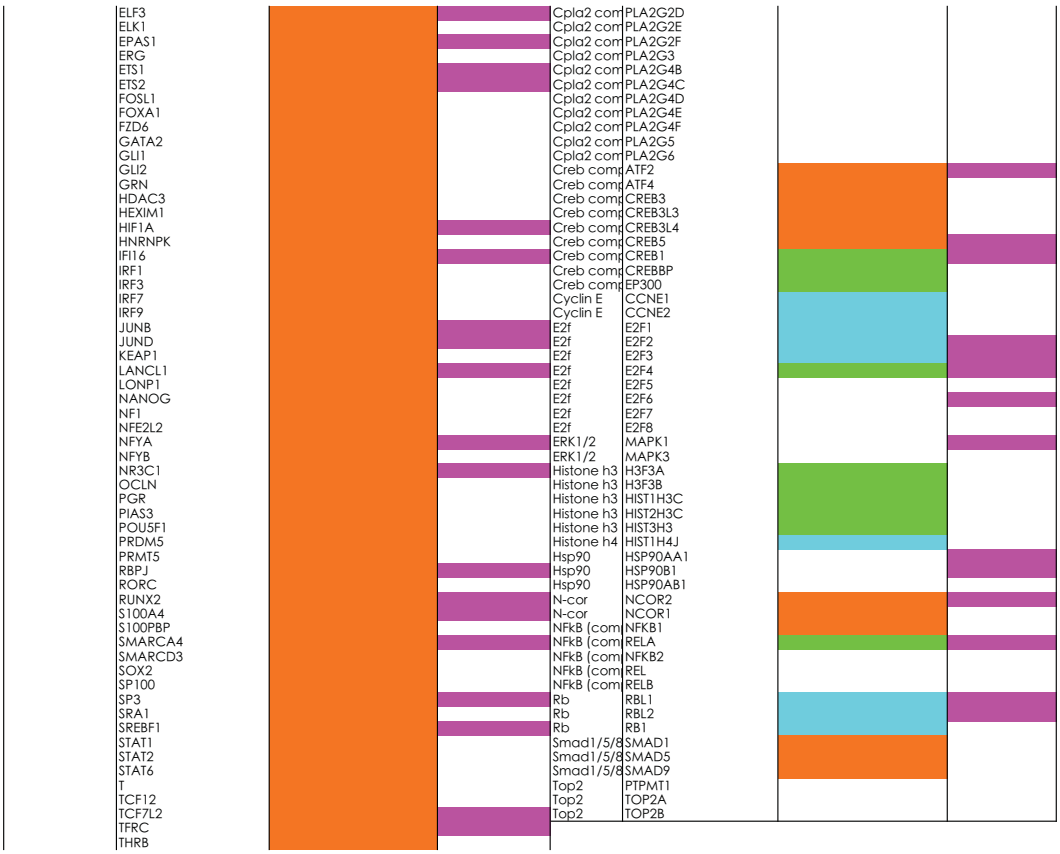
Appendix B.1: Literature evidence for the role of cluster 1 genes in TME and CIN.
Appendix B.2: The upstream regulator of cluster 1 genes, along with their association with CIN or TME and hypoxia responsiveness.
Appendix B.3: The causal network genes of cluster 1 genes, along with their association with CIN or TME and hypoxia responsiveness.

B.1

Section	Cluster 1 Gene	Process	Hypoxia Response?	PubmedID
TME	HMMR	Spindle Formation, Migration	Yes	12808028, 8634437
TME	CA9	Metabolism, Hypoxia	Yes	24146383
TME	CEMIP	Migration, Hypoxia	Yes	28819426
TME	CTHRC1	Extracellular Matrix, Angiogenesis, Hypoxia	Yes	16778098, 27686285, 25263696
TME	CXCL11	Immune Response	Yes	24940911
TME	MMP1	Extracellular Matrix, Hypoxia	Yes	26343184
TME	TNFRSF9	Lymphocyte Activation	Yes	8262389
TME	COL11A1	Extracellular Matrix	Yes	10486316
TME	DEPDC1B	Adhesion		25458010
TME	S100P	DNA Damage Response		26967060, 25614008, 19187537
TME	ADAMDEC1	Immune Response		12037602
TME	ADAMTS14	Extracellular Matrix		25636539
TME	CARMIL2	Migration		19846667, 26466680
TME	CBS	Angiogenesis, Metabolism, Hypoxia		24730679, 24328859
TME	CCL11	Migration, Inflammation		19351767
TME	CCR8	Migration, Inflammation		23363815
TME	CLEC5A	Immune Response		12570670, 10449773
TME	CLEC6A	Immune Response		16423983
TME	COL10A1	Extracellular Matrix		22894674
TME	CXCL10	Immune Response, Hypoxia		28592115, 24799675
TME	HAPLN1	Extracellular Matrix		22159717
TME	IL4I1	Metabolism, Immune Response		28405502, 19436310
TME	KIF26B	Adhesion, Migration		28581513, 25652119
TME	LHX2	Transcription		24423492
TME	MMP11	Extracellular Matrix		26892540
TME	MMP13	Extracellular Matrix		22992737, 20371345
TME	TDO2	Metabolism, Immune Response		25628622
TME	PITX1	Transcription, Hypoxia		25558831
TME	TFR2	Angiogenesis, Hypoxia		20360937
TME	COL9A3	Extracellular Matrix		8660302
TME	COMP	Extracellular Matrix		16051604
TME	CST1	Protease Inhibitors		UniProt
TME	CST4	Protease Inhibitors		3488317
TME	EPYC	Extracellular Matrix		11292372
TME	LRRC15	Extracellular Matrix		16098915
TME	MFAP2	Extracellular Matrix		26167404
TME-R	GJB2	Cell-cell Communication		22514560
TME-R	IBSP	Extracellular Matrix		24103036
TME-R	SCT	Endocrine Signalling		11060443
TME-R	VGF	Endocrine Signalling		19194657
MISC	B3GNT4	Glycolsylation	Yes	
MISC	ACTL8	NA		
MISC	CHST6	Glycosaminoglycan metabolism		
MISC	CRYBG2	NA		
MISC	ENTHD1	NA		
MISC	GBX2	Transcription		
MISC	HSD17B6	Metabolism		
MISC	INAVA	NA		
MISC	MKRN3	Ubiquitinaton		
MISC	OR2B6	Olfactory Receptor		
MISC	PLPP4	Metabolism		
MISC	SIX2	Transcription		
MISC	ZIC1	Transcription		
CIN-R	HIST1H2AM	Chromatin	Yes	Uniprot
CIN-R	HIST2H2AA3	Chromatin	Yes	Uniprot
CIN-R	HIST1H2BO	Chromatin		Uniprot
CIN-R	KREMEN2	Signalling		12050670
CIN-R	MYBL2	Transcription		9840932, 23842645
CIN-R	STRA8	Meiosis		18799751
CIN	ASF1B	Chromatin	Yes	11897662
CIN	CDC20	Spindle Assembly Checkpoint	Yes	19528295, 15525512
CIN	ANLN	Cytokinesis		16040610, 20732437
CIN	BUB1	Spindle Assembly Checkpoint		21646403, 19487456, 16760428
CIN	CDC45	DNA Replication		23643534
CIN	CDCA2	Chromatin		21820363, 16998479
CIN	CDKN3	Cell Cycle		23775190
CIN	CEP55	Cytokinesis		16198290, 16406728, 25915844
CIN	CKAP2L	Cell Cycle		24260314, 17376772
CIN	CLSPN	DNA Replication		12766152, 11090622
CIN	DLGAP5	Spindle Formation		17118403, 15987997
CIN	FOXM1	Transcription, Hypoxia		23768511, 19097132
CIN	KIF4A	Cytokinesis, Spindle Organization		15625105, 15326200, 15297875
CIN	KNL1	Kinetochore		24344188
CIN	MELK	Mitotic Progression		15908796, 24844244
CIN	PBK	DNA Damage Response		17482142
CIN	PKMYT1	Mitotic Progression		12738781, 10504341, 7569953
CIN	PLK1	Spindle Assembly Checkpoint		18615013, 17376779, 16837776
CIN	RRM2	DNA Damage Repair		20159953
CIN	SHCBP1	Cytokinesis		27129942, 25486361
CIN	SPC24	Kinetochore		14699129, 15961401, 27713128
CIN	UHRF1	DNA Methylation		24486181
CIN	TUBB3	Cytoskeleton, Hypoxia		23321512, 18178340
CIN	KIF2C	Chromosome Segregation		23098759, 19060894, 14960279

B.2


Complex
Upstream regulator gene
Interacts with CIN/TME/Both
Hypoxia Response
AJUBA
ATF3
BARX2
BRCA1
BUB1B
BUB3
CASP7
CCNB1
CDC27
CDC7
CDK1
CDK2
CELF1
CKS1B
DRAP1
DVL2
FBXO5
FBXW5
FZR1
HDAC1
HERC2
LATS1
LIN28A
MAD1L1
MAD2L1
MAD2L1BP
MAP2K1
MBD2
MIIP
MIS12
MSH6
MXI1
NCAPG
NDC80
NEDD1
NINL
NOTCH3
NPM1
NUF2
NUPR1
ORC2
OTUB1
PAK1
PARP1
PMS2
POLR2A
PPP1CC
PRC1
PSMA1
PSMA2
PSMA3
PSMA4
PSMA5
PSMA6
PSMA7
PSMB1
PSMB3
PSMB4
PSMC2
PTEN
RAE1
RARA
RASSF1
REST
RRM1
RRM2B
RUNX1
SATB1
SEPT9
SKP2
SMAD2
SMAD4
SMAD7
SMARCA2
SMC2
SOX11
SPAG9
SRSF1
STEAP3
STK10
SUMO1
SUZ12
TAL1
TERF1
TERT
TOPBP1
TTK
UBE2S
UHRF1BP1
USP28
USP9X
WDHD1
WEE1
YAP1
YY1
ZBED6
ZNF24
ZNF622
ZWINT
ARNT
BACH1
BCL6
BHLHE41
BMPR1A
CDKN2A
CEBPB
CITED2
CLTC
CTNNB1
CXCR3
DNAJB6
Complex
Upstream regulator gene
Interacts with CIN/TME/Both
Hypoxia Response
TNFSF9
TPX2
TWIST1
TYROBP
USF1
WNT5A
YWHAZ
ZNF217
CBX5
CCND1
ESR1
ESR2
EZH2
FOXO3
MYC
SMAD3
STAT3
STAT5A
TP53
TP63
WT1
ADAMTS7
AKAP12
ANAPC2
ANAPC7
APC
AR
ARX
ATF6
ATRX
AURKA
AURKB
BARD1
BCL2
BCL2L1
BCL2L14
BIRC5
BRCA2
CCL1
CCNA1
CCNA2
CD44
CDC14A
CDC25B
CDC25C
CEP170
CHEK1
CHUK
CUEDC2
CUL1
CYLD
DAXX
DSN1
DUSP1
EGFR
EIF4B
EPOR
ERCC6L
ETV1
FBXL14
FOXO1
HAT1
HAUS8
HDAC2
HFE
HNF4A
HSPA8
IPO4
IRF4
ITGA2
ITGB1
KAT5
KDM5B
KIF23
KIFC1
KLF4
KMT2E
LIN37
LIN9
LRRK2
MAPK10
MAPK8
MAVS
MCM7
MDM2
MITF
MMP25
NASP
NEUROG1
NFKBIA
ODF2
PIN1
PLEKHG6
PRKDC
RAF1
RBBP4
SIN3B
SP1
SPC25
TIAL1
XRCC5
XRCC6
Ap1 comp
Ap1 comp
Cpla2 com
Cpla2 com
Cpla2 com
Cpla2 com
Cpla2 com
Cpla2 com
Cpla2 com
FOS
JUN
PLA2G4A
JMJD7-PLA2G4B
PLA2G10
PLA2G12A
PLA2G12B
PLA2G1B
PLA2G2A


ELF3
ELK1
EPAS1
ERG
ETS1
ETS2
FOSL1
FOXA1
FZD6
GATA2
GLI1
GLI2
GRN
HDAC3
HEXIM1
HIF1A
HNRNPK
IFI16
IRF1
IRF3
IRF7
IRF9
JUNB
JUND
KEAP1
LANCL1
LONP1
NANOG
NF1
NFE2L2
NFYA
NFYB
NR3C1
OCLN
PGR
PIAS3
POU5F1
PRDM5
PRMT5
RBPJ
RORC
RUNX2
S100A4
S100PBP
SMARCA4
SMARCD3
SOX2
SP100
SP3
SRA1
SREBF1
STAT1
STAT2
STAT6
T
TCF12
TCF7L2
TFRC
THRB
Cpla2 com
Cpla2 com
Cpla2 com
Cpla2 com
Cpla2 com
Cpla2 com
Cpla2 com
Cpla2 com
Cpla2 com
Cpla2 com
Cpla2 com
Creb comp
Creb comp
Creb comp
Creb comp
Creb comp
Creb comp
Creb comp
Creb comp
Cyclin E
Cyclin E
E2f
E2f
E2f
E2f
E2f
E2f
E2f
ERK1/2
ERK1/2
Histone h3
Histone h3
Histone h3
Histone h3
Histone h3
Histone h4
Hsp90
Hsp90
Hsp90
N-cor
N-cor
NFkB (com
NFkB (com
NFkB (com
NFkB (com
Rb
Rb
Rb
Smad1/5/8
Smad1/5/8
Smad1/5/8
Top2
Top2
Top2
PLA2G2D
PLA2G2E
PLA2G2F
PLA2G3
PLA2G4B
PLA2G4C
PLA2G4D
PLA2G4E
PLA2G4F
PLA2G5
PLA2G6
ATF2
ATF4
CREB3
CREB3L3
CREB3L4
CREB5
CREB1
CREBBP
EP300
CCNE1
CCNE2
E2F1
E2F2
E2F3
E2F4
E2F5
E2F6
E2F7
E2F8
MAPK1
MAPK3
H3F3A
H3F3B
HIST1H3C
HIST2H3C
HIST3H3
HIST1H4J
HSP90AA1
HSP90B1
HSP90AB1
NCOR2
NCOR1
NFKB1
RELA
NFKB2
REL
RELB
RBL1
RBL2
RB1
SMAD1
SMAD5
SMAD9
PTPMT1
TOP2A
TOP2B

B.3

Complex	Upstream causal gene	Entrez Gene ID	Interacts with CIN/TME/Both	Hypoxia Response
	CD24	100133941		
	CDKN1A	1026		
	DDR2	4921		
	ETV1	2115		
	IRF4	3662		
	KDM5B	10765		
	LILRB4	11006		
	NAMPT	10135		
	PTGS2	5743		
	RAC1	5879		
	RUNX2	860		
	SUMO3	6612		
	ACVR1	90		
	ARX	170302		
	DPAGT1	1798		
	FAS	355		
	GBP1	2633		
	GPX4	2879		
	HFE	3077		
	HIF1AN	55662		
	HNF4A	3172		
	IFNL4	101180976		
	IL11	3589		
	IL18	3606		
	LAMA3	3909		
	RARRES3	5920		
	S100A9	6280		
	SAFB	6294		
	SDR9C7	121214		
	TBK1	29110		
	TIAL1	7073		
	TICAM1	148022		
	ATF6	22926		
	ERBB2	2064		
	MITF	4286		
	SP1	6667		
	TRAF2	7186		
	BSG	682		
	FOXO1	2308		
	HDAC2	3066		
	RABL6	55684		
	RAF1	5894		
	E2F6	1876		
	MDM2	4193		
	OGT	8473		
	SEL1L	6400		
	AR	367		
	AREG	374		
	BTRC	8945		
	DVL2	1856		
	FOXM1			
	NELFE	7936		
	NUPR1	26471		
	PPP2CB	5516		
	REST	5978		
	SVIL	6840		
E2f complex	E2F3	1871		
E2f complex	E2F4	1874		
E2f complex	E2F1	1869		
E2f complex	E2F2	1870		
E2f complex	E2F5	1875		
E2f complex	E2F7	144455		
E2f complex	E2F8	79733		
IL1 complex	IL1A	3552		
IL1 complex	IL1RN	3557		
IL1 complex	IL33	90865		
IL1 complex	IL36G	56300		
IL1 complex	IL1B	3553		
IL1 complex	IL1F10	84639		
IL1 complex	IL36A	27179		
IL1 complex	IL36B	27177		
IL1 complex	IL36RN	26525		
IL1 complex	IL37	27178		
Jnk complex	MAP2K4	6416		
Jnk complex	MAPK10	5602		
Jnk complex	MAPK12	6300		
Jnk complex	MAPK8	5599		
Jnk complex	MAPK9	5601		
NFkB complex	NFKB1	4790		
NFkB complex	NFKB2	4791		
NFkB complex	REL	5966		
NFkB complex	RELA	5970		
NFkB complex	RELB	5971		
PLC complex	PLCB4	5332		
PLC complex	PLCE1	51196		
PLC complex	PLCL1	5334		
PLC complex	PLCL2	23228		
PLC complex	NOTUM	147111		
PLC complex	PDIA3	2923		
PLC complex	PLCB1	23236		
PLC complex	PLCB2	5330		
PLC complex	PLCB3	5331		
PLC complex	PLCD1	5333		
PLC complex	PLCD3	113026		
PLC complex	PLCD4	84812		
PLC complex	PLCG1	5335		
PLC complex	PLCG2	5336		
PLC complex	PLCH1	23007		
PLC complex	PLCH2	9651		
PLC complex	PLCZ1	89869		
Ras complex	KRAS	3845		
Ras complex	RRAS2	22800		
Ras complex	HRAS	3265		
Ras complex	MRAS	22808		
Ras complex	NRAS	4893		
Ras complex	RRAS	6237		
SAA complex	SAA1	6288		
SAA complex	SAA2	6289		
SAA complex	SAA3	6290		
SAA complex	SAA4	6291		