IDENTIFICATION OF GENE SETS THAT PREDICT ACUTE MYELOID LEUKEMIA PROGNOSIS USING INTEGRATIVE GENE NETWORK ANALYSIS

by

Hanie Samimi B.S.

A thesis submitted to the Graduate Council of
Texas State University in partial fulfillment
of the requirements for the degree of
Master of Science
with a Major in Computer Science
August 2018

Committee Members:

Vangelis Metsis, Chair

Habil Zare

L. Kevin Lewis

FAIR USE AND AUTHOR'S PERMISSION STATEMENT

Fair Use

Duplication Permission

ACKNOWLEDGEMENTS

This study was in collaboration with the Dr. Aly Karsan lab at the British Colombia Cancer Agency. It was supported by an internal grant from Texas State University. I acknowledge Dr. Tobias Herold for providing clinical data of the Herold *et al.* study.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF ABBREVIATIONS

Abbreviation	Description
AFT	accelerate failure time
AML	Acute Myeloid Leukemia
ELN	European Leukemia Net
FEM	Functional Epigenetic Modules
GEO	Gene Expression Omnibus
HCC	Hepatocellular Carcinoma
HDAC	Histone deacetylases
LASSO	Least Absolute Shrinkage and Selection Operator
Pathway ORA	Pathway Over-Representation Analysis
PCA	Principal Component Analysis
POI	Phenotype of Interest
PPI	Protein-protein Interaction Network
PTM	Post–Translational protein Modification
RRBS	Reduced Representation Bisulfite Sequencing
TCGA	The Cancer Genome Atlas
TSS	Transcription Start Sites
WGBS	Whole–Genome Bisulfite Sequencing

ABSTRACT

Orthogonal data types can potentially provide new opportunities to pinpoint the underlying molecular mechanisms of diseases. However, currently-available techniques to capitalize on information from different data types suffer from a substantial loss of statistical power. Therefore, there is urgent need to develop algorithms to integrate data types. In this thesis, I have developed a data integration approach based on multi-view clustering. I demonstrate the usefulness of my approach in prognostication of Acute Myeloid Leukemia (AML), a particular type of blood cancer. AML accounts for 1.2% of cancer deaths per year in the USA. AML patients are categorized into low, medium and high-risk groups. The variable survival rate for medium-risk patients leads to difficulties in deciding on the appropriate treatment for these patients. Current methods of prognostication of AML use only gene expression, mutations and molecular cytogenetic abnormalities. However, the DNA methylation data, which have valuable information that would be useful for prognostication, have not yet been effectively used in the existing clinical tests. In this project, I have used The Cancer Genome Atlas (TCGA) dataset and developed a method that analyzes both gene expression and DNA methylation data in a single model using network analysis. The model based on this methodology correctly classified 13 out of 90 patients as high-risk, whereas they were previously labeled as medium-risk using current clinical methods. All 13 of these cases died within two years after diagnosis. To validate these results, I tested the method using an independent dataset. The model labeled 11 out of 228 patients as high-risk, whereas they were previously labeled as medium-risk based on the European Leukemia Net (ELN) 2010 criteria. All 11 patients died within two years of diagnosis, and their risk group is not predictable with other currently used methods.

I. INTRODUCTION

Motivation

Hematologic malignancies are types of cancer that initiate in the cells of the blood-forming tissue, (i.e., the bone marrow), or in the cells of the immune system (Forman et al., 2015). AML is an example of hematologic malignancy. It is an aggressive type of blood cancer and accounts for 1.2% of cancer deaths in the United States and, if not treated, it can lead to death within months after diagnosis (Jemal et al., 2002). Based on current methods, AML patients are classified as low, medium and high-risk.

There are practical treatments for high- and low-risk patients such as chemotherapy and bone marrow transplant surgery. Medium risk patients are a group of high- and low- risk patients but, their actual risk level is not detectable based on current methods. Since the survival rate of this cohort of patients is not clear, it is not possible to choose an appropriate treatment for them. So, there is a demand to find a way to reclassify these patients into either high or low-risk groups based on their clinical data.

In most methods, gene expression is the main data source to find risk groups. Some research groups used mutation and molecular abnormalities along with gene expression and found risk groups of more patients. I decided to use genes' methylation level since it is more robust and stable than gene expression (Paziewska et al., 2014).

Epigenetics studies the chemical changes in DNA molecules without any alteration in the nucleotide structure. Such changes can alter the expression of genes and how the resulting proteins function. DNA methylation is an example of epigenetic changes (Bird, 2007). A methyl group, CH_3, contains one carbon atom bonded to three hydrogen atoms. During the DNA methylation process, methyl groups are added to DNA and bind to thymine or cytosine. It usually represses

gene transcription. This procedure regulates gene transcription in human cells. The levels of methylation differ in most cancerous and noncancerous cells.

At the beginning, I used DNA methylation solely to categorize AML patients and the results were promising. Therefore, I decided to use DNA methylation and gene expression information together. There are other approaches that analyze each data type separately and then combine the results (Li et al., 2009) (Gevaert et al., 2013). My developed approach is based on analyzing different datatypes together based on multi-view clustering. It is a novel approach because:

- It found new information about AML disease which is not obtainable by analyzing different data types solely.
- It not only does not deal with the loss of statistical power issue but, the power increased compared to using datasets separately.

Related work

Cox proportional hazard model

In the datasets that I used in this project, the clinical information of some cases was censored (e.g., survival time, vital status, etc.). Therefore, I only included those cases that had gene expression, DNA methylation and clinical information at the same time. To fit a model using inferred eigengenes as the features, I needed to rank and select the most significant ones. To rank eigengenes based on their impact on predicting survival time, I used the Cox proportional hazard model (Cox, 1992):

$$h(t) = h_0(t) * e^{b_1 x_1 + b_2 x_2 + ... + b_n x_n}$$

- $x_1, x_2, ..., x_n$ are model features, eigengene values in this case.
- $b_1, b_2, ..., b_n$ are coefficients that measures the impact of features on the survival.

- t represents the survival time
- $h(t)$ is the hazard function based on time. The occurrences of hazard on each time is determined based on input features and their covariates.
- $h_0(t)$ is the baseline hazard that shows the value of $h(t)$ when the value of all input features is 0.

Current methods

FEM

Functional Epigenetic Modules (FEM) is an R package based on an algorithm that integrates and analyzes Infinium 450k DNA methylation and matched or unmatched gene expression data (Jiao et al., 2014). The aim of this analysis is to find the epigenetically regulated gene modules or molecular pathways that play a key role in cellular differentiation and disease. The procedure of this pipeline is based on four main steps, which are:

1. Integration of DNA methylation, mRNA expression, and Protein-protein Interaction Network (PPI)
2. Construction of the weighted integrated network
3. Inference of FEM
4. Identification of top targets within an FEM

Here is a brief summary of the above-mentioned steps: Since the DNA methylation data is defined for each loci of the genome, it is important to convert these values to gene level. To do so, a method conducted by West *et al.* (West et al., 2013) which is based on the distance from the Transcription Start Sites (TSS) is used (Jones et al., 2013). It creates a DNA methylation profile matrix for all patients. It creates another profile matrix for the gene expression data, too. To define a single statistics value for each gene-based DNA

methylation and gene expression data separately, it computes the t-statistics of gene expression or DNA methylation with the Phenotype of Interest (POI), which in this case is AML disease (Smyth, 2004) (Zhuang et al., 2012). The next step is to define a network based on genes and their interaction. To find which genes interact with each other, this method uses the PPI network (Cerami et al., 2010). The network consists of 8,434 genes which are labeled with the NCBI Entrez ID and their 303600 documented interactions (edges). Thus, any other genes that are in either gene expression or DNA methylation dataset but not defined in the PPI network are omitted in this step.

To find heavy subnetworks from the main network of genes, it uses the spin-glass algorithm (Reichardt and Bornholdt, 2006). This algorithm selects a limited number of genes (originally 100 genes) as seeds of heavy subnetworks. Then, it looks for neighbors of these selected genes to increase the average weight and size of each subnetwork. At the end, those subnetworks whose weight is greater than a defined threshold are selected as FEM subnetworks.

I have used this package to try and find subnetworks with the aim of doing survival analysis using genes of top selected subnetworks. There were several issues with this package:

1. Since a greedy algorithm is used in this pipeline to find subnetworks, running this pipeline multiple times results in a different number of subnetworks each time despite using the same dataset as its input data.

2. It is not possible to find which loci are mapped to a gene. Consequently, it is not possible to analyze the distance of selected loci to the mapped gene.

3. The number of genes that are used in the PPI network is limited. To define a network based on both gene expression and DNA methylation data, I had to omit those genes which were not included in the PPI network, which results in loss of data and unreliable final stratification of patients.

LSC17

Ng *et al.* introduced a new score-based method to classify AML patients into different risk groups (Ng et al., 2016). The score is defined as the weighted sum of gene expression of 17 specific genes. The weight of genes is modified and varies between -0.070400 and 0.087400 (Table 1.1). It is a rapid method which can be used in large datasets, since it classifies patients based on the average sum of the expression of 17 genes among patients. Those patients with a score higher than the average are classified as high-risk patients while those that have a rating lower than the average are low-risk patients.

Table 1.1: Genes and their weight used in the LSC17 method

Gene Symbol	**Weight (coefficient)**
DNMT3B	0.0874
ZBTB46	-0.0347
NYNRIN	0.00865
ARHGAP22	-0.0138
LAPTM4B	0.00582
MMRN1	0.0258
DPYSL3	0.0284
KIAA0125	0.0196
CDK6	-0.0704
CPXM1	-0.0258
SOCS2	0.0271
SMIM24	-0.0226
EMP1	0.0146
NGFRAP1	0.0465
CD34	0.0338
AKR1C3	-0.0402
GPR56	0.0501

ELN 2017

The European LeukemiaNet is a publicly funded research network with the aim of curing several types of leukemia by integration of European leukemia research and spread of excellence (Döhner et al., 2010). ELN 2017 is an update to the recommendations published by the ELN network in 2010. Clinicians and

researchers have widely used these recommendations regarding AML. ELN 2017 provides updates based on progress and developments in the prognosis and diagnosis of AML during these years (Döhner et al., 2016). The ELN 2017 methodology is based on finding a classifier using mutations and molecular abnormalities that are highly detected in patients with AML. It stratifies patients into three different risk categories, which are favorable, intermediate and adverse risk patients. Patients are stratified into four risk groups as favorable, intermediate-risk I, intermediate-risk II and adverse, based on ELN 2010.

Gerstung

AML is a kind of heterogeneous disease. Most of the methods for analyzing risk are based on recognizing specific mutated cancer genes, while there are only a few cancer genes which are therapeutic targets. Others are mutated in the tumor type, with a very low probability. Each tumor type has several driver genes, and different combinations of these driver genes are observed in different patients. Hence the demand for studying cancer genes and their clinical info in a large dataset simultaneously. The focus of the Gerstung *et al.* method is mainly about analyzing a large bank of information which contains 1,540 AML patients' clinical data and cytogenetic profiles of 111 of their cancer genes (Papaemmanuil et al., 2016). The methodology involves finding subcategories of AML to improve the specificity of advised treatments for patients. The researchers also validated their results using the TCGA dataset.

Overview of the implemented methodology

The purpose of this study was to provide a method that can find the group of genes effective in AML development by using both DNA methylation and gene expression data. R is the programming language used to implement functions and for constructing the networks. R is a high-level statistical programming language. It is a powerful tool for analyzing big data in different projects, since it

provides statistical and graphical functions (Team et al., 2018). Here is a summary of the steps followed in constructing the pipeline for the project:

First, I implemented a network based on DNA methylation data, where nodes are genes and edges are weighted based on the DNAm correlation between two different genes. Since DNA methylation data contains the amount of methylation on different loci of the human genome, I transformed these data to gene level methylation data to get a unique value of DNA methylation for each gene. I therefore applied network analysis on the mapped loci of each gene and then selected the cluster which shows loci highly correlated to each other. Then I defined a value for these chosen clusters to be the DNA methylation value for the related gene. I used Principal Component Analysis (PCA) as a tool to compute the weighted sum of the DNA methylation value of loci of each cluster.

Secondly, I had to define another network of genes, where nodes are genes and edges are the correlation of their gene expression and DNA methylation level. I created this network using `WGCNA` package in R (Langfelder and Horvath, 2008). This generated gene modules out of 12,000 genes. For each gene module, I defined a weighted sum value using PCA, based on their gene expression (Abdi and Williams, 2010). I used the `Pigengene` package in R to compute these values (Zare et al., 2016).

To select top modules that are highly correlated with survival time, I used the Cox proportional hazard model (Cox, 1992). It takes survival time, features and vital status of patients, as its input and ordering features are based on their effect on estimating survival time.

To create a model based on the selected modules, I used the accelerate failure time (AFT) model (Kalbfleisch and Prentice, 2011). It takes gene modules and their values for each patient, vital status of patients and their survival time as its input and estimates patients' survival time. In addition, it classifies patients into high, medium and low-risk groups based on their estimated survival time.

To validate results, I constructed the model for a different dataset using genes of previously selected modules. It can recognize a group of high-risk patients who were previously labeled as medium-risk patients but died within 2 years. In addition, I compared the results with the results of current popular methods as well as with recently published methods.

II. METHOD

Each section of the following pipeline is a separate R script. To run all scripts together or individually, I wrote the runall.R script as a driver and defined different tasks in it (Noble, 2009). Each task is related to a unique section of the pipeline. In this chapter, I explain the implemented methodology in detail.

The datasets used in this study

In this project, I used two different datasets to train and test the pipeline. For the training phase, I used the dataset which is provided by TCGA (Wan et al., 2015). This is publicly available from their website. It contains clinical, genomic and epigenomic information of 200 AML patients. The TCGA DNA methylation dataset contains the level of methylation of 485,577 loci on the human genome for 194 AML patients. The TCGA gene expression dataset contains the expression of 20,442 genes for 179 AML patients. To test the pipeline, I used the dataset provided by Herold *et al.* (Herold et al., 2014). It is publicly accessible through Gene Expression Omnibus (GEO) (accession number: GSE37642) (Barrett et al., 2013). It contains the genomic, epigenomic and clinical information of 494 patients. I downloaded this data using `getGeo` function from the `GEOquery` R package (Davis and Meltzer, 2007). The following paragraphs discuss in detail the process of preparing gene expression and DNA methylation data for the network analysis.

Preparing DNA methylation data

The DNA methylation dataset contains loci on the genome whose methylation value is not available for more than half of the patients. Besides, there are loci on the genome which have low variation. To filter out this information from the dataset, I used the `RnBeads` package (Assenov et al., 2014).

`RnBeads` is an R package that contains tools for analyzing DNA methylation data of different assays such as Infinium 450k microarray, Whole–Genome Bisulfite Sequencing (WGBS), Reduced Representation Bisulfite Sequencing (RRBS), other forms of enrichment bisulfite sequencing and any other large-scale method that can provide DNA methylation data at single base-pair resolution (e.g., MeDIP-seq after suitable preprocessing). I cleaned the DNA methylation data using some of the `RnBeads` package functions, in three steps:

1. There were loci in the dataset for which a methylation value was not available for more than 50% of patients. I omitted such loci from the rest of the normalization process.

2. The amount of methylation across the genome is required to be normalized before using it for the rest of the downstream analysis. To do this, I used the `rnb.run.preprocessing` function from the `RnBeads` package. It takes an unfiltered `RnBeads` object which contains DNA methylation information and returns an RnBeads object containing the normalized DNA methylation values. The amount of normalized DNA methylation varies from 0 to 1.

3. After normalization, I tried to find those probes whose methylation values highly correlate with patients' survival time. I excluded those probes whose absolute value of Spearman correlation with the survival time of dead patients was less than 0.2 (Daniel, 1990). I chose this threshold based on the distribution of the correlation value of DNA methylation with the survival time of dead AML patients. Those excluded probes would be added later for the network analysis if they were mapping to a gene whose expression level correlates highly with the survival time of dead patients.

There were 367,979 probes excluded during those steps, as mentioned above. Finally, there were 24,649 probes remaining from the preprocessing steps.

Preparing gene expression data

The gene expression dataset also contains some genes which have low variation, i.e., their degree of expression of standard deviation is less than 10^{-8}. I filtered out these genes using a function called `load.data`. It takes expression data as its input and returns filtered genes and their expression values as the output. After that, I computed the absolute value of the Spearman correlation between gene expression and survival time of dead patients and sorted them decreasingly (Daniel, 1990). Based on experience in gene expression data cleaning for network analysis in other projects of the lab, I kept 1/3 of those genes that highly correlated with the dead patients survival time (Zainulabadeen et al., 2017).

Union of gene expression and DNA methylation data

In this project, I analyzed the expression and methylation data at the same level. To do this, I made a combined set of genes, of which either the expression or methylation value highly correlated with dead patients' survival time. There were some probes which were mapped to multiple genes. Also, some genes were mapped by multiple probes. I resolved these issues after this step using network analysis, which I have explained in detail in the next section. By taking a union of genes that were selected during the last two steps, I had a set of genes which had valuable information about the AML disease, based on either their gene expression or DNA methylation amount.

Network analysis

After selecting genes based on their gene expression or DNA methylation value, I had a considerable number of genes which I assumed correlated with the AML disease. To find biomarker genes from this massive set, I used network analysis. This procedure consists of two main steps:

1. Trim: Network analysis for probes of mapped genes
2. Network: Network analysis for selected genes
3. Eigengenes: Computing eigengenes for subnetworks of the genes

The steps are explained in detail in the following paragraphs.

Trim

As I mentioned in the previous section, it is challenging to determine a single methylation value for every gene since there were some genes which have multiple probes mapped to them (Koestler et al., 2014) (Jones, 2012). In this case, I had to define a value as the DNA methylation for these genes. To do this, I used the hierarchical clustering and a greedy approach for these specific genes (Yip and Horvath, 2007) (Langfelder and Horvath, 2012). Here is a brief summary of the algorithm.

For each gene:

- If the number of related mapped probes are less than or equal to 6, the amount of DNA methylation for the gene g_i is the average of methylation amount on the loci of mapped probes.
- Otherwise:
    1. Use the `cluster_fast_greedy` function from the `igraph` package to cluster mapped probes (Csardi and Nepusz, 2006).
    2. Take an average of the edge weights of each cluster and rank clusters according to these values.
    3. If there is only one cluster which has the maximum amount of average edge weights, select this cluster as the representative of the gene g_i.. Assign the sum of DNA methylation level of the probes of this cluster to the DNA methylation level of gene g_i.

4. Otherwise, select patients that are originally labeled as high or low-risk based on the dataset. Compute eigengenes based on the DNA methylation of these patients for the probes of clusters that had maximum average edge weight. Then, project these values for originally medium-risk labeled patients, too.

The `cluster_fast_greedy` function from the `igraph` package tries to find dense subgraphs. It measures the optimization of the density based on the modularity score which is defined for each subgraph. The goal of this function is to find subnetworks with maximum score (Csardi and Nepusz, 2006).

To compute eigengenes, I used the `compute.pigengene` function from the `Pigengene` package (Foroushani et al., 2017). It uses PCA to compute eigengenes for the DNA methylation value of probes. To project eigengenes for originally medium-risk patients too, I used the `project.eigen` function from the `Pigengene` package. This function projects DNA methylation of medium risk patients onto the eigengenes of modules from high and low-risk patients.

Network

In this step, I found groups of genes which are highly correlated with each other based on their methylation levels and gene expression values. To do so, I defined two separate networks based on gene expression and DNA methylation data. The nodes in both networks are genes, and edges are defined as the Pearson correlation between every two genes (Benesty et al., 2009). I created these two networks using the `adjacency` function from the `WGCNA` package. `WGCNA` is an R package containing a comprehensive collection of R functions for performing various aspects of weighted correlation network analysis. The package includes functions for network construction, module detection, gene selection, calculations of topological properties, data simulation, visualization, and interfacing with external software (Langfelder and Horvath, 2008). To combine the weight matrix of these two networks and make a combined network of genes based on both

DNA methylation and gene expression values, I used the following formula:

$$\mathcal{W}(g_i, g_j) = (1 - \lambda)\left|\text{cor}_E(g_i, g_j)\right| + \lambda\left|\text{cor}_M(g_i, g_j)\right| \tag{II.1}$$

where $\left|\text{cor}_M(g_i, g_j)\right|$ and $\left|\text{cor}_E(g_i, g_j)\right|$ are weights of the previously constructed networks based on DNA methylation and gene expression values. For each i and j index, the value of the $W(g_i, g_j)$ shows the weight of the edge between the g_i and g_j in the combined network. It is a combination of both $\left|\text{cor}_E(g_i, g_j)\right|$ and $\left|\text{cor}_M(g_i, g_j)\right|$. The λ value is a factor which determines the degree of combination of the weight matrix for each of these two networks. The value of the λ coefficient varies between 0 to 1. Based on several experiments, I selected 0.6 as the value of the λ coefficient, since the final model based on this λ was the best model. I have explained in detail about the method of comparing models and selecting the best one in the survival analysis section.

The degree of correlation between some pairs of genes is very low. To decrease the impact of these pairs of genes on the next step of the analysis, I utilized a soft-threshold using the `pickSoftThreshold.fromSimilarity` function from the `WGCNA` package (Zhang and Horvath, 2005) (Horvath and Dong, 2008). It takes the network and other related parameters as it's inputs and returns the best power of the networks with regards to its connectivity as the soft threshold. The best soft threshold for my network was 6.

Then, I found highly correlated subnetworks of genes. To do so, I used the `blockwiseModules` function from the `WGCNA` package. It takes the $Network$ matrix, whose values are raised to the power of the amount of the soft-threshold, which is 6 in this case, and other related parameters such as:

- `maxBlockSize` : The maximum size of subnetworks. I used the number of genes in the W matrix as the value of maxBlockSize.
- `corType` : This parameter shows the correlation method which can be either *Pearson* or *bidweight midcorrelation*. I used the *Pearson* correlation.

- `minModuleSize`: The minimum size of subnetworks. Based on several experiments, I got the best final model in the survival analysis phase while using at least 5 genes in each subnetwork.

Eigengenes

In the previous section, I tried to classify a vast number of genes into several clusters of highly correlated genes. The next step was computing an eigengene as a weighted average value based on either gene expression, DNA methylation or a combination of both of them for each of the clusters. I used the `compute.pigengene` function from the `Pigengene` package to compute these values (Zare et al., 2016). The `compute.pigengene` function takes as the input a defined value for each gene and the related cluster assignments, and computes an eigengene for each cluster using PCA. For each gene in the cluster, there were three possible ways to define an eigengene:

- $dnam_{i,j}$: DNA methylation of the $gene_{i,j}$
- $expr_{i,j}$: gene expression value of the $gene_{i,j}$
- $\lambda * dnam_{i,j} + (1 - \lambda) * expr_{i,j}$: a combination of gene expression and DNA methylation of the $gene_{i,j}$

I decided to use only the gene expression values, due to the fact that the number of datasets that contain both gene expression and DNA methylation values is too low, while most datasets contain gene expression values. Thus, the pipeline will be useful for many datasets.

Survival analysis

The methodology that I used for performing survival analysis is based on the survival analysis that the other collaborators performed for a melanoma dataset (Zainulabadeen et al., 2017). I improved their scripts and added some extra features to their functions.

I used the `glmnet` function from the `glmnet` R package (Simon et al., 2011) to perform a penalized Cox analysis (Gui and Li, 2005). I set the $\alpha = 1$ for using the Least Absolute Shrinkage and Selection Operator (LASSO) in `glmnet` function (Tibshirani and Efron, 2002). The LASSO set the coefficients of most eigengenes in the Cox proportional hazards model to be zero. Consequently, it identified the modules that highly associate with the survival time (Zainulabadeen et al., 2017).

Using the output of this function, I selected the top three eigengenes. Using three important features guarantees that the final model will not over-fit the training data. I fitted an AFT model, which is a survival regression model to inferred eigengene values (Kalbfleisch and Prentice, 2011). I used the `survreg` function from the `survival` package. It took eigengene values, actual survival time of patients and their vital status as its input (Therneau and Grambsch, 2000). I used this model to predict the survival time of patients. I defined two time thresholds to classify patients into risk groups. The first threshold is the maximum time that a high-risk patient is alive. The second threshold is the minimum time that a low-risk patient is alive. These thresholds are computed based on the minimum recall for low-risk and high-risk patients. I defined 0.2 and 0.05 as the minimum recall for low and high-risk patients. Those patients whose survival times are in the range of the two thresholds are counted as medium-risk patients.

I used the `survdiff` function from the `survival` package to find whether the low and high-risk patients differ significantly. It computed a log-rank p-value base on Mantel-Haenszel test for low and high-risk patients (Mantel and Haenszel, 1959). To show the results graphically, I used the `survfit` function from the `survival` package and drew a Kaplan-Meier survival curve for each of the risk groups (Habib et al., 2008). It contains a curve for each risk group. Each point on the curve shows the survival probability of patients at a certain point of time.

Pathway analysis

To check if the genes of top selected modules are participating in any known pathway, I performed pathway analysis using the innateDB website. InnateDB is a web-based tool which has been developed to facilitate systems level investigations of mammalian physiology such as human, mouse and bovine innate immune response. It provides functions for analyzing information of the manually-curated knowledge-base of the genes, proteins. In particular, functions are focused on the interactions and signaling responses involved in mammalian innate immunity (Breuer et al., 2012). I used KEEG (Kanehisa and Goto, 2000) and Reactome (Croft et al., 2014) repositories on InnateDB to determine the biological pathways associated with each of the gene modules. I converted the gene symbols of top selected modules to EntreZ IDs. Then I uploaded these gene IDs on the InnateDB website, the pathway analysis section. To find significant pathways, I performed Pathway Over-Representation Analysis (Pathway ORA). I used the Hyper-geometric test as the analysis algorithm and the Benjamini & Hochberg approach as the correction method.

Validation

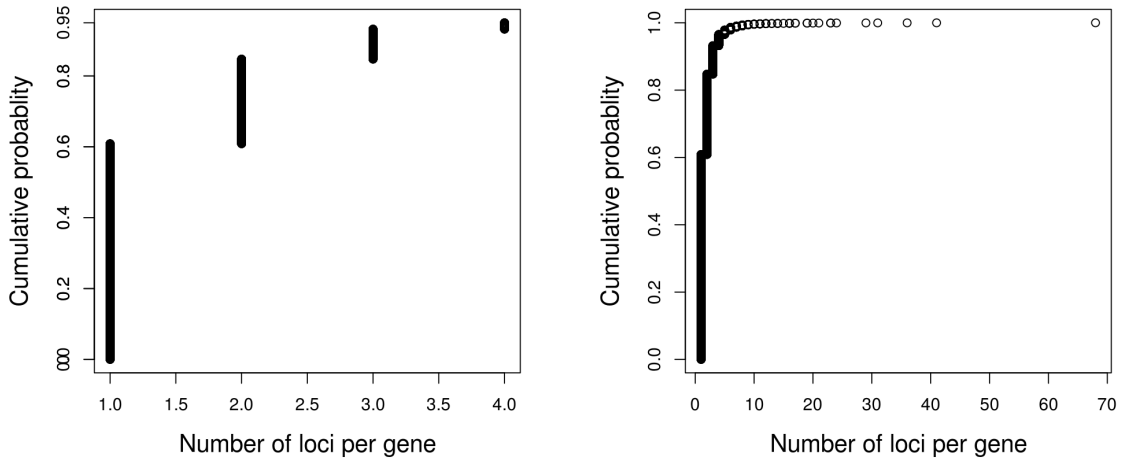
To validate results from the previously described constructed pipeline, I wrote some additional scripts that can prepare datasets for being analyzed during the test phase. These scripts are `integrate` , `valid_prep` and `valid_survival` . They are explained in detail in the following paragraphs. To save key information about the model that I constructed using the TCGA data, I used `integrate script` . I saved the following information about the top selected modules as an R object:

- Gene names and their weight in computing the eigengene of each module
- Loci mapped to each gene and their weight in computing the DNA methylation value of the gene.

- The λ value that was used in construction of the network.
- The DNA methylation and gene expression data combination method.

Different clinical labs use different methods and protocols in their projects. There are different methods for creating a gene expression profile in labs. Using key factors of the model based on the TCGA dataset, it is possible to create a model that fits different sets of gene expression information. The above-mentioned items are the minimum required information in order to build a model that fits with the test dataset information. I included significant information of DNA methylation values in the process of network construction. To build a model that fits the new datasets, DNA methylation values are an optional input. To prepare gene expression of new datasets, I wrote the `valid_prep` script. The dataset that I used to test the model had information for two cohorts of AML patients.

The gene expression values of these two cohorts were computed using three different platforms, GPL96, GPL97 and GPL570. The `valid_prep` script integrated this information at the very first step. Next, it normalized the gene expression values and prepared them for computing eigengenes. Using the integrator object from the integrate script, it computed new eigengene values using the gene expression information of the test dataset. These values are mandatory inputs for survival analysis. The outputs of the `valid_prep` script were inferred eigengene values. I wrote the `valid_survival` script to perform survival analysis using eigengene values. In this phase, there is no need for a Cox analysis, since we already have the top modules and their related eigengene values. The rest of the procedure for fitting a survival regression model is the same as I carried out for the TCGA dataset.



(a) Genes that are mapped from at most 4 highly correlated loci

(b) All genes mapped from highly highly correlated loci

Figure 3.1: Genes and mapped loci. These loci are those whose DNA methylation value highly correlates with the survival time of dead patients.

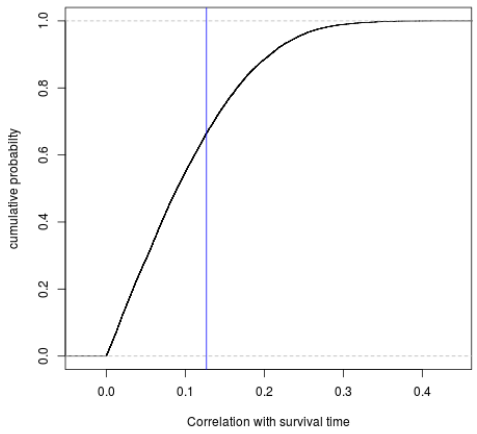
III. RESULTS

Network analysis

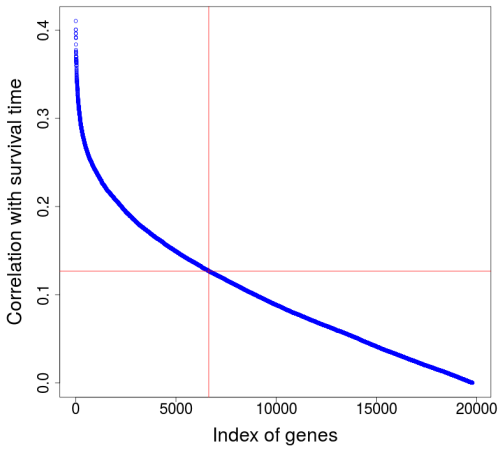
In the process of preparing DNA methylation data for the network analysis, there were some loci that had low correlation with the survival time of the dead AML patients. I excluded those loci but kept them for the union section of the pipeline. Each gene is mapped from at most 4 loci with the probability of 95% i.e., there are a few genes that are mapped from more than 4 loci. The maximum number of loci mapped to a gene is 70 (Figure 3.1).

In the process of preparing gene expression profile matrix, I filtered 19911 genes. Then, I kept only 6637 of them, which were the top third of genes that highly correlated with the dead patients survival time (Figure 3.2).

Module sizes had a mean, median and standard deviation of 127, 25 and 310, respectively (Figure 3.3). For each of the modules, I computed an eigengene, which is a weighted average of expression levels of the genes in that module.



(a) Cumulative probability of genes and their gene expression correlation with the survival time of dead patients.



(b) Sorted genes based on the absolute value of their gene expression correlation with dead patients time. Those genes that are at the top left of the diagram are selected and passed to the network analysis.

Figure 3.2: The absolute correlation value of the gene expression and survival time of dead patients. These genes are from the expression profile matrix.

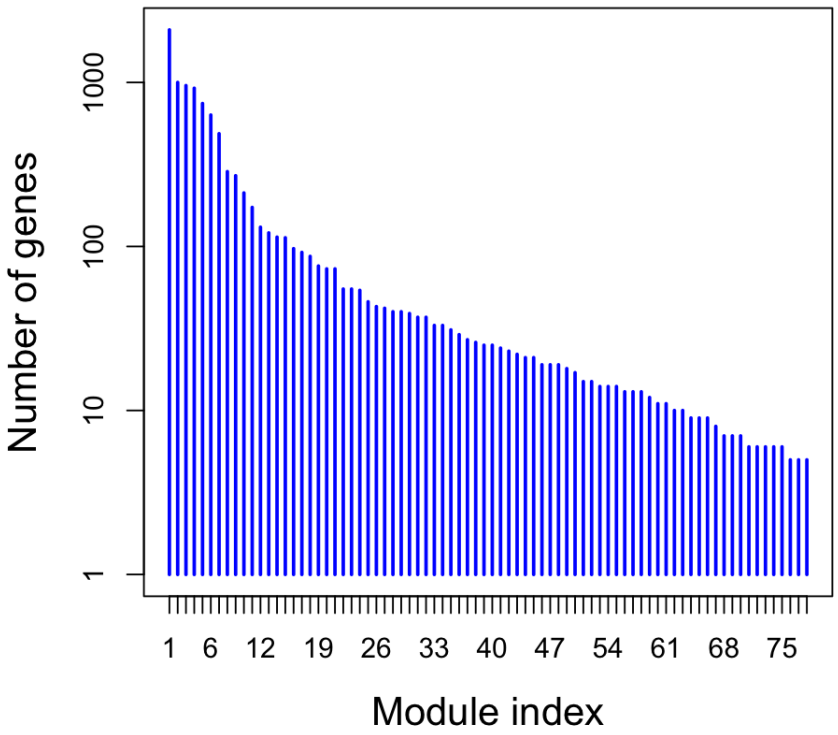


Figure 3.3: Number of genes in different clusters. The largest and smallest modules consist of 2,092 and 5 genes, respectively.

Module 56 has the eigengene most correlated with the survival time, with a Pearson correlation of 0.3, and the most anticorrelated eigengene corresponds to module 51, with a correlation of 0.4.

Survival analysis

Based on the results of the network analysis, there were 78 inferred eigengenes as covariates (prognostic features). Since the information on some patients was censored, I included only the 154 AML cases for which their vital status, DNA methylation and gene expression data were available from the TCGA dataset. Ordering these features based on their impact to predict the survival time, using `glmnet` function to sort features based on their impact on the survival time, showed that the top three modules are 55, 51 and 46. These modules contain 14, 15 and 19 genes, respectively (Table 3.1).

Using the accelerated failure time model, it was revealed that the best model to predict survival time was the combination of eigengenes of modules 46, 51 and 55. Using this classifier, 26 of the cases were predicted as high-risk, 27 as low-risk and 101 as medium-risk. Out of 26 high-risk patients, one left the study in less than a year and 25 of them died within two years. This indicates the high sensitivity of the network classifier model for high risk patients. The p-value 10^{-11} shows that the survival curves that correspond to high-risk and low-risk groups differ significantly (Figure 3.4 and Figure 3.5).

While 93 patients died of AML (mean = 1.1, median = 0.8, and standard deviation = 1 years), 61 cases were alive at the last follow up time (mean = 2.5, median = 2, and standard deviation = 2 years).

Based on cytogenetic abnormalities criteria, the TCGA cohort was previously classified as 31 low, 90 medium, and 31 high-risk cases (Network et al., 2013). The p-value 10^{-3} shows that the survival curves that correspond to high-risk and low-risk groups do not differ significantly from each other. High and low-risk curves seem to have the same patterns in the initial years of

Table 3.1: Gene symbols of the top three modules and their related weight in the computation of eigengenes. These modules are selected using the Cox proportional hazard model

Module 46	**Weight**	**Module 51**	**Weigth**	**Module 55**	**Weight**
CCDC64	0.65	KCTD17	0.87	PCSK4	-0.77
BCL2L11	0.56	PMM1	0.8	COQ10A	-0.74
PPDPF	0.55	SFXN3	0.67	MAP6D1	-0.74
GLTSCR1	0.55	PDZD7	0.56	YWHAH	0.7
IL6ST	0.54	PHGDH	0.47	LRP10	0.64
RIMS3	0.52	SEC14L2	0.47	CALHM2	0.6
SLC35E4	0.51	LOC285830	0.44	REEP6	-0.58
TUBB2A	0.48	RDH13	0.39	GPX1	0.57
SQLE	0.46	TSHZ3	-0.3	LCOR	0.26
C1orf204	0.41	NKX3-1	-0.22	TRIM65	0.2
C11orf84	0.4	PLAUR	-0.18	UBC	0.16
CAMK4	0.39	CCDC85C	0.17	C22orf24	0.12
ABHD11	0.35	KCNJ5	-0.05	TMEM65	-0.1
PHLDA3	-0.34	NT5C3L	0.03	PGPEP1	0.06
TSPAN14	0.33	AP1S1	-0.01		
LRRN2	0.32				
DUSP3	-0.31				
SOCS2	0.12				
RAB3IP	-0.04				

survival (Figure 3.6).

Comparison of the results using a confusion matrix shows that there were 13 patients whom both network classifier and cytogenetic information labeled as low-risk patients, while there were 3 patients that were classified as low-risk using network classifier and high-risk using cytogenetic information (Figure 3.7). In addition, there were 11 patients whom both network classifier and cytogenetic information labeled as high-risk patients. Also, there were 2 patients classified as high-risk using network classifier and low-risk using cytogenetic information. Based on the advice of Dr. Aly Karsan, the lab collaborator from British Colombia Cancer Agency, I focused on patients that were predicted as

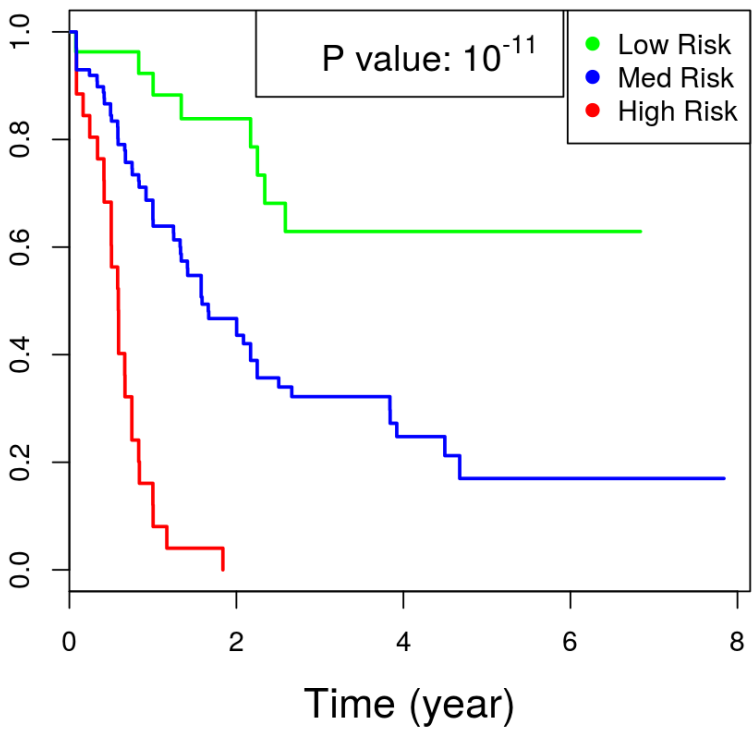


Figure 3.4: Km-plots for TCGA data using network analysis. The X-axis shows the time in year and the Y-axis shows the probability of being alive at a certain point of time. Curves shows the probability of being alive during the time for low, medium and high-risk AML patients.

medium-risk group with classifiers other than network classifier. These cases are more clinically interesting to prognosticate.

Then, I focused on the 90 patients who were labeled as medium-risk patients using cytogenetic information. Based on network classifier, 11 of them were low-risk, 66 were medium-risk and 13 were high-risk patients. All 13 high-risk patients died (or left the study) within two years (Figure 3.8).

Gerstung *et al.* recently reanalyzed 111 cancer genes, cytogenetic profiles and clinical data from 1,540 AML cases, and showed that their integrative approach provides considerably more informative and accurate statements than the current standards in clinical practice (Gerstung et al., 2017; Papaemmanuil et al., 2016). In particular, validation using data from independent patients in the TCGA cohort revealed that the Gerstung approach is superior to the

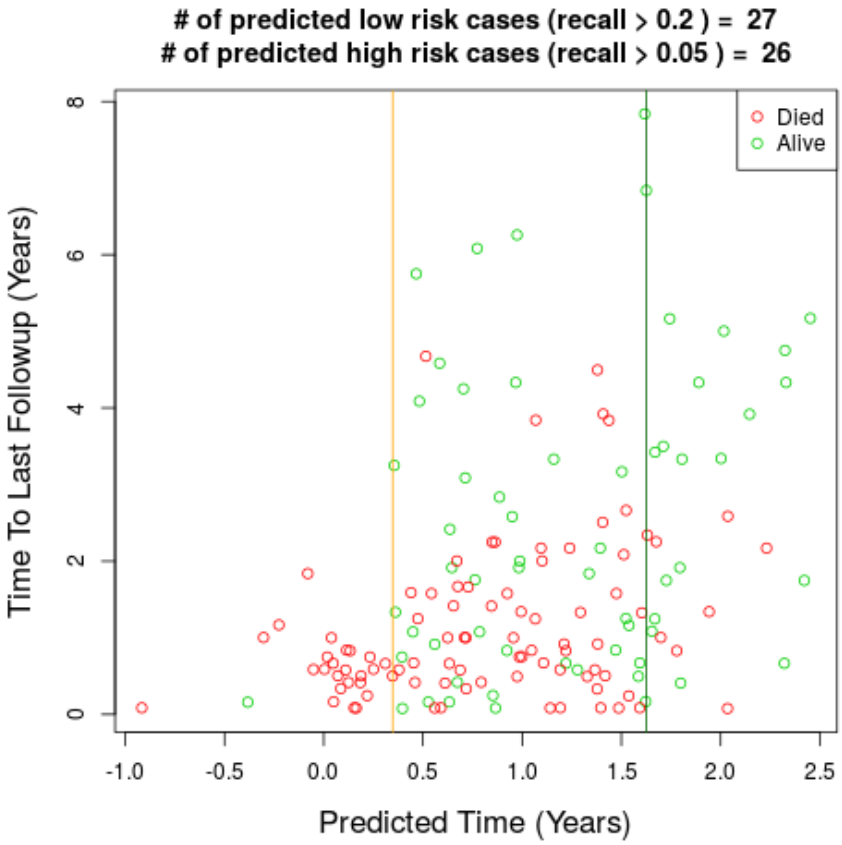


Figure 3.5: Classification of AML patinets in TCGA dataset; The X axis shows the predicted time based on network analysis while the Y axis shows the patients actual survival time. The vertical orange and green lines show the maximum and minimum predicted survival time based using network analysis.

prognostication based solely on cytogenetics. Their results using the TCGA dataset showed that 39 of cases were grouped as low-risk, 81 as medium-risk and 33 as high-risk patients.

The group of 81 patients that Gerstung *et al.* classified as medium-risk included a mix of high and low-risk cases. That is, in this group, 49 (60%) cases died of AML while there were 14 (17%) other cases that were followed for at least two years after diagnosis and were alive at the last time of contact. It is thus clinically critical to further assess the risk for this subset of cases. Network classifier labeled this cohort of 81 AML cases as 10 low, 60 medium and 11 high-risk patients (Figure 3.9). All 11 high-risk labeled patients died or left the study within two years. The subset of 11 (14%) patients reclassified as high-risk cases lived for significantly shorter periods than other cases in this group

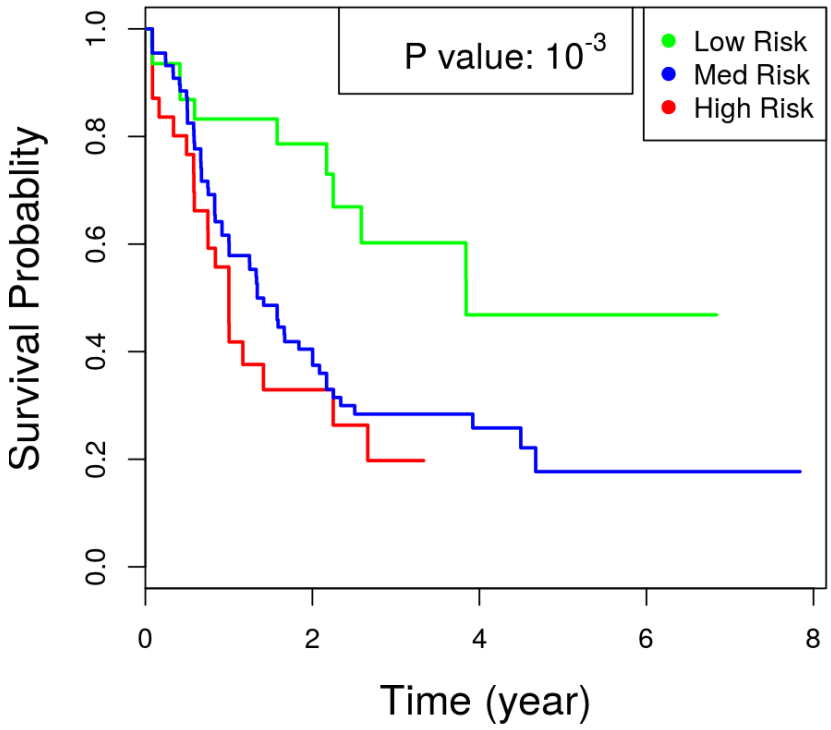


Figure 3.6: Km-plots for TCGA data using network analysis. The X-axis shows the time in year and the Y-axis shows the probability of being alive at a certain point of time. Curves shows the probability of being alive during the time for low, medium and high-risk AML patients.

	Low	Med	High	rowSum
Low	13	11	3	27
Med	16	66	17	99
High	2	13	11	26
colSum	31	90	31	152

Figure 3.7: Confusion matrix for TCGA dataset patients using network and cytogenetic informations as classifiers. Rows are results based on the network classifier and columns are results based on the cytogenetic abnormalities information.

(p-value$\leq 10^{-4}$) (Figure 3.10).

Pathway analysis

I listed the main information I gained using pathway analysis:

- Top pathways related to the genes of module 46 are described below and

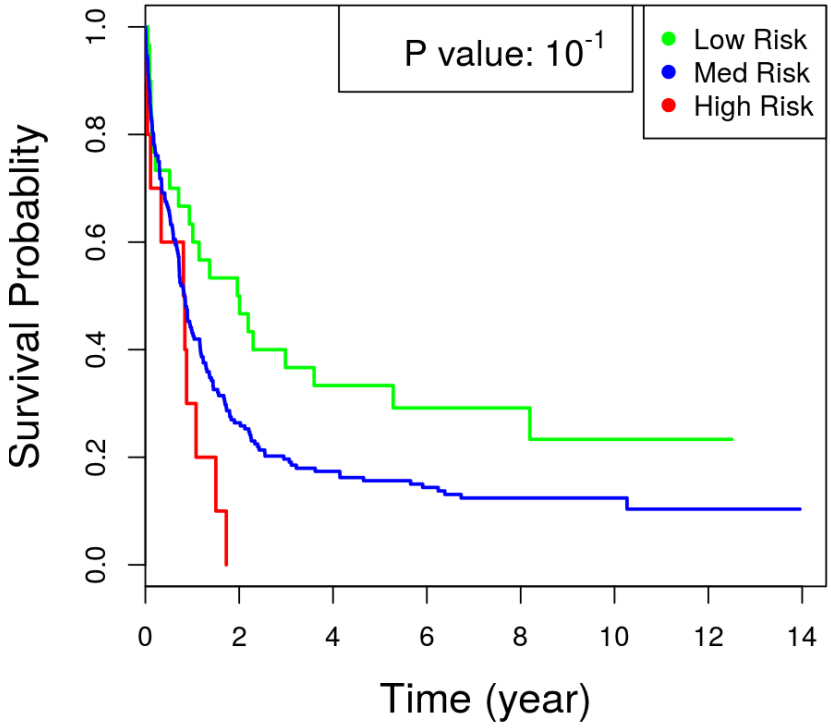


Figure 3.8: km-plots for a subset of TCGA data classified based on cytogenetic abnormalities criteria, reclassified using network classifier. This cohort of patients were previously classified as medium-risk based on the cytogenetic abnormalities criteria.

	Low	Med	High	rowSum
Low	16	10	1	27
Med	22	60	18	100
High	1	11	14	26
colSum	39	81	33	153

Figure 3.9: Confusion matrix for TCGA dataset patients using network and Gerstung *et al.* methods as classifier. Rows are results based on the network classifier and columns are results based on the Gertung *et al.* method.

shown graphically in Figure 3.11.

1. Immune System: It is made up of a network of cells, tissues, and organs that work together as the human body's defense. It protects human body from infectious organisms and other invaders through a series of steps called the immune response (Parham, 2014). Important genes found in this pathway that are highly correlated with the

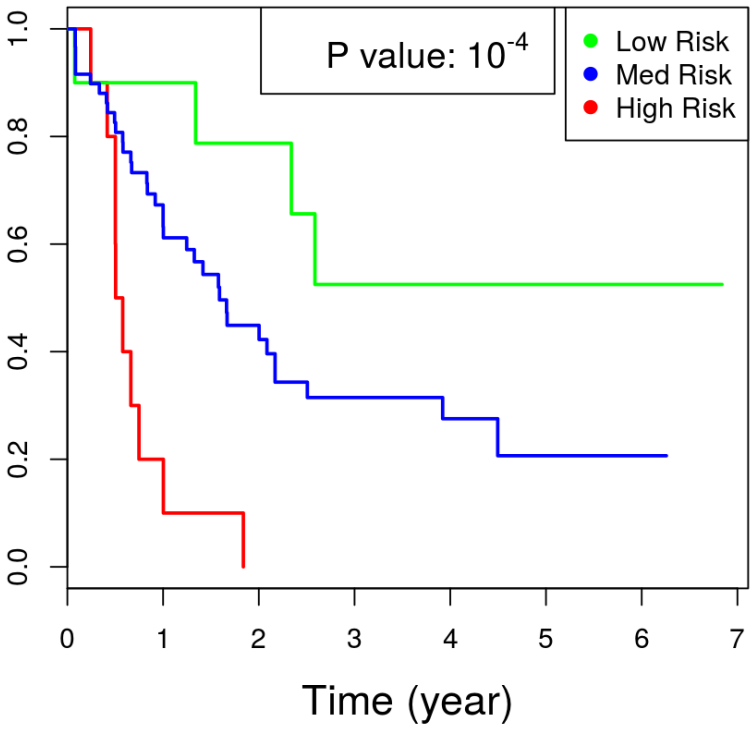


Figure 3.10: km-plots for a subset of TCGA data classified using the Gertsung *et al.* classifier, reclassified using network classifier. This cohort of patients was previously classified as medium-risk by Gertsung *et al.* .

survival time are CAMK4(+, positively correlate with survival time), DUSP3 (-, negatively correlate with survival time), IL6ST (+), SOCS2 (+), TUBB2A (+).

2. Signaling events mediated by HDAC Class II: In general, DNA is wrapped around histones, and its expression is regulated by acetylation and deacetylation process. Histone deacetylases (HDAC) are a group of enzymes that mediate the expression of DNA by removing acetyl groups from an amino acid on a histone (Seto and Yang, 2010). Important genes that are founded in this pathway are CAMK4 (+), TUBB2A (+)

3. Signalling by NGF: Neurotrophins (NGF, BDNF, NT-3, NT-4/5) are a family of proteins that play important roles in survival,

differentiation, functionality of neurons (Hempstead, 2006) (Reichardt, 2006). They can send signals about survival, differentiation or growth to a particular cells (Allen and Dawbarn, 2006). Important founded genes in this pathway are: CAMK4 (+), DUSP3 (-), BCL2L11 (+).

- Top pathways related to the genes of module 51 are described below and shown graphically in Figure 3.12.:

    1. Post–Translational protein Modification (PTM): It refers to a set of covalent and enzymetix modifications of proteins which occurs after their biosynthesis process (Walsh, 2006). Important genes are: PLAUR (-) and PMM1 (+).

- Top pathways related to the genes of module 55 are described below and shown graphically in Figure 3.13.:

    1. Apoptosis: It is a process that occurs in multicellular organisms. During this process, cell goes into a programmed death which can be initiated through an intrinsic or extrinsic pathway (Karam, 2009). Important founded genes in this pathway are: UBC (+) and YWHAH (+).

    2. Membrane trafficking: Macromolecules such as proteins are distribute from the extracellular space throughout the cell during the membrane trafficking pathway (Sadler, 2011). Important founded genes in this pathway are UBC (+) and YWHAH (+).

    3. Cellular responses to stress: It is a group of molecular changes that happens in response to unusual changes in the cell environment such as extremes of temperature. The goal of these changes is to protect cells against those unfavorable conditions (Welch, 1993). Important founded genes in this pathway are UBC (+) and YWHAH (+)

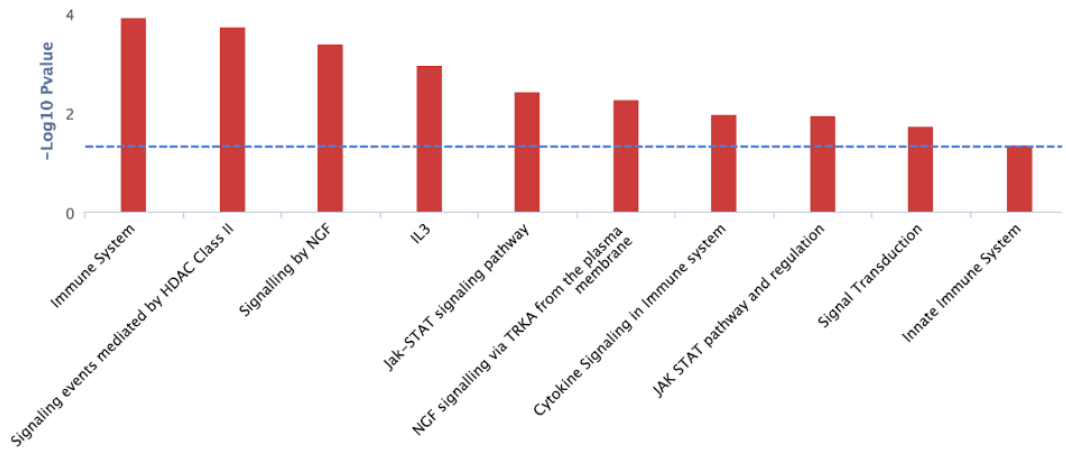


Figure 3.11: Pathway analysis for the genes of module 46. The genes of module 46 were highly associated in 10 known pathways with -$\log_{10}(p-value)$>1.3. The blue horizontal line shows the threshold for -$\log_{10}(p-value)$, which is 1.3. The very top related pathway with module 46 is the Immune System pathway.

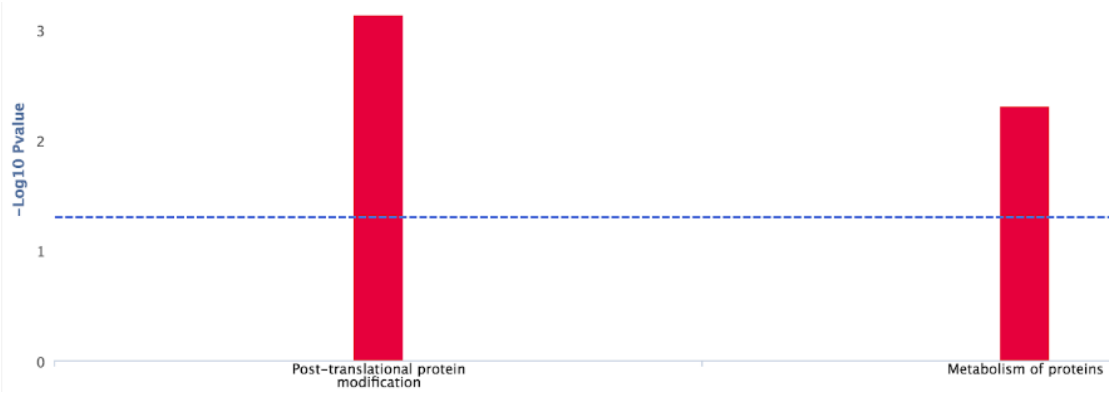


Figure 3.12: Pathway analysis for the genes of module 51. The genes of module 51 were highly associated in 1 known pathways with -$\log_{10}(p-value)$>1.3. The blue horizontal line shows the threshold for -$\log_{10}(p-value)$, which is 1.3.

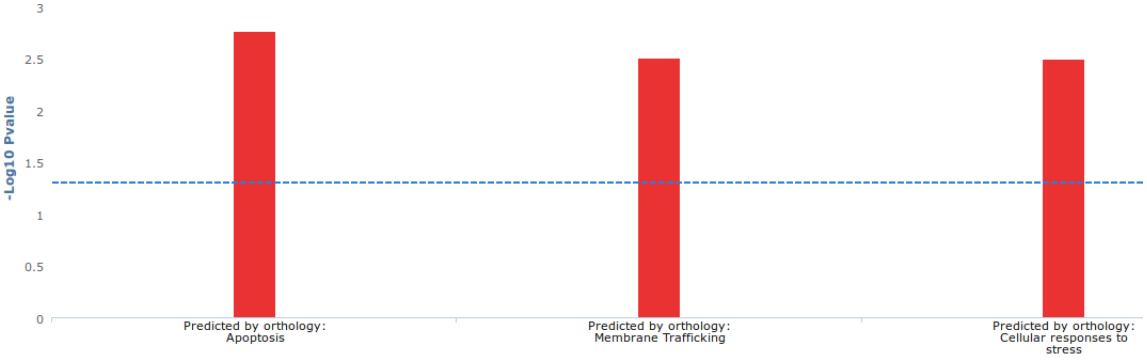


Figure 3.13: Pathway analysis for the genes of module 55. The genes of module 55 were highly associated in 3 known pathways with -$\log_{10}(p-value)$>1.3. The blue horizontal line shows the threshold for -$\log_{10}(p-value)$, which is 1.3.

Validation using an independent dataset

In order to validate the results, I used the dataset which was used by Herold *et al.* , and contains gene expression, DNA methylation and clinical information of total 562 AML patients (Herold et al., 2014).

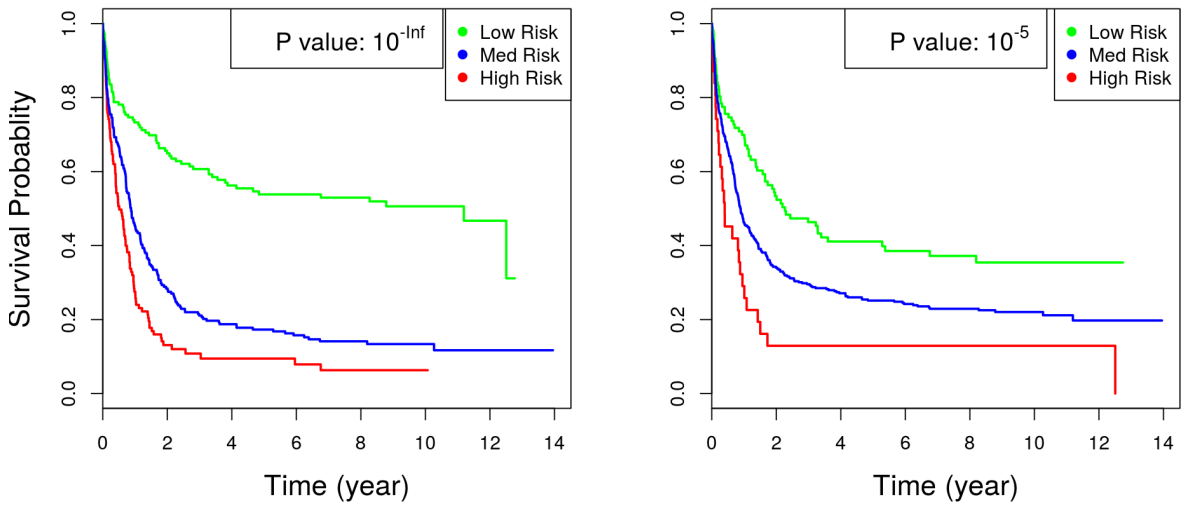
To prepare these data for the survival analysis, I used the `valid_prep` script. Since the information of the Herold dataset is available in two sets, the script also integrates this information into a unified object, too. The results of this script are computed eigengenes of the genes which were in modules 51, 55 or 46. These modules were the top three modules among the others in the previous constructed network using the TCGA dataset.

Using the `valid_survival` , I performed survival analysis for the Herold dataset. The model reclassified 553 AML patients into new risk groups. It classified 415 patients as medium risk, 107 patients as low risk and 31 patients as high-risk. These results are based on using minimum recall equals to 0.2 for low risk groups and 0.05 for high risk groups. The maximum predicted survival time is 3 years.

From 553 patients, 494 of them had uncensored data. These 494 patients were previously classified based on the ELN method: 120 patients as high risk, 146 patients as low risk and 228 patients as medium risk groups (Figure 3.14 and Figure 3.15).

The cohort of 228 previously medium risk patients classified into 30 low risk ,188 medium risk and 10 high risk patients using network classifier. All of these 10 reclassified patients were died before 2 years, which is a strong proof for the correctness of network classifier (Figure 3.16).

Since the patients mutation info of this dataset is not publicly accessible, Dr. Zare and I asked Dr. Tobias Herold for it. He performed analysis on those above mentioned 10 high-risk reclassified patients. Based on his email on March 6, 2018: "After adjustment for multiple hypothesis testing there was no significant (p-value$<$0.05) association of the variable included in the European



(a) ELN classifier

(b) network classifier

Figure 3.14: km-plots for Herold dataset. The X axis shows the actual survival time for patients and the Y axis shows the probability of survival. Each class of patients has a related curve. The p-value shows the probability of patients being classified randomly.

	Low	Med	High	rowSum
Low	41	30	14	85
Med	100	188	91	379
High	5	10	15	30
colSum	146	228	120	494

Figure 3.15: Confusion matrix for Herold dataset patients using network and ELN methods as classifier. Rows are results based on the network classifier and columns are results based on the ELN classifier.

Leukimia Net 2017 generic risk classifier nor clinical variables like WBC, hemogolobin, platelets, LDH or ECOG with the high-risk subgroup as predicted by the network classifier." It showed that the risk group of this cohort of 13 AML patients is only predictable by using the combination of DNA methylation and gene expression information.

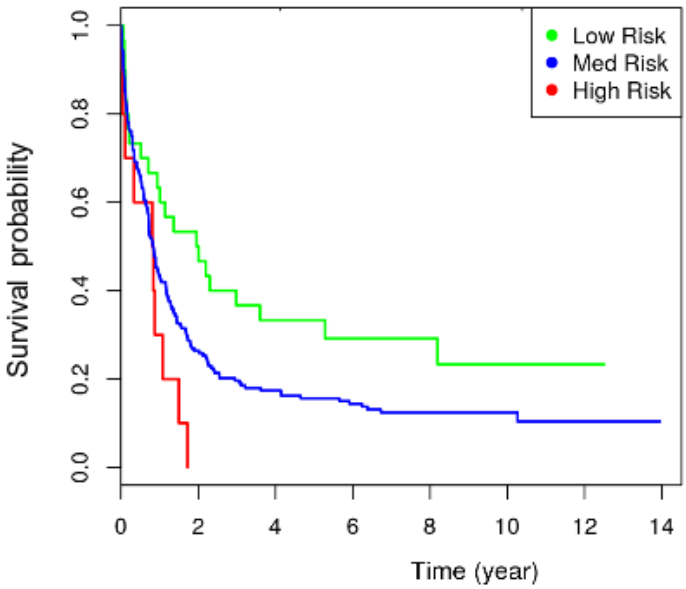


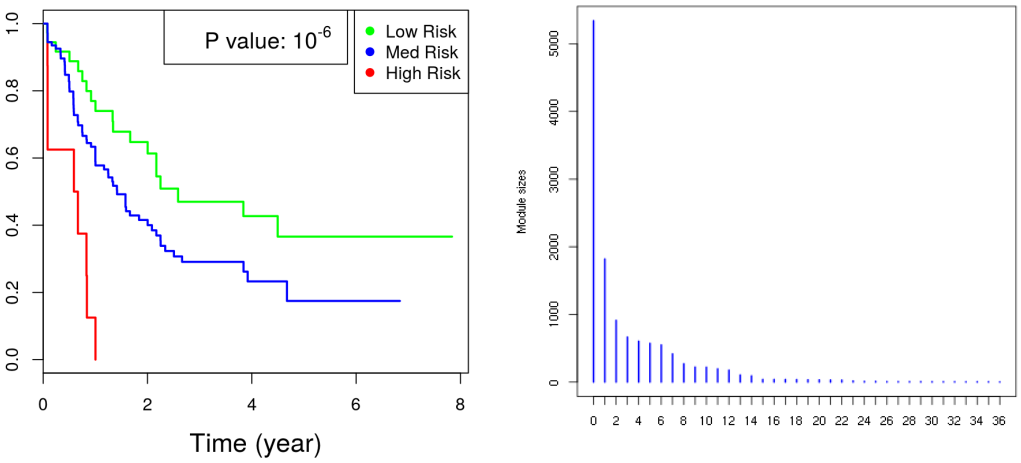
Figure 3.16: Km-plots for a subset of Herold dataset classified based on ELN classifier, reclassified using network classifier. This cohort of patients were previously classified as medium-risk based on the ELN methodology.

IV. CONCLUSIONS

n this project, I developed a novel approach to integrating different data-types based on multi-view clustering. The approach works well in data combination and improves clustering performance. I used this approach in prognostication of AML disease risk. This is a novel method since:

- Other popular methods use only gene expression data and some other mutation information. This information was not adequate to identify the actual risk group of medium-risk patients which my method identified as high-risk patients with a high sensitivity.
- The new model can work well even without DNA methylation values of its cases. This is the most important characteristic of my pipeline, since the number of datasets that have the information of both DNA methylation and gene expression values is quite low. Most of the datasets contain only gene expression values.
- The approach used in the FEM package combines gene expression and DNA methylation with a different methodology. Their method was not accurate for the AML patients dataset.

In the future, I am going to test the impact of some other clinical factors of AML patients such as sex and age on the prognostication. Also, I am going to define a statistical test to find those groups of patients that lived relatively longer than others who left the study. I will identify and add those genes whose gene expression or DNA methylation highly correlate with this cohort of patients' survival times. In addition, I will test this method on other disease datasets such as Hepatocellular Carcinoma (HCC).



(a) Km-plot for the TCGA dataset (b) Size of genes modules

Figure A.1: Survival results for TCGA dataset using $\lambda = 0$. The model is based on modules 14, 24 and 3.

APPENDIX SECTION

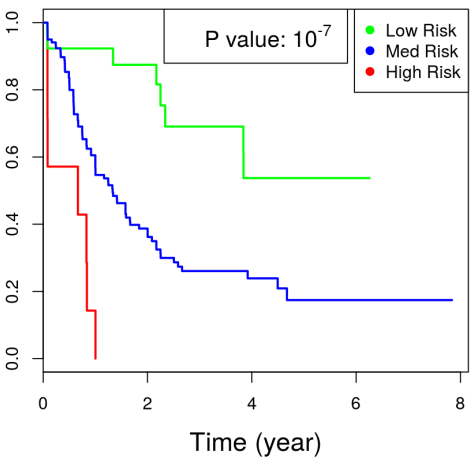
Appendix A: Optimizing the weight of each data type

To combine the gene expression values and DNA methylation levels in network construction phase, I used the following formula:

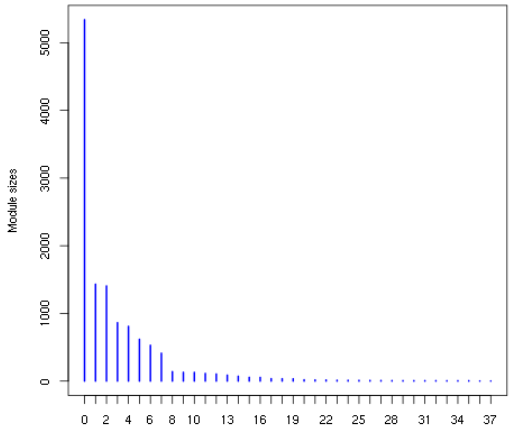
$$\mathcal{W}(g_i, g_j) = (1 - \lambda)\left|\text{cor}_E(g_i, g_j)\right| + \lambda\left|\text{cor}_M(g_i, g_j)\right| \qquad (.1)$$

I used λ as a hyper-parameter to specify the weight of each dataset. To find the best λ value, I tested the model several times with different values from 0 to 1. In each iteration, I increased the λ value by 0.1 (Figures 1 to 10).

Comparing final survival results based on different lambda values showed that the model had the best performance using $\lambda = 0.6$.

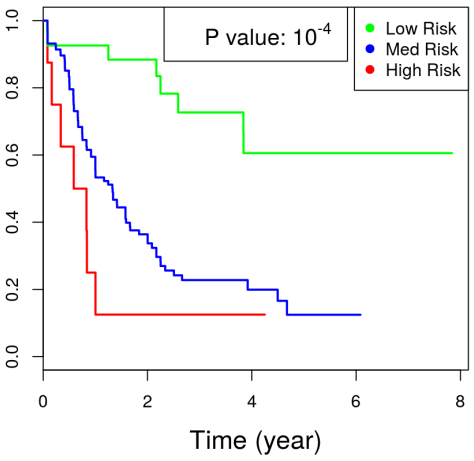


(a) Km-plot for the TCGA dataset

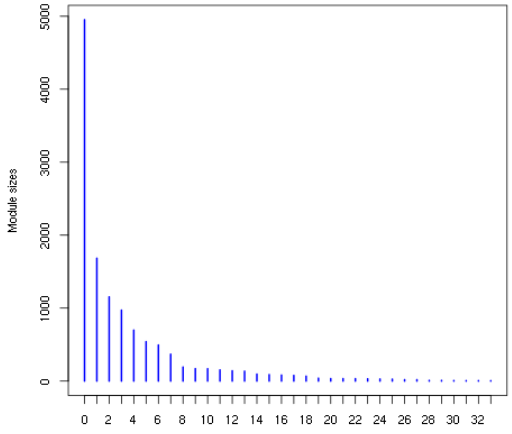


(b) Size of genes modules

Figure A.2: Survival results for TCGA dataset using $\lambda = 0.1$. The model is based on modules 14, 9 and 19.

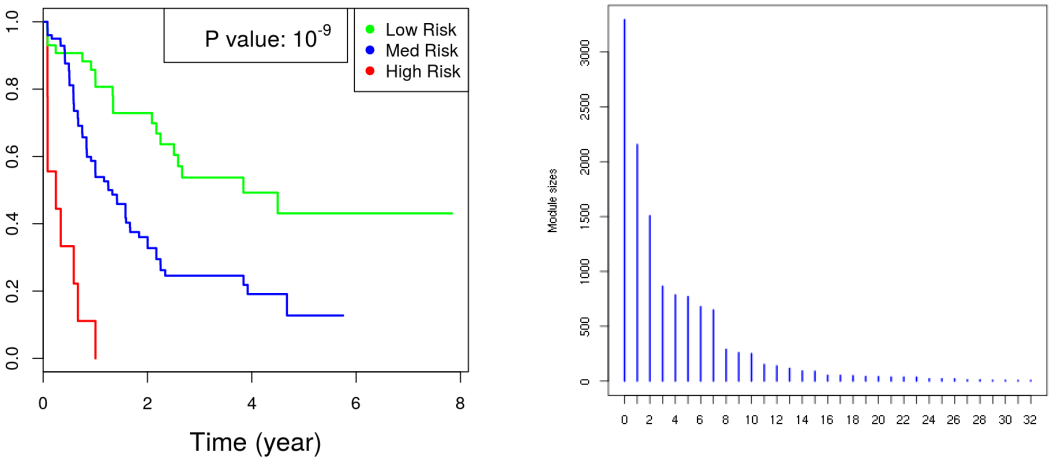


(a) Km-plot for the TCGA dataset



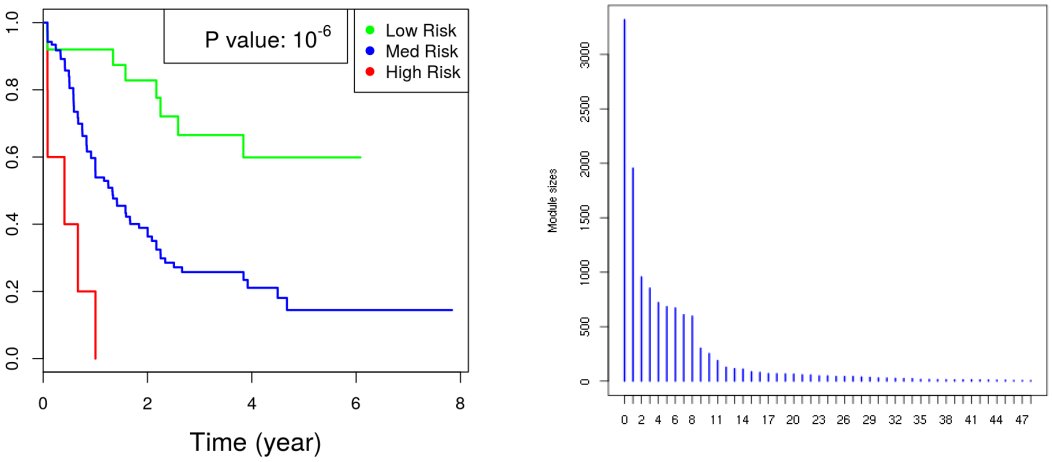
(b) Size of genes modules

Figure A.3: Survival results for TCGA dataset using $\lambda = 0.2$. The model is based on modules 9 and 27.



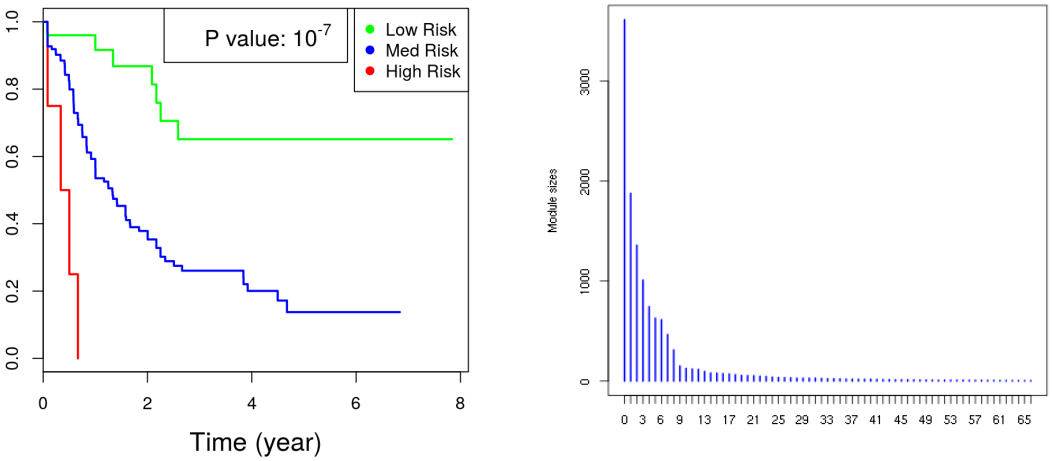
(a) Km-plot for the TCGA dataset (b) Size of genes modules

Figure A.4: Survival results for TCGA dataset using $\lambda = 0.3$. The model is based on modules 9, 22 and 13.



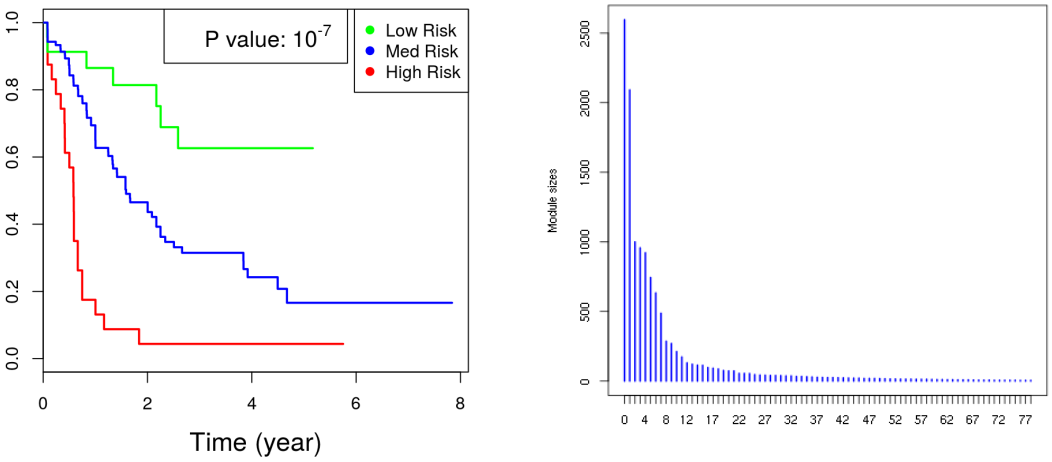
(a) Km-plot for the TCGA dataset (b) Size of genes modules

Figure A.5: Survival results for TCGA dataset using $\lambda = 0.4$. The model is based on modules 30 and 8.



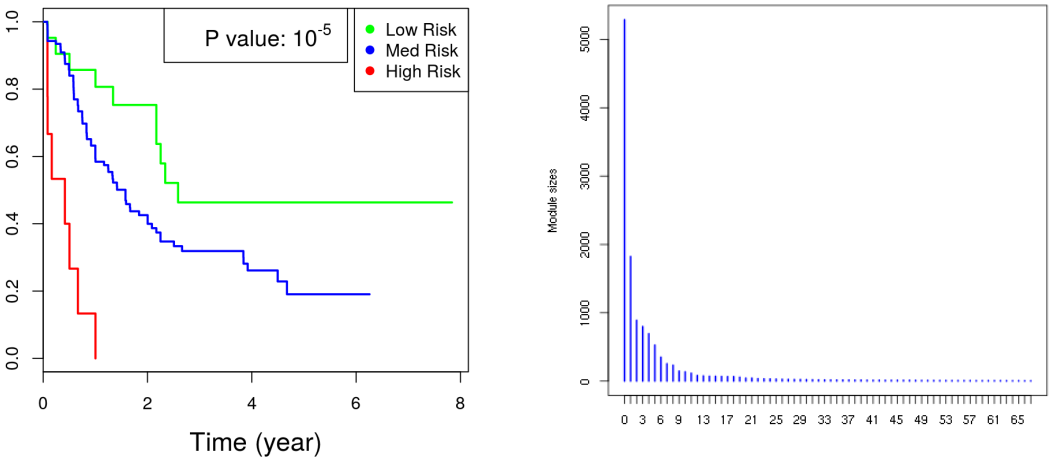
(a) Km-plot for the TCGA dataset (b) Size of genes modules

Figure A.6: Survival results for TCGA dataset using $\lambda = 0.5$. The model is based on modules 33 and 30.



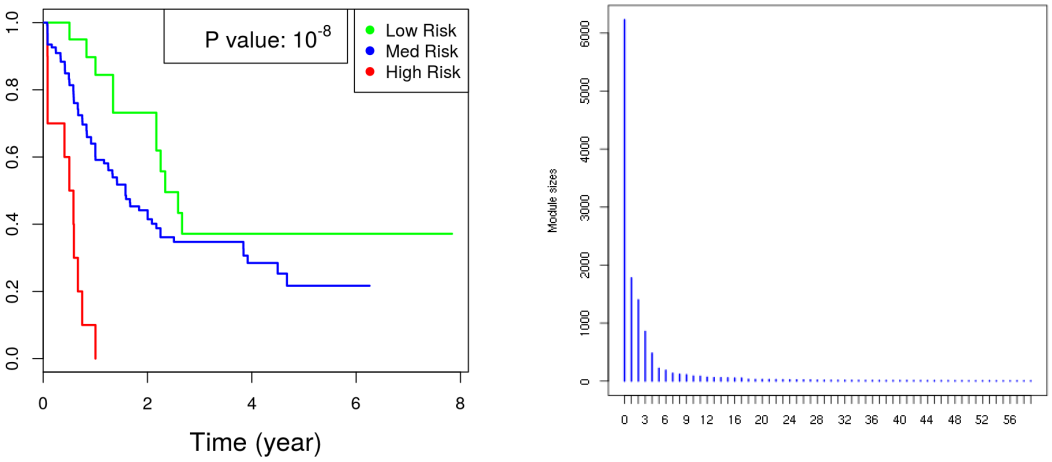
(a) Km-plot for the TCGA dataset (b) Size of genes modules

Figure A.7: Survival results for TCGA dataset using $\lambda = 06$. The model is based on modules 51 and 55.



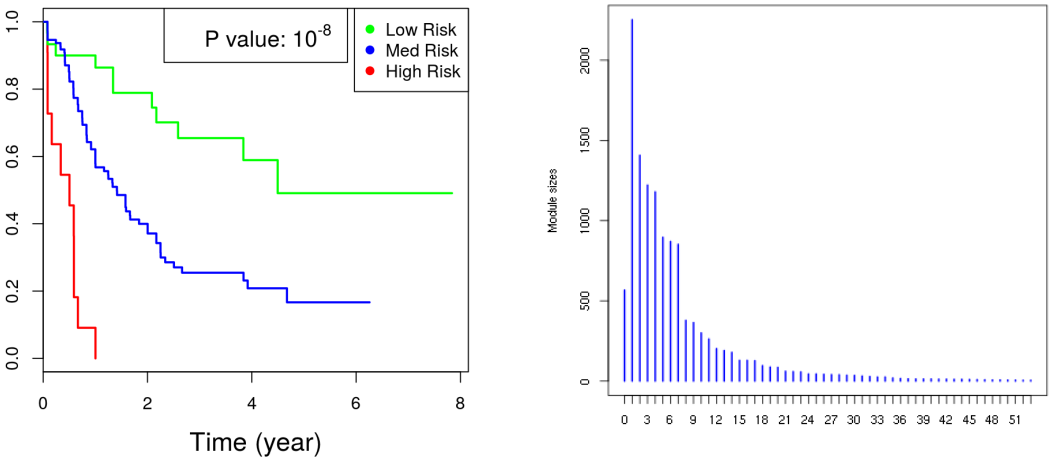
(a) Km-plot for the TCGA dataset (b) Size of genes modules

Figure A.8: Survival results for TCGA dataset using $\lambda = 0.7$ The model is based on modules 8, 33 and 31.



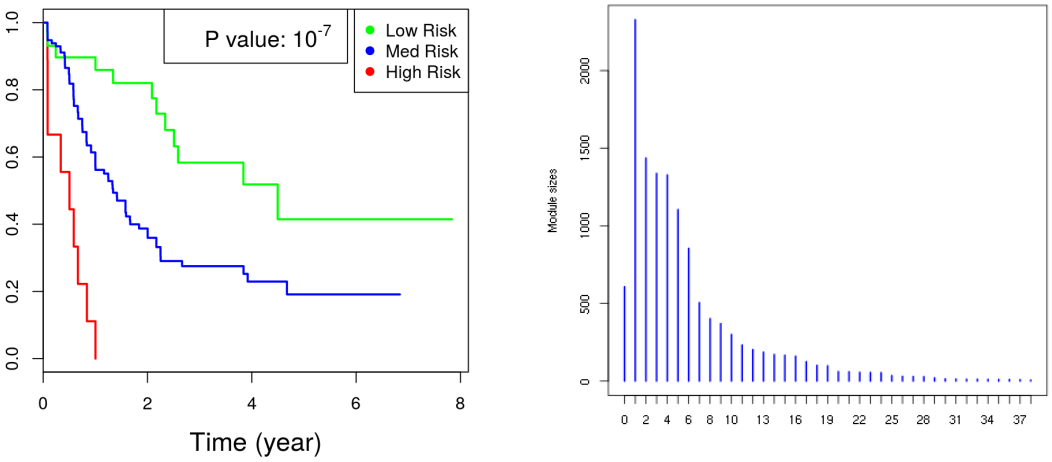
(a) Km-plot for the TCGA dataset (b) Size of genes modules

Figure A.9: Survival results for TCGA dataset using $\lambda = 0.8$. The model is based on modules 44 and 57.



(a) Km-plot for the TCGA dataset (b) Size of genes modules

Figure A.10: Survival results for TCGA dataset using $\lambda = 0.9$. The model is based on modules 28, 20 and 17.



(a) Km-plot for the TCGA dataset (b) Size of genes modules

Figure A.11: Survival results for TCGA dataset using $\lambda = 1$. The model is based on modules 24, 19 and 31.

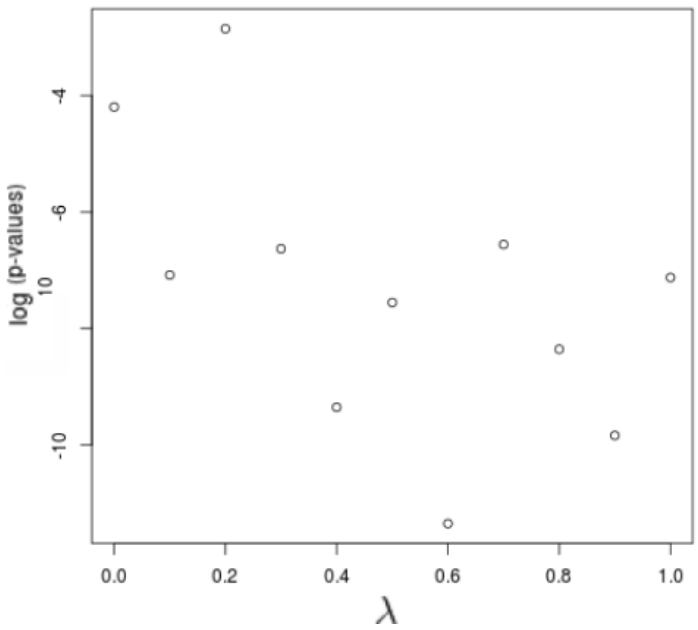


Figure A.12: Comparing the performance of the network classifier using different λ values. The X-axis shows the λ values and the Y-axis shows the log of p-value of the survival model.

REFERENCES

Abdi, H. and Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459.

Allen, S. J. and Dawbarn, D. (2006). Clinical relevance of the neurotrophins and their receptors. *Clinical Science*, 110(2):175–191.

Assenov, Y., Müller, F., Lutsik, P., Walter, J., Lengauer, T., and Bock, C. (2014). Comprehensive analysis of dna methylation data with rnbeads. *Nature methods*, 11(11):1138.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., et al. (2013). Ncbi geo: archive for functional genomics data sets–update. *Nucleic acids research*, 41(D1):D991–D995.

Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer.

Bird, A. (2007). Perceptions of epigenetics. *Nature*, 447(7143):396.

Breuer, K., Foroushani, A. K., Laird, M. R., Chen, C., Sribnaia, A., Lo, R., Winsor, G. L., Hancock, R. E., Brinkman, F. S., and Lynn, D. J. (2012). Innatedb: systems biology of innate immunity and beyond–recent updates and continuing curation. *Nucleic acids research*, page gks1147.

Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, Ö., Anwar, N., Schultz, N., Bader, G. D., and Sander, C. (2010). Pathway commons, a web resource for biological pathway data. *Nucleic acids research*, 39(suppl_1):D685–D690.

Cox, D. R. (1992). Regression models and life-tables. In *Breakthroughs in statistics*, pages 527–541. Springer.

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M. R., et al. (2014). The reactome pathway knowledgebase. *Nucleic acids research*, 42(D1):D472–D477.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5):1–9.

Daniel, W. W. (1990). Spearman rank correlation coefficient. *Applied nonparametric statistics, 2nd ed. PWS-Kent, Boston*, pages 358–365.

Davis, S. and Meltzer, P. S. (2007). Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor. *Bioinformatics*, 23(14):1846–1847.

Döhner, H., Estey, E., Grimwade, D., Amadori, S., Appelbaum, F. R., Büchner, T., Dombret, H., Ebert, B. L., Fenaux, P., Larson, R. A., et al. (2016). Diagnosis and management of aml in adults: 2017 eln recommendations from an international expert panel. *Blood*, pages blood–2016.

Döhner, H., Estey, E. H., Amadori, S., Appelbaum, F. R., Büchner, T., Burnett, A. K., Dombret, H., Fenaux, P., Grimwade, D., Larson, R. A., et al. (2010). Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the european leukemianet. *Blood*, 115(3):453–474.

Forman, S. J., Negrin, R. S., Antin, J. H., and Appelbaum, F. R. (2015). *Thomas' hematopoietic cell transplantation: stem cell transplantation*. John Wiley & Sons.

Foroushani, A., Agrahari, R., Docking, R., Chang, L., Duns, G., Hudoba, M., Karsan, A., and Zare, H. (2017). 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*, 10(1):16.

Gerstung, M., Papaemmanuil, E., Martincorena, I., Bullinger, L., Gaidzik, V. I., Paschka, P., Heuser, M., Thol, F., Bolli, N., Ganly, P., et al. (2017). Precision oncology for acute myeloid leukemia using a knowledge bank approach. *Nature genetics*, 49(3):332–340.

Gevaert, O., Villalobos, V., Sikic, B. I., and Plevritis, S. K. (2013). Identification of ovarian cancer driver genes by using module network integration of multi-omics data. *Interface focus*, 3(4):20130013.

Gui, J. and Li, H. (2005). Penalized cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, 21(13):3001–3008.

Habib, N., Kaplan, T., Margalit, H., and Friedman, N. (2008). A novel Bayesian DNA motif comparison method for clustering and retrieval. *PLoS Computational Biology*, 4(2):e1000010.

Hempstead, B. L. (2006). Dissecting the diverse actions of pro-and mature neurotrophins. *Current Alzheimer Research*, 3(1):19–24.

Herold, T., Metzeler, K. H., Vosberg, S., Hartmann, L., Röllig, C., Stölzel, F., Schneider, S., Hubmann, M., Zellmeier, E., Ksienzyk, B., et al. (2014). Isolated trisomy 13 defines a homogeneous aml subgroup with high frequency of mutations in spliceosome genes and poor prognosis. *Blood*, 124(8):1304–1311.

Horvath, S. and Dong, J. (2008). Geometric interpretation of gene coexpression network analysis. *PLoS computational biology*, 4(8):e1000117.

Jemal, A., Thomas, A., Murray, T., and Thun, M. (2002). Cancer statistics, 2002. *CA: a cancer journal for clinicians*, 52(1):23–47.

Jiao, Y., Widschwendter, M., and Teschendorff, A. E. (2014). A systems-level integrative framework for genome-wide dna methylation and gene expression data identifies differential gene expression modules under epigenetic control. *Bioinformatics*, 30(16):2360–2366.

Jones, A., Teschendorff, A. E., Li, Q., Hayward, J. D., Kannan, A., Mould, T., West, J., Zikan, M., Cibula, D., Fiegl, H., et al. (2013). Role of dna methylation and epigenetic silencing of hand2 in endometrial cancer development. *PLoS medicine*, 10(11):e1001551.

Jones, P. A. (2012). Functions of dna methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484–492.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data*, volume 360. John Wiley & Sons.

Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30.

Karam, J. A. (2009). Apoptosis in carcinogenesis and chemotherapy. *Netherlands: Springer*.

Koestler, D. C., Jones, M. J., and Kobor, M. S. (2014). The era of integrative genomics: more data or better methods? *Epigenomics*, 6(5):463–467.

Langfelder, P. and Horvath, S. (2008). Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559.

Langfelder, P. and Horvath, S. (2012). Fast r functions for robust correlations and hierarchical clustering. *Journal of statistical software*, 46(11).

Li, M., Balch, C., Montgomery, J. S., Jeong, M., Chung, J. H., Yan, P., Huang, T. H., Kim, S., and Nephew, K. P. (2009). Integrated analysis of dna methylation and gene expression reveals specific signaling pathways associated with platinum resistance in ovarian cancer. *BMC medical genomics*, 2(1):34.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies. *J natl cancer inst*, 22(4):719–748.

Network, C. G. A. R. et al. (2013). Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *The New England journal of medicine*, 368(22):2059.

Ng, S. W., Mitchell, A., Kennedy, J. A., Chen, W. C., McLeod, J., Ibrahimova, N., Arruda, A., Popescu, A., Gupta, V., Schimmer, A. D., et al. (2016). A 17-gene stemness score for rapid determination of risk in acute leukaemia. *Nature*, 540(7633):433.

Noble, W. S. (2009). A quick guide to organizing computational biology projects. *PLoS computational biology*, 5(7):e1000424.

Papaemmanuil, E., Gerstung, M., Bullinger, L., Gaidzik, V. I., Paschka, P., Roberts, N. D., Potter, N. E., Heuser, M., Thol, F., Bolli, N., et al. (2016). Genomic classification and prognosis in acute myeloid leukemia. *New England Journal of Medicine*, 374(23):2209–2221.

Parham, P. (2014). *The immune system*. Garland Science.

Paziewska, A., Dabrowska, M., Goryca, K., Antoniewicz, A., Dobruch, J., Mikula, M., Jarosz, D., Zapala, L., Borowka, A., and Ostrowski, J. (2014). Dna methylation status is more reliable than gene expression at detecting cancer in prostate biopsy. *British journal of cancer*, 111(4):781–789.

Reichardt, J. and Bornholdt, S. (2006). Statistical mechanics of community detection. *Physical Review E*, 74(1):016110.

Reichardt, L. F. (2006). Neurotrophin-regulated signalling pathways. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 361(1473):1545–1564.

Sadler, T. W. (2011). *Langman's medical embryology*. Lippincott Williams & Wilkins.

Seto, E. and Yang, X.-J. (2010). Regulation of histone deacetylase. *Handbook of cell signaling*, 3:2379.

Simon, N., Friedman, J., Hastie, T., Tibshirani, R., et al. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software*, 39(5):1–13.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1):1–25.

Team, R. C. et al. (2018). R: A language and environment for statistical computing.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling survival data: extending the Cox model*. Springer Science & Business Media.

Tibshirani, R. J. and Efron, B. (2002). Pre-validation and inference in microarrays. *Statistical Applications in Genetics and Molecular Biology*, 1(1):1–18.

Walsh, C. (2006). *Posttranslational modification of proteins: expanding nature's inventory*. Roberts and Company Publishers.

Wan, Y.-W., Allen, G. I., and Liu, Z. (2015). Tcga2stat: simple tcga data access for integrated statistical analysis in r. *Bioinformatics*, page btv677.

Welch, W. J. (1993). How cells respond to stress. *Scientific American*, 268(5):56–64.

West, J., Beck, S., Wang, X., and Teschendorff, A. E. (2013). An integrative network algorithm identifies age-associated differential methylation interactome hotspots targeting stem-cell differentiation pathways. *Scientific reports*, 3:1630.

Yip, A. M. and Horvath, S. (2007). Gene network interconnectedness and the generalized topological overlap measure. *BMC bioinformatics*, 8(1):22.

Zainulabadeen, A., Yao, P., and Zare, H. (2017). Underexpression of specific interferon genes is associated with poor prognosis of melanoma. *PloS one*, 12(1):e0170025.

Zare, H. et al. (2016). Pigengene: Computing and using eigengenes.

Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1).

Zhuang, J., Widschwendter, M., and Teschendorff, A. E. (2012). A comparison of feature selection and classification methods in dna methylation studies using the illumina infinium platform. *BMC bioinformatics*, 13(1):59.