공학박사 학위논문

Computational analysis of biological pathways in breast cancer subtypes: a visual exploration system and a probabilistic framework

유방암 아형의 생물학적 경로 : 시각적 탐색 시스템과 확률적 프레임워크

2020 년 8 월

서울대학교 대학원

컴퓨터 공학부

김 인 영

Computational analysis of biological pathways in breast cancer subtypes: a visual exploration system and a probabilistic framework

유방암 아형의 생물학적 경로 : 시각적 탐색 시스템과 확률적 프레임워크

지도교수 김 선

이 논문을 공학박사 학위논문으로 제출함

2020 년 7 월

서울대학교 대학원

컴퓨터 공학부

김 인 영

김인영의 공학박사 학위논문을 인준함

2020 년 8 월

위 원 장	박근수
부위원장	김 선
위 원	장병탁
위 원	채희준
위 원	안홍렬

Abstract

Computational analysis of biological pathways in breast cancer subtypes: a visual exploration system and a probabilistic framework

Inyoung Kim
Department of Computer Science & Engineering
College of Engineering
Seoul National University

Breast cancer is cancer that develops from breast tissue and it is the leading type of cancer in women, accounting for 25% of all cases (Bray *et al.*, 2018). Although breast cancer has been studied extensively for several decades, there are a number of issues that remain to be answered. With the rapid development of instrument technologies, a huge amount of molecular data from breast cancer has been produced. Molecular data measured inside breast cancer cells can be very useful to investigate many unresolved issues in breast cancer. However, analysis of molecular data such as genetic mutations and gene transcripts

is very difficult since the number of dimensions is huge (over 20,000 up to several millions) and the number of samples or patients is only a few thousands. This is one of the unresolved machine learning problems, that is, analysis of high dimension low sample data. Thus, new computational methods are much needed to study breast cancer at the molecular level.

In this thesis, I addressed this computational challenge by utilizing biological pathways that can be used to explain cancer mechanisms in terms of biological functions, such as cell growth, cell death, and metastatic potentials.

In my doctoral study, I developed two computational methods. A web-based system was developed for exploring pathways in terms of genetic mutations, gene copy number variations, and gene expression levels. A probabilistic framework was developed for determining driver genes from genes in biological pathways.

The first study was to integrate TCGA breast cancer data onto the KEGG pathway to visualize the multi-omics data of breast cancer patients. Pathway based multi-omics analysis system is necessary but challenging due to larger sample sizes and higher dimension. BRCA-Pathway, a web-based interactive exploration and visualization system of TCGA breast cancer data on KEGG pathway, was developed to address these difficulties and provide broad perspective of TCGA breast cancer data. Through the first study, it was confirmed that the multi-omics data of breast cancer patients appeared differently for each subtype from the perspective of the pathway. In particular, gene expression data could identify different expression patterns for each subtype in several biologically important pathways.

The second study was to solve the problem of selecting genes specific to the subtype by using different expression patterns from the viewpoint of the KEGG pathway for each breast cancer subtype. The difference in the gene expression pattern at the pathway level was represented to a numerical value

of the degree of activation in the pathway. The difference in gene expression level for each subtype was quantified and defined as a Gene factor, and the difference in the degree of pathway activation for each subtype was defined as a Pathway factor. Likelihood of gene given subtype and posterior probability of subtype given gene were defined using Gene factor and Pathway factor. Then, genes were ranked by likelihood and posterior probability. It can be seen that the problem of selecting subtype specific gene corresponds to feature selection in the subtype classification model. For this reason, we evaluated the performance of the predictive model with selected genes as features of the classification problem. We also analyzed the biological implications of the selected genes.

In summary, my doctoral thesis proposed how biological pathways that are important domain knowledge can be used to characterize breast cancer subtypes by visually exploring molecular data and by selecting genes in a probabilistic framework to show difference in pathway activation among breast cancer subtypes.

Keywords: High dimensional data, Biological pathways, Gene expression, Machine learning, Bayesian approach

Student Number: 2014-30325

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Breast cancer refers to cancer that develops from breast tissue, but each breast subtype has heterogeneous characteristics. Therefore the prognosis for each subtype is also different. Although breast cancer has been studied extensively for several decades, there are still a number of issues that remain to be answered.

With the rapid development of instrument technologies, a huge amount of molecular data from breast cancer has been produced. Molecular data measured inside breast cancer cells can be very useful to investigate many unresolved issues in breast cancer.

However, analysis of molecular data such as genetic mutations and gene transcripts is very difficult since the number of dimensions is huge (over 20,000 up to several millions) and the number of samples or patients is only a few thousands. This is one of the unresolved machine learning problems, that is, analysis of high dimension low sample data. Thus, new computational methods are much needed to study breast cancer at the molecular level.

In this thesis, I addressed this computational challenge by utilizing bio-

logical pathways that can be used to explain cancer mechanisms in terms of biological functions, such as cell growth, cell death, and metastatic potentials. In my doctoral study, I developed two computational methods. A web-based system was developed for exploring pathways in terms of genetic mutations, gene copy number variations, and gene expression levels. A probabilistic framework was developed for determining driver genes from gene expression data in biological pathways.

Chapter 2 introduces previous studies related to this thesis. It includes description of the pathway-based omics analysis system and gene expression signatures for breast cancer prognosis.

Chapter 3 describes a pathway-based biological information retrieval system, BRCA-Pathway: a structural integration of TCGA breast cancer data and an interactive visualization on KEGG pathway. We developed BRCA-Pathway to interpret breast cancer data at the pathway level, a biological functional unit. We could see how gene expression differences between breast cancer subtypes differ at the pathway level by BRCA-Pathway.

Chapter 4 describes a probabilistic framework that discovers subtype specific driver gene sets, guided by pathway information. The visual difference at the pathway level in BRCA-Pathway was transformed to pathway activation score and this value was introduced into the problem of selecting genes that distinguish subtypes. In the probabilistic framework, the likelihood and posterior probability of each gene were calculated to select genes specific to the subtype.

Chapter 5 summarizes my doctoral study and discusses how this study contributed to the gene selection problem specific to the breast cancer subtypes.

Chapter 2

Related works

2.1 Driver gene and passenger gene

Through next-generation sequencing, thousands of mutations have been reported in large cohorts (TCGA, ICGC). However, a fraction of millions of mutations that are observed in cancer are directly related to cancer developmental process, and a majority of mutations are passenger mutations that are either simply difference among individuals or consequences of cancer development (Vogelstein *et al.*, 2013). In many cancer studies, large efforts have been made to determine driver genes involved in the cancer process, especially in mutation studies (Kris *et al.*, 2011) and in epigenetic studies (Kalari and Pfeifer, 2010).

In my doctoral study, I am interested in classifying subtypes of breast cancer patients using gene expression data. To provide a solution to the problem of selecting key genes that are essential in each of breast cancer subtypes based on gene expression level information.

The concept of the driver gene and the passenger gene is essential in my

doctoral thesis. However, a definition of the driver gene and the passenger gene requires a very large cohort study over a long period of time. Thus, the concept is yet to be precisely defined in the cancer science community.

In my thesis, I define as follows. Among many genes that differ in the expression level, a key gene that has a pivotal role in distinguishing breast cancer subtypes is a driver gene. In my doctoral study, this concept is implemented in a probabilistic framework that utilizes biological pathway information to select genes that are specific to subtypes. Thus, in my thesis, *a driver gene* is considered as *a subtype specific gene* that is ranked high by the probabilistic framework. All other genes that are not driver genes are considered as passenger genes.

From a computer science perspective, this is a problem of distinguishing between true positives and false positives. It is difficult to solve the problem when the number of patient samples are much smaller than the size of dimensions, that is genes.

To address this computationally infeasible problem, I leveraged biological pathway information that is curated manually based on the scientific literature. In the probabilistic framework that I designed and implemented, genes are ranked in terms of likelihood and posterior probabilities that are weighted by Pathway factors. In other words, selection of genes are guided by pathway information to discover subtype specific driver gene sets.

2.2 Breast cancer

Breast cancer refers to cancer that develops from breast tissue, but each breast subtype has heterogeneous characteristics. Therefore the prognosis for each subtype is also different. Although breast cancer has been studied extensively for several decades, there are still a number of issues that remain to be answered. Worldwide, there are about 2.1 million newly diagnosed female breast cancer cases every year, accounting for almost 1 in 4 cancer cases among women (Bray *et al.*, 2018).

2.2.1 Breast cancer subtypes

The studies conducted by Sørlie et al. reported a distinctive molecular portrait of breast cancer, according to which tumors were classified into five intrinsic subtypes with distinct clinical outcomes, i.e., Luminal A, Luminal B, HER2 overexpression, Basal and Normal-like tumors (Perou *et al.*, 2000; Sørlie *et al.*, 2001).

Patients diagnosed with breast cancer are internally divided into five subtypes according to three molecular characteristics: estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2). Each of the five intrinsic subtypes is mapped to an Immunohistochemistry (IHC)-defined subtype in Table 2.1 (Dai *et al.*, 2015). IHC markers including ER, PR and HER2 are classically used for breast tumor subtyping (Dai *et al.*, 2016).

When breast cancer cells have a significant amount of ER, they are called ER+. In other words, the ER positive type means that cancer cells grow in response to the estrogen hormone. Interestingly, the patient's prognosis varies depending on the subtype. Among them, the triple negative type, called Basal subtype, is more aggressive than other subtypes, and is likely to have already

Table 2.1: Breast cancer molecular subtypes. From "Breast cancer intrinsic subtype classification, clinical use and future trends," by X. Dai et al, 2015, *American journal of cancer research*, 5.10:2929.

Intrinsic subtype	IHC status	Outcome	Prevalence
Luminal A	[ER+|PR+]HER2-KI67-	Good	23.7%
Luminal B	[ER+|PR+]HER2-KI67+	Intermediate	38.8%
	[ER+|PR+]HER2+KI67+	|Poor	14%
HER2	[ER-|PR-]HER2+	Poor	11.2%
Basal	[ER-|PR-]HER2-	Poor	12.3%
Normal-like	[ER+|PR+]HER2-KI67-	Intermediate	7.8%

metastasized to other parts at the time of discovery. The likelihood of recurrence is high, and the likelihood of dying within 5 years is high. Figure 2.1 specifies molecular subtypes of breast cancer and their relative prognosis outcome.

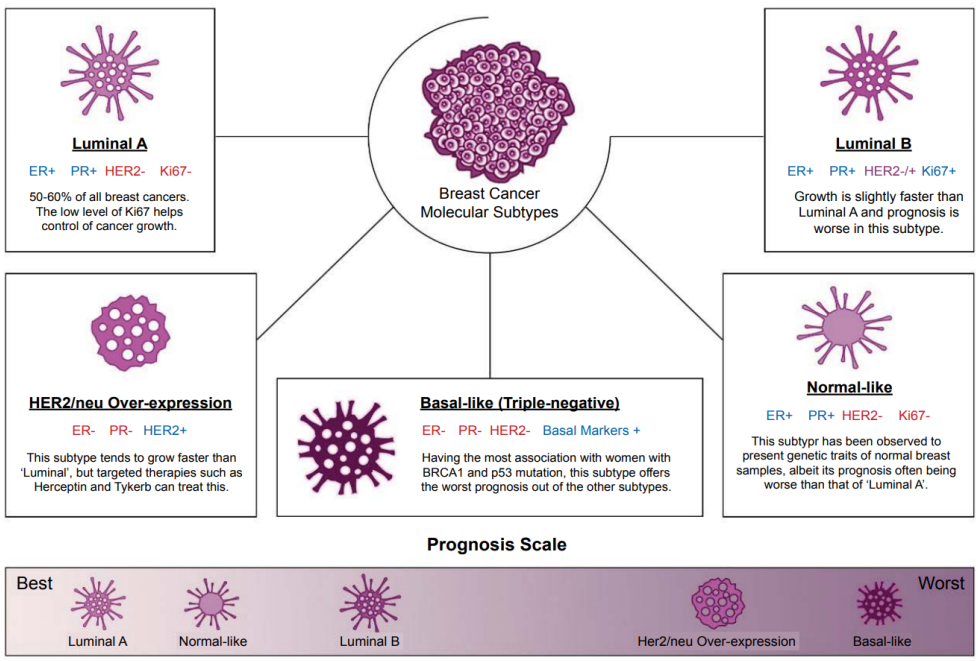


Figure 2.1: Molecular subtypes of breast cancer and their relative prognosis outcome. Patients diagnosed with breast cancer are internally divided into five subtypes according to three molecular characteristics: estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2). Each of the five intrinsic subtypes is mapped to an Immunohistochemistry (IHC)-defined subtype. From "Organoids as reliable breast cancer study models: an update," by A. Sasmita and Y. Wong, 2018, *Int J Oncol Res*, 1.008.

2.3 KEGG pathway

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from genomic and molecular-level information (Kanehisa and Goto, 2000). It is a computer representation of the biological system, and KEGG pathway is a a collection of manually drawn pathway maps representing knowledge on the molecular interaction, reaction and relation networks.

Figure 2.2 shows Cell cycle pathway. The cell cycle is the series of events that takes place in a cell leading to its division and duplication (Schwartz and Shah, 2005). Regulation of the cell cycle involves processes crucial to the survival of a cell, including the detection and repair of genetic damage as well as the prevention of uncontrolled cell division (Maddika *et al.*, 2007). Like Cell cycle pathway, each map of KEGG pathway is a collection of biologically closely related genes and can be considered as a unit that performs biological functions.

2.3.1 Pathway-based analysis systems

A number of systems have been developed to utilize KEGG pathway. Pathview (Luo and Brouwer, 2013) is an R/Bioconductor package for pathway-based data integration and visualization. Pathview automatically downloads the pathway graph, parse, maps and integrates user data onto the pathway and renders pathway graphs with the mapped data. Pathview is useful and easy to use. But, installation and preparation of data files are remained to users. Pathway Inspector (Bianco *et al.*, 2017) is a web application for finding patterns of gene expression in complex RNAseq experiments. Pathway Inspector combines identification of differentially expressed genes (DEGs) and a topology-based

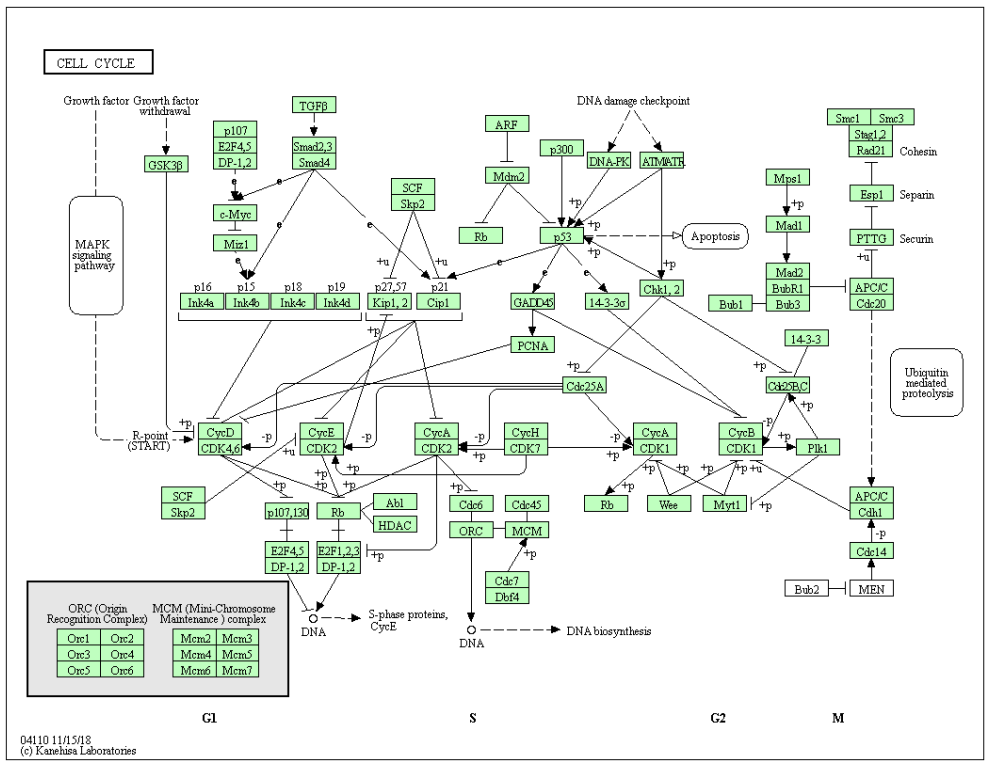


Figure 2.2: Example of KEGG pathway for representing gene interactions. KEGG pathway is a collection of manually drawn diagrams called the KEGG reference pathway diagrams (maps), each corresponding to a known network of functional significance.

analysis of enriched pathways. However, existing pathway-based analysis tools are not powerful enough to investigate complex diseases such as cancer with multi-omics data. Therefore, we need more powerful and flexible information system for integrating and analyzing different types of multi-omics data in cancer.

Many tools have been developed for TCGA data analysis. TCGA2STAT (Wan *et al.*, 2016) is one of the tools, but these tools require knowledge of the detailed specifications of TCGA data and programming skills. Meanwhile,

there are web-based multi-omics data analysis services such as cBioPortal (Gao *et al.*, 2013), OASIS (Fernandez-Banet *et al.*, 2016), and NetGestalt (Zhu *et al.*, 2014). This frameworks provide easy access and interpretation of multi-omics data. OASIS provides multi-omics data such as mutations, CNV and gene expression in selected cancer types and selected oncogenic pathways. But, there is no information on the relationship between genes because multi-omics data is displayed using a table. NetGestalt adopts visualization at the horizontal dimension and extends it to a large network, providing multi-omics data from a network perspective. In addition, by zooming into specific gene, multi-omics data of each gene is displayed and protein-protein interactions around the selected gene are provided. KeyPathwayMinerWeb (Pandey *et al.*, 2004) provides an online multi-omics network. It receives gene expression data and active gene list from users and provides the maximum connected sub-network using protein-protein interaction (PPI). This system displays the relationship of genes in the network, but does not show multi-omics data in the network because it uses only the gene expression data.

2.4 Gene expression signatures for breast cancer prognosis

Breast cancer is a complex disease encompassing multiple tumor entities, each characterized by distinct morphology, behavior and clinical implications (Dai *et al.*, 2016). According to the study by Sørlie et al., breast cancer is classified into five intrinsic subtypes with distinct clinical outcomes: Luminal A, Luminal B, HER2 overexpression, Basal and Normal-like tumors.

These five intrinsic subtypes have been repeated by other studies with varying numbers of genes included in the signature. For example, Hu et al. found a signature containing 306 genes that can distinguish these subtypes with significant differences observed on relapse-free and overall survival (Hu *et al.*, 2006). Parker et al. reported a 50-gene classifier (PAM50, which contains mostly hormone receptor and proliferation related genes, and genes exhibiting myoepithelial and basal features), which has significant prognostic and predictive values on breast tumors (Parker *et al.*, 2009).

However, these previous researches are not sufficient for functional interpretation that can be linked to therapeutic options. To overcome this shortcoming, Gatza et al. classified breast cancer patients making use of patterns of pathway activity based on intrinsic gene expression signatures (Gatza *et al.*, 2010). Gene expression data was normalized using Bayesian Factor Regression Modeling (Carvalho *et al.*, 2008) and analyzed by unsupervised hierarchical clustering to reveal complex patterns of gene expression. According to the clustering results, 18 pathways were defined and validated with independent biochemical or genetic analyses.

Chapter 3

Pathway-based biological information retrieval system

3.1 Motivation

3.1.1 Biological pathway analysis

Use of multi-omics data that are observed inside cancer cells can be very useful to reveal biological mechanisms in cancer. Transcriptome data measured at the whole genome level require a higher level of interpretation rather than simply measuring the expression level of each gene in order to grasp its biological meaning. Single nucleotide variation (SNV) and copy number variation (CNV), as well as mutations, are often measured for cancer research. Different types of data as such are often termed as *multi-omics data*.

Direct interpretation of multi-omics data is very difficult since the number of genes are large, $20,000$, and number of mutations observed in cancer cells is huge, up to several millions. Thus, multi-omics data are often mapped to KEGG pathway since pathways are much easier for interpretation.

A number of systems have been developed to utilize KEGG pathway.

Pathview (Luo and Brouwer, 2013) is an R/Bioconductor package for pathway-based data integration and visualization. Pathview automatically downloads the pathway graph, parse, maps and integrates user data onto the pathway and renders pathway graphs with the mapped data. Pathview is useful and easy to use. But, installation and preparation of data files are remained to users. Pathway Inspector (Bianco *et al.*, 2017) is a web application for finding patterns of gene expression in complex RNAseq experiments. Pathway Inspector combines identification of differentially expressed genes (DEGs) and a topology-based analysis of enriched pathways. However, existing pathway-based analysis tools are not powerful enough to investigate complex diseases such as cancer with multi-omics data. Therefore, we need more powerful and flexible information system for integrating and analyzing different types of multi-omics data in cancer.

3.1.2 Pathway based analysis tools

Many tools have been developed for TCGA data analysis. TCGA2STAT (Wan *et al.*, 2016) is one of the tools, but these tools require knowledge of the detailed specifications of TCGA data and programming skills.

Meanwhile, there are web-based multi-omics data analysis services such as cBioPortal (Gao *et al.*, 2013), OASIS (Fernandez-Banet *et al.*, 2016), and NetGestalt (Zhu *et al.*, 2014). This frameworks provide easy access and interpretation of multi-omics data.

OASIS provides multi-omics data such as mutations, CNV and gene expression in selected cancer types and selected oncogenic pathways. However, there is no information on the relationship between genes because multi-omics data is displayed using a table.

NetGestalt adopts visualization at the horizontal dimension and extends it to a large network, providing multi-omics data from a network perspective.

In addition, by zooming into specific gene, multi-omics data of each gene is displayed and protein-protein interactions around the selected gene are provided.

KeyPathwayMinerWeb (Pandey *et al.*, 2004) provides an online multi-omics network. It receives gene expression data and active gene list from users and provides the maximum connected sub-network using protein-protein interaction (PPI). This system displays the relationship of genes in the network, but does not show multi-omics data in the network because it uses only the gene expression data.

Our system, BRCA-Pathway (Kim *et al.*, 2018), is designed to simultaneously represent the relationship between genes in the KEGG pathway and the corresponding multi-omics data. In summary, a pathway-based multi-omics analysis system is required, but it is difficult due to the larger sample size and high-dimensional problems of multi-omics. For this reason, we developed BRCA-Pathway, a web-based interactive exploration and visualization system of TCGA breast cancer data on the KEGG pathway to provide a broad perspective of TCGA breast cancer data.

3.2 Methods

3.2.1 System design of BRCA-Pathway

BRCA-Pathway consists of three parts - Database system, Representational State Transfer Application Programming Interface (REST API), and Web front-end. As a first technical base, Database system which structured omics data such as gene expression RNASeq normalized counts, mutation, Copy Number Variation (CNV) data, clinical data, and KEGG pathway data. And its management system is MySQL.

Second part of BRCA-Pathway is REST API. This interface given by Django web framework provides structured omics data from database in response to client's requests.

The last part of BRCA-Pathway is front-end. This part calls omics data with ajax, one of the useful asynchronous communication method while dealing with various repositories of biology databases (Aravindhan *et al.*, 2009). Once the web browser which is in side of clients requests data through URL, REST API would send data after throwing a query to database system and returned query result (queryset) by database virtual view. Clients could approach those data by not only BRCA-Pathway web service, but also web browser or terminal.

As shown in Figure 3.1, the system design of BRCA-Pathway consists of a web server on the right and a web front end that the user encounters on the left. The server consists of a database system, REST API.

The left part is client side, and the right part is BRCA-Pathway server side. BRCA-Pathway server builds the database which contains structured TCGA multi-omics data and KEGG pathway data. Database is abstracted by virtual view which simplifies table join functions. REST API could get the data set when it throws query to database. These data set will be provided by

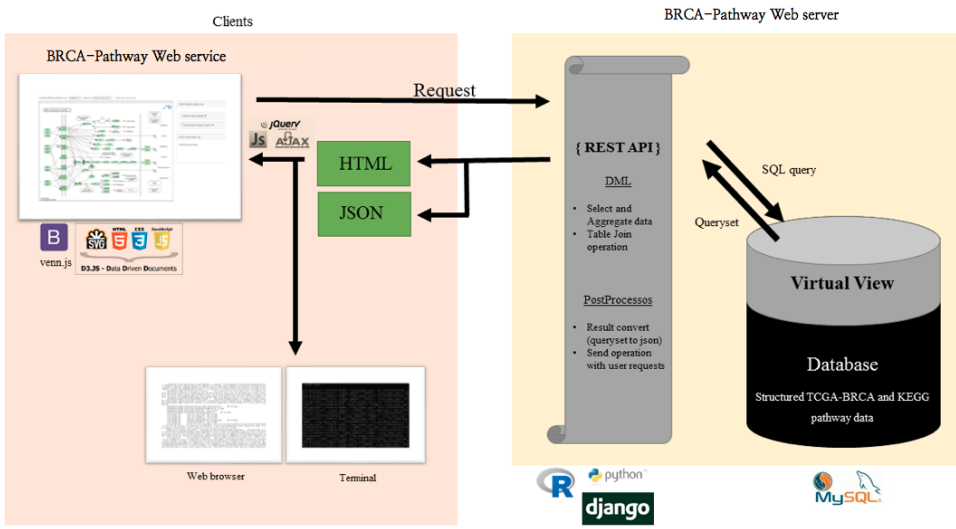


Figure 3.1: BRCA-Pathway system structure overview. The left part is client side, and the right part is BRCA-Pathway server side. BRCA-Pathway server builds the database which contains structured TCGA multi-omics data and KEGG pathway data. Database is abstracted by virtual view which simplifies table join functions. REST API could get the data set when it throws query to database. These data set will be provided by Django framework so that the clients can access the data set by web browser or terminal. BRCA-Pathway web front-end receives the data set and visualizes for clients.

Django framework so that the clients can access the data set by web browser or terminal. BRCA-Pathway web front-end receives the data set and visualizes for clients.

3.2.2 Data resources and database architecture

BRCA-Pathway integrates multiple resources in a relational database and provides web-based interactive interface and REST API. A database system using MySQL is designed for the structured integration of multi-omics data of TCGA-BRCA, KEGG pathway data, Hallmark gene sets, transcription factors, driver genes, and PAM50 subtype.

BRCA-Pathway system can update KEGG Pathway and TCGA-BRCA data by initiating the update software module to incorporate the most recent information. The current configuration is based on KEGG Pathway released on October 1, 2016 and the latest version of GDAC Firehose on January 28, 2016.

Table 3.1 shows the data resources used in the BRCA-Pathway system. It uses TCGA data, KEGG pathway data, transcription factors, characteristic gene sets, and driver gene information. In the TCGA data, the patient's clinical information, gene expression data, mutation data and CNV data were used. TCGA data is taken from the latest version of the BROAD Institute GDAC firehose. Much of the data provided by the BRCA-Pathway is from TCGA, a cancer genomics program that began in 2006 in collaboration with the National Cancer Institute (NCI) and the National Human Genome Institute (NHGRI). It started with the purpose of integrating and accumulating mutation data and analyzing biological information. The relational database schema of BRCA-Pathway is shown in Figure 3.2.

Table 3.1: Description of data provided by BRCA-Pathway

Data type	Data size	Data source
Clinical	22 attributes 1,098 patients 1,098 rows	Standardized analysis-ready TCGA data, Broad Institute TCGA Genome Data Analysis Center (2016): Firehose stddata__2016_01_28 run Broad Institute of MIT and Harvard doi:10.7908/C11G0KM9
Gene Expression	1,093 patients 20,531 genes 22,440,383 rows	
Mutation	977 patients 17,280 genes 86,765 rows	
CNV	1,080 patients 24,776 genes 26,758,080 rows	SNP6 Copy number analysis (GISTIC2) Broad Institute of MIT and Harvard doi:10.7908/C1NP23RQ
KEGG Pathway	307 pathways	Kyoto Encyclopedia of Genes and Genomes
Transcription Factors	85,314 TF-TG pairs	Human Transcriptional Regulation Interaction Database (HTRIdb) Molecular Signatures Database (MSigDB)
Hallmark gene sets	50 Hallmark gene sets	Molecular Signatures Database (MSigDB)
Driver genes	486 driver genes	Cancer Gene Census in COSMIC database Vogelstein, Bert, et al. “Cancer genome landscapes.” science 339.6127 (2013): 1546-1558. Table S2A

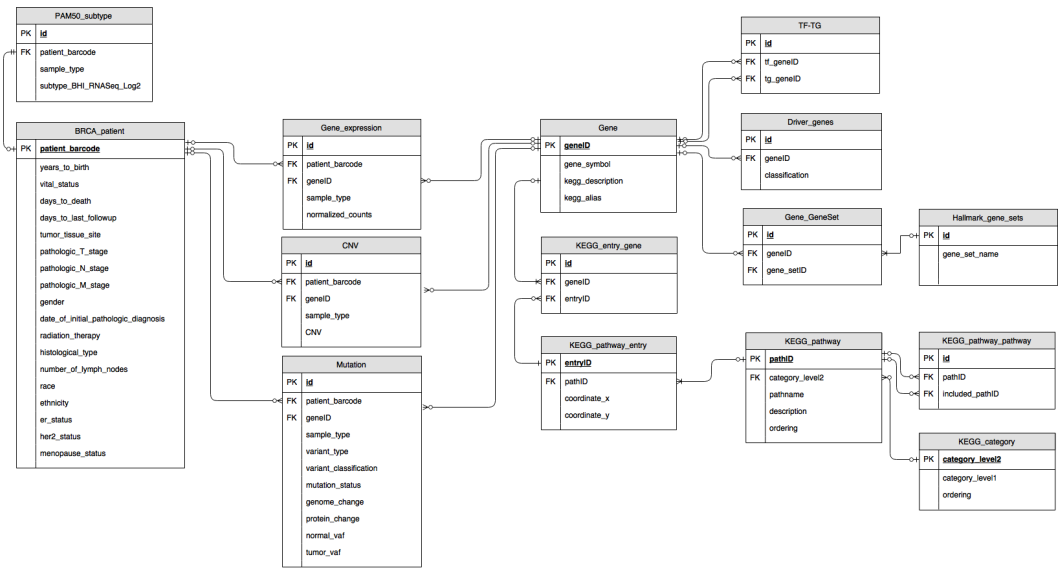


Figure 3.2: Entity Relationship Diagram of BRCA-Pathway : ERD shows the relationship between tables. Patient table contains clinical information of the individual breast cancer patient. PAM50 subtype table provides the subtype of the patient by PAM50 method. Three tables are for TCGA omics data, Gene expression, Copy Number Variation, and Mutation. TF-TG table provides the regulation information between genes, Transcription Factor and Target gene. Driver gene table contains the important gene list in cancer progression. Hallmark gene sets summarize and represent specific well-defined biological states or processes and display coherent expression. There are 5 tables to represent the KEGG pathway information.

3.2.3 Workflow of BRCA-Pathway

Figure 3.3 shows an example workflow of BRCA-Pathway. To map the data of breast cancer patients to the KEGG pathway, select the pathway and the patients to be mapped.

First, select the pathway of interest. If user doesn't know which pathway to choose, user can choose one of the several options below. User can choose from the hallmark gene set, a set of genes that are important in cancer, enter the gene symbol of interest, or choose from genes that show differences in expression between breast cancer subtypes. Then, the pathways containing the selected gene are shown as a list, and the user can select a pathway from the list. Next, we select the patient population data to map to the pathway. Clinical information and PAM50 subtypes allow user to select a subpopulation of breast cancer patients in any combination.

In this example, we will show the procedure of analyzing patients of Basal subtype starting with genes (DEG) expressed differently than normal tissue samples. Calculation of Wilcoxon rank-sum test reveals DEGs from Basal subtype patient group and lists the KEGG pathway including these DEGs. In addition, when the user selects a pathway from among them, multi-omics data mapped to the selected pathway is visualized. In other words, gene expression, mutation information, and CNV information are mapped and visualized on the pathway.

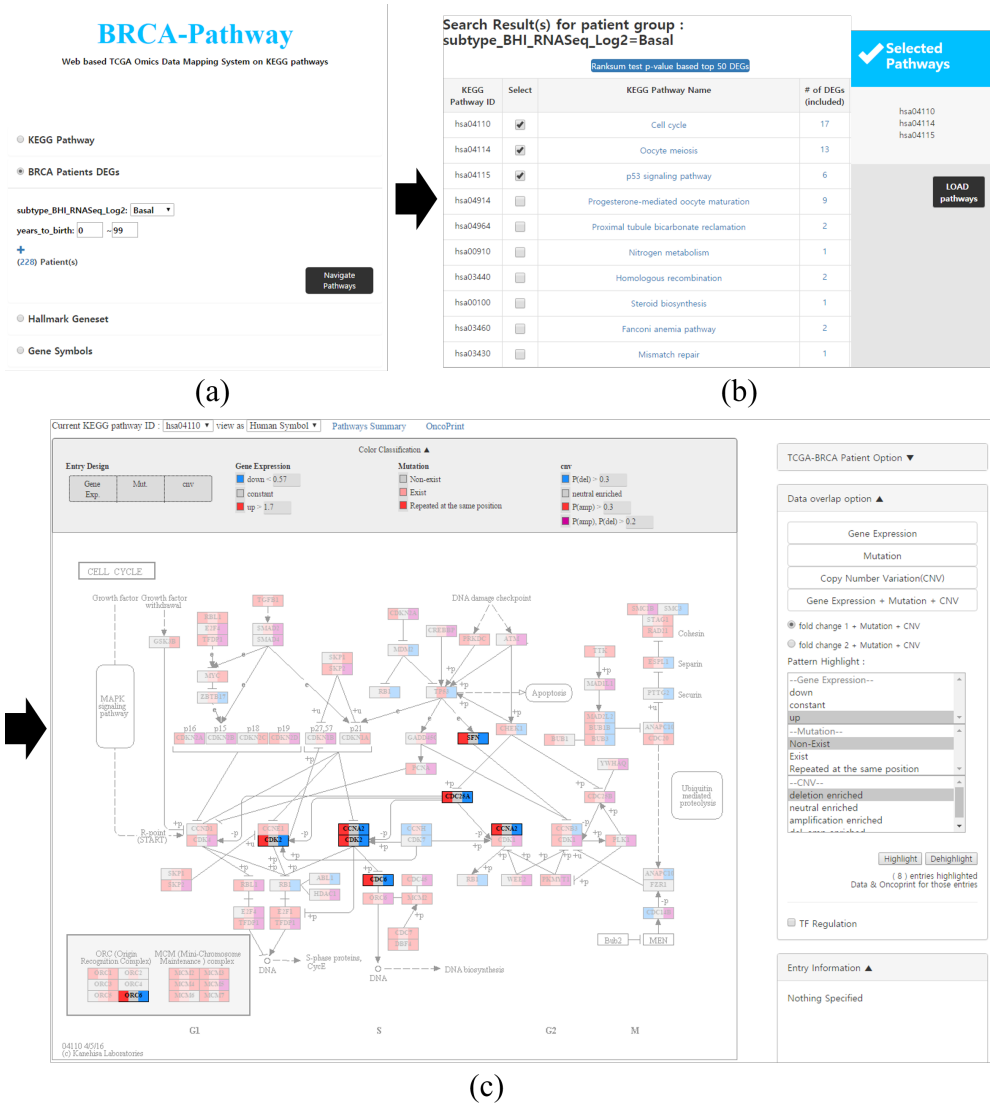


Figure 3.3: Workflow of BRCA-Pathway : (a) Selection of subpopulation by subtype and clinical features. In this example, patients with Basal subtype are selected and (b) pathways including differential expressed genes in Basal subtype are listed as the result of computation of Wilcoxon rank-sum test. (c) Visualization of multi-omics data from patients with Basal subtype. Genes with specific pattern are highlighted by selecting the condition.

There are several ways to select KEGG pathway. In this example, pathway selection by DEG is described. Figure 3.3(a) shows the selection of patients to map on pathway by breast cancer subtype and clinical features. If patients with Basal subtype are selected, then pathways including differentially expressed genes in Basal subtype are listed as the result of computation of Wilcoxon rank-sum test as shown in Figure 3.3(b). Select pathways from the list then BRCA-Pathway visualizes multi-omics data of patients whose subtype is Basal. Users can view the pathway by three different data types. Gene expression, mutation, and CNV data are visualized in separate way or integrated way. Genes having specific omics-pattern are highlighted by selecting the omics condition. For example, genes having no mutation and of which CNV is Loss (-1) or Del (-2) but overexpressed are selected for highlighting then the pathway turns shaded except the highlighted genes as in Figure 3.3(c).

3.3 Results

The major features of BRCA-Pathway:

- Multi-omics data such as SNV, CNV, and gene expression can be visualized simultaneously on KEGG pathway maps, together with TF-TG correlation and relationships among cancer driver genes.
- Users can perform comparative analysis of BRCA data, including selection of differentially expressed genes (DEGs) in arbitrary patient groups, mutual exclusivity module (MEMo) summary of genomic alterations (SNV and CNV).
- Data can be downloaded by REST API.

3.3.1 Pathway-based exploration of TCGA BRCA

BRCA-Pathway can visualize gene expression level, mutation, and CNV data, respectively, or combine three omics and show them in an integrated view. On the left of Figure 3.4, we can see the pathways showing gene expression, mutation, and CNV in Cell cycle pathway, respectively. On the right, it is a screen that shows three omics combined. One square can be regarded as one gene, and each square is divided into three parts. The first part indicates the gene expression level. Red color means overexpression of the gene. The middle part represents the frequency of mutation. Gray color means there are no mutations. The last part represents the CNV, blue color means copy number deletion. And by highlight function, only genes that satisfy the conditions set by the user are highlighted and displayed.

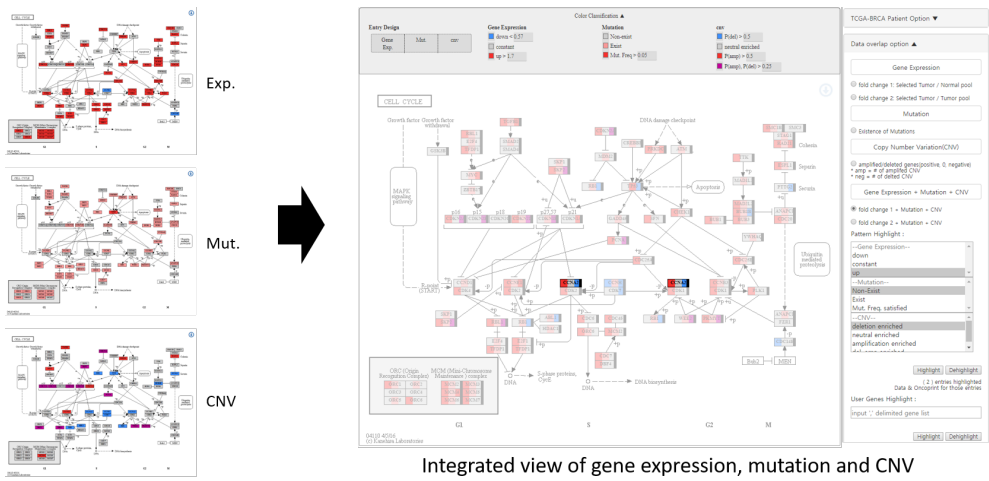


Figure 3.4: Integrated view of multi-omics data using BRCA-Pathway : BRCA-Pathway can visualize gene expression level, mutation, and CNV data, respectively, or combine three omics and show them in an integrated view. On the left side, we can see the pathways showing gene expression, mutation, and CNV in Cell cycle pathway, respectively. On the right side, it is a screen that shows three omics combined. One square can be regarded as one gene, and each square is divided into three parts. The first part indicates the level of gene expression. Red color means overexpression. The middle part represents the frequency of mutation. Gray color means there are no mutations. The last part represents the CNV mutation, blue color means copy number deletion. And through the highlight function, genes satisfying the conditions set by user are highlighted and displayed.

Visualization of multi-omics on KEGG pathway

Figure 3.5 shows visualization of TCGA BRCA using BRCA-Pathway. Multi-omics data of patients whose subtype is Basal is mapped on hsa04110 (Cell cycle pathway). User can view pathway maps in three different data types. Gene expression, mutation and copy number variation (CNV) data are visualized either separately or integrated. Genes with specific patterns are highlighted by selecting the conditions. For example, genes, having no mutation and of which CNV is Loss (-1) or Del (-2) but that are overexpressed, are selected. Genes in the pathway except the highlighted genes become shaded.

Investigation of multi-omics

Figure 3.6 shows the multi-omics of Basal subtype patients mapped on Cell cycle pathway. When user clicks each gene entry on the pathway, multi-omics of that gene are displayed such as gene expression level, mutation count ratio and CNV. CNV value -2 is copy number deletion, -1 is loss, 0 is neutral, +1 is gain, +2 is copy number amplification. For example, CCNA2 gene of Basal subtype patients is overexpressed. Fold-change of CCNA2 gene expression level of Basal subtype patients is 13.04 compared to normal tissues. Thus the gene expression level is colored in red. And there is no mutation in CCNA2. CNV is almost Neutral (0) or Loss (-1,-2). Among 223 Basal subtype patients, 191 patients were Neutral or Loss in CNV.

Transcription factor - Target gene correlation

Transcription is the process of making messenger RNA from DNA sequence. Gene expression data in BRCA-Pathway is the amount of messenger RNA. The amount of DNA sequence corresponding CCNA2 gene is not amplified, but the amount of messenger RNA of CCNA2 is 13 times more than normal tissue. In other words, the overexpression of CCNA2 gene in Basal subtype patients

is not due to the large amount of gene copy of CCNA2, but some regulators probably promoted the transcription of CCNA2. For more investigation, user can check the correlation coefficient of the transcription factor and target gene expression level, and user can see that the correlation coefficient between FOXP3 and CCNA2 is significantly different from the normal sample (0.58) and the tumor sample (0.16). The correlation coefficient value at (-0.07) is displayed in gray. As p-value is 0.05 or higher, it is displayed in gray. We found that the correlation breaks down in tumor samples.

Therefore, these findings can be the starting point for Basal subtype breast cancer research. We believe that BRCA-Pathway can provide insight into breast cancer research.

Mutual exclusivity in Breast cacner patients

Figure 3.7 is Oncoprint showing the mutual exclusive pattern of genes in Cell cycle pathway. Selecting genes, Oncoprint shows copy number variation (Amp or Del) and mutation in the mutual exclusive way between selected genes. In this case, some extremely overexpressed genes (fold change>10 by normal sample pool) and RB1 gene are selected. Oncoprint shows mutual exclusive pattern between CCNE2, RB1, CCNE1 and CDKN2A at the top 4 rows respectively. Each column represents each patient. Basal subtype patients are mutually exclusive in alteration of four genes. That is, patients with genomic alteration in CCNE1 gene have little alteration in other genes (CCNE2, RB1, CDKN2A).It is not easy to predict a key variation based on the frequency of variation. Because genomic alteration (i.e., mutation and CNV) on the same pathway is often mutually exclusive and pathways are disrupted by different combinations of these variations. Therefore, it is helpful to identify mutually exclusive alteration patterns by Oncoprint.

Figure 3.5: Visualization of TCGA BRCA using BRCA-Pathway : Multi-omics data of patients whose subtype is Basal is mapped on hsa04110 (Cell cycle pathway). Users can view pathway maps in three different data types. Gene expression, mutation and copy number variation (CNV) data are visualized either separately or integrated. Genes with specific patterns are highlighted by selecting the conditions. For example, genes, having no mutation and of which CNV is Loss (-1) or Del (-2) but that are overexpressed, are selected. Genes in the pathway except the highlighted genes become shaded.

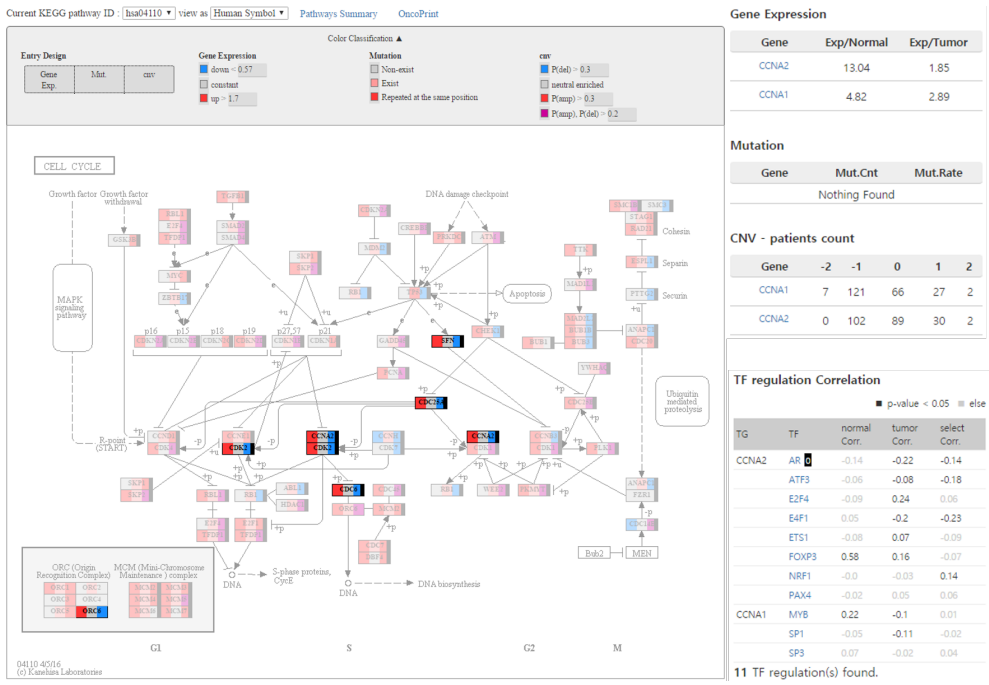


Figure 3.6: Multi-omics and TF-TG correlation : Clicking a gene entry on the pathway displays multi-omics data of the gene such as gene expression level compared with normal sample pool and with tumor sample pool, mutation count and ratio, CNV as -2 (Del), -1 (Loss), 0 (Neutral), 1 (Gain), 2 (Amp) with the count of patients having each CNV value and Pearson correlation coefficients of TF-TG gene expression. Three columns represent the correlation of TF-TG in normal sample pool, tumor sample pool and selected subpopulation pool, respectively. Correlation with p-value over 0.05 are shaded.

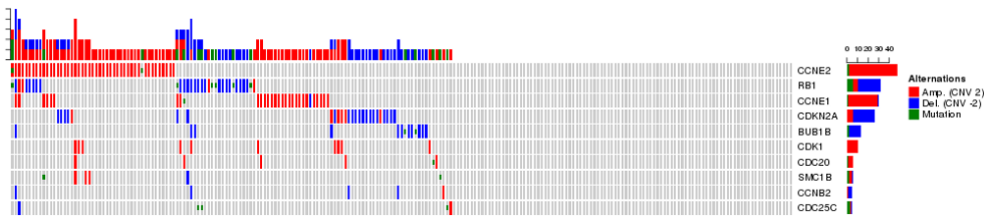


Figure 3.7: Mutual exclusivity by Oncoprint : selecting genes, Oncoprint shows copy number variation (Amp or Del) and mutation in the mutual exclusive way between selected genes. In this case, some extremely overexpressed genes (fold change> 10 by normal sample pool) and RB1 are selected. Each column represents each patient. Oncoprint shows mutual exclusive pattern between CCNE2, RB1, CCNE1 and CDKN2A at the top 4 rows respectively. Basal subtype patients are mutually exclusive in alteration of four genes. That is, patients with genomic alteration in CCNE1 gene have little alteration in other genes (CCNE2, RB1, CDKN2A).

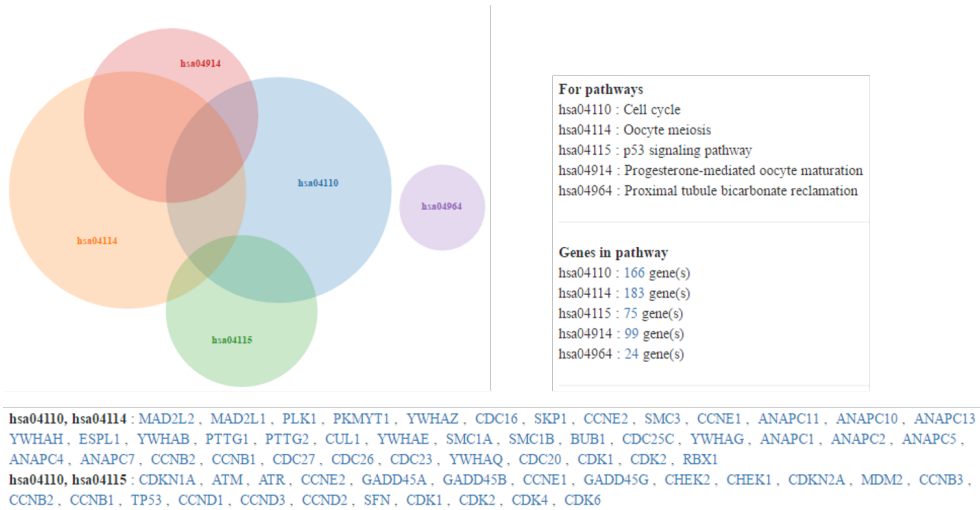
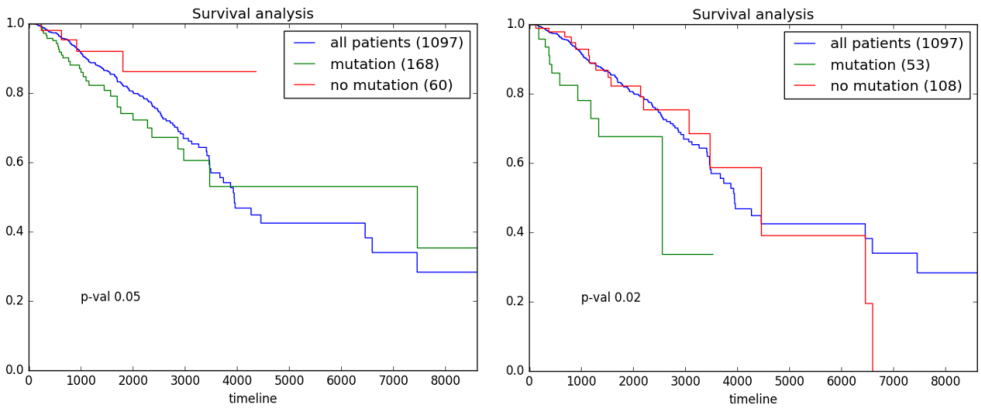


Figure 3.8: Pathway relations are shown in Venn diagram. A list of pathways and the number of genes in each pathway are described on the right side. Genes shared between pathways are listed under the Venn diagram.

Survival analysis

In addition to the visualization of multi-omics data, BRCA-Pathway provides survival plots in the context of pathway. Figure 3.9 is the survival plot of Basal subtype patient group. Basal subtype patient group can be divided into two groups by the presence of mutations in genes belonging to Cell cycle pathway. One is a group without mutations in Cell cycle pathway and the other is a group with mutations in Cell cycle pathway. Figure 3.9(a) Basal subtype and Cell cycle pathway are selected, patients with at least one mutation in the genes involved in Cell cycle pathway will belong to the mutation group (green line). On the other hand, patients without a mutation in Cell cycle pathway genes will belong to the mutation free group (red line), and all breast cancer patients are depicted as blue line for the comparison. (b) Her2 subtype and Oocyte meiosis pathway are selected, Her2 patients having at least one mutation in Oocyte meiosis genes are depicted as green line, and the rest Her2 patients are depicted as red line.

P-value by the log rank test is provided for the significance in the difference of two groups. P-value of 0.05 shows that there is a significant difference between the mutation group and mutation free groups. It can be seen that the survival plot of the mutation group is significantly different from the mutation free group. Therefore, we can say that mutations in genes belonging to the pathway affect survival, and based on this fact, we can get guide for selecting genes predicting prognosis in the subgroup of patients.



(a) Basal subtype patients, cell cycle pathway

(b) Her2 subtype patients, oocyte meiosis pathway

Figure 3.9: Survival analysis (a) Basal subtype and Cell cycle pathway are selected, patients with at least one mutation in the genes involved in Cell cycle pathway will belong to the mutation group (green line). On the other hand, patients without a mutation in Cell cycle genes will belong to the mutation free group (red line). (b) Her2 subtype and Oocyte meiosis pathway are selected, Her2 patients having at least one mutation in Oocyte meiosis genes are depicted as green line, and the rest Her2 patients are depicted as red line. P-value by the log-rank test is provided for the comparison of two groups, and all breast cancer patients are depicted as blue line for the comparison.

User data visualization on KEGG pathway

BRCA-Pathway provides visualization of user data to extend the usability of the system. Figure 3.10 shows the visualization of user data. After switching to User data mode, input a text file consisting of Entrez geneId and fold-change value, and then the gene expression level is mapped and colored on KEGG pathway.

Adjusting color or threshold value helps to customize pathway visualization. In the figure above, genes with fold-change over 1.7 in the gene expression level are colored in red, and blue when fold-change is less than 0.57. In the figure below, genes with fold-change over 5 in the gene expression level are colored in red, and blue when fold-change is less than 0.3. However, unlike TCGA mode, only a single gene selected by the entry label is considered and all other genes belonging to the entry are ignored.

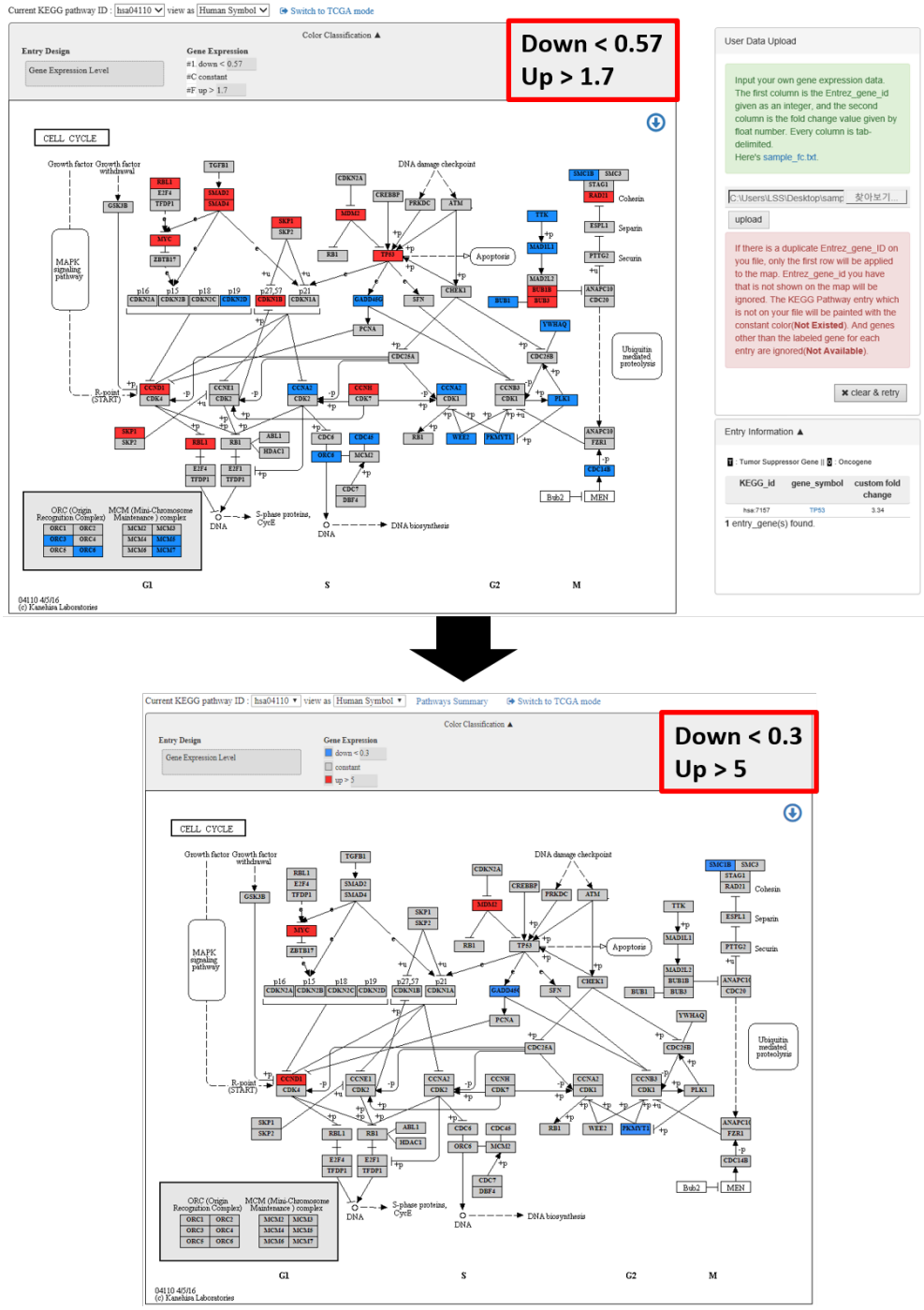


Figure 3.10: User data visualization on KEGG pathway

3.3.2 REST API

REST API separates data extraction from the developmental environment so that users can easily extract data without understanding the internal system. Given a patient option in the web page, a front-end program creates a URL and sends the URL to REST API. This allows the system to aggregate omics data sets for a subset of patients to create a dynamic user view.

By using REST API, it is possible to extract genes contained in KEGG pathway maps and to aggregate TCGA data after getting query result from MySQL. Users can access to data with simple endpoint coding.

REST API examples

The domain address is the server URL that BRCA-Pathway is configured on. After the slash (/) mark, at least one argument should be given. The first argument specifies the data to retrieve and the argument value can be `landscape, search, genes, pathways, TCGA-BRCA`.

In the example, `tcga-brca.bhi2.snu.ac.kr/api/landscape`, 'landscape' represents the current status of TCGA data and KEGG pathway. The argument 'search' means that the pathway list will be provided by searching gene names or pathway names, and 'genes' provides the gene list in pathways specified by argument 2. Furthermore, 'pathways' returns pathway information specified by argument 3's endpoint filtered by argument 2. The last example means that it will provide the result of aggregating CNV data from TCGA-BRCA data given the patients option is male and the genes filtered by the pathway 'hsa00010'. REST API arguments and parameters are shown in Table 3.2.

- tcga-brca.bhi2.snu.ac.kr/api/landscape
- tcga-brca.bhi2.snu.ac.kr/api/search?keyword=erbb1

- tcga-brca.bhi2.snu.ac.kr/api/genes/hsa00010+hsa00030
- tcga-brca.bhi2.snu.ac.kr/api/pathways/hsa00010/related_pathways
- tcga-brca.bhi2.snu.ac.kr/api/TCGA-BRCA/hsa00010/CNV?gender=male

Patient_options are listed below:

- subtype_BHI_RNASeq_Log2 : all|Basal|Her2|LumA|LumB|Normal
- years_to_birth_from : integer & years_to_birth_to : integer
- er_status : all|indeterminate|negative|positive
- pr_status : all|indeterminate|negative|positive
- her2_status : all|indeterminate|negative|positive|equivocal
- vital_status : all|0|1 *0: alive, 1:dead
- pathologic_stage : all|stage_i|stage_ii|stage_iii|stage_iv|stage_iv|stage_tis|stage_x
- pathologic_T_stage : all|t1|t2|t3|t4|tx
- pathologic_N_stage : all|n0|n1|n2|n3|nx
- pathologic_M_stage : all|cm0_|m0|m1|mx
- gender : all|female|male
- radiation_therapy : all|no|yes
- histological_type : all|infiltrating_carcinoma_nos|infiltrating_ductal_carcinoma|infiltrating_lobular_carcinoma|medullary_carcinoma|metaplastic_carcinoma|mixed_history(please_specify)|mucinous_carcinoma|other__specify
- number_of_lymph_nodes : all|0|1|2 *1: #of node less than or equal to 10, 2: greater thane 10

- race : all|american_indian_or_alaska_native|asian|black_or_african_american|white
- ethnicity : all|hispanic_or_latino|not_hispanic_or_latino
- menopause_status : all|indeterminate|peri|post|pre

arg1	arg2	arg3	Parameters	Description
landscape	-	-	type=[JSON | LIST] fields=[KEGGPathway_update | KEGGPathways | Pathway_Pathway_Membership | TCGAGenes | TCGA_BRCA_Patients | TCGA_BRCA_update | TFTG_relation | driver_genes]	KEGG pathways, TCGA-BRCA data status
search	-	-	keyword type=[JSON | LIST] fields=[keyword | gene_labels | rate | pathID | pathname]	Get search result by keyword. Gene or pathway names with input_type.
genes	(pathIDs, sep="+")	- count	type=[JSON | LIST] fields=[pathID | gene_symbol | pathname | geneID | count**] **given only arg3 related_pathways	Retrieve pathways information. related_pathways operation supported.
pathways	(pathIDs, sep="+")	- related_pathways	type=[JSON | LIST] fields=[pathID | pathname | category_level2_id | description | related_pathID**] **given only arg3 related_pathways	Retrieve pathways information. related_pathways operation supported.
TCGA-BRCA	patient_list patient_count (pathIDs, sep="+")	- - (given pathIDs,) fold_change mutation CNV	type=[JSON | LIST] ** arg2 patient_list fields=[patient_barcode] ** arg2 patient_count fields=[count] ** arg3 fold_change fields=[geneID|gene_symbol | exp_select | exp_tumor | exp_normal | fc_select_tumor| fc_select_normal] ** arg3 mutation fields=[geneID | gene_symbol |count] ** arg3 CNV fields=[geneID | gene_symbol | cnv_pos | cnv_zero | cnv_neg] ** And patient_options described_above	Retrieve pathways and TCGA-BRCA omics data. Gene expression with fold_change value, mutation with the number of patients having mutation, and CNV with GISTIC2 result range from -2 to 2. Each omics result will be given by gene level aggregating patients group omics data patient_list, patient_count operation supported.

Table 3.2: REST API arguments and parameters supported by BRCA-Pathway

Chapter 4

Pathway-guided gene set selection

In order to select gene sets specific to each subtype guided by pathway information, we conducted studies according to the following strategy.

Strategy:

- Select pathways that are important in breast cancer
- Show how each of these pathways are different in subtypes
- Show how to quantify the difference in the pathway level of each subtype
- Discuss about probability framework to select genes using ML techniques
- Show how activation of these pathways can show difference among subtypes

4.1 Motivation

The information through analysis of each gene is limited. This is because the genes do not work alone, but through interactions between genes in a network. Therefore, it is necessary to analyze and interpret data from a biological process, pathways and networks. And with this approach, we can have a global perspective on the data.

In bioinformatics research, pathway analysis is used to identify related genes within a pathway. This is helpful when studying differential expression of a gene in a disease. By examining the changes in gene expression in a pathway, its biological causes can be explored. Pathway analysis helps to understand or interpret omics data from the point of view of canonical prior knowledge structured in the form of pathways diagrams (García-Campos *et al.*, 2015).

4.1.1 Important pathways in breast cancer

We selected important pathways in breast cancer from previous studies (Huang *et al.*, 2003; Bild *et al.*, 2006; Gatza *et al.*, 2010) as shown in Table 4.1.

- Breast cancer pathway summarizes the major cancer progression process for each subtype of breast cancer. The molecular subtypes of breast cancer, which are based on the presence or absence of hormone receptors (estrogen and progesterone subtypes) and human epidermal growth factor receptor-2 (HER2). Hormone receptor positive breast cancers are largely driven by the estrogen/ER pathway. In HER2 positive breast tumours, HER2 activates the PI3K/AKT and the RAS/RAF/MAPK pathways, and stimulate cell growth, survival and differentiation. In patients suffering from TNBC, the deregulation of various signalling pathways (Notch and Wnt/beta-catenin), EGFR protein have been confirmed.

- Apoptosis is a genetically programmed process for the elimination of

damaged or redundant cells by activation of caspases (aspartate-specific cysteine proteases).

- Cell cycle pathway is a unidirectional process that governs cell division, and cells respond to DNA damage by activating signaling pathways that promote cell cycle arrest and DNA repair.

- P53 activation is induced by a number of stress signals, including DNA damage, oxidative stress and activated oncogenes. This results in three major outputs; cell cycle arrest, cellular senescence or apoptosis.

- Circadian rhythm is an internal biological clock, which enables to sustain an approximately 24-hour rhythm in the absence of environmental cues. In mammals, the circadian clock mechanism consists of cell-autonomous transcription-translation feedback loops that drive rhythmic, 24-hour expression patterns of core clock components.

- The ErbB protein family or epidermal growth factor receptor (EGFR) family is a family of four structurally related receptor tyrosine kinases. Excessive ErbB signaling is associated with the development of a wide variety of types of solid tumor. ErbB-1 and ErbB-2 are found in many human cancers and their excessive signaling may be critical factors in the development and malignancy of these tumors.

- Estrogen receptor refers to a group of receptors which are activated by the hormone estrogen. The main function of the estrogen receptor is as a DNA binding transcription factor which regulates gene expression.

- The Notch pathway is an evolutionally conserved signaling pathway which plays an important role in diverse developmental and physiological processes. These include cell-fate determination, tissue patterning and morphogenesis, cell differentiation, proliferation and cell death.

Table 4.1: Important pathways in breast cancer

KEGG ID	KEGG pathway name
hsa04210	Apoptosis pathway
hsa05224	Breast cancer pathway
hsa04110	Cell cycle pathway
hsa04710	Circadian rhythm pathway
hsa04012	ERBB signaling pathway
hsa04915	Estrogen signaling pathway
hsa04010	MAPK signaling pathway
hsa04330	Notch signaling pathway
hsa04114	Oocyte meiosis pathway
hsa05200	Pathways in cancer
hsa04151	PI3K-Akt signaling pathway
hsa04115	P53 signaling pathway
hsa04014	Ras signaling pathway
hsa04310	Wnt signaling pathway

- Ras proteins are small GTPases and are involved in transmitting signals within cells. In this way, Ras signaling controls many downstream processes, including cell proliferation, survival, growth, migration and differentiation.

- WNT signal, through the canonical pathway, controls cell fate determination and through the non-canonical pathway controls cell movement and tissue polarity (Kandasamy *et al.*, 2010).

- Pathways in cancer pathway shows the overview of cancer.

4.1.2 Difference between each subtype on the pathways

Through the previous study BRCA-Pathway, we visualised the expression level of the gene for each subtype of breast cancer as a pathway level. And it was found that the difference in expression level clearly distinguished each subtype in biologically important pathways.

The cell cycle is the series of events that takes place in a cell leading to its division and duplication. Regulation of the cell cycle involves processes crucial to the survival of a cell, including the detection and repair of genetic damage as well as the prevention of uncontrolled cell division (Vermeulen *et al.*, 2003). Cancer is unchecked cell growth. Mutations in genes can cause cancer by accelerating cell division rates or inhibiting normal controls on the system, such as cell cycle arrest or programmed cell death (O'Connor *et al.*, 2010).

For example, in LumA subtype patient, it can be seen that the expression level of the gene belonging to Cell cycle pathway is lower than the average expression of all breast cancer patients (Figure 4.1 (a)). In contrast, it can be seen that the gene expression level of the patients of Basal subtype is higher than the average (Figure 4.1 (c)). This is consistent with the previous research that cell cycle genes are known to be associated with proliferation, whose overexpression is prognostic of poor clinical outcome (Dai *et al.*, 2016). While identifying visual differences at the pathway level, we became interested in the problem of quantifying the differences in the pathway level and using the quantified values to select genes specific to each subtype.

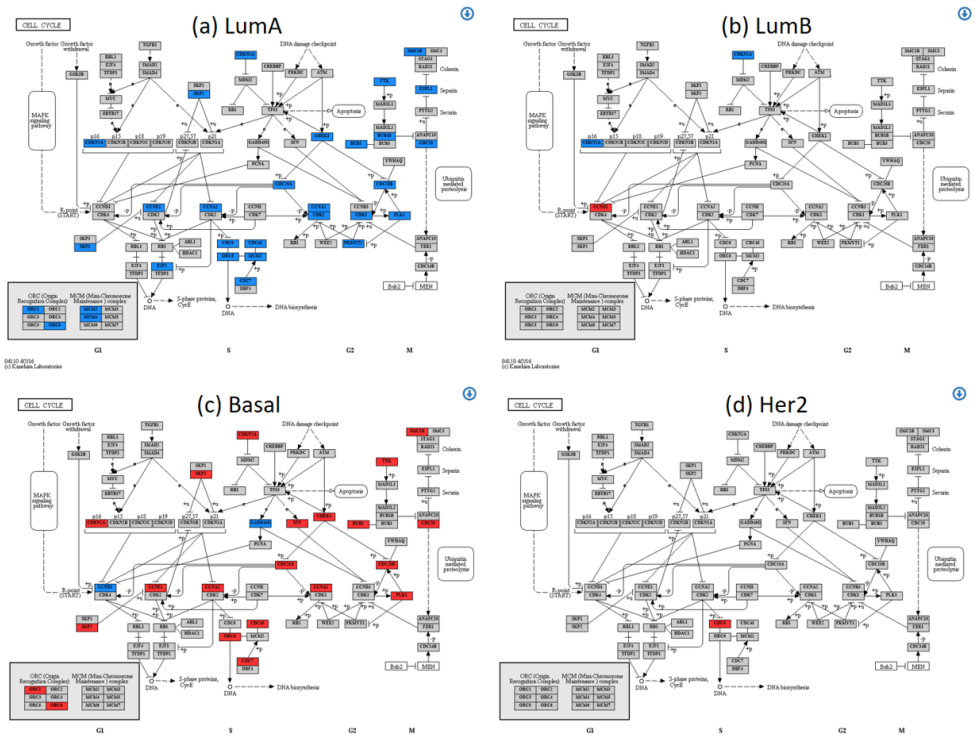


Figure 4.1: Difference in expression level between breast cancer subtype in Cell cycle pathway. This shows the gene expression level mapped on Cell cycle pathway using BRCA-Pathway. It can be seen that the expression level of genes belonging to Cell cycle pathway is different for each subtype.

4.1.3 Challenge in gene selection from pathways

However, our method is challenging because only one third of the 20,000 genes are included in KEGG pathway. Nevertheless, if the pathway-based gene selection is successful, it is solving two problems with one solution. In other words, the selection of subtype-specific genes and the interpretation of biological functions of the selected genes are performed simultaneously. This is because the KEGG pathway is a collection of manually drawn pathway maps representing expert knowledge on the molecular interaction, reaction and relation networks (Kanehisa and Goto, 2000).

There have been many gene expression signatures to predict prognosis in breast cancer (Sotiriou and Pusztai, 2009; Kwa *et al.*, 2017). However, there has been no attempt to calculate the probability for each gene based on a probabilistic framework using the expression distribution of genes and the pathway activation score, and to select subtype specific gene by the probability value.

4.2 Methods

4.2.1 Gene factor and Pathway factor

Difference in gene expression level between subtypes

Figure 4.2 shows a histogram and kernel density estimation of GATA3 gene. The x-axis is the normalized expression level, and the y-axis is the probability density. Kernel density estimation is used to show Basal subtype and other subtypes. It shows that the expression level of GATA3 gene differs between breast cancer subtypes. Gene expression was expressed on the x-axis and probability density on the y-axis using the number of patients belonging to the bin. It can be seen that the expression level of GATA3 gene in patients corresponding to Basal subtype is small compared to other subtypes.

Definition of Gene factor

Gene factor is defined using the area of the area where the probability density functions of other types do not overlap among the total area of Basal subtype probability density function. In case of GATA3 gene, since the area of the overlapping part is 0.2, Gene factor of GATA3 gene in Basal subtype patient group can be said to be 0.8 (Figure 4.3).

$$\textit{Gene factor}_{s_m}(g_i) := \int_{mn}^{mx} \mathrm{pdf}_{s_m}(g_i) - \int_{mn}^{mx} \min(\mathrm{pdf}_{s_m}(g_i), \mathrm{pdf}_{others}(g_i)) \tag{4.1}$$

where mn, mx is the minimum and maximum value of the probability density function of the target subtype on the x-axis, respectively.

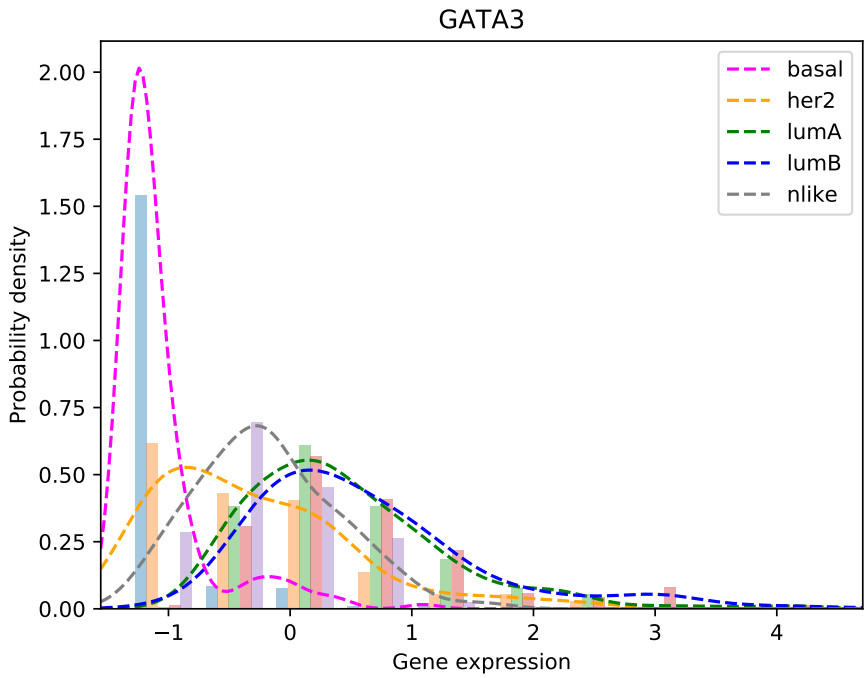


Figure 4.2: Histogram and kernel density estimation of GATA3 gene : the x-axis is the normalized expression level, and the y-axis is the probability of patients in the corresponding section. The kernel density estimation is used to show the distributional difference between subtypes. The expression level of GATA3 gene in Basal subtype patient group is lower than that of other subtypes. Pink dashed line indicates Basal subtype patient group.

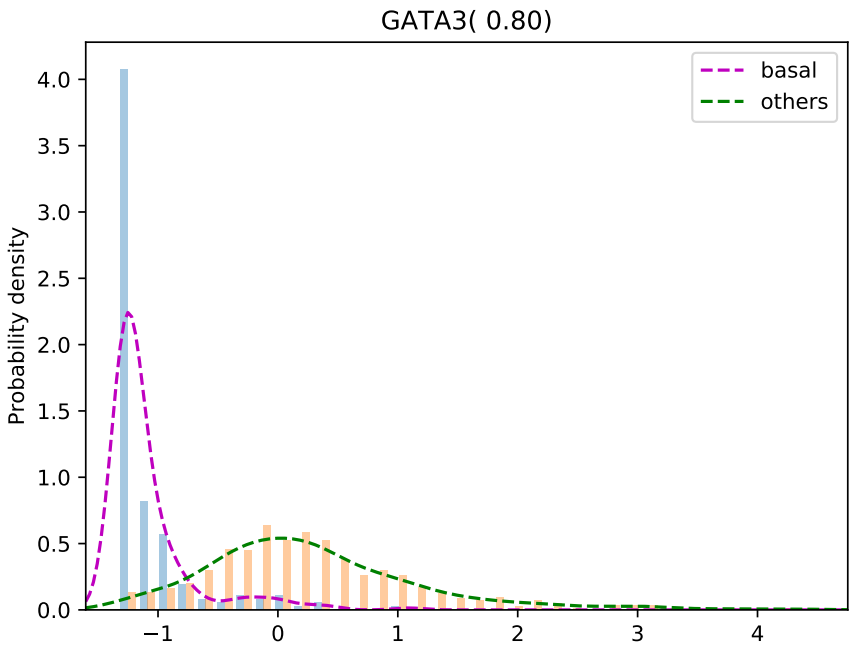


Figure 4.3: Gene factor of GATA3 gene in Basal subtype. Histogram of GATA3 gene expression. The x-axis represents normalized gene expression, and the y-axis is probability of patients in the interval. Distribution of gene expression level shows that GATA3 gene expression of Basal subtype is lower than that of other subtype. Gene factor is defined as not overlapping area under the PDF of GATA3 in Basal subtype. In this case, Gene factor of GATA3 in Basal subtype is 0.8.

Difference in pathway activation level between subtypes

Just as a gene has an expression level, a pathway has a degree of activation by aggregating the gene expression level belonging to the pathway. There have been many studies that quantify the pathway activation levels (Lim *et al.*, 2020), and among them, we used Lim et al.'s approach (Lim *et al.*, 2016) to quantify the activation level of each pathway. Each patient has 215 pathway activation level data, as if each patient has 20,000 gene expression data.

Definition of Pathway factor

Similar to Gene factor, Pathway factor is defined using non-overlapping area under the probability density function. Basal subtype group and other subtype groups were represented as histogram using the pathway activation score (Figure 4.4). In this case, Pathway factor of hsa05321 in Basal subtype patient group is 0.37.

$$\textit{Pathway factor}_{s_m}(p_j) := \int_{mn}^{mx} \mathrm{pdf}_{s_m}(p_j) - \int_{mn}^{mx} \min(\mathrm{pdf}_{s_m}(p_j), \mathrm{pdf}_{others}(p_j)) \tag{4.2}$$

where mn, mx is the minimum and maximum value of the probability density function of the target subtype on the x-axis, respectively.

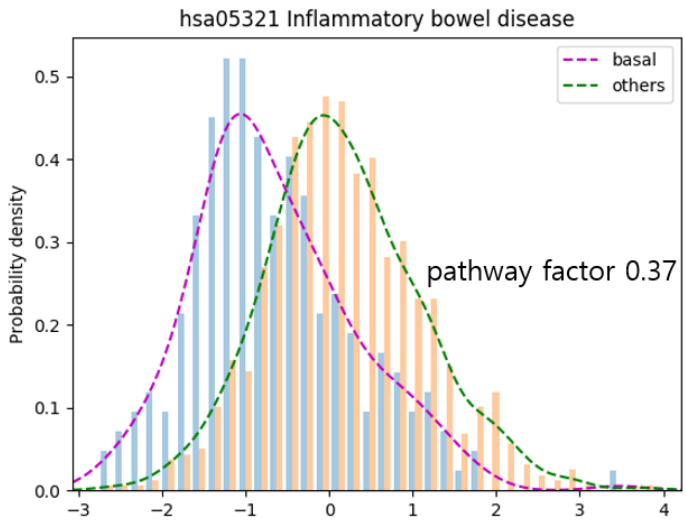


Figure 4.4: Pathway factor of Inflammatory bowel disease pathway. Histogram of pathway activation score in hsa05321 : the x-axis represents normalized pathway activation score, and the y-axis is the probability density. Distribution of pathway activation score shows that hsa05321 pathway of Basal subtype patients has lower activation score than that of other subtype. And Pathway factor of hsa05321 in Basal subtype is 0.37.

4.2.2 Likelihood and Posterior probability

Posterior probability

Since the purpose of our research is to select genes specific to each subtype, the problem of selecting genes specific to each subtype can be considered as a problem of selecting genes with high posterior probability of each subtype by Bayes theorem.

Bayes theorem:

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)} \tag{4.3}$$

- $P(h \mid D)$ is the posterior probability of hypothesis h
- $P(h)$ is the prior probability of hypothesis h, background knowledge about the chance that h is a correct hypothesis.
- $P(D \mid h)$ is the likelihood of the data D given h
- $P(D)$ is the prior probability of data D, that training data D will be observed.

The most probable hypothesis h given the data D is called a *maximum a posteriori* (MAP) hypothesis.

$$\begin{aligned} h_{MAP} &\equiv \underset{h \in H}{\operatorname{argmax}} P(h \mid D) \\ &= \underset{h \in H}{\operatorname{argmax}} \frac{P(D \mid h)P(h)}{P(D)} \\ &= \underset{h \in H}{\operatorname{argmax}} P(D \mid h)P(h) \end{aligned} \tag{4.4}$$

Equation 4.3 is applied to our problem as follows.

Problem definition:

$$P(s_m \mid g_i) = \frac{P(g_i \mid s_m)P(s_m)}{P(g_i)} \tag{4.5}$$

- $P(s_m \mid g_i)$ is the posterior probability of subtype s_m given gene g_i
- $P(s_m)$ is the prior probability of hypothesis s_m. The proportion of patients with each subtype among all breast cancer patients will be used as this prior probability.
- $P(g_i \mid s_m)$ is the likelihood of the data g_i given s_m
- $P(g_i)$ is the prior probability of data g_i, that training data g_i will be observed.

$$P(g_i) = \sum_{m=1}^{5} P(g_i \mid s_m)P(s_m) \tag{4.6}$$

Likelihood

To utilize the pathway information, the likelihood $P(g_i \mid s_m)$ is decomposed as follows.

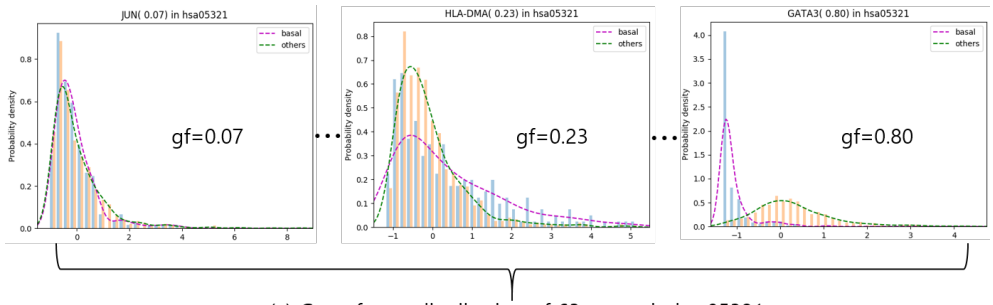
$$P(g_i \mid s_m) \colon = \frac{\sum_{j=1}^{215} P(g_i \mid s_m, p_j)P(p_j \mid s_m)}{n} \tag{4.7}$$

where n is the number of pathways containing the gene g_i

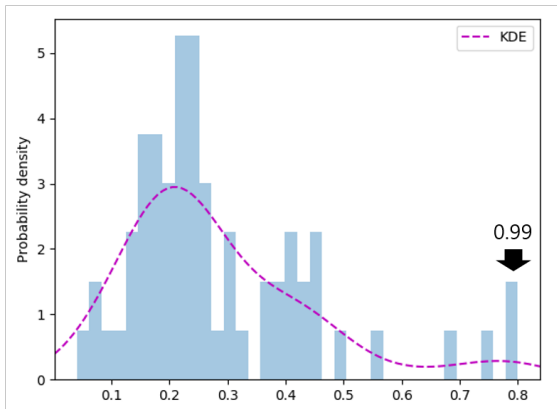
$$P(g_i \mid s_m, p_j) \colon = \text{percentile of } \mathit{Gene\ factor}_{s_m}(g_i) \text{ in the pathway } p_j \tag{4.8}$$

$$P(p_j \mid s_m) \colon = \text{percentile of } \mathit{Pathway\ factor}_{s_m}(p_j) \text{ in all the pathway set} \tag{4.9}$$

Figure 4.5 shows $P(g_i \mid s_m, p_j)$, the conditional probability of gene g_i given subtype s_m and pathway p_j. Figure 4.6 shows $P(p_j \mid s_m)$, the conditional probability of pathway p_j given subtype s_m.

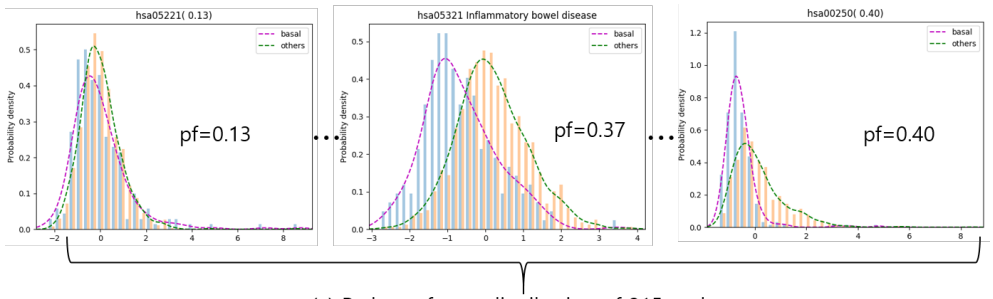


(a) Gene factor distribution of 63 genes in hsa05321

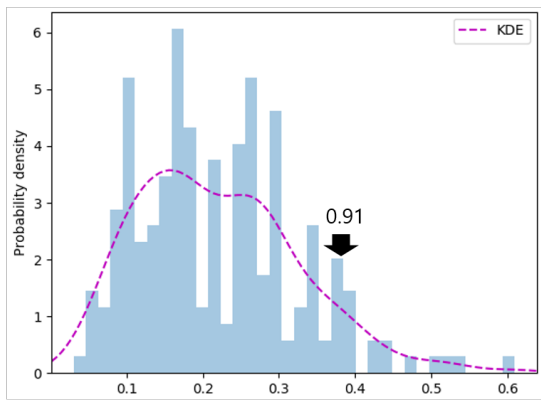


(b) Percentile of gf 0.8 among 63 genes in hsa05321

Figure 4.5: Distribution of Gene factors for all genes in hsa05321 pathway. The x-axis represents Gene factor value, and y-axis is based on the counts of genes in the interval. $P(g_i|s_k, p_j)$ is defined as percentile of Gene factor of gene g_i in pathway p_j. Percentile of Gene factor of GATA3 in hsa05321 Inflammatory bowel disease (IBD) pathway is 0.99.



(a) Pathway factor distribution of 215 pathways



(b) Percentile of pf 0.37 among 215 pathways

Figure 4.6: Distribution of Pathway factor : the x-axis represents Pathway factor of 215 KEGG pathways, and the y-axis is based on the counts of pathways in the interval. Pathway factor of hsa05321 of Basal subtype (0.37) is located in higher in the overall distribution. $P(p_j|s_k)$ is percentile of Pathway factor of pathway p_j in all the pathway set : percentile of Pathway factor of hsa05321 inflammatory bowel disease (IBD) pathway is 0.91

Why posterior probability?

As described above, likelihood is a value calculated by reflecting the pathway information in the process of selecting a gene having a difference in expression level for each subtype. That is, it is basically a value that serves as a criterion for selecting genes specifically expressed in each subtype. Meanwhile, the posterior probability is a normalized value for each of the five subtypes, reflecting the background knowledge we know about the subtype ratio of breast cancer patients.

For example, certain genes have very high likelihood in the normal-like subtype patients. However, if this group of normal-like subtype has less than 10 percent of all the breast cancer patients, the posterior probability of this gene is not very high compared to that of high likelihood. Actually, the posterior probability has a low value no matter how high the likelihood is.

The purpose of our study is to select genes for solving the problem of classifying all breast cancer patients into each subtype, while being specific to each subtype. Therefore, it can be said that even though the likelihood in a small number of patient groups is very high, it does not satisfy our purpose.

4.3 Results

We calculated the likelihood and posterior probability of each gene using Gene factor and Pathway factor based on the probabilistic framework. Then, genes were ranked by its likelihood and posterior probability in each subtype. Table 4.2 shows the list of genes by likelihood. Table 4.3 shows the list of genes by posterior probability.

We evaluated the selected gene set in two ways. The first evaluation was performed by measuring breast cancer subtype classification accuracy in the context of feature selection problem. The second evaluation was performed as Gene Ontology analysis and Pathway enrichment analysis from a biological point of view.

To see if the selected genes distinguish subtypes of breast cancer, we visualized patient data in the geometric space by likelihood. In the same way, breast cancer patient were shown in the geometric space by posterior probability. Finally, we performed Gene Ontology (GO) analysis and Pathway enrichment analysis to understand the biological meaning of the selected gene set.

Table 4.2: The list of top 20 genes for each subtype by likelihood

Basal	HER2	lumA	lumB	normal-like
XBP1	SYT1	SEPHS1	VNN1	TXNIP
VNN1	PGK1	RORB	TTK	TAC1
TMBIM6	NRG2	PSAT1	SFRP1	SGOL1
SUOX	NME1	NRG2	ROR1	CRY2
SPOPL	MED24	LASS3	ORC6L	BRCC3
PLOD1	MED1	CRY2	KLK2	STAG3
MTHFD1L	DNAH8	CDC6	DNALI1	SMC1B
ME2	CES1	CDC45	DNAL1	PENK
GPR161	ACSM3	TYMS	CSGALNACT1	NLRP1
DEGS2	CTPS	TTK	CHST3	GBP2
BHLHE40	ACSM1	ORC1L	CDC6	AVPR2
PRNP	HCCS	NRG1	CDC45	AQP1
NEU4	NRG1	ORC6L	MAML2	ALDH1L1
GNMT	DHRS11	MTHFD1L	LGR6	SOX17
CRY2	CRY2	GBP1	EPHB6	SFRP1
SUV39H2	DNAH3	DSG1	DTX1	MGAT3
PCBD2	SRD5A1	BRCC3	DNAH17	IFI16
PAX6	MTHFD1L	GBP5	DKK1	QRSL1
LASS3	DNAH12	MGAT3	DNAI1	SSTR1
QDPR	NME2	DEFB1	PIGR	MTHFD1L

Table 4.3: The list of top 20 genes for each subtype by posterior probability

Basal	HER2	lumA	lumB	normal-like
ZNF768	QPRT	PABPN1	MMP14	POLR3H
PHKG2	NMNAT2	TNFRSF18	MLST8	LTB4R
ZNF267	NMNAT1	SHARPIN	CEBPE	UBE2G1
SETD1B	NAPRT1	CFHR3	REL	PPP1R3F
MBOAT7	PGM2L1	ZNF708	CLIP1	ZNF256
GORAB	NNT	ZNF426	ZBTB17	UBB
CPT2	NMNAT3	ZNF26	POU2F1	TELO2
CARD14	FN1	SNRPB	NRCAM	TAOK3
ZNF570	SIRT7	P2RY10	KRIT1	NDUFS2
WWC1	NUDT12	MOBKL1A	GPS2	HDAC10
TAZ	NT5M	IL1RN	DAPK3	H3F3A
SHISA5	NFE2L2	ENTPD4	CCNT1	GK
PARK7	FLOT2	ELOVL7	AGPAT6	DOCK4
ETNK2	GMPPA	NUP98	IP6K3	CACNA1I
MLL3	NADK	TNFRSF19	CHST13	ZNF419
ROCK2	PLA2G15	DDX3X	ENDOG	FNIP2
GALNTL4	PTDSS2	ZNF556	SIX4	OR51E1
EHMT2	IDH1	SIVA1	PPIP5K2	DEPDC5
EEA1	TWIST1	AMH	AGPAT2	ZNF8
B3GALT5	TWIST2	ARSA	APBB1IP	ABCA1

4.3.1 Visualization of patients in the geometric space

Genes are ranked for each subtype using probability, and patient data is represented as vector using the top ranked genes. For example, if top 10 genes are selected, each patient is represented as a 10-dimensional vector. To visualize multi dimensional data, we used t-SNE (Maaten and Hinton, 2008).

Geometric space by likelihood

Figure 4.7 shows the visualization of breast cancer patient data using the selected 10 gene vectors. Top ranked 10 genes by the likelihood of gene given Basal subtype are selected. And all the patients are represented with these 10 genes as a 10-dimensional vector. Since the selected genes are specific to Basal subtype, it can be seen that Basal subtype versus the rest subtype is well divided. As Figure 4.7(b) shows, patients in the test data set are also well divided into Basal vs. others. Figure 4.8 shows the visualisation of breast cancer patient in the training data set using union gene set. Selecting top 10 genes for each subtype, 42 genes are used except for the duplicate genes. In Figure 4.7, the patients are divided into Basal versus the others, while in Figure 4.8 we can see that they are divided into subtypes. It is a natural result to be distinguished by each subtype because genes specific to each subtype are selected and patient data is represented with those union genes.

Geometric space by posterior probability

Figure 4.9 shows patient data with genes selected using posterior probability. Compared to Figure 4.7, it is not well classified by subtypes than when patients are represented with genes selected using likelihood. Gene sets by posterior probability did not distinguish each subtype well.

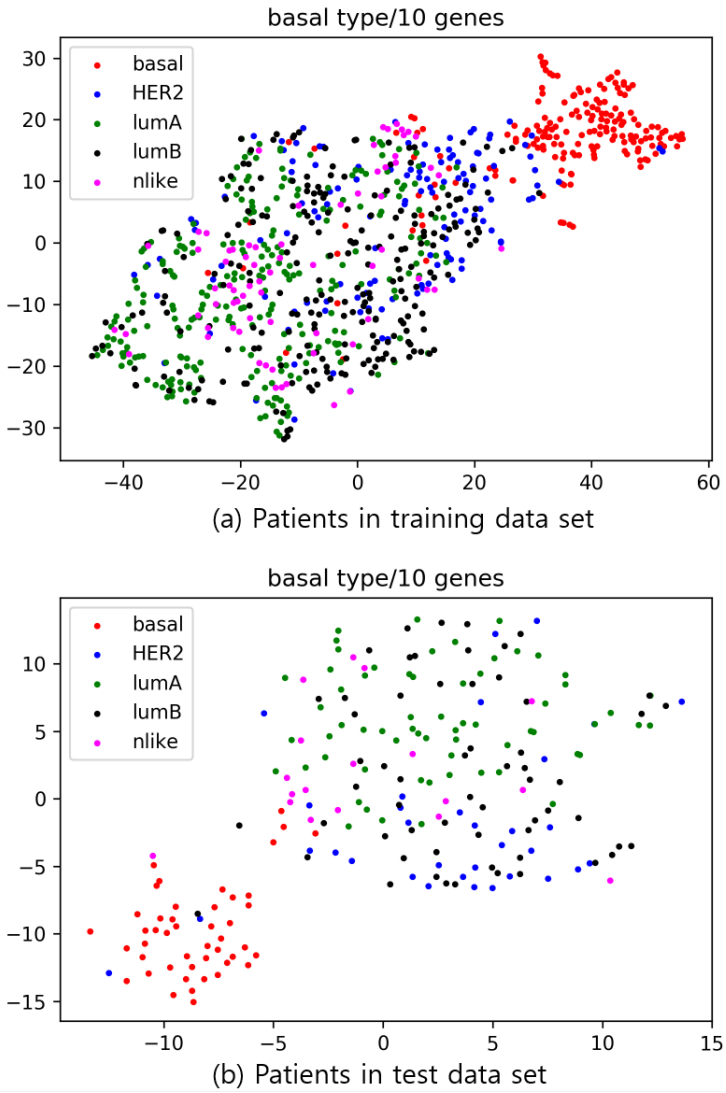


Figure 4.7: Breast cancer patient in the geometric space using top 10 genes by likelihood. This is the result of selecting the top 10 genes based on the likelihood of gene given Basal subtype and visualization of breast cancer patient data using the selected 10 gene vectors. Since the selected 10 genes are specific in Basal subtype, it can be seen that Basal subtype versus the rest subtype is well divided.

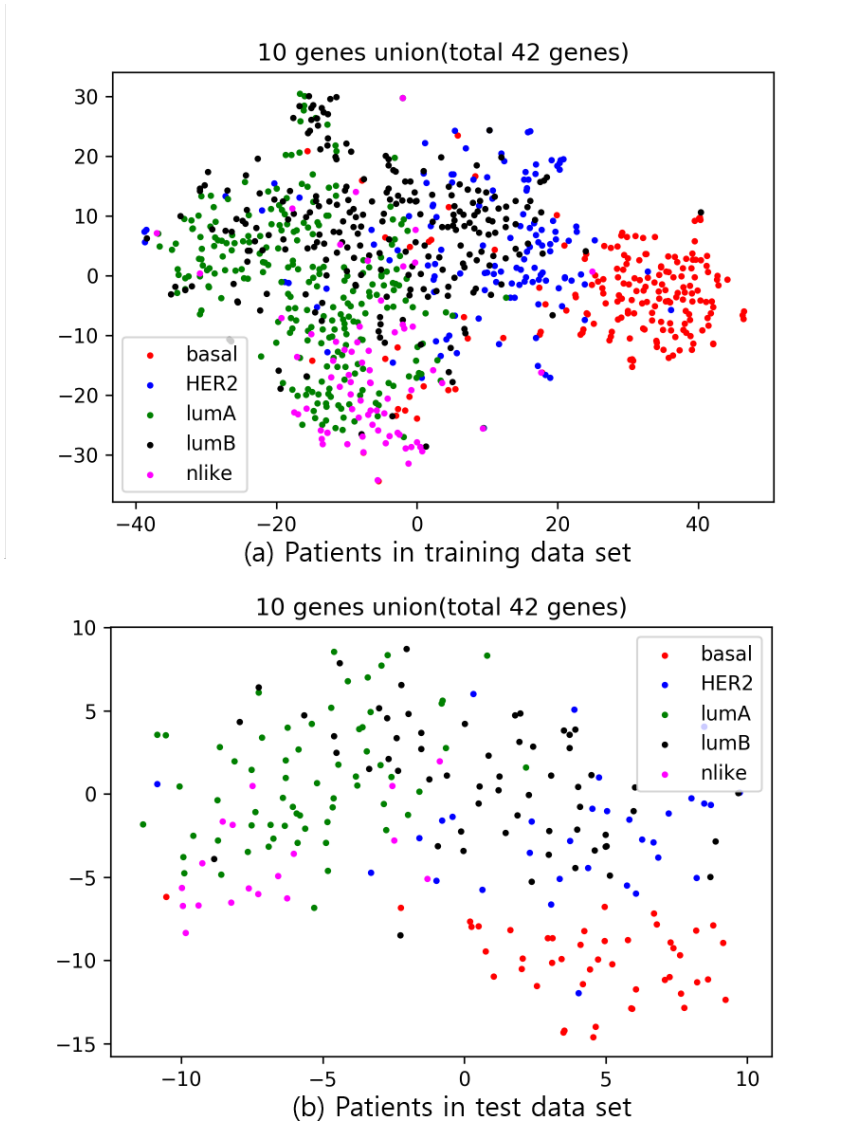


Figure 4.8: Breast cancer patients in the geometric space using 42 union genes by likelihood : visualisation of breast cancer patient in the training data set using union gene set. Selecting top 10 genes for each subtype, 42 genes are used except for the duplicate genes. Compared to Figure 4.7, patients are divided by each subtype.

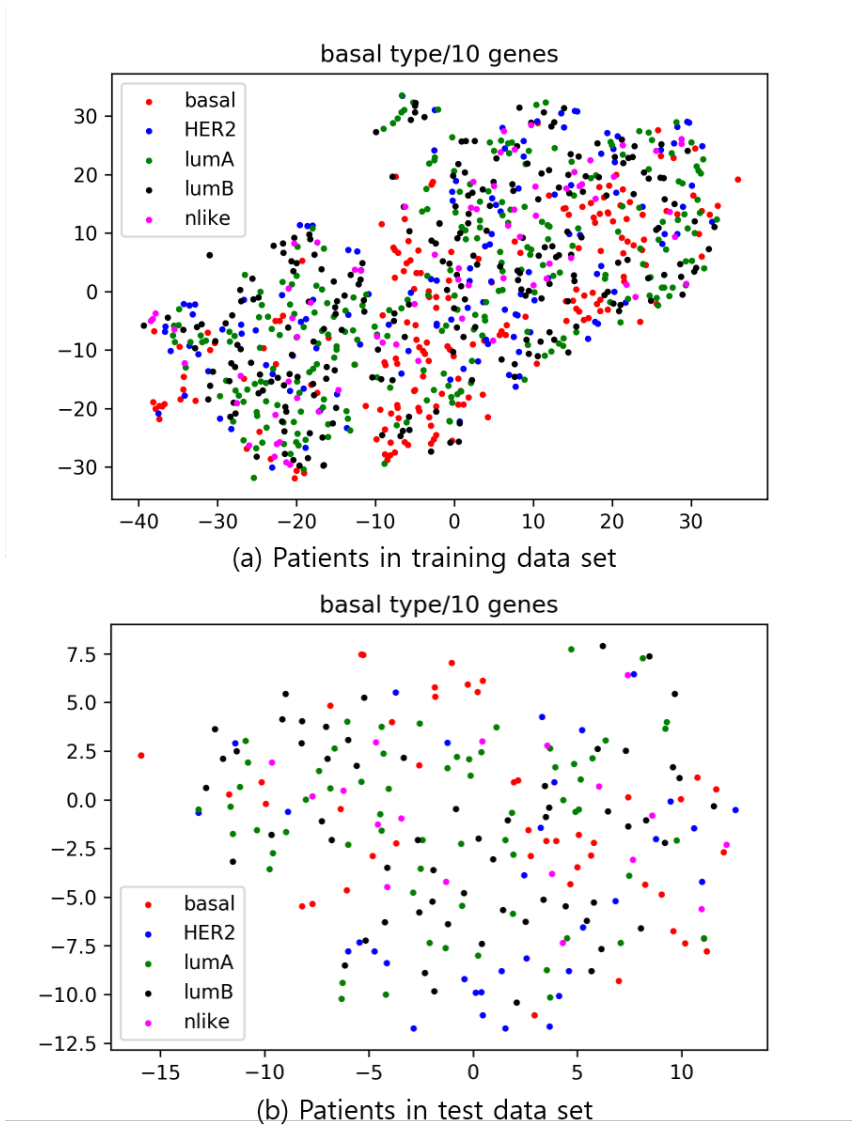


Figure 4.9: Breast cancer patient in the geometric space by posterior probability criteria : visualisation of breast cancer patient in the training data set using top 10 Basal subtype specific genes.

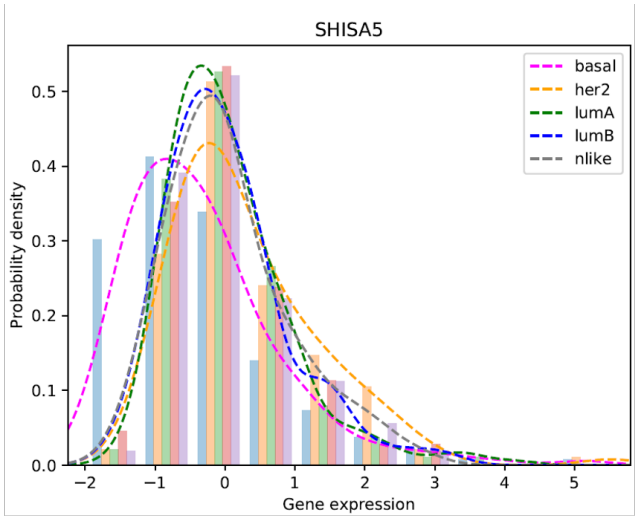
High posterior probability

As we can see in Figure 4.10, genes with a high posterior probability generally have little difference in expression levels between subtypes, and thus have a low likelihood. Therefore, there is not much ability to distinguish subtypes, which is why it is difficult to distinguish subtypes in t-SNE as Figure 4.9 shows.

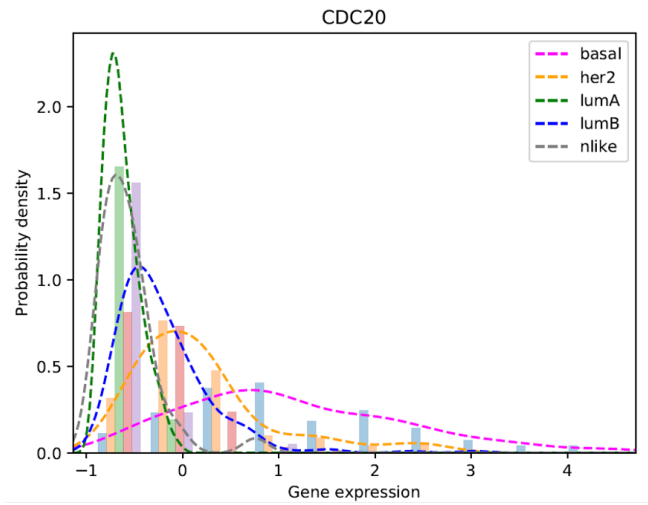
Instead, they are distinct from other subtypes so that we can consider other subtypes in one subtype. In that sense, genes having high posterior probability are really specific to that subtype. We can revise the calculation of the posterior probability in Equation 4.10, it can be seen that the expression of a gene with a high posterior probability differs only in the corresponding subtype as in the case of the SHISA5 gene. And there is little difference in between other subtypes. This is because the likelihood of the subtype is normalized with the sum of likelihood of other subtypes. If the likelihood of each subtype is large, the likelihood of the corresponding subtype is divided by a large value, so the posterior probability decreases.

$$\begin{aligned} P(s_m \mid g_i) &= \frac{P(g_i \mid s_m)P(s_m)}{P(g_i)} \\ &= \frac{P(g_i \mid s_m)P(s_m)}{\sum_{m=1}^{5} P(g_i \mid s_m)P(s_m)} \end{aligned} \tag{4.10}$$

$$\begin{aligned} P(Basal \mid \text{SHISA5}) &= \frac{P(\text{SHISA5} \mid Basal)P(Basal)}{P(\text{SHISA5})} \\ &= \frac{P(\text{SHISA5} \mid Basal)P(Basal)}{\sum_{m=1}^{5} P(\text{SHISA5} \mid s_m)P(s_m)} \end{aligned} \tag{4.11}$$



(a) Gene expression level of SHISA5 gene



(b) Gene expression level of CDC20 gene

Figure 4.10: Differences in gene expression levels between breast cancer subtypes. (a) Gene expression level of SHISA5 gene that has a high posterior probability of Basal subtype. Genes with a high posterior probability generally have little difference in expression level between subtypes, and thus have a low likelihood. (b) Gene expression level of CDC20 gene that shows a step-by-step difference in expression level by breast cancer subtype. CDC20 gene is included in PAM50.

4.3.2 Gene selection by combination of likelihood and posterior probability

Feature selection for subtype classification

For the breast cancer subtype classification problem, we used the PAM50 genes as features and predicted subtype of patients in the test data set using Support Vector Machine (SVM) model. The SVM model was trained using 870 out of 1091 breast cancer patients, and the performance of the classifier was evaluated using 221 patients' data as a test set. When predicted using the PAM50 genes as features, 185 out of 221 patients were correctly classified.

As shown in Figure 4.10, genes with a high posterior probability generally have little difference in expression levels between subtypes, and thus have a low likelihood. Therefore, there is not much ability to distinguish subtypes, which is why it is difficult to distinguish subtypes in t-SNE as Figure 4.9 shows. Instead, they are distinct from other subtypes so that we can consider other subtypes in one chunk. In that sense, genes having high posterior probability are really specific to that subtype.

Based on these differences in likelihood and posterior probability, we selected genes under various conditions and applied them to the subtype classification problem using the selected genes.

Criteria:

1. Select genes with high likelihood so as to represent a difference between subtypes

2. Select genes with high posterior probability so as to represent a specificity of each subtype

3. Select genes with $th_{pn} <= P(s_m|g_i) <= th_{px}$ in 3 or more subtypes so as to represent a common specificity to multiple subtypes

Table 4.4: Criteria of gene set selection

Condition	Criteria	Change step
Likelihood	$0.5 <= P(g_i\|s_m) <= 0.8$	0.05
Posterior probability	$0.25 <= P(s_m\|g_i) <= 0.4$	0.01
Common	$0.2 - \alpha <= P(s_m\|g_i) <= 0.2 + \alpha$	α step 0.01 to 0.05

Where th_{pn} and th_{px} refer to the minimum threshold and the maximum threshold in the posterior probability, respectively.

In order to select a gene set that satisfies the above criteria, we tried to select the gene set with the best performance in SVM under the conditions shown in Table 4.4. Total of 679 cases were tested, among which the top 4 conditions with the best SVM prediction performance are shown in Table 4.5.

The best gene set was GS2 selected from genes with a likelihood of 0.7 or higher and a posterior probability of 0.27 or higher, or a likelihood of 0.7 or higher and a posterior probability between 0.15 and 0.25 in three or more subtypes. The test accuracy was the highest and the number of genes was smallest. As can be seen from the results in Table 4.5, when comparing GS2 and GS4, the number of genes in GS2 is less than GS4 only satisfying the likelihood criteria, and there is a slight improvement in the accuracy of the SVM. When comparing GS4 and GS5, the likelihood condition is the same as 0.7 or more, but in the case of GS5 with the additional post-probability condition, the number of genes was reduced. But, it can be seen that the SVM prediction accuracy was not good as GS4.

In order to see how effective the gene set we chose was to predict the subtype of breast cancer patients, we trained it using the SVM model and predicted it with a test data set. Table 4.6 shows the results of comparing the PAM50 gene set with the best prediction results to predict the breast cancer

Table 4.5: Breast cancer subtype classification by SVM

Gene set	LE	PO	Common	genes	train	test
GS1	0.7	0.26	$0.15 \sim 0.25$	493	91.61% (797/870)	81.90% (181/221)
GS2	0.7	0.27	$0.15 \sim 0.25$	488	91.95% (800/870)	81.90% (181/221)
GS3	0.7	0.25	$0.15 \sim 0.25$	495	91.95% (800/870)	81.44% (180/221)
GS4	0.7			580	91.72% (798/870)	81.45% (180/221)
GS5	0.7	0.27		462	92.30% (803/870)	79.64% (176/221)

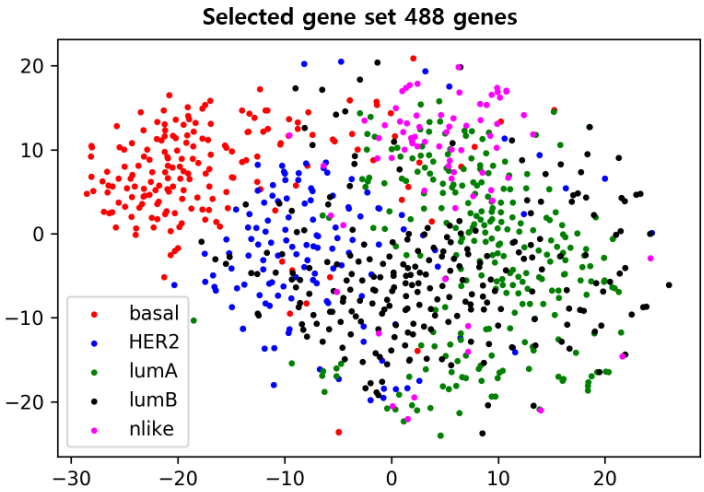
Table 4.6: Comparison of classification accuracy

Classifier	Random genes	Our method	PAM50
Features	488	488	50
Accuracy	75.56% (167/221)	81.9% (181/221)	84.61% (187/221)

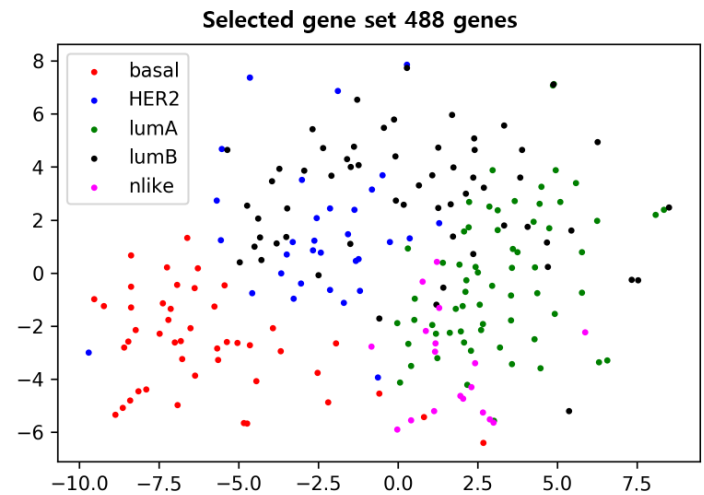
subtype. Also, for comparison, the same experiment was conducted on the random gene set 100 times and the results are shown in Table 4.6. Random genes are selected from genes in KEGG pathway.

Patients in the geometric space using selected gene set

We selected 488 genes according to the above conditions and represented the data of breast cancer patients using the selected genes. For the comparison, we showed t-SNE plot with the same number of random genes. Figure 4.11 represent the breast cancer patients using selected genes by our method. Figure 4.12 shows breast cancer patient using 488 randomly selected genes. In Figure 4.11, it is well distinguished between subtypes when compared to Figure 4.12. Figure 4.13 shows the patient data using PAM50 genes, and it is well classified for each subtype.



(a) Patients in training data set



(b) Patients in test data set

Figure 4.11: Breast cancer patients in the geometric space using selected gene set

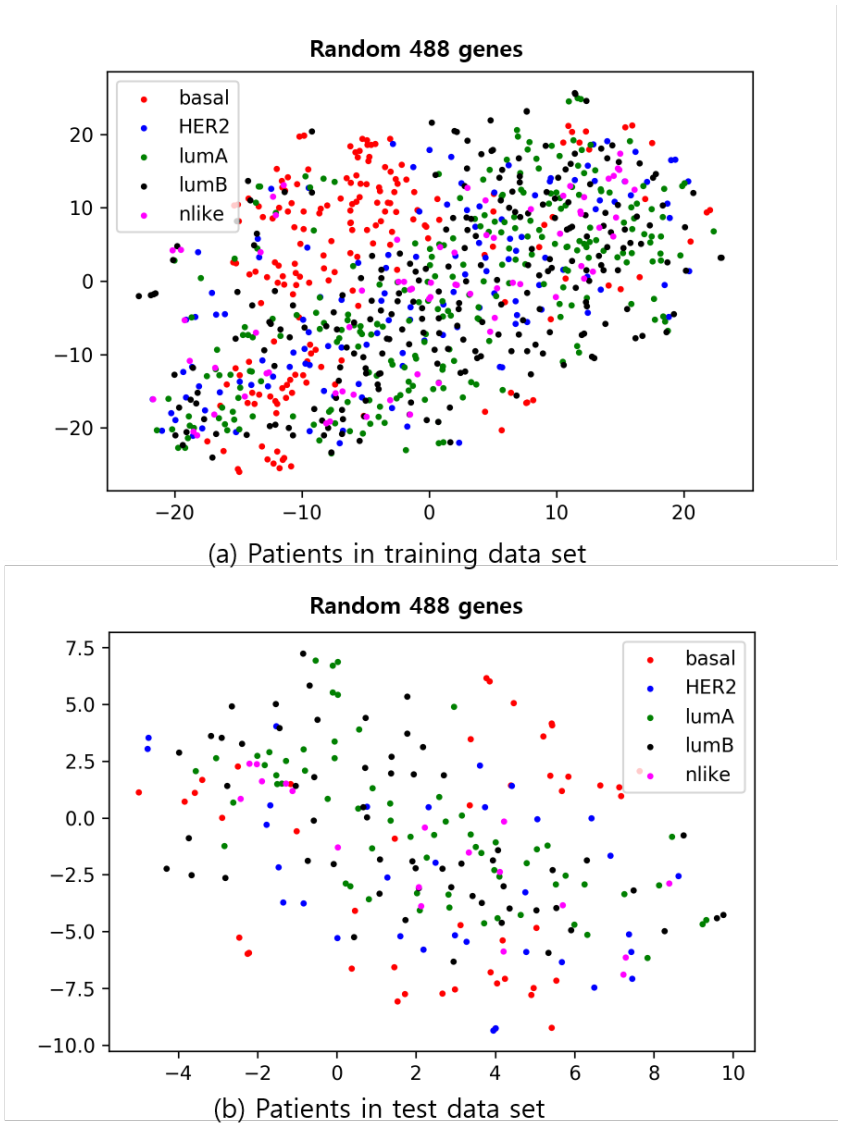


Figure 4.12: Breast cancer patients in the geometric space using random gene set

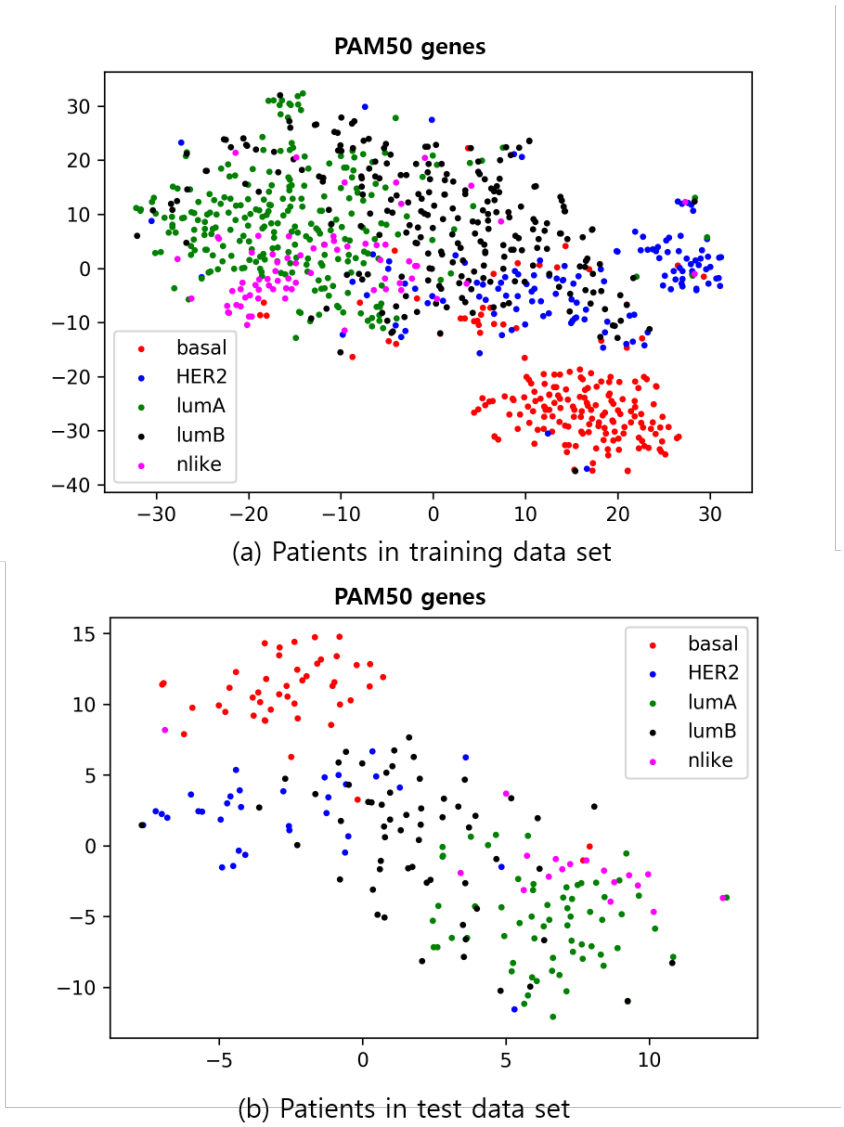


Figure 4.13: Breast cancer patients in the geometric space using PAM50 gene set

Difference in pathways among subtypes

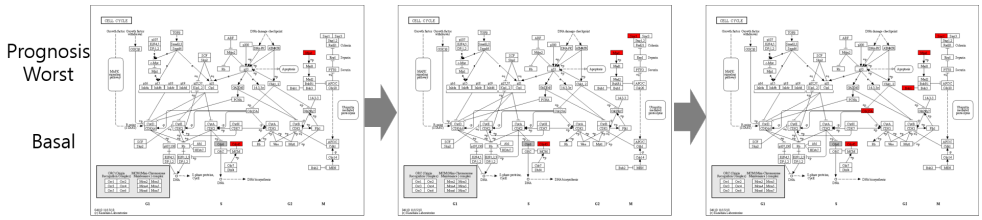
To show how activation of the important pathways can show difference between subtypes, we added genes by ranking and mapped gene expression level onto the corresponding pathways.

Figure 4.14 shows the gene expression level of each subtype patient mapped to Cell cycle pathway. Genes included in the pathway increase from left to right. We added the genes according to the rankings to see how the pathway changes. Cell cycle pathway is an important pathway in breast cancer, and the expression pattern of each subtype was shown using BRCA-Pathway in Figure 4.1. Figure 4.14 shows not only the difference in Cell cycle pathway among subtypes, but also which gene shows the most difference between subtypes and how the overall Cell cycle pathway changes as genes are added. Figure 4.15 shows how the expression levels differ for each subtype in Circadian rhythm pathway, another important pathway in breast cancer. As shown in Figure 4.14, Basal subtype with the worst prognosis is shown at the top, and LumA subtype with the best prognosis is at the bottom.

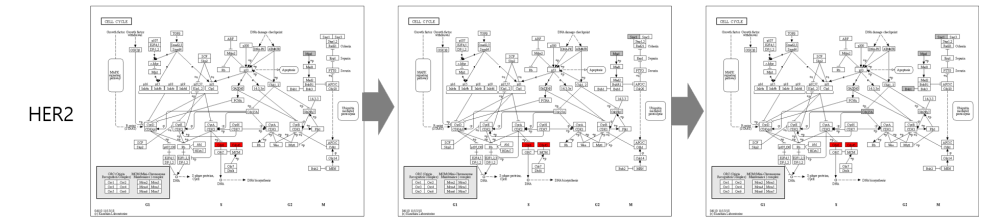
One of the processes regulated by the circadian clock is the cell cycle. Disruption of circadian rhythms can therefore be associated with abnormal cell divisions that occur in cancer (Filipski *et al.*, 2002). Indeed, there are links between altered circadian clocks and tumorigenesis in most notably, breast cancer (Filipski *et al.*, 2002).

Interestingly, the expression level of important genes in Circadian rhythm pathway is the opposite of the expression level in Cell cycle pathway (Figure 4.15). In other words, many genes in Cell cycle pathway of Basal subtype are overexpressed, whereas many genes in Circadian rhythm pathway are under-expressed. On the contrary, LumA subtype with best prognosis show under-expressed genes in Cell cycle pathway and overexpressed genes in Circadian rhythm pathway.

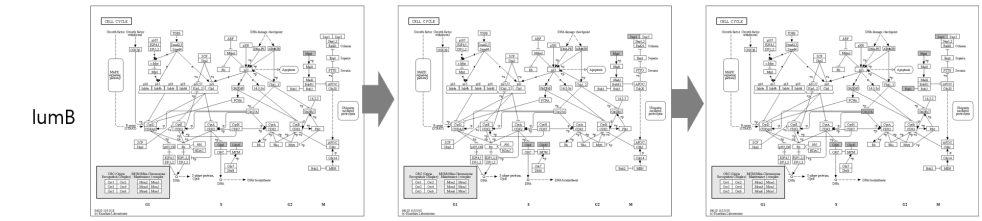
Other important pathways also show different gene expression levels for each subtype (Figure 4.16, 4.17, 4.18, 4.19, 4.20, 4.21, 4.22). And it is possible to know which gene included in the pathway can best distinguish the subtype. This is because colored genes on the left have higher likelihood values than genes on the right.



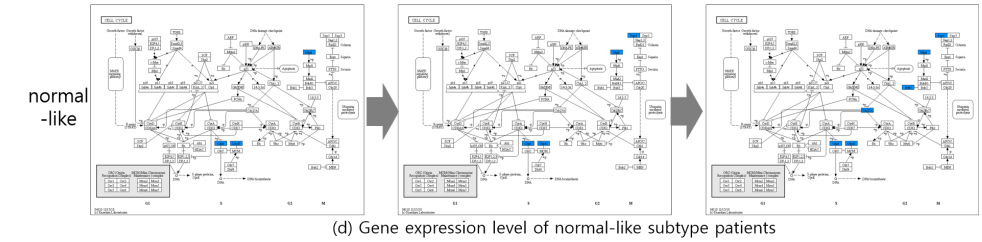
(a) Gene expression level of basal subtype patients



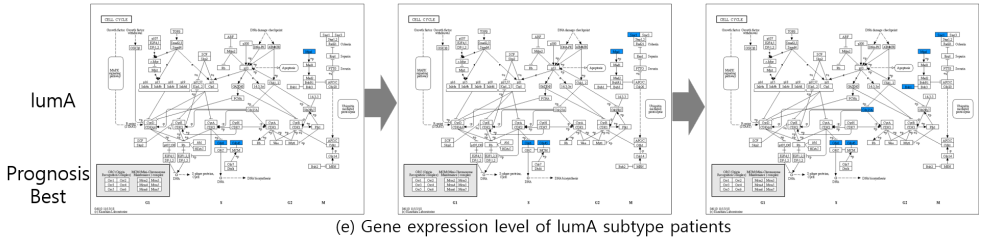
(b) Gene expression level of HER2 subtype patients



(c) Gene expression level of lumB subtype patients

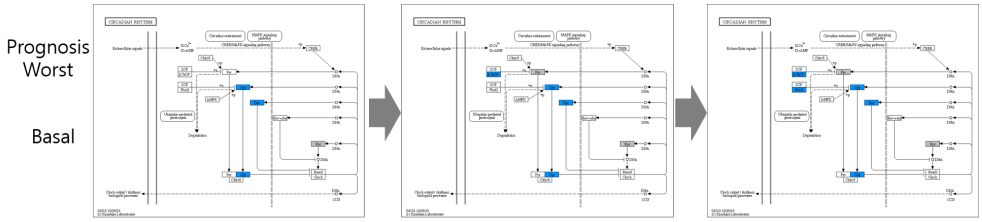


(d) Gene expression level of normal-like subtype patients

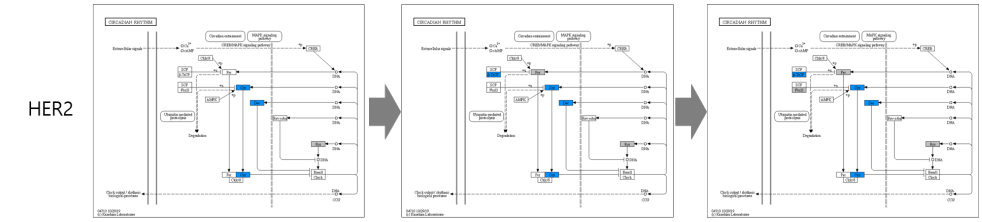


(e) Gene expression level of lumA subtype patients

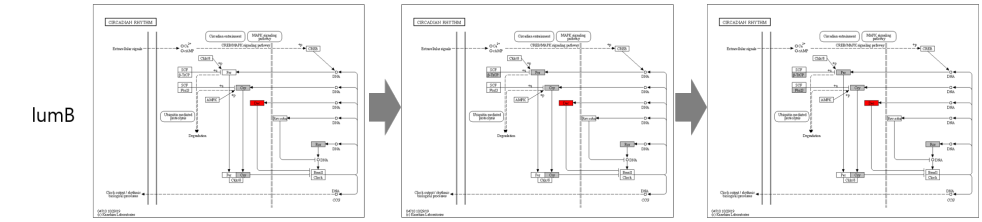
Figure 4.14: Difference between breast cancer subtypes on Cell cycle pathway



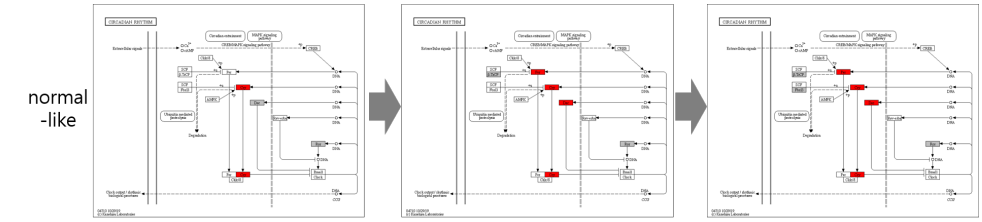
(a) Gene expression level of basal subtype patients



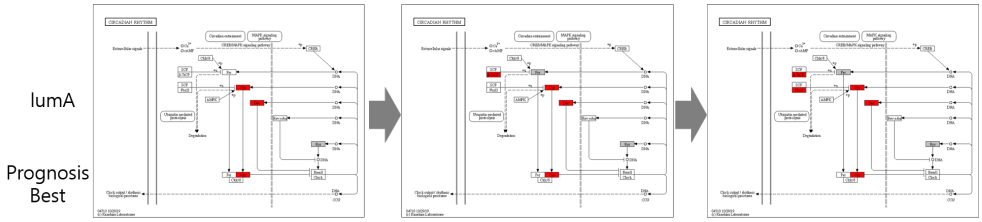
(b) Gene expression level of HER2 subtype patients



(c) Gene expression level of lumB subtype patients

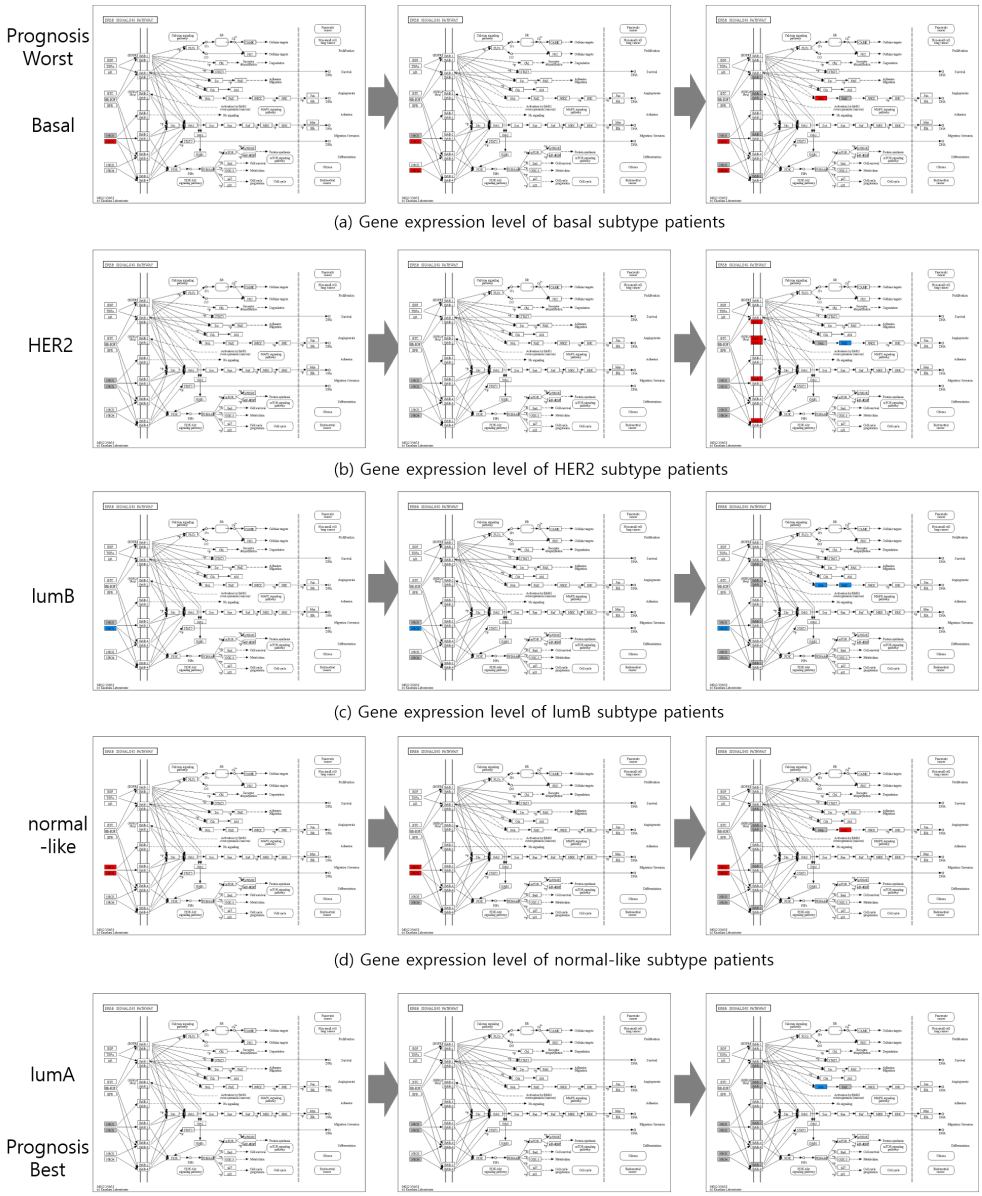


(d) Gene expression level of normal-like subtype patients



(e) Gene expression level of lumA subtype patients

Figure 4.15: Difference between breast cancer subtypes on Circadian rhythm pathway



(a) Gene expression level of basal subtype patients

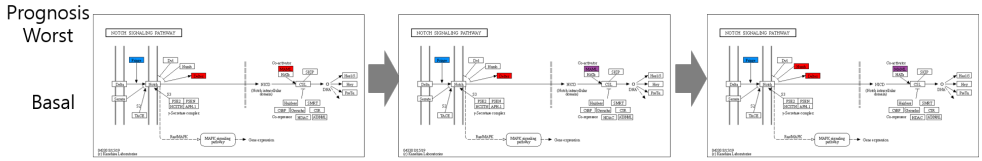
(b) Gene expression level of HER2 subtype patients

(c) Gene expression level of lumB subtype patients

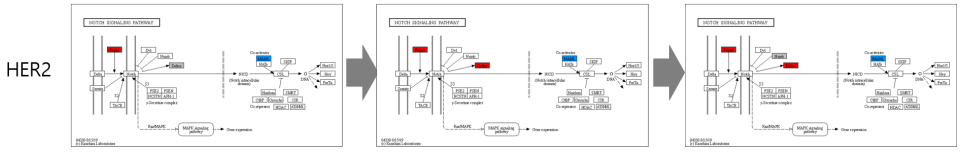
(d) Gene expression level of normal-like subtype patients

(e) Gene expression level of lumA subtype patients

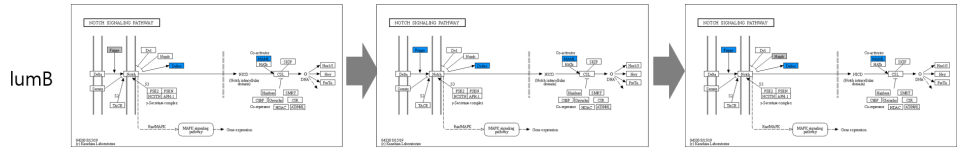
Figure 4.16: Difference between breast cancer subtypes on ERBB signaling pathway



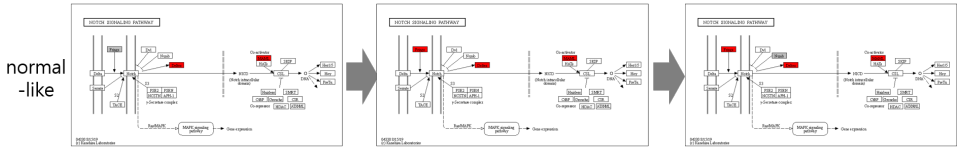
(a) Gene expression level of basal subtype patients



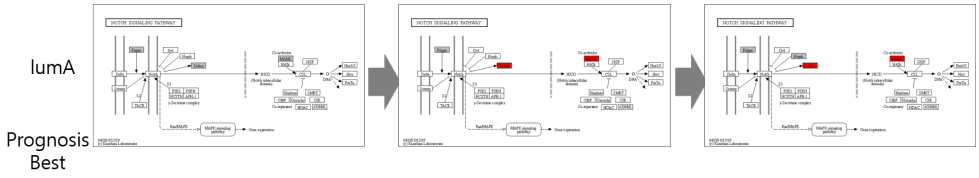
(b) Gene expression level of HER2 subtype patients



(c) Gene expression level of lumB subtype patients

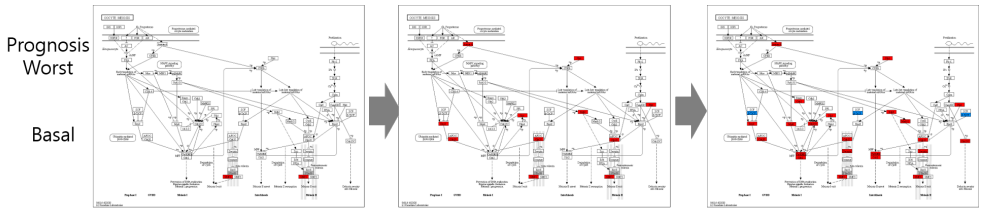


(d) Gene expression level of normal-like subtype patients

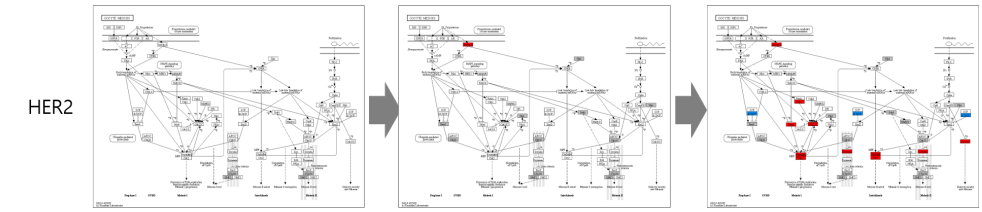


(e) Gene expression level of lumA subtype patients

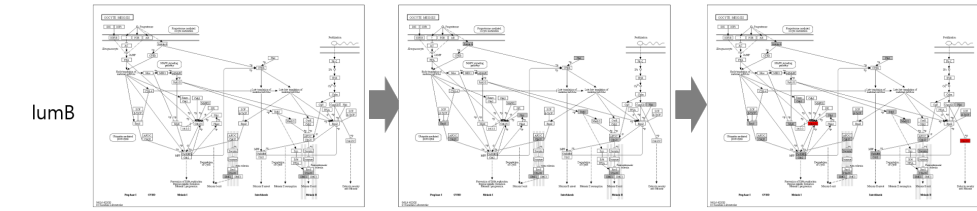
Figure 4.17: Difference between breast cancer subtypes on Notch signaling pathway



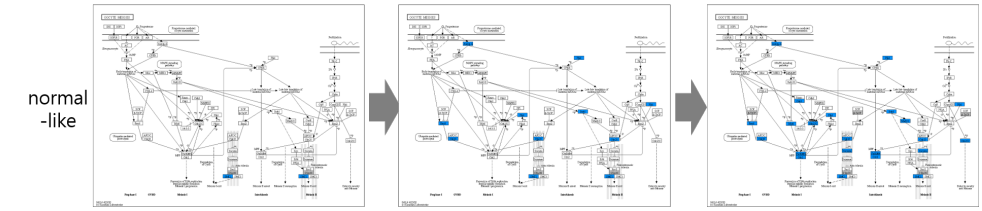
(a) Gene expression level of basal subtype patients



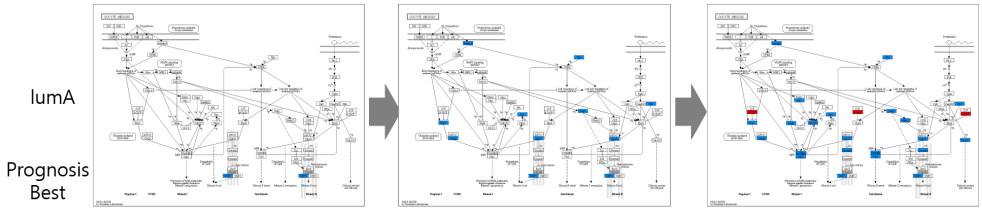
(b) Gene expression level of HER2 subtype patients



(c) Gene expression level of lumB subtype patients

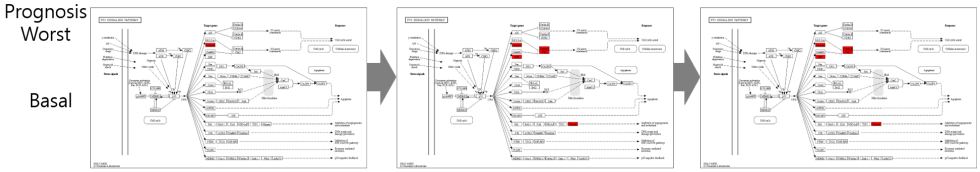


(d) Gene expression level of normal-like subtype patients

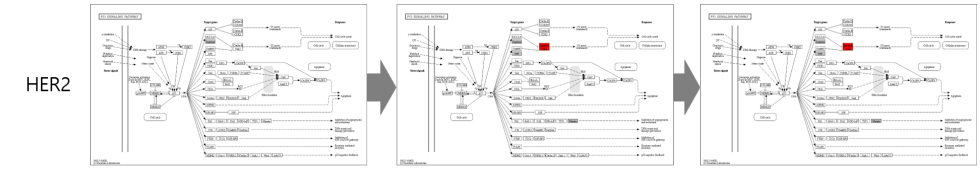


(e) Gene expression level of lumA subtype patients

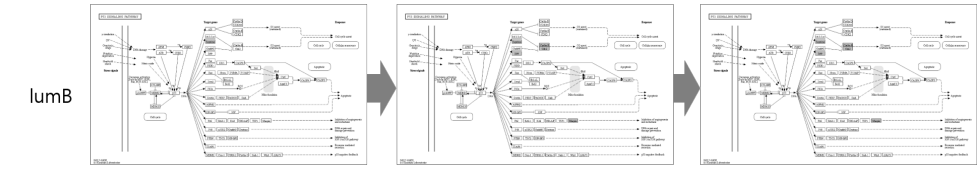
Figure 4.18: Difference between breast cancer subtypes on Oocyte meiosis pathway



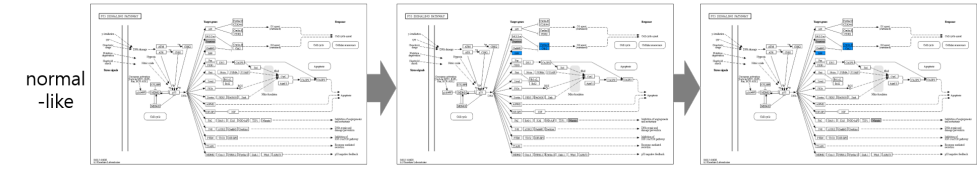
(a) Gene expression level of basal subtype patients



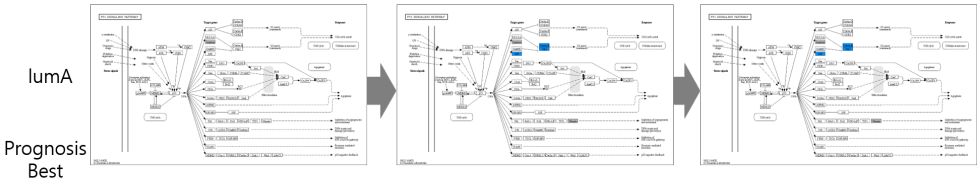
(b) Gene expression level of HER2 subtype patients



(c) Gene expression level of lumB subtype patients

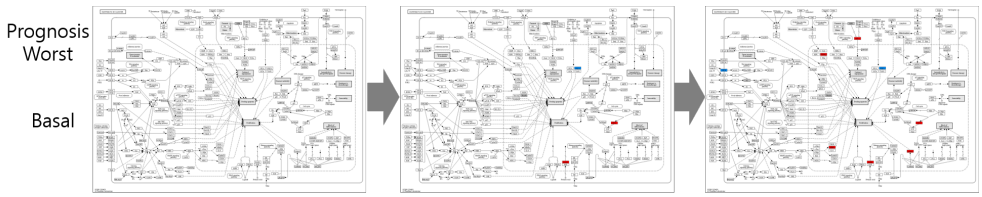


(d) Gene expression level of normal-like subtype patients

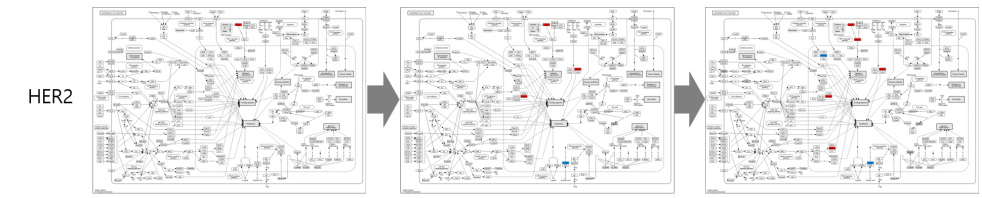


(e) Gene expression level of lumA subtype patients

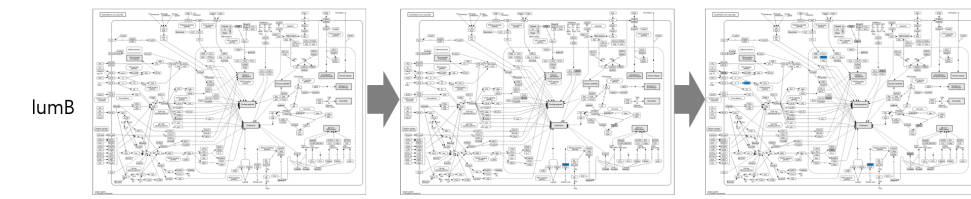
Figure 4.19: Difference between breast cancer subtypes on P53 signaling pathway



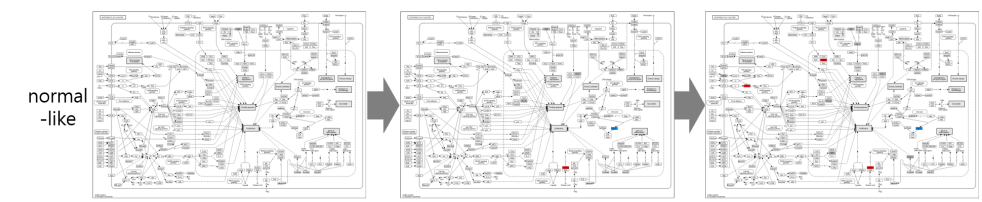
(a) Gene expression level of basal subtype patients



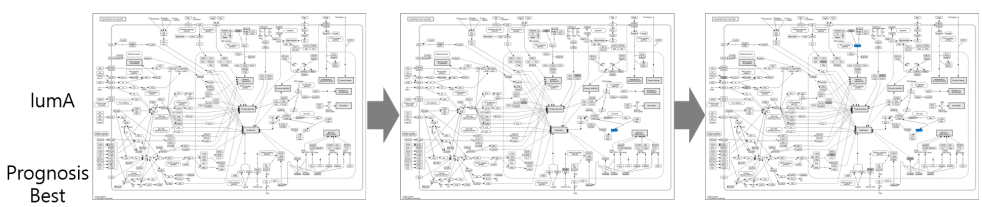
(b) Gene expression level of HER2 subtype patients



(c) Gene expression level of lumB subtype patients

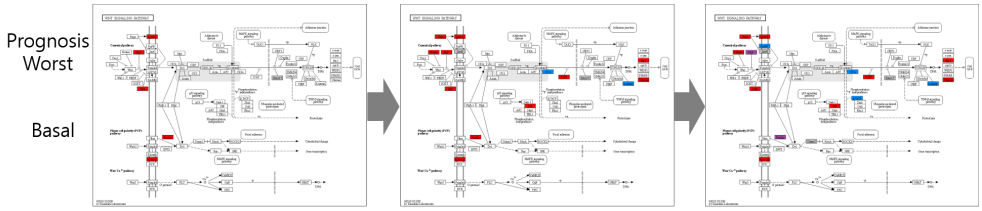


(d) Gene expression level of normal-like subtype patients

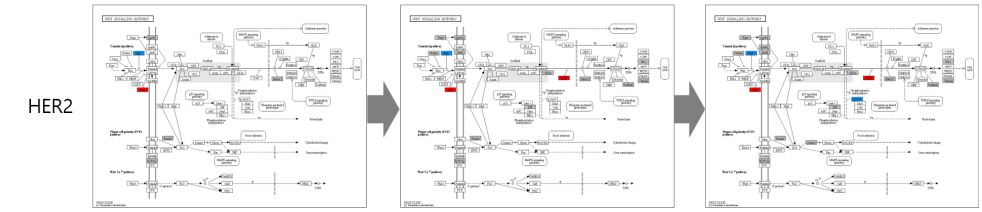


(e) Gene expression level of lumA subtype patients

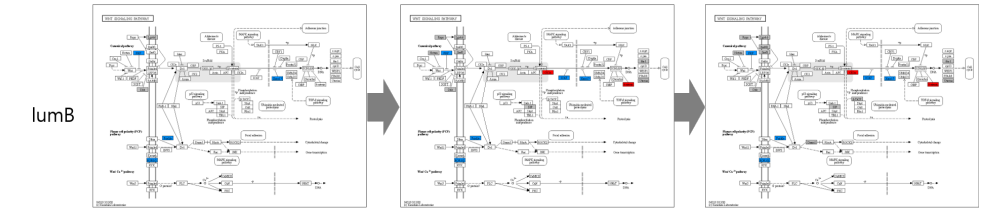
Figure 4.20: Difference between breast cancer subtypes on Pathways in cancer pathway



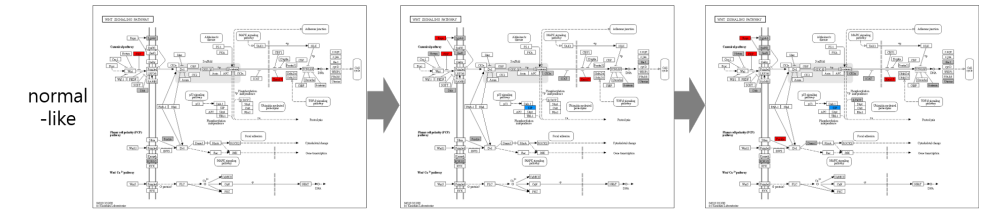
(a) Gene expression level of basal subtype patients



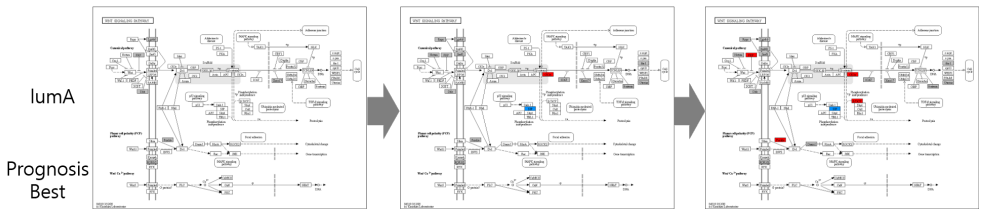
(b) Gene expression level of HER2 subtype patients



(c) Gene expression level of lumB subtype patients

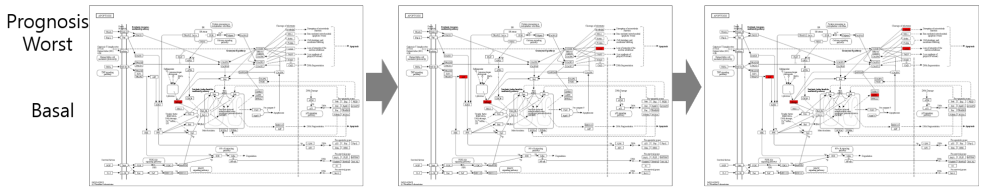


(d) Gene expression level of normal-like subtype patients

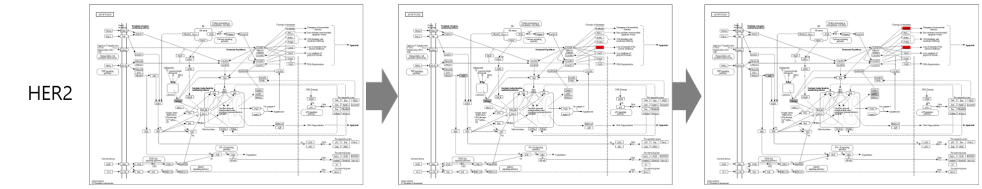


(e) Gene expression level of lumA subtype patients

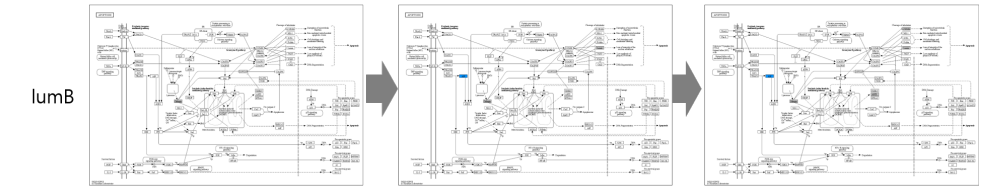
Figure 4.21: Difference between breast cancer subtypes on Wnt signaling pathway



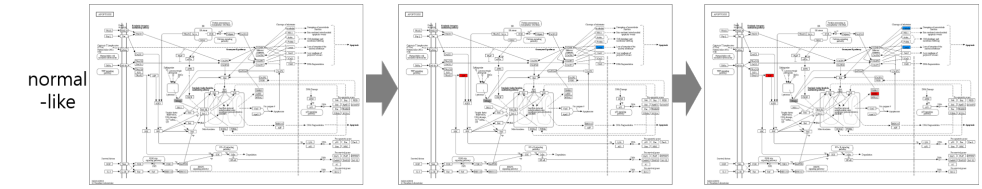
(a) Gene expression level of basal subtype patients



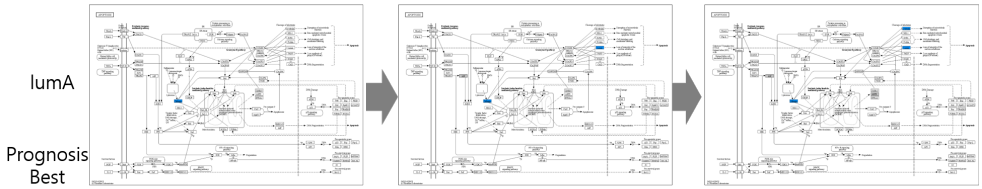
(b) Gene expression level of HER2 subtype patients



(c) Gene expression level of lumB subtype patients



(d) Gene expression level of normal-like subtype patients



(e) Gene expression level of lumA subtype patients

Figure 4.22: Difference between breast cancer subtypes on Apoptosis pathway

4.3.3 Biological meaning of the selected gene set

In order to find out biological meaning of the genes, we analyzed the results in the context of KEGG pathway and Gene ontology using Enrichr (Chen *et al.*, 2013; Kuleshov *et al.*, 2016).

Gene set by likelihood

We selected top 20 genes for each 5 of the subtypes by likelihood (Table 4.2), and 87 genes are left out of the 100 genes collected, excluding duplicates. We performed a gene set enrichment analysis using 87 genes.

Pathway enrichment analysis: There have been many studies on the relationship between breast cancer and circadian rhythm. Disrupted expression of circadian genes can alter breast biology and may promote cancer (Blakeman *et al.*, 2016). One of the processes regulated by the circadian clock is the cell cycle. Disruption of circadian rhythms can therefore be associated with abnormal cell divisions that occur in cancer (Filipski *et al.*, 2002). Indeed, there are links between altered circadian clocks and tumorigenesis in metastatic colorectal cancer, osteosarcoma, pancreatic adenocarcinoma and, most notably, breast cancer (Filipski *et al.*, 2002).

Table 4.7 (a) shows a result of enrichment analysis of 87 genes, the KEGG pathway enrichment analysis. Among the top ranked pathways, we mapped the expression levels of genes belong to Circadian rhythm pathway. Since each gene expression value was converted to z-score, if the z score of a gene in a patient is negative, it means that the expression level of that gene is lower than the average expression level of the gene. And if it is positive, the expression level is higher than the average. It is colored in blue when the average z score of patients of a specific subtype is less than -0.3, red when the average z score is greater than 0.3, and gray when the average z score has a value between

-0.3 and 0.3. Therefore, the blue color indicates that the expression level is low compared to the average expression level of all breast cancer patients, and the red color indicates that the expression level is higher than the average expression level of all breast cancer patients.

Figure 4.23 shows Circadian rhythm pathway. Of the 87 genes selected by each likelihood, 5 genes belong to Circadian rhythm pathway, (a) is the gene expression level of patients corresponding to Basal subtype, and (b) is the gene expression level of patients corresponding to LumA subtype. What this figure shows is that the genes selected by the likelihood show well the differences between subtypes at the pathway level. The results in Table 4.7 (a) show that Circadian rhythm pathway and Cell cycle pathway were listed together, which is consistent with the previous study mentioned before, that is, circadian rhythm is closely related to the progression of breast cancer and particularly affects cell cycle (Blakeman *et al.*, 2016).

Gene Ontology enrichment analysis: Table 4.7 (b) shows the GO enrichment analysis of 20 Basal subtype genes selected by likelihood. Negative regulation of intrinsic apoptotic signaling pathway is ranked at the top. Apoptosis is a tightly regulated cell suicide program that plays an essential role in the maintenance of tissue homeostasis by eliminating unnecessary or harmful cells. Defects in this native defense mechanism promote malignant transformation and frequently confer chemoresistance to transformed cells. Indeed, the evasion of apoptosis has been recognized as a hallmark of cancer (Plati *et al.*, 2008). The result is consistent with the poor prognosis in Basal subtype patients.

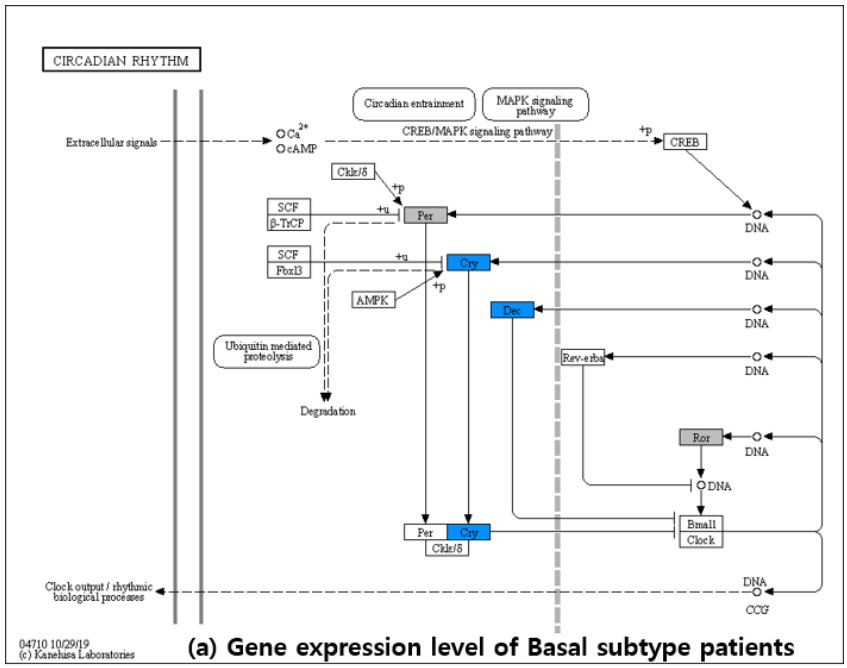
Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	Circadian rhythm	2.155e-7	0.00006637	37.08	569.17
2	One carbon pool by folate	0.000001534	0.0002363	45.98	615.52
3	Huntington disease	0.00002128	0.002185	8.34	89.69
4	NOD-like receptor signaling pathway	0.001082	0.08333	6.46	44.10
5	Notch signaling pathway	0.001194	0.07354	14.37	96.70
6	Cell cycle	0.002107	0.1082	7.42	45.70
7	Renin-angiotensin system	0.004460	0.1962	19.99	108.20
8	Butanoate metabolism	0.006570	0.2529	16.42	82.52
9	HIF-1 signaling pathway	0.009491	0.3248	6.90	32.12
10	Tyrosine metabolism	0.01071	0.3298	12.77	57.94

(a) KEGG pathway enrichment analysis of 87 union genes by likelihood

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	negative regulation of intrinsic apoptotic signaling pathway (GO:2001243)	0.00003115	0.1589	48.39	502.11
2	negative regulation of embryonic development (GO:0045992)	0.00004255	0.1086	200.00	2012.99
3	negative regulation of endoplasmic reticulum stress-induced intrinsic apoptotic signaling pathway (GO:1902236)	0.00008583	0.1460	142.86	1337.59
4	regulation of endoplasmic reticulum stress-induced intrinsic apoptotic signaling pathway (GO:1902235)	0.0001786	0.2278	100.00	863.06
5	entrainment of circadian clock by photoperiod (GO:0043153)	0.0001786	0.1822	100.00	863.06
6	photoperiodism (GO:0009648)	0.0001972	0.1678	95.24	812.48
7	negative regulation of response to endoplasmic reticulum stress (GO:1903573)	0.0002168	0.1581	90.91	766.94
8	regulation of calcium ion transport into cytosol (GO:0010522)	0.0002168	0.1383	90.91	766.94
9	positive regulation of T cell differentiation (GO:0045582)	0.0004628	0.2624	62.50	479.89
10	response to amino acid (GO:0043200)	0.0005228	0.2668	58.82	444.49

(b) Gene Ontology analysis of basal 20 genes by likelihood

Table 4.7: Gene set enrichment analysis by Enrichr



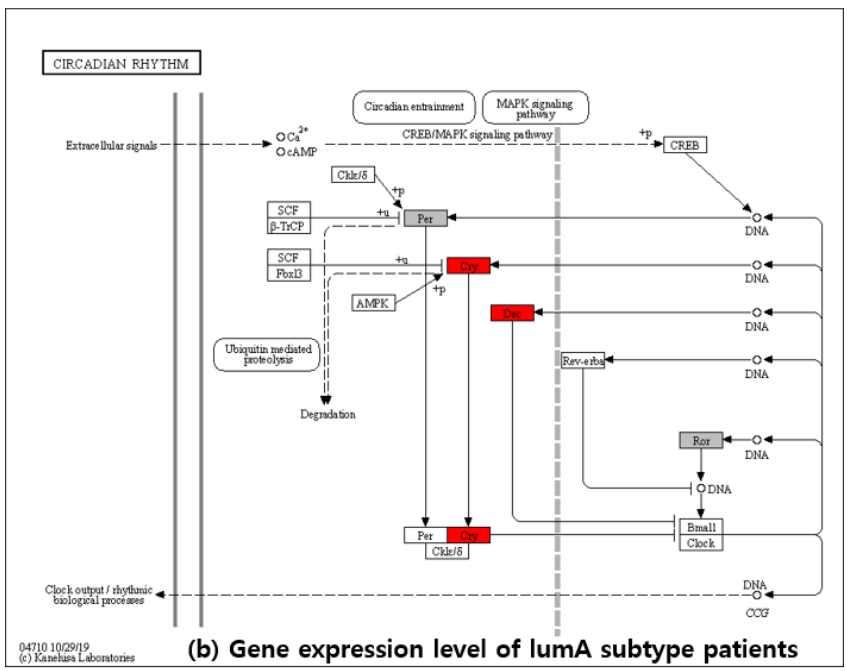


Figure 4.23: Difference in gene expression level in Basal subtype and LumA subtype on Circadian rhythm pathway

Gene set by posterior probability

We selected top 20 genes for Basal subtype by posterior probability (Table 4.3), and performed a gene set enrichment analysis using this Basal subtype specific genes. As shown in Table 4.8,

Pathway enrichment analysis: Table 4.8 (a) shows the KEGG pathways containing genes specific to Basal subtype. There are many researches that have studied the activation of nuclear factor (NF)-κB signaling in triple negative breast cancer (TNBC) (Pan *et al.*, 2012; Zhu *et al.*, 2013). And the aberrant activation of nuclear factor (NF)-κB signaling is a frequent characteristic of TNBCs (Poma *et al.*, 2017). The top ranked NF-KB pathway is consistent with the previous studies.

Gene Ontology enrichment analysis: Negative regulation of intrinsic apoptotic signaling pathway is ranked at the top. Apoptosis is a tightly regulated cell suicide program that plays an essential role in the maintenance of tissue homeostasis by eliminating unnecessary or harmful cells. Defects in this native defense mechanism promote malignant transformation and frequently confer chemoresistance to transformed cells. Indeed, the evasion of apoptosis has been recognized as a hallmark of cancer (Plati *et al.*, 2008). The result is consistent with the poor prognosis in Basal subtype patients.

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	NF-kappa B signaling pathway	0.0001116	0.03438	31.58	287.38
2	Glycerophospholipid metabolism	0.0001187	0.01829	30.93	279.54
3	Herpes simplex virus 1 infection	0.001281	0.1316	8.13	54.14
4	Lysine degradation	0.001571	0.1210	33.90	218.85
5	p53 signaling pathway	0.002328	0.1434	27.78	168.40
6	Parkinson disease	0.008746	0.4490	14.08	66.75
7	Apoptosis	0.008865	0.3901	13.99	66.09
8	Epstein-Barr virus infection	0.01695	0.6527	9.95	40.57
9	Fatty acid degradation	0.04311	1.000	22.73	71.45
10	Malaria	0.04790	1.000	20.41	62.01

(a) KEGG pathway enrichment analysis of 20 basal genes by posterior probability

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	regulation of cysteine-type endopeptidase activity involved in apoptotic signaling pathway (GO:2001267)	0.00003406	0.1738	222.22	2286.11
2	negative regulation of oxidative stress-induced intrinsic apoptotic signaling pathway (GO:1902176)	0.0001440	0.3673	111.11	982.88
3	histone lysine methylation (GO:0034968)	0.0002168	0.3688	90.91	766.94
4	negative regulation of oxidative stress-induced cell death (GO:1903202)	0.0002588	0.3301	83.33	688.30
5	regulation of autophagy of mitochondrion (GO:1903146)	0.0003285	0.3353	74.07	594.15
6	positive regulation of sequence-specific DNA binding transcription factor activity (GO:0051091)	0.001220	1.000	13.95	93.61
7	negative regulation of intrinsic apoptotic signaling pathway (GO:2001243)	0.001733	1.000	32.26	205.09
8	regulation of neuron death (GO:1901214)	0.002328	1.000	27.78	168.40
9	negative regulation of cell death (GO:0060548)	0.002936	1.000	24.69	143.97
10	negative regulation of neuron death (GO:1901215)	0.003080	1.000	24.10	139.35

(b) Gene Ontology analysis of 20 basal genes by posterior probability

Table 4.8: Gene set enrichment analysis of 20 Basal subtype genes by posterior probability. There are many researches that have studied the activation of nuclear factor (NF)-κB signaling in triple negative breast cancer (TNBC). And the aberrant activation of nuclear factor (NF)-κB signaling is a frequent characteristic of TNBCs. Top ranked NF-KB pathway is consistent with the previous studies.

4.4 Discussion

XGBoost is an effective approach in classification, which produces a prediction model in an ensemble of weak models, typically decision trees (Chen and Guestrin, 2016). We evaluated the XGBoost model in subtype classification of breast cancer patients. The prediction accuracy was 86.87% by matching 192 subtypes out of 221 of the total test patients. The accuracy of the model was higher than that obtained using the PAM50 features. However, it can be seen that the accuracy is much superior because the prediction is performed using the results ensembled by multiple decision trees, rather than classifying them using a specific features like PAM50 gene set. To compare the prediction accuracy in the context of feature selection, we performed the experiment as shown in Table 4.10 using gene set by feature importance provided by XGBoost. Feature importance for 625 genes was provided and subtype classification was performed with genes having higher importance (Figure 4.24). As can be seen in Table 4.10, the best accuracy was lower than the PAM50 gene set and was similar to our method.

Rhee et al. presented a high accuracy model in predicting subtypes of breast cancer patients using gene expression data (Rhee *et al.*, 2017). They utilized protein-protein interaction (PPI) network to represent the interaction between genes and applied graph convolutional neural network (graph CNN) to learn the localized gene expression patterns of cooperative gene cluster. Then, Relational Network (RN) was applied to infer the relationship between previously learned expression patterns. This hybrid approach showed an accuracy of 86.29%. However, this study is not a feature selection model, but a classification model using PPI network as well as gene expression data. Internally, each gene is represented as a latent vector by the expression of it's neighbor genes on the PPI.

Table 4.9: Comparison of classification accuracy

Classifier	Random genes	Our method	PAM50	XGBoost	Rhee et al.
Features	488	488	50		
Accuracy	75.56%	81.9%	84.61%	86.87%	86.29%

Table 4.10: Subtype classification using XGBoost features

Gene set	Features	Accuracy
XG_{10}	10	78.73% (174/221)
XG_{30}	30	79.63% (176/221)
XG_{50}	50	80.09% (177/221)
XG_{100}	100	82.35% (182/221)
XG_{200}	200	81.44% (180/221)
XG_{300}	300	80.09% (177/221)
XG_{400}	400	80.54% (178/221)
XG_{500}	500	81.90% (181/221)
XG_{600}	600	81.90% (181/221)
XG_{625}	625	81.44% (180/221)

The biggest limitation of our method is that we select a gene set for only one third of all genes. Nevertheless, classification performance with our gene set was slightly lower than that of the gold standard PAM50, and much better than randomly selected gene set. In addition, although the accuracy of the subtype prediction was lower than the prediction in XGBoost, which is often used for its excellent performance, it is difficult to directly compare with our gene set because XGBoost does not perform feature selection.

The significance of our gene set is that it is a biologically interpretable gene set. Because gene selection was performed within the KEGG pathway, a domain knowledge of biological mechanism.

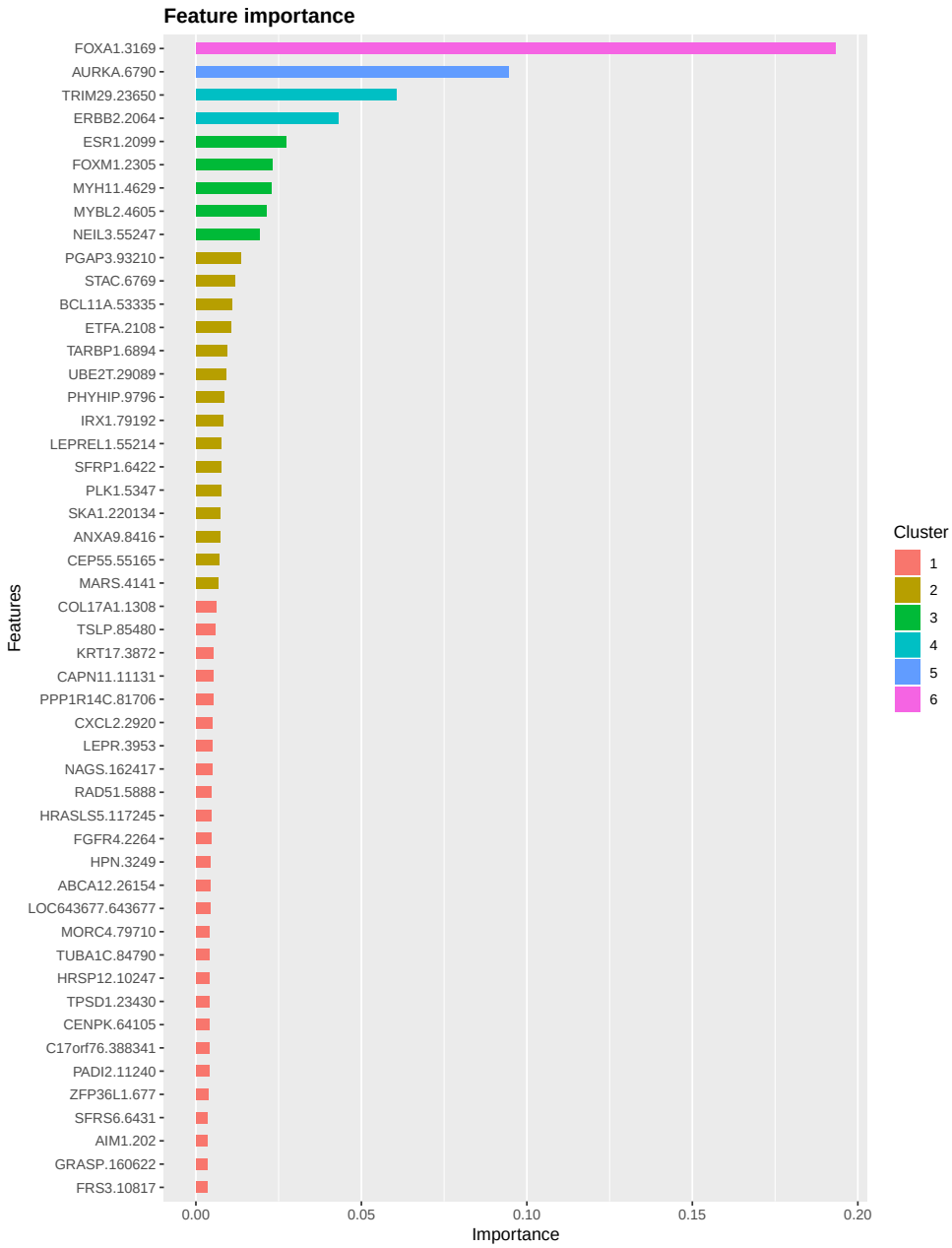


Figure 4.24: Feature importance for classification of breast cancer subtypes by XGBoost.

Table 4.11: Gene factor of PAM50 genes

Gene	Basal	HER2	lumA	lumB	nlike
ACTR3B	0.61	0.29	0.3	0.24	0.28
BCL2	0.42	0.42	0.4	0.34	0.36
BIRC5	0.54	0.42	0.59	0.44	0.47
BLVRA	0.46	0.2	0.18	0.21	0.21
CCNB1	0.4	0.35	0.52	0.36	0.48
CCNE1	0.61	0.49	0.54	0.48	0.61
CDC20	0.67	0.49	0.54	0.44	0.48
CDC6	0.49	0.56	0.57	0.44	0.49
CDH3	0.57	0.33	0.43	0.52	0.39
EGFR	0.81	0.47	0.29	0.43	0.51
ERBB2	0.4	0.62	0.4	0.34	0.35
ESR1	0.69	0.48	0.48	0.54	0.44
FGFR4	0.29	0.49	0.21	0.23	0.21
KRT14	0.7	0.58	0.59	0.41	0.62
KRT17	0.58	0.49	0.48	0.62	0.55
MAPT	0.6	0.43	0.48	0.39	0.39
MDM2	0.38	0.27	0.29	0.42	0.25
MYC	0.32	0.29	0.24	0.22	0.29
NAT1	0.7	0.38	0.42	0.34	0.39
ORC6L	0.58	0.44	0.55	0.43	0.53
PGR	0.47	0.38	0.36	0.29	0.35
PHGDH	0.7	0.4	0.42	0.39	0.42
PTTG1	0.48	0.39	0.53	0.36	0.43
RRM2	0.41	0.45	0.52	0.39	0.47
SFRP1	0.64	0.56	0.54	0.7	0.57
TYMS	0.49	0.32	0.45	0.35	0.39

Table 4.12: Likelihood of PAM50 genes

Gene	Basal	HER2	lumA	lumB	nlike
ACTR3B	0.64	0.62	0.62	0.24	0.09
BCL2	0.39	0.48	0.45	0.47	0.39
BIRC5	0.54	0.42	0.51	0.37	0.21
BLVRA	0.88	0.48	0.02	0.14	0.32
CCNB1	0.44	0.4	0.7	0.56	0.71
CCNE1	0.41	0.5	0.5	0.59	0.48
CDC20	0.4	0.28	0.74	0.81	0.76
CDC6	0.5	0.28	0.98	0.97	0.76
CDH3	0.74	0.34	0.3	0.8	0.01
EGFR	0.45	0.55	0.39	0.46	0.43
ERBB2	0.38	0.72	0.54	0.48	0.21
ESR1	0.33	0.64	0.35	0.42	0.43
FGFR4	0.24	0.43	0.16	0.25	0.19
KRT14	0.41	0.36	0.59	0.29	0.74
KRT17	0.39	0.35	0.57	0.32	0.73
MAPT	0.58	0.41	0.39	0.33	0.39
MDM2	0.31	0.37	0.34	0.44	0.22
MYC	0.27	0.31	0.26	0.26	0.26
NAT1	0.49	0.51	0.35	0.33	0.54
ORC6L	0.61	0.27	0.96	0.94	0.82
PGR	0.46	0.57	0.54	0.46	0.67
PHGDH	0.84	0.56	0.84	0.63	0.5
PTTG1	0.33	0.25	0.72	0.72	0.71
RRM2	0.31	0.64	0.5	0.37	0.46
SFRP1	0.69	0.25	0.32	1	0.82
TYMS	0.52	0.72	0.88	0.5	0.6

Table 4.13: Posterior probability of PAM50 genes

Gene	Basal	HER2	lumA	lumB	nlike
ACTR3B	0.27	0.2	0.38	0.14	0.02
BCL2	0.18	0.17	0.3	0.29	0.07
BIRC5	0.25	0.14	0.34	0.23	0.04
BLVRA	0.56	0.23	0.02	0.12	0.08
CCNB1	0.16	0.11	0.36	0.27	0.1
CCNE1	0.16	0.15	0.29	0.32	0.08
CDC20	0.13	0.07	0.35	0.35	0.1
CDC6	0.13	0.06	0.38	0.35	0.08
CDH3	0.29	0.1	0.17	0.43	0
EGFR	0.2	0.19	0.26	0.28	0.08
ERBB2	0.16	0.22	0.32	0.26	0.03
ESR1	0.16	0.23	0.24	0.28	0.08
FGFR4	0.2	0.27	0.19	0.28	0.06
KRT14	0.19	0.12	0.38	0.17	0.13
KRT17	0.18	0.12	0.38	0.2	0.13
MAPT	0.28	0.15	0.28	0.22	0.08
MDM2	0.17	0.16	0.28	0.34	0.05
MYC	0.2	0.18	0.28	0.26	0.08
NAT1	0.24	0.19	0.25	0.22	0.11
ORC6L	0.16	0.05	0.37	0.33	0.09
PGR	0.18	0.17	0.31	0.24	0.1
PHGDH	0.24	0.12	0.34	0.24	0.06
PTTG1	0.12	0.07	0.37	0.35	0.1
RRM2	0.14	0.22	0.33	0.23	0.08
SFRP1	0.23	0.06	0.15	0.45	0.11
TYMS	0.16	0.17	0.39	0.21	0.07

Chapter 5

Conclusions

In my doctoral study, I conducted the following studies by integrating the breast cancer multi-omics data and the KEGG pathway.

1. a structural integration of TCGA breast cancer data and an interactive visualization on KEGG pathway

2. a probabilistic framework for biologically explainable gene set selection

In the first study, I developed a system that maps TCGA breast cancer multi-omics data onto the KEGG pathway to integrate and visualize breast cancer patient data. Cancer scientists need to utilize multi-omics data fully to gain more insight into biological mechanisms in cancer. For example, when conducting research using transcriptome data, a higher level of abstraction is required than simply measuring the expression level of each gene in order to understand the biological meaning. Therefore, mapping gene expression data to the KEGG pathway is a widely used approach in bioinformatics. There have been quite a number of studies on developing systems that automatically map gene expression data to the KEGG pathway. However, existing tools are not

powerful enough to fully utilize multi-omics data in an integrated fashion for cancer research. In addition, pathway-based multiomics analysis systems have difficulties due to larger dimensions and small sample sizes. For this reason, we developed BRCA-Pathway, a web-based interactive exploration and visualization system of TCGA breast cancer data on KEGG pathway to provide a broad perspective of TCGA breast cancer data.

The data provided by BRCA-Pathway includes multi-omics data TCGA such as gene expression data, Copy Number Variation (CNV) data, mutation data, and clinical information of patients (22 features including age, race, hormone receptor information by IHC). Also, cancer subtype of a patient is provided using the PAM50 method. In addition, the hallmark gene set known to be important in cancer research, a transcription factor database that regulates gene expression, and a driver gene database provided by Cancer gene census are also provided. It is possible to freely select a patient group using clinical information and subtype information of breast cancer patients. BRCA-Pathway provides a variety of functions such as survival plot and mutual exclusivity sort, as well as visualization of data, which is of great help to breast cancer researchers. In addition, BRCA-Pathway provides user data visualization onto KEGG pathway. And data contained in the BRCA-Pathway can be downloaded using the REST API.

Using BRCA-Pathway, it was shown that the multi-omics data of breast cancer patients appeared differently for each subtype at the pathway level. In particular, gene expression data could identify different expression patterns for each subtype in several biologically important pathways such as Cell cycle pathway.

I conducted the second study to provide a solution to the problem of selecting genes specific to the subtype using pathway activity level for each breast cancer subtype. Differences in expression patterns at the pathway level were

transformed into pathway activation scores using methods by previous study. The difference in gene expression level for each subtype was quantified by unique area under probability density function and defined as a Gene factor. The difference in the degree of pathway activation for each subtype was defined as a Pathway factor in a similar way to Gene factor definition. Likelihood and posterior probability were defined using Gene factor and Pathway factor, and important genes for each subtype were selected based on these two probabilities. The selected genes were also analyzed in terms of biological meaning and used as features for subtype classification.

I developed a probabilistic framework based on the Bayesian approach. Genes with high likelihood are the genes with large differences in expression level for each subtype. Therefore, when t-SNE visualization was performed using genes with high likelihood, it was clear that the distinction between subtypes was good. However, when t-SNE visualization using genes with high posterior probability, the distinction between subtypes was not good. Since the posterior probability is obtained by normalizing each likelihood as the sum of the total likelihood, the posterior probability is high when the likelihood in other subtype is low and the likelihood in that subtype is relatively high. Therefore, other subtypes are not distinguished enough to be regarded as one subtype, and in the case of a gene in which only the subtype has a difference in expression level, the posterior probability has a high value. Therefore, a gene with a high posterior probability can be said to be a specific gene only for the subtype.

Therefore, in order to increase accuracy in the subtype classification problem, genes were selected by considering both criteria. I selected genes that show difference in the expression level among each subtypes by measuring likelihood and genes specific to each subtype be computing posterior probabilities. According to our criteria, 488 genes were selected, and the results were

significantly higher than the prediction accuracy using 488 random genes and slightly lower than the prediction accuracy using a manually selected gene set, PAM50.

To show how well the selected genes explain the difference between each subtype in the pathway level, we mapped the gene expression level for each subtype on the important pathways in breast cancer. The results were consistent with previous research that disruption of circadian rhythm can be associated with abnormal cell divisions that occur in cancer.

The expression level in Circadian rhythm pathway is the opposite of the expression level in Cell cycle pathway, and this tendency of the two pathways showing reverse gene expression levels was related to the subtype. In Basal subtype with poor prognosis, Circadian rhythm pathway was underexpressed and Cell cycle pathway was overexpressed. In LumA subtype with best prognosis, Circadian rhythm pathway was overexpressed and Cell cycle pathway was underexpressed.

In this thesis, I presented a machine-learning method that enables comparison between subtypes of breast cancer by interpreting biological information contained in KEGG pathway and multi-omics data. Omics data is high-dimensional data with a significantly smaller number of samples than observed features. Biological knowledge and effective machine learning methods are required to extract meaningful information from such data. Therefore, this thesis aimed to reduce the high-dimensional features of breast cancer data using the biological knowledge such as KEGG pathway database. The first study was to map the breast cancer multi-omics to the pathway and show differences between breast cancer subtypes through the pathway, which is a higher level interpretation that includes biological meanings rather than individual gene level. The second study used a probabilistic framework to select gene set specific to each subtype of breast cancer using pathway information.

Two studies are interrelated in that gene selection is performed using a high level interpretation. The first study provided interpretation of gene expression data at the pathway level that can explain the biological function, and showed the differences at the pathway level. The second study used the differences at the pathway level to reduce the high dimensional features into subtype specific gene sets.

The significance of this study is that important genes, possibly driver genes, for each subtype of breast cancer can be selected with a machine learning method, rather than manually by cancer experts.

In addition, the genes selected by the proposed method are directly from pathways, thus difference in pathway activations among breast cancer subtypes is clear, which is much more interpretable than other methods for gene selection, including the manual selection by cancer experts.

Bibliography

Aravindhan, G., Kumar, G. R., Kumar, R. S., and Subha, K. (2009). Ajax interface: a breakthrough in bioinformatics web applications. *Proteomics Insights*, **2**, PRI–S2261.

Bianco, L., Riccadonna, S., Lavezzo, E., Falda, M., Formentin, E., Cavalieri, D., Toppo, S., and Fontana, P. (2017). Pathway inspector: a pathway based web application for rnaseq analysis of model and non-model organisms. *Bioinformatics*, **33**(3), 453–455.

Bild, A. H., Yao, G., Chang, J. T., Wang, Q., Potti, A., Chasse, D., Joshi, M.-B., Harpole, D., Lancaster, J. M., Berchuck, A., *et al.* (2006). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, **439**(7074), 353–357.

Blakeman, V., Williams, J. L., Meng, Q.-J., and Streuli, C. H. (2016). Circadian clocks and breast cancer. *Breast Cancer Research*, **18**(1), 89.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, **68**(6), 394–424.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and

West, M. (2008). High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association*, **103**(484), 1438–1456.

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC bioinformatics*, **14**(1), 128.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Dai, X., Li, T., Bai, Z., Yang, Y., Liu, X., Zhan, J., and Shi, B. (2015). Breast cancer intrinsic subtype classification, clinical use and future trends. *American journal of cancer research*, **5**(10), 2929.

Dai, X., Xiang, L., Li, T., and Bai, Z. (2016). Cancer hallmarks, biomarkers and breast cancer molecular subtypes. *Journal of Cancer*, **7**(10), 1281.

Fernandez-Banet, J., Esposito, A., Coffin, S., Horvath, I. B., Estrella, H., Schefzick, S., Deng, S., Wang, K., AChing, K., Ding, Y., *et al.* (2016). Oasis: web-based platform for exploring cancer multi-omics data. *Nature methods*, **13**(1), 9–10.

Filipski, E., King, V. M., Li, X., Granda, T. G., Mormont, M.-C., Liu, X., Claustrat, B., Hastings, M. H., and Lévi, F. (2002). Host circadian clock as a control point in tumor progression. *Journal of the National Cancer Institute*, **94**(9), 690–697.

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., *et al.* (2013). Integrative analysis of

complex cancer genomics and clinical profiles using the cbioportal. *Science signaling*, **6**(269), pl1–pl1.

García-Campos, M. A., Espinal-Enríquez, J., and Hernández-Lemus, E. (2015). Pathway analysis: state of the art. *Frontiers in physiology*, **6**, 383.

Gatza, M. L., Lucas, J. E., Barry, W. T., Kim, J. W., Wang, Q., Crawford, M. D., Datto, M. B., Kelley, M., Mathey-Prevot, B., Potti, A., *et al.* (2010). A pathway-based classification of human breast cancer. *Proceedings of the National Academy of Sciences*, **107**(15), 6994–6999.

Hu, Z., Fan, C., Oh, D. S., Marron, J., He, X., Qaqish, B. F., Livasy, C., Carey, L. A., Reynolds, E., Dressler, L., *et al.* (2006). The molecular portraits of breast tumors are conserved across microarray platforms. *BMC genomics*, **7**(1), 1–12.

Huang, E., Ishida, S., Pittman, J., Dressman, H., Bild, A., Kloos, M., D'Amico, M., Pestell, R. G., West, M., and Nevins, J. R. (2003). Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nature genetics*, **34**(2), 226–230.

Kalari, S. and Pfeifer, G. P. (2010). Identification of driver and passenger dna methylation in cancer by epigenomic analysis. In *Advances in genetics*, volume 70, pages 277–308. Elsevier.

Kandasamy, K., Mohan, S. S., Raju, R., Keerthikumar, S., Kumar, G. S. S., Venugopal, A. K., Telikicherla, D., Navarro, J. D., Mathivanan, S., Pecquet, C., *et al.* (2010). Netpath: a public resource of curated signal transduction pathways. *Genome biology*, **11**(1), 1–9.

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

Kim, I., Choi, S., and Kim, S. (2018). Brca-pathway: a structural integration and visualization system of tcga breast cancer data on kegg pathways. *BMC bioinformatics*, **19**(1), 42.

Kris, M., Johnson, B., Kwiatkowski, D., Iafrate, A., Wistuba, I., Aronson, S., Engelman, J., Shyr, Y., Khuri, F., Rudin, C., *et al.* (2011). Identification of driver mutations in tumor specimens from 1,000 patients with lung adenocarcinoma: The nci's lung cancer mutation consortium (lcmc). *Journal of Clinical Oncology*, **29**(18_suppl), CRA7506–CRA7506.

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., *et al.* (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, **44**(W1), W90–W97.

Kwa, M., Makris, A., and Esteva, F. J. (2017). Clinical utility of gene-expression signatures in early stage breast cancer. *Nature reviews Clinical oncology*, **14**(10), 595–610.

Lim, S., Park, Y., Hur, B., Kim, M., Han, W., and Kim, S. (2016). Protein interaction network (pin)-based breast cancer subsystem identification and activation measurement for prognostic modeling. *Methods*, **110**, 81–89.

Lim, S., Lee, S., Jung, I., Rhee, S., and Kim, S. (2020). Comprehensive and critical evaluation of individualized pathway activity measurement tools on pan-cancer data. *Briefings in bioinformatics*, **21**(1), 36–46.

Luo, W. and Brouwer, C. (2013). Pathview: an r/bioconductor package for pathway-based data integration and visualization. *Bioinformatics*, **29**(14), 1830–1831.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, **9**(Nov), 2579–2605.

Maddika, S., Ande, S. R., Panigrahi, S., Paranjothy, T., Weglarczyk, K., Zuse, A., Eshraghi, M., Manda, K. D., Wiechec, E., and Los, M. (2007). Cell survival, cell death and cell cycle pathways are interconnected: implications for cancer therapy. *Drug Resistance Updates*, **10**(1-2), 13–29.

O'Connor, C. M., Adams, J. U., and Fairman, J. (2010). Essentials of cell biology. *Cambridge, MA: NPG Education*, **1**, 54.

Pan, H., Zhou, W., He, W., Liu, X., Ding, Q., Ling, L., Zha, X., and Wang, S. (2012). Genistein inhibits mda-mb-231 triple-negative breast cancer cell growth by inhibiting nf-κb activity via the notch-1 pathway. *International journal of molecular medicine*, **30**(2), 337–343.

Pandey, R., Guru, R. K., and Mount, D. W. (2004). Pathway miner: extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data. *Bioinformatics*, **20**(13), 2156–2158.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., *et al.* (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*, **27**(8), 1160.

Perou, C. M., Sørlie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., *et al.* (2000). Molecular portraits of human breast tumours. *nature*, **406**(6797), 747–752.

Plati, J., Bucur, O., and Khosravi-Far, R. (2008). Dysregulation of apoptotic

signaling in cancer: molecular mechanisms and therapeutic opportunities. *Journal of cellular biochemistry*, **104**(4), 1124–1149.

Poma, P., Labbozzetta, M., D'Alessandro, N., and Notarbartolo, M. (2017). Nf-κb is a potential molecular drug target in triple-negative breast cancers. *OMICS: A Journal of Integrative Biology*, **21**(4), 225–231.

Rhee, S., Seo, S., and Kim, S. (2017). Hybrid approach of relation network and localized graph convolutional filtering for breast cancer subtype classification. *arXiv preprint arXiv:1711.05859*.

Schwartz, G. K. and Shah, M. A. (2005). Targeting the cell cycle: a new approach to cancer therapy. *Journal of clinical oncology*, **23**(36), 9408–9421.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., *et al.* (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, **98**(19), 10869–10874.

Sotiriou, C. and Pusztai, L. (2009). Gene-expression signatures in breast cancer. *New England Journal of Medicine*, **360**(8), 790–800.

Vermeulen, K., Van Bockstaele, D. R., and Berneman, Z. N. (2003). The cell cycle: a review of regulation, deregulation and therapeutic targets in cancer. *Cell proliferation*, **36**(3), 131–149.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., and Kinzler, K. W. (2013). Cancer genome landscapes. *science*, **339**(6127), 1546–1558.

Wan, Y.-W., Allen, G. I., and Liu, Z. (2016). Tcga2stat: simple tcga data access for integrated statistical analysis in r. *Bioinformatics*, **32**(6), 952–954.

Zhu, H., Bhaijee, F., Ishaq, N., Pepper, D. J., Backus, K., Brown, A. S., Zhou, X., and Miele, L. (2013). Correlation of notch1, pakt and nuclear nf-κb expression in triple negative breast cancer. *American Journal of Cancer Research*, **3**(2), 230.

Zhu, J., Shi, Z., Wang, J., and Zhang, B. (2014). Empowering biologists with multi-omics data: colorectal cancer as a paradigm. *Bioinformatics*, **31**(9), 1436–1443.

국문초록

유방암은 유방 조직에서 발생하는 암이며 여성에서 발생하는 암의 25%를 차지하는 여성의 주요 암 유형이다. 유방암은 수십 년 동안 광범위하게 연구되어 왔지만, 여전히 해결해야 할 문제가 많이 있다. 생명공학 기술의 급속한 발전으로 유방암으로부터 대량의 분자 데이터가 생성되고 있다. 유방암 세포 내에서 측정된 분자 데이터는 유방암에서 해결되지 않은 많은 문제를 연구하는 데 매우 유용하게 사용될 수 있다. 그러나, 유전자 돌연변이 및 유전자 전사체와 같은 분자 데이터의 분석은 고려해야 할 특성의 수가 많고 (20,000개에서 수백만까지) 환자의 수가 수천에 불과하기 때문에 매우 어렵다. 이러한 고차원 저샘플 문제는 기계학습에서 해결되지 않은 문제 중 하나이다. 따라서 분자 수준에서 유방암을 연구하기 위해서는 새로운 방법이 필요하다.

본 박사학위 논문에서는 세포 성장, 세포 사멸, 암의 전이와 같은 생물학적 기능의 관점에서 유방암 메커니즘을 설명하기 위해 생물학적 경로를 이용하여 유방암과 그 아형에 대한 문제를 해결하는 방법을 다루었다. 첫 번째는 유전자 돌연변이, 유전자 복제수 변이 및 유전자 발현 수준의 측면에서 생물학적 경로를 탐색하기 위한 웹 기반 시스템을 개발하였고 두 번째는 생물학적 경로를 이용하여 각 유방암 아형 특이적인 유전자를 결정하기 위한 확률적 프레임 워크를 개발하였다.

첫 번째 연구는 TCGA 유방암 데이터를 KEGG 생물학적 경로에 통합하여 유방암 환자의 멀티오믹스 데이터를 시각화하는 것이다. 생물학적 경로 기반의 다중 오믹스 분석 시스템이 필요하지만, 더 큰 샘플 크기와 더 큰 차원으로 인해 어려움이 있다. 이러한 어려움을 해결하고 TCGA 유방암 데이터에 대한 생물학적으로 통합적인 관점을 제공하기 위해 KEGG 생물학적 경로에 대한 TCGA 유방암 데이터의 웹 기반 대화형 탐사 및 시각화 시스템인 BRCA-Pathway를 개발하였다. 첫 번째 연구를 통해, 유방암 환자의 멀티오믹스 데이터가 생물학적

경로 수준에서 각 아형에 대해 다르게 나타나는 것을 확인할 수 있었다. 특히, 유전자 발현량 데이터는 몇몇 생물학적으로 중요한 경로에서 각각의 유방암 아형에 대해 상이한 발현 패턴을 보이는 것을 확인할 수 있었다.

두 번째 연구는 KEGG 생물학적 경로 수준에서 각 유방암 아형별 상이한 발현 패턴을 이용하여 아형 특이적인 유전자를 선택하는 문제를 해결하려고 하였다. 각 유방암 아형이 보이는 생물학적 경로 수준에서의 발현 패턴의 차이는 생물학적 경로의 활성화 점수를 통해 표현되었다. 각 유방암 아형에 대한 유전자 발현량의 차이를 정량화하고 이 값을 유전자 인자로 정의하고, 각 유방암 아형에 대한 생물학적 경로 활성화 정도의 차이를 경로 인자로 정의하였다. 우도와 사후 확률은 유전자 인자와 경로 인자를 사용하여 정의되었으며, 유전자는 각각 우도와 사후 확률로 순위가 매겨진다. 생물학적 경로 정보를 사용하여 각 유방암 아형 특이적인 유전자를 선택하는 문제는 유방암 아형 분류 모델에서의 특징 선택에 해당함을 알 수 있다. 이러한 이유로, 분류 문제의 특징으로서 선택된 유전자를 갖는 예측 모델의 성능을 평가하였다. 또한, 선택된 유전자의 생물학적 의미를 분석하였다.

이 연구의 중요성은 유방암 각 아형 특이적인 유전자가 생물학적 경로 정보에 의해 유도된 기계학습 방법에 따라 선택되었다는 것이다. 이는 생물학적 기능과 밀접한 관련이 있는 생물학적 경로 정보가 유전자 선택 과정에 사용되기 때문에 우리의 방법으로 선택된 유전자는 생물학적으로 해석 가능한 유전자라고 말할 수 있다.

주요어: 고차원 데이터, 생물학적 경로, 유전자 발현량, 기계학습, 베이지안 방법
학번: 2014-30325