이학박사 학위논문

Quantification of pathway activity using RNA-seq data

RNA-seq 데이터를 활용한 패스웨이 활성도의 정량화에 관한 연구

2019 년 8 월

서울대학교 대학원

협동과정 생물정보학

임 상 수

이학박사 학위논문

Quantification of pathway activity using RNA-seq data

RNA-seq 데이터를 활용한 패스웨이 활성도의 정량화에 관한 연구

2019 년 8 월

서울대학교 대학원

협동과정 생물정보학

임 상 수

Quantification of pathway activity using RNA-seq data

RNA-seq 데이터를 활용한 패스웨이 활성도의 정량화에 관한 연구

지도교수 김 선

이 논문을 이학박사 학위논문으로 제출함

2019 년 5 월

서울대학교 대학원

협동과정 생물정보학

임 상 수

김민수의 이학박사 학위논문을 인준함

2019 년 6 월

위 원 장	이병재
부위원장	김선
위 원	손현석
위 원	황대희
위 원	김광수

Abstract

Quantification of pathway activity using RNA-seq data

Sangsoo Lim
Interdisciplinary Program in Bioinformatics
College of Natural Sciences
Seoul National University

Measuring the dynamics of RNA transcripts using RNA-seq data has become routine in bioinformatics analyses. However, RNA-seq produces high-dimensional transcriptome data on more than 20,000 genes in humans. This makes the interpretation of the data extremely difficult given a relatively small set of samples. Therefore, it is desirable to use well-summarized and widely-used information such as biological pathways for better biological comprehension. However, summarizing transcriptome data in terms of biological pathways is a very challenging task for several reasons. First, there is a huge information loss when transforming transcriptome data to pathway space. For example, in humans, only one third of the entire set of genes being analyzed are present in KEGG pathways. Second, each pathway consists of many genes; thus, measuring pathway activity requires a strategy to summarize expression profiles of component genes into a single value, while considering relationship among the constituent genes.

My doctoral study aimed to develop a new method for pathway activity measurement, and to perform extensive evaluation experiments on existing pathway measurement tools in terms of multiple evaluation criteria. In addition, a cloud-based system was constructed to deploy such tools, which facilitates users analyzing their own data easily.

The first study is to develop a new method to summarize transcriptome data in terms of pathways by using explicit transcript quantity information and considering relationship among genes in terms of their interactions. In this study, I propose a novel concept of decomposing biological pathways into subsystems by utilizing protein interaction network, pathway information, and RNA-seq data. A subsystem activation score (SAS) was designed to measure the degree of activation for each subsystem and each patient. This method revealed distinctive genome-wide activation patterns or landscapes of subsystems that are differentially activated among samples as well as among breast cancer subtypes. Next, we used SAS information for prognostic modeling by classification and regression tree (CART) analysis. Eleven subgroups of patients, defined by the 10 most significant subsystems, were identified with maximal discrepancy in survival outcome. Our model not only defined patient subgroups with similar survival outcomes, but also provided patient-specific decision paths determined by SAS status, suggesting functionally informative gene sets in breast cancer.

The second study aimed to systematically compare and evaluate thirteen different pathway activity inference tools based on five comparison criteria using a pan-cancer data set. Although many pathway activity tools are available, there is no comparative study on how effective these tools are in producing useful information at the cohort level, enabling comparison of many samples. This study has two major contributions. First, this study provides a comprehensive survey on computational techniques used by existing pathway activity inference

tools. Existing tools use different strategies and assume different requirements on data: input transformation, use of labels, necessity of cohort-level input data, use of gene relations and scoring metrics. Second, extensive evaluations were conducted using five comparison criteria concerning the performance of these tools. Starting from measuring how well a tool maintains the characteristics of an original gene expression profile, robustness was also investigated by introducing noise into gene expression data. Classification tasks on three clinical variables were performed to evaluate the utility of tools.

The third study is to build a cloud-based system where a user provides transcriptome data and measures pathway activities using the tools that were used for the comparative study. When a user uploads input data to the system and selects which preferred analysis tools are to be run, the system automatically generates pathway activity values for each tool as well as a summary of performance comparison for the selected tools. Users can also investigate which pathways are significant in terms of the given sample information and visually inspect genes within a pathway-linked KEGG rest API.

In conclusion, in my thesis, I sought to develop an analysis method regarding biological pathways using high throughput gene expression data to compare different types of tools with comprehensive criteria, and to arrange the tools in a cloud-based system that is easily accessible. As pathways aggregate various molecular events among genes in to a single entity, the set of suggested approaches will aid interpretation of high-throughput data as well as facilitate integration of diverse data layers such as miRNA or DNA methylation profiles being taken into consideration.

Keywords: biological pathway, pathway activity, protein-protein interaction, biological network, gene expression, RNA-seq

Student Number: 2014-30099

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The phenotype of an organism is the outcome of the complex nature of biological components such as genes, proteins or metabolites. Reflecting the central dogma of molecular biology, it is now common practice to investigate biological entities in a more comprehensive way such that their complex relationships are well explained (Khatri *et al.*, 2012). Their cooperative mechanisms work in a highly correlated manner that together build several coordinated units. To understand how biological entities are coordinated, it is necessary to use functional annotations of them to improve the interpretability of given molecular data sets (Mattson, 2004; Vogelstein and Kinzler, 2004; Reynard and Loughlin, 2013). One favorable approach is to utilize biological pathways.

A pathway is a series of relations involving molecules closely related in terms of a certain biological context (Figure 1.1). Pathways contain not just molecular entities, but also their regulatory/interactive information from biological findings integrated over decades (Cary *et al.*, 2005). Some pathways contain highly complex relationships involving single entities of multiple biologically represen-

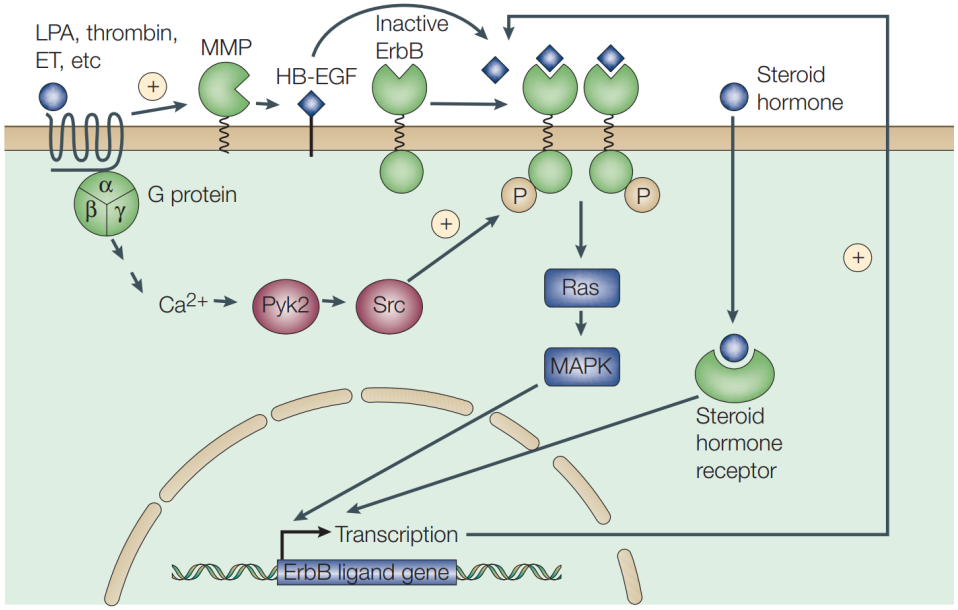

Figure 1.1: Example illustration of a biological pathway - ErbB signaling pathway (Yarden and Sliwkowski, 2001). Pathways contain heterogeneous biological constituents as well as their regulatory/hierarchical relationships to explain biological functions.

tative functions. Despite many researches on characterizing the role of genes in terms of pathways, only one-third of all genes reside in pathway databases (Khatri *et al.*, 2012). Therefore, it is a challenging task to effectively utilize pathway databases to compensate for such information loss.

One of the solutions suggested was to summarize component gene expression values of each pathway to a single value, called pathway activity (Rahnenführer *et al.*, 2004). Calculating pathway activity is a quantitative way of explaining the dynamics of a pathway using gene expression data. There are a variety of such tools adapting external resources to better reflect complicated relationships within a pathway (Mitrea *et al.*, 2013; Jaakkola and Elo, 2015; Bayerlová *et al.*, 2015). However, some of the tools fail to reflect structural information among

the genes and comparative elaboration with previous tools.

1.1 Biological background

1.1.1 Biological pathways

Efforts have been made to integrate the rich information accumulated over the decades to categorize genes according to their functional or molecular characteristics (Haeussler *et al.*, 2018; Sayers *et al.*, 2019). This categorization was dedicated to design a series of genes or constituents that are closely related so as to build de novo entities – biological pathways. There are many pathway databases available including the famous KEGG and Reactome (Kanehisa and Goto, 2000; Croft *et al.*, 2013) (Figure 1.2). They share the same basic idea of building a single pathway from biological components that are closely related in a certain context. In the meantime, the databases differentiate themselves by focusing on specific biological context such as signaling or disease (Figure 1.3) (Paz *et al.*, 2010; Caspi *et al.*, 2013).

1.1.2 Gene expression

The level of gene expression is in general considered to be the amount of messenger RNA (mRNA) in a sample. It is a crucial molecular signature to understand the dynamics of organisms (Ahr *et al.*, 2001; Sotiriou and Pusztai, 2009; Liberzon *et al.*, 2015). However, the fact that there are more than 20,000 genes to be measured in human makes analyzing each gene, one at a time, difficult to investigate the genome-wide landscape of mRNA profiles.

One of the breakthroughs in gene expression measurement was to use array-based measurement techniques (TAUB *et al.*, 1983). Genome-wide measurement of mRNA levels became affordable since array-based measurement entered the

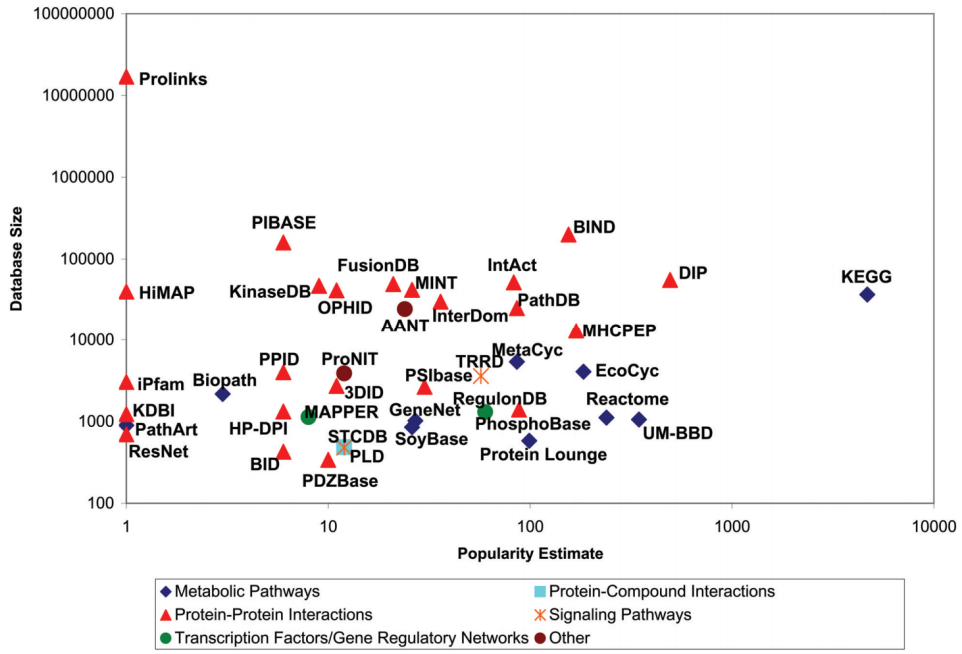

Figure 1.2: Pathway databases in terms of their popularity and size (Bader *et al.*, 2006). As databases have their own context, they were categorized into six different types: Metabolic Pathways, Protein-Protein Interactions, Transcription Factors/Gene Regulatory Networks, Protein-Compound Interactions, Signaling Pathways, and others.

mainstream. Microarray technology gained its popularity due to its rapid measurement of a large number of samples (Barrett *et al.*, 2010). However, it requires predesigned complementary DNA (cDNA) fragments matched to target genes (or mRNAs). This becomes the bottleneck in finding new discoveries regarding the nature of RNA, as extensive splicing events cannot be detected (Mortazavi *et al.*, 2008; Sultan *et al.*, 2008). It also has difficulty in measuring absolute amounts of mRNA molecules. Reduced sensitivity to low abundance mRNAs also sets another hurdle for its application to further genome-wide studies. Meanwhile, DNA sequencing techniques, first released in 1977 (Sanger

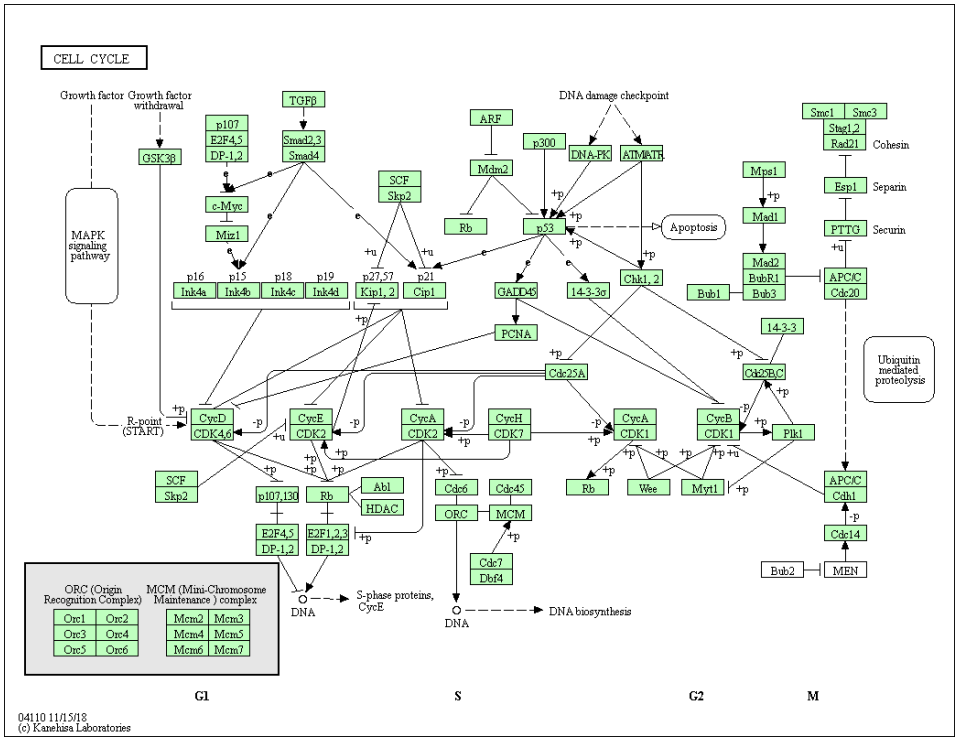

Figure 1.3: Cell cycle pathway of KEGG database (Kanehisa and Goto, 2000). There are proteins (genes) with relational information between them. There are also links to other pathways as many important genes are shared among pathways.

et al., 1977), were considered as a counterpart of array-based technologies as sequencing does not require template DNA fragments to read mRNA (or cDNA) sequence. Once a primer is prepared against the target DNA fragment to be sequenced, sequential light signals at four different wavelengths indicate which character is present at the corresponding position.

Since the advent of high-throughput sequencing of genomic materials, measuring gene expression profiles using RNA sequencing (RNA-seq) has become routine in genome-wide molecular analysis of diseases (Kawashima *et al.*, 2012;

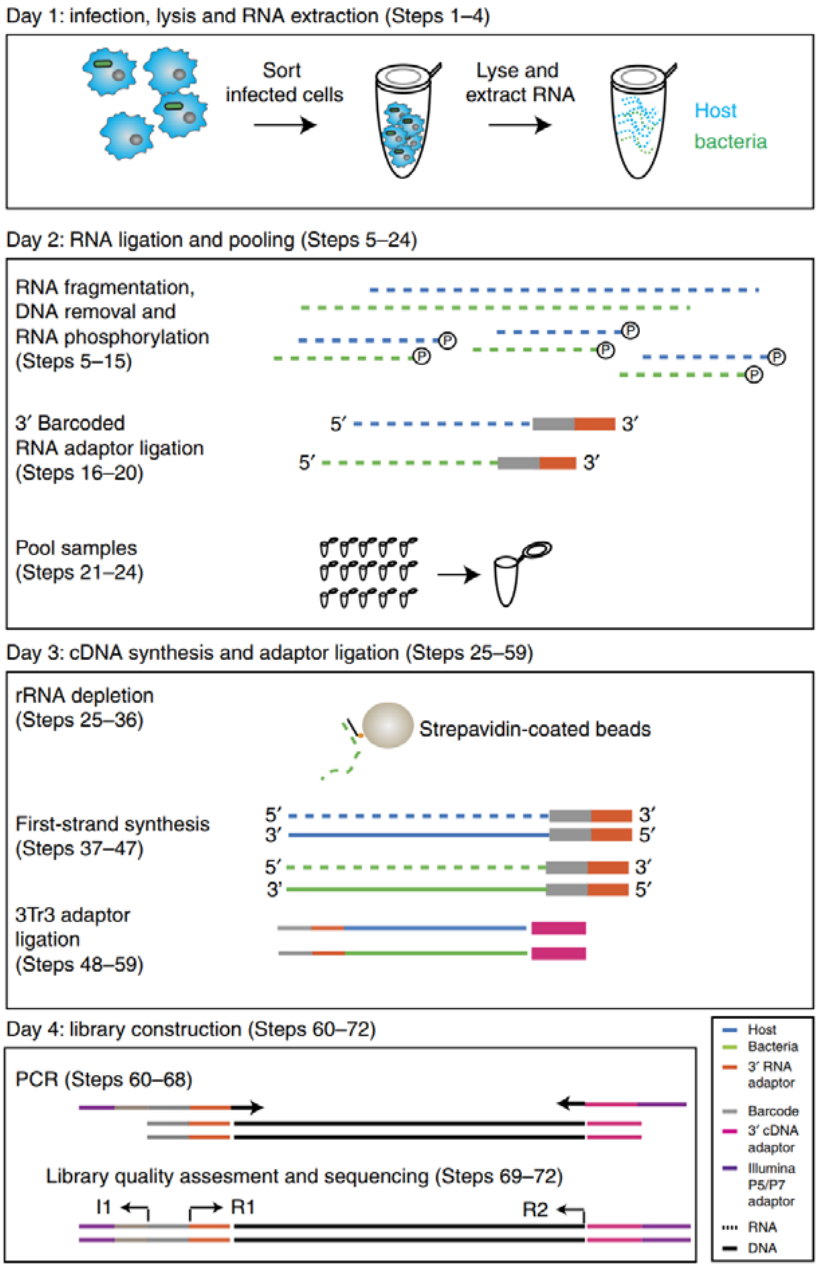

Figure 1.4: Overview of RNA-seq analysis protocol (Avraham *et al.*, 2016).

Wang *et al.*, 2009; Trapnell *et al.*, 2010). RNA-seq generates sequence reads from captured mRNA molecules in a massively parallel manner that results in a huge number of sequence records (Figure 1.4). Many advantages of RNA-seq over previous array-based measurements include higher dynamic range, greater sensitivity, being unbiased and no limitations because of prior genomic knowledge (Marioni *et al.*, 2008; Zhao *et al.*, 2014). RNA-seq provides unprecedented opportunities in genome-wide projects, with greater numbers of samples being analyzed at reduced cost (Hrdlickova *et al.*, 2017). Several studies have also demonstrated the usefulness of RNA-seq on large data sets (Park *et al.*, 2012; Lonsdale *et al.*, 2013; Leiserson *et al.*, 2015).

1.1.3 Pathway-based analysis

Most genome-wide studies focus on identifying differentially expressed genes (DEGs) in gene expression data sets. However, there has been a paradigm shift in interpreting the data from gene to pathway levels since the release of tailored pathway databases (Slonim, 2002; Ravasz *et al.*, 2002; Ge *et al.*, 2003; Wagner *et al.*, 2007). This is because using pathways can empower identification of risk factors for complex diseases by measuring the aggregate effects of individual genes (Huang *et al.*, 2008).

Pathway-based analysis is to identify pathways that are statistically significantly enriched with genes of interest at a certain confidence level. This type of analysis gained its popularity in elucidating mechanisms of complex diseases (Emmert-Streib and Glazko, 2011). Such methods, in general, adapt statistical or machine learning-based techniques to investigate significance. GSEA (Subramanian *et al.*, 2005) and DAVID (Huang *et al.*, 2007) are two major examples of tools employed to investigate enrichment of certain pathways or ontologies from given gene list. Both tools measure significance based on how many genes

are statistically enriched in a pathway. PARADIGM (Vaske *et al.*, 2010) is a tool aimed at modeling actual known relationships within a pathway by using Bayesian network approaches.

1.1.4 Pathway activity measurement

In addition to simply identifying pathways enriched with DEGs or genes of interest, other approaches have focused on assigning a score for each pathway (Rahnenführer *et al.*, 2004). Compared to conventional pathway-based analysis, calculating pathway activity values throughout the samples and pathways can identify individual variation in the samples. For example, GSVA uses a non-parametric approach to identify the degree of perturbation to pathways (Hänzelmann *et al.*, 2013). GSVA further divides itself into two subsequent tools (GSVAmax and GSVAdif) to better reflect the nature of input gene expression data. Meanwhile, PLAGE (Tomfohr *et al.*, 2005) simply uses the first principal component of each pathway genes as pathway activity. These methods can be extended to build a machine learning model for disease classification problems (Gatza *et al.*, 2010).

Some tools were revisited in previous reviews to help understand the differences between them, and focusing on the ability to identify known significant pathways (Mitrea *et al.*, 2013; Jin *et al.*, 2014; Jaakkola and Elo, 2015; Bayerlová *et al.*, 2015). These reviews mainly focus on describing calculation processes and usefulness in classification problems, or by simply cataloging their features. However, there needs to be an unbiased review that provides extensive evaluation on such tools using a large number of data sets with thorough criteria.

1.2 Challenges in pathway activity measurement

This thesis is dedicated to addressing three main problems (Figure 1.5). The first problem concerned how to effectively leverage pathway information to assign a single representative pathway activity. The second problem was to systematically evaluate pathway activity inference tools by suggesting several comparative criteria for the tools through extensive elaboration. The final issue was to provide users who are not familiar with but interested in using such tools with easy access, and to help navigate the results arising from their own data.

1.2.1 Calculating effective pathway activity values from RNA-seq data

Even though it is valuable to use pathways in analyzing gene expression data, most pathway databases envelope a smaller number of genes than the number of genes being analyzed. This is because the curation of most databases has been biased toward the most studied genes. Therefore, it is natural that documentation is inevitably biased to certain well-studied findings. Nevertheless, describing a pathway with a single value is still a desirable practice since the abstraction of gene level information can also be regarded as enrichment of the essential biological information (Khatri *et al.*, 2012; Ramanan *et al.*, 2012).

One of the applications of pathway activity is to build a machine learning model, regarding pathways as features for disease classification (Gatza *et al.*, 2010; Lee *et al.*, 2008). Calculating pathway activity for each sample from gene expression data can then be considered as a powerful way of transforming gene dimensions into pathway dimensions to avoid potential over-fitting issues. Directly using genes as features from currently available large-scale gene expression data suffers from significantly insufficient number of samples. However,

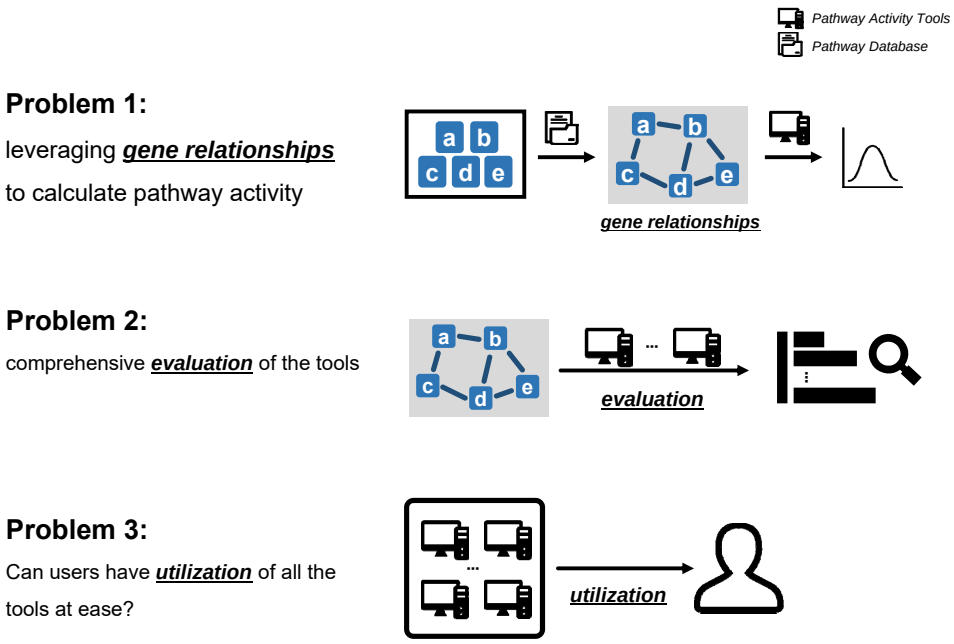

Figure 1.5: Technical problems to be addressed in this thesis. As there are complex gene relationships present in pathways, it is important to utilize them to summarize component gene expression profiles into a single value. Such challenge was addressed by several tools. Therefore, it requires an extensive comparison of them to understand which of the tools are more useful by evaluating on five different criteria. Finally, users should be able to have an easy access to the tools through a single unified system.

biological pathways as features can provide more precise and straightforward biological essence, while meeting the sufficiency for building a machine learning model to solve many classification problems. A model based on such features also should be able to solve more difficult classification problems such as survival outcome or cancer subtypes (Kim *et al.*, 2012).

1.2.2 Lack of comparative criteria to evaluate pathway activity tools

Previous reviews on pathway activity tools focus on simply cataloging or comparing tools in terms of whether a tool incorporates other information such as relationships between genes (Mitrea *et al.*, 2013; Jaakkola and Elo, 2015; Bayerlová *et al.*, 2015). For example, as there are various types of approach for pathway activity inference, some techniques are sensitive to subtle perturbations in gene expression data. Reliable outcomes are not guaranteed, especially when the perturbation is not from biological factors but rather from technical issues (Goncalves *et al.*, 2011; Wang *et al.*, 2012; DeLuca *et al.*, 2012; Zwiener *et al.*, 2014; Danielsson *et al.*, 2015). In addition, it is desirable to know to what extent a tool sustains the structure of the original input data when transforming to pathway activity. Therefore, a systematic criteria for the comparison of pathway activity inference tools in various aspects needs to be established.

1.2.3 Absence of a user-friendly environment of pathway activity inference tools

Regardless of the significance a tool to a particular field of study, it is almost impossible for a user to implement all the available tools alone, especially when not familiar with bioinformatic tools, or without sufficient computational resources. One such examples is Enrichr (Chen *et al.*, 2013), a web-based suite

of tools that accepts a list of genes provided by the user. Simple submission of a gene list provides a wide range of results, including the enrichment of various ontology-based databases. It was spotlighted as being user-friendly, with interactive visualization of results and the capability of analyzing genes from species other than human. However, it introduces only a single analysis strategy for various database libraries. One is also not aware of which tools are valid for the relevant research without thorough comparisons. Therefore, it is important for a set of tools to be easily accessible online.

1.3 Outline of the thesis

In this thesis, a series of solutions are suggested throughout three chapters to tackle the above problems using biological pathway databases with gene expression data and other resources.

In chapter 2, I developed a method to calculate pathway activity using the relational information between genes within each pathway. Both gene expression values and the relational information of the genes within each pathway are taken into consideration to reflect that nearby genes share more in common. This work was extended to decompose a pathway that has more than a single biological significance into several components in terms of the given gene expression data.

In chapter 3, a comprehensive evaluation of pathway activity inference tools was undertaken to demonstrate their usefulness against several performance criteria. Five criteria were introduced that covers different aspects of the tools, including distance preservation and robustness to noise.

In chapter 4, I deployed previously compared pathway activity tools on a cloud-based web platform. Users can upload their own gene expression data to run selected tools, visualize the results in terms of significance, and direct

the result to the pathway database. The comparative analysis criteria from a previous study were adapted in this platform to help users comprehend which of the tools are appropriate for their own data.

Collectively, chapter 5 summarizes both the significance and potentials of scoring pathway activity using biological pathways in interpreting RNA-seq gene expression data. The thesis is concluded with a bibliography of references and appendices.

Chapter 2

Measuring pathway activity from RNA-seq data to identify breast cancer subsystems using protein-protein interaction network

2.1 Related works

Prognostication and prediction of patients' survival are one of the major goals in breast cancer research. Practical decision making of the breast cancer treatment plan is based on clinicopathological features such as tumor size, lymph-node metastasis, histological grade and three receptor (ER, PR, and HER2) responses to endocrine therapy (Reis-Filho and Pusztai, 2011). Although these methods have been widely and successfully used since 1970s, they are not effective for diagnosis of the cancer at earlier stages and precise clinical decisions requires more than the clinicopathological features (EBCTCG *et al.*, 2005). Thus, investigation on the genome-wide landscape of molecular features in breast cancer

has been extensively performed (Perou *et al.*, 2000; Sotiriou *et al.*, 2006; TCGA *et al.*, 2012; Ross *et al.*, 2015). These efforts initiated a new paradigm of clustering patients followed by annotating characteristic labels on the clusters of patient groups in terms of survival outcome (Curtis *et al.*, 2012).

In an effort to develop a model for clustering patients, there were several array-based gene expression studies grouping patents based on a set of genes that are differentially expressed among the cohort, yielding molecular subtypes based on patient clusters (Perou *et al.*, 2000; Sorlie *et al.*, 2001; Van De Vijver *et al.*, 2002; Van't Veer *et al.*, 2002; Sørlie *et al.*, 2003; Paik *et al.*, 2004; Ma *et al.*, 2004; Chang *et al.*, 2005; Hu *et al.*, 2006).

Surprising discovery from these studies was that only a small number of genes were sufficient to characterize patient groups at the molecular level. In addition, genes selected by different studies show similarities in terms of gene expression levels. These gene expression signatures proved themselves as a determinant to survival outcome without resorting to anatomical prognostic variables such as tumor size or nodal status (Reis-Filho *et al.*, 2010). Most of the methods showed equipotent performances in terms of prognostic modeling with a high concordance rate (Fan *et al.*, 2006). Among them, PAM50 method became standardized as the fundamental requirement for molecular diagnosis of breast cancer, of which assigns subtypes by incorporating microarray expression values to the centroids of 50 genes (Parker *et al.*, 2009).

However, even PAM50 subtypes remained heterogeneous in receptor status; for example, among basal-like subtype patients, 17 % of the samples were in neither ER-negative or HER2-negative statuses, despite that being accepted as typical clinical-pathological features of basal-like subtype (Prat and Perou, 2011). In another study by Parker *et al.* (2009), it was suggested that luminal B subtype can be divided into at least five subgroups. One reason for this would

be from the fact that the selection of genes in PAM50 was not guided by accurate gene expression profiles that are measured by microarray technologies. This can be resolved by leveraging RNA-seq technologies as demonstrated in a study by Wang *et al.* (2014). In comparison with microarray data, RNA-seq produced more accurate gene expression measurements at the whole transcriptome level by showing that RNA-seq data had much higher concordance rate with expression profiles measured by qRT-PCR and also that RNA-seq achieved much better sensitivity for low-abundant genes.

Another technical issue for characterizing biological mechanisms underlying breast cancer is to consider relational nature of deregulated genes with context. Pang *et al.* (2006) used random forests for prioritizing important pathways in several diseases such as breast and lung cancer, rather than simply listing important genes for the diseases. Another popular technique is to use network. Recently, a consortium of network biology was launched to analyze multi-dimensional genomic data (Krogan *et al.*, 2015). PIN is one of the most widely used network-based analysis techniques to cover true relational characteristics (Han, 2008). For example, Hofree *et al.* (2013) used PIN as a template to diffuse the significance of somatic mutation profiles and discovered biological modules crucial for identifying patient clusters of several cancers. This was consistent with previous studies that mutational events are localized to certain area (modular structure) of a network, hardly perturbing the whole biological structures (Jeong *et al.*, 2000; Yook *et al.*, 2004).

2.2 Motivation

Importance of pathway and network utilization Wang *et al.* (2008) classified module identification methods into three categories: expression-based,

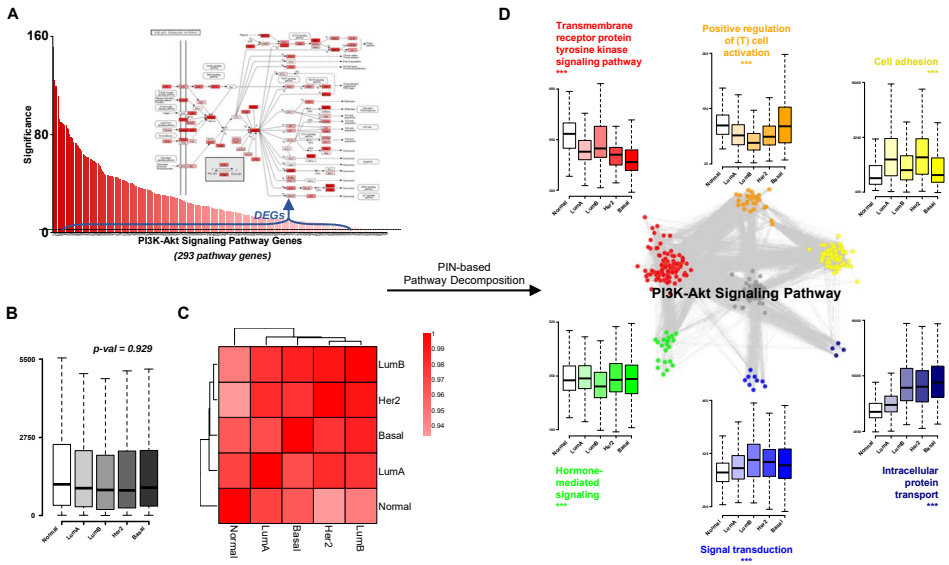

Figure 2.1: Motivating example of the PI3K-Akt signaling pathway. (A) Analysis on DEGs among the breast cancer subtypes and the mapping of the identified DEGs to the KEGG pathway. Horizontal dotted line indicates Bonferroni correction threshold for multiple comparison. (B) Boxplot of average gene expression values for each subtype of all the genes in the pathway. (C) Heatmap showing correlation of gene expression values among the subtypes. (D) Pathway decomposition revealed six subsystems of the pathway. Each subsystem shows distinct biological functions (represented in `GO:Biological Process`). Boxplots of the subsystems show discriminant patterns among the subtypes. p-values were calculated by ANOVA and asterisks at the end of the GO terms showed ANOVA significance ($0 <^{***} <0.001 <^{**} <0.01 <^{*} <0.05$).

pathway-based, and network-based approaches and this categorization was recently revisited and well summarized by Creixell *et al.* (2015).

As biological knowledge discovery moving toward deciphering the functions of cooperative machinery rather than individual DEGs, identifying the cluster or gene set modules became one of the popular research topics (Ravasz *et al.*, 2002; Ge *et al.*, 2003; Wagner *et al.*, 2007). These methods mostly used machine learning or statistical techniques to identify systems of coordinated genes. In addition, several studies focused on the measurement of activity or level of perturbation using pathway information and expression profiles (Kristensen *et al.*, 2014). It is desirable to use multi-dimensional omics data to precisely measure the activity of a pathway as performed by Vaske *et al.* (2010). However, the integrated analysis of multi-omics data needs to be further developed.

Fortunately, there are many studies using only gene expression data to measure the degree of distorting the original (trained) distribution of gene set or metagene scores (West *et al.*, 2001; Huang *et al.*, 2003; Ahn *et al.*, 2014). For example, a Bayesian regression model introduced by West *et al.* (2001) used a set of 100 genes that maximally discriminates the ER status of breast cancer. This approach was extended to examine the status of several oncogenic pathways by using metagene concept (Huang *et al.*, 2003). A further analysis of 18 representative pathways was successful to classify human breast cancer subtypes (Gatza *et al.*, 2010).

In addition, there are a number of studies that utilized well curated networks other than biological pathways. Among biological networks, PIN is widely used. As PIN covers a lot more number of genes than biological pathways, there were several studies that initiated the identification of prognostic signatures (Cheng *et al.*, 2013; Wu and Stein, 2012). Wu and Stein (2012) did a seminal work that incorporated gene expression information to PIN. In their analyses, edge

weights in PIN were defined by using microarray-based gene expression data and then network modules were identified by using the MCL clustering method. This study produced many false positives because using the microarray data do not have accurate gene expression information and co-expression information was not explicitly used. Furthermore, activation status of a module was simply calculated by averaging the gene expression values in the module without incorporating the relationships among the genes. This drawback can be remedied by utilizing RNA-seq expression data to use more accurate gene expression information and also by defining network modules in a stringent way (Wang *et al.*, 2009).

Necessity of Subsystems As discussed in the previous subsection, gene expression or transcriptome data can be better analyzed in terms of biological pathways. Commonly used pathway databases are KEGG (Kanehisa and Goto, 2000), REACTOME (Joshi-Tope *et al.*, 2005) and NCI cancer pathway (Schaefer *et al.*, 2009). A pathway is defined to model a series of actions among molecules in a cell that leads to a certain product or a change in a cell. As a result, some pathways consist of multiple complex biochemical functions, rather than a single biological function. This led to several research efforts to define multiple coherent units of a pathway. Overbeek *et al.* (2005) pioneered the use of a subsystems approach to annotate genomes by categorizing genes into single functional groups. Chang *et al.* (2009) proposed a strategy of decomposing pathway information into smaller modular structures. All these studies assure that defining functional units of a pathway is desirable and useful. However, there is no systematic study on defining subsystems of a specific disease using transcriptome data measured from many samples.

The goal of this study is to reveal biological mechanisms underlying breast

cancer in terms of pathways. To achieve this goal, a computational method needs to be developed to define functional units or subsystems of a pathway using transcriptome data. This approach is illustrated using PI3K-Akt Signaling pathway that consist of 293 genes in Figure 2.1. The widely used DEG analysis results in too many statistically significant genes that can be mapped to many pathways, so the DEG approach does not distinguish core pathways from many activated pathways when expression values of all DEGs are mapped to pathways. To measure the activation status of a pathway, when expression values of all genes in the pathway were simply averaged to a single value, the difference in the activation status of the pathway was not clear among cancer subtypes (Figure 2.1A to 2.1C). However, the approach of decomposing the pathway into a set of distinct subsystems was effective to explain the differential activation status of the pathway among cancer subtypes (Figure 2.1D).

2.3 Methods

2.3.1 Breast cancer subsystems

The overview of the subsystem identification is illustrated in Figure 2.2A. PIN from STRING (`ver.9.1`; Franceschini *et al.* (2013)) was used as a template network. For the edges in the template network, Spearman's correlation coefficient values were calculated by using the gene expression values from RNA-seq data of breast invasive carcinoma from TCGA (`http://cancergenome.nih.gov/`). Then, the template PIN network was instantiated by multiplying the weight (`combined score`) specified in STRING and the absolute Spearman's correlation coefficient for each pair of genes in the network. This instantiated the PIN of 2,004,213 edges (16,807 vertices) as BRCAPIN. To generate clusters from the network, BRCAPIN was divided into clusters by using Markov cluster

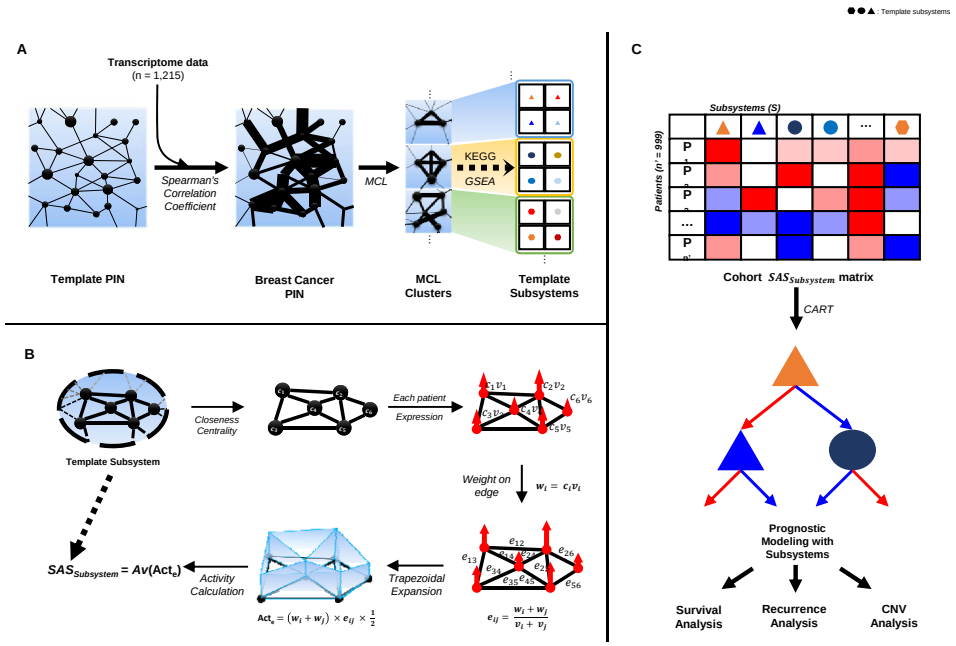

Figure 2.2: Overview of the research protocol in this paper. (A) A template PIN was instantiated by weighting Spearman's correlation coefficient calculated from breast cancer transcriptome data to the edges of PIN. In this way, a breast cancer PIN (BRCAPIN), a breast cancer PIN, was generated. (B) SAS calculation. When subsystems were generated, topological significance (closeness centrality) of each gene in each MCL cluster was set as a template value to be inflated by transcriptome data. SAS calculation is geometrically defined for a pair of genes where topological significance. i.e., closeness centrality, and gene expression of each gene is used as the two bases of a right trapezoid and relationship between genes is used as the height of the trapezoid. Then SAS is defined naturally as an average of area of the trapezoids. (C) For each subsystem and each patient, SAS was calculated and the matrix of n patients and subsystems was utilized for the prognostic modeling by CART followed by relapse and CNV analyses.

algorithm (MCL; `http://micans.org/mcl/`; Dongen (2000)). For each MCL cluster, a modified Fisher's exact test (EASE; Hosack *et al.* (2003)) followed by the Bonferroni correction was performed to investigate whether there was an enrichment of KEGG pathways (Kanehisa and Goto, 2000) or not (adjusted p-value < 0.01). The largest connected component between MCL cluster and its enriched KEGG pathway was defined as a subsystem. In total, 855 subsystems were identified.

2.3.2 Subsystem Activation Score

It is not trivial to measure the activation status of a subsystem since there are quite a number of genes in a subsystem. Simply using the average expression level of all genes in a subsystem as a value for subsystem activity is not correct. Thus, I developed a new scoring scheme for the activation status of a subsystem, called Subsystem Activation Score (SAS; Figure 2.2B). The scheme was designed to utilize both topological importance and transcriptional abundance of genes. The topological importance of a gene was determined by the closeness centrality value of a gene, which was to weight a gene in terms of the shortest distance to all other genes within MCL cluster from the gene. The transcriptional abundance was to utilize co-expression of two adjacent genes in the network. Then two factors were combined together to define a right trapezoid between two genes (Equation 2.1), where two parallel bases are weighted gene expressions and the height defined 'edge centrality' as Equation 2.2.

$$Act_e = (w_i + w_j) \times e_{ij} \times \frac{1}{2} \tag{2.1}$$

$$e_{ij} = \frac{w_i + w_j}{v_i + v_j} \tag{2.2}$$

where w is the gene expression (v) weighted by the closeness centrality (c) of a gene i and j. In this way, the area of the trapezoid between two genes was defined. Finally, SAS was calculated as an average of areas of all trapezoids of a subsystem (Equation 2.3).

$$SAS = \frac{\sum_{edges} Act_e}{\sum_{edges}} \tag{2.3}$$

Then, a matrix of subsystems vs. breast cancer samples was generated using SAS (Figure 2.2C).

2.3.3 Prognostic modeling

To determine subsystems related to the patient survival outcome, a classification and regression tree analysis (CART; Breiman *et al.* (1983)) was performed by using the `rpart` library of `R` package (Therneau *et al.*, 2010). The overview of the CART analysis is shown in (Figure 2.2C). CART analysis produced a tree where branching at each node, i.e., subsystem, was determined by the SAS value. CART was used to select a set of subsystems that characterize the patient survival and also predict the hazard ratio by using a regression model. Parameter setting for the CART analysis was guided by two prior large-scale analysis of breast cancer transcriptome data. Gatza *et al.* (2010) reported that the number of clusters of breast cancer patients remained stable near 20 as the number of patients increase. Curtis *et al.* (2012) showed that there were at least 10 clusters of breast cancer patients with distinct molecular characteristics and survival outcome. To incorporate these findings, a parameter for the minimum number of patients in each terminal node of the tree was set to be

50, which approximately corresponds to a maximum of 20 theoretical groups for 999 patients.

2.3.4 Hierarchical clustering of patients and subsystems

Hierarchical clustering of patients was performed by `ward.D2` method using Manhattan distance of SAS values. To find the optimum number of patient clusters, we used Adjusted Rand Index (ARI) implemented in `mclust` library of `R` package. ARI is to measure the level of similarity between the two clustering objects, here hierarchical clustering object and PAM50 subtypes.

The number of SCs was determined by Normalized Mutual Information (NMI). As NMI monotonically increased as the number of clusters to examine increased from 1 to 100, we performed local linear regression using the seven consecutive NMI values. To visualize the relationship among the subsystems by setting 855 subsystems as nodes and their relationships as edges (365,085 edges), PCC of SAS values among the subsystems within and outside the SCs were calculated. To maximally gain the edges within the SCs and remove the edges outside SCs, we utilized accuracy varying the PCC threshold. The accuracy was defined here as the proportion of the sum of remaining edges within SCs and removed edges outside SCs when applied PCC threshold to the all possible 365,085 edges. The maximum accuracy was obtained when this threshold was set as 0.39.

For each SC, a set of subsystems with their ANOVA F-value of SAS with respect to the subtypes were calculated. To summarize the SAS and F-values for the subsystems into a single value for each patient, we averaged SAS weighted by the corresponding F-values. This means that the subsystem with more significant difference among the subtypes contribute more to the final value.

2.3.5 Tools used in this study

MCL clustering algorithm followed by `Python` with `stats` of `scipy` library was used to generate subsystems. Network view generation was done by `Cytoscape 3.0.1`. Survival and CNV analyses were done by R.

2.4 Results

2.4.1 Pathways were decomposed into coherent functional units - subsystems

MCL divided the instantiated BRCAPIN into smaller subnetworks (MCL clusters), by utilizing the edge weights. Enrichment analysis for the MCL clusters was performed to identify the intersection between each MCL cluster and KEGG pathways. As a result, a total of 855 breast cancer subsystems with non-zero *SAS* values were identified from 186 MCL clusters. As 855 subsystems were identified out of 269 KEGG pathways, there were some KEGG pathways decomposed into more than a single subsystem. Positive correlation was found between the size of a pathway and the number of subsystems in a pathway ($r = 0.52$; Figure 2.3A). However, regardless of the number of genes in a pathway, pathways related to signaling (e.g. PI3K-Akt signaling pathway and Toll-like receptor signaling pathway) tended to have more subsystems than other pathways. Pathways of specific metabolic pathways such as Glutathione metabolism and Sphingolipid metabolism contained only a single subsystem. Among the KEGG pathways, 9 pathways including Lysosome, Adrenergic signaling in cardiomyocytes, Hippo signaling pathway were decomposed into 7 or more subsystems. As shown in Figure 2.1, average gene expression profile of a big pathway is not distinct among breast cancer subtypes. However, when decomposed into subsystems, the activation status is distinct among breast cancer

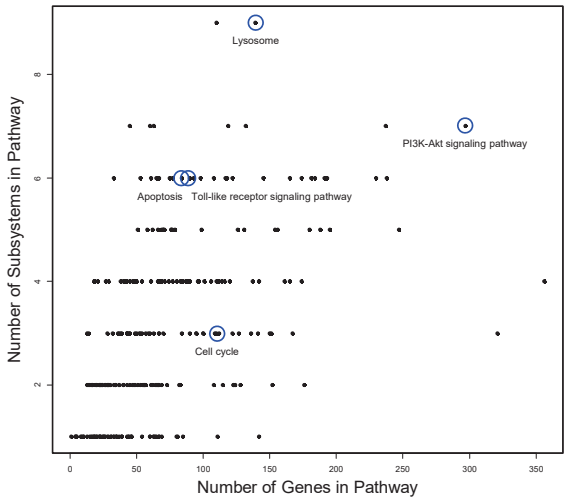

Figure 2.3: Scatterplot of the number of genes and the number of subsystems in KEGG pathways. Larger pathways share more number of subsystems.

subtypes and also at the patient level, representing distinct biological functions.

2.4.2 Landscape of subsystems reflect the breast cancer biology

To see the landscape of subsystems through their relationships, the 855 subsystems were divided into clusters called subsystem clusters (SCs). To determine the number of SCs, investigation was performed varying the numbers (Figure 2.5B). As the number of clusters increases, marginal increase in positiveness gradually converged to zero. For example, using the first 5 points, the rate of increase in NMI with respect to the number of clusters (regression coefficient) was 0.093 (red line), and it decreased to 0.061 when using the next 5 points (blue line). Here, we set the number of clusters to be the point when the regression coefficient becomes right before below 0.01. This means that the increase in the number of clusters gives no more marked improvement in the clustering perfor-

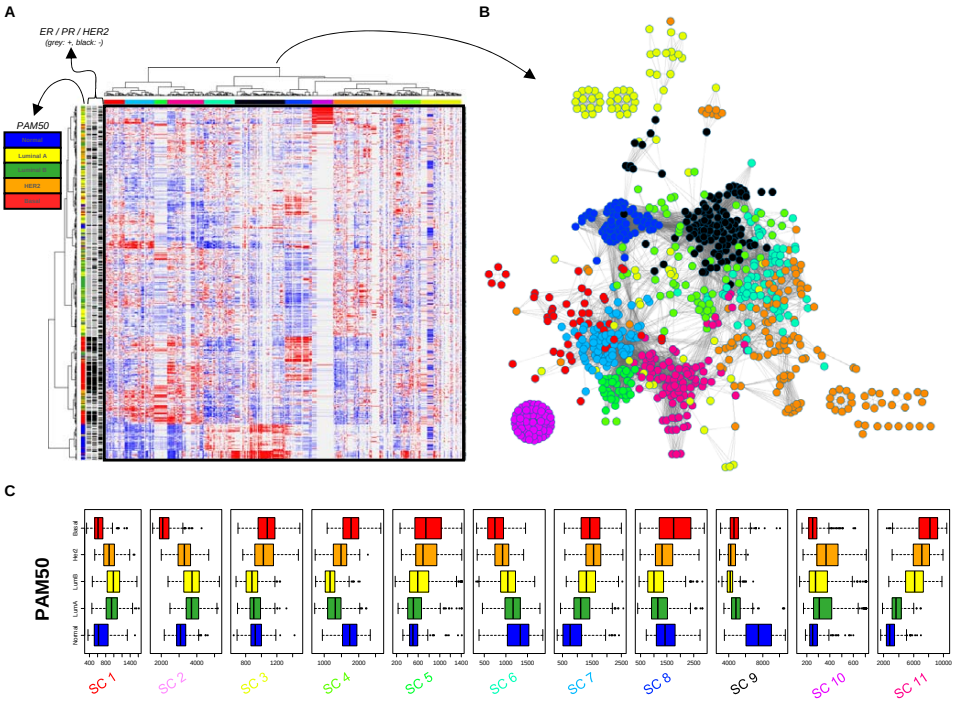

Figure 2.4: Landscape of breast cancer subsystems. (A) Heatmap of *SAS* in a matrix of subsystems vs. breast cancer samples. Labels for ER, PR, and HER2 status along with PAM50 subtypes were shown in the left panel of the heatmap. (B) Network visualization shows co-activation of the subsystems and their relationships. 11 colors indicate the subsystem clusters identified from (A). (C): Boxplots for each of 11 SCs in terms of PAM50 subtypes.

mance, which was determined to be 11. A graph was built where subsystems are nodes and weighted edges are defined by Pearson's correlation coefficient (PCC) of SAS values (Figure 2.4). In this graph, subsystems are naturally formed into clusters since correlated subsystems are connected by edges. To visualize this relationship among the subsystems, a threshold was set for PCC at 0.39 (Figure 2.5A). As a result, Figure 2.4B displayed that subsystems within SC were densely localized while maintaining distant positions to the subsystems of other SCs.

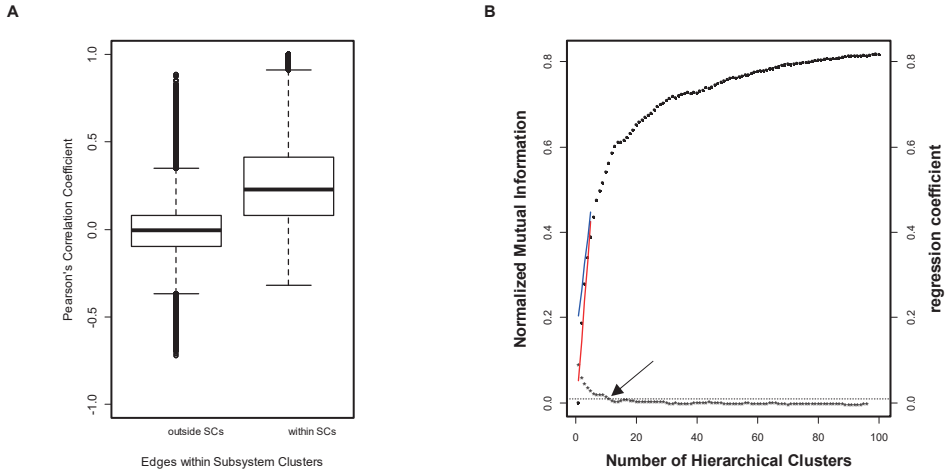

Figure 2.5: Subsystem Cluster Determination. Subsystem Cluster Determination. (A) For the 855 subsystems, there were edges connecting these subsystems within and outside 11 SCs. Pearson's correlation coefficient of SAS between two subsystems were calculated and visualized as within and outside SCs. (B) Varying the cluster numbers, normalized mutual information was calculated between the generated clusters and MCL clusters. Localized linear regression was performed using 5 consecutive points and the threshold was set at 0.01 for the regression coefficient.

Since subsystems are functional units and they are grouped into clusters

when correlated in terms of SAS, we now compare SCs and investigate the difference in 11 SCs among the subtypes (Figure 2.4C). Note that the PAM50 subtypes were well classified using SAS of subsystems (see Section 2.4.3). Thus it is interesting how well each of the 11 SCs can classify the PAM50 subtypes. Since each SC has distinct biological functions by design, if a certain SC classifies the PAM50 subtypes well, we are able to explain how much different the biological mechanisms are. To investigate the difference both among the PAM50 subtypes and among subsystems, representative value for each SC was calculated as the weighted sum of SAS. We found that SC#11 with 89 subsystems was the most effective, as higher SAS in more aggressive subtypes. Subsystems included in SC11 were mostly from the pathways related to cellular proliferation such as Cell cycle, DNA replication, and DNA repair mechanisms (Figure 2.6), of which SAS values among the subtypes share highly similar pattern. This pattern was also observed in SC5, of which 31 subsystems were related to the regulation of cyclase activity.

2.4.3 SAS revealed patient clusters associated with PAM50 subtypes.

In addition to the subsystem clusters, SAS also should be able to mirror conventional clinical annotations such as the PAM50 subtypes. By varying the number of sample clusters from 1 to 100, ARI values were calculated and visualized in Figure 2.7. As a result, the maximum number of ARI value was obtained when the number of clusters was set as 8. We call these 8 clusters as Patient Clusters (PC1 8). The distribution of PAM50 subtypes for the PCs was summarized in Table 2.1. There were eight Patient Clusters (PCs) identified by hierarchical clustering of the SAS matrix (Figure 2.4A and Table 2.1). In general, each PC was correlated with the PAM50 subtypes. For example, 89.4% of the PC6 were

There are ***89*** subsystems in SC11.

Subsystems enriched to…	# Genes	ANOVA *p*-value
Progesterone-mediated oocyte maturation	15	483.7
Oocyte meiosis	23	474.9
Cell cycle	54	439.3
HTLV-I infection	25	418.6
Metabolic pathways	61	349.4
DNA replication	28	348.8
p53 signaling pathway	13	338.7
Mismatch repair	16	310.0
Pyrimidine metabolism	30	308.9
Fanconi anemia pathway	30	307.8
One carbon pool by folate	8	304.4
Purine metabolism	31	288.5
RNA transport	36	280.3
Ribosome biogenesis in eukaryotes	39	250.4
Base excision repair	17	242.8
Nucleotide excision repair	19	235.7
…		…

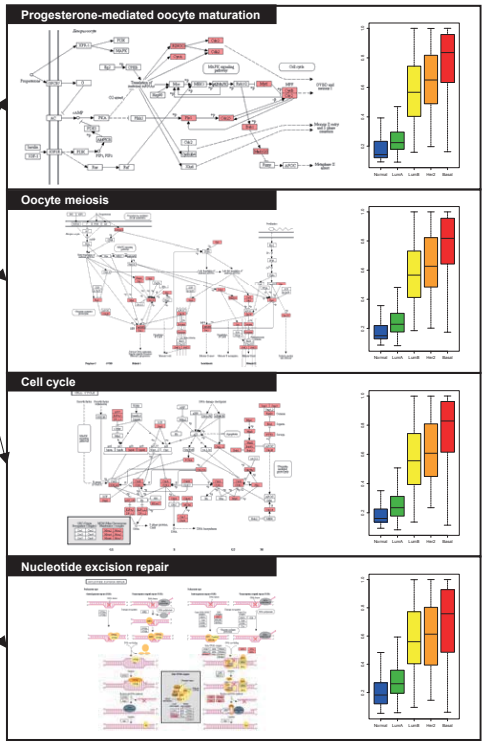

Figure 2.6: Subsystems within SC11. There are 89 subsystems enriched to KEGG pathways with corresponding ANOVA p-value of SAS among the subtypes. Right panel shows the mapping of genes in selected subsystems to their enriched KEGG pathways and boxplot of SAS among the subtypes.

Table 2.1: Association of Patient Clusters (PCs) from hierarchical clustering of SAS values with PAM50 subtypes Numbers are the number of patients.

	Normal	Luminal A	Luminal B	HER2-enriched	basal-like
PC1	0	30	22	5	0
PC2	20	52	57	28	9
PC3	27	153	64	16	3
PC4	18	35	22	8	30
PC5	4	35	125	23	4
PC6	118	8	0	0	6
PC7	6	11	13	72	12
PC8	0	0	1	5	173

Normal subtypes, while 96.6% of the PC8 were basal-like subtypes. These PCs also showed strong association with 3 receptor statuses - ER, PR, and HER2. In summary, subsystems and their activation status in SAS was able to classify the PAM50 subtypes with functional explanation.

2.4.4 Prognostic modeling by subsystems showed 11 patient subgroups with distinct survival outcome

When the survival information was set as response variable, CART generated a subsystem tree to separate the patients to maximize the discriminatory power in survival outcome. In Figure 2.8A, the whole cohort was separated into 11 patient subgroups with 10 selected subsystems out of 855 subsystems. There were five subsystems each that improved survival outcome when activated or deactivated, respectively. Survival outcome of the 11 subgroups were shown in Kaplan-Meier plot (Figure 2.8B; $p <$ 1e-16).

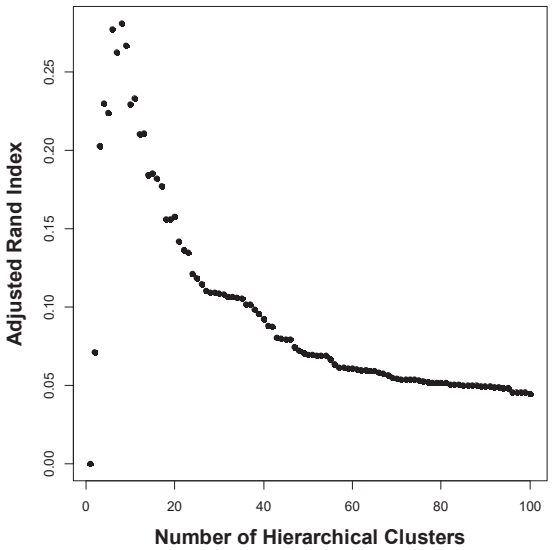

Figure 2.7: Adjusted Rand Index (ARI) varying the number of patient clusters (PCs). ARI was calculated between the generated PCs and PAM50 subtypes.

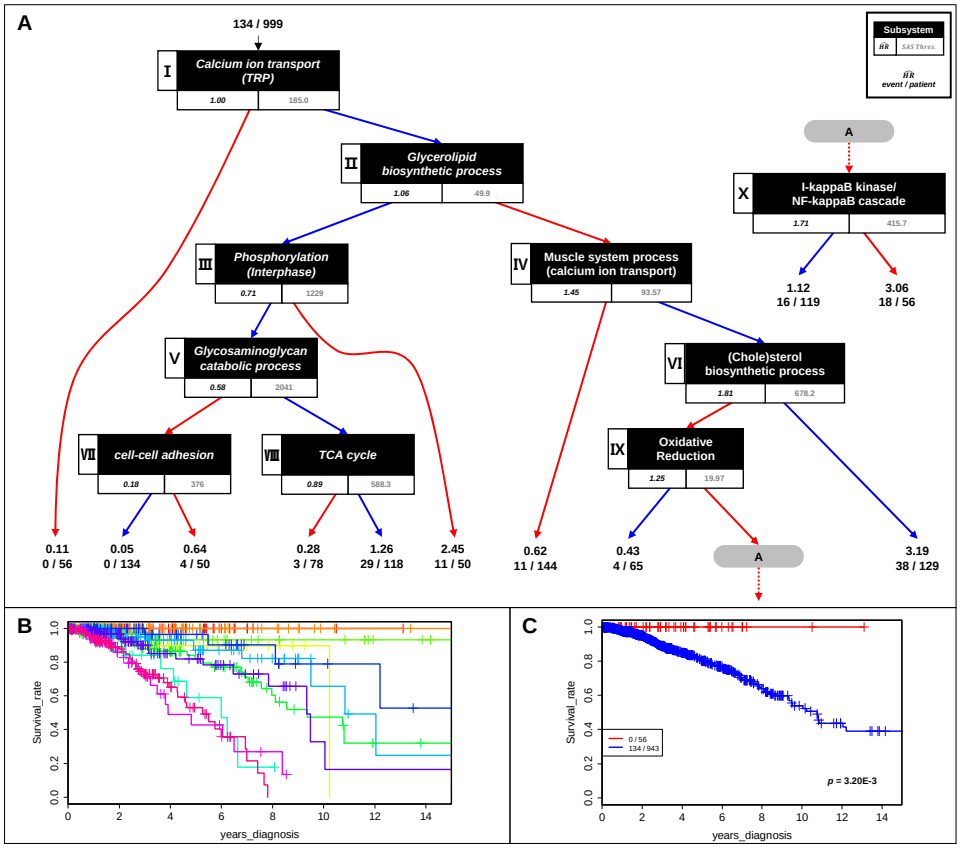

Figure 2.8: Classification and regression trees (CART) on breast cancer patients' survival prediction. (A) A branch is followed using a patient's SAS value. if it is above a specified threshold value, a red arrow branch is followed, otherwise a blue arrow branch. Each black rectangle indicates a decision point (subsystem). $\hat{HR}$ is the predicted hazard ratio of the patients in the corresponding subgroup, which is calculated for the patient subgroups at terminal nodes. The number in the bottom-left box in a subsystem node is a hazard ratio and the number in the bottom-right box is SAS threshold. (B) Kaplan-Meier plot is displayed for the 11 patient subgroups identified in (A). (C) The first subsystem identified as significant in CART divided the cohort into two patient sub-cohorts.

A subsystem "Calcium ion transport (TRP)" was chosen to be the primary target of survival classification in the subsystem tree (Figure 2.8A). This subsystem consisted of four genes (*TRPA1*, *TRPV2*, *TRPV3*, *TRPV4*). When the SAS of this subsystem was above 185.0, the hazard ratio of the corresponding patient subgroup (n = 56) was predicted as 0.11, while the hazard ratio of the other patient group (n = 943) was 1.06. To see how much degree the two subgroups differ in survival outcome, Kaplan-Meier plot was shown in Figure 2.8C ($p = 0.0032$). Kaplan-Meier plots shows discrimination power of the other 9 subsystems in Figure 2.9. Previously, up-regulation of *TRPV2* was shown to be crucial to the induction of apoptotic cell death in bladder cancer (Yamada *et al.*, 2010; Morelli *et al.*, 2012; Nabissi *et al.*, 2013). Those *TRP* genes would be the candidates of survival outcome determinant of breast cancer, which corresponds to the implications from previous studies (Prevarskaya *et al.*, 2007; Shapovalov *et al.*, 2011; Ouadid-Ahidouch *et al.*, 2013). This finding, the importance of "Calcium ion transport (TRP)" for breast cancer, needs serious further investigation since its importance is already reported in bladder cancer and its discriminatory power for survival outcome is great in TCGA data.

To evaluate the prediction power on patient survival, area under curve ($iAUC$; Song and Zhou (2008)) on receiver operating characteristic curve was measured followed by 10-fold cross-validation. Since the label for classification was survival outcome that is the combination of both days to initial diagnosis and event occurrence, a time-dependent AUC should be considered in this case. An AUC developed by Song and Zhou (2008) was chosen in this work to assess the time-dependent classification performance. It measures AUC for each user-defined time point and integrates all AUC values over the times to generate $iAUC$. `R` library of caret (Kuhn, 2015) was utilized for the 10-fold cross-validation. The performance of this system based on BRCAPIN was superior

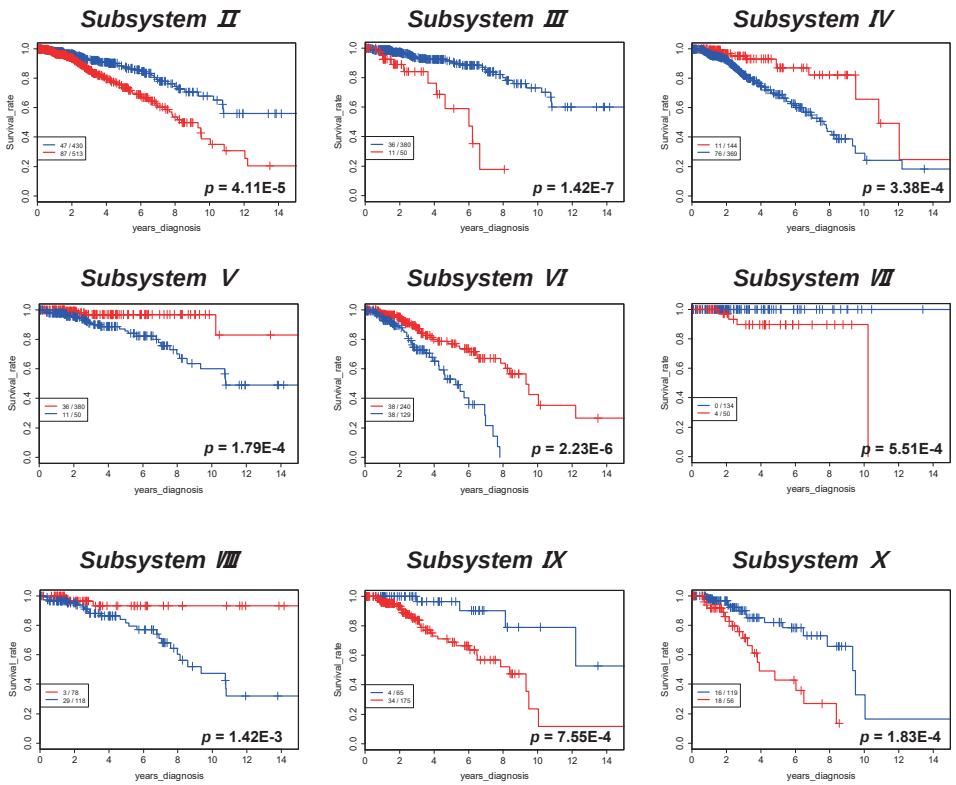

Figure 2.9: Kaplan-Meier plots for the subsequent subsystems (slowromancapii@ slowromancapx@) in Figure 2.8

Table 2.2: Comparison of predictive power of subsystems with BRCAPIN, permuted PIN, and PAM50 subtypes.

Method	*iAUC*
`SAS` (breast cancer PIN)	0.750
`SAS` (permuted PIN)	0.720
`Pathifier`	0.720
PAM50 subtypes	0.527

in classification performance to the permuted one and the PAM50 subtypes in terms of the *iAUC* (Table 2.2).

2.4.5 Relapse rate and CNVs were enriched to worse prognostic subgroups

We observed that patients' survival outcome became worse under relapse, thus the subsystem tree should also be able to significantly differentiate the relapse rate as well. There were 5 subsystems (I, IV, V, VI, X) that enriched relapse rate more than 1.5-fold (Relapse Enrichment Score; RES > 1.5) to the worse prognostic side (Figure 2.10). This generated five *relapse paths*, each navigating from the top of the subsystem tree to the corresponding subsystems.

Copy number variations (CNVs) are one of the major drivers of transcriptional perturbations. Recently, copy number loss of CDH1 gene of invasive lobular carcinoma (ILC) was cataloged as an underlying genetic driver that accelerates the depletion of Cadherin-1, which was further implicated as a major characteristic of ILC-type breast cancer (Ciriello *et al.*, 2015). Here, we examined whether there was a CNV that occurred more frequent along the paths and the occurrence of some CNVs even increased the relapse rate. We found 26 CNVs that were more likely to be present in each of the four paths. Twelve of

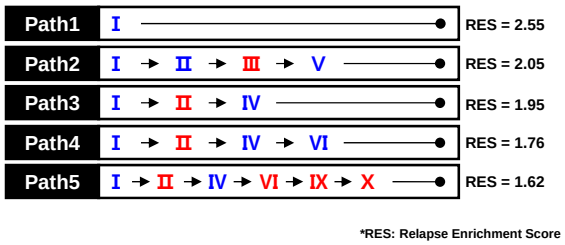

Figure 2.10: Five *relapse path*s that enriched patients' relapse rate. RES was calculated to the ratio of relapse rate of the patients who followed the full paths over the patients who followed the paths except the last subsystem.

these CNVs enriched the relapse rate within the path. In addition, fifteen CNVs increased the relapse rate under CNV occurrence while following the path.

2.5 Discussion

Gene expression information is typically interpreted in terms of biological pathways to characterize biological mechanisms underlying phenotypes. However, a pathway is a model for series of actions among molecules in a cell that leads to a certain product or a change in the cell (`https://www.genome.gov/27530687#al-1`). Thus, a pathway is often not a single functional unit. This fact is a major hurdle to perform functional analysis from transcriptome data. In this paper, we proposed a novel concept of subsystems of a pathway that can be a functional unit as shown in Figure 2.1. To define the activation score of a subsystem, we introduce a novel concept called `SAS` by combining gene expression information and the relationship between genes defined by manipulating PIN. `SAS` calculation is geometrically defined as an average area of the trapezoids for a set of pairs of genes where topological significance, i.e., closeness centrality, of each gene is used as the weight and the relationship between adjacent genes is used

as the height of a right trapezoid. Then `SAS` is defined naturally as an average of areas of trapezoids. In this way, we identified 855 subsystems out of 269 KEGG pathways. With `SAS`, we showed that the landscape of subsystem activation is distinguishable among breast cancer subtypes. We also show that the subsystems can be further grouped to eleven subsystem clusters (SCs) and each cluster represents biological functions important to breast cancer. In particular, SC11 of 89 subsystems had different `SAS` values among cancer subtypes, especially higher `SAS` values in more aggressive subtypes. Biological functions related to SC11 were cellular proliferation such as Cell cycle, DNA replication, and DNA repair mechanisms (Figure 2.6), which demonstrates the utility of subsystems and `SAS`.

The subsystems with `SAS` information can be used to model patient survival at cohort level. With the patient survival information, we built a classification and regression tree (CART). At the leaves of the tree, eleven patient groups were defined, each of which has a distinct survival outcome. In particular, there were two subgroups of which no deaths were presented (subgroups 1 and 2). This modeling leverages the tree structure; thus the tree model explains how the survival classification is done with which subsystems. In Figure 2.8, we showed which subsystems are important for patient survival prediction, thus subsystem paths were produced. Since different survival outcomes are observed at the leaves of the tree, this model can be useful for explaining heterogeneity of cancer patients and their underlying biological landscape. In the subsystem path, some subsystems were well documented in the literature and others such as a subsystem with transient receptor proteins are subsystems that suggest further investigation. Note that these subsystems in the decision path show clear difference in survival outcomes, thus subsystems in the decision path are surely important for survival prediction, just yet to be characterized. Subsys-

tems were also explanatory for relapse and CNV in addition to the patient survival outcomes.

The tree model with subsystems and SAS have several advantages over existing patient survival models. First, most existing models make prognostic decisions as either good or bad while my model provides quantitative decisions as a hazard ratio. My tree model also explain which biological mechanisms are associated with the decision in terms of subsystem decision path, which is not provided by any of the existing models including the widely used PAM50 model.

Chapter 3

Comprehensive evaluation of pathway activity measurement tools on pan-cancer data

3.1 Related works

Biological pathway is a series of actions among molecules in a cell that leads to a certain product or a change in the cell (NHGRI, 2015). Accumulation of biological knowledge over the years has produced a rich set of pathway databases (Kanehisa and Goto, 2000; Romero *et al.*, 2004; Joshi-Tope *et al.*, 2005; Schaefer *et al.*, 2008; Elkon *et al.*, 2008; Pico *et al.*, 2008; Mi *et al.*, 2012; Yang *et al.*, 2014). Then, an important question is how to utilize biological pathway information to explain transcriptome data measured from biological experiments that are designed to investigate scientifically specific questions. The most widely used approach is to investigate enriched pathways by identifying DEGs (Huang *et al.*, 2008; Rahmatallah *et al.*, 2015). However, selection of DEGs is often subjective and, more importantly, DEGs are mapped to a small fraction of

pathways. This often results in many highly expressed genes being excluded at the pathway level analysis, thus this approach does not explain pathway activities as a whole.

There are other issues when it comes to a cohort level. The steep decrease in sequencing cost accelerated the generation of cohort-level gene expression data to elucidate sample-wise molecular characteristics (Weinstein *et al.*, 2013; Lonsdale *et al.*, 2013). At the cohort level, variations among samples are high, even for experiments performed under the same condition. Thus, rather than investigating at the gene level, it is much easier to investigate sample-wise variations at the pathway level (Pang *et al.*, 2006; Bild *et al.*, 2006; Gatza *et al.*, 2010). The most important issue is how to measure the activity of a pathway in a single value and how to utilize the pathway activity values for further analyses. In fact, a number of computational tools were developed to generate abstract quantification of pathways and used them as features for characterizing underlying biological mechanisms (Efroni *et al.*, 2007; Lee *et al.*, 2008).

3.2 Motivation

There are several comparative studies of existing pathway activity measurement tools (Tarca *et al.*, 2013; Bayerlová *et al.*, 2015; Jaakkola and Elo, 2015). The main focus of these studies is to evaluate how different pathway activities are between tumor and normal samples or between different cancers. This approach of comparing pathways has the same problem as for the DEG selection task since selection of pathways can be different, depending on the criteria being used. To make comparative study robust and meaningful, it is necessary to measure the pathway activities from the transcriptome data consistently, regardless of the

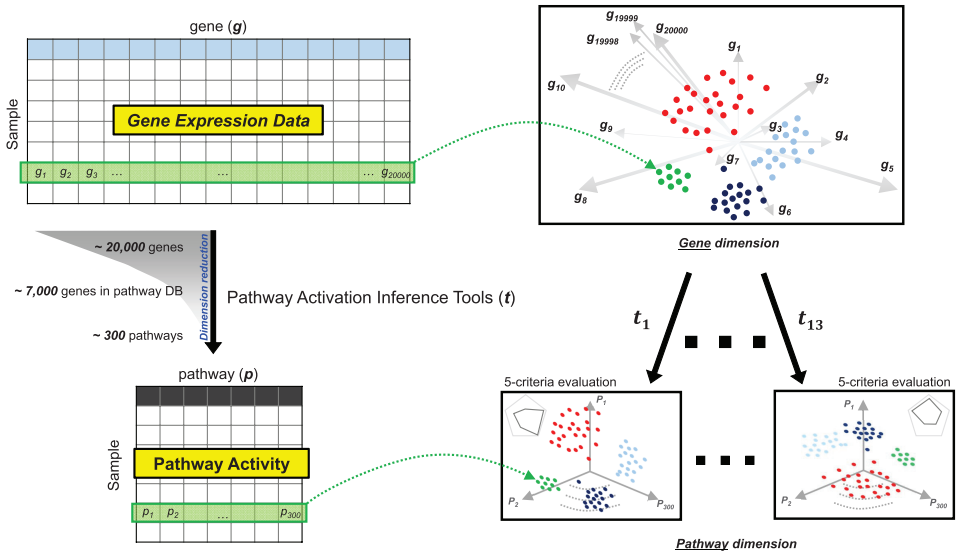

Figure 3.1: Illustration of the strategy for the comparative evaluation of pathway activity inference tools. Each tool takes gene expression data as input and produces pathway *vs* sample matrix, which can be interpreted as a mapping of cohort-level samples to the pathway dimension created by the tool, $t_i, 1 \leq i \leq 13$. Because different tools use different methods to map samples to the pathway dimension, the tools are evaluated at the pathway level using five comparison criteria.

Table 3.1: Pathway activity inference tools investigated in this study. They are listed with their abbreviated name and categories followed by reference. `Input transformation`: how to transform input gene expression data, `Labels`: use of tumor *vs* normal information into calculation of activity, `Cohort`: necessity of cohort-level data to calculate pathway activity, `Gene relations`: use of data-driven or prior gene-gene relationships within a pathway, `Scoring type`: strategy of measuring pathway activity values.

Tools	Input transformation	Labels	Cohort	Gene relations	Scoring type	Reference
CORG	z-score	N	Y	-	arithmetic	Lee *et al.* (2008)
DART	z-score	Y	Y	correlation Network	arithmetic	Jiao *et al.* (2011)
ESEA	mutual information	Y	Y	pathway structure	enrichment	Han *et al.* (2015)
GSVAdif	rank	N	Y	-	enrichment	Hänzelmann *et al.* (2013)
GSVAmax	rank	N	Y	-	enrichment	Hänzelmann *et al.* (2013)
IndividPath	explicit	Y	Y	stable pairs	enrichment	Wang *et al.* (2015)
LLR	log likelihood	Y	Y	-	arithmetic	Su *et al.* (2009)
PADOG	t-score	Y	Y	-	arithmetic	Tarca *et al.* (2012)
PathAct	median polishing	N	Y	-	arithmetic	Mogushi and Tanaka (2013)
Pathifier	SD-based normalization	Y	Y	-	PCA	Drier *et al.* (2013)
PLAGE	explicit	Y	Y	-	PCA	Tomfohr *et al.* (2005)
SAS	explicit	N	N	pathway structure	arithmetic	Lim *et al.* (2016)
ssGSEA	rank	N	Y	-	enrichment	Barbie *et al.* (2009)

PCA: principal component analysis, SD: standard deviation

evaluation criteria, and then evaluate the measured pathway activities in terms of multiple criteria. What this study outlines can be explained in two steps:

- Step 1: Mapping each sample from the gene dimension to the pathway dimension (the left panel of Figure 3.1).
- Step 2: Using multiple evaluation criteria, investigate how well samples are distinguished in the pathway dimension that is defined by each pathway activity measurement tool (the right panel of Figure 3.1).

By mapping samples or patients from the gene dimension to the pathway dimension, the transcriptome data can be interpreted more easily in a biologically meaningful way. This is because the number of dimensions reduced dramatically from about 20,000 genes to about 300 pathways that are well curated biological knowledge. Then, what are the challenges? The main challenge is only 1/3 of genes are mapped to the pathway dimension. Thus, mapping the transcriptome data of each sample to the pathway dimension is achieved with a huge information loss. Though interpretation of the transcriptome data at the pathway level is desirable, can we distinguish samples in the pathway dimension defined by each pathway activity measurement tools? Since the pathway dimension is high, it is not practical, maybe not feasible, to evaluate whether samples are correctly mapped in the pathway dimension. A practical approach is to evaluate the tools in terms of biological and clinical significance. *Therefore, the ultimate question is how the pathway activity space produced by each tool is valuable at the cohort level, e.g., patient survival and cancer subtype classification.*

In this study, we systematically compared and evaluated 13 different pathway activity inference tools (Table 3.1) based on five comparison criteria using pan-cancer data sets. Starting from measuring how well a tool maintains the

characteristics of original gene expression values, robustness was also investigated by adding noise into gene expression data. Classification tasks on three clinical variables (tumor *vs* normal, survival, and cancer subtypes) were performed to evaluate the utility of tools for their clinical applications. In addition, the inferred activity values were compared between the tools to see how much similar they are along with the scoring schemes they use.

3.3 Materials and methods

3.3.1 Pathway activity inference Tools

The list of tools compared is shown in Table 3.1. The tools were categorized based on techniques used for computing pathway activity at the cohort level: data transformation, measurement of pathway activity from gene expression data, and then evaluation of the resulting matrix of pathway activities *vs* samples.

Different transformation techniques are used to process input gene expression data: rank- or statistic-based methods in `CORG` (Lee *et al.*, 2008), `DART` (Jiao *et al.*, 2011), `GSVAdif` (Hänzelmann *et al.*, 2013), `GSVAmax` (Hänzelmann *et al.*, 2013), `LLR` (Su *et al.*, 2009), `PADOG` (Tarca *et al.*, 2012), `PathAct` (Mogushi and Tanaka, 2013), and `ssGSEA` (Barbie *et al.*, 2009); mutual information based transformation in `ESEA` (Han *et al.*, 2015); and standard deviation based transformation in `Pathifier` (Drier *et al.*, 2013). Instead of using data transformation, `IndividPath` (Wang *et al.*, 2015), `PLAGE` (Tomfohr *et al.*, 2005), and `SAS` (Lim *et al.*, 2016) used explicit gene expression values directly to pathway activity calculation.

After the data transformation step, an important issue is how to measure the activity of a pathway from gene expression data. This task can be viewed as

Table 3.2: TCGA gene expression data sets for pathway analysis. Data set is selected when the number of corresponding normal samples is more than 30.

Code	Source	Samples (normal + tumor)
BRCA	breast invasive carcinoma	1212 (112 + 1100)
COAD	colorectal adenocarcinoma	326 (41 + 285)
HNSC	head and neck squamous cell carcinoma	566 (44 + 522)
KIRC	kidney renal clear cell carcinoma	606 (72 + 534)
KIRP	kidney renal papillary cell carcinoma	323 (32 + 291)
LIHC	liver infiltrate hepatocellular carcinoma	423 (50 + 373)
LUAD	lung adenocarcinoma	567 (50 + 517)
LUSC	lung squamous cell carcinoma	552 (51 + 501)
PRAD	prostate adenocarcinoma	550 (52 + 498)
STAD	stomach adenocarcinoma	450 (35 + 415)
THCA	thyroid carcinoma	568 (59 + 509)

aggregating expression values of genes in a pathway to a single activity value of the pathway. Most of the tools use arithmetic aggregation of gene level values, or enrichment of gene level perturbations. Tools, such `PLAGE` and `Pathifier`, use PCA to create a feature space in calculation of pathway activity values.

3.3.2 Data sets

We used The Cancer Genome Atlas (TCGA) RNA-seq data sets and corresponding clinical information. The data sets were downloaded from Firebrowse (`http://firebrowse.org/`). We used the RSEM-processed normalized gene expression data sets from each of the cancer types by the name '`illuminahiseq_rnaseqv2-RSEM_genes_normalized`'. Eleven TCGA cancer projects

(BRCA, COAD, HNSC, KIRC, KIRP, LIHC, LUAD, LUSC, PRAD, STAD, and THCA) with sufficient number of normal samples (n $\geq$ 30) were chosen for this study (Table 3.2). Subtype information for the cancers (COAD, PRAD, and STAD) were from the original articles. A list of BRCA subtypes is generated in this study by the PAM50 classification method (Parker *et al.*, 2009) using log2-transformed RNA-seq data (Table 3.3) since the original article of BRCA did not include subtype information for many samples (TCGA *et al.*, 2012).

3.3.3 Pathway database

We used 314 pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa and Goto (2000)) database. Graph structural information of the pathways were retrieved using `ESEA` library in `R`.

3.3.4 Notations

The following notations will be used throughout this paper.

gene expression matrix We used a gene expression matrix for each cancer type to run each of the methods to infer pathway activity. The data sets are denoted as $\mathbf{D}$ = { BRCA, COAD, ..., THCA }. The gene expression matrix for cancer $d \in \mathbf{D}$ is denoted as $M_d \in \mathbb{R}^{n_d} \times \mathbb{R}^{g_d}$ where n_d is the number of samples and g_d is the number of genes in data d.

tools A set of tools are defined here as $\mathbf{T}$ = {CORG, DART, ..., ssGSEA} and each tool as t.

pathways We denoted a set of KEGG pathways as $\mathbf{P}$ = { p_k, $1 \leq k \leq 314$ }. Since the number of pathways that can be inferred by tools varies, each was

Table 3.3: Cancer data set with subtypes. The number assigned to each subtype indicates the number of samples for the corresponding subtype. Subtypes with few assigned samples were removed from the scope of analysis.

Cancer	Total	List of subtypes	Source
BRCA	1208	LumA (321), LumB (298), HER2 (161), Basal (234), Normal (194)	[1]
COAD	259	CMS1 (42), CMS2 (82), CMS3 (38), CMS4 (72), NOLBL (25)	Guinney *et al.* (2015)
PRAD	317	ERG (152), ETV1 (28), ETV4 (14), SPOP (37), other (86)	Network *et al.* (2014)
STAD	280	CIN (139), EBV (25), GS (54), MSI (62)	Network *et al.* (2015)

[1] Breast cancer subtype was obtained for each sample in this study by using RNA-seq data and PAM50 (Parker *et al.*, 2009).

defined as p_t.

pathway activity matrix We define an inferred pathway activity matrix for data set $\mathtt{d} \in \mathbf{D}$ by a set of tools as $A_{d,t} \in \mathbb{R}^{n_d} \times \mathbb{R}^{p_t}$.

3.4 Comparative approach

3.4.1 Radar chart criteria

This study introduces five criteria of comparison to evaluate the tools:

- Preservation of distance between samples
- Robustness against noise
- Classification on tumor *vs* normal information
- Classification on survival information
- Classification on cancer subtypes

Scores from each of the criteria in the radar chart are min-max normalized.

Criterion 1: Preservation of distance between samples

Since all tools are designed to use RNA-seq data, it will be interesting to see how well sample differences at the gene level are preserved at the pathway level (Vitali *et al.*, 2017). For $d \in \mathbf{D}$ and $\mathrm{t} \in \mathbf{T}$, the distance at the gene level is defined as d_{M_d} in terms of gene expression values M_d and the distance at the pathway level is defined as $d_{A_{d,t}}$ in terms of activity values $A_{d,t}$ that were produced by each tool t. We then summarize these values as distance preservation (DP_t) for each tool as the reciprocal of the weighted average of Mean Squared Error (MSE) between d_{M_d} and $d_{A_{d,t}}$ as follows:

$$d_{M_d}(x,y) := Sim(M_d^x, M_d^y), \tag{3.1a}$$

$$d_{A_{d,t}}(x,y) := Sim(A_{d,t}^x, A_{d,t}^y), \tag{3.1b}$$

$$MSE(d,t) := \frac{1}{\binom{n_d}{2}} \sum_{1 \le x,y \le \mathtt{n_d}} \left\{ d_{M_d} - d_{A_{d,t}} \right\}^2, \tag{3.1c}$$

$$DP_t := \frac{n_d}{\sum\limits_{d \in D} n_d \cdot MSE(d,t)}, \tag{3.1d}$$

where $M_d = [M_d^1 \dots M_d^{n_d}]$, $A_{d,t} = [A_{d,t}^1 \dots A_{d,t}^{n_d}]$ and Sim is cosine similarity such as

$$Sim(u,v) := \frac{u \cdot v}{\|u\| \, \|v\|}, \tag{3.2}$$

for non-zero vectors u, v.

Criterion 2: Robustness against noise

Generation of perturbed data : All data contain some degree of errors, thus we performed experiments on how much tolerant each tool is to noise or error in the gene expression data each tool is. When pathway activity values are measured using BRCA data, impact of noise on RNA-seq data was partly assessed in previous studies (Li *et al.*, 2014; Jia *et al.*, 2017). First, gene expression value m:= M_{BRCA}^{ij} from $M_d = (M_{BRCA}^{ij})$ was log2-transformed to generate z (Equation 3.3a). Then perturbed expression value z' was generated based on the normal distribution with a mean value of z and standard deviation of $ze/100$, where e is the perturbation factor to inflate the level of applied noise ranging from 1 to 200 ($e \in \{1, 2, 3, 4, 5, 10, 30, 50, 100, 200\}$; Equation 3.3b). This is based on the previous research showing that log2-transformed RNA-seq FPKM or RPKM values are shown to be normally distributed (Bengtsson *et al.*, 2005). The error generation was performed 30 times for each value of e to generate 300 matrices in total. For example, perturbed gene expression values of m

with the factor e at iteration i is termed as $m'(e,i)$ (Equation 3.3c). Thirteen tools were then run for each of the matrices to calculate perturbed pathway activity values. We define perturbed activity value matrix of tool t as $A_t^{'}(e,i)$ (abbreviation: $A_t^{'}$).

$$z := log_2(m+1), \tag{3.3a}$$

$$z'(e,i) \sim \mathcal{N}(z, (ze/100)^2) \tag{3.3b}$$

$$m'(e,i) := 2^{z'(e,i)} - 1. \tag{3.3c}$$

Measuring robustness : Robustness to noise of a tool was defined as the degree of correlation of pathway activity values from between noisy data and the original data. Given an original activity matrix and a perturbed activity matrix A'_t, a mean Spearmans's correlation coefficient value $Spearman^{p_k,t}(e,i)$ was computed (Equation 3.4a). $Spearman^{p_k,t}(e,i)$ was averaged for i and then weighted-summed up over e followed by averaging on the set of pathways to generate $Robustness(t)$ (Equation 3.4b).

$$Spearman^{p_k,t}(e,i) = \frac{\mathbf{cov}\Big(\text{rank}(A_t^{\cdot j}(e,i)), \text{rank}(A_t^{'\cdot j}(e,i))\Big)}{\sigma_{\text{rank}(A_t^{\cdot j}(e,i))}\sigma_{\text{rank}(A_t^{'\cdot j}(e,i))}} \tag{3.4a}$$

$$Robustness(t) = \frac{1}{p^t}\sum_{p^t}\sum_{e}\frac{\sum_{i=1}^{30} Spearman^{p_k,t}(e,i)}{\sum_{i=1}^{30} e} \tag{3.4b}$$

where rank(v) is a rank-transformation of a given vector v.

Criterion 3: Classification - Tumor *vs* Normal

The next experiment was to measure how well the inferred pathway activity values can be used to classify between tumor and normal samples (Table 3.2). We performed tumor *vs* normal classification experiments using Gaussian Naive Bayes (`GaussianNB`) and Random Forest (`RF`) classifiers using `sklearn` library (Pedregosa *et al.*, 2011) in `Python`. Since there are a lot more tumor samples than normal samples, i.e., highly unbalanced data, weighted F1 score was used for performance evaluation. Weighted F1 score calculates F1 score (also called F1 measure) weighted by the number of samples in each of subtypes. Meanwhile the original F1 score is a measure of accuracy by calculating the harmonic mean of precision and recall.

Criterion 4: Classification - Survival Information

Survival data from clinical information were also used to evaluate the inferred pathway activity values. We utilized `rfsrc` library to measure concordance index in `R` followed by a 5-fold cross-validation. The folds were generated by using `createFolds` function of `caret` library (Kuhn *et al.*, 2008) in `R`.

Criterion 5: Classification - Cancer Subtypes

Subtype classification of cancer is a difficult classification task even when gene expression information of all genes are used. Thus, measuring subtype classification accuracy using the inferred pathway activity values will be challenging and meaningful. Cancer subtypes to be analyzed are listed in Table 3.3. These subtypes are mostly based on molecular signatures from high-throughput sequencing data. We performed subtype classification experiments as in tumor *vs* normal classification task.

3.4.2 Similarity among the tools

It is also interesting to see how much the inferred activity values by each tool `t` are similar to each other. We calculated similarity using the pathway activity values on multiple cancer data sets (Figure 3.2). Since some tools produce activity values for a subset of pathways, we used consensus pathways as the intersection of the pathway sets for all the activity matrices $A^{d,t}$ (Figure 3.2). We measured pairwise cosine similarity $Sim(t_1, t_2)$ of two tools for each cancer data *d*. We measured pairwise cosine similarity $Sim(t_1, t_2)$ of two tools for each cancer data *d*. Aggregating for all cancer data to compute the similarity of two tools, $Association(t_1, t_2)$, is a weighted sum of $Sim(t_1, t_2)$ according to the data size (Equation 3.5). To see how similar the 13 tools are in terms of pathway activity values, hierarchical clustering was performed by using the `Average` clustering option in `hclust` library of `R`.

$$Association(t_1, t_2) = \frac{\sum_d Sim(t_1, t_2)_d \times n^d}{\sum_d n^d} \tag{3.5}$$

3.5 Results

3.5.1 Distance preservation

The distance between samples was defined as a measure to characterize the activity values. The two distance metrics, d_{M_d} from the raw expression data and $d_{A_{d,t}}$ generated by a tool t, were compared for each cancer d. By taking a reciprocal of the weighted average of MSE for each tool, termed as DP_t, we estimated how much a tool maintains the original characteristic from RNA-seq data (Equation 3.1). The smaller the distance between RNA-seq data and activity values, the greater the DP_t value is. This means that the sample space of the inferred pathway activity values is closer to that of the original gene expres-

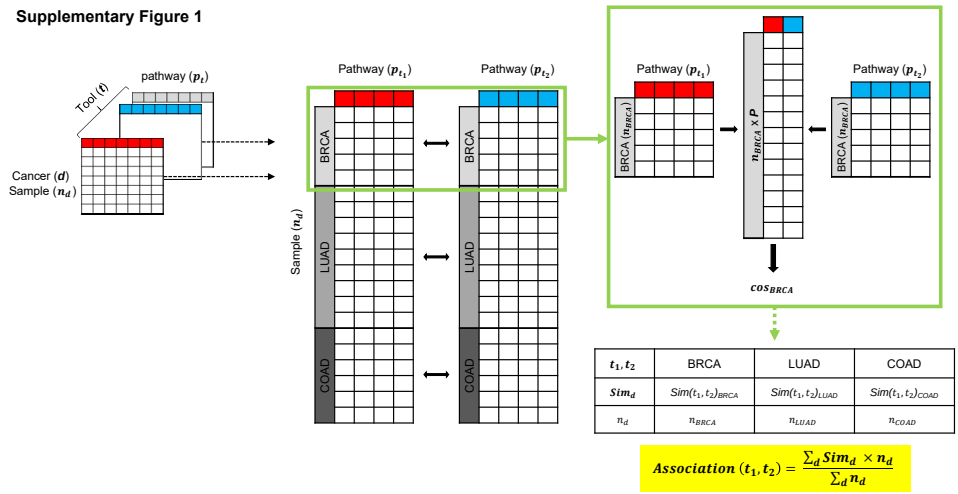

Figure 3.2: Schematic diagram of measuring similarity between pathway activity inference tools. Cosine similarity among the tools was first calculated for each cancer data set d. These cosine similarity values Sim_d are then averaged over all cancer data sets weighted by sample size n_d to calculate Association(t_1, t_2).

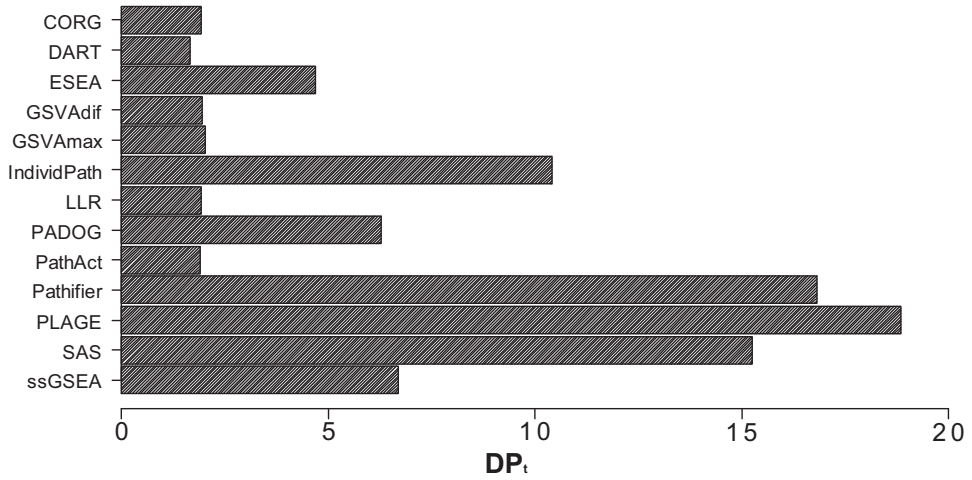

Figure 3.3: DP_t values for the tools. Distance preservation (DP) was calculated by measuring the difference between the distances between samples in gene expression data and in pathway activity values of a tool. The greater the DP_t value, the more data characteristic was sustained in pathway activity inference.

sion. Note that only a small portion of genes, approximately 1/3 of all genes, are included in pathways. Then, when tools measure the activity of pathways, it is interesting to see how well distance between samples in terms of pathway activity values can preserve the characteristics of the original data, i.e., all genes.

As it can be seen in Figure 3.3, DP_t for `PLAGE` is the largest among all the tools ($RC_{PLAGE} = 18.20$). It seems that the use of PCA by `PLAGE` is effective in preserving characteristics of the original gene expression data while transforming the transcriptome data at the pathway level. `Pathifier` performed the second best. `Pathifier` is also in line with `PLAGE` in using PCA for the data transformation which is likely effective in preserving the characteristics of the original data. `SAS` that uses the explicit gene expression information and performed next best.

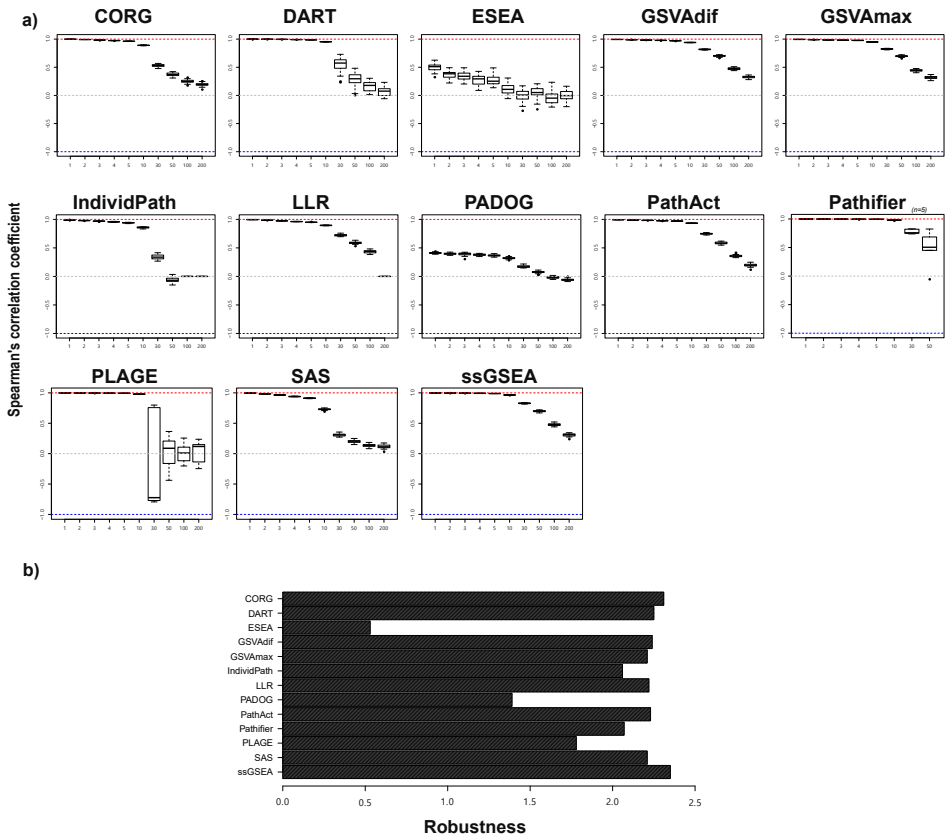

Figure 3.4: Effect of noise in input gene expression data. (a) Boxplots of Spearman's correlation coefficient ($Spearman^{p_k,t}(e,i)$) values for `hsa04110:Cell cycle` pathway across the perturbation factor e. Each subplot was drawn for 30 repeated experiments on i. Dotted lines with red and blue at $Spearman^{p_k,t}(e,i)$ = 1 and -1 indicate theoretical maximum and minimum, respectively. (b) Barplot of $Robustness(t)$ values across the tools.

Table 3.4: Pathway databases used in this study

Database	Number of unique entities			Source	
	pathway	gene	interaction	gene	interaction
KEGG	314	7,200	47,589	https://www.kegg.jp/kegg/rest	
NCI	165	2,495	24,391	http://www.ndexbio.org	
REACTOME	675	6,025	11,929	(Liberzon *et al.*, 2011)	https://reactome.org/download-data

Additional comparative analyses of the tools were performed to investigate whether there is any dependency on using other pathway databases such as NCI and Reactome (Table 3.4). Marginal but statistically insignificant changes in DP_t values were observed between the databases (Table 3.4). As KEGG database includes more genes and interactions than the other two databases do, comparative analysis throughout the study was performed using KEGG (Table 3.4).

3.5.2 Robustness against noise

The goal of this experiment is to measure the tolerance of each tool against noise in the data. The test metric is how similar the perturbed activity values (A'_t) are to the original activity values (A_t) in terms of ranks. The tools were similar in $Spearman^{hsa04110,t}(1,1)$ values at low level of noise. Most of the tools show almost perfect correlation to the original data at 0.99, `PADOG` showed relatively low correlation. Steady decrease in $Spearman^{p_k,t}(e,i)$ was observed in most of the tools, while it is the most dramatic between $e \geq$10. $Spearman^{hsa04110,SAS}(e,1)$ showed a decreasing pattern as the level of applied noise increases from 0.996 at e=1 to 0.173 at e=200. The overall trend over different perturbation levels (e) is depicted in Figure 3.4a. This validates that whichever details in pathway activity inference uses can sustain a certain level of robustness to random introduction of noise in input gene expression data.

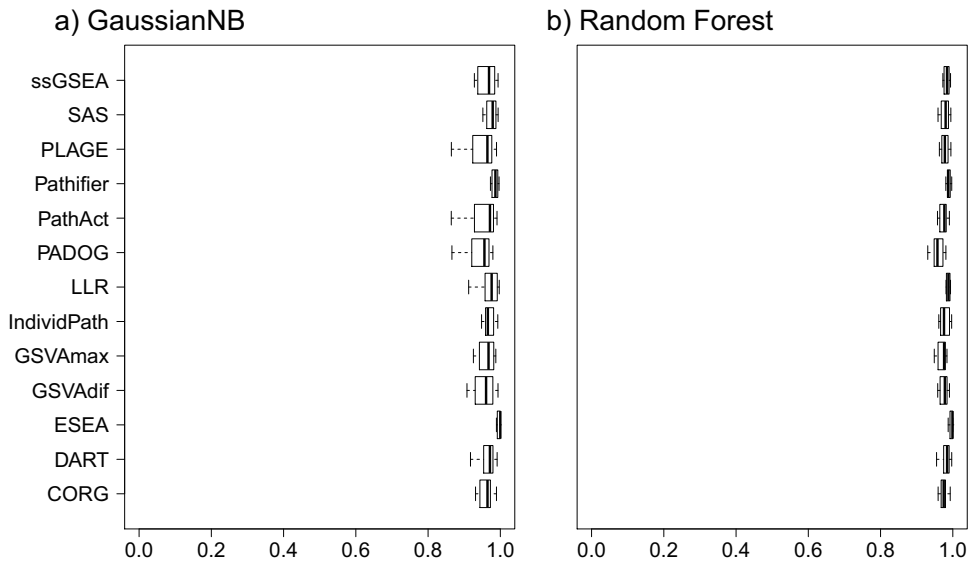

Figure 3.5: Performance Comparison of the tools to classify tumor *vs* normal samples by classifiers: `GaussianNB` and `RF`. The metric to evaluate the performance is weighted F1 score. Each box in boxplot was built on across the cancer data sets.

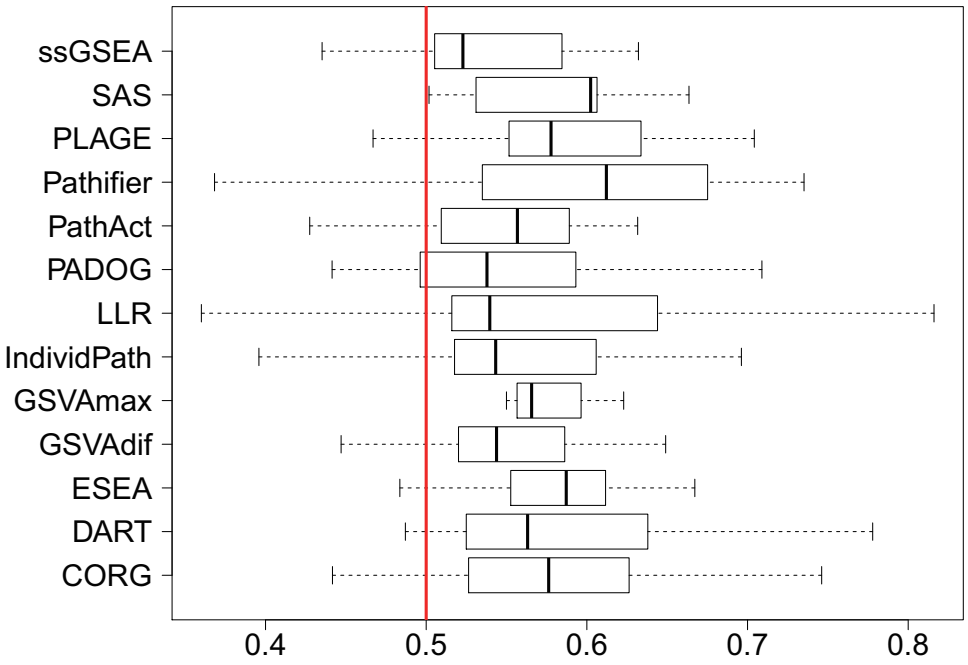

Figure 3.6: Performance comparison of tools to classify survival information by c-index. The vertical red line indicates the performance at random. Each subplot was depicted for the performance of a tool on pan-cancer data set.

To collectively summarize the above experiment results as a single measurement value, we measured $Robustness(t)$ throughout the pathways as in Equation 3.4 (Figure 3.4b). As the theoretical upper bound of $Robustness(t)$ is approximately 2.45, eight tools (`CORG`, `DART`, `GSVAdif`, `GSVAmax`, `LLR`, `PathAct`, `SAS`, `ssGSEA`) are shown to be highly robust to noise. This indicates that using raw expression values in the inference pipeline is advantageous to stay robust against introduction of noise. Meanwhile, tools such as `IndividPath`, `PADOG`, and `PLAGE` are below average to be stable. `PADOG` seems suffering from down-weighting genes that are assigned to more than a single pathway, thus letting itself sensitive to noise.

We note here that both `DART` and `IndividPath` suffer from decrease in the number of inferred pathways as the level of applied noise increases (Figure 3.7). This is due to their own mechanism of leveraging pathway database that uniquely apply denoising algorithm of template relational structure within a pathway. As noise increases, the algorithm would drive itself to remove elevated noise and less relationships would survive. This can be considered as both minor drawback and advantage for both tools by losing information from removed pathways, in the meantime by refining robust pathways. `LLR`, at extreme level or noise (here at e=200), forces all the perturbed activity values either 1 or 0 as it uses cumulative probability of each gene expression value to calculate log-likelihood (data not shown).

3.5.3 Classification: Tumor *vs* Normal

Here, the task is to measure how well tumor *vs* normal samples can be distinguished by using the inferred pathway activity values. We used two machine learning classifiers (`NaiveBayes` and `RF`) from `sklearn` library in `Python`. All tools performed very well, achieving a weighted F1 score of 0.9 or higher. The

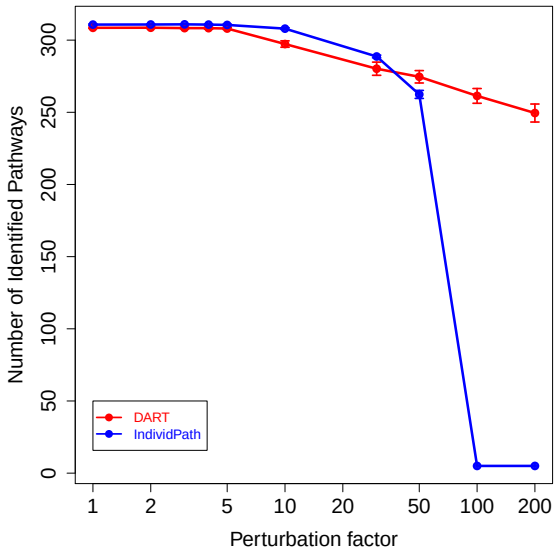

Figure 3.7: The number of pathways to infer activity values in `DART` (red) and `IndividPath` (blue) with respect to the varying noise level e.

performance of all tools is summarized in Figure 3.5. As aforementioned, only about 1/3 of all genes are included in pathways, thus the high performance in the tumor *vs* normal classification task indicates that all tools are successful in summarizing transcriptome data at the pathway level.

`ESEA` performed the best with a weighted F1 score of 0.996. It is also interesting that `ESEA` showed the best performance at perfect classification for both of the classifiers in cancer data sets (BRCA, KIRC, LIHC, LUAD, LUSC, and THCA).This seems because `ESEA` projects the distance between samples with respect to normal samples in terms of the difference in mutual information. In addition, variable importance (VI) from RF classifier was measured to rank pathways in order of significance to classification for each cancer type and tool. This revealed that many pathways commonly detected by multiple tools were discovered to be valuable (Data not shown).

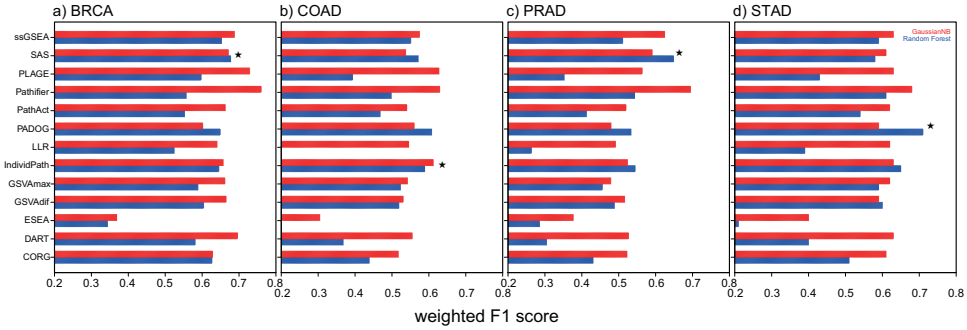

Figure 3.8: Performance Comparison of the tools to classify a) BRCA, b) COAD, c) PRAD, d) STAD cancer subtypes. Red bar shows the the performance of `GaussianNB` classifier, while blue one shows that of `RF` classifier. For a tool at maximum performance on average for both classifiers is marked with $\star$. Note that the scale ranges from 0.2 to 0.8 to maximally distinguish the comparison (Performance of `LLR` in COAD was 0.197.)

3.5.4 Classification: survival information

The task is to measure how well the inferred pathway activity values predict patient survival. We built a RF model using pathway activity values and the goodness of fit of the survival model was measured by concordance index (c-index).

Figure 3.6 shows boxplot of c-index values of the tools across pan-cancer data set. Overall, all tools were successful in predicting patient survival for most of patients since the first quartile is above the random performance, except `PADOG`. In terms of median, `Pathifier` performed best in this task. In terms of all cancer data sets, `SAS` performed best since no survival prediction was below the random performance. It was notable that 7 tools nominated 'Cell cycle pathway (hsa04110)' in KIRP data at an average rank of 3.7.

3.5.5 Classification: cancer subtypes

The presence of heterogeneity among the samples within a single cancer data hampers the interpretation of genuine characteristic of a disease under study. As such, there have been several studies to develop sub-classes to better understand and provide treatment of the disease called 'subtypes'. Since they are partially built on genome-wide omics data such as gene expression, we investigated whether there is further capability of pathway activity on differentiating subtypes within each cancer. As a result, it was more distinctive among the tools in classifying subtypes than previous tasks - tumor *vs* normal and survival information.

The performance of all 13 tools is summarized in Figure 3.8. `SAS` performed best in classification of BRCA and PRAD cancer subtypes on average of both classifiers. `IndividPath` and `PADOG` showed the best performance in COAD and STAD cancer subtypes, respectively. `ESEA` showed quite low classification performance for all the cancer types, which we conjecture that `ESEA` highly focus on the difference between tumor and normal information, thus is unable to capture the distance between tumor samples.

We note that the performances of the tools vary significantly for the subtype classification task, compared to the other four evaluation experiments. This is because the subtype classification task is the most challenging. More study such as on simulation is needed to understand why the performance variations are high for the subtype classification task.

3.5.6 Similarity among the tools

Additional investigation was performed on how much the tools are similar in their pathway activity values, we calculated cosine similarity on each cancer

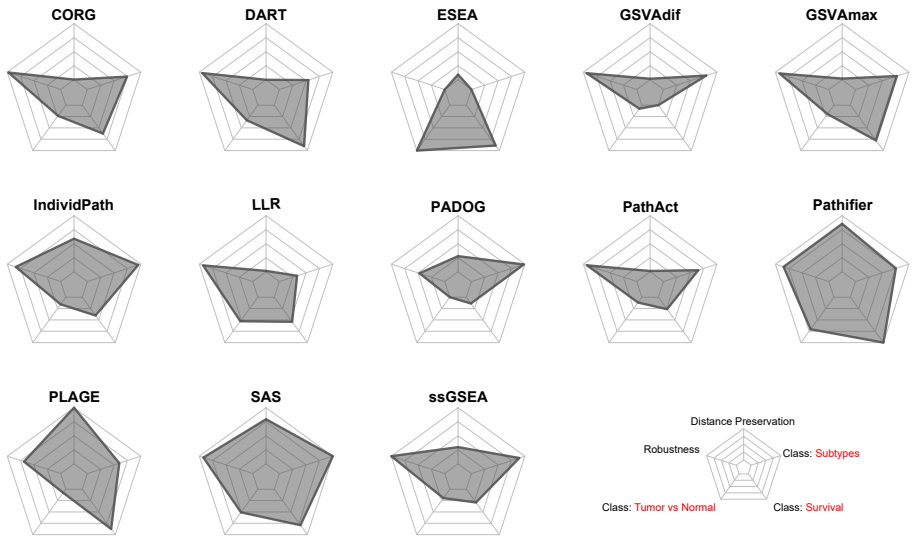

Figure 3.9: Radar charts of the pathway activity inference tools for 5 comparative criteria. Each vertex counter-clockwise from 12 o'clock of pentagonal radar chart indicates distance preservation, robustness against introduction of noise, classification performance on tumor *vs* normal information, classification performance on survival information, and classification performance on subtype information. The highest scorer for each standard was marked with blue star at corresponding vertex.

data set (Figure 3.10). It is reasonable that the similarity between `GSVAmax` and `GSVAdif` was the greatest as they share the same background framework. Their difference arises from the point where the former focus on the maximum of whether deviation from zero, while the latter focus on both positive and negative deviations (Hänzelmann *et al.*, 2013). The tools (`GSVAdif, GSVAmax, CORG`, and `PathAct`) were clustered together in a greater scale, except for `ssGSEA`, as they share the same characteristic of both using tumor *vs* normal information and cohort-level input data.

3.6 Discussion

In this study, we compared pathway activity inference tools in terms of five evaluation criteria. The performance comparison results are summarized as a radar chart (Figure 3.9).

`Pathifier`, `PLAGE` and `SAS` performed the best in terms of preserving the characteristics of the original gene expression data. It seems that using PCA was effective in preserving characteristics of the original gene expression data while transforming the transcriptome data at the pathway level. Transformation of raw input data (e.g. rank transformation) was able to sustain sample distance even by distorting the original data structure.

For the tolerance to noise, `ssGSEA` was the most robust. Many other tools were also comparable in robustness to `ssGSEA`. This reinforces the utility of pathway activity inference tools as subtle fluctuation in the input gene expression data can even out when genes of rich biological context are grouped into a set of pathways.

In terms of tumor vs normal classification, `ESEA` performed best, achieving the perfect classification on all six cancer data sets. All tools achieved very good performance, which shows that pathway activity inference from gene expression data is done well.

For the patient survival prediction, all tools made reasonably good prediction above random prediction performance. In average, `Pathifier` performed best for all patients. In terms of all cancer data sets, `SAS` performed best since no survival prediction was below the random performance.

For the cancer subtype classification, `SAS` and `PADOG` showed best. This is because both tools take into account the gene importance in terms of pathway structural context for pathway activity inference.

As we can see in the three classification tasks, majority of the tools show subtle difference in performance. However, for the preservation of the original data structure and robustness to the noise, it seems reasonable to choose the tools with balanced radar in Figure 3.9, such as `SAS`, `Pathifier`, and `IndividPath`.

There are pathway entities representing the same biological phenomena in different pathway databases. Five representative pathways from each pathway database were chosen and compared both on the constituent genes and their pathway activities (Data not shown). Since the pathways commonly share few genes for the 3 pathway databases, their activity values were also significantly different (p-value $<$ 2e-16 by ANOVA). It was also interesting that `IndividPath`, `Pathifier`, and `SAS` managed to preserve distance profile whichever pathway database was used (Data not shown).

In addition, 'PI3K/AKT/mTOR signaling pathway' was chosen from (Ersahin *et al.*, 2015) to compare with the corresponding pathway in KEGG: hsa04151. They shared some genes in common (Jaccard Index = 0.18). Activity values from 10 tools revealed that enrichment based tools (`GSVAdif`, `GSVAmax`, `ssGSEA`) produce significantly different activity values between the two pathways (Data not shown). This indicates that such enrichment-based techniques are highly sensitive to the way how a pathway is defined. Other three tools (`ESEA`, `IndividPath`, `PADOG`) were not applicable to the given pathway as they require a collection of pathways in their analysis pipeline but Ersahin *et al.* (2015) contains only a single entity.

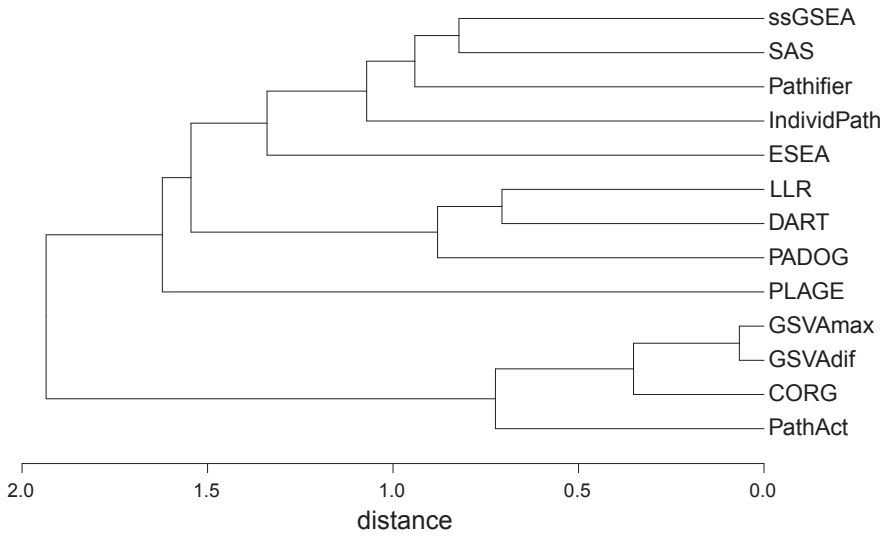

Figure 3.10: Similarity among the tools. Pairwise distance was calculated from pan-cancer cosine similarity values (See Approach).

Chapter 4

A cloud-based system of pathway activity inference tools using high-throughput gene expression data

4.1 Related works

Biological pathway is information accumulated over the decades to categorize genes according to the functional or molecular characteristics Haeussler *et al.* (2018); Sayers *et al.* (2019). Pathway databases have been extensively used along with high-throughput gene expression profiles to investigate biological functions of a disease Schadt *et al.* (2005); Draghici *et al.* (2007).

There are several web-based platforms addressing how to identify significant pathways out of user-provided list of DEGs or genes of interest Huang *et al.* (2007); Liberzon *et al.* (2011); Xie *et al.* (2011); Wang *et al.* (2013, 2017). They extensively integrated various databases to help users find the significance from the list of genes. Such platforms perform statistical analysis on input genes

against database using their basis strategy. Hosack *et al.* (2003); Subramanian *et al.* (2005).

It becomes more important to build a platform or a tool with easy access and visual inspection of the results. Such examples include Cytoscape Kohl *et al.* (2011) and Enrichr Chen *et al.* (2013); Kuleshov *et al.* (2016). Cytoscape offers a new type of platform offering various graphical visualization techniques. It can be extended by using plugin software to perform additional analysis of the entities. ClueGO and CluePedia are examples that investigate the genes in in terms of biological pathways Bindea *et al.* (2009, 2013). Enrichr is a platform that significantly extended the scope of knowledge-based analysis by integrating a large number of gene sets from more than a hundred of libraries. Its highly efficient environment with user-friendly layout paved a new way of demonstrating pathway-level analysis.

4.2 Motivation

Despite popularity in identifying significant pathways, current pathway databases still recruit approximately one-third of the whole genes being analyzed in expression profiles (Table 3.4). To address the issue, pathway activity measurement was suggested. Pathway activity values inferred from gene expression values were used as input for machine learning model that considers pathway as feature to learn characteristics of the given data (Figure 3.5).

Moreover, individual pathway activity inference tools use their own libraries which often discourages its widespread use. Some tools built their standalone libraries in Bioconductor Gentleman *et al.* (2004), users still need to follow thorough instructions to get final outcome. It is more difficult for those who are not prepared with computational language skills or prior knowledge on

bioinformatics, demanding a great deal of time and effort. It is also impossible to be aware which of the tools are more appropriate to their analysis without systematic comparisons on the tools. Therefore, a framework is necessary that help users utilizing pathway activity tools as well as comparing the results in the context of input data.

In this work, a web-based system called PathwayCloud is introduced to provide an environment where users can calculate pathway activity inference. The platform imports both gene expression profiles and sample information from a user. PathwayCloud runs the user-selected tools on a cloud/web server, analyzes the data and visualizes the results on a web-based platform. It performs pathway activity calculation and visualization of the results under interactive environment, also directing the results to KEGG pathway images of involving genes. Comparative analysis on the selected tools is also available on user-provided data to help users understand from which of the tools the result is more reliable.

4.3 Implementation

Web-based system PathwayCloud is based on an environment where users can easily use pathway activity inference tools. The system is written in JAVA and JSP based on Spring framework and deployed on Apache web server. Visualization of the results and graphs are based on D3 JavaScript library. To resolve the limitation on physical computational resources, PathwayCloud provides a cloud environment image. This enables the inference of pathway activities from RNA-seq data with improved flexibility and on-demand access without considering detailed configurations.

Input data Users can use PathwayCloud using their own data or public data including both gene expression profiles and sample information. The data file should be in a csv-formatted structure. so that the system can process the input to the tools. The labels in sample information should be in binary format terms as 'CASE' or 'CONTROL'. All the genes should be in official gene symbols.

Pathway activity inference tools There are ten pathway activity inference tools in PathwayCloud. All the pathway activity inference tools except `IndividPath` are written either in `R` or `Python`. `IndividPath` was employed in this work using both `shell script` and `awk` based on the strategy described in the original paper (TABLE 3.1).

Comparison of the tools PathwayCloud provides a comparative analysis of the tools using the input data. After calculating pathway activity from each tool, distance preservation described in Chapter 3.4.1.

4.4 Results

The results will be illustrated by using TCGA bladder cancer (BLCA) RNA-seq data set as an example. The data set has 326 number of samples (285 cancer and 41 normal samples) downloaded from Firebrowse (`http://firebrowse.org/`).

4.4.1 Calculating pathway activity values

When TCGA BLCA gene expression and sample information data are provided to PathwayCloud, both files are transferred to selected pathway activity inference tools on the cloud server. Pathway activity values are calculated and then returned back to the web server. PathwayCloud will automatically send an e-mail to the user with a notice and a direct link to the result page. Users can

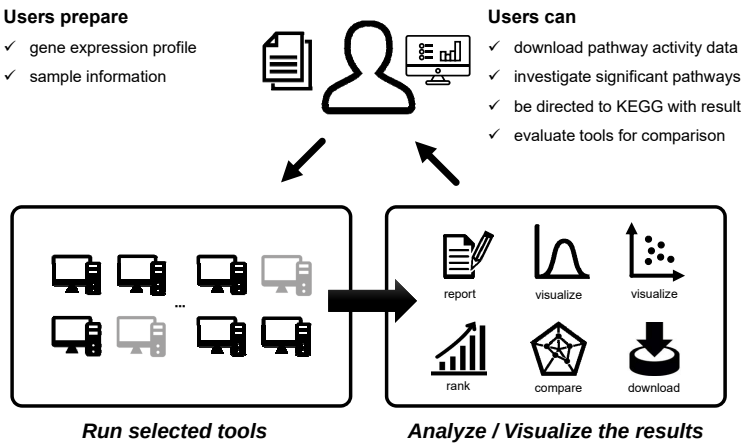

Figure 4.1: An overview of PathwayCloud. When input data (gene expression matrix and sample information) is ready, PathwayCloud runs the selected tools to generate pathway activity values.

also download the pathway activity matrices for further analysis.

4.4.2 Identification of significant pathways

After the calculating step, PathwayCloud uses the pathway activity itself to investigate which pathway is significantly perturbed between the given class information. Student's t-test followed by FDR correction is performed for each pathway across the tools. The resulting significance is depicted to a graph in descending order so that the user can easily recognize which pathway to further investigate Figure 4.2.

4.4.3 Visualization in KEGG pathways

When investigating which pathway is significant from given data, it is also useful to further see which genes contributed more to such significance (Figure 4.3). PathwayCloud provides a direct link to KEGG pathway database for each path-

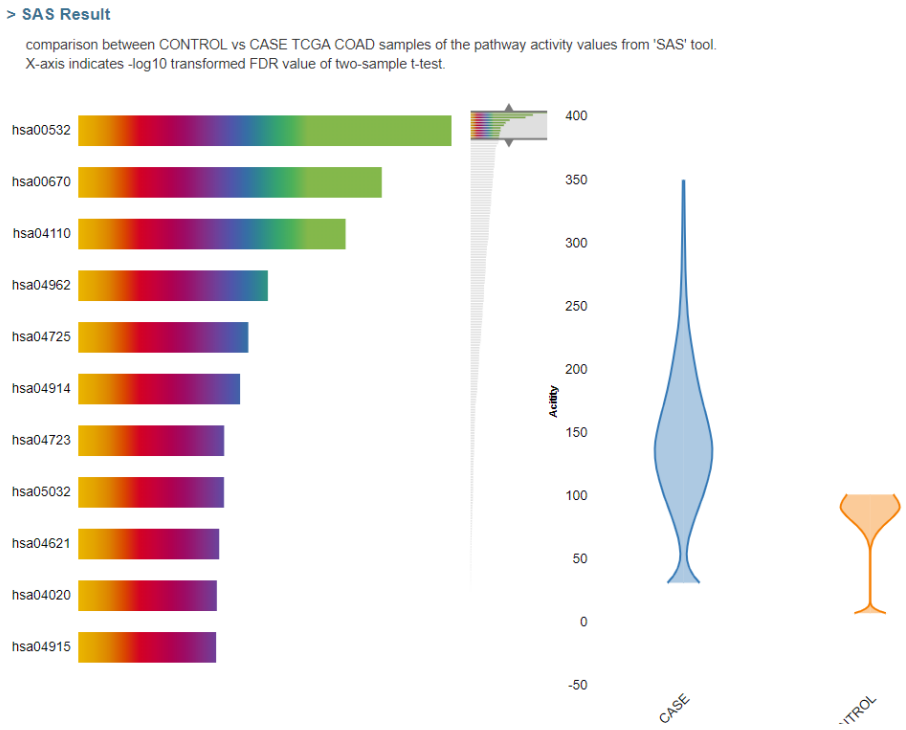

Figure 4.2: Identification of significant pathways from pathway activity. Pathway activity values from `SAS` for each pathway were compared between the sample groups and visualized as a graph. Users can investigate the name of pathways and its significance when placing the cursor on each bar in the graph.

way with individual fold-change as significance. Fold-change value is represented as color-coded format, blue for down-regulated and red for up-regulated.

It is different from the previous method in showing information of all the genes in a pathway. While DAVID shows only the intersection between user-provided genes and pathway genes, PathwayCloud provides the whole landscape of a pathway in a color-coded format to visually inspect how many and where the genes are expressed.

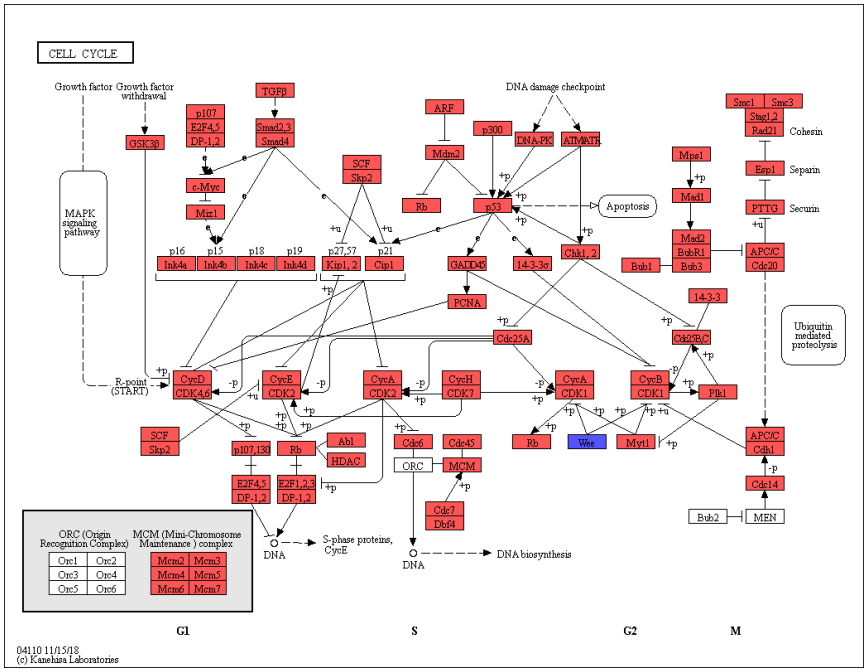

Figure 4.3: Example direction to KEGG pathway - Cell cycle from pathway activity values SAS. Colors are automatically assigned to genes using the input gene expression data.

4.4.4 Comparison of the tools

Since there are eleven pathway activity tools to be selected in PathwayCloud, comparative analysis of the pathway tools are provided (Figure 4.4). Reflecting Chapter 3, one of the criteria to evaluate the tools was to calculate the distance between pathway activity and input gene expression values. This sample-wise metric can be applied regardless of the number of features as it compares the pairwise distance between the samples.

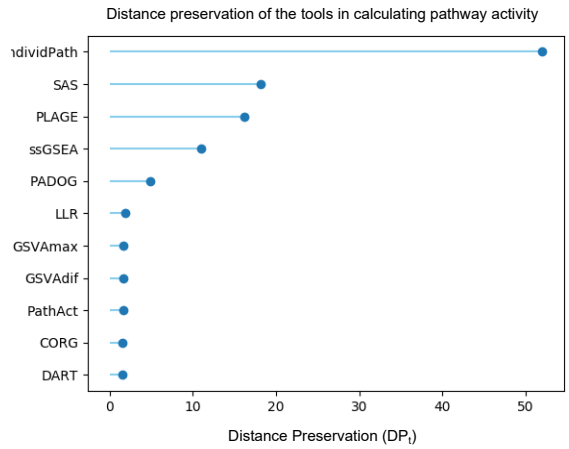

Figure 4.4: Comparison of the tools using Distance Preservation (DP) PathwayCloud performs an evaluation of the tools using Equation 3.1.

4.5 Discussion

PathwayCloud displays the output of pathway activity results using various ways. The result can be seen similar to previously mentioned tools such as DAVID or Enrichr, as both also show the list of significant pathways given data. However, what makes PathwayCloud different is that the analysis resulted from using actual genome-wide expression values unlike other tools mentioned

previously such as DAVID or Enrichr solely using filtered list of genes. This makes PathwayCloud more comprehensive as a whole set of genes in each pathway is considered before investigating the significance in terms of given class information.

One of the features of PathwayCloud is that users can download pathway activity value matrices from the tools that is the very beginning of the result. In addition to list up significant pathways, users can further utilize other machine learning tools out of the pathway activity matrix, considering pathways as features. Revisiting that using pathway activity was very effective in reducing dimension of the given gene expression data (Chapter 2.3), various machine learning models that can aid interpretation of the data are useful such as Random Forests generating variable importance.

Chapter 5

Conclusion

Long time accumulation of biological knowledge lets us learn from the past to make valuable discoveries out of new technologies. This means that addressing biological or technical challenges at that time can provide a different way of interpretation on data. This thesis suggests a pathway-based interpretation of high-throughput molecular data as one of such contributions to be valuable for further researches. There are three challenges addressed in this thesis:

1. Summary of gene expression profiles to a single representative value for each pathway
2. Criteria on comparison of pathway activity inference tools
3. Lack of a platform to run and compare pathway activity tools all together.

In the first study, a method to infer pathway activity, `SAS`, using explicit gene expression data was developed. Many pathways consist of multiple biological functions. To characterize the complex biological mechanisms underlying disease, use of well curated biological pathways was an effective approach.

However, it was challenging how to summarize gene expression values into single pathways. Thus, to perform the pathway-based characterization of disease, subsystems are defined by decomposing biological pathways into multiple functional units by utilizing PIN and defining a scoring scheme. This showed that the landscape of subsystem activity was distinctive among breast cancer subtypes.

The second study provided a comprehensive survey on pathway activity inference tools in terms of input transformation, use of labels, necessity of cohort-level input data, use of gene relations, and scoring metric. Though there are several approaches to represent a pathway for each sample with a single value from gene expression data, there is no comprehensive evaluation on such tools with systematic criteria. Thus, I performed extensive evaluation on the performance of these tools by introducing five criteria. Overall, `SAS` performed best among 13 tools. In addition, `IndividPath` can also be considered a useful tool since it measures relative orderings of the relations within a pathway to detect perturbed pathways in each sample. Both `SAS` and `IndividPath` are favorable than other tools since they performed better than other tools when all five criteria were considered.

In the last study, a web-based system called PathwayCloud was developed to help users understand and utilize pathway activity tools. PathwayCloud takes gene expression profiles and corresponding sample information as input to calculate pathway activity values and evaluate them on a cloud server. The results are shown on a web-based system to provide users easy access and comprehension on pathway activity tools.

In conclusion, my doctoral study addressed three challenges in utilizing biological pathways along with high-throughput gene expression data. This contributed to the field of bioinformatics by making an effort to solve the challenges

and provide convenient way of using pathway activity tools.

Bibliography

Ahn, T., Lee, E., Huh, N., and Park, T. (2014). Personalized identification of altered pathways in cancer using accumulated normal tissue data. *Bioinformatics*, **30**(17), i422–i429.

Ahr, A., Holtrich, U., Solbach, C., Scharl, A., Strebhardt, K., Karn, T., and Kaufmann, M. (2001). Molecular classification of breast cancer patients by gene expression profiling. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, **195**(3), 312–320.

Avraham, R., Haseley, N., Fan, A., Bloom-Ackermann, Z., Livny, J., and Hung, D. T. (2016). A highly multiplexed and sensitive rna-seq protocol for simultaneous analysis of host and pathogen transcriptomes. *Nature protocols*, **11**(8), 1477.

Bader, G. D., Cary, M. P., and Sander, C. (2006). Pathguide: a pathway resource list. *Nucleic acids research*, **34**(suppl_1), D504–D506.

Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F., Schinzel, A. C., Sandy, P., Meylan, E., Scholl, C., *et al.* (2009). Systematic rna interference reveals that oncogenic kras-driven cancers require tbk1. *Nature*, **462**(7269), 108.

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., *et al.* (2010). Ncbi geo: archive for functional genomics data sets—10 years on. *Nucleic acids research*, **39**(suppl_1), D1005–D1010.

Bayerlová, M., Jung, K., Kramer, F., Klemm, F., Bleckmann, A., and Beißbarth, T. (2015). Comparative study on gene set and pathway topology-based enrichment methods. *BMC bioinformatics*, **16**(1), 334.

Bengtsson, M., Ståhlberg, A., Rorsman, P., and Kubista, M. (2005). Gene expression profiling in single cells from the pancreatic islets of langerhans reveals lognormal distribution of mrna levels. *Genome research*, **15**(10), 1388–1392.

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.

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W.-H., Pagès, F., Trajanoski, Z., and Galon, J. (2009). Cluego: a cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, **25**(8), 1091–1093.

Bindea, G., Galon, J., and Mlecnik, B. (2013). Cluepedia cytoscape plugin: pathway insights using integrated experimental and in silico data. *Bioinformatics*, **29**(5), 661–663.

Breiman, L., Friedman, J., Stone, J. C., and Olshen, R. (1983). *Classification and Regression Trees*. Chapman and Hall/CRC.

Cary, M. P., Bader, G. D., and Sander, C. (2005). Pathway information for systems biology. *FEBS letters*, **579**(8), 1815–1820.

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., Holland, T. A., Keseler, I. M., Kothari, A., Kubo, A., *et al.* (2013). The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic acids research*, **42**(D1), D459–D471.

Chang, H. Y., Nuyten, D. S., Sneddon, J. B., Hastie, T., Tibshirani, R., Sørlie, T., Dai, H., He, Y. D., van't Veer, L. J., Bartelink, H., *et al.* (2005). Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(10), 3738–3743.

Chang, J. T., Carvalho, C., Mori, S., Bild, A. H., Gatza, M. L., Wang, Q., Lucas, J. E., Potti, A., Febbo, P. G., West, M., *et al.* (2009). A genomic strategy to elucidate modules of oncogenic pathway signaling networks. *Molecular Cell*, **34**(1), 104–114.

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.

Cheng, W.-Y., Yang, T.-H. O., and Anastassiou, D. (2013). Development of a prognostic model for breast cancer survival in an open challenge environment. *Science Translational Medicine*, **5**(181), 181ra50–181ra50.

Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., Zhang, H., McLellan, M., Yau, C., Kandoth, C., *et al.* (2015). Comprehensive molecular portraits of invasive lobular breast cancer. *Cell*, **163**(2), 506–519.

Creixell, P., Reimand, J., Haider, S., Wu, G., Shibata, T., Vazquez, M., Mustonen, V., Gonzalez-Perez, A., Pearson, J., Sander, C., *et al.* (2015). Pathway and network analysis of cancer genomes. *Nature Methods*, **12**(7), 615–621.

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M. R., *et al.* (2013). The reactome pathway knowledgebase. *Nucleic acids research*, **42**(D1), D472–D477.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., *et al.* (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**(7403), 346–352.

Danielsson, F., James, T., Gomez-Cabrero, D., and Huss, M. (2015). Assessing the consistency of public human tissue rna-seq data sets. *Briefings in bioinformatics*, **16**(6), 941–949.

DeLuca, D. S., Levin, J. Z., Sivachenko, A., Fennell, T., Nazaire, M.-D., Williams, C., Reich, M., Winckler, W., and Getz, G. (2012). Rna-seqc: Rna-seq metrics for quality control and process optimization. *Bioinformatics*, **28**(11), 1530–1532.

Dongen, S. v. (2000). Graph clustering by flow simulation. *Ph.D. Thesis*.

Draghici, S., Khatri, P., Tarca, A. L., Amin, K., Done, A., Voichita, C., Georgescu, C., and Romero, R. (2007). A systems biology approach for pathway level analysis. *Genome research*, **17**(10), 1537–1545.

Drier, Y., Sheffer, M., and Domany, E. (2013). Pathway-based personalized analysis of cancer. *Proceedings of the National Academy of Sciences of the United States of America*, **110**(16), 6388–6393.

EBCTCG *et al.* (2005). Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. *The Lancet*, **365**(9472), 1687–1717.

Efroni, S., Schaefer, C. F., and Buetow, K. H. (2007). Identification of key processes underlying cancer phenotypes using biologic pathway analysis. *PloS one*, **2**(5), e425.

Elkon, R., Vesterman, R., Amit, N., Ulitsky, I., Zohar, I., Weisz, M., Mass, G., Orlev, N., Sternberg, G., Blekhman, R., *et al.* (2008). Spike–a database, visualization and analysis tool of cellular signaling pathways. *BMC bioinformatics*, **9**(1), 110.

Emmert-Streib, F. and Glazko, G. V. (2011). Pathway analysis of expression data: deciphering functional building blocks of complex diseases. *PLoS computational biology*, **7**(5), e1002053.

Ersahin, T., Tuncbag, N., and Cetin-Atalay, R. (2015). The pi3k/akt/mtor interactive pathway. *Molecular Biosystems*, **11**(7), 1946–1954.

Fan, C., Oh, D. S., Wessels, L., Weigelt, B., Nuyten, D. S., Nobel, A. B., van't Veer, L. J., and Perou, C. M. (2006). Concordance among gene-expression–based predictors for breast cancer. *New England Journal of Medicine*, **355**(6), 560–569.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., *et al.* (2013). String v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, **41**(D1), D808–D815.

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.

Ge, H., Walhout, A. J., and Vidal, M. (2003). Integrating 'omic'information: a bridge between genomics and systems biology. *TRENDS in Genetics*, **19**(10), 551–560.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., *et al.* (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, **5**(10), R80.

Goncalves, A., Tikhonov, A., Brazma, A., and Kapushesky, M. (2011). A pipeline for rna-seq data processing and quality assessment. *Bioinformatics*, **27**(6), 867–869.

Guinney, J., Dienstmann, R., Wang, X., De Reyniès, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P., *et al.* (2015). The consensus molecular subtypes of colorectal cancer. *Nature medicine*, **21**(11), 1350–1356.

Haeussler, M., Zweig, A. S., Tyner, C., Speir, M. L., Rosenbloom, K. R., Raney, B. J., Lee, C. M., Lee, B. T., Hinrichs, A. S., Gonzalez, J. N., *et al.* (2018). The ucsc genome browser database: 2019 update. *Nucleic acids research*, **47**(D1), D853–D858.

Han, J., Shi, X., Zhang, Y., Xu, Y., Jiang, Y., Zhang, C., Feng, L., Yang, H., Shang, D., Sun, Z., *et al.* (2015). Esea: discovering the dysregulated pathways based on edge set enrichment analysis. *Scientific reports*, **5**, 13044.

Han, J.-D. J. (2008). Understanding biological functions through molecular networks. *Cell Research*, **18**(2), 224–237.

Hänzelmann, S., Castelo, R., and Guinney, J. (2013). Gsva: gene set variation analysis for microarray and rna-seq data. *BMC bioinformatics*, **14**(1), 7.

Hofree, M., Shen, J. P., Carter, H., Gross, A., and Ideker, T. (2013). Network-based stratification of tumor mutations. *Nature Methods*, **10**(11), 1108–1115.

Hosack, D. A., Dennis Jr, G., Sherman, B. T., Lane, H. C., Lempicki, R. A., *et al.* (2003). Identifying biological themes within lists of genes with ease. *Genome Biology*, **4**(10), R70.

Hrdlickova, R., Toloue, M., and Tian, B. (2017). Rna-seq methods for transcriptome analysis. *Wiley Interdisciplinary Reviews: RNA*, **8**(1), e1364.

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), 96.

Huang, D. W., Sherman, B. T., Tan, Q., Collins, J. R., Alvord, W. G., Roayaei, J., Stephens, R., Baseler, M. W., Lane, H. C., and Lempicki, R. A. (2007). The david gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome biology*, **8**(9), R183.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2008). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, **37**(1), 1–13.

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.

Jaakkola, M. K. and Elo, L. L. (2015). Empirical comparison of structure-based pathway methods. *Briefings in bioinformatics*, **17**(2), 336–345.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabási, A.-L. (2000). The large-scale organization of metabolic networks. *Nature*, **407**(6804), 651–654.

Jia, C., Hu, Y., Kelly, D., Kim, J., Li, M., and Zhang, N. R. (2017). Accounting for technical noise in differential expression analysis of single-cell rna sequencing data. *Nucleic acids research*, **45**(19), 10978–10988.

Jiao, Y., Lawler, K., Patel, G. S., Purushotham, A., Jones, A. F., Grigoriadis, A., Tutt, A., Ng, T., and Teschendorff, A. E. (2011). Dart: Denoising algorithm based on relevance network topology improves molecular pathway activity inference. *BMC bioinformatics*, **12**(1), 403.

Jin, L., Zuo, X.-Y., Su, W.-Y., Zhao, X.-L., Yuan, M.-Q., Han, L.-Z., Zhao, X., Chen, Y.-D., and Rao, S.-Q. (2014). Pathway-based analysis tools for complex diseases: a review. *Genomics, proteomics & bioinformatics*, **12**(5), 210–220.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G., Wu, G., Matthews, L., *et al.* (2005). Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, **33**(suppl 1), D428–D432.

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

Kawashima, E. H., Farinelli, L., and Mayer, P. (2012). Method of nucleic acid amplification. US Patent 8,143,008.

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*, **8**(2), e1002375.

Kim, S., Kon, M., and DeLisi, C. (2012). Pathway-based classification of cancer subtypes. *Biology direct*, **7**(1), 21.

Kohl, M., Wiese, S., and Warscheid, B. (2011). Cytoscape: software for visualization and analysis of biological networks. In *Data mining in proteomics*, pages 291–303. Springer.

Kristensen, V. N., Lingjærde, O. C., Russnes, H. G., Vollan, H. K. M., Frigessi, A., and Børresen-Dale, A.-L. (2014). Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer*, **14**(5), 299–313.

Krogan, N. J., Lippman, S., Agard, D. A., Ashworth, A., and Ideker, T. (2015). The cancer cell map initiative: Defining the hallmark networks of cancer. *Molecular Cell*, **58**(4), 690–698.

Kuhn, M. (2015). A short introduction to the caret package. *R Found Stat Comput*, pages 1–10.

Kuhn, M. *et al.* (2008). Caret package. *Journal of statistical software*, **28**(5), 1–26.

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.

Lee, E., Chuang, H.-Y., Kim, J.-W., Ideker, T., and Lee, D. (2008). Inferring pathway activity toward precise disease classification. *PLoS computational biology*, **4**(11), e1000217.

Leiserson, M. D., Vandin, F., Wu, H.-T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., Papoutsaki, A., Kim, Y., Niu, B., McLellan, M., *et al.* (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature genetics*, **47**(2), 106.

Li, S., Łabaj, P. P., Zumbo, P., Sykacek, P., Shi, W., Shi, L., Phan, J., Wu, P.-Y., Wang, M., Wang, C., *et al.* (2014). Detecting and correcting systematic variation in large-scale rna sequencing data. *Nature biotechnology*, **32**(9), 888–895.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (msigdb) 3.0. *Bioinformatics*, **27**(12), 1739–1740.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. *Cell systems*, **1**(6), 417–425.

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.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz,

R., Walters, G., Garcia, F., Young, N., *et al.* (2013). The genotype-tissue expression (gtex) project. *Nature genetics*, **45**(6), 580.

Ma, X.-J., Wang, Z., Ryan, P. D., Isakoff, S. J., Barmettler, A., Fuller, A., Muir, B., Mohapatra, G., Salunga, R., Tuggle, J. T., *et al.* (2004). A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell*, **5**(6), 607–616.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, **18**(9), 1509–1517.

Mattson, M. P. (2004). Pathways towards and away from alzheimer's disease. *Nature*, **430**(7000), 631.

Mi, H., Muruganujan, A., and Thomas, P. D. (2012). Panther in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic acids research*, **41**(D1), D377–D386.

Mitrea, C., Taghavi, Z., Bokanizad, B., Hanoudi, S., Tagett, R., Donato, M., Voichita, C., and Draghici, S. (2013). Methods and approaches in the topology-based analysis of biological pathways. *Frontiers in physiology*, **4**, 278.

Mogushi, K. and Tanaka, H. (2013). Pathact: a novel method for pathway analysis using gene expression profiles. *Bioinformation*, **9**(8), 394.

Morelli, M. B., Nabissi, M., Amantini, C., Farfariello, V., Ricci-Vitiani, L., di Martino, S., Pallini, R., Larocca, L. M., Caprodossi, S., Santoni, M., *et al.* (2012). The transient receptor potential vanilloid-2 cation channel impairs

glioblastoma stem-like cell proliferation and promotes differentiation. *International Journal of Cancer*, **131**(7), E1067–E1077.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, **5**(7), 621.

Nabissi, M., Morelli, M. B., Santoni, M., and Santoni, G. (2013). Triggering of the trpv2 channel by cannabidiol sensitizes glioblastoma cells to cytotoxic chemotherapeutic agents. *Carcinogenesis*, **34**(1), 48–57.

Network, C. G. A. R. *et al.* (2014). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, **513**(7517), 202–209.

Network, C. G. A. R. *et al.* (2015). The molecular taxonomy of primary prostate cancer. *Cell*, **163**(4), 1011–1025.

NHGRI, N. H. G. R. I. (2015). Biological pathways. https://www.genome.gov/27530687/biological-pathways-fact-sheet. [Online; accessed 29-March-2018].

Ouadid-Ahidouch, H., Dhennin-Duthille, I., Gautier, M., Sevestre, H., and Ahidouch, A. (2013). Trp channels: diagnostic markers and therapeutic targets for breast cancer? *Trends in Molecular Medicine*, **19**(2), 117–124.

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H.-Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., *et al.* (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research*, **33**(17), 5691–5702.

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F. L., Walker, M. G., Watson, D., Park, T., *et al.* (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine*, **351**(27), 2817–2826.

Pang, H., Lin, A., Holford, M., Enerson, B. E., Lu, B., Lawton, M. P., Floyd, E., and Zhao, H. (2006). Pathway analysis using random forests classification and regression. *Bioinformatics*, **22**(16), 2028–2036.

Park, E., Williams, B., Wold, B. J., and Mortazavi, A. (2012). Rna editing in the human encode rna-seq data. *Genome research*, **22**(9), 1626–1633.

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), 1160–1167.

Paz, A., Brownstein, Z., Ber, Y., Bialik, S., David, E., Sagir, D., Ulitsky, I., Elkon, R., Kimchi, A., Avraham, K. B., *et al.* (2010). Spike: a database of highly curated human signaling pathways. *Nucleic acids research*, **39**(suppl_1), D793–D799.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

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.

Pico, A. R., Kelder, T., Van Iersel, M. P., Hanspers, K., Conklin, B. R., and Evelo, C. (2008). Wikipathways: pathway editing for the people. *PLoS biology*, **6**(7), e184.

Prat, A. and Perou, C. M. (2011). Deconstructing the molecular portraits of breast cancer. *Molecular Oncology*, **5**(1), 5–23.

Prevarskaya, N., Zhang, L., and Barritt, G. (2007). Trp channels in cancer. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, **1772**(8), 937–946.

Rahmatallah, Y., Emmert-Streib, F., and Glazko, G. (2015). Gene set analysis approaches for rna-seq data: performance evaluation and application guideline. *Briefings in bioinformatics*, **17**(3), 393–407.

Rahnenführer, J., Domingues, F. S., Maydt, J., and Lengauer, T. (2004). Calculating the statistical significance of changes in pathway activity from gene expression data. *Statistical applications in genetics and molecular biology*, **3**(1), 1–29.

Ramanan, V. K., Shen, L., Moore, J. H., and Saykin, A. J. (2012). Pathway analysis of genomic data: concepts, methods, and prospects for future development. *TRENDS in Genetics*, **28**(7), 323–332.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *science*, **297**(5586), 1551–1555.

Reis-Filho, J. S. and Pusztai, L. (2011). Gene expression profiling in breast cancer: classification, prognostication, and prediction. *The Lancet*, **378**(9805), 1812–1823.

Reis-Filho, J. S., Weigelt, B., Fumagalli, D., and Sotiriou, C. (2010). Molecular profiling: moving away from tumor philately. *Science Translational Medicine*, **2**(47), 47ps43–47ps43.

Reynard, L. N. and Loughlin, J. (2013). Insights from human genetic studies into the pathways involved in osteoarthritis. *Nature Reviews Rheumatology*, **9**(10), 573.

Romero, P., Wagg, J., Green, M. L., Kaiser, D., Krummenacker, M., and Karp, P. D. (2004). Computational prediction of human metabolic pathways from the complete human genome. *Genome biology*, **6**(1), R2.

Ross, J. B., Huh, D., Noble, L. B., and Tavazoie, S. F. (2015). Identification of molecular determinants of primary and metastatic tumour re-initiation in breast cancer. *Nature Cell Biology*, **17**(5), 651–664.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). Dna sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, **74**(12), 5463–5467.

Sayers, E. W., Agarwala, R., Bolton, E. E., Brister, J. R., Canese, K., Clark, K., Connor, R., Fiorini, N., Funk, K., Hefferon, T., *et al.* (2019). Database resources of the national center for biotechnology information. *Nucleic acids research*, **47**(Database issue), D23.

Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., GuhaThakurta, D., Sieberts, S. K., Monks, S., Reitman, M., Zhang, C., *et al.* (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nature genetics*, **37**(7), 710.

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and

Buetow, K. H. (2008). Pid: the pathway interaction database. *Nucleic acids research*, **37**(suppl_1), D674–D679.

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K. H. (2009). Pid: the pathway interaction database. *Nucleic Acids Research*, **37**(suppl 1), D674–D679.

Shapovalov, G., Lehen'kyi, V., Skryma, R., and Prevarskaya, N. (2011). Trp channels in cell survival and cell death in normal and transformed cells. *Cell Calcium*, **50**(3), 295–302.

Slonim, D. K. (2002). From patterns to pathways: gene expression data analysis comes of age. *Nature genetics*, **32**(4s), 502.

Song, X. and Zhou, X.-H. (2008). A semiparametric approach for the covariate specific roc curve with survival outcome. *Statistica Sinica*, **18**(947-965), 84.

Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M., Van de Rijn, M., Jeffrey, S., *et al.* (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(19), 10869–74.

Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., *et al.* (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(14), 8418–8423.

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

Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., *et al.* (2006). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute*, **98**(4), 262–272.

Su, J., Yoon, B.-J., and Dougherty, E. R. (2009). Accurate and reliable cancer classification based on probabilistic inference of pathway activity. *PloS one*, **4**(12), e8161.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., *et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, **102**(43), 15545–15550.

Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., *et al.* (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**(5891), 956–960.

Tarca, A. L., Draghici, S., Bhatti, G., and Romero, R. (2012). Down-weighting overlapping genes improves gene set analysis. *BMC bioinformatics*, **13**(1), 136.

Tarca, A. L., Bhatti, G., and Romero, R. (2013). A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PloS one*, **8**(11), e79217.

TAUB, FLOYD, E., DeLEO, J. M., and Thompson, E. B. (1983). Sequential

comparative hybridizations analyzed by computerized image processing can identify and quantitate regulated rnas. *Dna*, **2**(4), 309–327.

TCGA *et al.* (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, **490**(7418), 61–70.

Therneau, T. M., Atkinson, B., and Ripley, B. (2010). rpart: Recursive partitioning. r package version 3.1-46. *Computer software program retrieved from http://CRAN. R-project. org/package= rpart*.

Tomfohr, J., Lu, J., and Kepler, T. B. (2005). Pathway level analysis of gene expression using singular value decomposition. *BMC bioinformatics*, **6**(1), 225.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, **28**(5), 511.

Van De Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., *et al.* (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, **347**(25), 1999–2009.

Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., *et al.* (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**(6871), 530–536.

Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., Haussler, D., and Stuart, J. M. (2010). Inference of patient-specific pathway activities

from multi-dimensional cancer genomics data using paradigm. *Bioinformatics*, **26**(12), i237–i245.

Vitali, F., Li, Q., Schissler, A. G., Berghout, J., Kenost, C., and Lussier, Y. A. (2017). Developing a 'personalome' for precision medicine: emerging methods that compute interpretable effect sizes from single-subject transcriptomes. *Briefings in Bioinformatics*.

Vogelstein, B. and Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nature medicine*, **10**(8), 789.

Wagner, G. P., Pavlicev, M., and Cheverud, J. M. (2007). The road to modularity. *Nature Reviews Genetics*, **8**(12), 921.

Wang, C., Gong, B., Bushel, P. R., Thierry-Mieg, J., Thierry-Mieg, D., Xu, J., Fang, H., Hong, H., Shen, J., Su, Z., *et al.* (2014). The concordance between rna-seq and microarray data depends on chemical treatment and transcript abundance. *Nature Biotechnology*, **32**(9), 926–932.

Wang, H., Cai, H., Ao, L., Yan, H., Zhao, W., Qi, L., Gu, Y., and Guo, Z. (2015). Individualized identification of disease-associated pathways with disrupted coordination of gene expression. *Briefings in bioinformatics*, **17**(1), 78–87.

Wang, J., Duncan, D., Shi, Z., and Zhang, B. (2013). Web-based gene set analysis toolkit (webgestalt): update 2013. *Nucleic acids research*, **41**(W1), W77–W83.

Wang, J., Vasaikar, S., Shi, Z., Greer, M., and Zhang, B. (2017). Webgestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic acids research*, **45**(W1), W130–W137.

Wang, L., Wang, S., and Li, W. (2012). Rseqc: quality control of rna-seq experiments. *Bioinformatics*, **28**(16), 2184–2185.

Wang, X., Dalkic, E., Wu, M., and Chan, C. (2008). Gene module level analysis: identification to networks and dynamics. *Current Opinion in Biotechnology*, **19**(5), 482–491.

Wang, Z., Gerstein, M., and Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, **10**(1), 57.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., Network, C. G. A. R., *et al.* (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, **45**(10), 1113.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J. A., Marks, J. R., and Nevins, J. R. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(20), 11462–11467.

Wu, G. and Stein, L. (2012). A network module-based method for identifying cancer prognostic signatures. *Genome Biology*, **13**(12), R112.

Xie, C., Mao, X., Huang, J., Ding, Y., Wu, J., Dong, S., Kong, L., Gao, G., Li, C.-Y., and Wei, L. (2011). Kobas 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic acids research*, **39**(suppl_2), W316–W322.

Yamada, T., Ueda, T., Shibata, Y., Ikegami, Y., Saito, M., Ishida, Y., Ugawa, S., Kohri, K., and Shimada, S. (2010). Trpv2 activation induces apoptotic

cell death in human t24 bladder cancer cells: a potential therapeutic target for bladder cancer. *Urology*, **76**(2), 509–e1.

Yang, Y., Han, L., Yuan, Y., Li, J., Hei, N., and Liang, H. (2014). Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nature communications*, **5**, 3231.

Yarden, Y. and Sliwkowski, M. X. (2001). Untangling the erbb signalling network. *Nature reviews Molecular cell biology*, **2**(2), 127.

Yook, S.-H., Oltvai, Z. N., and Barabási, A.-L. (2004). Functional and topological characterization of protein interaction networks. *Proteomics*, **4**(4), 928–942.

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., and Liu, X. (2014). Comparison of rna-seq and microarray in transcriptome profiling of activated t cells. *PloS one*, **9**(1), e78644.

Zwiener, I., Frisch, B., and Binder, H. (2014). Transforming rna-seq data to improve the performance of prognostic gene signatures. *PloS one*, **9**(1), e85150.

초록

RNA-seq 데이터를 사용하여 RNA 전사체의 변화량을 측정하는 것은 생물정보학 분야에서 필수적으로 수행하고 있는 분석 방법 중 하나이다. 그러나 RNA-seq은 인간의 2만개 이상의 유전자를 포함하는 고차원의 전사체 데이터를 생성하기 때문에, 상대적으로 적은 양의 샘플들을 분석하고자 할때는 데이터 해석에 있어서 어려움이 있다. 따라서, 더 나은 생물학적 이해를 위해서는 생물학적 패스웨이와 같이 잘 요약되고 널리 사용되는 정보를 사용하는 것이 유용하다. 그러나 전사체 데이터를 생물학적 패스웨이로 요약하는 것은 몇 가지 이유로 매우 어려운 작업이다. 첫째, 전사체 데이터를 패스웨이 차원으로 변환할 때 엄청난 정보 손실이 발생한다. 예를 들어, 인간에 존재하는 전체 유전자의 1/3만이 KEGG 패스웨이 데이터베이스에서 보고되고 있다. 둘째, 각 패스웨이는 많은 유전자로 구성되어 있으므로 패스웨이의 활성도를 측정하려면 구성하고 있는 유전자 간의 관계를 고려하면서 유전자 발현 값을 단일 값으로 요약해야 한다.

본 박사 학위 논문은 패스웨이 활성도 측정을 위한 새로운 방법을 개발하고 여러 비교 기준에 따라 기존에 보고된 패스웨이 활성도 도구들에 대한 광범위한 평가 실험을 수행하고자 한다. 또한 일반 사용자가 자신의 데이터를 쉽게 분석할 수 있도록 앞서 언급한 도구들을 웹 기반 시스템 구축을 통해 쉽게 사용할 수 있도록 하였다.

첫 번째 연구에서는 전사체 유전자 발현양 정보를 그대로 사용하고, 상호작용 네트워크 측면에서 유전자 간의 관계를 고려하여 패스웨이의 관점으로 전사체 데이터를 요약하는 새로운 방법을 개발하였다. 이 연구에서는 단백질 상호 작용 네트워크, 패스웨이 데이터베이스 및 RNA-seq 전사체 데이터를 활용하여 생물학적 패스웨이를 여러 개의 시스템으로 구분하는 새로운 개념을 제안하고자 한다. 각 시스템 및 각 샘플마다의 활성화 정도를 측정하기 위해 SAS (Subsystem Activation

Score)를 개발하였다. 이 방법은 샘플 들간 및 유방암 아형들 사이에서 차별적으로 활성화되는 특유의 유전체 상에서의 활성화 패턴 또는 서브 시스템을 표현할 수 있었다. 그런 다음, 분류 및 회귀 트리 (CART) 분석을 수행하여 예후 모델링을 위해 SAS 정보를 사용했습니다. 그 결과, 10 개의 가장 중요한 하위 시스템으로 정의 된 11 개의 환자 하위 그룹은 생존 결과에 있어 최대 불일치로 확인되었다. 이 모델은 유사한 생존 결과를 가진 환자 하위 그룹을 정의했을뿐만 아니라 기능적으로 유익한 유방암 유전자 세트를 제안하는 하위 시스템의 활성화 상태에 따라 결정되는 샘플 특이적인 상태의 판단 경로를 제공한다.

두 번째 연구는 전 암 (pan-cancer) 데이터 세트를 사용하여 다섯 가지 비교 기준에 따라 13 가지의 패스웨이 활성도 측정 도구를 체계적으로 비교 및 평가하는 연구이다.현존하는 패스웨이 활성도 측정 도구가 많이 있지만, 이러한 도구가 코호트 수준에서 유용한 정보를 제공하는지에 대한 비교 연구는 없다. 이 연구는 크게 두 가지 부분에 대해서 의미가 있다. 첫째, 이 연구는 기존의 패스웨이 활성도 측정 도구에서 사용되는 계산 기법에 대한 포괄적인 정보를 제공한다. 패스웨이 활성도 측정은 다양한 접근법을 사용하고, 입력 데이터의 변환, 샘플 정보의 사용, 코호트 수준의 인풋 데이터의 필요성, 유전자 관계 및 점수체계의 사용 등에서 다양한 요구 사항을 가정해야 한다. 둘째, 이러한 도구의 성능에 대한 다섯 가지 비교 기준을 사용하여 광범위한 평가가 수행되었다. 도구가 원래의 유전자 발현 프로파일의 특성을 얼마나 잘 유지하는지를 측정하는 것부터, 유전자 발현 데이터에 노이즈를 임의로 도입하였을 때 얼마나 둔감한지 등을 조사했다. 임상 적용을 위한 도구의 유용성을 평가하기 위해 세가지 변수 (종양 대 정상, 생존 및 암의 아형)에 대한 분류 작업을 수행했다.

세 번째 연구는 사용자가 전사체 데이터를 제공하고, 앞선 연구에서 비교한 활성도 측정 도구를 사용하여 패스웨이 활성도를 측정하는 클라우드 기반 시스템 (PathwayCloud)을 구축하는 것이다. 사용자가 데이터를 시스템에 업로드하고 실행할 분석 도구를 선택하면, 이 시스템은 각 도구에 대한 패스웨이 활성도 값과 선택한 도구에 대한 성능 비교 요약을 자동으로 수행한다. 사용자는 또한 주어진

샘플 정보의 측면에서 어떤 패스웨이가 중요한지 조사 할 수 있으며, KEGG rest API를 통해서 직접 패스웨이의 어떤 유전자의 변화가 유의미한지를 시각적으로 분석할 수 있다.

결론적으로, 본 학위 논문은 고용량의 유전자 발현 데이터를 사용하여 생물학적 패스웨이에 대한 분석 방법을 개발하고, 다른 유형의 도구를 포괄적인 기준으로 비교하고, 사용자가 이 도구들에 쉽게 접근할 수 있는 웹 기반 시스템을 제공하는 것을 목표로 한다. 이 전반적인 접근 방식은 생물학적 패스웨이 측면에서 유전자 발현 데이터를 이해하는 데 중요했다.

주요어: 패스웨이, 패스웨이 활성도, 생물학적 네트워크 분석, 유전자 발현, RNA 시퀀싱

학번: 2014-30099