Optimisation Models for Pathway Activity Inference in Complex Diseases

Chen, Yongnan

Awarding institution:
King's College London

Optimisation Models for Pathway Activity Inference in Complex Diseases

Yongnan Chen

Department of Informatics, Faculty of Natural, Mathematical and Engineering Sciences
King's College London

This dissertation is submitted for the degree of
Doctor of Philosophy

December 2022

Acknowledgements

I would like to thank my supervisor, Dr. Sophia Tsoka, and my second supervisor, Dr. Konstantinos Theofilatos, for their invaluable support, guidance and trust throughout my Ph.D. Thanks to Professor Lazaros Papageorgiou from University College London and Dr. Songsong Liu from Harbin Institute of Technology for their help and advice in implementing the methodology.

Many thanks to the members of our group who provided invaluable advice in work. Thanks to Roman Laddach and Pedro Henrique Da Costa Avelar for providing ideas and suggestions for some of the chapters. Thanks to Dr. Elmira Amiri Souri for providing much help in the first year of my Ph.D. Thanks to Yutong Li for the many enjoyable hours we spent together during the final year. I am also very grateful to the other Ph.D. students in the Bush House office, and it has been a pleasure to share my time with you.

I am also very grateful to my friends for their support and companionship over the past four years, without whom life would have been much harder. I am grateful to Jiarui Jin for helping me proofread this thesis and to my roommate Yuxin Mo for being so helpful in life. Thanks to Yu Sang, Wuyue Guan and Xin Zhang, three of my dearest friends who have always provided me with sincere support and advice. Special thanks to my boyfriend, Sizhan Chen, for his help with a few dirty work in data processing and his unwavering companionship.

Finally, I would like to thank my mother for her selfless love and support.

Abstract

With advances in high-throughput technologies, there has been enormous increase in data related to profiling the activity of molecules in disease. While such data provide more comprehensive analysis of cellular actions, their large volume and complexity pose difficulty in accurate disease phenotype classification. Therefore, novel modelling methods that can not only improve accuracy but also offer interpretable means of analysis, are important. In this respect, biological pathways (i.e. gene sets that reflect related functional cascades) can be used to incorporate a-priori knowledge of biological interactions, so as to decrease the data dimensionality from gene-level to pathway-level and increase biological interpretability of related methodologies. Methods to infer pathway activity values across high-throughput data have shown good potential towards better understanding of the regulation pattern of gene expression values. This thesis focuses on the application and development of mathematical programming models for pathway activity inference in disease classification and gene signature identification in gene profiling data.

First, an optimisation model, known as DIGS, for pathway activity inference toward precise disease phenotype prediction is implemented on Microarray datasets of ischemic stroke and RNA-Seq datasets of colorectal cancer. DIGS is a mixed integer linear programming (MILP) mathematical optimisation model aiming at separating the different cancer subtypes to the largest extent. In supervised manner, DIGS defines pathway activity as the linear combination of the member gene expression values multiplying the inferred gene weights. Inside the DIGS model, gene weights are optimised to maximise the discriminative power of the inferred pathway activity and the optimisation objective is set to minimise the number of incorrect sample allocation. Comparative analysis shows that DIGS model outperforms other up-to-date methods in three pathway activity evaluation metrics, classification accuracy, robustness against noisy data and survival outcome prediction accuracy of patients.

Next, the model is improved to form a more efficient MILP model (DIGS2). This model avoids a large number of binary decision variables in the original model and is thus easier to be solved to global optimality. The assessment of DIGS2 model on two

RNA-Seq datasets shows improvements on solution qualities and better performance on the evaluation metrics compared with other pathway activity inference methods.

These models exhibit outstanding contribution on identifying disease relevant pathways and genes, which are also verified on relevant findings in the literature. Finally, the effectiveness of the proposed MILP models is explored in the noisier and sparser scRNA-Seq data. In addition to the classification effect, following the up-to-date research interests for this type of data, the clustering ability of pathway activity value is emphasised in this work to see whether the pathway activity values can clustering the cells of same label through dimension reduction methods. A comparison made with methods from literature shows that the proposed method achieves competitive results for separating the cells.

Overall, this thesis demonstrates that the flexible nature of mathematical programming lends itself well to developing solution procedures for pathway activity inference. The evaluation metrics show the proposed methods to outperform other methods from literature, as well as to provide explainable means of modelling. Also, the proposed methods show the potential to reveal meaningful biological interpretations for complex diseases such as cancer.

Publications

Chen, Y., Theofilatos, K., Papageorgiou, L. G., & Tsoka, S. (2020, May). *Identification of Important Biological Pathways for Ischemic Stroke Prediction through a Mathematical Programming Optimisation Model-DIGS.* In Proceedings of the 2020 12th International Conference on Bioinformatics and Biomedical Technology (pp. 25-31). Best paper award.

Chen, Y., Liu, S., Papageorgiou, L., Theofilatos, K., & Tsoka, S. (Submitted). *Optimisation models for Pathway Activity Inference in Cancer.*

Chen, Y., Laddach, R.,Karagiannis. S, Papageorgiou, L. G., & Tsoka, S. (In preparation). *Optimisation-based Pathway Activity Inference Method for Single-cell RNA Sequence Data.*

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Pathway activity inference, a solution for Systems Biology

Systems biology is an integrated discipline that links the molecular analysis of components at a single biological scale or different scales (e.g. cells, tissues and organ systems) to physiological functions and organism phenotypes through quantitative reasoning, computational modelling and high-throughput experimental techniques (transcriptome profiling via microarray or sequencing) [1]. In brief, Systems Biology employs computational and mathematical analyses to decode complex underlying mechanisms in biological systems. In recent decades, the rapid development of high-throughput technologies led to acquisition of large omics datasets of the basic entities forming these complex systems. Moreover, these advancements through these technologies facilitated the development of Systems Biology and precision medicine for complex diseases.

In detail, omic data (including genomics, transcriptomics, proteomics and metagenomics profiling) reflect the molecules expression levels in whole-tissue or across single cells. For example, the bulk RNA Sequencing technologies enable the transcriptome quantification of tissues. Genomic variants may occur, including single nucleotide variations, small insertions, deletions and structural variations. These variations may influence the susceptibility and pathogenesis in diseases [2, 3]. Therefore, the importance of analyzing omics data lies in capturing the effects of these variants.

A generalised workflow for analysis of omic datasets [4] is indicated in Figure 1.1. First, the omic data is generated by bulk or single cell profiling via sequencing technologies. Then, integrative analysis of omic data as well as non-omics data can be conducted to identify novel biomarkers, pathways, to cluster samples or cells into groups of similar properties, to predict sample states (e.g. health or disease), to unravel the phenotype associations or to reveal underlying mechanisms [4].

The objectives of analysing the omic datasets are not identical but related to each other. Biomarker discovery aims to the identification of significant genes that may express differently between patient groups or between health and disease. Also, identification of genes that express differently between different complex disease phenotypes (e.g. across molecular subtypes in cancer) is important. The identification of gene biomarkers can be extended to the identification of pathway biomarkers, as genes that belong to the same pathway work together for a cellular process and pathway products involve molecules that directly affect the biological processes inside the cell. Therefore, recent methods for identification of pathway biomarkers are popular. Another popular objective for analysing the omic data is sample outcome prediction. One of the key purposes of omic data analysing is to improve precision medicine and personalised treatment. Therefore, accurate disease outcome and phenotype prediction occupies an important place when designing the computational methods for analysing the omic data.

In integrative analyses, key aspects refer to the complexity of the data. Firstly, omic data has the "large-p small-n" nature (p stands for the number of predictors,n for the number of samples in a dataset) [5]. The large number of features for each sample necessitates finding important features through appropriate feature selection and dimensionality reduction methods with statistical analysis. Secondly, the technical noise of the omics data render the selected features facing the challenge of uncertainty. Therefore, the challenge presented here is to design a computational method that can withstand noise and enable valuable insights from the large volume of omics datasets. Various machine learning algorithms have been applied for identifying biomarkers and disease classification [6]. Moreover, research directions that combine pathways with the analysis of gene expression data have demonstrated usefulness in improving robustness of the analysis, as well as being biologically interpretable. Compared to genetic biomarkers, pathway biomarkers are less unstable across patient populations [7]. For further related background please refer to Chapter 2.

Therefore, methods that can reduce the gene-level data to pathway-level data are popular. The resulting pathway-level data are termed as pathway activity. Pathway activity is defined as a value that can represent the collective activity (e.g. gene expression level) of all pathway member genes. The function of pathway activity is to aggregate the expression levels of constituent entities in a particular pathway for a sample. Using pathway activity values to replace gene-level values is an effective way to reduce the high dimensionality while introducing more biological interpretability.

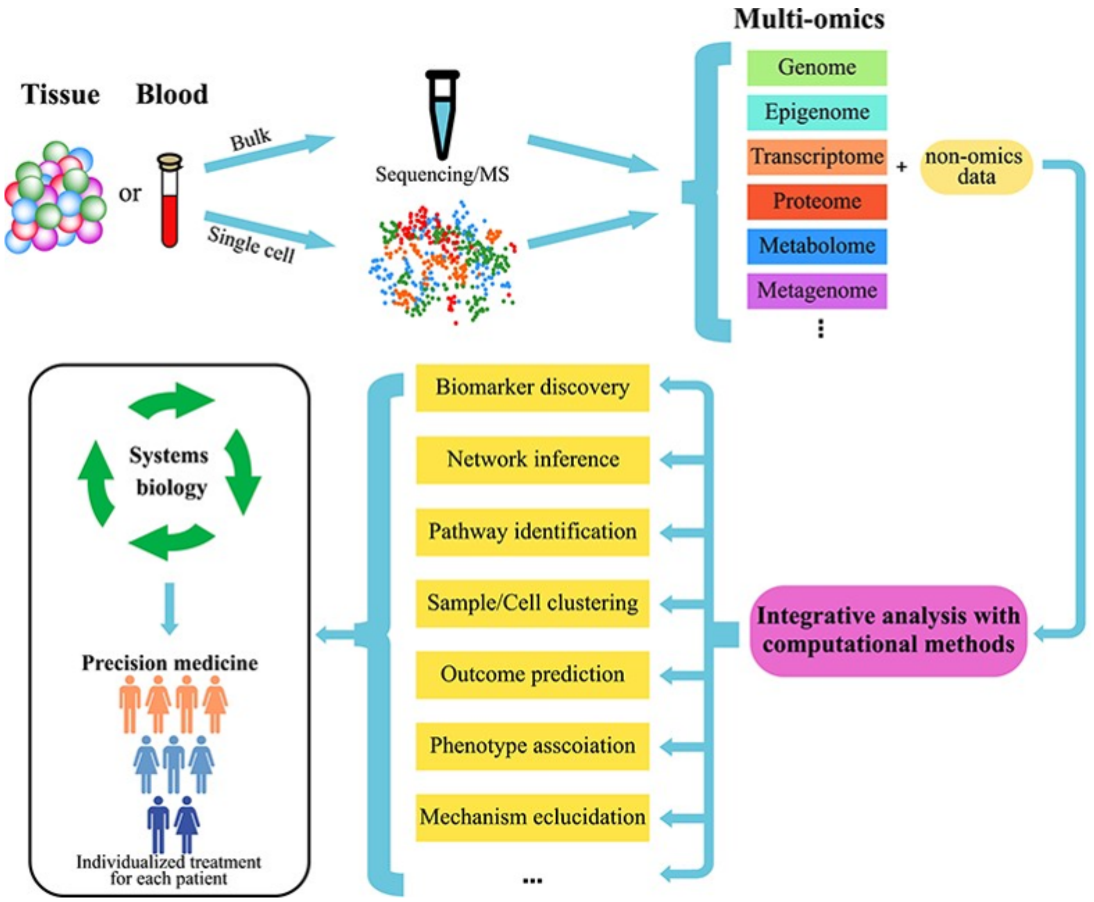


Fig. 1.1 General workflow of multi-omics studies. First, the multi-omics data can be generated at the bulk or single-cell level using sequencing-and/or MS-based technologies. Then, integrative analysis of the multi-omics data (e.g. genomics, epigenomics, transcriptomics, proteomics and metagenomics) and non-omics (i.e. the clinical information of patients) data can be conducted to identify novel biomarkers, networks, pathways, or cluster samples/cells, or predict health/disease states, or unravel the phenotype associations and underlying mechanisms. Finally, one crucial application of multi-omics studies is to realise precision medicine with systems biology approaches. Figure taken from [4]

.

Basic statistical methods (e.g. mean, median) and advanced computational methods (e.g. regression model, correlation score and principle components) have been designed for pathway activity inference. The improvements offered by the pathway activity inference methods are (i) the reduced dimension in pathway activity values so as to better describe the dataset and (ii) suitability to achieve the objectives listed in the omics dataset workflow (Figure 1.1). Specifically, the inferred pathway activity values are expected to have better ability to discriminate among samples or cells of different groups, and improve identification of significant pathways, sample or cell clustering and outcome prediction.

1.2 Mathematical programming optimisation approach

Mathematical programming, or mathematical optimisation, is applied throughout this thesis to conduct the pathway activity inference task. Mathematical programming is an optimisation approach that models a problem as mathematical equations and finds the best answer matching to the maximised or minimised the objective function value [8].

A general representation for mathematical programming model is:

$$min \quad f(x)$$

$$s.t. \quad g(x) \leq 0$$

$$h(x) = 0$$

$$x \in X$$

where $x \in X$ are the decision variables, $f(x)$ is the objective function, $g(x)$ and $h(x)$ are inequality and equality constraints. $s.t.$ refers to as "subject to".The objective function is the quantity I wish to maximise or minimise. Constraint uses mathematical way to restrict the possible values that the variables can take. To build a model for a specific problem, I need to identify the decision variables and formulate the objective function and constraints.

Typical mathematical programming models include linear programming models (LP), non-linear programming models (NLP), mixed integer linear programming models (MILP) and mixed integer non-linear programming models (MINLP).

- LP model consists of linear objective function and constraints, where all decision variables are continuous;

- NLP model is non-linearity in objective function and/or constraints, where all decision variables are continuous;
- MILP model consists of linear objective function and constraints, while some of the decision variables are restricted to take discrete values, i.e. 0, 1;
- MINLP model is non-linearity in objective function and/or constraints, while some of the decision variables are restricted to take discrete values, i.e. 0, 1.

A **feasible solution** for a mathematical optimisation problem is a set of values for the decision variables that satisfies all of the constraints. The set of all feasible solutions defines the feasible region of the problem. Most optimization algorithms operate by first trying to locate any feasible solution, and then attempting to find another (better) feasible solution that improves the value of the objective function. This process of trying to find improving feasible solutions repeats until either no further improvement is possible or some other stopping criteria is met. Then the **optimal solution** is one where there is no other feasible solution with a better objective function value [8].

Overall, mathematical programming is a modelling framework with the ability to meet the objectives of pathway activity inference problem. This thesis focuses on the application and development of mathematical programming optimisation pathway activity inference models. The mathematical programming objective function can directly optimise the categorisation of samples, and the flexibility of the model lends itself well to the development of versatile constraint formulas.

Different gene profiling technologies for transcriptional activity and various cancer datasets are used. Each stage of method development is discussed and illustrates the evolution and adaptation of the implementation pipeline and the optimisation models. The evaluation procedures undertaken show the methods to be comparable to other approaches in the literature and are able to yield meaningful biological insights for these complex diseases. The optimisation models described in this thesis have the potential to provide new insights into the disease-related significant biological pathways and help to enrich pathology knowledge of target disease via systems biology approaches.

1.3 Research Aims

The overall aim of this thesis is to develop mathematical programming methods to address pathway activity inference for different gene profiling technologies. The inferred pathway activity values are expected to obtain higher ability to separate samples of different classes compared with using individual genes, with a view to aiding the extraction of biologically meaningful results for complex diseases. This aim can be further decomposed into five core research goals:

- To build on existing pathway activity inference method to evaluate its ability on analysing bulk gene expression profiling data.
- To build and implement a new pathway activity inference method that has higher solution quality and higher prediction accuracy.
- To extend the application of pathway activity inference method from bulk data to single cell data.
- To evaluate the methodology to show comparability with existing methods from literature.
- To demonstrate the potential of such methods to find meaningful results in biological applications.

1.4 Thesis Outline

The thesis unfolds as follows. Chapter 2 provides background information on the topic of transcriptome profiling technologies, pathway activity inference and its significance in biological analyses. Related work is discussed followed by a review of existing pathway activity inference methods. In particular, a mixed integer linear programming (MILP) model is reported. This model represents the starting point for the methodologies derived in this thesis.

Chapter 3 works as a transitional chapter, which applies the MILP model given in Chapter 2 on microarray datasets of Acute Ischemic Stroke. In this chapter, the MILP model is incorporated with an integrated dataset that is a combination of three gene expression profiles. A robust validation pipeline for implementing the pathway activity inference methods is built in this chapter and will be used throughout the thesis. The pathway activity inferred by the MILP model aims to separate the samples belonging to binary labels, stroke and healthy.

In Chapter 4, the MILP method is applied on a more advanced gene expression profiling technology, RNA Sequencing (RNA-Seq). The pathway activity inference is conducted on a colorectal cancer RNA-Seq dataset for both binary classification problem (tumor and normal) and multi-class classification problem (four cancer molecular subtypes). In addition to the evaluation on sample separation, two more metrics are included in the pathway activity evaluation process, which are testing the robustness of the pathway inference methods on noisy datasets and using pathway activity values for survival analysis. Comparative studies are carried out in these two chapters to evaluate the method's performance across the baseline methods.

Chapter 5 proposes a new MILP model, which is modified based on the model given in chapter 2, to reduce the computation complexity and achieve better performance. The former model is found to suffer on obtaining global optimal solutions because of the large amount of binary decision variables. Therefore, in the design of the new model, the number of binary variables is reduced to allow better solution quality and better performance on the pathway activity evaluation metrics. This chapter uses two RNA-Seq datasets, colorectal cancer and breast cancer. The validation pipeline and evaluation metrics are similar to the ones reported in Chapter 4. Besides, this chapter explores the outputs of the MILP model in in-depth analysis across individual pathways.

Chapter 6 discusses the application of the new proposed model on single cell RNA sequencing data (scRNA-Seq). Challenges posed by scRNA-Seq technology are the sparsity of the datasets and the heterogeneity among individual cell, which cause uncertainty to the quality of the inferred pathway activity values. Therefore, this chapter first tests the performance of using pathway activity values to separating cell types. After the method is proven to be promising, the pathway activity values are used to separate the tumor subtypes of breast cancer. In this chapter, the evaluation of the pathway activity values focuses on cell clustering and low-dimensional visualisation. Also, this chapter discusses the application of pathway activity for integrating different scRNA-Seq datasets.

Finally, chapter 7 concludes the thesis by first giving a brief overview of each chapter. Next, research aims outlined above are revisited in order to indicate where they are addressed in the thesis and to ascertain to what degree they are fulfilled. Major contributions of the thesis are then given, followed by discussions regarding the limitations of the work and the potential avenues of future research.

Chapter 2

Background and Related work

The previous chapter outlined the rationale behind the work that will be presented in this thesis and set out several key goals. The central aim of this thesis is to develop mathematical programming methods to tackle the pathway activity inference problem with a view to aiding gene signature identification and extracting biologically meaningful results from high-throughput profiling data. In order to carry out this task, a thorough understanding is required of the problem, the potential applications of existing methodologies as well as the possible limitations.

In this chapter, a review is given of the essential background and related work that underpins this thesis. This includes introducing three kinds of technologies for profiling data with their properties, and describing the biological pathways that can be used to incorporate the a-priori knowledge of biological interactions with the profiling datasets for functional analysis. From this biological background, the pathway activity inference framework is then taken forward and discussed in detail followed by the introduction to some baseline and up-to-date pathway activity inference methods. Finally, an overview of an optimisation model that uses a weighted sum of gene expression values as newly constructed pathway activity features is given. Overall, the information provided in this chapter sets the scene for the research described in the remainder of this thesis.

2.1 Methodologies for profiling data, applications and limitations

Gene expression is the process by which the information encoded in a gene is used in the synthesis of a functional gene product, e.g. proteins. The end products of gene expression would ultimately affect the sample phenotype as the final effect. Genome-wide profiling of gene expression is the study of transcriptions at a genomic scale at a single time [9]. The major application of transcriptome profiling is to determine what genes or transcripts are enriched in a particular disease, disease phenotype, tissue type or cell type. The emerging and rapidly developing genome profiling technologies in recent decades make the transcriptome analysis an integral part of many genomic studies of disease and biological processes [10].

Transcriptome profiling data are typically obtained from microarray or sequencing technologies. Although all of these techniques are designed to measure the number of transcripts in a sample, there are many differences between them in terms of profiling principles and data pre-processing methods. Therefore, it is better not to generalise when analysing the data generated by the different techniques. It is important to establish the methodological characteristics. Therefore, essential concepts of the different gene expression profiling technologies are described below.

2.1.1 Microarray

Typically microarrays simultaneously monitor the expression levels of thousands of genes in an organism [11]. The microarray chip is a special glass or silicon chip with thousands, or tens of thousands of nucleic acid probes attached. The gene expression levels of a tissue are measured by the signal intensity of fluorescent molecules (genomic DNA or mRNA) that are bound to complementary probes on the surface of the chip [12].

As one of the earliest technologies that allow the acquisition of vast amount of complex digital data, the application of microarrays has shown its potential as a medical diagnostic tool [13, 14]. Given a set of training samples with relevant labels (e.g. “healthy” and “disease”), the label or condition of a new patient can be predicted. Feature (gene) selection consequently become an intuitive approach to conduct the task. This process aims to identify molecular gene markers, i.e. a subset of genes that can be associated to a specific disease condition [15]. These identified gene sets, which are called Differentially Expressed Genes (DEGenes), are considered to contribute in distinguishing the samples across two conditions. Various methods have been developed to detect DEGenes from gene expression data [13, 14] using traditional means such as statistical t-test, regression models or mixture models [15]. Although various methods exist, the key idea for detecting DEGenes is to identify the expression level differences between two conditions. DEGene detection is used frequently in the analysis of high-throughput data and will be mentioned frequently in the following sections.

Microarray technology is limited due to the following reasons. First, the measured expression level values are related to the concentration of molecules in the solution, rather than the true number of molecules in the sample. Second, because of the shared sequence between genes, it is difficult for a single probe on the chip to detect only the gene that it is designed for. Finally, the microarray chip can only detect the sequences that the array is designed to detect, as it is a hybridisation-based approach. If there are nucleic acids in the solution but there are no complementary sequences on the chip, those genes would not be detected [16, 17].

2.1.2 RNA Sequencing

In contrast to microarrays, sequence-based RNA sequencing (RNA-Seq) provides a relatively unbiased perspective to simultaneously map and quantify nucleic acids in solution. RNA is first converted to a library of cDNA (complementary DNA) fragments with sequencing adaptors attached to one or both ends of the cDNA molecules. Then the molecules are sequenced in a high-throughput manner to obtain the short sequences. After sequencing, the resulting reads are either aligned to a reference genome or reference transcripts to obtain the expression level of each gene [17, 18].

In principle, RNA-Seq can determine the absolute quantity of the molecules in bulk tissues with appropriate read count normalization methods. The prevailing normalization methods for RNA-Seq data is TPM (transcript per million) [19], which deduces the specific expression level of a gene from the total number of reads falling into the exons of the gene, normalised by the length of the exons. The particular advantage of RNA-Seq is that it can capture the dynamic transcriptome across different tissues or conditions, which enables RNA-Seq to track the gene expression variations during cell development or among different tissue conditions. In this perspective, RNA-Seq provides robust support for investigating the molecular differences in biomedical samples across multiple phenotype labels in complex diseases.

2.1.3 Single-cell RNA Sequencing (scRNA-Seq)

The RNA-Seq technology relies on the macroscopic level measurements from cell populations. However, the observed macroscopic “average” cannot be directly used as an accurate evaluation of the “average” of the cells, as it results from the interactions between individual cells [20]. Thus, based on RNA-Seq technology, scRNA-Seq achieves a comprehensive way of measuring the transcriptomes of individual cells.

The development of microfluidics technology improves the efficiency and reliability of isolating single cells from tissue. For example, the commercial 10X Genomics platform [21] incorporates the concept of separating single cells by controlling the droplet volume and flow rate. Also, the application of unique molecular identifiers (UMI) [22] is a crucial step in delivering scRNA-Seq. With the UMI bonded to the mRNAs released upon cell analysis, the followed sequencing step, which is taken on a pool of cells, is able to track transcript counts at a single-cell level.

In summary, the rapidly evolving technologies of gene expression profiling makes it possible to improve understanding of mechanisms at genomic level. Furthermore, the availability of such datasets poses a need for advanced computational methods to model the large volume datasets and provide comprehensive mathematical and biological interpretations. In the profiling data, the amount of features (i.e. genes) for each sample could reach tens of thousands, depending on the size of genome species and the profiling technology. Therefore, designing a practical solution for dimensionality reduction is essential for obtaining useful information from the large number of features. In this field, DEGenes [13, 14] has become a mature method for identifying the important genes that can differentiate sample labels at gene-level and has become a primary approach for analysing the gene expression profiling data. However, based on the fact that genes work together to perform various cellular functions though biological pathways, incorporating pathway information in the analysis of genome expression data, rather than focusing on individual genes, is popular. Therefore, a brief introduction is given next to exhibit key background to typical pathway analysis.

2.2 Analysis of profiling data incorporating biological pathways

2.2.1 Motivation for pathway approaches

Various machine learning-based methods were designed to deal with large high-throughput molecular datasets. These computationally advanced methods aim to identify a group of disease-related genes from the tens of thousands of genes in profiling data. The identified gene groups are expected to be used as gene biomarkers and benefit disease diagnosis or prognosis. Disease diagnosis was normally achieved by using clinical indicators. However, the gene biomarkers show no significant advances compared to traditional clinical indicators [23–25].

A concrete example of using machine learning methods to analyse high-throughput data and obtain a prognostic gene list, namely 70-gene classification profile [26]. The 70-gene list is derived from the microarray data of breast cancer patients. They used machine learning approaches to search for the 70 genes correlating with follow-up clinical outcomes. Little biological knowledge is involved in this process. The expression pattern of the 70-gene among the sample cohort is shown in Figure 2.1.

The gene signatures is selected by calculating the correlation score of the expression levels and the outcome indicators. Then genes are ranked according to their correlation score and the top 70 genes are selected as the marker genes. A series of classifiers was constructed to verify the prediction accuracy of the selected genes. However, this 70-gene list shared small overlap with the other identified lists of genes [27, 28], and suffered from inherent instability when repeating the selection procedure [29].

This phenomenon can be explained in two aspects. First, from a mathematical perspective, due to the large-p, small-n nature of the high-throughput data, gene ranking can shift by hundred with a small change in the dataset. Second, from biological perspective, when changing the training patient set, a completely new 70-gene signature may be generated because of the heterogeneous nature of disease. Although this problem can be overcome by increasing the number of training samples, thousands of additional samples are needed to produce a 50% overlap rate of significant genes [7]. Therefore, the drawbacks of these purely statistical and machine learning approaches motivated the emergence of incorporating biological knowledge into the analysis pipeline to gain more robust results.

A biological pathway is a series of interactions among molecules in a cell that leads to the production of a certain product, a signalling event or a change in a cell (Figure 2.2) [30]. The functions of pathways include triggering the assembly of new molecules(i.e. proteins), turning on or turning off genes, or spurring a cell to move. Most complex diseases are caused by a series of genetic mutations rather than a single gene mutation. These mutated genes disrupt the pathways they are associated with, thereby driving the occurrence of disease.

Taken further, the identification of disease related pathways simplifies the complex task of focusing on multiple gene-level mutations. The identification of a few altered pathways can consequently facilitate the analysis of the progression, diagnosis, and treatment of the disease. For example, a well-known key tumor suppressor gene, *tp53*, is linked to cellular immortalization, which is a crucial step in the tumor transformation of

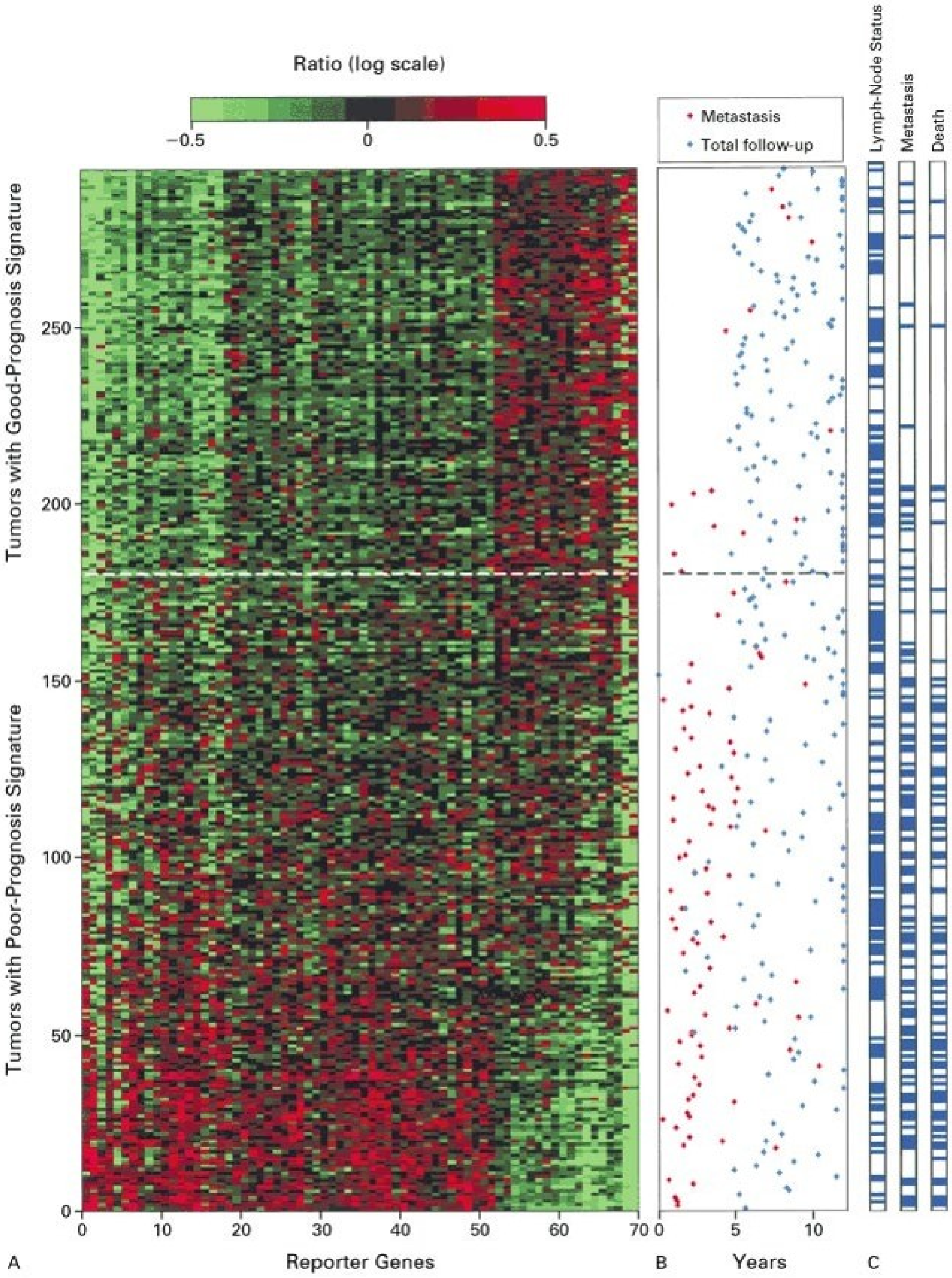


Fig. 2.1 The pattern of expression of the 70 marker genes (columns) in a series of 295 consecutive patients (rows) with breast carcinomas. Figure taken from [26].

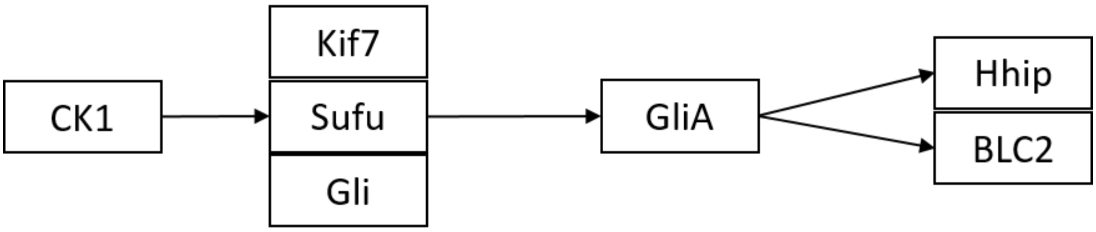


Fig. 2.2 An example of biological pathway diagram (simplified Hedgehog Signalling Pathway). Each box represents a member gene of this pathway, arrows linking genes can activate or inhibit.

a cell [31]. However, the loss of *tp53* is a necessary but not sufficient component for this process. There are additional gene mutations that are required for the immortalization process. Several studies identified a few pathways, such as the cell cycle pRB/p53 pathway [32], the cytoskeletal genes [33] and the MAP kinase pathway [34], which are enriched in key immortalization regulation genes. The identification of the *tp53* relevant pathways deepens the understanding of biological processes of tumor transformation. In the following paragraphs, pathway analysis approaches that focus on deriving the relevant pathways from gene expression data are introduced.

2.2.2 Pathway analysis approaches

As explained in the last section, the identification of disease-related pathways is more efficient than the identification of disease-related genes, pathway analysis has become a widely used approach for analysing the disease gene expression data. Similar to the identification of differentially expressed genes (DEGenes), Pathway Analysis (also known as functional enrichment analysis) means the detection of relevant pathways altered in disease samples compared to control samples [35]. The earliest pathway analysis approaches were extended from DEGenes analysis. In detail, pathway analysis finds the pathways that the DEGenes are enriched in.

West et al. [36] used microarray data to predict the clinical status of breast cancer, where they used validation of classifications with out-of-sample cross-validation methods to select relevant genes. The expression levels of this group of selected genes are most highly correlated with the classification of samples in tumor versus control classes. Based on this group of genes, regression models were used to assign the relative probability of the pathway deregulation in the sample cohort. The methodology reported in this work not only yields genes that are involved in the clinical phenotype (ER-regulated pathway genes), but also genes for which the connections are not immediately clear, such as such as HNF3α and androgen receptor. Thus, this work provided the potential for identifying pathways underlying an observed breast cancer phenotype. An extension of this work [37] further showed that genes that reflected the activation status of five oncogenic pathways can be identified using the same analysis pipeline. This type of research made the projection of genes onto pathways. Similar works that used DEGenes to identify the disease relevant pathways are [38, 39].

Despite the popularity and simplicity of the DEGene-based pathway analysis approaches,its power is limited by the pre-selected genes, which ignores other genes in pathways. A classic approach that does not rely on pre-selected gene lists is Gene Set Enrichment Analysis (GSEA) [40]. The working pipeline of GSEA can be simplified as follows: (i) Rank all genes in the dataset by their expression difference between two phenotypes. (ii) For each pre-defined gene set, e.g. GO terms or pathways, calculate the enrichment score (ES). The ES is a running summary statistic reflecting the spread of the member genes of a pathways among all genes. (iii) Estimate the significance of ES by calculating the nominal p value using an empirical phenotype-based permutation test. In summary, GSEA allows users to choose a pathway and determine its relative statistical significance to the sample phenotypes. The pathway score is measured to assess the combined contributions of the pathway member genes in the ranked list.

Since the appearance of GSEA, many approaches have been proposed for the analysis of entire gene sets [41–43]. A few reviews [44, 45] reported that these approaches can be divided into two groups, competitive tests and self-contained tests. The difference between these two groups is the definitions of the null hypothesis. The null hypothesis for Competitive tests is the pathway member genes are as often differentially expressed as the complement genes. The null hypothesis for Self-contained tests is no gene in the pathway is differentially expressed. To be more specific, Competitive test (e.g. GSEA) compares the differential expression of the member genes to the complement genes in the genome. Self-contained test compares the member genes to a fixed standard that does not depend on the of genes outside the gene set.

A representative self-contained tests approach is global test [46], which includes linear regression and a simple random effects model. The model gives a covariance matrix of the random effects that is then used to construct the pathway score test to test whether the null hypothesis, which is the regression coefficients of pathway genes are of zero variance, is true.

A review [46] compared the performance of these two types of pathway analysis approaches by using them to identify pathways involved in apoptosis process. The comparison between three self-contained approaches, global test [47], ANOVA global test [48] and SAM-GS [49], showed similar performance; only SAM-GS was slightly superior. In contrast, three competitive approaches, GSEA, SAFE [50] and Fisher's exact test [51], showed low power for identifying differential expression in phenotypes. To summarise, the five pathways involved in apoptosis were identified by the three self-contained approaches. For the three competitive approaches, SAFE and GSEA missed three of them, and Fisher's exact test did not identify any of them.

In summary, the analysis of pathways that are differentially expressed between phenotypes, has become a routine approach for the analysis of gene expression data. This analysis is intuitively appealing for several reasons. First, by projecting genes into pathways, the analysis complexity is reduced. Second, as it is pathways, not genes, that lead to a certain change in a cell, identifying differentially expressed pathways is more appealing than identifying genes from the biological point of view. Third, the genes in the list of differentially expressed genes are often highly correlated, and consequently cause a large number of false positives. Considering pathways leverages the correlation problem to some extent. More importantly, it is believed that, in many diseases, the changes of the expression values of individual genes are not significant enough to be detected. Therefore, the complex task of focusing on individual genes can be improved by looking into relevant pathways.

2.3 Pathway activity inference framework

2.3.1 Motivation and definition for pathway activity inference

Although pathway analysis approaches have been sufficiently explored and widely used in the analysing process of gene profiling data, they are limited in two aspects. First, most pathway analysis approaches are only applicable to binary experimental conditions, e.g. disease versus control. Second, the diverse range of rapidly expanding data produced by modern molecular biology has fueled a need for accurate classification and prediction toward individual samples [6], where the need cannot be satisfied by classic pathway analysis approaches. Most pathway analysis approaches provide only a list of pathways with their p-values to show the extent of the pathway significance. They point out the important pathways towards specific questions, but it is hard to indicate the exact degree to which the pathway is deregulated. In brief, results given by pathway analysis approaches are coarse, and there is no quantitative analysis at pathway-level.

Therefore, pathway activity inference is proposed to fulfil these needs, where gene expression data are first aggregated at pathway level to yield a compact representation

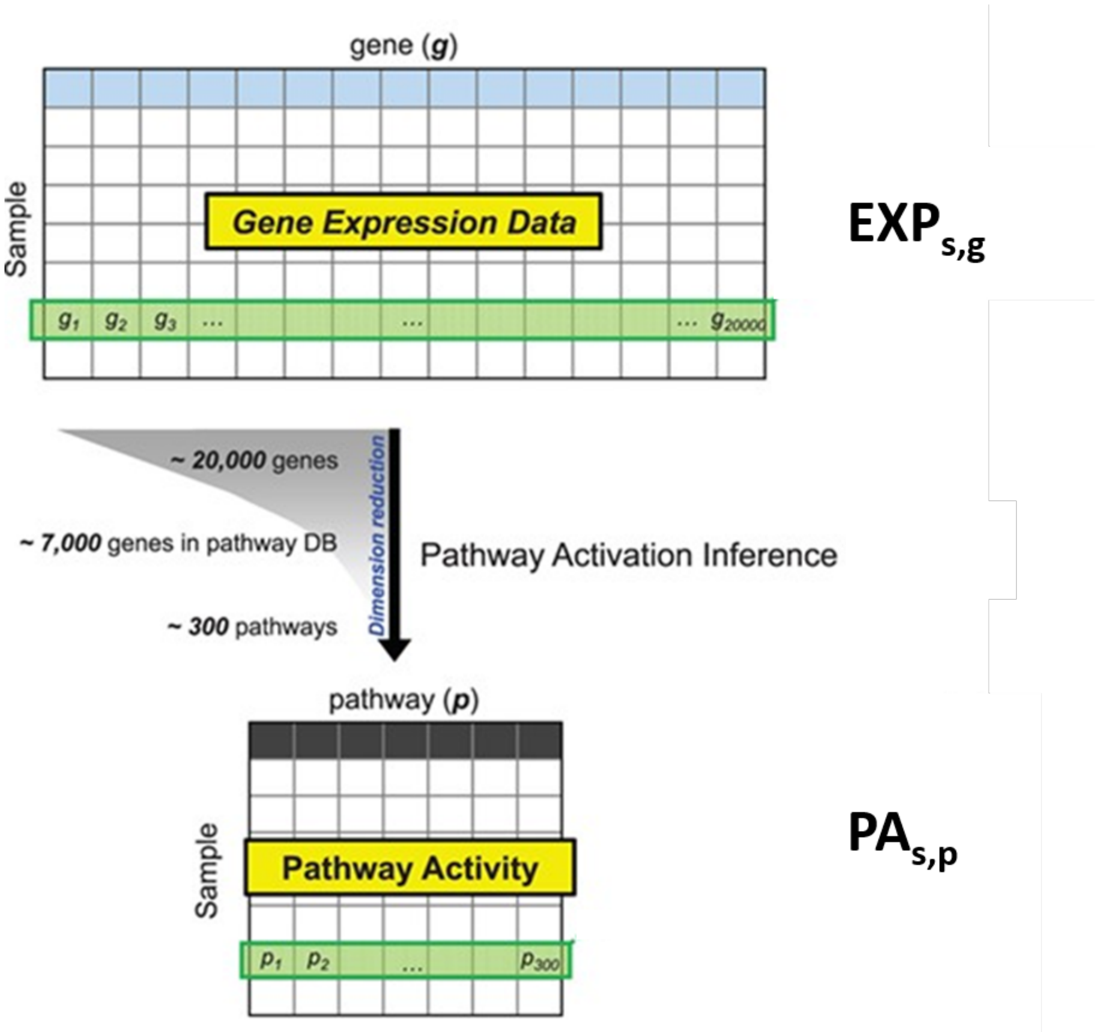


Fig. 2.3 Illustration of the strategy for pathway activity inference. The pathway activity inference method takes gene expression data as input and produces pathway activity matrix, which can be interpreted as a mapping of genes to the pathway dimension. Figure taken from [52].

of the original data. The columns in the gene expression profile are replaced with pathways. Figure 2.3 shows the schematic of pathway activity inference.

By aggregating gene-level data into pathway-level data, the shortcomings of pathway analysis approaches can be overcome and a boarder range of analysing aims can be pursued. First, significant pathways can be identified by applying statistical methods on the pathway-level data. Secondly, the pathway space data can be used for sample classification and prediction using machine learning classifiers. More importantly, the "large-p small-n" problem of the profiling data can be solved this way. After the aggregation, the extremely large dimensionality of genes is reduced to a few hundred of pathways, which allows the number of features to become comparable to the number of samples. At the same time, the dimension-reduced data is better for biological interpretations, as the cellular functions are reflected better by the pathways rather than single genes. Last, this transformation is expected to yield a more robust representation of the data that can reduce the intrinsic technology and biological variance across samples [53]. Because the expression values of the member genes in a pathway could vary considerably across samples of the same phenotype, the summarised pathway activity values can become consistent across samples.

The pathway activity inference represents a well-established framework, which follows three steps, data collection, pathway activity profile calculation and pathway activity evaluation. The following sections introduce the general procedure for these three steps in detail.

2.3.2 Data collection

2.3.2.1 Profiling dataset collection

Pathway activity inference procedure starts by collecting the gene expression profiling datasets. The profiling datasets can be obtained from publicly available databases. The most famous database is Gene Expression Omnibus (GEO) [54], a public high-throughput functional genomics data repository that accepts both array-based and sequence-based data. Another landmark cancer genomics database is The Cancer Genome Atlas (TCGA), which characterised molecularly over 20,000 primary cancer, matched normal samples spanning 33 cancer types and generated over 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data[55].

Table 2.1 List of datasets

Name	Database	Data Type	Disease	Tissue	Chapter
GSE22255	GEO	Microarray	Stroke	PBMCs	3
GSE16561	GEO	Microarray	Stroke	Whole Blood	3
GSE58294	GEO	Microarray	Stroke	Whole Blood	3
COAD	TCGA	RNA-Seq	Colorectal Cancer	Tumor	4 and 5
BRCA	TCGA	RNA-Seq	Breast Cancer	Tumor	5
PBMC 3k	10X Genomics	scRNA-Seq	-	PBMCs	6
PBMC 10k	10X Genomics	scRNA-Seq	-	PBMCs	6
GSE114725	GEO	scRNA-Seq	Breast Cancer	PBMCs	6

Table 2.1 summaries all the profiling datasets used in this thesis. The three data types, Microarray, RNA-Seq and scRNA-Seq and three kinds of diseases, breast cancer, colorectal cancer and stroke are analysed in different Chapters of this thesis.

The main concern with dataset usage is the availability of sample annotations. For example, if the research question deals with analysing the differences between breast cancer subtypes (dataset BRCA in Table 2.1), the samples labels can be Luminal A, Luminal B, HER2, etc.

2.3.2.2 Prepossessing imbalanced data

A dataset is imbalanced if the classification categories are not approximately equally represented [56]. Particular sampling technique is needed for training a classification model with imbalanced data. Unbalanced data is not uncommon for profiling datasets. For example, the dataset COAD in Table 2.1 contains 41 normal tissue samples and 480 tumor tissue samples. This issue can be solved either by oversampling the minority class and/or under-sampling the majority class. Under-sampling the majority class is not suitable for the COAD dataset, therefore, in this thesis, I deal with such imbalanced data with an oversampling algorithm called Synthetic Minority Oversampling Technique (SMOTE) [57].

SMOTE over-samples the minority class by taking each minority class sample and introducing synthetic samples along the line segments joining any/all of the k nearest minority class neighbors. Depending upon the amount of over-sampling required, The number of k are randomly chosen. Synthetic samples are generated in the following way: Take the difference between the samples under consideration and its nearest neighbor. Multiply this difference by a random number between 0 and 1, and add it to the samples under consideration. This approach effectively forces the decision region of the minority class to become more general. [56]

SMOTE has been applied throughout this thesis to deal with the imbalanced sample classes, especially for binary classification problems.

Table 2.2 List of pathway collections

Name	Database	No. Pathway	No. Gene	Avg. Size	Chapter
KEGG	MsigDB C2	186	5267	50-150	3
KEGG	KEGG	279	6761	50-150	4 and 5
Biocarta	MsigDB C2	292	4383	30	6
Hallmark	MsigDB C2	50	1509	200	6

2.3.2.3 Biological pathway collections

Biological pathways form another part of input data for pathway activity inference. Academic and commercial institutions have been generating and maintaining pathway databases [58], for example KEGG [59], Reactome [60], WikiPathways [61], NCIPathways [62], MSigDB [63] and Pathway Commons [64]. These databases are open-sourced and well-established. The pathway databases can be categorised into two types, the primary pathway databases and the integrative databases. The pathways that are discovered and assembled in the experimental biology labs are collected in the primary databases; the integrative databases assemble pathways resources from multiple primary databases, and provide a more comprehensive view of the pathway landscapes.

The major primary pathway database is KEGG and the major integrative database is MSigDB, according to the number of citations in studies investigating pathways associated with variable gene expression patterns in different sets of conditions (27,317 publications using KEGG and 2,892 publications using MSigDB by 2019) [65]. The number of pathways in KEGG is 316 in 2022, and the number keeps increasing as new pathways are found. MSigDB contains over 1000 gene sets and contains the older versions of KEGG and Reactome gene sets. The number of KEGG pathways in the C2 collection of MSigDB is 186, which is the most widely used KEGG pathway gene sets. To gain a better representation of these wide range of pathways, increase the utility and decrease the redundancy across the pathways, MSigDB developed a pathway collection called Hallmark [66]. The Hallmark collection contains 50 gene sets, each consisting of a “refined” gene set, derived from multiple “founder” sets, which conveys a specific biological state or process and displays coherent expression.

In Chapter 3, Chapter 4, Chapter 5 and Chapter 6 the latest version of KEGG pathways, MSigDB Hallmark collection, and Biocarta collection are used for analysis. Table 2.2 lists the details of these pathway collections. "No. Pathway" and "No. Gene" represent the number of pathways and number of unique genes in the pathway collection respectively. The "Avg. Size" means the approximate average number of genes for the pathways in the collection.

2.3.3 Pathway activity profile calculation

2.3.3.1 Pathway expression matrix construction

After the gene expression profiles and the pathway gene sets are pre-processed and ready for use, the first step is to construct the pathway-specific expression matrix $PWM_{s,m}$, where the rows are samples s, the columns are the member genes m of pathway p, and the values are the gene expression values. In other words, the gene expression profile $EXP_{s,g}$ is split into hundreds of pathway expression matrices $PWM_{s,m}$. This step is illustrated in Figure 2.4.

The dimensionality of $PWM_{s,m}$ is usually between tens of genes and hundreds of genes. The constructed $PWM_{s,m}$ is the actual input data for the pathway activity inference methods.

2.3.3.2 Pathway activity inference

Pathway activity inference refers to calculating the pathway activity values PA_s from the $PWM_{s,m}$ (Figure 2.5). The PA_s represents the overall performance of the member genes on the sample s. This is the core step for the entire process, where pathway activity inference methods are proposed for the PA_s calculation.

A straightforward method called MEAN [53] is introduced to describe this process more concretely. The MEAN method takes the mean value of the expression levels of all member genes in a pathway as the pathway activity value for sample s, as shown in Equation (2-1). This method has been widely applied in literature [67–70]. A variant of the MEAN method uses the median of the member gene expression values as the pathway activity values (referred to as MEDIAN) [70]. Both methods represent intuitive ways to summarise the pathway activity values.

$$PA_s = \overline{PWM_{s,m}} \qquad (2-1)$$

The inferred pathway activity values aim to yield better means to distinguishing sample phenotypes (sample labels) when compared to using gene expression values. High-quality pathway activity can cluster samples, predict phenotypes and reveal biological insights. Therefore, more sophisticated and specialised methods of pathway activity inference are more reasonable than using statistical methods (i.e. the mean or median

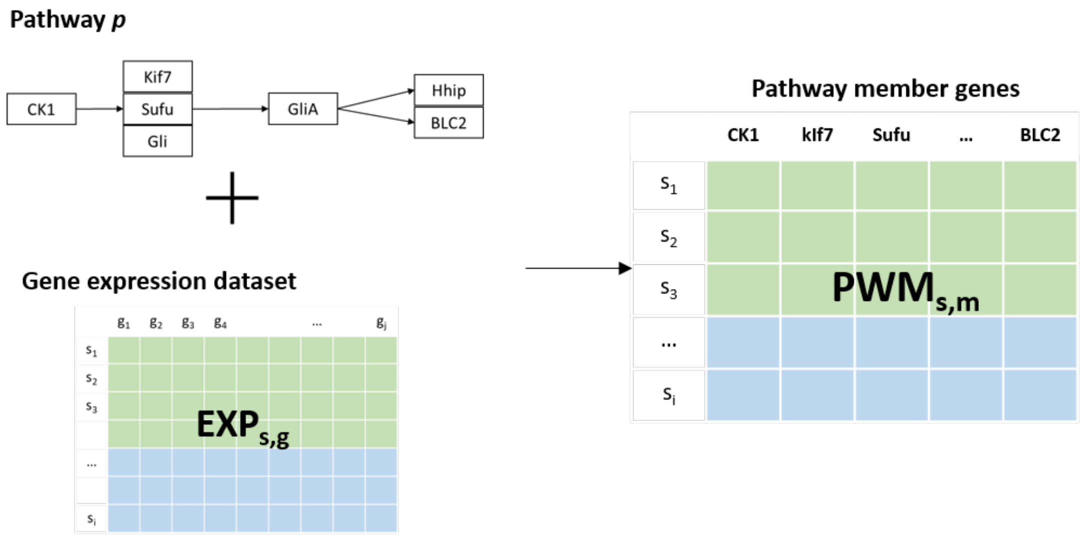


Fig. 2.4 Pathway expression matrix construction.

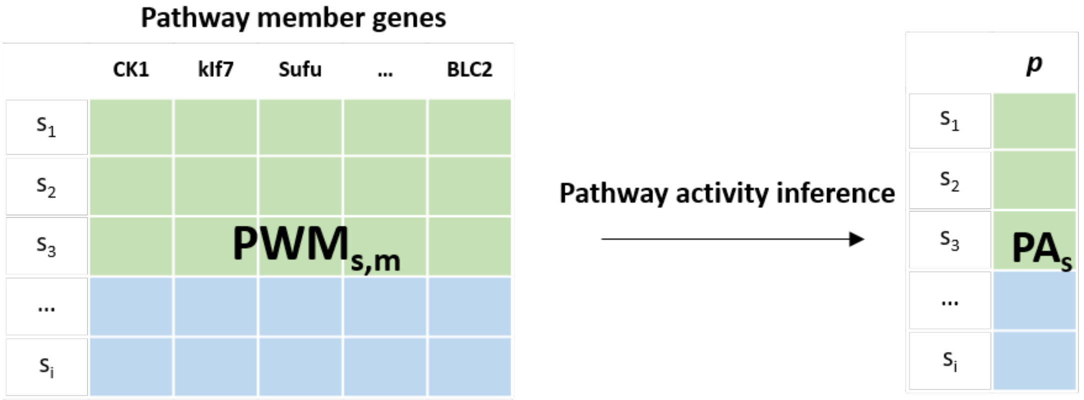

Pathway member genes
CK1 klf7 Sufu ... BLC2
s_1 s_2 s_3 ... s_i
$PWM_{s,m}$
Pathway activity inference
p
s_1 s_2 s_3 ... s_i
PA_s

Fig. 2.5 Pathway activity inference.

of expression values). In the Section 2.4, a review of the representative pathway activity inference methods will be delivered.

2.3.3.3 Pathway activity profile

The final step in the pathway activity inference procedure is to assemble the PA_s vectors to form the pathway activity profile $PA_{s,p}$. Until here, the pathway activity inference procedure is completed. Through the pathway activity inference methods, the gene expression profile $EXP_{s,g}$ is converted to pathway activity profile $PA_{s,p}$. The dimensionality of the gene expression profile dataset is significantly reduced. Figure 2.3 summarises this process from a dimension reduction perspective. By mapping samples from the gene space to the pathway space (the clustering plots on the right side), the expression data can be interpreted more efficiently since the number of dimensions is reduced from over 20,000 genes to about 300 pathways [52]. Also, as pathways are well-curated knowledge, biological interpretability is introduced via the pathway activity profile.

2.3.4 Pathway activity evaluation

A variety of pathway activity inference methods, including arithmetic, enrichment, statistical, machine learning approaches and mathematical programming are used to calculate PA_s. It is, therefore, necessary to use uniform criteria to evaluate the quality of the pathway activity values inferred by these methods. This section introduces the widely used pathway activity evaluation criteria.

2.3.4.1 Machine learning classifiers for sample classification

The priority evaluation criteria of pathway activity inference method is sample classification. To be more specific, it means using the pathway activity profile (PA_{sp}) to train a machine learning classifier and calculate the classification accuracy.

The term *machine learning* refers to the automated detection of meaningful patterns in data. In the past couple of decades it has become a common tool in almost any task that requires information extraction from large datasets [71]. Machine learning algorithms are designed for performing multiple tasks, including classification, regression, clustering, etc. Section 2.2.1 has given an example of applying machine learning algorithms on the biological field. For the pathway activity inference methods, machine learning classifiers are adopted to evaluate the sample prediction accuracy, as this is an efficient way to measure how the pathway activity can separate the samples of different labels.

The widely used machine learning classifiers include Decision Tree [72], Logistic Regression [73], Random Forest (RF) [74], and K-Nearest Neighbor (KNN) [75]. I mainly use KNN and RF in the following chapters. KNN is a type of instance-based learning, who does not attempt to construct a general internal model but stores instances of the training data. Classification is computed from a majority vote of the k nearest neighbors of the query point. RF is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control overfitting.

The classifier performance is usually evaluated in terms of two measures: accuracy and F1-score, which are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \qquad (2-2)$$

$$F1 - score = \frac{2 * (precision * recall)}{precision + recall} \qquad (2-3a)$$

$$Precision = \frac{TP}{TP + FP} \qquad (2-3b)$$

$$Recall = \frac{TP}{TP + FN} \qquad (2-3c)$$

TP, TN, FP and FN denote true positive, true negative, false positive and false negative, respectively. Positive or negative indicates the predicted classifier outcome; true or false indicates whether the prediction is correct. For example, for a sample labelled as "1", the prediction of this sample is "1", then it is a true positive; if its prediction is "0", then it is a false negative. In binary classification, the default positive label is "1".

Accuracy is a percentage quantity showing the number of times that the classifier is correct in its classification, and it conveys the right intuition when the positive and negative populations are roughly equal in size. Precision is the percentage of times that the classifier is correct in its classification of positive samples. Recall is the percentage of known positive samples that the classifier would classify as being positive. The F1 score is the harmonic mean of the precision and recall [76].

In the multiclass case, TP, FP and FN calculate the true positive, false positive and false negative of the target class versus the rest classes [77]. For example, when there are three classes, "0", "1" and "2", the FP value for "0" is the number of samples predicted to be "0" but with a true class of "1" or "2". Further, the calculation of precision and recall are defined as weighted averaged value of each class:

$$Precision = \sum P_i W_i \qquad (2-4a)$$

$$Recall = \sum R_i W_i \qquad (2-4b)$$

where i is the index of classes, W_i is the weight of class i (percentage of the number of samples in i), P_i and R_i are the precision and recall of class i, respectively.

2.3.4.2 Validation pipeline

A reliable validation pipeline is a vital component to gaining precise classification accuracies. The validation process consists of internal and external validation. The internal validation is assessed by K-fold cross-validation [78]. First, the gene expression profile is divided into k subsets of approximately equal size sets. Usually, k is in the range of 5 to 10. Also, the k-fold split should be stratified (i.e. dividing members of the population into homogeneous subgroups before sampling) to keep the sample label proportions in each training fold the same as in the original dataset.

Then the pathway activity profiles are calculated for the $k-1$ folds and the remaining fold separately. $k-1$ folds are used as the training samples, and the remaining fold is used as the testing set of samples. Consequently, k training pathway activity profiles and k testing pathway activity profiles are generated. Finally, the machine learning classifier is trained on the training pathway activity profiles and tested on the testing profiles.

The application of K-fold cross-validation avoids the over-fitting problem caused by full training [79]. The classifier is trained and tested several times on different samples to ensure the reliability and robustness of the results. The final classification accuracies (and other measurements) are averaged from the k times classification.

Although internal cross-validation is a convenient solution to assess the method performance within a single dataset, it may lead to optimistically higher performance estimates. Thus, it is increasingly being noted that it is also meaningful to externally validate the performance of the pathway activity inference method from an independent test dataset [80, 81]. In the external validation process, the machine learning classifier is trained on one dataset and tested on the other. The two datasets are independent but have identical sample labels. The data split strategy for both datasets is the same.

However, external validation is not mandatory. Although external validation can gain a more realistic estimation of the generalization of the pathway activity inference method, the technical differences (e.g., batch effect) between the datasets and the biological differences between phenotypic classes can significantly impact the classifiers' effectiveness by making classification results lose fairness and underestimated the performance of the pathway activity inference methods. Therefore, external validation is often used in binary classification problems. [82].

2.3.4.3 Survival Prediction

In addition to the main purpose of pathway activity inference, another important application is using the pathway activity values to perform survival analysis. Survival analysis is an important analysis aspect in the research of complex disease, especially cancer study. Literature proves that the activity score of specific pathways relate well with the tumor degree and patient survival rate [83, 84]. Also, the basic pathway analysis approaches also put forward prognostic models based on gene expression values [26, 85, 86]. Therefore, as an advanced way of summarising the pathway activity values for individual patient, pathway activity inference methods are expected to provide higher survival prediction accuracy compared with gene-level methods.

Survival analysis is the analysis of time-to-event data. Such data describe the length of time from a time origin to an endpoint of interest. For example, individuals might be followed from birth to the onset of some disease, or the survival time after the diagnosis of some disease might be studied. In summary, survival analysis methods are usually used to analyse data collected prospectively in time [87]. Therefore, the clinical information needed for survival analysis includes two parts, (i) survival status (e.g. 0 for alive or 1 for the dead), (ii) survival time (in months or years). The most widely used survival model is The Cox proportional hazards model [88], which builds a regression model between the survival time and clinical features (e.g. race, age, gender, etc.) to examine whether survival times are related to the features. With the appearance of gene expression profiling technologies, the Cox model has been applied to gene features for predicting survival time [85, 89–91].Besides the Cox model, many binary classifiers are also used for survival analysis. These classifiers distinguish samples into "low-risk" and "high-risk" categories. Such classifiers include support vector machine (SVM), random forest classification (RF) algorithms, and logistic regression.

A prevalent approach is the Random Survival Forest [92], which combines the advantages of the above two approaches, accepts many features and returns a risk score. Random Survival Forests is an ensemble tree-based method and is an extension of the random forest method. Survival trees are built by recursively partitioning the feature space using binary splits to form groups of samples who are similar according to the survival outcome. The final predictor ensemble is formed by aggregating the results of many survival trees [93]. The Random Survival Forest has become the most common survival prediction model for survival analysis with pathway activity values as inputs.

The concordance index (c-index) is used to evaluate the predictions made by the survival model. It is defined as the proportion of concordant pairs divided by the total number of possible pairs. C-index=0.5 indicates poor predictive accuracy, and c-index=1.0 indicates good predictive accuracy [85].

2.3.4.4 Robustness metric

No matter which high-throughput technology is used for profiling, background noise is unavoidable in gene expression profile data [94]. However, the pathway activity inference methods should be able to counteract noise in data. Therefore, the third pathway activity measurement evaluates the robustness against the noisy data. This process follows two steps, (i) adding noise to the expression data to perturb the dataset and (ii) evaluating the classification accuracies on the pathway activity inferred from the perturbed gene expression data.

There are many approaches to add noise to gene expression data. For example, multiplying the log-transformed expression data with different levels of perturbation factor [52] or randomly selecting genes and replacing the data with values sampled from a normal distribution with the same mean and variance as the original gene [95]. The most simple and efficient approach consists of randomly permuting the order of the samples in the expression matrix for the selected genes. During noise-adding processes, the proportion of the perturbation is controlled to simulate different noise levels in the data.

The robustness of the pathway activity inference method is then determined by the accuracy of the classification of sample types in terms of the pathway activity values inferred from the perturbed gene expression data. The pathway activity inferred from noisy data is expected to keep the sample separation capacity as the original data. Thus, as the percentage of perturbation increases, the more stable the classification accuracy indicates that the method is more tolerant to noise.

In conclusion, the evaluation of pathway activity is tightly related to the purposes of the pathway activity. First, methods that have higher sample classification accuracy are better. Then, robustness against noisy data is also important to ensure the method's feasibility. Lastly, the application of pathway activity for survival analysis is widely discussed and has shown promising performance in literature.

2.4 Pathway activity inference methods in literature

The previous section shows the value and efficiency of pathway activity inference as a dimension-reduction tool for analysing high-throughput profiling data. Moreover, they demonstrate the need for developing methodologies that can achieve accurate sample prediction accuracies and important pathway identification. Such methods have been studied since 2005 [53] and have been in continuous evolution. This section introduces the pathway activity inference methods from literature in order of complexity.

2.4.1 Baseline pathway activity inference methods

2.4.1.1 Mean-based methods

Mean-based pathway activity inference methods are the most straightforward method, which has frequently appeared in the literature in one form or another. Except the MEAN method (mean of all pathway member genes) introduced in Section 2.3.3.2, two other widely used mean-based methods exist. A simple variant of the MEAN method is calculated by averaging only the top half of the member genes with larger t-statistics (Mean top 50%), which is proposed by [82] and is used in their study as a comparison.

Another variant is called CORGs (condition-responsive genes) [96]. This method calculates the pathway activity by the mean expression of key member genes, instead of all the member genes. CORGs are the genes that, upon aggregating their expression profiles by averaging, yield a pathway activity vector PA_s that is the most discriminative between two classes in the data. Identifying the CORGs of a pathway in a given two-class dataset begins with a t-test on z-scaled $PWM_{s,m}$. Then, the expression direction (that is, up or down) of the pathway is determined as the sign of the averaged t-statistic values of its member genes ($\bar{t}$). Next, all the member genes are sorted by their t-statistic values in accordance with the overall regulatory direction of the pathway; the most strongly up-regulated genes are arranged to the top for an overall up-regulated pathway, whereas the most strongly down-regulated genes are arranged to the top for an overall down-regulated pathway, as illustrated in Equation (2-4).

$$T_m \geq t_{m+1}, \qquad if \qquad \bar{t} \geq 0 \qquad (2-4a)$$

$$T_m \leq t_{m+1}, \qquad if \qquad \bar{t} < 0 \qquad (2-4b)$$

Then, the CORG set is initially set to contain only the top-ranked gene and iteratively expanded. At each iteration, the gene of the next rank is added to the candidate CORG set, PA_s is obtained by taking the mean of expression profiles of the candidate CORGs, and a t-test is then performed to give the pathway t-statistic. The iteration stops when the pathway's t-statistic $S(G_k)$ no longer improves, at which point the final CORG set is obtained. This process is captured in Equation (2-5):

$$PA_s = \sum_k \frac{EXP_{s,m}}{sqrt(k)} \qquad ,k \in G_k = m_1, m_2, ..m_k \qquad (2-5)$$

where G_k is the GORG set of pathway p, if k is the smallest number satisfying $S(G_{k+1}) <= S(G_k)$.

2.4.1.2 Projection-based methods

Principle component analysis (PCA) has long been applied to the analysis of gene expression data, especially for exploratory data visualization to discriminate between sample groups. Using PCA, the correlation matrix is first computed from z-scaled gene expression data. Then, through the Eigen-decomposition of the correlation matrix, major directions in the data with the largest variability are identified as eigenvectors corresponding to the largest eigenvalues of the correlation matrix. The eigenvectors are called the principal components (PCs). In addition to its use in exploratory data visualization, PCA has also been used as a pathway-level aggregation method in several literature studies [43, 97, 98]. In calculating pathway activity, PCs are found by applying PCA to the z-scaled $PWM_{s,m}$. The projection of the $PWM_{s,m}$ onto the first PC is taken as the PA_s of that pathway.

Another projection-based pathway activity inference method is Partial least squares Regression (PLSR) [98], a regression method that combines properties of multiple regression and PCA. PLSR analysis consists of a data matrix X and a response matrix Y, which contain values of the independent and the dependent variables, respectively. Unlike the standard regression method, which builds a regression model between the data matrix X and response matrix Y, PLSR seeks to build a regression model between the latent component scores of X and those of Y. The latent variables are expected to best summarise the variance in X and are the most relevant for the response Y. [98] used PLSR as a pathway activity inference method. In this approach, data consist of a matrix X of the z-scaled $PWM_{s,m}$ and a numeric sample label vector Y. Each element in the label vector indicates class membership of the samples using 0 or 1. The first latent component is taken as the PA_s of that pathway.

However, the latent components obtained under this dummy coding scheme lack meaningful biological interpretation. Since "1" is a larger numeric value than "0", regression on these two dummy numeric values makes the inferred pathway activity values larger in “1” samples and smaller in “0” samples. Although the differentiation ability of the inferred pathway activity values increased compared with PCA, the subsequent pathway level analysis would falsely reveal relationships among the sample labels because numeric values are assigned to nominal labels.

2.4.2 Extensions of GSEA

The Analysis of Sample Set Enrichment Scores (ASSESS) method [99] can be considered as a sample-level extension of the GSEA (introduced in Section 2.2.2). The difference between GSEA and ASSESS is that GSEA provides an overall enrichment score of a pathway for two sample groups, and ASSESS provides an enrichment score for each sample. For this reason, ASSESS employs the random walk computations twice. The first use of random walk is applied at the individual gene level. Given an expression level of a gene in a sample, ASSESS calculates the sample’s log-likelihood ratio of belonging to one class instead of the other. The second use of the random walk is at the level of each pathway. Using the log-likelihood ratio values obtained for its member genes, it computes the enrichment score of a pathway for a sample by the maximum deviation of the random walk from zero.

Another more widely known extension of GSEA is Gene Set Variation Analysis (GSVA) [100], which estimates the variation of pathway activity over a sample population. Instead of ranking the genes using the whole set of samples, GSVA first ranks the genes for each individual sample using non-parametric kernel estimation of its cumulative density function. In the case of microarray data, a Gaussian kernel [101] is used, and in the case of RNA-Seq data, a discrete Poisson kernel [102] is employed. For each sample, the genes are ranked according to the expression-level statistic. The following step condenses expression-level statistics into pathway-level by calculating sample-wise enrichment scores. The enrichment score assessment is similar to the GSEA and ASSESS methods [40, 99] by using the Kolmogorov-Smirnov (KS) like random walk.

It should be noted that the ASSESS is a supervised method, and GSVA is an unsupervised method. Therefore, ASSESS is well-suited for assessing gene set variation across a dichotomous phenotype, while by omitting phenotypic information, GSVA enables more flexible downstream analyses and broader applications. There are also many other GSEA-based pathway activity inference methods, including ssGSEA [103] and ESEA [104]. A more detailed description of these methods can be found in their original publication.

2.4.3 Advanced arithmetic pathway activity inference methods

The basic arithmetic methods, including the MEAN, Mean top 50%, and the Median mentioned in Section 2.3.3.2, have been widely used because they are simple and convenient. More advanced arithmetic methods are also proposed to compensate for their shortcomings.

A representative method is the Log-likelihood ratio (LLR) [105]. The LLR first estimates the conditional probability density functions under different phenotypes for each gene in the pathway. Based on the conditional probability density functions, the method then transforms the expression values of the member genes into log-likelihood ratios to obtain an LLR matrix from the gene expression matrix. The process is summarised in the following formulas:

$$\lambda_i(x_j^i) = log\left[\frac{f_i^1(x_j^i)}{f_i^2(x_j^i)}\right] \qquad (2-6a)$$

where i is the index of genes, and j is the index of samples. f^1 and f^2 are the conditional probability density function of the expression level of gene i under phenotype 1 and 2, respectively. λ_i is the log-likelihood ratio (LLR) between the two phenotypes for the gene i. Then the gene expression value x_j^i is transformed into $\lambda_i(x_j^i)$ from the LLR matrix (PWM_j^i). The LLR matrix is then normalised, and the pathway activity for sample j is inferred by combining the normalised LLRs of its member genes using the:

$$A_j = \sum_i (\lambda_i(x_j^i)) \qquad (2-6b)$$

Another method called PathAct [106] uses the median polish (MP) algorithm [107] to summarise the pathway activity values. MP is an exploratory data analysis for extracting both row-wide and column-wide trends from a two-dimensional matrix. MP is an iterative procedure that consists of the following four steps: (i) calculating the median values of each row, (ii) subtracting the median values from each row, (iii) calculating the median values of each column, and (iv) subtracting the median values from each column. These steps are repeated using the residual matrix, and the median vectors for rows and columns are accumulated at each iteration. This procedure iterates until the reduction of the sum of absolute residual is less than a specified value or the maximum limit of iteration is exceeded. The MP method has been used for several bioinformatics tools, including the robust multi-array average (RMA) method [108], one of the most well-known normalization methods for DNA microarray data.

After the pre-defined number of iterations that take turns subtracting the median value of the rows and columns, the input matrix, $PWM_{s,m}$, is converged at a certain state by PathAct as shown in the following formula:

$$PWM_{s,m} = g_m + a_s + R_{s,m} \qquad (2-7)$$

where g_m is the gene effect that reflects a degree of bias, such as the hybridisation efficiency of each gene, a_s is the individual effect, which is regarded as the pathway activity values for each sample (individual), and $R_{s,m}$ is the residual matrix.

Some methods do not only use the gene expression values, but other information, e.g. the pathway topology [109] and post-translational control of signal transduction [110]. PROGENy [110] is a representative method. They pointed out that inferring the signalling activity of pathways from gene expression disregards the effect of post-translational modifications. Therefore, they used a large compendium of publicly available perturbation experiments to yield a common core of pathway-responsive genes and linear regression models were used to fit genes responsive to all pathways. However, PROGENy is available for 11 signalling pathways because of the availability of perturbation experiments.

2.4.4 An advanced Projection-based pathway activity inference method

The baseline projection-based pathway activity inference method, PCA, projects the samples into one dimension with the largest variance. A more complex and informational method, called Pathifier [111], takes a further step of PCA. Pathifier uses the principal curve to measure the deviation of a sample from the behaviour of the control sample group [112]. More specifically, in the d_p dimensional space, where d_p are the member genes of a particular pathway, the entire set of samples forms a cloud of points. Then, the nonlinear "principal curve" is calculated to capture the variation of this cloud. Each sample is projected onto this curve. The pathway activity value is defined as the distance $D_p(s)$, measured along the curve, of the projection of sample s from the projection of the other samples.

In the cloud of all samples, the normal (healthy) samples are defined as a reference set, and the centroid of the reference set is defined as a reference point. The reference set concentrates on one side of the curve due to the high similarity, and they have large differences from disease samples. Therefore, the reference set defines the curve's direction by ensuring the reference set is closer to the beginning of the curve. Then, the pathway activity score is defined as the distance along the curve between the projection point of the sample and the reference point.

To improve the accuracy of the estimated principle curve, Pathifier performs pre-processing on the input matrix PWM. Genes with higher variance over all samples are selected and normalised with mean and standard deviation. The reason for the gene selection is that many of the genes in the pathway might be highly correlated, i.e. conveying the same information. In contrast, more important information might reside in a single gene in the pathway.

2.4.5 Optimisation-based pathway activity inference method

The methods above are either relying on the statistical summation of the expression data or applying the decomposition approaches to the expression data to infer the pathway activity. Although nearly all the methods include some pre-processing operations on the expression data before the pathway activity calculation, these approaches still rely on the quality of the data. Moreover, their calculation processes are in a "black box" mode that can only provide the conclusions on the pathway-level but makes it hard to provide more gene-level details, which is more beneficial for diagnosis or prognosis value.

A mathematical programming model named Differential Gene Signatures (DIGS) is proposed by [113], which defines the pathway activity inference as a mixed integer linear programming (MILP) problem. This method is now described in detail.

First the indices, sets and parameters associated with the DIGS model are listed below:

Indices

s Sample ($s = 1, 2, \ldots, S$)

m Gene ($m = 1, 2, \ldots, M$)

c,k Class or phenotype ($c = 1, 2, \ldots, C$)

cs Class label for sample s

Parameters

A_{sm} Expression level of gene m on sample s

G_{sm} standardised gene expression profile

ε A small positive number

U A large positive number

NoG Number of member genes allowed to have non-zero weight in building pathway activity for each pathway, a user-specific value

Binary Variables

L_m 1 if effect of gene m on pathway activity inference is positive; 0 if negative effect

E_s 1 if pathway activity of sample s falls within the range of its class; 0 otherwise

Y_{kc} 1 if upper bound of pathway activity range for class k is lower than lower bound of that for class c; 0 otherwise

W_m 1 if gene m is active in pathway activity inference (have non-zero weight); 0 otherwise

Positive Variables

rp_m Positive influence of gene m towards pathway activity inference

rn_m Negative influence of gene m towards pathway activity inference

Unrestricted Variables

pa_s Pathway activity of sample s

LO_c Lower bound of range of class c on pathway activity

UP_c Upper bound of range of class c on pathway activity

Next, the objective function and the constraints that form the model are introduced. Two sets of positive variables rp_m and rn_m are introduced, quantifying the positive and negative weights of gene m towards pathway activity inference. For sample s, pathway activity, pa_s, is defined as the summation of the standardised gene expression values, G_{sm} multiplied by the gene weight (rp_m -rn_m) overall member genes:

$$pa_s = \sum_{m}^{M} G_{sm}(rp_m - rn_m) \qquad \forall s \qquad (2-8a)$$

where M is the total number of member genes for the pathway and S is the total number of samples. Both positive weights rp_m and negative weights rn_m of gene m are defined as positive continuous variables; their values are determined by the optimisation model. Using a set of binary variables, L_m, equations (2-8b) and (2-8c) below ensure that for each gene m at most one of rp_m and rn_m can take positive values:

$$rp_m \leq L_m \qquad \forall m \qquad (2-8b)$$

$$rn_m \leq (1 - L_m) \qquad \forall m \qquad (2-8c)$$

where $L_m = 1$, rp_m can take any value between 0 and 1 while rn_m is forced to be equal to 0; otherwise, when $L_m = 0$, rp_m is forced to be equal to 0 while rn_m can be between 0 and 1. In either case, both rp_m and rn_m can be equal to 0, which means this particular gene has zero weight in inferring pathway activity. Overall, a gene can have positive, negative or zero weight towards the composite feature construction. For normalization purposes, the summation of absolute gene weights should be equal to one:

$$\sum_{m}^{M} rp_m + rn_m = 1 \qquad (2-8d)$$

A set of binary variables, W_m, is introduced to the model to indicate whether a member gene m is active, i.e. having non-zero weights in constructing pathway activity or not:

$$rp_m + rn_m \leq W_m \qquad (2-8e)$$

If W_m takes the value of 0, then both positive weight (rp_m) and negative weight (rn_m) of gene m are forced to be equal to 0, while when W_m is equal to 1, gene m is allowed to take any weight (rp_m-rn_m) between -1 and 1. The following equation restricts the maximum number of genes allowed to have non-zero weights to a manually specified value (NoG):

$$\sum_{m=1}^{M} W_m \leq NoG \qquad (2-8f)$$

In the case where NoG is equal to or larger than the number of member genes, the constraint is redundant as all the member genes will be allowed to take any weight (rp_m-rn_m).

For each class c, two continuous variables have been introduced as LO_c and UP_c, denoting the lower and upper bound of the range of pathway activity for phenotype c. In addition, a set of binary variables, E_s, have been introduced together with the following constraints:

$$0 \leq pa_s - LO_c + U(1 - E_s) \forall s\, c_s \qquad (2-8g)$$

$$pa_s - UP_c - U(1 - E_s) \leq 0 \forall s\, c_s \qquad (2-8h)$$

where c s is the phenotype for sample s, and U is an arbitrarily large positive number. On the constructed pathway activity, ranges of different classes are not allowed to overlap. A set of binary variables, Y_{kc} , have been introduced. The additional two sets of constraints have been introduced to guarantee the non-overlapping condition:

$$UP_k + \varepsilon \leq LO_c + U(1 - Y_{kc}) \qquad \forall k < c \qquad (2-8i)$$

$$UP_c + \varepsilon \leq LO_k + UY_{kc} \qquad \forall k < c \qquad (2-8j)$$

where ε is an arbitrarily small positive number ensuring that pair-wise classes do not share a border. Equations (2-8i) and (2-8j) are generated for each pair of classes. The objective of the optimisation problem is to infer the pathway activity such that it is as discriminative as possible, i.e. as many samples as possible can fall within the range of its corresponding classes ($E_s = 1$). In other words, the objective function is to minimise the number of misclassified samples:

$$Min\ z = \sum_{s=1}^{S}(1 - E_s) \qquad (2-8k)$$

The resulting mathematical programming-based formulation for inferring pathway activity is summarised below:

Objective function: (2-8k)

Subject to:

Pathway activity definition (2-8a)

Positive and negative gene effect constraints (2-8b) (2-8c)

Normalization constraint (2-8d)

Restriction of the number of active genes (2-8e) (2-8f)

Pathway activity enclosing constraints (2-8g) (2-8h)

Non-overlapping constraints for ranges of different classes (2-8i) (2-8j)

The superiority of this model is reflected in deciding the gene weights. Most methods mentioned above take the gene weights as a priori data, e.g. average weights or using pre-calculated t-scores as the weights. In DIGS, the gene weights that the optimisation model decides, make the constructed pathway activity distinguish samples of different classes optimally. Furthermore, the mathematical framework of this method offers the user the ability to explicitly constrain the maximum number of constituent genes (parameter NoG) that contribute to pathway activity inference.

The uniqueness of the DIGS model is also reflected by the gene weights decided by the model. Most methods use gene expression values to calculate the pathway activity values. For example, using mean and median values to summarise the expression values, using projection approaches to transform the expression values or using regression models to fit the expression values. However, the gene weights produced by the DIGS model can be re-used on the other data. Therefore, the dataset can be split into training samples and testing samples. After the DIGS model is trained in the training samples, the pathway gene weights are obtained and can be used on the testing samples.

The flexibility in defining the weights of the member genes that participate in the pathway activity calculation favours the investigation inside an individual pathway. For each pathway, the optimisation model selects a group of genes that got the weights through the NoG parameter. This set of genes is considered the important genes for this pathway that best benefit the separation of the phenotypes. Secondly, the assigned weights to the genes provide the ranking for the selected important genes. The gene that is given higher weight indicates higher importance. Therefore, a novel contribution of the DIGS model is that it provides insights for identifying the pathway-specific and important genes, rather than providing only the important pathways towards phenotype classification. The DIGS model will be explored into further method development in Chapters 3 and 4.

2.4.6 Summary

In this section, a review of the most well-known pathway activity inference methods has been presented. This area of research has attracted a large amount of interest. As a result, a complete review of all existing pathway activity inference methods is beyond the scope of this thesis. Here the aim was to illustrate the main approaches taken to solve the problem of summarising the gene expression values on the pathway level and in particular, to describe the pioneering methods that have underpinned much of the method development. There are many other methods than those described in this section, and new methods always appear. However, well-established methods, such as Mean-based, PCA-based and GSEA-based methods, are still commonly used in practice and are often used to benchmark new methods.

When choosing a pathway activity inference method, the following two conditions need to be considered jointly. First, the existing methods can be divided into two categories according to whether the sample labels are already known, i.e. supervised and unsupervised methods. In the application scenario of finding the significant pathways, the samples are well-labeled to apply statistical tests or train classifiers to identify the

Table 2.3 List of Pathway activity inference methods

Method	Supervised or Unsupervised	Binary or Multiclass	Scoring type	Reference
MEAN	Unsupervised	Multiclass	Arithmetic	[53]
MEAN top 50%	Supervised	Binary	Arithmetic	[82]
CORGs	Supervised	Binary	Arithmetic	[96]
PCA	Unsupervised	Multiclass	Projection	[97, 98, 43]
PLSR	Supervised	Binary	Projection	[98]
ASSESS	Supervised	Binary	Enrichment	[99]
GSVA	Unsupervised	Multiclass	Enrichment	[100]
LLR	Supervised	Binary	Arithmetic	[105]
Pathifier	Supervised	Multiclass	Projection	[111]
DIGS	Supervised	Multiclass	Optimisation	[113]

differential expression pathways. In contrast, in the diagnosis scenario, the sample class is unknown and relies on the pathway activity results to help with diagnosis.

Second, the existing methods can be divided in two categories according to whether the method suits multi-class sample labels or only binary labels. As pathway analysis inspires the pathway activity inference, most pathway activity inference methods consider only the “healthy” and “disease” conditions that follow pathway analysis approaches. Moreover, the involvement of t-score for finding the DEGenes in the pathway activity inference procedure limits their suitability to the multi-class classification problem. As a result, most current supervised methods are only suited for binary sample labels. A summary of the pathway activity inference methods mentioned in this section is given in Table 2.3.

Whether the method is supervised or unsupervised, its advantage is ultimately determined by the classification performance. However, it should be noted that whether the method suits the multi-class problem requires more attention than the binary problem. This is because when studying complex diseases, one is often faced with multi-class problems. Overall, pathway activity inference is considered a more sensitive and in-depth way to analyse and understand the massive and noisy gene expression data. More and more methods are being proposed to provide an accurate and widely applicable result.

2.5 Conclusions

This chapter has presented the necessary preliminaries for the work described in this thesis. First are the main properties of three high-throughput profiling techniques, their technical principles, applications, advantages and drawbacks. With the advent of these techniques, interpreting these large volume and noisy datasets is still challenging, which requires joint consideration of various properties. After a short introduction of the intuitive analysis approaches for the profiling data, identifying the differently expressed genes directly from the data, the focus turned to combining the pathway information as a-priori knowledge for the gene expression profiling analysis. Rather than focusing on the gene-level differences, the pathway analysis approaches look for the differentially expressed pathways that can be easily interpreted with cellular functions. Finally, the pathway activity inference framework was introduced as a more advanced concept with better biological interpretability and flexibility than the pathway analysis approaches. The pathway activity inference methods conclude the expression values on the pathway-level for individual samples, which allows sample classification and prediction and is expected to be more reasonable and accurate. All of the above form the basis for subsequent work.

The pathway activity inference methods discussed in Section 2.4 have dealt with the problem: summarising the gene expression values of the pathway member genes into a single value to represent the activity level of the pathway for a particular sample. This problem statement represents the core issue of this thesis. This thesis continues from the optimisation-based model for solving this problem. The first stage discusses the fitness of the optimisation-based model on the more complex high-throughput profiling technologies and the performance on sample prediction and pathway or gene signatures identification. In the next stage, the model is further improved to obtain better computation efficiency and performance on evaluation metrics.

Chapter 3

Pathway Activity Inference for Acute Ischemic Stroke Microarray Data

Pathway activity inference has been widely used as a powerful framework to reveal the underlying significant biomarkers of complex diseases. Ischemic stroke ranks second after heart disease as a cause of disability in high-income countries and as a cause of death worldwide. Identifying the pathway biomarkers of ischemic stroke can help diagnose stroke from non-stroke cases, as well as advance the understanding of the underlying mechanisms of the disease. In this chapter, the mathematical programming optimisation model called DIGS is applied to build a phenotype classification and pathway inference model using stroke gene expression profile data. The DIGS model is specifically designed for pathway activity inference towards supervised multi-class disease classification and showed great performance among pathway activity inference methods in the original work [113].

However, an area rarely discussed is calculating pathway activity on multiple datasets of the same disease. The sample size in a single dataset can be limited for diseases where samples are difficult to obtain. Therefore, it is necessary to combine multiple data sets to increase the sample size. In extension to the original work, this chapter demonstrates good performance of using combining multiple datasets for pathway activity inference. The results show that the highest accuracy of the prediction on determining stroke or non-stroke samples from the combined dataset reaches 84.4%, which is much better than the prediction accuracy produced by currently found stroke gene biomarkers. Stroke-related significant pathways are also produced by the DIGS model in this chapter.

3.1 Introduction

Acute ischemic stroke (AIS) is a dangerous disease worldwide, which has multiple complications and is hard to cure [114]. According to reports [115], stroke is still the second leading cause of death and the third leading cause of disability after years of clinical treatment and relevant research analyses. Also, it is well known that the economic costs of treatment and post-stroke care of stroke patients are substantial. Therefore, looking for an effective way to diagnose or pathogenesis of AIS is important for both scientific analysis and clinical practice [116, 117].

Microarray technology has become a popular methodology in deriving a comprehensive view from gene expression data for certain conditions. Based on the development of the microarray technology, several researchers have identified molecular biomarkers from AIS blood samples [118–120]. However, most of these efficient - albeit simple - biomarker deriving approaches focused on independent genes and adopt basic statistical approaches. Therefore, they were suffering from low prediction accuracy and difficulty in biological interpretation. Following the principle that genes do not act in isolation but work in concert, in recent years independent gene editing therapeutic methods are increasingly replaced by simultaneously considering functional gene groups. Biological pathways are a key type of functional gene sets, which are available from public databases, for example, Reactome [60], Kyoto Encyclopedia of Genes and Genomes (KEGG) [59] and Gene Ontology (GO) [121]. Biological pathways provide the possibility of analysing groups of genes that belongs to same pathways and identifying the target-relevant pathways as biomarkers. A review of the pathway databases is given in Section 2.3.2.

In [113], a novel multi-class disease classification method, Differential Gene Signature (DIGS), which infers pathway activity in a supervised manner, is proposed. DIGS is a MILP mathematical programming formulation that consists of a linear objective function and several linear constraints. The general idea of DIGS is using weighted linear summation of the constitute genes expression values from same pathway as the pathway activity evaluation of that sample, where the weights of constitute genes are decided by the optimisation model so that the constructed pathway activity can optimally distinguish samples from different phenotype. DIGS has been tested on Psoriasis, Breast Cancer, Prostate Cancer and diffuse large B-cell lymphoma (DLBCL), and showed good performance on distinguishing their sub-phenotypes and detecting biomarkers of these diseases [113].

In this chapter, DIGS is applied on three acute ischemic stroke microarray profile datasets, aiming to reach high prediction accuracy between stroke and non-stroke phenotypes and deriving relative functional pathways as biomarkers of AIS. This chapter is structured as follows. In the next section, the acquisition of the three datasets and the approach of integrating them are described in detail. Then, the DIGS method implementation and validation pipeline is illustrated. Finally, the results are evaluated in two aspects, i.e. in terms of the sample classification performance and the relevance of identified significant pathways to the disease.

3.2 Identification of biological pathways for ischemic stroke through mathematical programming optimisation

The work in this chapter has been performed in collaboration with Dr. Konstantinos Theofilatos (KCL Cardiovascular division). The gene expression dataset sources, pre-processing and integration were carried out by Dr. Konstantinos Theofilatos. My contribution to this chapter is the implementation of the pathway activity inference model, the implementation of the validation pipeline, determine the appropriate approaches for the results analysis. Specifically, Section 3.2.2 is contributed by Dr. Konstantinos Theofilatos and the other sections can be ascribed to myself.

3.2.1 Gene expression datasets and KEGG pathway acquisition

Three publicly available gene expression datasets GSE22255 [120], GSE16561 [122] and GSE58294 [123] were obtained from Gene Expression Omnibus (GEO). All the experiments of these three GEO series were conducted on Affymetrix Human Genome U133 Plus 2.0 Array Platform. GSE22255 contains 20 stroke and 20 control (non-stroke) peripheral blood mononuclear cells (PBMCs); GSE16561 contains 39 stroke and 24 control peripheral whole blood samples; GSE58294 contains 23 control whole blood samples and 69 stroke samples. Because the 69 stroke samples of GSE58294 were analysed at three time points: less than 3 hours, 5 hours, and 24 hours following the onset of stroke, only samples collected at 3 hour time points were used in this work. After dataset integration, there are 82 stroke peripheral blood samples from stroke patients and 55 from control patients. The datasets information is summarised in Table 3.1.

Table 3.1 Microarray datasets of acute ischemic stroke

Session ID	Tissue type	No. Control	No. Stroke
GSE22255	PBMCs	20	20
GSE16561	whole blood	24	39
GSE58294	whole blood	11	23
Summary		55	82

Pathway data were acquired from MsigDB KEGG C2 functional gene sets [63], which include 186 curated pathways with in total of 5267 genes.

3.2.2 Dataset integration

Pooling data from different studies meant combining expression arrays derived from whole blood samples vs. those from peripheral blood mononuclear cells (PBMCs). Because PBMC excludes non-nucleate cells, one may expect somewhat different profiles with respect to whole blood, due to differences in cell/tissue type. Consequently, it was necessary to determine whether pooling PBMC and whole blood data may introduce artefacts due to profile differences between the different types of blood samples. This was achieved by comparing the Fold-Change (FC) per-gene in the whole blood-only datasets to the pooled blood/PBMC data using Spearman rank-based correlation analyses.

The datasets were merged by performing a second layer of joint normalization similar to the standard approaches used for qPCR data [124]. Initially, 8 commonly used housekeeping genes (ACTB, B2M, HMBS, HPRT1, RPL13A, SDHA, TBA, YWHAZ) that were expressed at comparable mean levels in both treatment (stroke) and control groups were selected and followed the procedures outlined in [124] to identify the minimal subset of genes that show the most between-experiment variability to use as normalisers. Because all of these genes showed relatively high FC across samples (Log2FC prior to median per-sample normalization was > 0.3 for all the examined housekeeping genes), normalization was performed based on median per-sample expression level rather than rescaling with respect to expression levels of housekeeping genes.

The pooled data were filtered so that only genes with less than 10% missing expression values would be retained for further analysis. For the remaining missing values, the KNN-Impute method [125] was applied to properly impute the missing data with $K = 20$. For the final stage of quality control, outliers were identified with a method based on principal components analysis (PCA) – retaining the principal components that accounted for 90% of covariation, and then applying the local outlier factor (LOF) approach [126] to cluster samples and detect outliers as the un-clustered samples. This analysis indicated that less than 5% of the data were marked as outliers, thereby passing a predefined threshold of fewer than 10% outliers for a dataset to be considered valid for further analysis.

In conclusion, dataset integration includes three steps: (1) determining whether the different types of blood samples would introduce artefacts, (2) using the standard normalization approaches to merge the datasets, (3) filtering genes with too many missing values. The integrated dataset contains 137 samples (82 stroke samples and 55 control samples) and 13243 genes. This merged dataset is used as a whole in the following pathway activity inference procedures.

3.2.3 Pathway activity inference using the DIGS model

An overview of the computational procedure of DIGS, which is developed for pathway-based sample phenotype classification, and mathematical details are illustrated in Chapter 2 Section 2.4.5. In this section, an overview of the model is summarised in Figure 3.1. Pathway-specific gene expression matrices $PWM_{s,m}$ consist of the standardised gene expression values of sample, s, across pathway member gene, m. The number of gene expression matrices is equal to the number of pathways, which in this work is 186 KEGG pathways. For each $PWM_{s,m}$ matrix, DIGS assigns a pathway activity score to each sample and pathway activity range for each phenotype through the optimization model. The objective function of the optimization model can be described as ensuring that as many pathway activity scores as possible fall within the corresponding phenotype range. The following paragraphs review the details of the DIGS model.

The mathematical programming based formulation of DIGS contains 10 constraints and an objective function (see Chapter 2.4.5). The first part of the formulations define how pathway activity values are calculated. For each sample s, pathway activity pa_s is defined as the summation of the gene expression values $PWM_{s,m}$ multiplied by a gene weight $(rp_m - rn_m)$, where rp_m represents the positive weight of gene m and rn_m represents the negative weight. Then, the first limitation (set by the Equations 2-8b and 2-8c) is applied on a pair of positive variables, rp_m and rn_m. For each m, neither rp_m nor rn_m can take positive value, which means one of them is forced to be zero. A binary variable L_m is introduced to construct these constraints.

The second limitation (set by the Equation 2-8d) restricts the number of genes that can be “active” genes among all member genes in a pathway. Active genes are defined

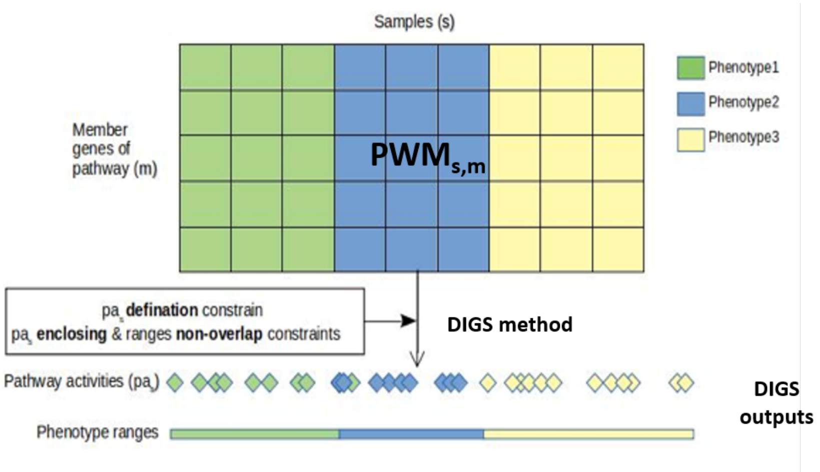

Samples (s)
Member genes of pathway (m)
PWM$_{s,m}$
Phenotype1
Phenotype2
Phenotype3
pa$_s$ defination constrain
pa$_s$ enclosing & ranges non-overlap constraints
DIGS method
Pathway activities (pa$_s$)
DIGS outputs
Phenotype ranges

Fig. 3.1 Schematic flow chart of DIGS pathway activity inference method. $PWM_{s,m}$ serve as the input data. The DIGS model produces the pathway activity values for samples s, and the ranges for phenotype c.

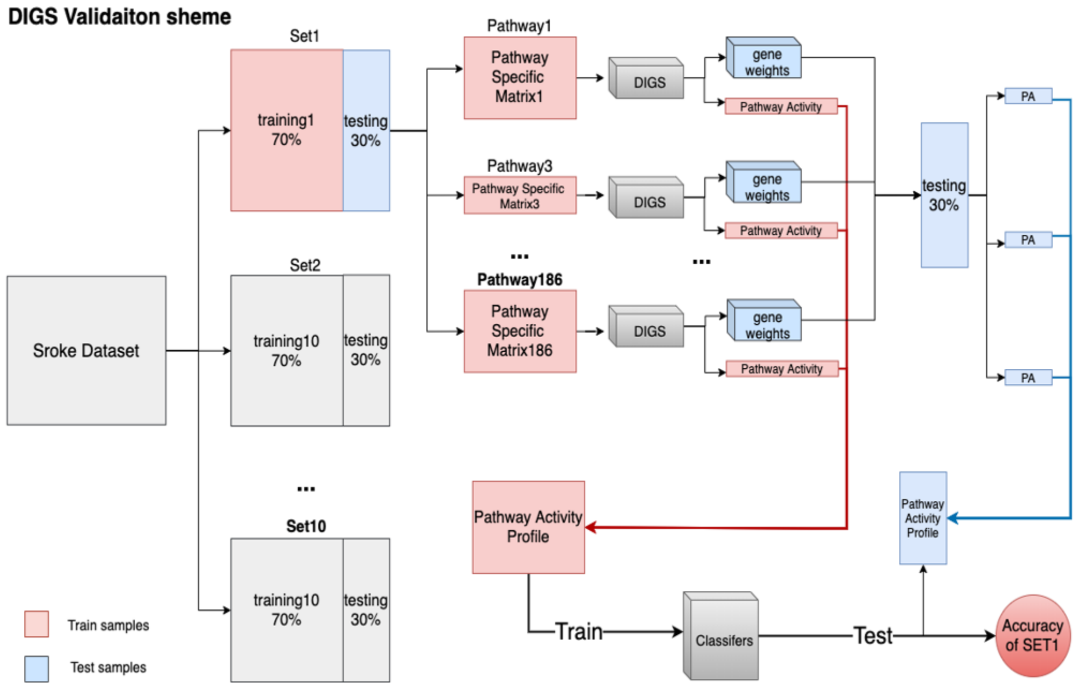


Fig. 3.2 Overview of the DIGS validation scheme using microarray gene expression profile for phenotype classification.

as genes that gain non-zero weights, while keeping the remaining non-active genes' weights equal to zero. A binary variable W_m is introduced to indicate whether a gene m is "active". When W_m takes the value of 1, gene m is an active gene and its weight $(rp_m - rn_m)$ would take a value between -1 and 1. Also, a user-defined variable NoG (used in the Equation 2-8e) is introduced to restrict the maximum number of active genes. For normalization purposes, the summation of absolute gene weights is equal to 1, as defined as in Equation (2-8f).

The following part of formulations set restrictions on the phenotype ranges. According to the Equations (2-8g) and (2-8h), the range for a phenotype c is defined by two continuous variables, lower bound LO_c and upper bound UP_c. A binary variable E_s is adopted to indicate whether the pathway activity value of a sample s falls within the LO_c and UP_c of its corresponded phenotype.

The last two constraints (Equations 2-8i and 2-8j) are introduced to guarantee that, for each pair of phenotypes (c, k), the ranges do not overlap. Here a binary variable Y_kc ensures this requirement. When $Y_kc = 1$, the relationship between c and k is $k < c$ and UP_k is lower than the LO_c; when $Y_kc = 0$ the condition is reversed ($c < k$, $UP_c < LO_k$). Also, ε, an arbitrarily small positive number, is designed to ensure pair-wise classes do not share borders. Finally, the objective function (Equation 2-8j) of this optimization problem can be defined by minimising the number of miss-classified samples $(1 - E_s)$.

In conclusion, all constraints of DIGS model are linear with a linear objective function and multiple binary or continuous variables. Therefore, DIGS is a mixed integer linear programming (MILP) model that can be solved to reach global optimum with standard algorithms.

3.2.4 Implementation and validation scheme

The implementation procedure of pathway activity inference and pathway activity-based disease classification is illustrated in Figure 3.2. To gain robust and objective prediction results, all samples of the integrated stroke dataset are randomly split into 70% training set and 30% testing set. This procedure is repeated 10 times to produce 10 training/testing sets. During the model training process, testing samples are always blind to the training procedure to ensure no information leakage. For every training gene expression matrix, gene sets from KEGG pathways are integrated with the gene expression matrix to create individual pathway-specific expression matrices ($PWM_{s,m}$). Therefore, 1860 pathway-specific expression matrices (10 training sets * 186 pathways) are generated and the DIGS models are trained on them. The $PWM_{s,m}$ serve as the input of the model and the output of the model consists of the weights of the NoG genes.

From model solving results, the composite features, which summarise the expression patterns of member genes m into a new feature PA_s, are constructed by the summation of the gene expression values multiplied by the gene weights. This newly constructed feature is the pathway activity values for this $PWM_{s,m}$. After obtaining the pathway activity vectors from all pathway specific expression matrices independently, pathway activity matrices (PA_{sp}) are formed by the ensemble of corresponded 186 pathway activity vectors for each training set.

For each training or testing set and each pathway, gene weights (rp_m and rn_m) are extracted when solving DIGS models on training samples for the pathway activity calculation on testing samples. Therefore, the testing samples pathway activity values are calculated using the same gene weights as the training samples. Similarly, pathway activity vectors from testing samples are combined into testing pathway activity matrix.

The DIGS model is implemented in the General Algebraic Modelling System (GAMS) [127] and solved using the CPLEX. CPLEX is a high-performance mathematical programming solver for linear programming, mixed integer programming, and quadratic programming. It was the first commercial linear optimiser on the market to be written in the C programming language [128].

According to the sensitivity analysis for parameter NoG in the original publication, the DIGS model is robust with respect to NoG in range of 5 to 20. Here, NoG is set to 10, which means the model allows 10 genes per pathway to participate in pathway activity calculation. The optimal gap (optcr) is set as 0.00 so as the attempt obtain globally optimal solutions. The computation time limit for solving a DIGS model is set as 200 seconds. The solving status of DIGS models includes both optimal solution and feasible solution obtained after time limitation.

The pathway activity matrices are then used to train machine learning classifiers. Overall, the procedure showed in Figure 3.2 produces 10 sets of training and testing pathway activity matrices corresponding to original random division of training and testing sets on the stroke dataset. To avoid contingency, six commonly used machine learning classifiers, K-nearest-neighbour (KNN), Logistic Regression (LR), Random Forest, Neural Network (NN), Naive Bayes and Support Vector Machine (SVM), are employed in this study. The Python package sklearn version 0.22.1 is used to implement these classifiers and produce the classification accuracies towards the sample phenotypes. These six classifiers were trained on 10 training pathway activity matrices and tested on testing pathway activity matrices with the following parameters: for NN, hidden

Table 3.2 Averaged prediction accuracy on testing sets

Classifier	ACC	AUC
5-NN	0.844	0.895
Logistic Regression	0.833	0.932
Random Forest	0.836	0.923
Neural Network	0.840	0.917
SVM	0.830	0.927
Naive Bayes	0.843	0.907

layer is 2, learning rate is 0.1, training time is 10000; for KNN, the number of clusters is 5. For the other classifiers, default settings are retained.

3.3 Results

In this section, the performance of DIGS is demonstrated though a comparative analysis with other pathway activity inference methods. Also, the DIGS selected important pathways are compared with the DEGenes and the superiority of using genes within pathways as a whole for phenotypic classification was confirmed.

3.3.1 Evaluation of Prediction performance

To rigorously evaluate the prediction accuracy of various implemented classification approaches, prediction accuracy are averaged over 10 training sets and testing sets for each classifier. Due to the inherent problem of unbalanced numbers of samples across two phenotypes (Table 3.1), both classification accuracy (ACC) and area under curve (AUC) [76] are used as metrics to measure the prediction accuracy of a classification model. ACC is defined as the fraction value of the number of correctly classified samples divided by the number of all samples. Higher AUC values correspond to better prediction performance, with AUC of 1 indicating perfect prediction, 0.5 indicating performance equal to random. Overall, Table 3.2 shows the averaged ACC and AUC values across 10 testing sets produced by six different classifiers on stroke dataset.

Generally, all six classifiers have produced relatively high accuracies (83.5%) and high AUC scores (91.5%) towards the classification on stroke and control samples. Among six classification methods, 5-Nearest-Neighbours reached the highest prediction accuracy (84.4%) and Logistic Regression got the best AUC score (93.2%).

To further validate the superiority of DIGS, other three widely used pathway activity inference methods were implemented on stroke dataset for comparison. In overview, these three methods are: (i) Mean method [53] that takes the mean gene expression values of all genes within a pathway for each sample; (ii) the second method, referred as Median method [129], has exactly same procedure as Mean method, replacing the mean expression values across genes with the median expression values across genes; and (iii) the third method is called PCA method, built by [96], which uses the first principal component of the pathway specific expression matrix as representation of pathway activity scores for each sample. These three methods are the most widely used baseline pathway activity inference methods, their details can be found in Section 2.4. To make the prediction results comparable, the validation scheme for these other three pathway activity inference method is same as DIGS and the exact same ten training and testing sets used for DIGS were applied to Mean, Median and PCA method. The arrangement of the resulting 10 pathway activity matrices and same classifier training procedures are adopted. The output prediction accuracy for these three methods are also averaged across 10 testing sets and all results are plotted in Figure 3.3.

Specifically, the DIGS model can be trained and tested (described in Section 2.4), whereas the other methods cannot. The implementations of the other methods are separate for the training and testing samples. In detail, after the dataset has been split, the Mean, Median and PCA methods first calculate the pathway activity values for the training samples and then for the testing samples. Therefore, the training and test pathway activity profiles are the same for all methods. This pipeline is used in the following chapters for methods comparison.

In Figure 3.3, x-axis is labelled with six classification approaches and y-axis represents the prediction accuracy values for each pathway activity inference methods across each classifier. From the figure, it is obvious that DIGS-based classification approach achieves higher classification rates than other pathway inference methods. The performance of Mean and Median methods is similar (accuracies range from 60% to 80%), and PCA methods gets the lowest prediction accuracies (range from 50% to 60%). It can be concluded that DIGS is the most effective method among four methods for deriving pathway activity scores towards classification of stroke phenotypes.

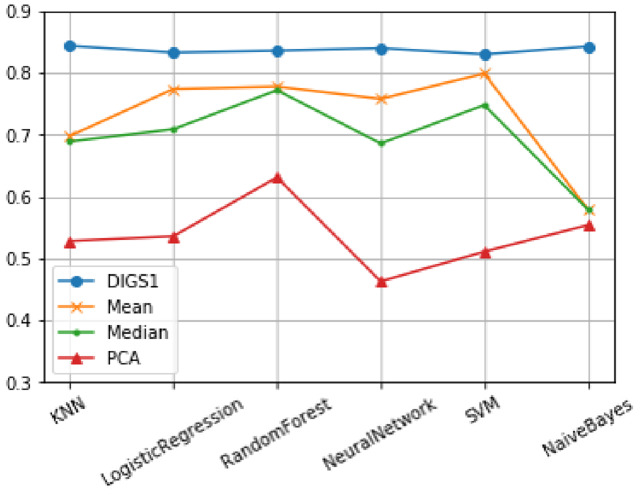


Fig. 3.3 Classification accuracy comparison of four pathway activity inference methods.

3.3.2 AIS Relevant Pathway Identification

3.3.2.1 Pathway relevance ranking

Not only promising classification rates can be achieved by DIGS model, but also several pathways are identified as significant, which may indicate pathway biomarkers. To rank the pathways, Point-biserial correlation coefficient ranking method from Python SciPy package (Version 1.3.0) is employed in this work. Point-biserial Correlation Coefficient is a statistical measure of the relationship between a binary variable and a continuous variable, and it is mathematically equivalent to the Pearson correlation. In Machine Learning, Point-biserial correlation can be used to calculate the similarity between features and categories. In other words, it is adopted to judge whether the extracted features are positively correlated, negatively correlated, or not correlated with the corresponding categories. The range of Point-biserial Coefficient is [-1, 1] and the greater the absolute value is, the stronger the correlation is. In this section, this intuitive way of the pathway selection is used, and in the next two chapters, further approaches for pathway selection are introduced.

To gain an ultimate pathway activity value for each pair of sample and pathway, 10 pathway activity matrices (combination of the corresponded training samples and testing samples of each training/testing set) were merged into one pathway expression matrix by averaging. Then, for each pathway, the Point-biserial Correlation Coefficient is calculated using the pathway activity vector across all samples and the phenotype vector that consists of sample phenotypes (stroke or control). The absolute value of

Table 3.3 Average prediction accuracy on testing sets

Pathway Name	Coef.[1]
CELL_CYCLE	0.782
B_CELL_RECEPTOR_SIGNALING_PATHWAY	0.744
UBIQUITIN_MEDIATED_PROTEOLYSIS	0.743
LEISHMANIA_INFECTION	0.739
PYRIMIDINE_METABOLISM	0.739
SPLICEOSOME	0.737
CELL_ADHESION_MOLECULES_CAMS	0.736
RNA_DEGRADATION	0.727
TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY	0.725
EPITHELIAL_CELL_SIGNALING_IN_HELICOBACTER_PYLORI_INFECTION	0.721

[1]Point-biserial correlation coefficient

the calculated correlation metrics for each pathway were ranked in descending order and top 10 pathways were selected as the most discriminative pathways.

The selected ten discriminative pathways are listed in Table 3.3. Apart from pathways that have obvious links to cancer pathways, for example the well-known signalling pathway (B cell receptor signalling pathway), and pathways involved in cell metabolism procedures and genetic information processing (Cell cycle, Pyrimidine metabolism, RNA degradation and Spliceosome), I note a piece of research concluding that the immunoblockade or genetic deletion of adhesion molecules showed to reduce infarction volume, edema, behavioural deficits and/or mortality in different animal models of ischemic stroke [130]. Also, in [131] indicates that the ubiquitin-mediated proteolysis pathway, especially TRAF6, may be the most vital molecules among TLR downstream pathways in incidences of ischemic stroke, which proves that two of our top ten pathways (Ubiquitin mediated proteolysis and Toll-like receptor signalling pathway) are strongly related to the diagnosis of ischemic stroke.

In conclusion, these DIGS top ranked pathways are highly related to AIS, and thereby can be treated as pathway biomarkers.

3.3.2.2 Top ranked pathway evaluation

To intuitively display the performance of significant pathways, two heat-maps were drawn using gene expression data and pathway activity data for significant pathways. In Figure 3.4, rows are gene names and pathway names for the upper plot and lower plot, respectively. Columns are sample phenotypes. Samples are hierarchically clustered based on similarity. In horizontal colour bar, different colour (green and blue) represents "stoke" and "control" respectively. From the comparison between these two plots, most of the samples belonging to the same phenotype are indeed assigned to the same cluster with the significant pathways data found by DIGS.

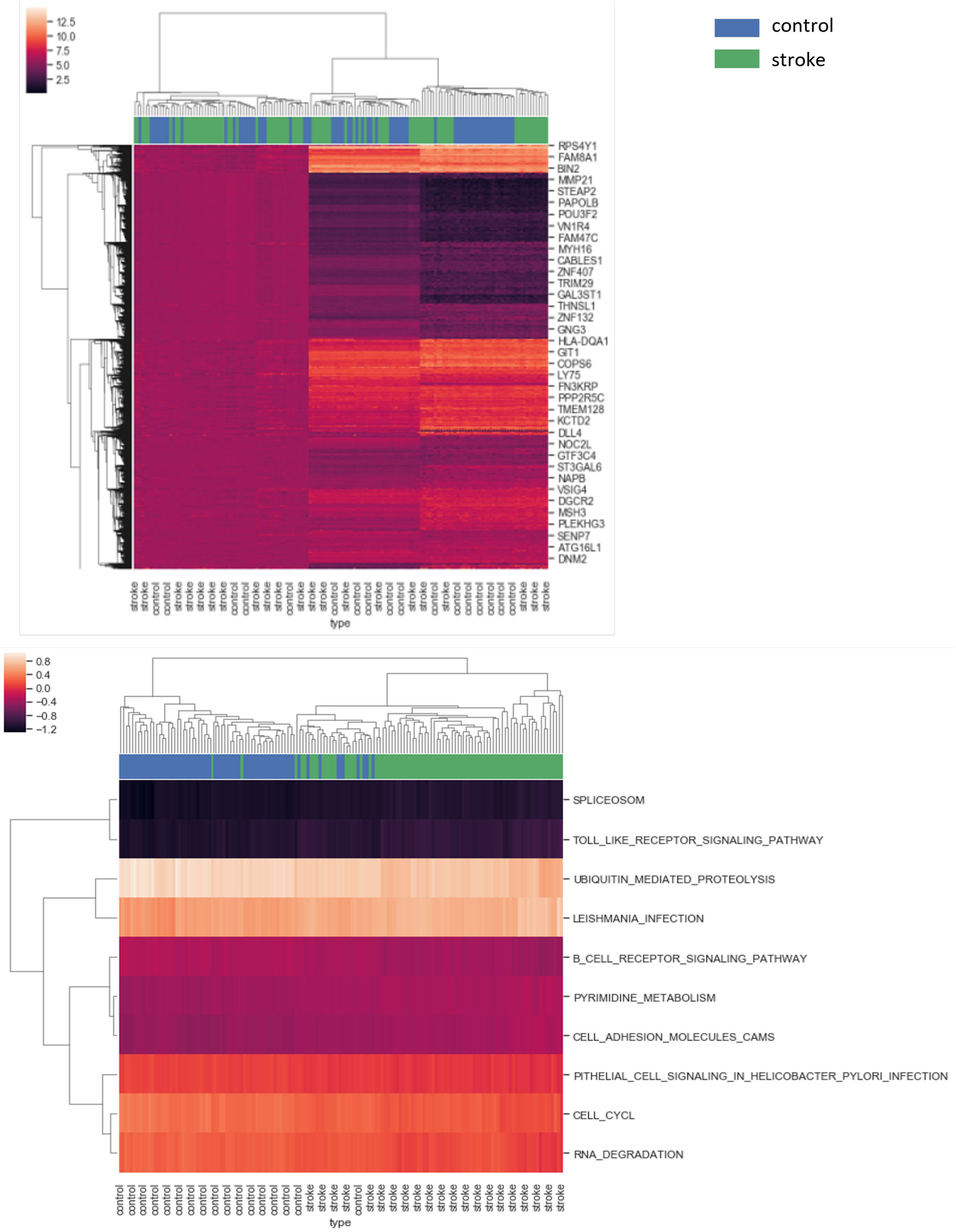


Fig. 3.4 Hierarchical clustering for gene expression profiles and significant pathway activity in the stroke dataset.

To further explain to what extent the significance has been reached by the top ranked pathways, 10 box plots in Figure 3.5 show the distribution of pathway activity values of the two different sample phenotypes for each significant pathway. The left box plot in each subplot represents the activity values of stoke samples and the right box plot represents the control samples. It is obvious that the ranges between upper quartiles and lower quartiles of the two phenotypes are well-separated in each subplot. Even using only one of these top discriminate pathways to classify the phenotypes, the accuracy would be at least 75%. Similar analysis was done in [132], where the 10 most statistically significant differentially expressed genes, illustrating the extent of Fold Change and the separation of median expression levels between the two phenotypes, were selected from the stroke dataset and their logarithmic relative expression values were plotted. The box plots of these significant differentially expressed genes in [132] presents the distribution of the expression ranges of "stroke" and "control" phenotypes. However, Figure 3.5 shows better distribution differences between the two phenotypes. The ranges between upper quartiles and lower quartiles of the two phenotypes in the box plots of the selected differential expressed genes overlap in [132], which indicates that the prediction accuracy of the significant genes is lower than the top ranked pathways. Also, according to their study, the prediction accuracy produced by log FC expression level significance genes (557 genes) is less than 71%, whereas the classification accuracy of DIGS reaches 84.4%.

It can be concluded that the pathways found by DIGS model have stronger discriminate power in separating stroke samples from control samples than the differentially expressed biomarkers genes. Also, these results have proven that using genes belonging to the same functional group is better than using single genes independently for phenotype classification in gene expression profiles .

3.4 Conclusions

This chapter applies optimisation-based pathway activity inference, DIGS, on a merged ischemic stroke gene expression profile for the purpose of inferring pathway activity values for classifying ischemic stroke samples from the healthy. The classification results using six machine learning classifiers show promising accuracy rates (83.5%) and relatively high AUC values (91.5%) over 10 repetitions of training and testing data splits. To the authors' best knowledge, the classification accuracy reached by DIGS is higher than most current gene-based stroke phenotype prediction methods. Also, the biological pathways identified by DIGS model are proved relevant to the cause of ischemic stroke and can be regarded as pathway biomarkers for ischemic stroke.

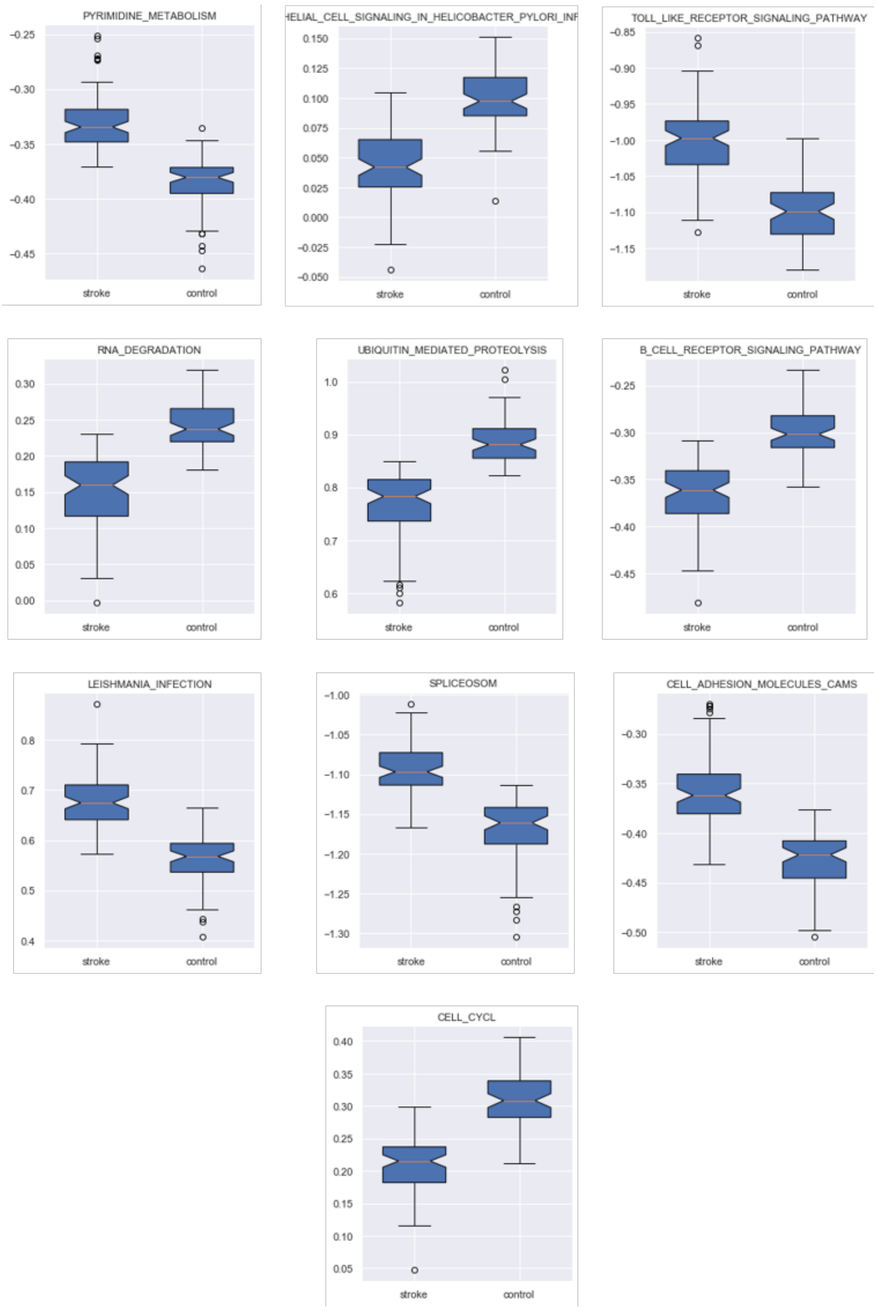


Fig. 3.5 Boxplots: pathway activity Distributions of different phenotypes for the top pathways.Y-axis represents pathway activity values.

The main contribution of this chapter is to provide a validation of some of the theories from Chapter 2. First, this chapter proves that incorporating biological pathways with gene expression profiles improves the classification accuracy compared with former gene-based stroke studies. The advancement of pathway activity inference is approved by the higher classification accuracy achieved by using pathway activities in this work compared with using differentially expression genes. Second, the advancement by the optimization-based methods DIGS is proven by comparing the classification accuracy with other baseline pathway activity inference methods. DIGS provides a more flexible way for inferring stroke-related biomarkers. By modifying the number of active genes (parameter NoG) inside DIGS model, different levels of biological information can be extracted from the running outputs. The final contribution is that this chapter verifies the DIGS model is highly adaptable to different datasets. This is due to the fact that optimization models are not dependent to the quality of the data itself as arithmetic-based or projection-based methods. Therefore, in the next chapter, the adaptability of DIGS model on other datasets by RNA-Seq is explored.

Chapter 4

Pathway Activity Inference Applied in Cancer RNA Sequencing Data

Computational methods for aggregating gene-level into pathway-level data have become a mature approach for many applications, such as gene signature identification and drug discovery. With the emergence of high-throughput gene profiling technologies, the volume and complexity of the transcriptomics data keep increasing. When analysis them, gene-based dimension reduction and analysis methods suffer from low prediction accuracy and difficulty in biological interpretation [7]. There is a need to address such limitations with a well-built method that is not only adapted to multiple kinds of high-throughput data but also one that is robust in accuracy and biological interpretability.

As an optimisation-based, supervised pathway activity inference method, DIGS has shown its ability to classify multi-class complex diseases using the microarray datasets in the previous work [113]. This chapter applies DIGS to an RNA-Seq dataset of colorectal cancer. The samples are labelled with four molecular subtypes, CMS1, CMS2, CMS3 and CMS4. By comparing DIGS with baseline and newly proposed pathway activity inference methods, I illustrate that the high prediction accuracy is also achieved in the RNA-Seq profiling technology. Additionally, DIGS enables the identification of colorectal cancer-related gene signatures. Further, the inferred pathway activity values are robust against noisy data and enhance the survival prediction accuracy.

The application on the RNA-Seq dataset proves that DIGS is not limited to microarray data. The high prediction accuracy toward multi-class classification problems was maintained in RNA-Seq data, and the literature supported the significant pathways detected by DIGS. Also, DIGS kept its discriminative power on the highly noisy data in robustness evaluation, and for the survival analysis, DIGS showed good prognostic power. Overall, this chapter highlights the potential of applying DIGS on other kinds of high-throughput data with high-quality classification performance and biological interpretability, contributing to one of the main goals of this thesis.

4.1 Introduction and related work

Identifying gene signatures is important for complex disease classification, diagnosis, prognosis, and prediction of treatment outcome [133–135]. The appearance of the various high-throughput technologies, which allow the analysis for complete genomic or molecular profiles, has yielded a good base to tackle disease classifications and offer a stable platform to correlate patient phenotypes to latent genetic alterations [136]. Genome-wide analyses have frequently focused on identifying differentially expressed genes across phenotypes that can accurately classify patients into their corresponding disease status [137].

To date, various multivariate gene signatures have been successfully proposed in the literature, which outperform traditional pathological variables [138–140]. However, the inherent "large-p small-n" nature of high-throughput genomic data - whereby the number of samples is usually two orders of magnitudes smaller than the number of genes in a single transcriptomic profile - causes lack of robustness in determining gene signatures. It is suggested that several thousand patient samples may be needed to achieve a desired level of robustness when deriving gene signatures [7]. Consequently, reducing gene dimensionality can pave the way for more precise analyses.

To search for more robust and biologically relevant biomarkers, various methods recently have employed analysing expression patterns of gene sets by incorporating a-priori biological knowledge, typically in the form of expertly curated biological pathways [7, 141]. It is increasingly recognised that complex diseases are associated with the deregulation of pathways [142–145]. The fast-accumulating knowledge of pathways is being deposited in public databases, including Reactome [60], the Kyoto Encyclopedia of Genes and Genomes (KEGG) [146], and Gene Ontology (GO) [121]. Integrating pathways with context-specific genomic data can significantly facilitate the identification of disease-perturbed pathways and the construction of pathway signatures for more robust classification [96, 147].

Therefore, pathway activity inferencing methods are popular for improving disease classification accuracy. Pathway activity is defined as inferring one score for each sample-pathway pair based on context-specific profiling data. As mentioned in Chapter 3, baseline pathway activity inference methods use mean or median values [53] or the first principal component [148–150, 111] as the pathway activity scores. A more advanced method is proposed by [105], which presents a probabilistic inference method that computes pathway activity as the aggregate of the log-likelihood ratios between two phenotypes over all the pathway constituent genes. However, a common disadvantage of these methods is the difficulty in interpreting pathway activity values biologically, as they focus solely on capturing differences in gene expression values between sample groups. Also, robustness of the results is limited by the data quality.

More advanced methods were proposed to bring more interpretability. [96] proposed a heuristics-based method that ranks pathway constituent genes according to their discriminative power. It then searches for a small set of highly differentially expressed genes and uses their averaged expression values to produce pathway activity. [111] proposed a method (referred to as Pathifier) which uses a principal curve to infer the pathway activity. The principal curve is calculated with all samples. Then the pathway activity of a disease sample is calculated as its distance to the control samples along the curve. According to [52], which reviews 13 pathway activity inference methods, Pathifier was proven to be the method that achieved the best performance across multiple evaluation criteria. Although these methods focus more on biological interpretability, they are either suitable for only binary problems or relying on capturing the variances of the data. The adaptability of these methods to multiple types of profiling data is still being determined.

In comparison to these methods, DIGS has the best adaptability to the different types of genomic data. DIGS calculates the pathway activity as a weighted linear combination of pathway constituent genes. In DIGS, weights of the genes are optimised to maximise the discriminative power of the disease outcomes, in contrast to other inference methods where gene weights are given a-priori (usually equal weights are assumed, e.g. [53, 151, 37]). The supervised nature of DIGS makes it more suitable in disease classification framework than unsupervised methods that summarise variance in the genomic data [53, 148, 152]. Lastly, DIGS is not limited to binary classification but is applicable to multi-phenotype disease classification problems.

In conclusion, this chapter presents the next application of the optimisation-based pathway activity inference method, DIGS. In this chapter, the adaptability of the DIGS model is tested on RNA-Seq dataset. RNA-Seq is a high-throughput sequencing technology that has shown strong potential to replace microarrays for genome-wide transcriptome analysis [153–157] and has been adopted by several pathway analysis or pathway activity inference methods [158–161, 100].

The dataset used in this chapter is the colorectal cancer RNA-Seq dataset obtained from the TCGA database[162]. The molecular subtypes in colorectal cancer were used as sample labels. The performance of DIGS was compared with three pathway activity inference methods that are either widely used or have outstanding performance and results illustrate promising results regarding molecular subtype classification, survival analysis and robustness.

4.2 Optimisation-based pathway activity scoring applied on cancer phenotype prediction and survival prognosis

In this work, the general-purpose solution algorithm CPLEX in the General Algebraic Modelling System (GAMS) [127] is used to solve the optimisation models. The locally optimal solutions are identified within a user-specified time limit (200 seconds). The maximum number of the genes assigned non-zero weights was set as 10 [113].

4.2.1 Data Preparation

Raw counts of colorectal cancer RNA-Seq dataset (COAD) was downloaded from The Cancer Genome Atlas Program (TCGA) database [162]. Labelling the samples in COAD was done in two ways. The first is to classify the samples as tumour and normal, which creates a two-class classification problem; the second is labelling the samples with the molecular subtypes of colorectal cancer. For decades, determining the molecular subtype of colorectal cancer has been difficult. Many researchers [163–166, 136, 138] have attempted to build classification models to classify the molecular subtypes of colorectal cancer tissue biopsy samples. CMS1, CMS2, CMS3 and CMS4 are the newest molecular subtypes defined by recent research, which combines the previous six techniques [167]. This chapter uses this newest labelling strategy as the molecular subtypes for the samples in the COAD dataset to address the multi-class classification problem.

For pre-processing, the raw read counts of RNA-Seq data is normalised by upper quartile FPKM (FPKM-UQ) [168]. Then, genes with over 30% zero expression values across the sample cohort are removed. Table 5.1 shows the summary of the TCGA-COAD dataset. The “Tumour or Normal label” indicates the number of samples used for

Table 4.1 Averaged prediction accuracy on testing sets

Dataset	Disease	Tumour or Normal Label	Molecular Subtype label
COAD	Colon Cancer	Tumor: 480 Normal: 41	CMS1: 85 CMS2: 165 CMS3: 58 CMS4: 120

binary classification, and the "Molecular subtype label" indicates the number of samples used for multi-class classification.

To rigorously evaluate the pathway activity inference method, I applied the stratified 10-fold cross-validation strategy [169] three times on the gene expression profile so that 30 training data and testing data pairs were created. Further, to avoid the effect of the highly unbalanced tumour-normal labels, SMOTE (described in Section 2.3.2.2) [57] is applied to the expression profile data to balance the number of normal and tumour samples. It should be noted that SMOTE analysis was conducted on the training sets and did not affect the testing sets to avoid biased results. As a result, 30 training and testing datasets for the multi-class problem and 30 training and testing sets for the 2-class problem were created.

Biological pathways were obtained through the KEGG API [170]. After eliminating pathways with fewer than three member genes, 279 Homo sapiens pathways with a total of 6761 unique genes were collected.

4.2.2 Validation Pipeline

Figure 4.1 depicts the pipeline for pathway inference using RNA-Seq data. There are two procedures. The first is applying pathway activity inference methods on the RNA-Seq gene expression data to construct the pathway activity profile. The gene expression profile $exp_(s, g)$, where s represents the sample, and g represents the gene name, was split into pathway expression matrices. Each pathway expression matrix consists of samples as rows and member genes of a particular pathway as columns. Therefore, there are 279 pathway expression matrices (279 KEGG pathways). The DIGS model is applied to each pathway expression matrix in the next step to produce the pathway activity values PA_s. Finally, combining all the 279 PA_s vectors forms the pathway activity profile $pa_(s, p)$ that consists of samples across pathways. With three times 10-fold CV, 30 train $PA_(s, p)$ and 30 test $PA_(s, p)$ were created.

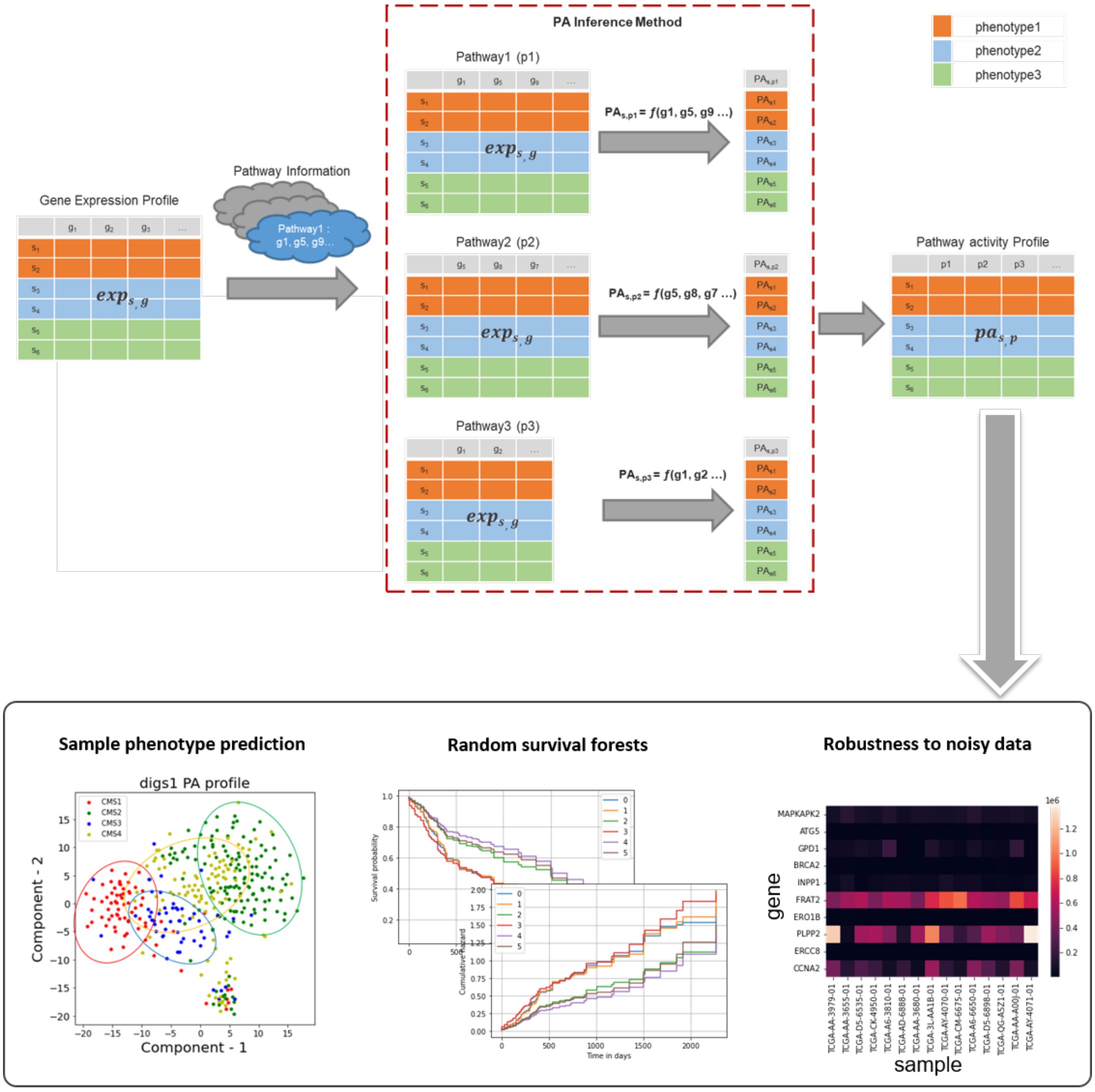


Fig. 4.1 WorkFlow for Pathway Activity Inference and Analysis. The normalised RNA-Seq gene expression profile is split into pathway specific expression matrix, where pathway activity inference methods were applied on. Combining the pathway activity vectors of pathways to form the pathway activity profile for further analysis. Pathway activity analysis consists of three directions, the first is to predict the disease phenotypes using Machine Learning Classifier and evaluate the prediction performance; the second is to fit the survival model using pathway activity profile together with dataset survival information; the third is to test the robustness of the pathway activity inference methods to noisy gene expression data.

The next part of this pipeline depicts three means of performance evaluation. The first is sample prediction accuracy, which is the traditional approach. Sample phenotype prediction uses a machine learning classifier to train the $pa_(s,p)$ matrix to predict sample subtypes. The following two parts explore the survival prediction performance and robustness against noisy data. Survival analysis fitted the $pa_(s,p)$ to random survival forest regression model to predict the survival possibility of samples; Robustness to noisy data evaluates how the perturbation on $exp_(s,g)$ would affect the classification performance of $pa_(s,p)$. Each experiment was conducted repeatedly on all training and testing dataset pairs. The following sections give detailed descriptions of these three evaluation approaches.

4.2.3 Sample Phenotype Prediction

A common strategy for testing the performance of pathway activity inference methods is to utilise machine learning classifiers to predict the sample labels on pathway activity values. Here I used Random Forest (RF) [171] and K-Nearest-Neighbor (KNN) [75] classifiers to this effect. RF has been well studied in the context of gene expression classifiers as it performs well with highly correlated, high-dimensional data and is less prone to overfitting. To obtain the best classification performance, the parameter n_estimators was tuned from 200 to 2000 using grid search optimisation on training sets. As opposed to RF, KNN stands as a basic and naïve classifier. The parameter k of the KNN algorithm was selected using a trial and error process on the training dataset testing values 3, 5, 10, 20, 30 and 50, and k=30 was selected as the highest-performing classification metric. RF and KNN were implemented using the Python library sklearn 0.24.0 [172]. The performance evaluation of classifier prediction performance on testing data used four metrics: accuracy, F1-score, precision, and recall.

4.2.4 Survival data Analysis

Applying pathway activity values to survival analysis has been popular in several cancer studies. This work adopted survival prediction as a metric for assessing prognosis. As a new composite feature that can summarise gene expression levels, pathway activity values are expected to predict the survival probability better than random data.

The clinical data of the COAD dataset was downloaded together with the raw RNA read counts from the TCGA database. Then the observed samples survival status and survival time were extracted from the clinical data. The alive samples are labeled as "0" and the dead samples are labelled as "1", as describe in Section 2.3.4.2. Considering the inaccuracy caused by the loss of clinical information, I removed samples where the time was less than one year and the status was alive. In remaining samples, "0" samples were randomly selected so as to keep the number of "0" and "1" samples the same. After sample selection, 196 samples were kept for survival analysis. The Survival Random Forest [92] model was implemented using the Python package scikit-survival 0.16.0 [173] to build the survival regression models, and the goodness was measured by concordance index (c-index) [174].

4.2.5 Robustness against noise in profile data

Data from high throughput technologies are susceptible to technical noise and biological variations. Even though it has been proposed that grouping genes into pathways reduces the noisiness of biological data [159], high predictive accuracy on a given dataset is not sufficient to validate the robustness of a method due to the limitations of the chosen dataset [95]. In this work, I assess the robustness of pathway activity inference methods in dealing with fluctuations in gene expression values. Pathway activity inference methods are expected to retain good prediction performance over the same dataset as noise level increases.

Addressing noise in data is described in Figure 4.2. Perturbed genes are randomly selected from the whole genome. The proportion of genes affected is set to 0%, 3%, 10% and 50%. Then, sample order is permuted randomly for the selected genes. In this way, three perturbed expression profiles are generated.

By passing these gene expression datasets into pathway activity inference methods, one original pathway activity profile and three perturbed pathway activity profiles of different noise levels were obtained. Robustness was assessed through 30 training and testing cycles using the sample multi-label prediction accuracy with the KNN classifier (described in Section 4.2.4).

4.2.6 Pathway activity inference methods from literature

In this chapter, comparison will be conducted with both baseline and advanced methods for pathway activity inference. Methods that are suit for multi-class task are required for this chapter. Therefore, according to the summary presented in Table 2.3, the following methods are used. The MEAN method [53] computes the mean expression value across all constituent genes within the pathway as pathway activity. The PCA method [148–150, 152, 95] uses the first principle component as pathway activity. The

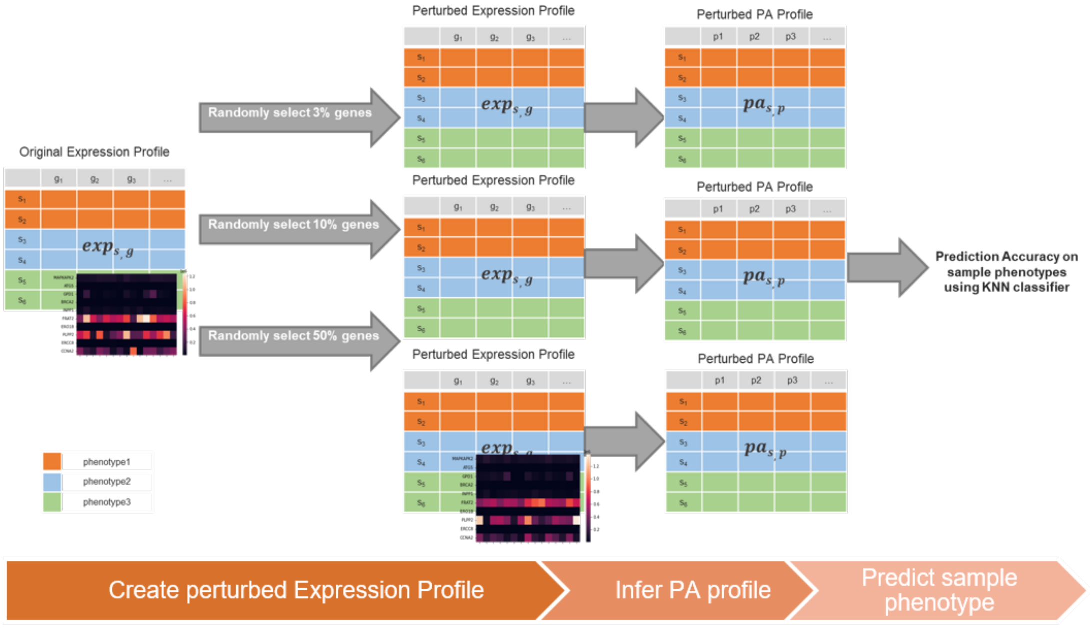

Original Expression Profile
Randomly select 3% genes
Randomly select 10% genes
Randomly select 50% genes
Perturbed Expression Profile
Perturbed PA Profile
$exp_{s,g}$
$pa_{s,p}$
Prediction Accuracy on sample phenotypes using KNN classifier
phenotype1
phenotype2
phenotype3
Create perturbed Expression Profile
Infer PA profile
Predict sample phenotype

Fig. 4.2 Robustness evaluation pipeline. 3%, 10% and 50% of genes are randomly selected without replacement from gene expression profile and the order of gene values is re-combined to form the perturbed expression profiles. Then pathway activity inference methods are applied on the perturbed expression profiles to get the pathway activity profiles. KNN (k = 30) is trained and tested on the pathway activity profiles to produce the prediction accuracy of sample phenotypes.

Pathifier [111] calculates the pathway activity by generating a principal curve. Details of these methods are described in Section 2.4.

In total, four pathway activity inference approaches are used for comparison: DIGS, MEAN, PCA, and Pathifier. The baseline methods are MEAN and PCA and the the more advanced method is Pathifier. Importantly, Pathifier has been proven to outperform most contemporary proposed methods [52], therefore its use in comparative tests is particularly powerful.

4.3 Results

In this chapter, RNA-Seq data is used to test the performance of the DIGS model for pathway activity inference. Sample label prediction accuracy, survival regression fitness, and robustness to noisy data are the metrics considered in the comparison. The results in these three metrics and the biological interpretations of the DIGS model are discussed below.

4.3.1 Two-class and multi-class classification comparison

The DIGS model's performance was compared to other well-established approaches. Comprehensive comparisons of three competing approaches on the TCGA-COAD dataset have been carried out, as stated in the previous section. All four methods' inferred pathway activity profiles were fed into the Random Forest (RF) and K-Nearest Neighbour (KNN) classifiers to calculate prediction accuracy for the 2-class and multi-class problems. The classification evaluation metrics were averaged across 30 testing sets. To achieve an objective evaluation of classification performance, the training process, including inferring pathway activity and training classifiers, is always blind to testing datasets. The outcomes of all techniques and classifiers combinations are shown in Table 4.2 (for multi-class problems) and Table 4.3 (for two-class problems).

In the multi-class scenario, the DIGS model inferring pathway activity produced higher prediction accuracy than other methods, as shown in Table 4.2. With the parameter-tuned classifier RF, DIGS enhanced accuracy by roughly 15% when compared to the two baseline methods, MEAN and PCA. When compared to Pathifier, the accuracy improves much more. As DIGS seeks to infer a pathway activity of optimal discriminative power [113], the sensible performance of DIGS is not a surprise. The simpler KNN classifier produced results that were similar to those of RF. In conclusion, DIGS performed

Table 4.2 Comparisons of multi-class prediction performance among methods.

	Random Forest							
	DIGS2		MEAN		PCA		Pathifier	
	Avg.	Std.	Avg.	Std.	Avg.	Std.	Avg.	Std.
Accuracy	**0.848**	0.071	0.710	0.050	0.700	0.036	0.681	0.064
F1	0.868	0.062	0.743	0.015	0.730	0.030	0.640	0.069
Precision	0.848	0.071	0.710	0.050	0.700	0.036	0.683	0.064
Recall	0.845	0.073	0.690	0.050	0.683	0.039	0.622	0.067
	KNN							
Accuracy	**0.762**	0.075	0.645	0.056	0.678	0.053	0.397	0.013
F1	0.745	0.080	0.699	0.050	0.735	0.051	0.231	0.058
Precision	0.804	0.078	0.654	0.056	0.678	0.053	0.397	0.013
Recall	0.762	0.075	0.626	0.050	0.658	0.050	0.263	0.029

Table 4.3 Comparisons of 2-class prediction performance among methods.

	Random Forest							
	DIGS2		MEAN		PCA		Pathifier	
	Avg.	Std.	Avg.	Std.	Avg.	Std.	Avg.	Std.
Accuracy	0.944	0.032	0.926	0.03	0.906	0.013	**0.973**	0.037
F1	0.95	0.028	0.937	0.023	0.92	0.011	0.966	0.045
Precision	0.969	0.014	0.964	0.008	0.957	0.005	0.975	0.037
Recall	0.944	0.032	0.926	0.03	0.906	0.013	0.973	0.037
	KNN							
Accuracy	0.904	0.035	0.916	0.032	0.886	0.014	**0.942**	0.047
F1	0.935	0.022	0.92	0.025	0.91	0.012	0.95	0.035
Precision	0.949	0.023	0.949	0.013	0.947	0.012	0.96	0.035
Recall	0.934	0.035	0.906	0.026	0.896	0.016	0.953	0.035

the best, MEAN and PCA took second place, while it is noteworthy that Pathifier performance was extremely poor for the multi-class scenario.

For the 2-class problem, the prediction performance of all four methods was similar. Regarding the actual prediction rates, all pathway activity inference methods achieved over 90% prediction accuracy in the 2-class problem. Therefore, in the more challenging multiclass classification problem for colorectal cancer, the superior performance of DIGS (84.8% prediction accuracy) further demonstrates the applicability and efficiency of the DIGS model on molecular subtype classification.

4.3.2 Survival analysis comparison

I examined pathway activity inference methods for predicting sample survival on the COAD dataset. The Random Survival Forest model was fitted to the pathway activity profiles over the 30 training splits of all methods. The predictive performance of those survival regression models was evaluated by calculating the concordance index.

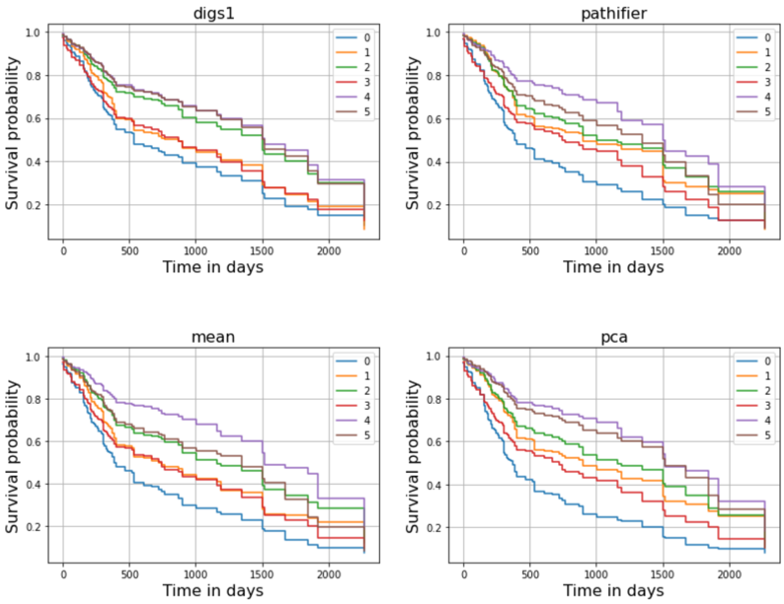
In Figure 4.3, the estimated survival function fitted by the Random Survival Forest of six training samples illustrated that the survival prediction of DIGS is more in line with expectations compared to the other three methods. For DIGS, it is obvious that the survival probability of samples labelled as "0" was higher than samples labelled as "1", and the differences in survival curves between "0" samples and "1" samples are distinct. The concordance index comparison further showed the superiority of the DIGS model.

The c-index distribution comparison between four pathway activity inference methods is shown in Figure 4.4. The c-index values are obtained from the three times 10-fold cross-validation using from the 147 training samples and the 49 testing samples. The dotted lines on the boxes indicates mean values. According to the figure, all methods perform well on training samples, as the mean values of the c-index are all above 0.8. For testing samples, except Pathifier, the mean and median c-index of the other three methods were all above 0.5 (c-index = 0.5 indicates random performance). However, DIGS attains higher mean concordance index values than other methods in both training and testing datasets, which shows that the survival prediction provided by DIGS is better than other methods.

4.3.3 Robustness comparison

I compared the sample phenotype prediction accuracy of the pathway activity inference methods in the presence of various levels of perturbation in transcriptome data. A KNN classifier with k = 30 was used on inferred pathway activity profiles to train and evaluate the models. Because the two-class problem was too basic to demonstrate distinctions, the multi-class problem scenario was used.

Figure 4.5 displays the prediction accuracies of KNN classifier. The values are averaged over 30 testing datasets. According to the figure, DIGS certainly provides the best overall performance. As more noise was introduced to the data, the DIGS model accuracy remained above 0.65. Although PCA and Pathifier performed the best in terms of decline degree (decreased from 0.66 in 0% to 0.60 in 50% and from 0.44 in 0% to 0.37 in 50%, respectively), the comparatively low accuracy in non-perturbed datasets



sample	Duration (days)	Event
0	535	Death
1	535	Death
2	1321	Alive
3	1	Death
4	1003	Alive
5	952	Alive

Fig. 4.3 Kaplan-Meier survival curves generated using the survival probabilities estimated by Random Survival Regression. Confidence interval is sets as 95%. Six CMS1 samples are randomly selected for plot.

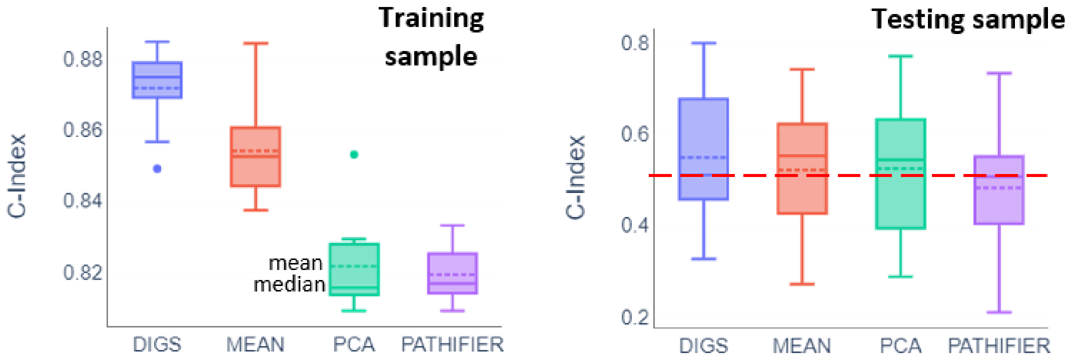


Fig. 4.4 The c-index distribution comparison between four pathway activity inference methods

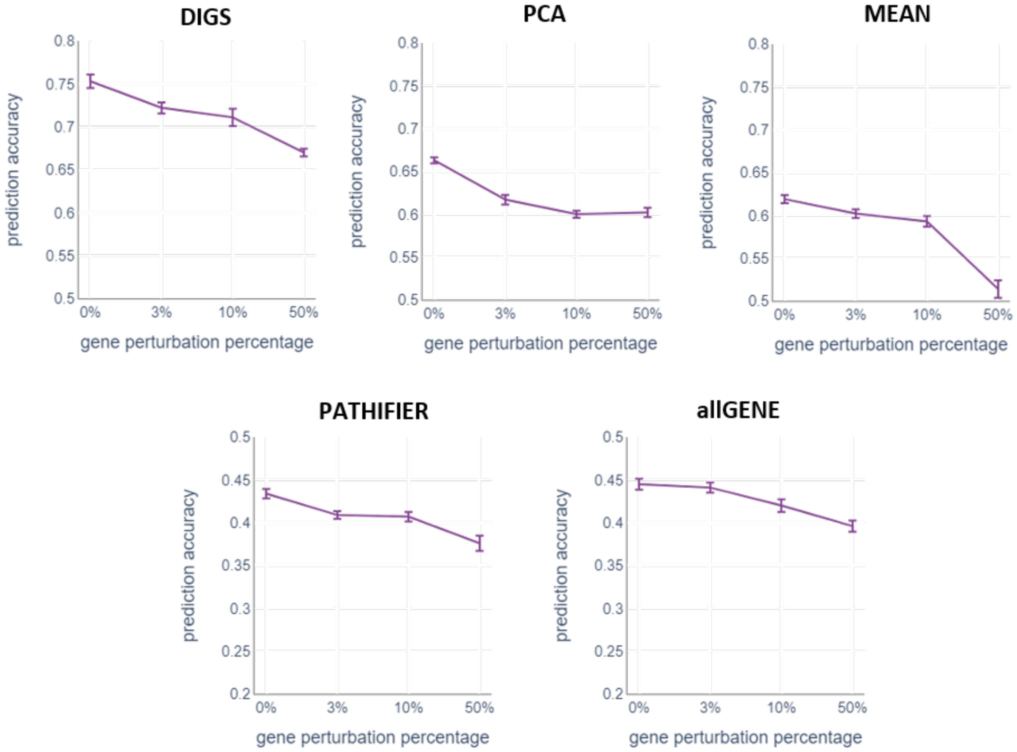


Fig. 4.5 Robustness comparison.

(0% perturbation rate) inversely corroborated the poor power in differentiating sample labels of these two methods.

4.3.4 Overall performance of DIGS model

In this work, the performance of the DIGS model is evaluated with three metrics and is compared against three other pathway activity inference methods. The performance over four methods on the three metrics is summarised in the radar chart (Figure 4.6).

Each vertex counter-clockwise from 12 o'clock in the radar chart indicates sample phenotype prediction accuracy, robustness against 50% perturbed genes and survival regression evaluated by c-index. The Multi-class Prediction vertex uses the RF prediction accuracies of molecular subtypes (multi-class problem). Robustness (50%) vertex is the KNN prediction accuracies of pathway activity values produced on 50% perturbed transcriptome dataset (under multi-class scenario). The Survival Regression vertex is the c-index value of testing samples. Overall, DIGS outperforms other methods on all vertices, especially in Multi-class Prediction. It is reasonable that DIGS performed well in predicting multiple sample classes as it is embedded in its objective function. The

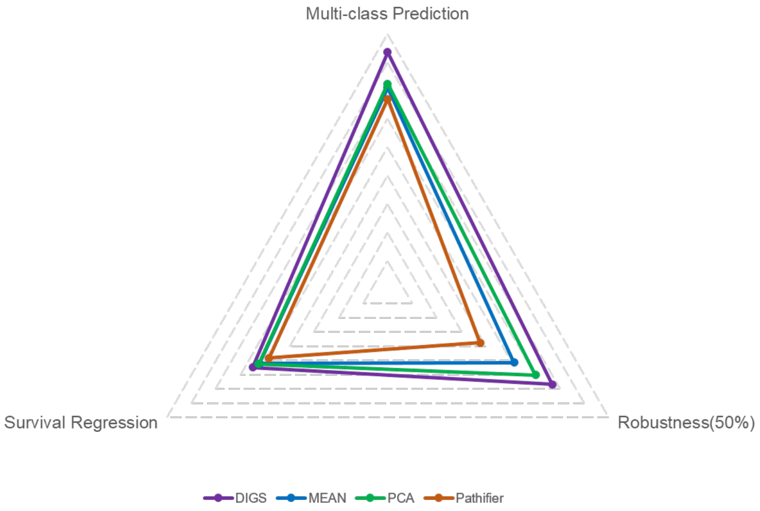


Fig. 4.6 Radar Chart for comparison between three pathway activity inference methods. Each vertex counter-clockwise from 12 o'clock of the radar chart indicates COAD molecular subtype prediction accuracy, robustness against 50% perturbed genes and survival regression evaluated by c-index.

tSNE dimension reduction visualisation also supports this finding (Figure 4.7), as the separation of the clusters of different sample subtypes was much clearer in DIGS than in other methods. These results show the adaptability of DIGS on RNA-Seq data. The exceptional performance of DIGS on the other two dimensions suggests the stability of this supervised mathematical optimisation model.

4.3.5 Pathway ranking

Most of the previous research used statistical methods to identify relevant pathways, such as ranking the p-value and information gain index [159, 161, 100, 175]. I present a self-based strategy for identifying significant pathways using the outputs of the DIGS model.

The DIGS model provides two kinds of outputs, the weights of pathway member genes (used to calculate pathway activity values) and the class intervals for each subtype. This procedure is illustrated in Figure 4.8. With each pathway expression matrix serving as input, the DIGS model is trained on training samples to get the member gene weights and subtype ranges. Then these results are applied to the testing samples. The gene weights are used to calculate pathway activity values, and subtype ranges

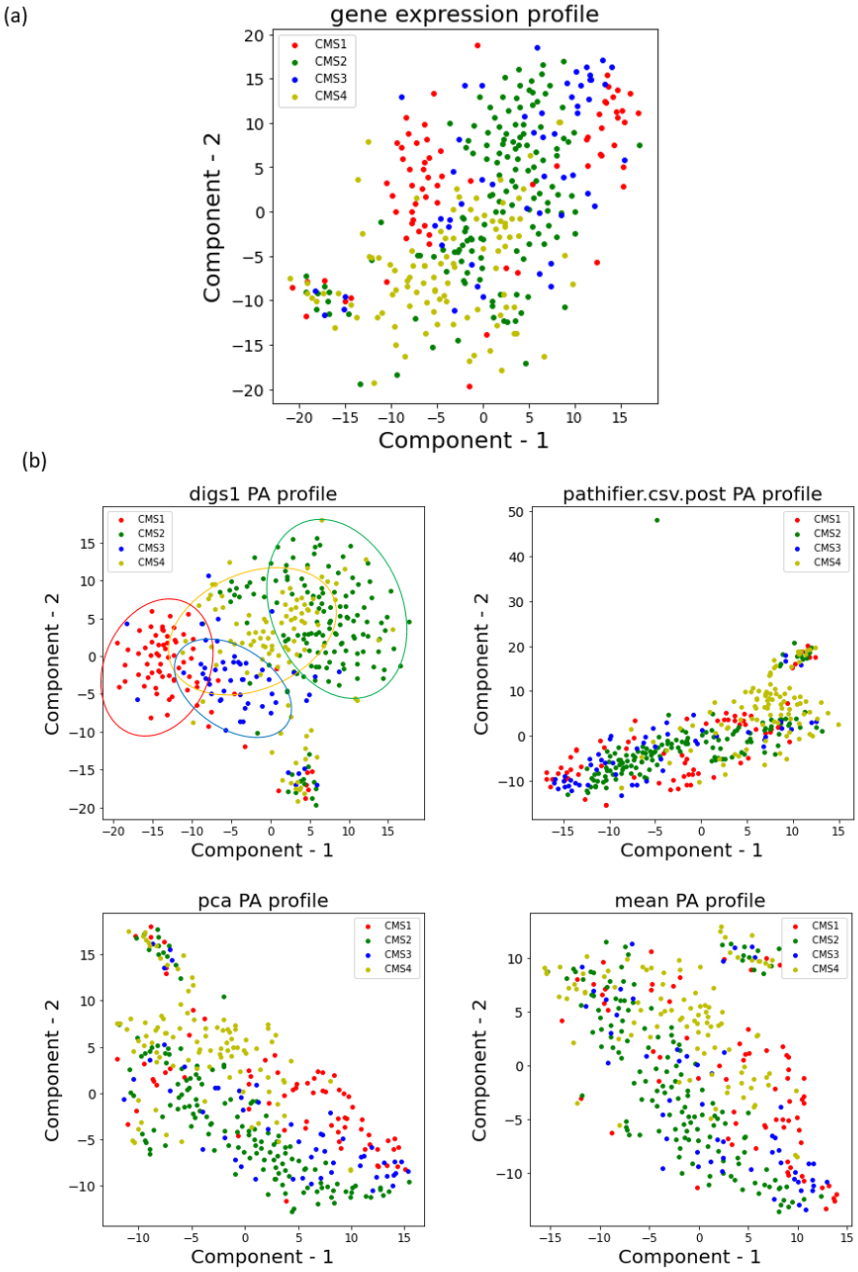


Fig. 4.7 Dimension reduction by TNSE. (a) scatter plot for the first two principal components of the RNA Seq data of COAD. (b) consists of the scatter plots for the first two principal components of pathway activity profiles of the four pathway activity inference methods.

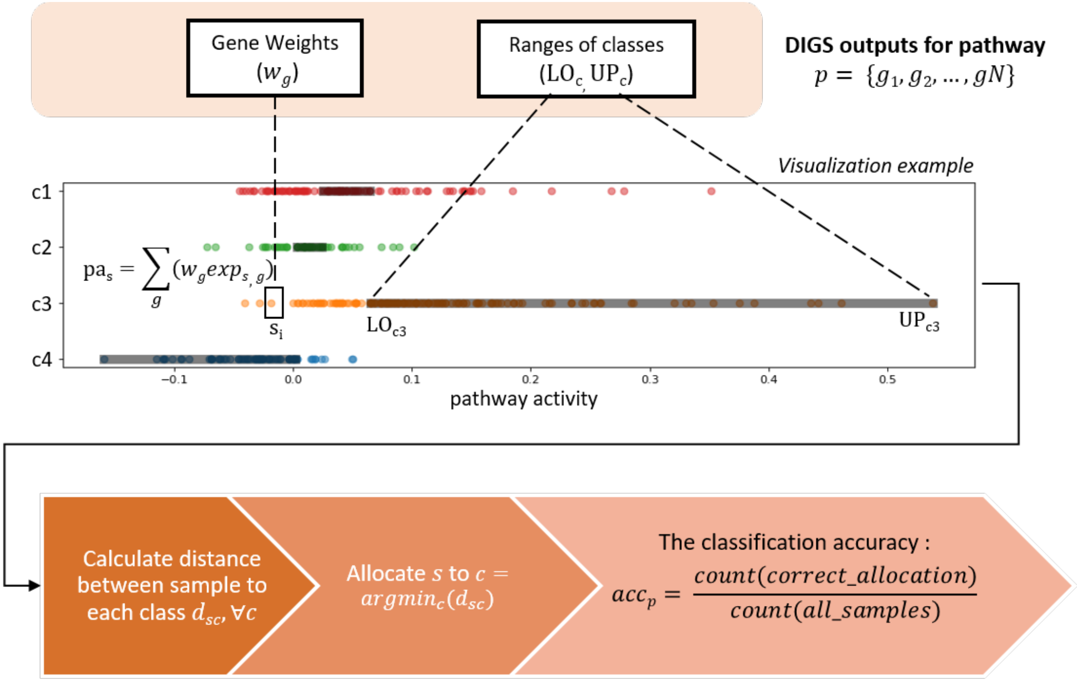

Gene Weights
(w_g)
Ranges of classes
(LO_c, UP_c)
DIGS outputs for pathway
$p = \{g_1, g_2, \ldots, gN\}$
Visualization example
c1
c2
c3
c4
$pa_s = \sum_g (w_g exp_{s,g})$
s_i
LO_{c3}
UP_{c3}
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
pathway activity
Calculate distance between sample to each class $d_{sc}, \forall c$
Allocate s to $c = argmin_c(d_{sc})$
The classification accuracy :
$acc_p = \frac{count(correct\_allocation)}{count(all\_samples)}$

Fig. 4.8 Sample allocation using DIGS outputs. For each individual pathway, DIGS returns gene weights W_g and the class ranges LO_c, UP_c. The pathway activity values pa_s (dots on the strip plot) are calculated using $W_g.d_{sc}$ is defined as the distance of pa_s to each class range. The sample s is allocated to c where d_{sc} is minimum. The classification accuracy for this specific pathway acc_p is calculated by how many samples are correctly allocated.

are used for allocating the pathway activity values. Specifically, testing samples can be allocated to their nearest subtype according to the distance between the samples to the class boundaries. There are three possible scenarios for d_{sc} calculation. When pa_s is inside the $[LO_c, UP_c]$, the $d_{sc} = 0$. When pa_s on the left of the $[LO_c, UP_c]$, the $d_{sc} = LO_c - pa_s$. When pa_s on the right of the $[LO_c, UP_c]$, the $d_{sc} = pa_s - UP_c$. The allocated class is selected with the minimum d_{sc}. Thus, the prediction accuracy for this pathway acc_p can be calculated as the percentage of samples allocated to their correct class. This percentage is referred to as **Individual Pathway Prediction Accuracy** and will be used throughout the rest of the thesis.

By ranking the acc_p for individual pathways, the top 30 pathways for the COAD dataset are selected and listed in Figure 4.9. The values presented in the heatmap were averaged pathway activity values across the samples belonging to the same molecular subtype. The values were scaled to 0-1 to indicate the pathway distinguishing power among different phenotypes. For example, the Mismatch Repair pathway shows great discriminative power among four phenotypes, while the Hedgehog Signalling pathway is better for distinguishing CMS4 from the other three phenotypes.

4.4 Discussion

4.4.1 DIGS reveals significant disease-relevant pathways

With the pathway ranking approach proposed in Section 4.3.4, the DIGS model provides pathway ranking of all 279 KEGG pathways for colorectal cancer. The Top 70 pathways related to the colorectal cancer molecular subtypes (CMS1 to CMS4) are listed in Appendix A. A large part of these Top 70 pathways was known as deregulation pathways in colorectal cancer or other types of human cancer.

In [136] a set of pathways that are associated with the molecular subtypes of colorectal cancer was reported. They identified these highly relevant pathways through gene set enrichment analysis. For some of these pathways, DIGS also yieds high ranking. For example, Renin-angiotensin System ranks 12, the Mismatch repair ranks 21, and the Hedgehog signalling pathway ranks 26. In [176] KEGG pathway analysis was done with differential expression genes in colorectal cancer. They indicated that the Vitamin Digestion pathway was among the most relevant. This pathway ranks 13 in the DIGS pathway. [177] investigated the involvement of signalling pathways in colorectal cancer and among nine pathways they reported, Hedgehog Signaling pathway and the Hippo Signaling pathway rank 26 and 22, respectively, in our analysis. The complete

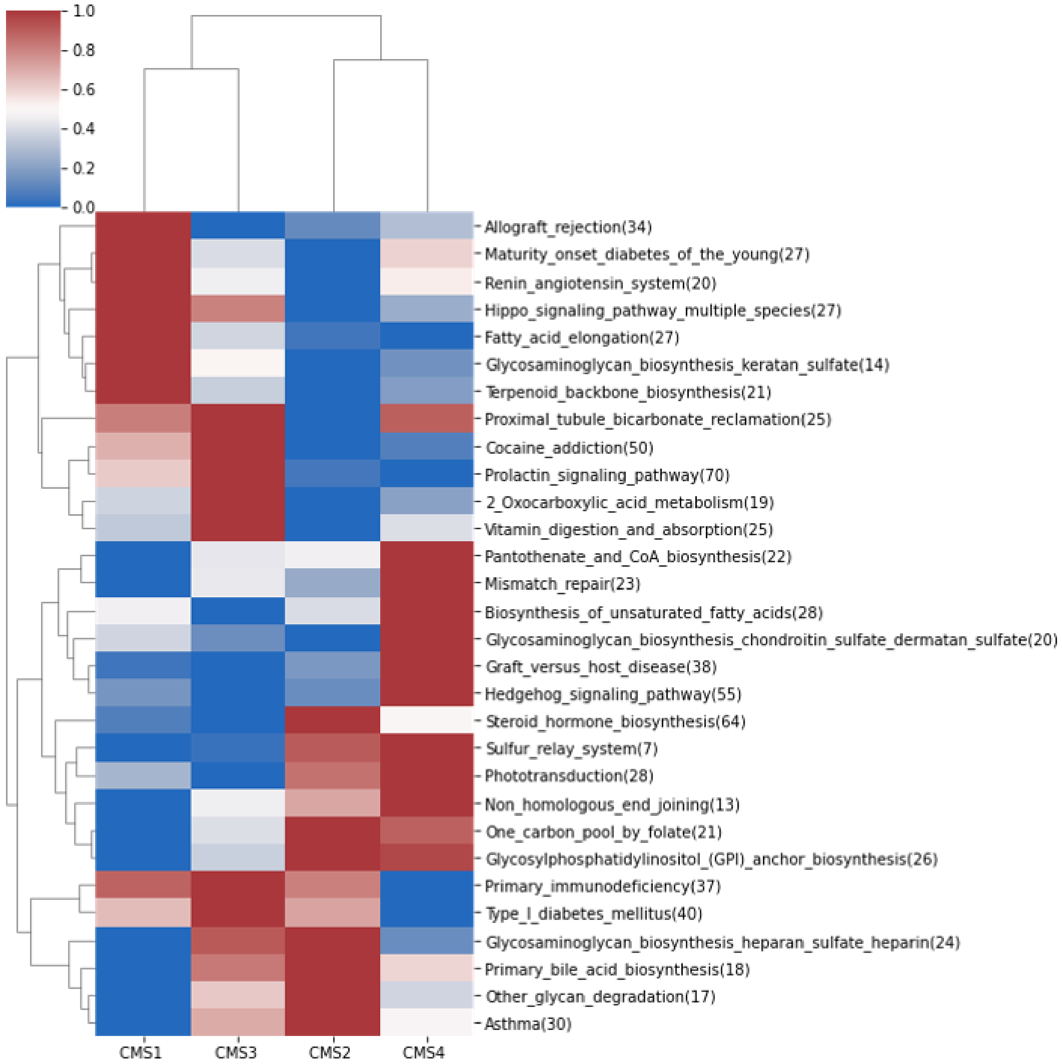


Fig. 4.9 Clustering Map of the averaged pathway activities (Top 30 pathways). Top pathways were selected by ranking the individual pathway prediction accuracy (acc_p). For each phenotype, pathway activities are averaged across constituent samples. For each pathway, the four averaged values are scaled to 0-1 to illustrate the distinguishing power between different colorectal cancer subtypes. The numbers in brackets indicates the number of member genes.

information on relevant literature pathways with DIGS ranking is provided in Appendix A.

As not all pathway member genes can obtain non-zero weights in the DIGS model, the weighted genes (no more than 10 genes) can be seen as important genes that contribute more to pathway activity levels and can thus be used as biomarkers to estimate the activity of the corresponding pathways. I further ranked the genes into three levels: high, medium and low, as follows. For each pathway, the DIGS model was trained 30 times and, as each training gives different weights to genes, the participating genes were first ranked inside each training, then the ranks of 30 pieces of training were combined to produce the final participating genes and their ranks for the corresponding pathways. Figure 4.10 shows the gene weights for top10 pathways. Figure 4.11 shows the KEGG pathway map with colors for important genes of the representative pathways: Hippo Signalling pathway. The KEGG maps of the other pathways mentioned in this paragraph can be found in Appendix A.

These genes marked by DIGS were also reported as frequently mutated genes. For example, [64] indicated FAT4 is the most compelling gene in colorectal cancers, and mutations in DCHS1 and DCHS2 were reported too. From DIGS results, the FAT4 and DCHS2 were the top two important genes marked in red in the Hippo Signaling pathway map. Also, in the Renin Angiotensin System pathway, the ENPEP that is ranked first was indicated by [178] as a risk factor associated with a high risk of developing colorectal cancer.

4.4.2 Conclusions

In this chapter, the adaptability of the optimisation-based pathway activity inference method on RNA-Seq data was investigated on a dataset of colorectal cancer from the TCGA dataset. Three state-of-art pathway activity inference methods, MEAN, PCA and Pathifier, were used for comparison. I used three metrics, sample subtype prediction, survival information prediction and robustness to noise to demonstrate the utility of the DIGS model. The results illustrated that the DIGS method, which summarises pathway activity values as a weighted linear combination of pathway constituent genes, produced the best performances in all three pathway activity evaluation metrics. It is worth mentioning that DIGS has an independent approach for finding important pathways and important genes, rather than relying on an out-of-frame statistical proof approach. The solution outputs of DIGS directly provide the ranges of the sample

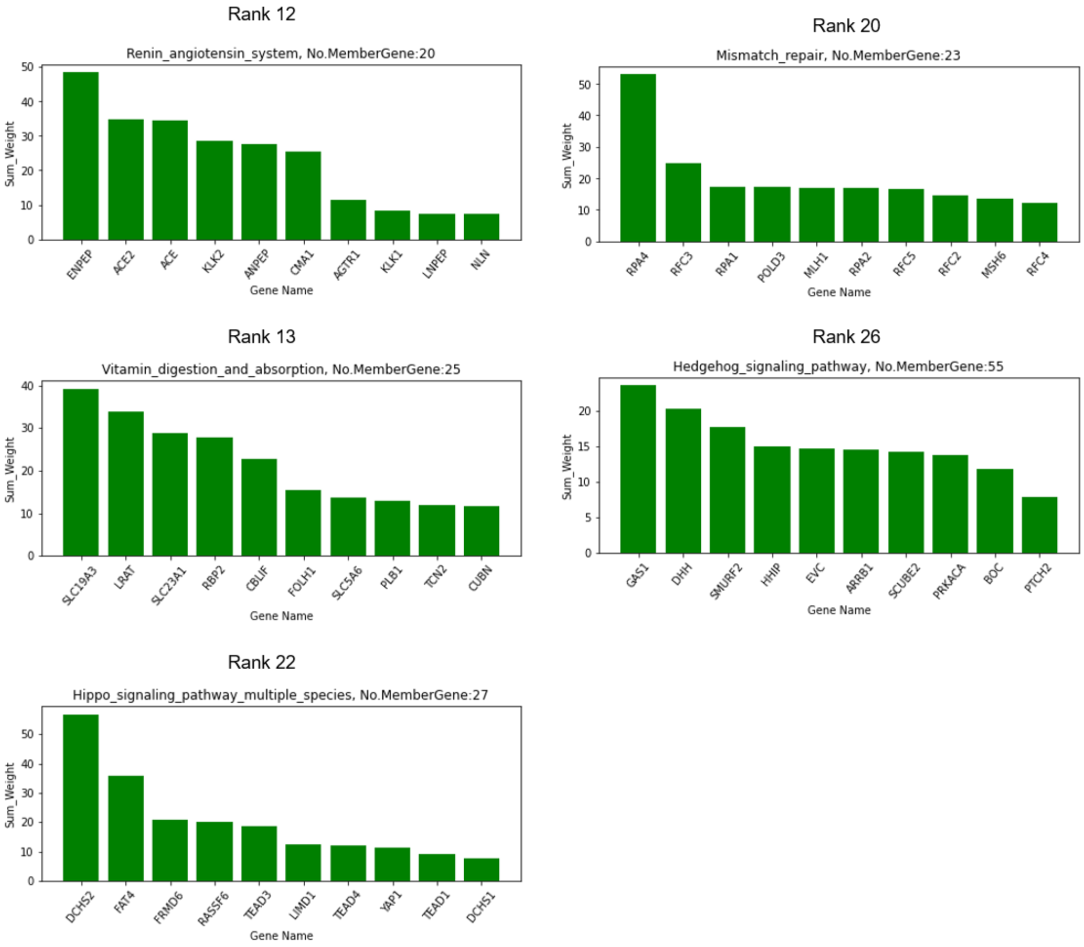


Fig. 4.10 Gene Weights for top 10 pathways. Top pathways were selected by ranking the individual pathway classification accuracy.

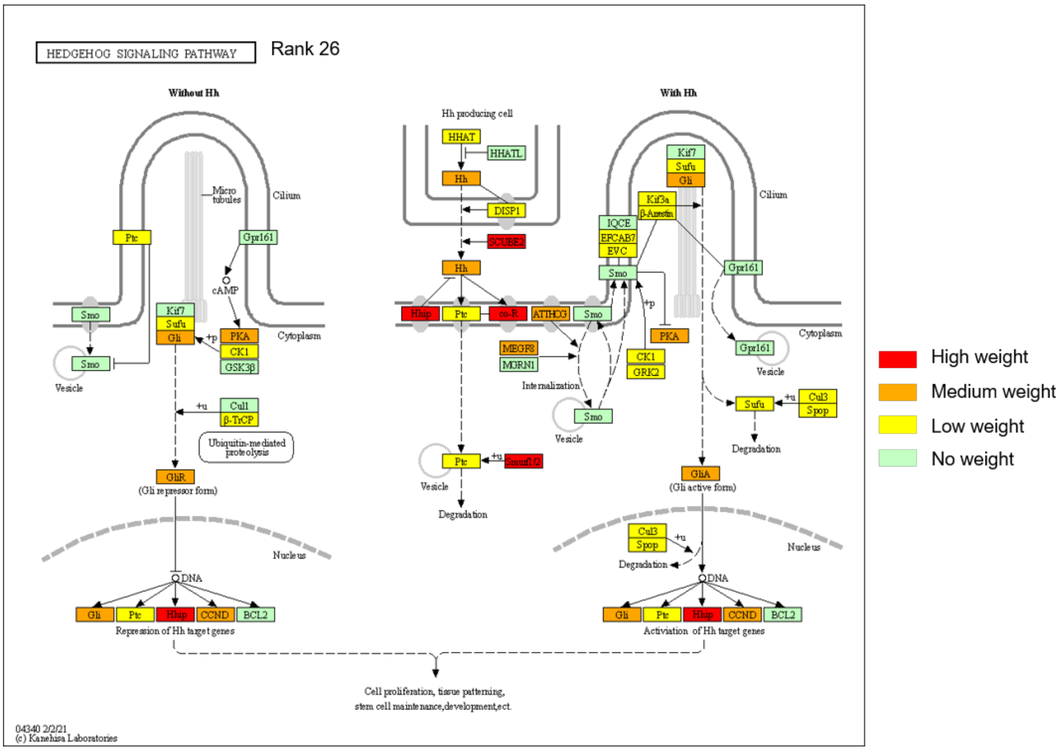


Fig. 4.11 KEGG pathway map with colors for significant genes identified.

labels and the weights of genes. Literature supports the pathways and genes deduced as significant by DIGS in complex diseases such as colorectal cancer.

Overall, the contribution of this chapter is two-fold. First, the DIGS model is applied to RNA-Seq data and validated using three metrics; the results prove the adaptability of the DIGS model to RNA-Seq data. Second, the special and specific way of interpreting the biological meaning from the pathway activity inference model makes DIGS easier to interpret than other methods. In the next chapter, the focus will shift from exploring the existing model to modifying the model for better computational efficiency.

Chapter 5

A Novel Optimisation Model for Pathway Activity Inference

Motivated by the excellent performance of DIGS, the model is refined to a more robust optimisation-based pathway activity inference method for multi-class disease classification tasks in this chapter, called DIGS2. Compared with the original model, DIGS2 reduces the number of binary decision variables to decrease computational complexity. Therefore, the DIGS2 model is more computationally efficient. This chapter adopts the evaluation pipeline from the former chapters to evaluate the performance of the new proposed model. In brief, three evaluation metrics (e.g. classification accuracy, survival prediction and robustness) are used for evaluation. Also, biological interpretation is derived from the output of the optimisation model. This chapter also extends the individual pathway evaluation process introduced in Chapter 4.3.4 to form a more powerful evaluation method named NaiveDIGS.

5.1 Introduction

Reducing the dimensionality of the gene expression profiles from gene-level data to pathway-level data is important for the diagnosing and prognosis of complex diseases like cancers [179–181]. A series of methods have been proposed to address this problem by constructing a one-dimension feature called Pathway Activity [53]. However, robustness and biological interpretability are challenging problems when trying to reduce the dimensionality from whole-genome profiling to a few hundreds of pathways.

Except for the baseline methods (e.g. MEAN [53], z-score [182] , PCA [148–150, 111] and Bayesian [183]), various methods have been introduced to improve interpretability and predictive power. A method called CORGs [96] ranks the pathway constituent genes according to their discriminative power, thereby identifying a small set of highly differentially expressed genes. Pathifier [111] uses a principal curve generated from non-tumour samples to calculate the pathway deregulation score of the tumour sample. The top pathways from Pathifier demonstrated high relevance to critical pathways of various diseases. Although these methods emphasise interpretability, most are still arithmetic, i.e. being based and relying on capturing the variance from the input dataset. Therefore, for these methods, the performance of disease classification and biological interpretation highly depends on the quality and volume of the input data. Also, most of the advanced methods are either applied to binary classification problems or are designed for a small number of signalling pathways [183, 184], as summarised in Table 2.1.

The mathematical programming optimisation model DIGS [113, 185] introduced in Chapter 2 aims to overcome these shortages. The superiority of the DIGS model has been demonstrated in Chapters 3 and Chapter 4, where the DIGS model is applied to two kinds of high-throughput data and explore its multiple applications. The advantages of the DIGS model derived from these two chapters are summarised below: (i) the DIGS model applies to multi-class disease classification problems, (ii) has been demonstrated to outperform other pathway activity inference methods in multi-classification tasks. (iii) DIGS offers interpretable rules and such biological interpretability is obtained at both pathway, as well as gene level (iv) DIGS can be applied on any type of high throughput profiling method.

It is noted that the computational complexity of the DIGS optimisation model prevents it from reaching a optimal solution in reasonable time. This is summarised from the model solutions of over one thousand runs from the former two chapters. Within the computation time limit that was set in the model implementation (200 seconds), the number of global optimal solutions achieved did not reach expectations. The optimal solution is the solution where there are no other feasible solutions with the better objective values [8]. It is reasonable to believe that the optimal solution represents a better objective value and, therefore, a better degree of sample separation at pathway activity level. More detailed statistics of the solutions will be presented in this chapter.

To achieve better solution quality, the DIGS model is refined to a more robust optimisation-based pathway activity inference method for multi-class disease classification tasks in this chapter, called DIGS2. The large number of binary decision variables prevented the DIGS model from getting more optimal solutions. Therefore, the model in DIGS2 was reformulated to include fewer number of binary decision variables. At the same time, DIGS2 retains the core concept of the original DIGS model for calculating the pathway activity values, i.e. the weighted summation of genes expression values. The mathematical model with descriptions of each constraint will be delivered in Section 5.2. Also, this chapter proposes a new phenotype prediction method NaïveDIGS, which can conduct sample phenotype prediction directly using the outputs of DIGS2 and does not rely on external machine learning classifiers.

In terms of experiment design, this chapter uses RNA-Seq datasets of colorectal cancer and breast cancer from The Cancer Genome Atlas (TCGA) database [162]. The colorectal cancer dataset is the same as the dataset used in Chapter 4. The results show that the newly proposed optimisation model significantly improves classification accuracy on multi-class problems compared to the original DIGS model and other comparative pathway activity inference methods.

5.2 A Novel Method for Pathway Activity Inference for Disease Classification

This section proposes a novel optimisation model for pathway activity inference. First, the indices, sets and parameters associated with the DIGS2 model are listed below:

Indices

s Sample (s=1,2,. . . ,S)

m Gene (m=1,2,. . . ,M)

c, k Class or phenotype (c=1,2,. . . ,C)

cs Class label for sample s

Parameters

G_{sm} Standardised gene expression profile

ε A small positive number

U A large positive number

Binary Variables

L_m 1 if effect of gene m on pathway activity inference is positive; 0 if negative effect

Y_{kc} 1 if upper bound of pathway activity range for class k is lower than lower bound of that for class c; 0 otherwise

Positive Variables

rp_m Positive influence of gene m towards pathway activity inference

rn_m Negative influence of gene m towards pathway activity inference

D_s Distance between sample pathway activity pa_s and pathway activity range of the phenotype that the sample belongs to

Unrestricted Variables

pa_s Pathway activity of sample s

LO_c Lower bound of range of class c on pathway activity

UP_c Upper bound of range of class c on pathway activity

Next, the objective function and the constraints that form the model are introduced.

In a similar manner as in DIGS, pathway activity value pa_s in the new model is also defined as a weighted linear summation of the expressions of pathway constituent genes:

$$pa_s = \sum_{m}^{M} G_s m(rp_m - rn_m) \forall s(5-1)$$

where $G_s m$ is the gene expression value for sample s and gene m, M is the total number of member genes for the current pathway, and S is the total number of samples. rp_m and rn_m are positive continuous variables that model the positive and negative weights of a gene m; the optimination model determines their values.

The following Equations (5-2), (5-3) and (5-4) set the restrictions on gene weights. A set of binary variables L_m, which takes values of either 0 or 1, is introduced. Equations (2) and (3) below ensure that, for each gene m, at most one of rp_m and rn_m can take positive values:

$$rp_m <= L_m \quad \forall m \quad (5-2)$$

$$rn_m <= (1-L_m) \quad \forall m \quad (5-3)$$

When $L_m = 1$, rp_m can take any value between 0 and 1 while rn_m is forced to be equal to 0; otherwise, when $L_m = 0$, $rp_m$is forced to be equal to 0 while rn_m can be between 0 and 1. In the other case, both rp_m and rn_m can equal 0, which means this particular gene gets zero weight towards its contribution to pathway activity. Overall, a gene can have positive, negative or zero weight toward the composite feature construction.

For normalination purposes, the summation of absolute gene weights should equal :

$$\sum_{m}^{M}(rp_m + rn_m) = 1 \quad (5-4)$$

Each phenotype occupies a unique interval on the pathway activity dimension and should not overlap with the intervals where samples of other phenotypes belong, which is implemented by Equations (5), (6) and (7):

$$UP_k + \varepsilon \leq LO_c + U(1 - Y_{kc}) \quad \forall k < c \quad (5-5)$$

$$UP_c + \varepsilon \leq LO_k + UY_{kc} \quad \forall k < c \quad (5-6)$$

$$LO_c \leq UP_c \quad (5-7)$$

where both c and k denote phenotype; LO_c and UP_c are continuous variables modelling the lower and upper bound of the activity interval of phenotype c. U and ε are respectively arbitrarily large and small positive constants. Y_{kc} is a set of binary variables used to ensure the non-overlapping property in pair-wise phenotype intervals. When $Y_{kc} = 1$, the lower bound of phenotype c is greater than the upper bound of phenotype k ($LO_c > UP_k$); when $Y_{kc} = 0$, the lower bound of phenotype k is greater than the upper bound of phenotype c ($LO_k > UP_c$). In either case, the phenotype ranges of k and c do not overlap.

For sample s, the violation distance D_s (positive continuous variable) is defined as the distance between sample pathway activity pa_s and pathway activity range of the phenotype that the sample belongs to $[LO_{c_s}, UP_{c_s}]$.

$$pa_s - UP_c \leq D_s \quad \forall sc_s \quad (5-8)$$

$$LO_c - pa_s \leq D_s \quad \forall sc_s \quad (5-9)$$

Figure 5.1b shows three possible scenarios between pa_s and$[LO_{c_s}, UP_{c_s}]$. When pa_s is outside its target phenotype interval and smaller than its lower bound LO_{c_s}, the violation distance becomes $D_s = LO_{c_s} - pa_s$. When pa_s is outside its target range and greater than its upper bound UP_{c_s}, the violation distance becomes $D_s = pa_s - UP_{c_s}$. When pa_s is inside its target range, violation distance is equal to 0.

The objective function (5-10) in the new model minimises the sum of violation distances D_s over all samples. The new objective function removes a large number of binary variables compared with DIGS (Equation 2-8k), which employs the binary variable E_s to decide whether a sample s falls within the correct range.

$$Min \quad z = \sum D_s \quad (5-10)$$

This novel optimisation model, named DIGS2, consists of a linear objective function and linear constraints. The presence of both binary variables and continuous variables makes it a mixed-integer linear programming (MILP) model. DIGS2 builds on our previous DIGS model but is formulated intentionally to lose most of the binary decision variables in DIGS, therefore, making it easily solvable. The mathematical formulation is summarised below:

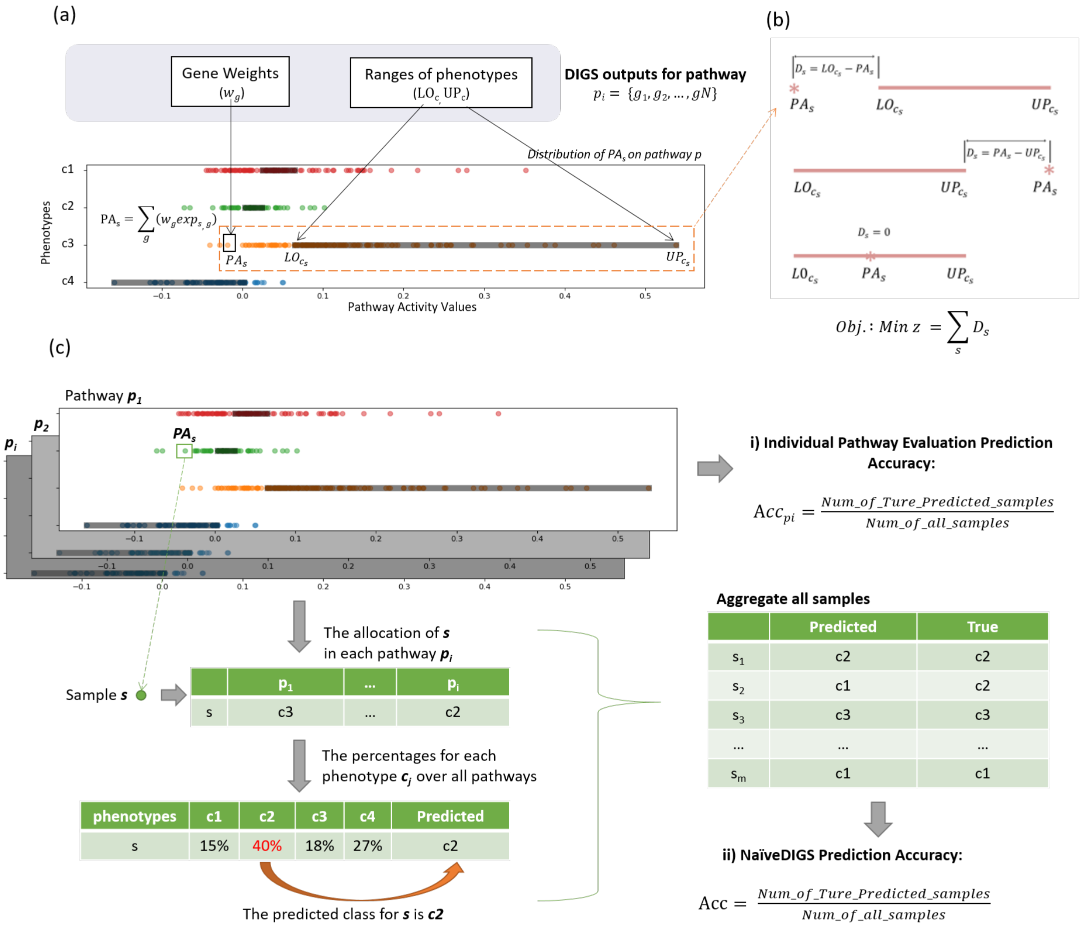
Objective function:

Minimising the summation of violation distance (5-10)

Subjected to:

Pathway activity definition (5-1)

Restrictions on positive and negative gene weights (5-2, 5-3 and 5-4)


(a)
Gene Weights
(w_g)
Ranges of phenotypes
$(\mathrm{LO}_c, \mathrm{UP}_c)$
DIGS outputs for pathway
$p_i = \{g_1, g_2, \ldots, gN\}$
Distribution of PA_s on pathway p
$\mathrm{PA}_s = \sum_g (w_g exp_{s,g})$
c1
c2
c3
c4
Phenotypes
PA_s
LO_{c_s}
UP_{c_s}
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
Pathway Activity Values
(b)
$D_s = LO_{c_s} - PA_s$
PA_s
LO_{c_s}
UP_{c_s}
$D_s = PA_s - UP_{c_s}$
LO_{c_s}
UP_{c_s}
PA_s
$D_s = 0$
LO_{c_s}
PA_s
UP_{c_s}
$Obj.: Min\ z = \sum_s D_s$
(c)
Pathway p_1
p_2
p_i
PA_s
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
i) Individual Pathway Evaluation Prediction Accuracy:
$\mathrm{Acc}_{pi} = \frac{Num\_of\_Ture\_Predicted\_samples}{Num\_of\_all\_samples}$
The allocation of *s* in each pathway p_i
Sample *s*
The percentages for each phenotype c_j over all pathways
The predicted class for *s* is *c2*
Aggregate all samples
ii) NaïveDIGS Prediction Accuracy:
$\mathrm{Acc} = \frac{Num\_of\_Ture\_Predicted\_samples}{Num\_of\_all\_samples}$

	p_1	…	p_i
s	c3	…	c2

phenotypes	c1	c2	c3	c4	Predicted
s	15%	40%	18%	27%	c2

	Predicted	True
s_1	c2	c2
s_2	c1	c2
s_3	c3	c3
…	…	…
s_m	c1	c1

Fig. 5.1 Individual pathway evaluation and naïveDIGS classification. (a) presents a motivating example of the DIGS2 model. Pathway activity is defined as the weighted linear summation of expressions of constituent genes. Gene weights W_g are modelled as free continuous variables in the DIGS2 model. Each phenotype occupies a distinct range (continuous variables $[LO_{c_s}, UP_{c_s}]$). The ranges are not overlap. (b) shows three possible scenarios between sample pathway pathway activities and the pathway activity range of its phenotype $[LO_{c_s}, UP_{c_s}]$.The distance between PA_s and range of its phenotype is termed as violation distance D_s.(c) Two applications of the outputs of the DIGS2 model are introduced. i) Individual pathway evaluation accuracy. ii) The NaïveDIGS accuracy evaluates the prediction accuracy by combining the results of all pathways.

Non-overlapping constraints for phenotype ranges (5-5, 5-6, and 5-7)

Sample violation distance calculation (5-8 and 5-9)

Figure 5.1a shows a motivating example of how DIGS2 separates the samples of different classes. The input data is a pathway expression matrix, where columns consist of 30 pathway member genes, and rows consist of samples from the gene expression profile. The model outputs consist of gene weights ($rp_m - rn_m$), which are further used for calculating the pathway activity values with equation (5-1), and class intervals (LO_c, UP_c). The points on the plot represent the inferred pathway activity values, and the black boxes on the four axes represent the class intervals. The colours of sample points represent the true class. All the intervals on the plot are non-overlapping, restricted by the Equations (5-5 to 5-7). Through Equations (5-8) and (5-9), the objective function (5-10) can perform the task of minimising the summation of D_s for all samples. As shown in Figure 5.1a, the pathway activity values for samples of the same class label are clearly clustered together, and the bounds of each class are close to the centres of the clusters, which verifies that the DIGS2 model can separate the sample labels to the largest extent.

The DIGS2 model is implemented using general-purpose solution algorithm CPLEX in the General Algebraic Modelling System (GAMS) [127], and the solutions are identified in a user-specified time limit (200 seconds by default). Note that simply increasing the time limit only marginally improves the quality of the solution. At the same time, the Relative Optimality criterion solver (optcr) is set to its default value of zero.

5.2.1 Other comparative pathway activity inference methods

A range of other pathway activity inference methods were implemented for comparison, including two baseline methods: i) MEAN [53] method that calculates the average gene expression values as pathway activity, which is the original work that proposed the concept of pathway activity inference. ii) PCA [97] approach that calculates the first principle component as a representation of pathway activity.

Three more advanced methods were used for comparison: i) our original DIGS model, which selects the best subset of genes to build pathway activity, minimising the number of misclassified samples; ii) The GSVA [100] approach, which is a variation of Gene set Enrichment analysis [40]; iii) The Pathifier [111] method, who investigates the extent to which the behaviour of a sample deviates from the control group. Pathifier was proven

Table 5.1 TCGA Datasets

Dataset	Tumour or Normal Label	Molecular Subtype label
COAD	Tumour: 480 Normal: 41	CMS1: 85 CMS2: 165 CMS3: 58 CMS4: 120
BRCA	Tumor: 480 Normal: 41	LumA: 579 LumB: 217 Basal: 191 Her2: 82 Normal: 22

to be the method that achieved the best performance in a recent review [52], which evaluated 13 pathway activity inference methods. Many other studies have employed these methods, and detailed descriptions are given in Chapter 2.4. The criteria for selecting these methods are (1) the method is widely used, and (2) the method is eligible for multi-class classification.

5.2.2 Dataset preparation

The experiments in this chapter use a widely referenced breast cancer dataset with standard molecular subtype annotations. As a representative cancer, breast cancer has been extensively studied and has been used as the testing dataset in many studies related to pathway activity inference [182, 52, 84]. The colorectal cancer dataset used in Chapter 4 is also used here.

The raw count RNA-Seq dataset was downloaded from The Cancer Genome Atlas (TCGA) [162] with their corresponding clinical information. Two publicly available TCGA projects (BRCA and COAD) were chosen as the experimental datasets for this work [111, 186, 187, 161]. The read count data is normalised by the Upper quartile FPKM (FPKM-UQ) [168] approach to obtain the gene expression profiles. Then, genes with very high missingness (over 30% zero expression values across the sample cohort) are removed. Table 5.1 shows details of these two RNA-Seq datasets.

The total number of samples in the COAD dataset is 521, of which 480 are tumour tissue samples, and 41 are normal tissue samples. The number of samples of four molecular subtypes is 85, 165, 58 and 120 for CMS1, CMS2, CMS3 and CMS4, respectively [167]. For the BRCA dataset, the total number of samples is 1211; among them, 1091 are tumour tissue samples, and 120 are normal tissue samples. PAM50 subtype [188] is used to determine the molecular subtype (retrieved using TCGAbiolinks package [55])

Table 5.2 Characteristics of breast cancer subtypes

Subtype	Gene Phenotype	Characteristics
Luminal A	ER+, PR+, HER2-	Grow slowly and good prognosis
Luminal B	ER+, PR-, HER- or ER+, PR+-, HER2+	Faster than Luminal A and slighted worse prognosis
HER2	ER-, PR-, HER2+	Faster than Luminal, worse prognosis but successful treated
TNBC	ER-, PR-, HER2-	More aggressive than Luminal with BRCA1 mutation

of BRCA. The subtype of breast cancer can be divided into Luminal A, Luminal B, Her2 and Basal (Table 5.2). Luminal A type is defined as the lack of HER2 expression; Luminal B is defined as the lack of progesterone receptor (PR); Her2 type is defined as the lack of estrogen receptor (ER) and PR; Basal type is defined as the lack of ER, PR, and HER2 expression [189]. In BRCA, there are 579 Luminal A samples, 217 Luminal B samples, 191 Basal samples, 82 Her2 samples and 22 Normal samples.

Biological pathway information was retrieved through the Kyoto Encyclopedia of Genes and Genomes (KEGG) API [170, 146] In total, 279 Homo sapiens pathways were used in this work, the number of genes in these pathways is 6761.

5.2.3 Pathway activity evaluation

The method implementation and validation pipeline in this chapter are similar to Chapter 4. In the following subsections, the implementation process for the three pathway activity evaluation metrics of classification, survival prediction and robustness are described, as well as the cross-validation process.

5.2.3.1 Classification

In this work, two classifiers, Random Forest (referred to as RF) and K-Nearest-Neighbor (referred to as KNN), are used. For each dataset, 10-fold cross-validation was applied on the normalised RNA-Seq data three times to gain robust results, creating 30 training sets and 30 testing sets. All the pathway activity inference methods were trained separately for each pair of training and testing sets. Then classifiers were trained on the inferred pathway activity of training samples and tested on the testing samples.

As shown in Table 5.1, 2-class classification uses tumour and normal as the sample labels; multi-class classification uses the molecular subtypes as the sample label. To avoid the effects of unbalanced 2-class labels for both datasets, SMOTE [57] (described in Section 2.3.2.2) is applied to the training sets to make the number of normal samples equal to the number of tumour samples. The RF parameter n_estimators was tuned from 200 to 2000 using grid search optimisation on the training sets to optimise the performance of the trained model. The parameter k of KNN was selected using a trial-and-error process on the training dataset testing values 3, 5, 10, 20, 30 and 50. k=30 was selected as it was the one performing the highest classification metrics. RF and KNN were implemented by the Python library sklearn 0.24.0 [172]. The performance of classifiers was evaluated with four metrics: Accuracy, F1-score, Precision, and Recall, averaging over 30 testing sets.

5.2.3.2 Robustness against noise in data

The robustness of the pathway activity inference methods when facing the unpredicted fluctuations in the gene expression values were assessed. Pathway activity inference methods are expected to retain good prediction performance as the noise level increases.

The Gene expression perturbation simulation process is performed by permuting the sample order for the randomly selected genes [95]. The proportion of genes affected was set as 0%, 3%, 10% and 50% (see Figure 4.2). Therefore, in addition to the original gene expression profile, three perturbed expression profiles were created for the two datasets. The five pathway activity inference methods were performed on all expression profiles with the 10-fold cross-validation strategy. Then the robustness of the methods was evaluated by the multi-class classification accuracy with the KNN classifier.

5.2.3.3 Survival analysis

Survival analysis is a key metric to express prognosis in cancer studies. As a new composite feature that can aggregate the gene expression values, pathway activity values are expected to perform better in predicting survival than on the random data.

Clinical data of both datasets were downloaded together with the raw read counts from the TCGA database, and the survival status and survival time for samples were extracted. Considering the inaccuracy caused by the loss of clinical information, samples with too short of a follow-up time (where the time was less than one year, and the status was alive) were removed. In the remaining cases, the "0" samples were randomly selected to keep the number of "0" and "1" samples the same. After sample selection, 196 samples were employed for COAD, and 206 were kept for BRCA.

Survival Random Forest [14] model (implemented using the Python package scikit-survival 0.16.0 [173]) is used for training and testing the survival model on 30 pairs of training and testing pathway activity matrices. Concordance index (c-index) [174] is used for evaluating the model. The c-index on random data is 0.5.

5.3 Comparative study

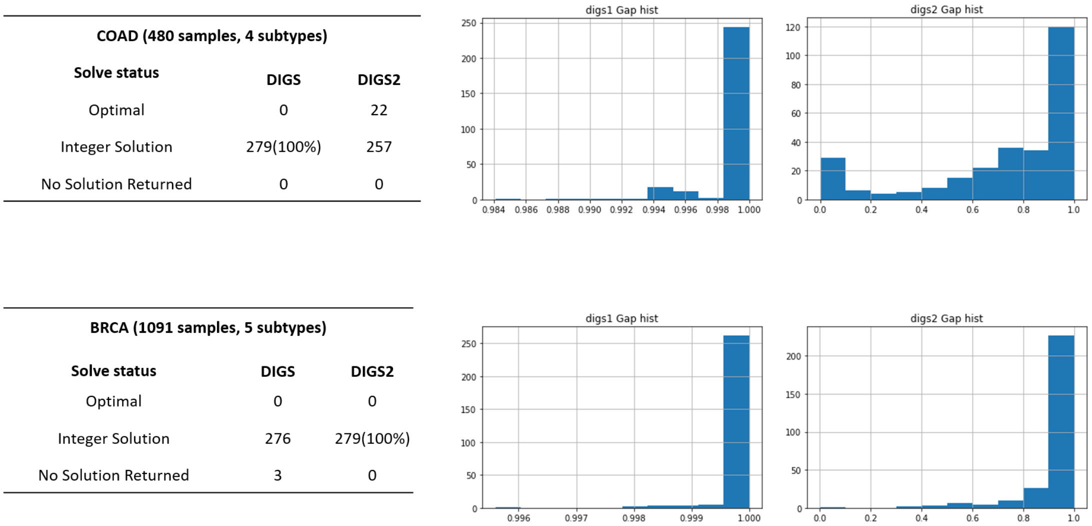
The comparative study consists of two parts. The first is to compare the computation efficiency between the original DIGS model and the newly proposed DIGS2 model. With fewer binary variables involved, the DIGS2 model is expected to have a higher optimal solution rate. The second comparison compares the three pathway activity evaluation metrics among all the methods, and DIGS2 is expected to perform well.

5.3.1 Efficiency improvement in the DIGS2 model

The solver status, model status and objective value can express the solution quality of the optimisation problems. GAMS contains many components for checking and comprehending a model through the output file (GAMS listing file) [127]. The "Solve Status" and "Gap" reported in the GAMS listing files are extracted to evaluate the solution qualities of the DIGS and DIGS2 models.

The "Solve Status" for all the solutions produced from the cross-validation process is summarised in Figure 5.2. For each dataset, DIGS2 and DIGS models were solved 30 times per pathway; the total number of solutions is 8370 (30 multiple 279 pathways) for each dataset. Although the optimisation models aim to gain global optimal solutions, feasible solutions are also acceptable for real problems because of the limitations, such as high dimensional data and the limited execution time [190]. The results showed that, in both datasets, DIGS2 gets more "Optimal" solutions and "Integer Solution" solutions than DIGS (22 more "Optimal" in COAD and three more "Integer" solutions in BRCA), which reflects that DIGS2 has capability of solving the same problems to better quality.

Another solution quality evaluation metric, "Gap", was plotted as histograms in Figure 5.2. Gap indicates the difference (in percentage) between the best potential and the best-found objective value. Instead of looking for optimal solutions that take a long time to compute, a solution guaranteed to be not worse than a certain percentage of the optimal solution is also acceptable in practical applications. Therefore, the Gap value



COAD (480 samples, 4 subtypes)		
Solve status	**DIGS**	**DIGS2**
Optimal	0	22
Integer Solution	279(100%)	257
No Solution Returned	0	0

BRCA (1091 samples, 5 subtypes)		
Solve status	**DIGS**	**DIGS2**
Optimal	0	0
Integer Solution	276	279(100%)
No Solution Returned	3	0

Fig. 5.2 Calculation efficiency Comparison

is in the range of 0 to 1, and a smaller Gap value represents better solution quality. It can be concluded from the figure that DIGS2 shows great improvements in decreasing the gap values. For smaller datasets (COAD), a large proportion of the solutions got gap values less than one among solutions of DIGS2; for a larger dataset that is more computationally consuming (BRCA), the observations of DIGS2 achieving less than one gap value still increased.

5.3.2 Evaluations on classification, survival analysis and robustness

The DIGS2 model performance was compared to other approaches in four aspects: Classification performance on 2-class labels, Classification performance on multi-class labels, Robustness against noise in data and survival analysis.

For Classification performance comparison on 2-class and multi-class problems, the inferred pathway activity profiles from 30 training sets of six methods (DIGS2, DIGS, MEAN, PCA, GSVA and Pathifier) were fed into the RF and KNN to train the classifiers. Pathifier is used for only 2-class classification problems as it is not designed for the multi-class problem. Then the classification evaluation metrics were averaged across 30 testing sets. To achieve an objective evaluation of classification performance, the training approach, i.e. inferring pathway activity and training a classifier, is always blind to testing sets. The outcomes of all methods and classifier combinations are shown in Figure 5.3a and 5.3b for multi-class and 2-class classification, respectively. The classification metrics are summarised in Table 5.3.

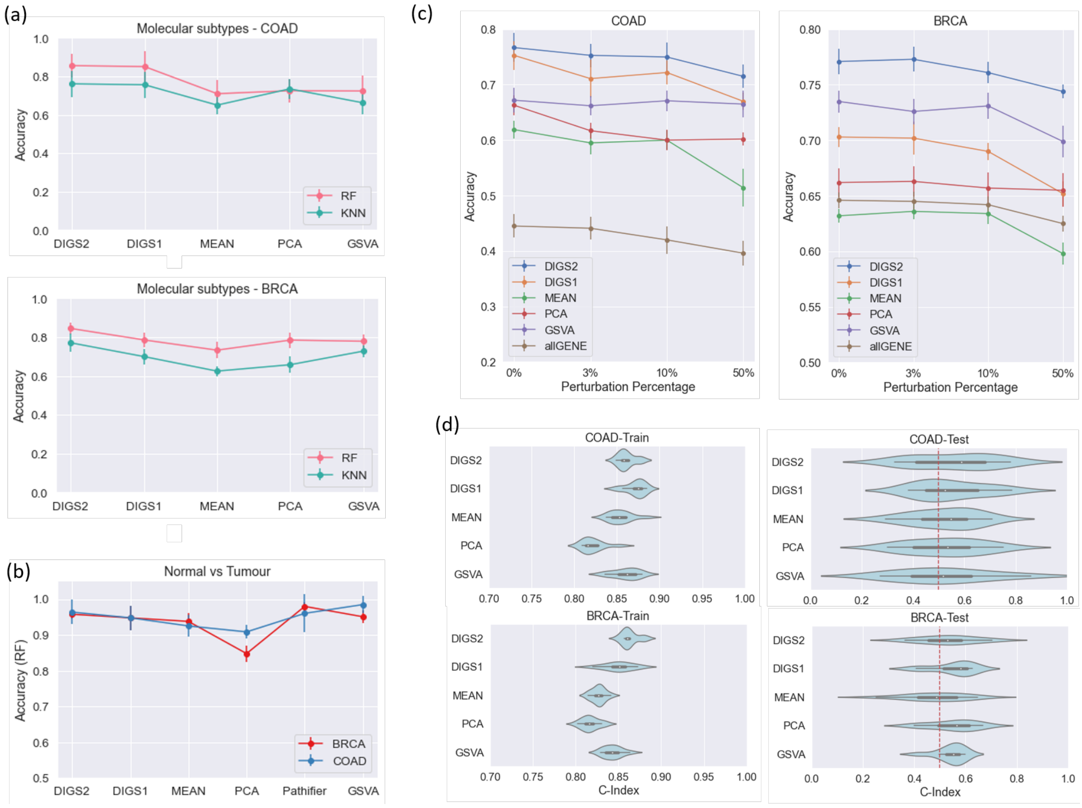


Fig. 5.3 Comparison between pathway activity inference methods. (a) Prediction performance comparison of five pathway activity inference methods (DIGS2, DIGS, MEAN, PCA, GSVA) on the multiclass scenario (molecular subtype) with two classification algorithm Random Forest (RF) and K-Neatest Neighbour (KNN). Prediction performance is summarised as the mean and standard error of classification accuracies achieved over three times of stratified 10-fold cross-validation. (b) follows the same pipeline as (a) with one more method, Pathifier, that is designed for inferring pathway activity towards binary sample labels. The prediction accuracy of RF on pathway activity matrices inferred under a 2-class scenario is plotted. (c) displays the prediction accuracy with noisy data added to RNA-Seq data. The x-axis is the percentage of perturbation performed on the RNA-Seq counts; the y-axis indicates the averaged prediction accuracy and standard errors of the KNN classifier over one time of 10-fold cross-validation. AllGENE refers to directly using the gene expression values to train and test the classifier. Pathway activity methods are implemented under a multiclass scenario. (d) shows the c-index for evaluating the fitted Random Survival Forest model on the pathway activity matrices (multiclass scenario). Both training and testing sets are included. c-index = 0.5 indicates the performance on the random dataset.

Table 5.3 Multi-class classification results

Random Forest								
	DIGS2		DIGS		MEAN		PCA	
	Avg.	Std.	Avg.	Std.	Avg.	Std.	Avg.	Std.
Accuracy	0.845	0.028	0.786	0.033	0.733	0.036	0.785	0.039
F1-socre	0.843	0.031	0.756	0.039	0.706	0.041	0.761	0.040
Precision	0.849	0.033	0.788	0.042	0.72	0.055	0.762	0.055
Recall	0.845	0.028	0.786	0.033	0.733	0.036	0.785	0.039
KNN								
Accuracy	0.771	0.034	0.703	0.026	0.632	0.018	0.662	0.038
F1-socre	0.74	0.045	0.643	0.038	0.581	0.026	0.623	0.043
Precision	0.797	0.043	0.748	0.049	0.63	0.04	0.662	0.034
Recall	0.771	0.034	0.703	0.026	0.632	0.018	0.662	0.038

From Figure 5.3a, it is concluded that the inferred pathway activity using the proposed DIGS2 model has resulted in more accurate predictions when compared to existing methods in the literature. DIGS2 achieved higher averaged accuracy for all datasets and classifiers in classifying molecular subtypes. As shown in Figure 5.3b, although GSVA performs best for BRCA and Pathifier performs best for COAD, DIGS2 is the second-best method in both datasets. Also, as all methods perform well in 2-class classification problems (over 90% accuracy, except for MEAN on BRCA), the slight differences between methods are negligible to some extent. Therefore, DIGS2 has shown its strong superiority in predicting the sample classes, especially in the multi-class scenario.

Robustness is evaluated by the sample prediction accuracies on the testing pathway activity profiles across different perturbation percentages (0%, 3%, 10% and 50%) of the RNA-Seq datasets. The classifier used for sample prediction is KNN (with k = 30) under the multi-class classification scenario. The results are shown in Figure 5.3c. AllGENE refers to the direct use of gene expression values to train and test KNN classifiers. It is used as a baseline to evaluate the extent to which noisy data will affect the accuracy of predictions without the participation of path information.

From Figure 5.3c, it is evident that DIGS2 provides the best overall performance. As more noise was introduced to the data, the DIGS2 accuracy stayed above 70% for both COAD and BRCA. Although PCA can keep a more stable performance while increasing the perturbation degree, the comparatively low accuracy in non-perturbed datasets inversely corroborated the poor power in differentiating sample types. For the performance of GSVA, while increasing the perturbation degree, the accuracies do not fluctuate in COAD but drops in BRCA. This phenomenon can be explained by the principle of GSVA, as it focuses on the ranking of the genes, not the exact count values. The distribution within the genes could only slightly affect the ranking between genes.

Survival regression was conducted for both training sets and testing sets. The c-index are presented in Figure 5.3d. For the training sets, all methods perform well on the two datasets. For the COAD dataset, DIGS takes the first place, and for the BRCA dataset, DIGS2 takes the first place. In the testing sets, nearly all methods yield c-index over 0.5 except MEAN on BRCA. Although the differences between the five pathway activity inference methods are not particularly distinct, DIGS2 and DIGS models still show more outstanding performance in predicting sample survival.

The Radar chart (Figure 5.4) compares the performance of the resulting pathway activity values of different methods on the three pathway activity evaluation metrics. The upper two plots compare the DIGS and DIGS2 models. For both datasets, the areas of DIGS2 are bigger than DIGS1, which proves that the DIGS2 model has better pathway activity quality. The other two plots summarise the performance over the three representative methods (DIGS2 for optimination, PCA for projection and GSVA for member gene ranking). The Multi-class Prediction vertex uses the RF prediction accuracies of molecular subtypes; Robustness (50%) vertex is the KNN prediction accuracies produced on 50% perturbed dataset; the c-index vertex is the c-index values of testing samples. The number on each vertex is normalised in a scale of 0-1 to illustrate the differences between methods. By comparing the areas in the radar chart, DIGS outperforms the other methods for nearly all the metrics for both datasets, especially on the multi-class Prediction vertex.

5.4 Biological interpretation

Another aspect of the analysis is how the new proposed DIGS2 model can be applied for deriving disease-related significant pathways and genes. In contrast to other methods, DIGS2 allows the user to explore the contribution of pathways in disease classification in greater depth. Extended from the works in the former chapters, a standard process for evaluating the pathway activity values of individual pathways is proposed in this section, called NaïveDIGS. Then its practical usage is applied to the BRCA dataset.

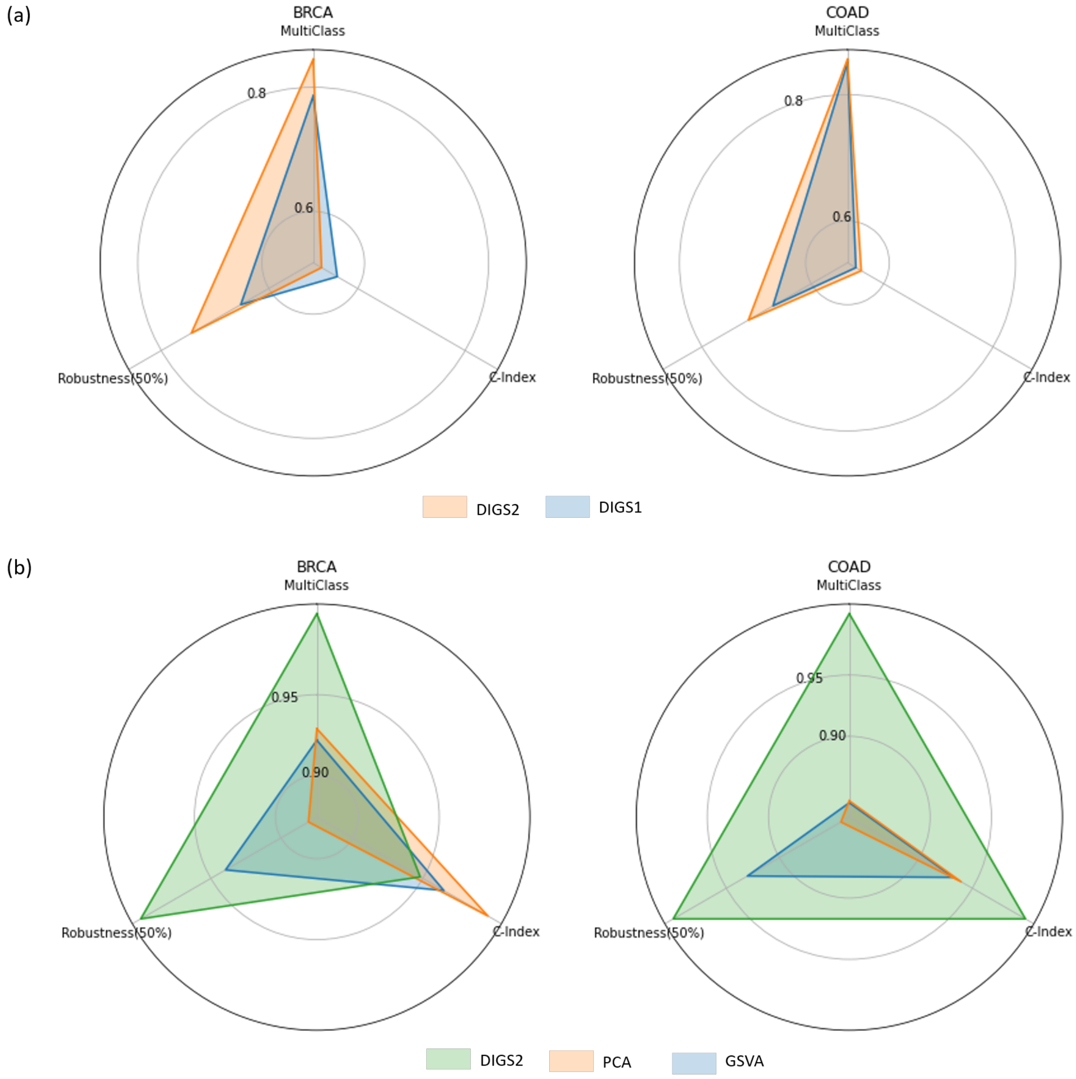


Fig. 5.4 Radar Charts. Each vertex counter-clockwise from 12 o'clock of the radar chart indicates sample phenotype prediction accuracy on molecular subtypes, robustness against 50% perturbed genes and survival regression evaluated by c-index. (a) compares the performance between three representative methods (DIGS2, PCA and GSVA). The number on each vertex is scaled to 0-1. (b) compares the performance between DIGS2 and original DIGS.

5.4.1 Individual pathway evaluation

One of the most important applications of pathway activity inference is identifying significant pathway signatures. Previous research uses computational or statistical approaches, such as p-value ranking or Information Gain index [161, 100, 159, 175], to select the pathways with relatively high significance. In this work, benefiting from the explainability of the optimisation model, we present a comprehensive method that can identify the important pathways directly from the outputs of DIGS2.

In Figure 5.1a and 5.1b, the mechanism of how the DIGS2 model explains the distribution of sample activity values for one pathway is illustrated. Based on that, another application of the outputs of the DIGS2 model is introduced in Figure 5.1c. For each pathway, the DIGS2 model provides the pathway activity value for each sample s and the class ranges for each class c. By calculating the distances of the samples to each class (Violation distance D_s), samples can be allocated to their nearest class. In other words, the allocation result is the predicted class of the sample s by pathway p_i. Therefore, the prediction accuracy of the pathway p_i can be calculated as the number of true positive samples divided by the total amount of samples. This accuracy has been mentioned in Chapter 4.3.5 and is termed as Individual Pathway Prediction Accuracy. Consequently, the pathways can be ranked according to their prediction accuracy, and the pathways with higher prediction accuracy are seen as significant pathways.

Except the ranking of pathways, DIGS2 also provides quantifiable evaluation of the constituent genes inside pathways. As the pathway activity value is defined as the weighted linear summation of the gene expression values, the gene that obtains higher weight from the DIGS2 model has a higher influence on the pathway activity values. Therefore, the member genes of a pathway can be ranked by their weights ($rp_m - rn_m$). Further, the importance of the genes can be quantified by the value of the weights. In this work, the final ranking of the genes is decided by accumulating the absolute weights from all the available results for the same pathway.

The power of classifying sample subtypes in a pathway activity inference method is often assessed using machine learning classifiers (e.g. use of the RF and KNN in this work). However, as the DIGS2 model by itself can be used to evaluate the pathway classification accuracy, here we propose a method to compute the overall classification accuracy for the DIGS2 model.

Continued from the calculation of Individual Pathway Prediction Accuracy in Figure 5.1c, for a specific sample s, the allocation results made by all the pathways are aggregated and transferred into percentage values. The percentage value for a class

Table 5.4 Prediction Accuracy between NaïveDIGS and ML classifiers.

Classifier	BRCA	COAD
NaiveDIGS	0.67 (0.064)	0.75 (0.043)
KNN	0.74 (0.046)	0.76 (0.068)
RF	0.84 (0.031)	0.85 (0.060)

c represents the number of pathways that allocate sample s into class c. Then, The final predicted class for sample s is the class that has the highest percentage. As the example shown in the figure, the highest percentage is 40%, which means 40% of the 279 KEGG pathways allocate the sample s into class $c2$. Therefore, the predicted class for s is $c2$ after aggregating all the pathways. In the next step, the predicted classes of all samples can be used for calculating the NaïveDIGS Prediction Accuracy, which is defined as the number of correctly predicted samples divided by the number of all samples. This method (referred to as NaïveDIGS) can be seen as an alternative to using machine learning classifiers.

In Table 5.4, the performance of NaïveDIGS is compared to Random Forest and KNN for multiclass scenarios (i.e. molecular subtypes). It is promising that the accuracy of NaïveDIGS is not significantly reduced compared to KNN, which implies that the relatively simple machine learning classifier brings little improvement to the classification capability of the model itself. The powerful classification ability of the DIGS2 model is further confirmed. Also, it is noticeable that the accuracy of NaïveDIGS is of the same level as the KNN accuracies for other pathway activity inference methods (referring to Figure 5.3a). This means that DIGS2 itself can provide high accuracy in predicting sample classes without the introducing of a classifier.

5.4.2 Biological pathway markers and gene marker identification

Beyond good prediction performance, the proposed DIGS2 model can also provide biological insights by identifying disease-relevant pathway markers. Pathway markers are the pathways most influential in the separation of molecular subtypes; therefore, based on the DIGS2 model, pathways that can predict the highest number of testing samples into the correct class are the pathways that should take the highest rank. Consequently, pathways are ranked according to the Individual Pathway Evaluation (Figure 5.1c) accuracies for both datasets.

Except ranking of pathways, DIGS2 also provides quantifiable evaluation for the constituent genes inside each pathway. According to the pathway activity definition in Equation (5-1), the gene gaining higher weight from the DIGS2 model means a stronger

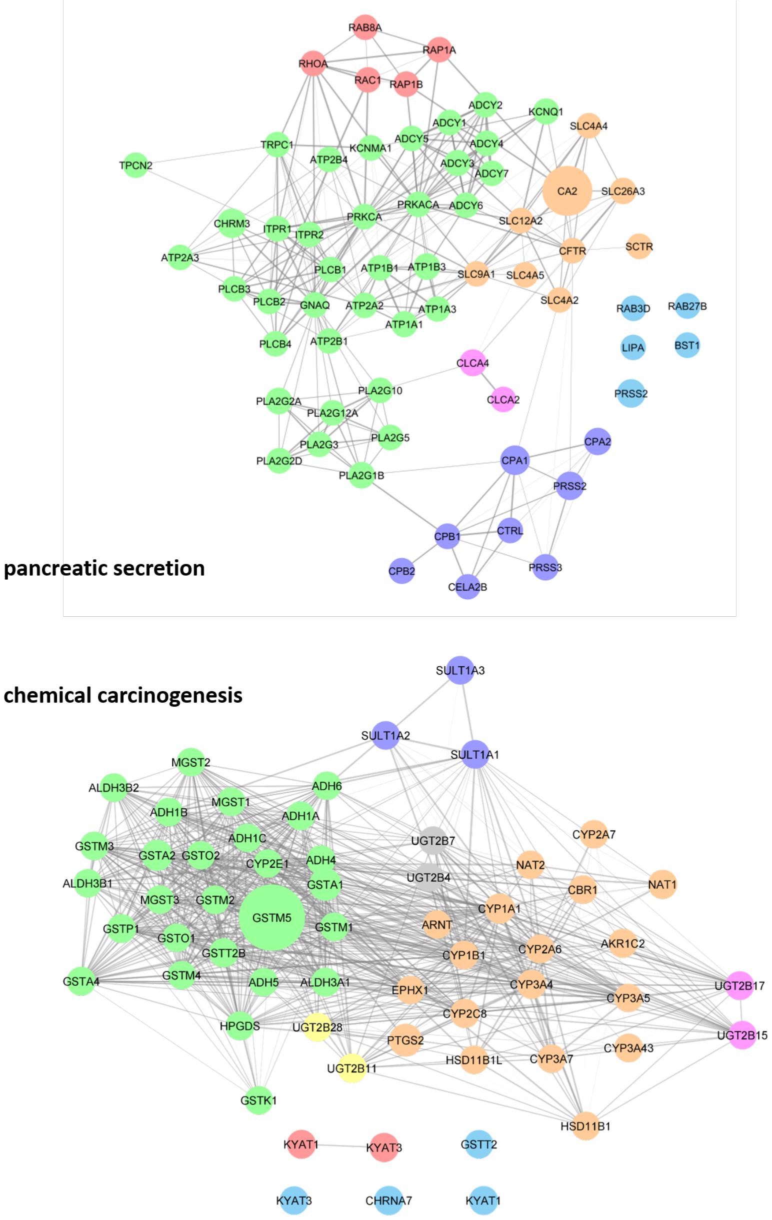
Table 5.5 Significant pathways and genes for BRCA

Pathway Name	No. Gene	Top Genes and Weights									
Pancreatic secretion	104	CA2	4.36	CPA1	0.72	CHRM3	0.69	PRSS2	0.53	CPA2	0.33
Circadian rhythm	32	RORB	3.24	ROR1	0.78	PRKAA2	0.49	RORA	0.49	CUL1	0.47
Peroxisome	87	AGXT	5.39	HAO2	3.21	ACSL6	0.23	PEX11A	0.10	IDH2	0.08
Chemical carcinogenesis	81	GSTM5	6.07	PTGS2	0.71	GSTA1	0.66	GSTA2	0.44	CYP1A1	0.30
Platinum drug resistance	73	GSTM5	6.71	GSTA2	1.30	GSTA1	0.82	CDKN2A	0.21	GSTT2B	0.15
Drug metabolism cytochrome P450	70	GSTM5	6.33	GSTA1	0.91	GSTA2	0.69	UGT2B11	0.22	FMO2	0.21
Folate biosynthesis	30	TPH1	4.11	PAH	3.88	ALPL	0.79	MOCOS	0.34	FPGS	0.20
Drug metabolism other enzymes	79	GSTM5	5.98	GSTA1	0.93	GSTA2	0.65	XDH	0.55	GSTT2B	0.21
Cocaine addiction	50	SLC18A2	3.56	DRD1	1.16	GRIN2A	0.93	CREB3L3	0.51	SLC18A1	0.41

influence on the inferred pathway activity values. Because the gene weights for DIGS2 are normalised by making the summation of all weights equal to 1 (Equation (5-4)), the ranking of constituent genes can be inferred by accumulating the absolute weights from 30 runs on the training samples for the same pathway. The gene ranking and corresponding weights of the top 70 pathways of BRCA are listed in Appendix B. Table 5.5 shows the top 10 pathways for BRCA with the top 5 pathway constituent genes and their accumulated weights.

I searched the literature to support the relationships between these pathways and genes and breast cancer. Several studies have demonstrated the role of Circadian Rhythm as an effective tumour suppressor [191]. Tumour development is triggered by the cellular level stimulation or disruption of signalling pathways due to the interaction between cells and environmental stimuli. The Circadian Rhythm pathway synchronizes the timekeeping in the peripheral tissues by integrating the light-dark input from the environment. It, therefore, has an impact on the development of tumours. The ROR family genes in Circadian Rhythm pathway regulate the secondary transcription/translational feedback loop and guide the rhythmic oscillation [192]. Peroxisome protein levels or enzymatic activities of peroxisome metabolism were largely reduced in breast tumor [193]. The top constituent gene AGXT is highly involved in colorectal cancer [194] and hepatocellular carcinoma [195]. Also, the relationship between folate and the risk of breast cancer has been massively investigated [196, 197]. It is noticeable that for Chemical carcinogenesis, Platinum drug resistance, Drug metabolism cytochrome P450, and Drug metabolism of other enzymes pathways, the GSTM5, GSTA1, and GSTA2 genes take important place in them. Several studies have verified the importance of GSTM family to the risk of breast cancer [198–201]. Our results are thus in agreement with existing reports, and the DIGS2 model was proven to possess the ability to identify new disease-gene associations.

For the two highly ranked pathways, Pancreatic secretion and Chemical carcinogenesis, that were not found in the literature to have a direct association with breast cancer, this work performs further analysis by reconstructing and clustering their corresponding protein-protein interaction (PPI) networks. Although the weight of a gene may be much higher than that of other member genes, the ranking of pathways results from the synergistic effect of all member genes. Therefore, finding the genes physically close to the highest weighted gene in the PPI network is another way to understand which part of the pathway is involved with the disease under study. In this chapter, the PPI data are collected from STRING [202], and MCL algorithm [203] was used to cluster them. The clusters are visualised in Figure 5.5 using Cytoscape software [204], with the size of nodes being proportional to the weight identified by DIGS and the edges proportional to the interaction confidence score. In principle, proteins in the same cluster either are part of the same protein complex or perform a similar molecular function.


pancreatic secretion

chemical carcinogenesis

Fig. 5.5 Protein-protein interaction network for pancreatic secretion and chemical carcinogenesis. The Protein-protein interaction networks are retrieved the protein-protein interaction network from the Stringdb for the proteins of each pathway. Proteins are clustered using MCL algorithm. The size of nodes is proportional to the weight identified by DIGS and the edges are proportional to the interaction confidence score.

For the Pancreatic secretion pathway, the gene with the highest weight, CA2, is the most active enzyme found in nature [205]. The cluster showed that CA2 communicates very closely with many SLC family member genes in the PPI network. Therefore, the role of SLC family genes is essential for the high ranking of the Pancreatic secretion pathway. In contrast, the SLC gene family was found to be highly correlated with breast cancer [206]. For the Chemical carcinogenesis pathway, the previously described GSTM family genes are abundantly present in the GSTM5-centered cluster. The ADH and ALDH gene families also presented in this cluster are risk factors for many types of cancer and have abnormal activity in stage IV breast cancer [207].

SHapley Additive exPlanations (SHAP), which is a unified framework for interpreting the predictions of ML methods by assigning importance values to the input features [208], also verify these DIGS2-identified important pathways. This work applies the Tree explainer from the python package SHAP [209] on our trained RF classifier. The results of SHAP provide the pathways with the highest impact on the prediction accuracy of the RF classifier. The top 30 pathways are presented in Figure 5.6; the top 10 pathways for each molecular subtype are presented in Appendix. Significant pathways identified by the DIGS2 model overlap highly with pathways that significantly impact the RF classifier. Over half of the top 30 pathways from SHAP overlapped with the top 70 pathways of the DIGS2 pathway ranking list.

The tSNE dimensionality reduction scatter plots and Hierarchical clustering heatmaps were created using the RNA count data and pathway activity data. In Figure 5.7, the two tSNE plots show how the pathway-level data can cluster the different sample labels compared to the gene-level RNA counts data. From the pathway activity scatter plot, it is clear that the subtypes with large samples, Luminal A, Luminal B and Basel, are well clustered. The heatmap using only significant pathways also verified that. The same analysis is repeated for the remaining pathway activity inference methods and COAD dataset (Appendix B).

5.5 Discussion and Conclusions

This chapter presents a new optimisation model, DIGS2, which is formulated to contain a much smaller number of binary variables than the DIGS model. DIGS2 has better solution quality than DIGS and outperforms several widely used models in binary classification, multi-class classification, survival analysis and robustness. Another outstanding contribution of DIGS2 is the nature of independent and in-depth insight into each pathway. Using the optimisation solution outputs, DIGS2 can calculate the prediction accuracy of individual pathways and provide rankings for the pathway

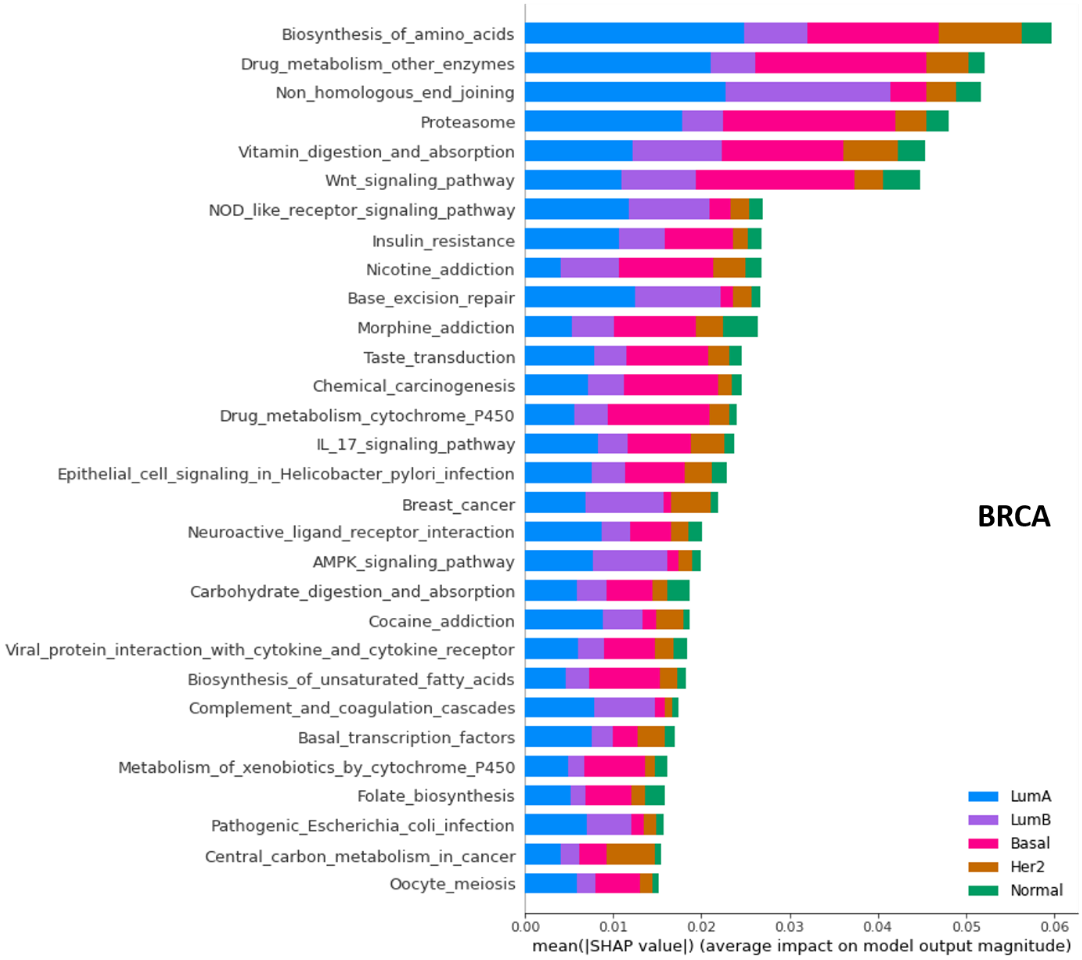


Fig. 5.6 Top 30 pathway from SHAP algorithm for BRCA.

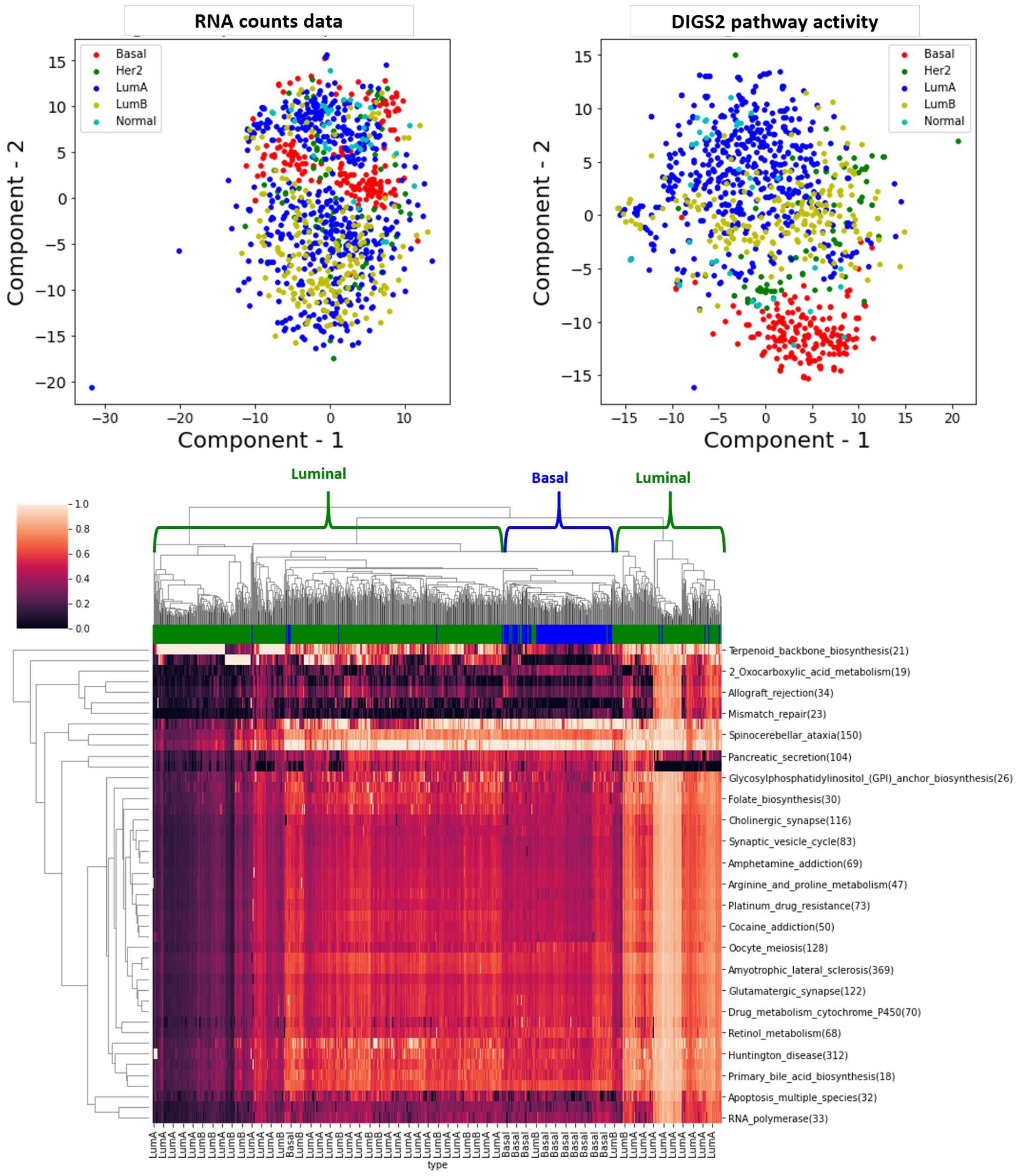


Fig. 5.7 tSNE dimensionality reduction scatter plot and Hierarchical clustering map of significant pathways in BRCA. Pathway activities are derived with the DIGS2 model using all samples. The left scatter plot is generated from the RNA count data with all KEGG genes; the right scatter plot is generated from the KEGG pathway activities derived from the DIGS2 model using all samples. The hierarchical clustering heatmap is developed with the pathway activities of the top 70 pathways of DIGS2.

member genes. In addition, the pipeline for analysing the individual pathway is standardised to formulate the NaiveDIGS method.

The biological interpretation ability of the DIGS2 model is verified in the breast cancer RNA-Seq dataset. DIGS2 ranked the pathways according to their degree of differentiation of disease subtypes and ranked the pathway member genes according to the gene weights inferred by the model. These deduced pathways and genes are validated in multiple ways: The top-ranked pathways are found to be related to the pathology of breast cancer in literature. These top pathways are found highly overlapped with the pathways identified by SHAP. The clustering using the DIGS2 inferred gene weights on the PPI connection network further verify the inferred gene weights play an important role in indicating the important genes.

In summary, the new proposed method DIGS2 is more robust and powerful than other pathway activity inference methods. In the next chapter, the performance of the DIGS2 model on single-cell RNA-Seq data is explored.

Chapter 6

Exploration of the Pathway Activity Inference Model on Single-cell RNA-Seq Data

The single-cell RNA-Seq (scRNA-Seq) technology offers a more detailed view of transcriptome, allowing to decipher the contribution of single cells to disease dynamics. However, the scRNA-Seq data generally suffer from high noise and sparsity, as heterogeneity exists among individual cells. Although methods and approaches designed for analysing bulk data may be unsuitable for single cell, some of these bulk methods have been shown to perform well on scRNA-Seq data. Therefore, it is worth seeing how the pathway activity inference method proposed in Chapter 5 can be applied to analyse scRNA-Seq data.

This chapter employs the pathway activity inference method DIGS2 to three scRNA-Seq datasets. The pathway activity values are expected to achieve three objectives. The first is identifying the cell types, e.g. B cells, T cells and monocyte cells, in human peripheral blood. The second is identifying breast cancer tumour subtypes using the tumour cells from breast cancer patients. The last is using the pathway activity values to integrate the scRNA-Seq datasets. As for evaluating the performance of the DIGS2 model, in addition to the traditional way of training machine learning classifiers and accessing prediction accuracy, this chapter also uses UMAP projection to assess how the pathway activity values can cluster the cells of the same labels. DIGS2 showed great performance on scRNA-Seq data when comparing it with the baseline method GSVA. Also, interesting results appeared when using DIGS2 for disease-related significant pathway identification.

6.1 Introduction and Background

The development of single-cell RNA-sequencing (scRNA-Seq) technologies has dramatically advanced our understanding of cellular states. Significantly, the scRNA-Seq technology reveals the heterogeneity of cellular populations at unprecedented resolutions [210–212]. This new technology is described in Section 2.1 in detail and is briefly summarised here to highlight how it differs from bulk technologies.

it is well established that what is observed at tissue level results from many interactions of several different single cells. Thus, the observable "average" cannot represent the "average cell". Instead, it is the resulting behaviour of the actions and interactions of many different cells. The appearance of scRNA-Seq enables the comparison of the transcriptomes of individual cells. Compared with the bulk technologies, the scRNA-Seq benefits research in two aspects. First, it allows addressing medical questions such as the role of cell populations contributing to disease progression and therapeutic potential. Also, it provides an understanding of context-specific dependencies, namely the behaviour and function that a cell has in a specific context, which can be crucial to understand some complex diseases, such as diabetes, cardiovascular disease and cancers [20]. In summary, scRNA-Seq analysis enables the uncovering of more refined and novel cell clusters [213], which have greatly advanced understanding of cellular states.

Although methods and principles previously developed for bulk data can be used, scRNA-Seq data analysis poses several unique challenges that require new strategies. In detail, scRNA-Seq data has characteristics such as drop-out events and low library sizes, consequently causing the data to suffer from low gene coverage, namely data sparsity. As a result, the robustness and applicability of the methods designed for analysing bulk gene expression data, to some extent, might be affected [214, 160]. From this perspective, pathway-level analysis also shows its power on this challenging data. Pathways are composed of a limited number of genes, so the proportion of missing genes can be reduced and low coverage difficulties can be avoided.

Many researchers have focused on incorporating prior gene set knowledge with the scRNA-Seq data. Specifically, [160] focuses on the possibility of applying the pathway analysis tools established for bulk sequencing data to scRNA-Seq data in a meaningful way. In this work, they investigated the performance of bulk data analysis tools (PROGENy [215], DoRothEA [216] and GSEA analysis) and tools designed for scRNA-Seq data (SCENIC/AUCell [217] and metaVIPER [218]) on simulated and real scRNA-Seq datasets. Their experimental results led to two conclusions. First, the bulk-based pathway analysis tools can be applied to scRNA-Seq data and partially outperform dedicated single-cell tools. Second, they further proved that the pathway analysis tools are more sensitive than statistical approaches.

Besides focusing on pathway analysis, many researchers also discussed the application of pathway activity inference on scRNA-Seq data. [218] uses the pathway activity values inferred using the Gene Ontology biological process gene sets to align the scRNA-Seq profiles of human and mice. They used a GSEA-based pathway activity inference method (see Chapter 2.4.2) to calculate the pathway activity values for humans and mice separately, then concatenate the two pathway activity profiles to see how the different cell types cluster on the joint dataset. Their results proved that transforming the gene-level data into pathway activity values can produce a dataset less influenced by common technical noises in scRNA-Seq profiles. Moreover, the transformed data preserve cellular state integrity and transitions.

Another work evaluates the performance of widely used pathway activity inference methods on scRNA-Seq data [219, 220]. This work follows the typical pathway activity inference and validation framework described in Chapter 2. More specifically, they concentrated on the performance of these tools on their ability to dissect meaningful cellular heterogeneity, which is expected to be retained in reduced dimensionality space. In other words, their research is answering whether the pathway activity values produced by these tools can classify cell types through either supervised or unsupervised classification. They used methods designed for bulk data, including methods such as GSVA, ssGSEA, and z-score, described or mentioned in Chapter 2.4, in their comparative study. Their study found that pre-processing of scRNA-Seq data (gene filtering or normalisation) is essential to obtain the desired performance, regardless of whether the method is designed for bulk or single cell data.

It can be concluded from the above background information that the analysis of scRNA-Seq data needs to incorporate functional biological pathways. Moreover, it is worth exploring whether pathway activity methods designed for the bulk data suit scRNA-Seq data. Therefore, this chapter discusses the adaptation of the DIGS2 model on scRNA-Seq datasets. The experiments in this chapter follow data collection, model implementation, cross-validation, and results evaluation. Because of the specificity of the scRNA-Seq data, there are some differences in the implementation pipelines compared to previous chapters.

First, the scRNA-Seq data provide unique analysis opportunities that allow novel biological discoveries, including the identification and characterisation of cell types and the study of their organisation in space. Therefore, a few experiments in the chapter use the DIGS2 model to separate cell types of healthy human peripheral blood mononuclear cells (PBMCs) . This part can also be used to validate that DIGS2 can analyse scRNA-Seq data. Then, the DIGS2 model is applied to breast cancer to discuss how the pathway activity values inferred by DIGS2 perform on separating the molecular subtypes of cancer. Second, concerning the evaluation of the inferred pathway activity values, a new metric is added to show the performance of clustering cells using pathway activity. In the general analysis pipeline, visualising low-dimensional representations with scatter plots of data is a step in analysing single-cell genomic data [221]. By feature selection, significant genes are selected and passed into dimensionality reduction algorithms, such as PCA, tSNE [222], and UMAP [223], to produce two-dimension vectors for scatter plot visualisation. Therefore, comparing the low-dimensional visualisation of pathway activity data with the gene expression data is meaningful for understanding how pathway activity boosts the analysis of scRNA-Seq data. (Figure 6.1)

This chapter explores the efficiency of pathway activity inference in analysing scRNA-Seq data using the DIGS2 model. Also, pathway activity inference is used not only for classifying disease subtypes but also for classifying the cell types. In parallel, the goal of extracting meaningful biological interpretations from DIGS2 results is preserved.

6.2 Application of optimisation-based pathway activity inference for dataset integration, cell clustering and breast cancer subtype prediction on scRNA Seq

6.2.1 Single-Cell RNA-Seq dataset acquisition

Using benchmarking datasets is essential for analysing scRNA-Seq data. Therefore, two typical scRNA-Seq datasets of human peripheral blood mononuclear cells are used in this work [224]. These datasets are available from 10X Genomics (https://www.10xgenomics.com/resources/datasets). The first dataset, refer to as PBMC 3k, contains 2700 immune cells. After filtering the cells and genes with too few reading counts, the dataset consists of 2,638 cells across 32,738 genes. The cells are allocated into nine subsets, Naïve CD4 T cells, Memory CD4 T cells, CD8 T cells, B cells, CD14+ Mono cells, Mono cells, DC cells, Platelet cells and Natural Kill cells. Two (DC and Platelet cells) of the nine cell types are removed from following analysis because of the highly unbalanced cell amount compared with the other seven cell types. The second PBMC dataset, refer to as PBMC 10k, has 10990 cells and 20292 genes after pre-processing. Cells are annotated into seven subsets: B cells, DC cells, Mono/Macro cells, Platelets cells, T cells, Duplicated and Unknown cells. The PBMC 10k dataset will be used for dataset merging, which is

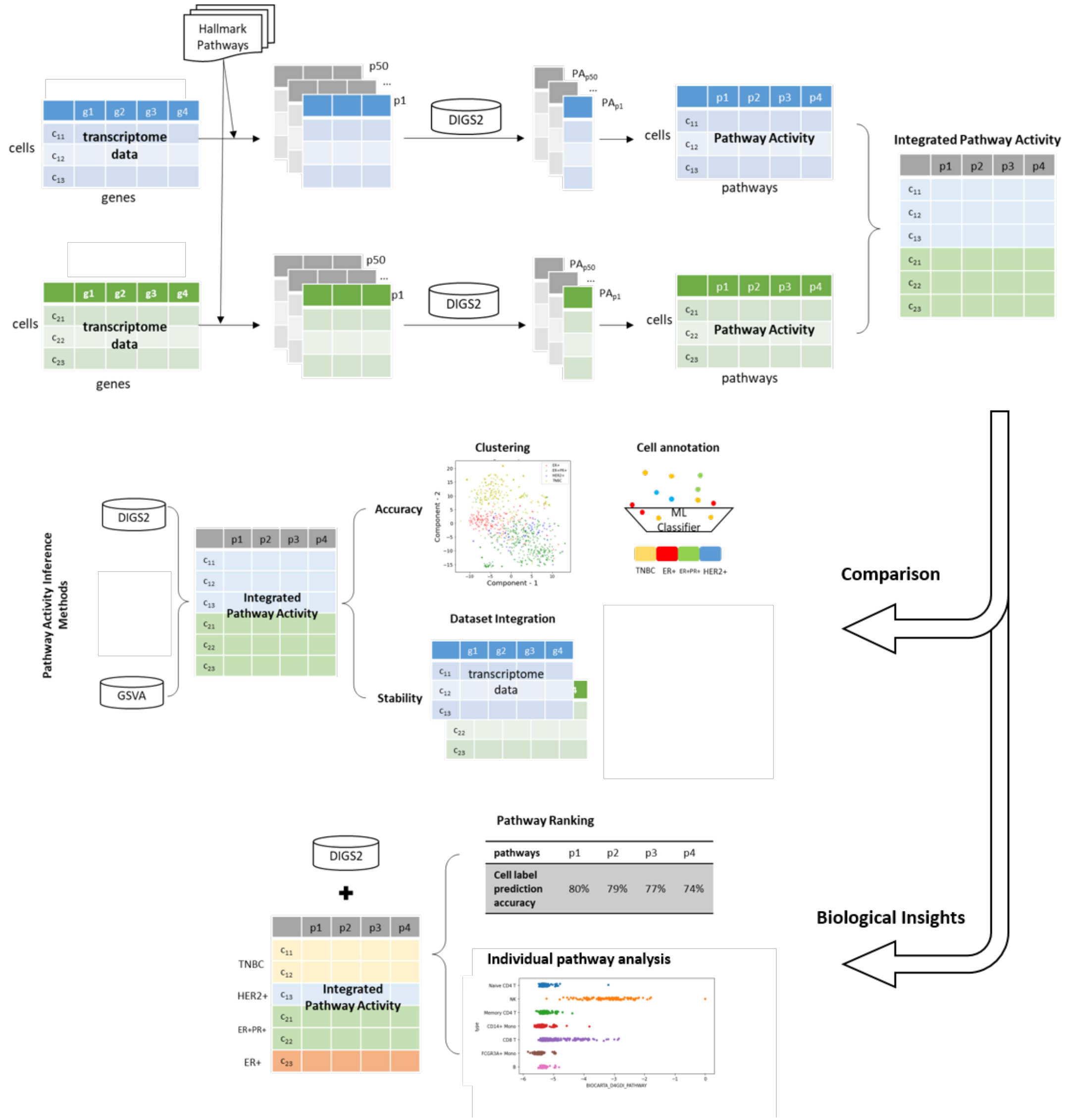


Fig. 6.1 Pathway activity implementation for scRNA-Seq data.

illustrated in Section 6.2.7. Therefore, only the three main cell population are kept for this work (B cell, T cell and Mono/Macro cell).

Except baseline datasets, this chapter also explores using scRNA-Seq data to analyse the breast cancer subtypes. The immune cells in the tumour microenvironment is essential for understanding the mechanism of immunotherapy response and pathology of breast cancer. Therefore, I downloaded the scRNA-Seq data of diverse immune phenotypes in breast cancer tumour microenvironment (refer to as the Azizi dataset) from the GEO data repository accession number GEO: GSE114725 [225]. The structure of this dataset is as follows: (i) this dataset contains 45,000 immune cells; (ii) these cells come from four tissues of eight breast cancer patients, breast carcinomas (TUMOR), normal breast tissue (NORMAL), blood (BLOOD) and lymph node (LYMPHNODE); (iii) the cells are annotated by seven cell types, Neutrophil cell, T cell, myeloid cell, Natural Kill T cell, Mast cell, B cell Natural Kill cell; (iv) the eight patients are annotated with four kinds of breast cancer subtypes, ER+PR+, ER+, HER2+ and Triple Negative (TNBC). Only the B cells (3,890 cells) in the Azizi dataset are used in this work, as the B cell population in the Azizi dataset has the cleanest clusters between TUMOR tissue and BLOOD tissue shown in Figure 6.2c. To perform different experimental aims, I applied corresponding sampling strategies for these three scRNA-Seq datasets, further illustrated in the following sections. A summary of the datasets after sampling is shown in Table 6.1.

6.2.2 Biological pathway acquisition

This chapter looks into the biological processes in single cells, which means more precise intra-cellular processes can be obtained than the bulk data. Therefore, except the large size pathways that describe the complete biological processes, smaller pathways that focus on more detailed procedures in a biological process are also used in this work.

Molecular Signature Database (MSigDB) groups the pathways into different collections. I selected two collections from MSigDB v2022.1 [66]. Hallmark gene sets collection (4,383 unique genes) is an initial release of 50 hallmarks which condense information from over 4,000 original overlapping gene sets from all the C1 to C6 collections of MSigDB v4.0. These gene sets summarise and represent specific, well-defined biological states or processes and display coherent expression. Each gene set in this collection contains around 200 genes. Another collection, the BioCarta gene sets collection (292 pathways, 1,509 unique genes), is also downloaded. The BioCarta collection contains canonical pathways gene sets derived from the BioCarta pathway database. This

Table 6.1 Summary of the scRNA-Seq datasets

Dataset	Coarse cell type	Cell type	Cell Amount
PBMC 3k	B cell	B cell	344
	Mono/Macro	CD14+ Mono	480
		FCGR3A+ Mono	162
	T cell	CD8 T	279
		Memory CD4 T	472
		Naive CD4 T	711
	NK cell	NK	144
Sum			2592
PBMC 10k	B cell	B cell	1485
	Mono/Macro	Mono/Macro	2890
	T cell	T cell	5159
Sum			9534

Dataset	Coarse cell type	Cell type
Azizi	ER+	203
	ER+PR+	470
	HER2+	78
	TNBC	378
Sum		1129
Azizi(B cells)	Tissue type	Cell Amount
	BLOOD	1486
	TUMOR	1128
Sum		2614

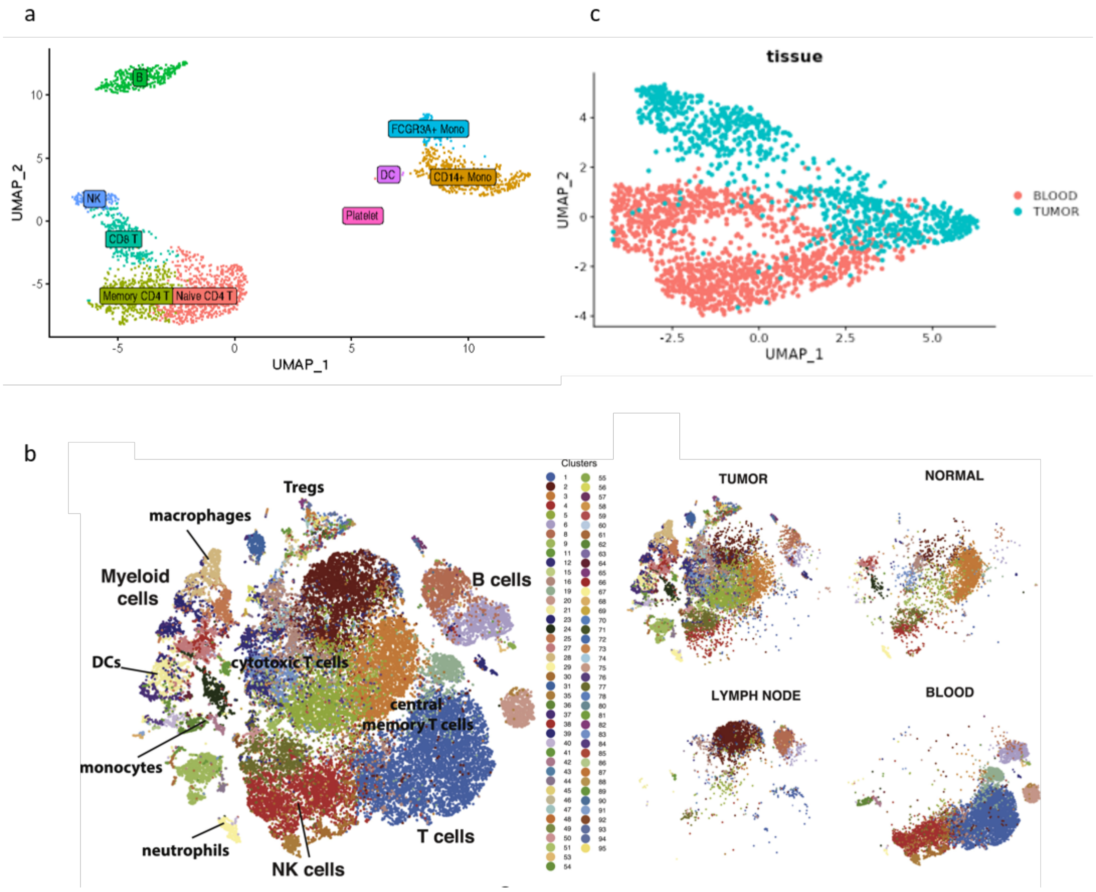


Fig. 6.2 Overview of the Azizi dataset and PBMC 3k dataset. (a) UMAP projection of PBMC 3k cell populations between nine cell types (b) UMAP projection of Azizi B cell populations between four tissues (c) UMAP projection of Azizi B cell populations between tissues: BLOOD and TUMOR. Figure is taken from [225]

collection highlights the common metabolic pathways, signal transduction pathways, and other biochemical pathways. The size of the BioCarta pathways is in the range of 10 to 50 genes, which are relatively small size pathways among various pathway collections.

6.2.3 Pathway activity inference methods

This chapter mainly discusses the application of the DIGS2 model for pathway activity inference on scRNA-Seq datasets. DIGS2 is an optimisation-based supervised pathway activity inference method, defined by a mixed linear programming (MILP) model that can be solved to global optimality using some of the standard algorithms like branch-and-bound [113]. Chapter 5 has provided that DIGS2 achieves high-quality solutions and is competitive with other pathway activity inference methods on RNA-Seq datasets. The DIGS2 model is implemented using general-purpose solution algorithm CPLEX in the General Algebraic Modelling System (GAMS) [127]. Time limit is set as 200 seconds.The Relative Optimality criterion solver (optcr) is set to its default value of zero.

A baseline method GSVA is used for comparison, which has been used in other studies for pathway activity inference on scRNA-Seq data. Gene Set Variation Analysis (GSVA) [100] is selected as the baseline method for comparison. As a variation to the GSEA [40], GSVA extends the pathway enrichment scores by calculating the sample-wise enrichment score. More details can be found in Section 2.4.2. GSVA is implemented using the Python package decoupler [226].

6.2.4 Pathway activity inference implementation

Four experiments of increasing complexity were applied on the PBMC 3k datasets and Azizi datasets to establish performance comparisons.

Separating seven immune cell types on PBMC 3k data. As a baseline dataset for scRNA-Seq analysis, the cells in PBMC 3k are well annotated with clear clusters and cell types (Fig 6.2a). Therefore, this experiment is designed to calculate the pathway activity values using the DIGS2 model in the PBMC dataset to compare the official results from 10X Genomics with the pathway activity values. From the perspective of dimension reduction, losing information is unavoidable when compressing high-dimensional data into low-dimension. However, the calculation behind the pathway activity contains a biological interpretation of the high-dimensional features (genes).

Consequently, the different biological processes behind the different cell types make it possible for the dimension reduction from gene-level to pathway-level to minimise the information loss and give similar clustering results.

Separating the BLOOD and TUMOR tissues on Azizi data. This experiment is designed to look into the differences between different tissues (blood and tumour). According to the UMAP plot for tissues (Fig 6.2b), BLOOD and TUMOR are the two most clearly separated clusters among four kinds of tissues. Also, from the UMAP plot for cell types (Fig 6.2c), B cells are well clustered with compact intra-group distance. Therefore, BLOOD B cells and TUMOR B cells are chosen for analysis. As there is a mixture of other types of cells (e.g., T cells, myeloid cells) in Azizi original cell annotations, I extracted the B cells from the raw data and got a cleaner separation between BLOOD B cells and TUMOR B cells (Fig 6.2c). After pre-processing, there are 1336 cells in the BLOOD and 944 cells in the TUMOR.

Separating four breast cancer subtypes on Azizi data. Next, I look into how pathway activity can separate the different breast cancer subtypes, which is a much more complex task compared with the separation of cell types. The eight breast cancer patients of Azizi datasets consist of four kinds of subtypes, ER+PR+, ER+, HER2+ and Triple negative breast cancer (TNBC). Only the TUMOR cells in Azizi dataset are used for this experiment. Since the number of TUMOR cells (21,253 cells) is too large, which could affect the computation efficiency, I did stratified sampling for the seven cell types and kept 500 cells for each subtype. The final input dataset contains 1,998 cells with the same amount of cells for each subtype.

Separating TNBC type breast cancer versus other subtypes on Azizi data. To further investigate the pathway activity values for the most aggressive type of breast cancer, I design this experiment using the same Azizi TUMOR dataset as the former one. In this experiment, the 1,998 cells are marked as TNBC and nonTNBC with 499 and 1,499 cells, respectively. TNBC, characterised by estrogen receptor (ER/ESR1)-negative, progesterone receptor (PR/PgR)-negative, and epidermal growth factor receptor 2 (HER2/ERBB2)-negative, is known to have high mutational burden. Among various breast cancer subtypes, TNBC is highly aggressive with a generally poor prognosis [227, 228]. The characteristics of four breast cancer subtypes are summarised in Chapter 5 Table 5.2. The pathway activity for distinguishing TNBC from the other three subtypes is expected to provide specific biological insights that can contribute to understanding the pathology and prognosis of TNBC.

In summary, the first two experiments focus on the widely discussed question for scRNA-Seq data, i.e. the identification of cell types. The following two experiments focus on how the scRNA-Seq data can provide biological insights into identifying breast cancer subtypes. To gain robust results, for each experiment, 5-fold cross-validation is applied. The DIGS2 model is trained on the training sets to calculate the training pathway activity profiles. Then the gene weights generated from the cells in training set are used on testing cells to compute the testing pathway activity profiles. For GSVA, the pathway activity calculation is conducted separately for the training and testing sets. Therefore, for each method, five pairs of training and testing pathway activity profiles are generated. In the whole pipeline, the testing cells are kept blind to the model training process. The data split from cross-validation are kept constant in the pathway activity calculation for both methods.

6.2.5 Pathway activity evaluation criteria

The first pathway activity evaluation criterion is dimension reduction. Dimension reduction is a common step in the standard analysis pipeline for scRNA-Seq, used to visualise the data. In the standard pipeline, datasets used for dimension reduction are already filtered to keep only the genes that have high variability; often, the number of genes after filtering is between 200 to 2,400 [229]. After gene selection, the dimension of the dataset is further reduced by dimension reduction approaches to two or three dimensions, which are combinations of the original features. The two or three dimensions are used for visualisation. Therefore, in this work, comparing the visualisation of the highly variable genes and the visualisation of pathways can be used to assess the performance of pathway activities. I use Uniform Approximation and Projection method (UMAP) [223] to achieve the reduced two dimensions to visualise the gene expression profiles and pathway activity profiles for each dataset. Then the averaged silhouette width [230] across all cells was used to evaluate the performance of dimensionality reduction for each dataset. The averaged silhouette width score is implemented by the python package sklearn [172].

The second evaluation criterion is sample label prediction (i.e. cell types for the first experiment, tissue types for the second experiment and cancer subtypes for the last two experiments). Since the appearance of pathway activity inference for high-throughput profiling data, precise disease phenotype prediction has been the main target to be achieved [53]. Machine Learning classifiers verify the prediction performance of pathway activity values. In this work, I used the Random Forest (Python package scikit-learn, n_trees set as 200) to evaluate the prediction accuracy of the pathway activity values produced by the two methods (DIGS2 and GSVA). The AllGENE method used in Chapter 5, which refers to the direct use of gene expression values to train and test

KNN classifiers, is also used in this chapter to show the prediction accuracy using gene-level data.

6.2.6 Single-cell RNA-Seq dataset integration

Except using the pathway activity values to cluster the cell populations and make predictions on cell labels, many researchers investigate merging the scRNA-Seq datasets using pathway activity values. Although individual experiments have expanded understanding of the properties of cell types, obtaining a comprehensive understanding of healthy and diseased cells requires integrating multiple datasets across datasets [231]. A study [232] uses the pathways activity values to do the cross-species integration for human and mouse scRNA-Seq datasets. Another study [233] did the integration of the different technologies (snDrop-seq and scTHS-seq) for human adult brain cells. Their results show that, when datasets integration are made, sparse scRNA-Seq datasets can be as informative as bulk sequencing of different cell populations, thus providing richer information for further analysis.

Figure 6.3 illustrates the aim for scRNA-Seq dataset integration. Because batch effects exist among different datasets, the initial dimension-reduced visualisation (Figure 6.3a) tends to cluster the cells of the same datasets together. However, after integration, the dataset borders disappear, and the cells are clustered by their cell types (Figure 6.3b). The overall purpose of scRNA-Seq dataset integration is to eliminate the batch effects of different datasets and cluster cells according to their type.

I primarily attempt to integrate the PBMC 3k and PBMC 10k datasets using their pathway activity values in this work. PBMC 3k and PBMC 10k datasets consist of the human peripheral blood mononuclear cells. Therefore, they have similar cell types and are comparable to each other. The integration process follows three steps: (i) calculating the pathway activity values for each dataset separately; (ii) scaling the pathway activity values in the range of 0 to 1; (iii) connecting the two pathway activity matrices and visualising it using UMAP.

DIGS2 is used to train the pathway activities with the Hallmark pathway collection. For PBMC 3k, the pathway activity values obtained from the experiment of Separating seven immune cell types (Section 6.2.4) are used. For PBMC 10k, three cell types (listed in Table 6.1) are used for training the pathway activity. As the size of the PBMC 10k dataset was too large and would affect the calculation efficiency, stratified sampling was applied to select a subset that contains 2,000 cells. The proportion of each cell type in the selected cells was consistent with the original dataset.

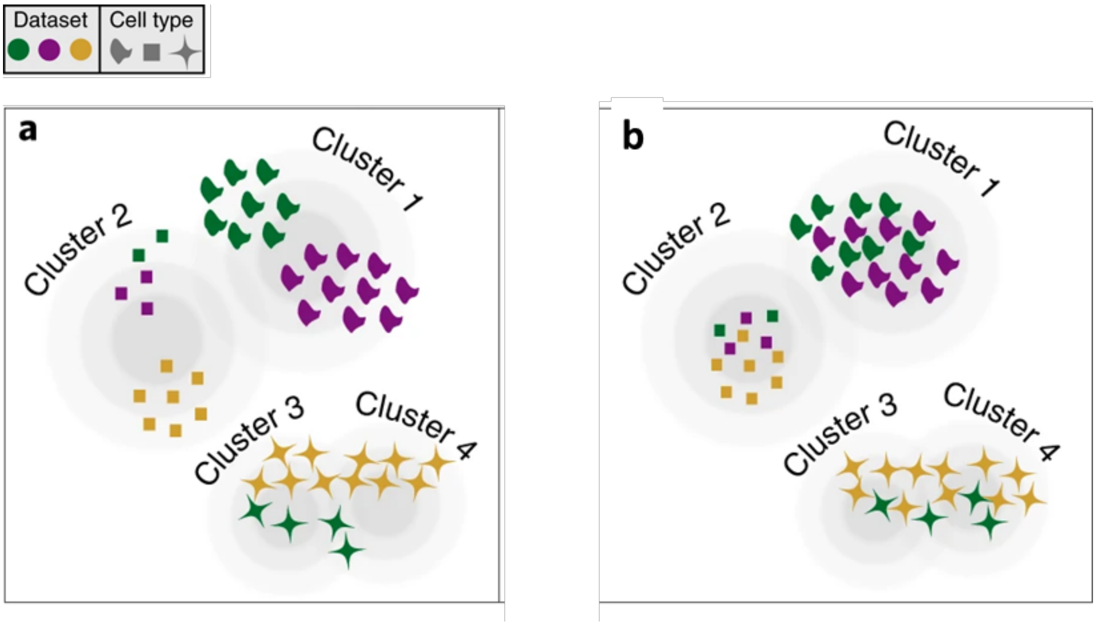


Fig. 6.3 Principle for integrating the scRNA-Seq datasets. (a) shows cell clusters before integration. (b) shows the cell clusters after integration. The purpose of integration is to eliminate the dataset borders and cluster cells by their cell types. Figure taken from [231]

The pathway activity profiles produced individually on the two datasets are then merged into one dataset. The integrated dataset consists of Hallmark pathways across cells from the two datasets. After normalising the pathway activity values to a 0-1 scale, the integrated pathway activity dataset is projected using UMAP into two-dimension representations for visualisation.

6.3 Comparison Results

The pathway activity profile is expected to preserve the integrity of cellular states and their transitions while increasing the separation degree between different cell types or disease types. Therefore, the three key results of the experimental studies on scRNA-Seq datasets in this work are (i) using pathway activity values to separate the cell types, (ii) using pathway activity values to separate the cells from different tissues and (iii) using pathway activity values to separate the breast cancer molecular subtypes. The performance of the inferred pathway activity is evaluated by the cell clustering and cell label prediction.

6.3.1 Clustering of cell types in PBMC 3k

Seven class separation is a complex task. Most pathway activity inference methods are designed for binary class separation, as shown in Table 2.3 and only a few of them are designed for multi-class tasks. Moreover, the most frequently used disease that the pathway activity inference methods are applied to, i.e. breast cancer, usually involves four or five subtypes. Therefore, applying pathway activity inference for separating seven classes is an advance in methodology.

Presentation of how the pathway activity values perform on separating the seven cell types in healthy human peripheral blood mononuclear cells (PBMC) is shown in Figure 6.4. The clustering of cells using pathway activity values produced with identical cell populations show significant differences between DIGS2 and GSVA. The edges of the clusters for DIGS2 are much more well-defined than GSVA, and almost no overlap between clusters is observed. The averaged silhouette score also reflects the improvement achieved by DIGS2. DIGS2 gets higher averaged silhouette scores for both pathway collections compared with GSVA. This observation proves that DIGS2 outperforms the widely used pathway activity inference methods GSVA and shows its power on a challenging problem.

Another fact that can be observed is that the cell clustering performance is better using Hallmark pathways (Figure 6.4a and 6.4b) than BioCarta (Figure 6.4c and 6.4d). Considering that pathways for irrelevant biological processes may not contribute to cell type separation but increase the noise instead, top pathways were selected according to the pathway ranking produced by the individual pathway analysis (see Chapter 5 Figure 5.1). The top 50 BioCarta pathways were used to plot the scatter plot figures. The higher the ranking, the more powerful the pathway in separating cell types. However, BioCarta pathways still suffer from the lower average silhouette score under such conditions. The possible explanation is the number of unique genes in the pathway collection and the size of each pathway. The number of unique genes in the Hallmark collection is twice as large as in the BioCarta collection, and the size of individual pathways is approximately ten times larger than in the BioCarta collection. Therefore, the information abundance is limited by the number of genes included in BioCarta pathways. Consequently, the performance of pathway activity is poorer than the pathway collection that contains a larger quantity of genes.

In the next step, comparisons are conducted between the UMAP visualisation generated from gene expression values (Figure 6.2a) with the UMAP visualisation of pathway activity values generated by DIGS2 using the Hallmark pathway collection Figure 6.4a). Although the datasets behind the plots employ different dimensions and biological

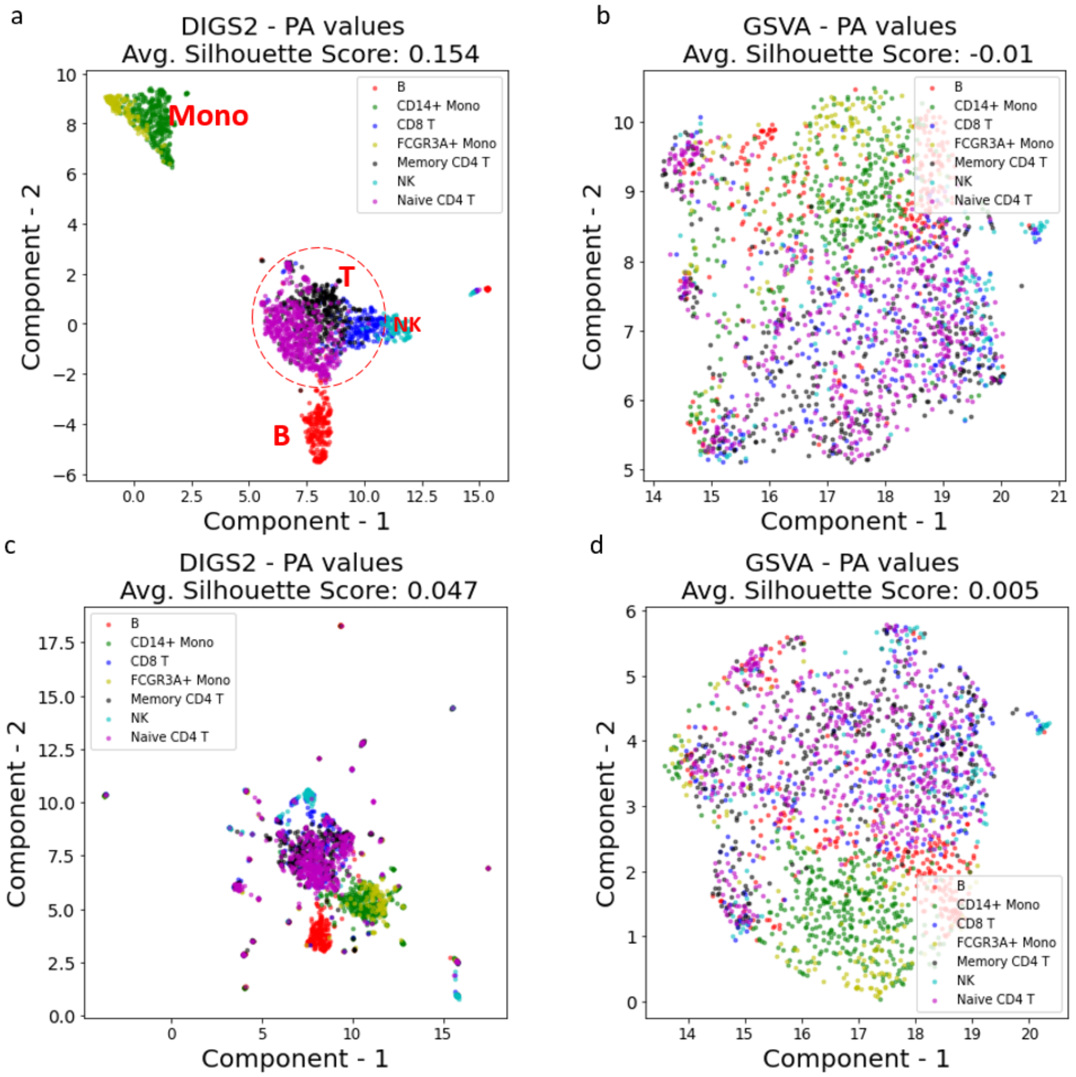


Fig. 6.4 UMAP projections for pathway activity values for PBMC 3k. The dimension of the pathway activity profiles calculated by DIGS2 and GSVA is reduced by UMAP and visualised as scatter plots. Each dot represents a cell. (a) and (b) are Hallmark collection pathway activity values, and (c) and (d) are BioCarta collection pathway activity values. The averaged Silhouette score is used to calculate the goodness of the clustering.

meanings, they show a similar relationship between the clusters. Specifically, the three types of T cells (CD8 T, Memory CD4 T and Naive CD4 T) are closely clustered together; The NK cell cluster is linked to the CD8 T cell cluster; B cell population and monocyte population separate in the plot away from the largest T cell population. In short, the distribution of the cell clusters generated by the best quality pathway activity profiles and gene expression profiles is identical. This fact provides important information that compressing the gene-level data into pathway-level data on scRNA-Seq data does preserve the integrity of cellular states. Therefore, results show the potential of applying pathway activity inference on scRNA-Seq data.

6.3.2 Clustering of tissues in Azizi B cell population

As the feasibility of pathway activity inference on the scRNA-Seq dataset has been demonstrated on the baseline dataset PBMC 3k, the next step shifts the focus to the analysis of the scRNA-Seq dataset for breast cancer (Figure 6.5). This section examines the discrimination of pathway activity to distinguish the same cell type (B cell) in different tissues (BLOOD and TUMOR).

The overall results of the performance for different pathway activity inference methods and pathway collections is the same as the separation of cell types on PBMC 3k data in the UMAP visualisations. In brief, DIGS2 performs better than GSVA and the Hallmark pathway collection performs better than the BioCarta collection. However, with the clustering task now only including two cell labels, the averaged silhouette scores increase substantially.

6.3.3 Clustering of breast cancer subtypes in Azizi TUMOR cell population

Previous results (i.e. using pathway activity separating cell types and separating tissue types) prove that larger pathways with more unique genes have better clustering performance for cell label separation than small-size pathways. Therefore, only Hallmark pathway collection is used for the much more challenging problem of separating the cells of different breast cancer subtypes. The UMAP visualisations for pathway activities produced by the DIGS2 and GSVA methods are presented in Figure 6.6.

In the clustering map of cancer subtypes (Figure 6.6a and 6.6b), GSVA could not find distinct cell clusters, and the cell labels were mixed. In contrast, DIGS2 produces better results. Although the cells of four subtypes overlap in the plot's middle area, the cells

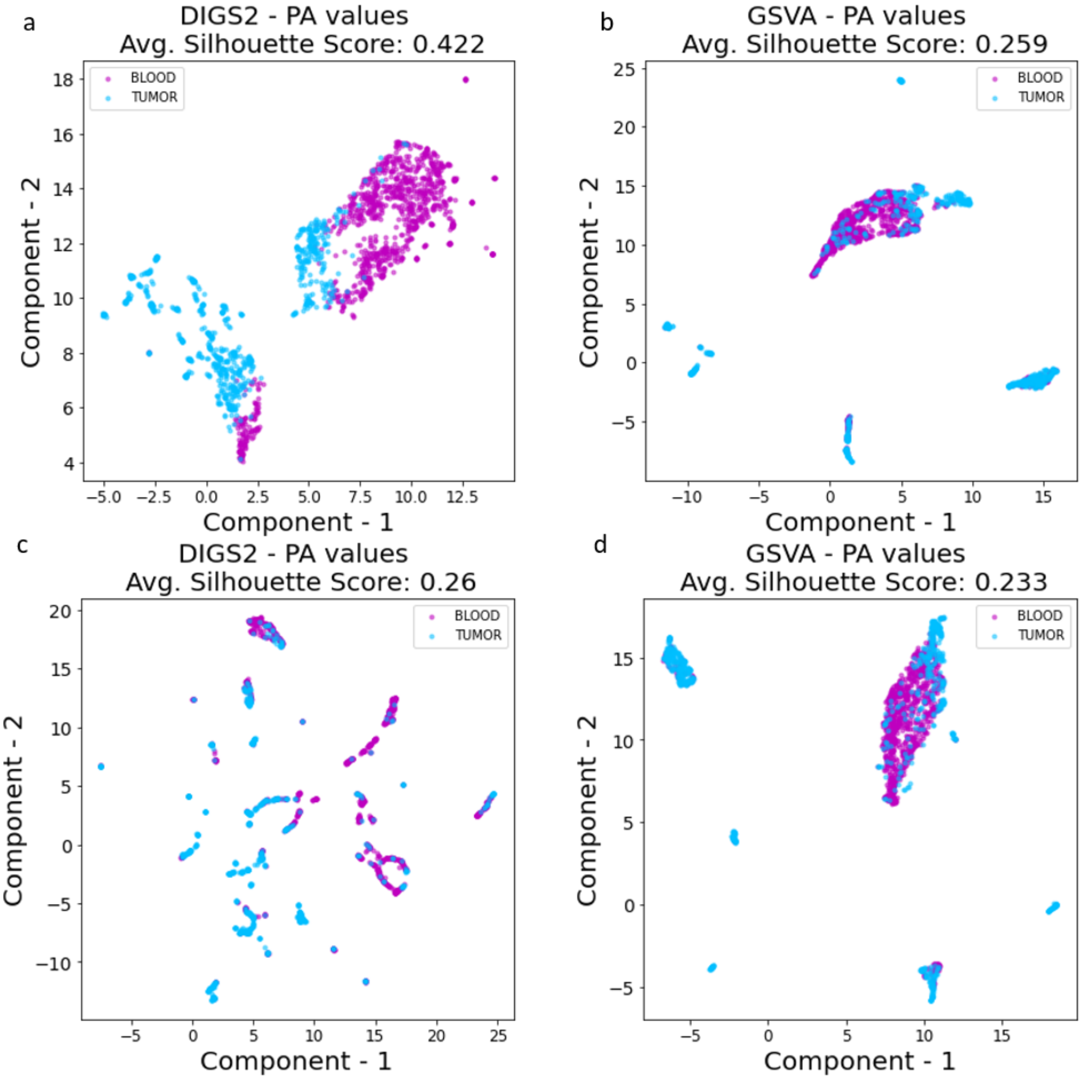


Fig. 6.5 UMAP projections for pathway activity values for Azizi B cell population. The pathway activity is calculated for cells in BLOOD and TUMOR from the B cells population in Azizi dataset. UMAP projections for two pathway activity inference methods and two pathway collections are visualised as scatter plots. (a) and (b) are Hallmark collection pathway activity values, and (c) and (d) are BioCarta collection pathway activity values. The averaged Silhouette score is used to calculate the goodness of clustering.

Table 6.2 RF prediction accuracy on the four experiments using Hallmark collection

Dataset	PBMC	Azizi		
Expriment	Cell Types(7 classes)	Tissue: BOOLD/TUMOR (binary)	Breast Cancer Subtypes (4 classes)	TNBC vs. Other subtypes(binary)
DIGS2	0.804	0.888	0.68	0.904
GSVA	0.536	0.86	0.652	0.888
AllGene	0.416	0.432	0.408	0.792

of different subtypes are clustered together. Moreover, two subtypes, ER+PR+ and TNBC, share a limited area with the other subtypes. Also, ER+ has a stand-alone area away from the mixed area. The condition is also reflected by the averaged silhouette score, with DIGS2 having a positive score and GSVA a negative one.

In particular, another experiment is conducted for the separation of TNBC versus the other three subtypes. As TNBC is the worst subtype of breast cancer, accurate identification of TNBC is beneficial. Figure 6.6c and 6.6d show the cell clustering results for this experiment.

Compared with UMAP projections for all four subtypes, both methods were significantly better at separating TNBC with other subtypes (Figure 6.6c and 6.6d). The two clusters are well separated in DIGS2, with a few numbers of cells overlapping. However, compared to another binary classification problem (separating TUMOR and BLOOD tissue), separating TNBC and the other three subtypes could not achieve the same mean silhouette scores. This observation also indicates the challenges faced by the identification of disease subtypes.

6.3.4 Cell label prediction accuracy

Following the traditional evaluation approach for pathway activity inference methods, using machine learning classifiers to predict sample labels is adopted to assess the quality of the pathway activity values inferred from scRNA-Seq data in this work. Random Forest (RF) is used for predicting the cell labels of all four experiments.

In addition to the two pathway activity inference methods (DIGS2 and GSVA), using the gene expression values to make the prediction is also included in this section to provide more information about how the pathway activity inference can boost the cell label prediction results. The gene-level classification approach is called AllGene, which uses all the unique genes of a pathway collection as the features to be trained by the RF classifier.

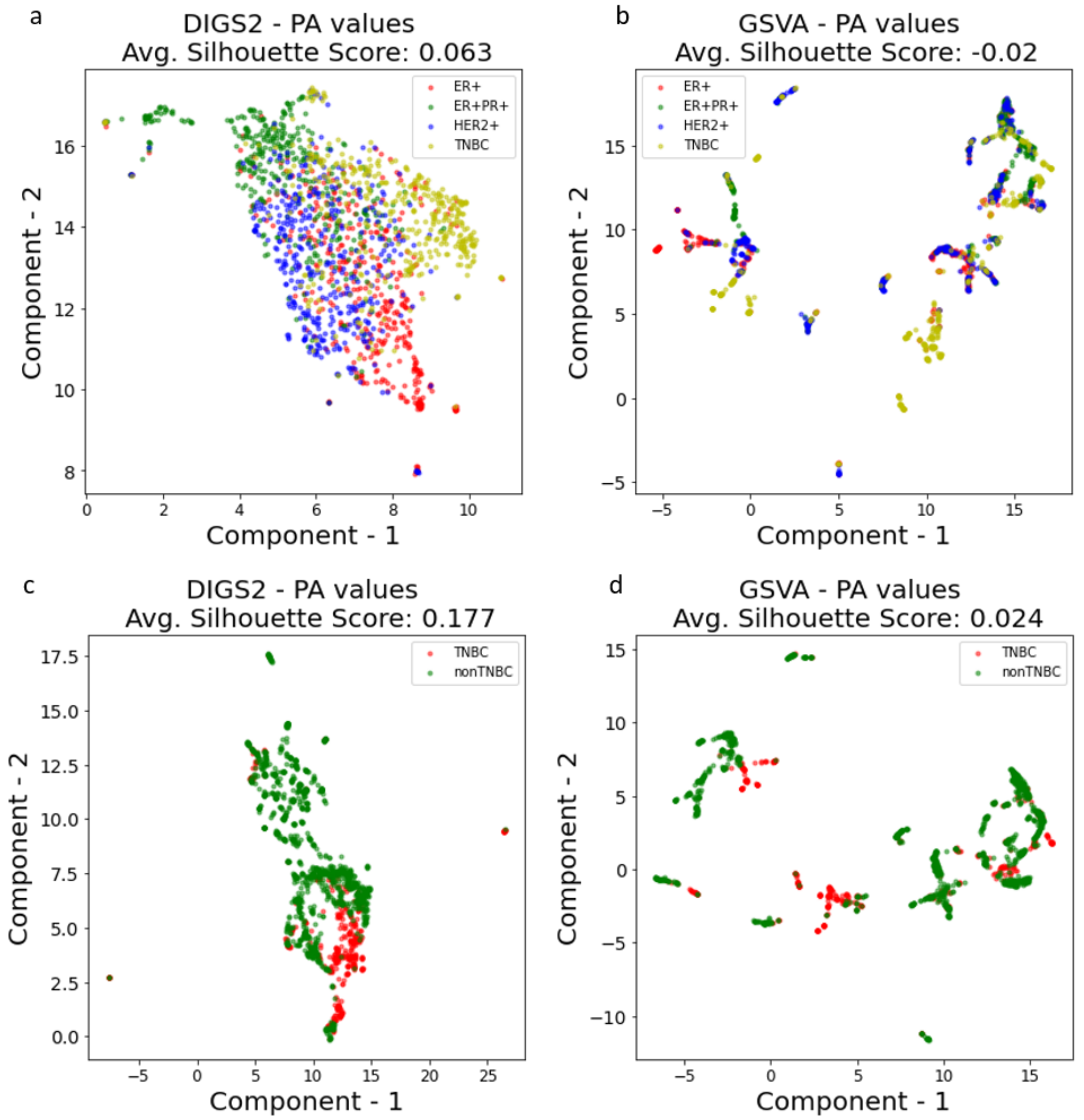


Fig. 6.6 UMAP projections for pathway activity values for Azizi TUMOR cells. (a) and (b) are the UMAP projections of pathway activity values calculated for four breast cancer molecular subtypes, ER+, ER+PR+, HER2+ and TNBC; (c) and (d) are the UMAP projections of pathway activity values calculated for the TNBC versus the other three subtypes. Hallmark pathway collection is used. All seven cell types of Azizi TUMOR population are used.

Table 6.3 RF prediction accuracy on the two experiments using Biocarta collection

Dataset	PBMC	Azizi
Expriment	Cell Types (7 classes)	Tissue: BOOLD/TUMOR (binary)
DIGS2	0.632	0.914
GSVA	0.556	0.898
AllGene	0.435	0.45

Table 6.2 and Table 6.3 present the results of the average accuracy over 5-fold cross-validation for each experiment and each method. The classifiers are trained on the pathway activity values/expression values of the training cells and tested on the testing cells. For every experiment, the data splits (e.g. 5-fold cross-validation) are done once for all the methods. In other words, the cells for training and testing are kept consistent. During the whole process, the testing data is kept blind to the training data. The prediction accuracy is calculated by the number of true positives divided by the number of all the cells. The final accuracy is the average number over 5-fold training and testing sets.

The first conclusion that can be derived is that the pathway activity methods increase the prediction accuracy to a large extent. Using individual genes to predict cell labels can hardly reach 50% accuracy. However, by altering and transferring the gene-level values to pathway-level values, the accuracies increased to a promising level, as the most accurate values of DIGS2 and GSVA in these two tables are over 80%. Secondly, DIGS2 outperforms the other two methods in all the experiments. Moreover, it is noticeable that the prediction accuracy of DIGS2 exceeds GSVA to a large extent in the experiment of separating cell types using Hallmark pathways (Table 6.2). In the use of Biocarta pathways, DIGS2 still keeps such improvement.

In summary, from the perspective of predicting the cell labels, pathway-level representation of gene expression values is superior to directly using gene-level expression values. This widely accepted concept for analysing bulk expression data is again proved in single cell expression data analysis. Also, compared with the baseline method GSVA, DIGS2 exhibits its power to achieve higher prediction accuracies. The significant improvements on the complex 7-class cell type prediction experiment provides confidence for the applicability of the DIGS2 method.

6.4 Biological interpretations

This section investigates the significant pathways and genes found by the DIGS2 method. The Individual Pathway Prediction Accuracy approach introduced in Chapter 5.4.1 (or see Figure 5.2) is applied to rank the pathways for each experiment in this work. The ranking of the constituent genes is implemented following the steps illustrated in Chapter 5.4.2.

6.4.1 Significant pathways and genes for identifying cell types in PBMC

The distribution of the pathway activity values for an individual pathway is shown in Figure 6.7 and Figure 6.8. Visualising the pathway activities using a strip plot (categorical scatter plot) provides an intuitive overview of how pathway activity values are differentiated for cells belonging to different classes. For example, in the Hallmark Allograft Rejection pathway strip plots, the NK cells and Mono cells (CD14+ Mono and FCGR3A+ Mono) are distinguished from the other cell types. The CD+ Mono cells are the most significant cell type in the Hallmark Apoptosis pathway. These results imply that the Allograft Rejection pathway acts differently in NK and Mono cells, and the Apoptosis pathway acts differently in Mono cells. The relationship between these two pathways and the corresponding cell types have been mentioned in the literature [234–237].

For the top 3 Biocarta pathways, related literature supporting the strip plot results can also be found. For example, [238] indicates that the expression of D4GDI in CD8+ T cells displayed a differential expression so that D4GDI could be involved in the functional differences between these cell subpopulations. Also, the caspase activation is well known for its relationship with Natural Killer (NK) cells [239–241].

For more information about the significant pathways and genes, Table 6.4 lists the top 5 pathways for Biocarta and Hallmark collections, with the top 5 constituent genes for each pathway. It can be seen that these top pathways from different pathway collections share many high-ranked genes, which also proves that the results produced by DIGS are kept consistent.

6.4.2 Significant pathways and genes for identifying breast cancer subtypes

Table 6.5 lists the set of Hallmark pathways and genes that are found as most discriminant with DIGS2. As the pathway activity for breast cancer cells is implemented using only the Hallmark pathway collection, the ranking of the pathways includes only the Hallmark pathways too. Interestingly, the top pathway for separating the four breast cancer subtypes, Hallmark Pancreas beta cells pathways, matches the top pathway in Chapter 5 Table 5.5, KEGG pathway Pancreatic secretion pathway. These two works are different in many aspects, i.e. Chapter 5 uses the bulk RNA-Seq breast cancer datasets whereas this chapter uses the scRNA-Seq datasets, Chapter 5 uses KEGG pathways, whereas this chapter uses the Hallmarks pathways. Although

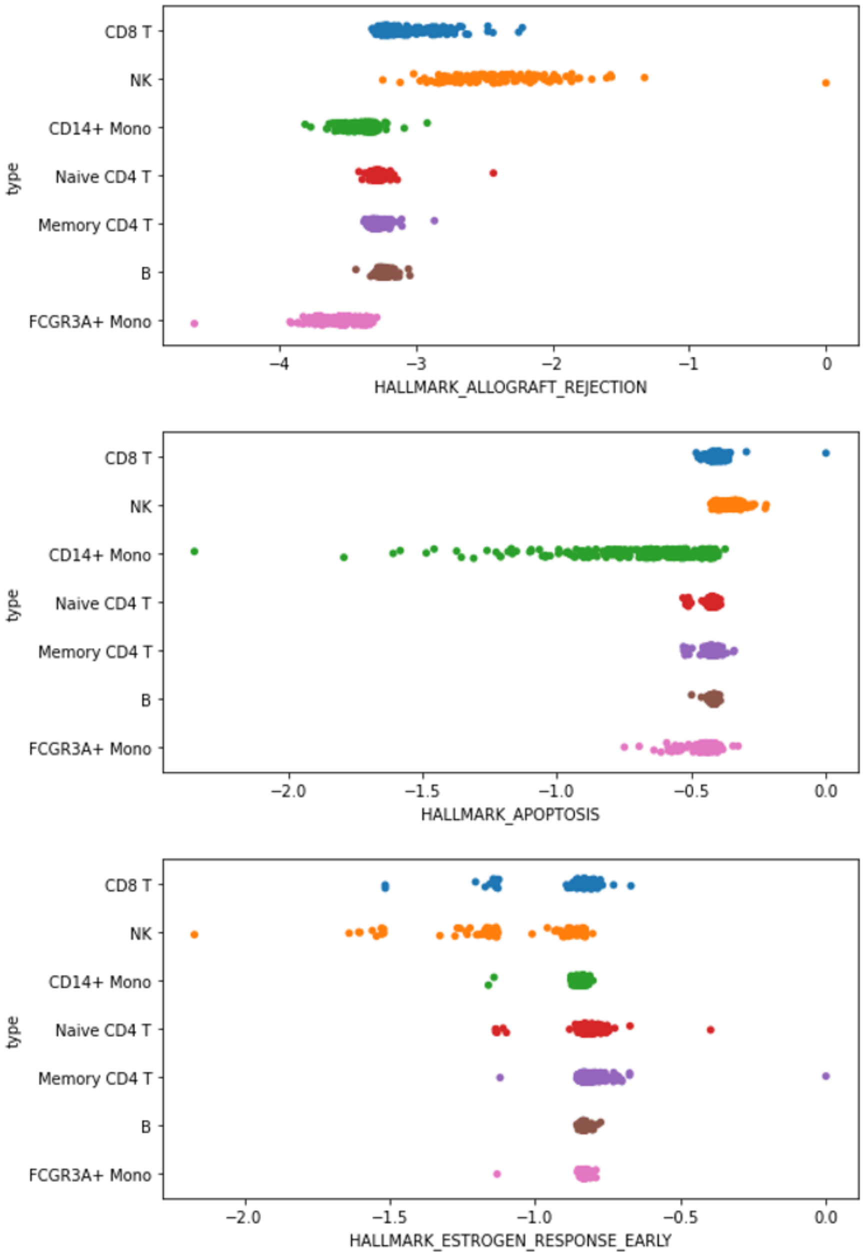


Fig. 6.7 Visualisation of the pathway activity values of the top three pathways in Hallmark. These three pathways are the top pathways for separating the cell types using PBMC 3k dataset. The top pathways are selected by ranking all the Hallmark pathways using the Individual Pathway Prediction Accuracy approach of DIGS2 method. Each dot represents the pathway activity value of a cell.

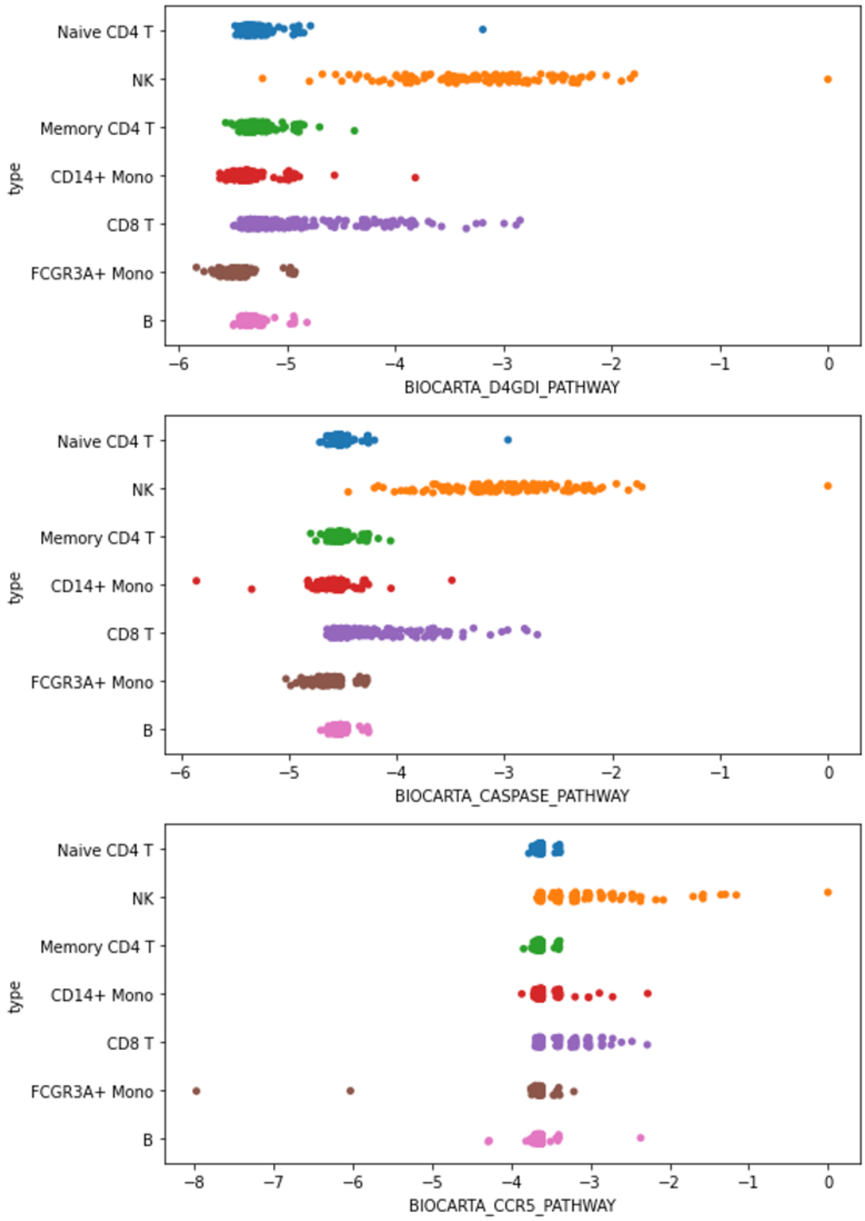


Fig. 6.8 Visualisation of the pathway activity values of the top three pathways in BioCarta. These three pathways are the top pathways for separating the cell types using PBMC 3k dataset. The top pathways are selected by ranking all the Biocarta pathways using the Individual Pathway Prediction Accuracy approach of DIGS2 method. Each dot represents a pathway activity value of a cell.

Table 6.4 Significant pathways and genes for identifying cell types in PBMC

Pathway name	Top 5 genes and their accumulated weights									
BIOCARTA_D4GDI_PATHWAY	GZMB	3.62	PRF1	1.18	CASP1	0.12	JUN	0.04	ARHGDIB	0.01
BIOCARTA_CASPASE_PATHWAY	GZMB	3.54	PRF1	1.13	CASP1	0.16	CASP4	0.13	ARHGDIB	0.02
BIOCARTA_CCR5_PATHWAY	CCL4	4.40	PRKCB	0.36	CALM3	0.07	CXCR4	0.05	CALM1	0.03
BIOCARTA_NO2IL12_PATHWAY	CD2	2.24	CD247	1.80	CD3D	0.48	CD3E	0.40	CD3G	0.06
BIOCARTA_STATHMIN_PATHWAY	CD247	1.81	CD2	1.76	HSPA5	0.72	CD3E	0.33	CD3D	0.29
HALLMARK_ALLOGRAFT_REJECTION	GZMB	2.02	CD79A	1.1	CCL4	0.39	IRF8	0.18	PRF1	0.14
HALLMARK_APOPTOSIS	CD14	1.46	PRF1	0.88	HMGB2	0.29	PPT1	0.21	TNFSF10	0.19
HALLMARK_ESTROGEN_RESPONSE_EARLY	CLIC3	2.72	PRSS23	1.23	AQP3	0.51	XBP1	0.13	CD44	0.12
HALLMARK_IL2_STAT5_SIGNALING	CTSZ	1.01	FGL2	0.86	IRF8	0.61	CD79B	0.42	IFITM3	0.41
HALLMARK_TNFA_SIGNALING_VIA_NFKB	CCL4	2.14	LITAF	0.25	MAP2K3	0.24	CEBPB	0.21	GADD45B	0.16

Table 6.5 Significant pathways and genes for breast cancer

Pathway name	Top 5 pathways and genes for identifying the subtypes									
HALLMARK_PANCREAS_BETA_CELLS	DCX	2.445	SYT13	2.35	PAK3	0.111	CHGA	0.023	ABCC8	0.021
HALLMARK_CHOLESTEROL_HOMEOSTASIS	GPX8	2.405	AVPR1A	2.316	MAL2	0.024	SEMA3B	0.018	ERRFI1	0.017
HALLMARK_APICAL_SURFACE	ATP8B1	1.937	EFNA5	1.565	SHROOM2	1.139	RTN4RL1	0.145	SCUBE1	0.032
HALLMARK_ANGIOGENESIS	KCNJ8	1.489	COL5A2	1.364	SERPINA5	1.01	PDGFA	0.712	MSX1	0.177
HALLMARK_SPERMATOGENESIS	CLGN	1.456	GPR182	1.444	NEK2	0.495	ZC2HC1C	0.323	TTK	0.311
Pathway name	Top 5 pathways and genes for identifying the TNBC vs. Other subtypes									
HALLMARK_COMPLEMENT	TMPRSS6	0.917	PCLO	0.696	F3	0.654	KCNIP3	0.373	F10	0.339
HALLMARK_TNFA_SIGNALING_VIA_NFKB	ICOSLG	2.278	F3	1.148	TNC	0.455	FJX1	0.354	CXCL11	0.145
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	FBLN2	0.815	COL6A3	0.774	DKK1	0.725	SNAI2	0.313	FOXC2	0.229
HALLMARK_ALLOGRAFT_REJECTION	RPL3L	1.401	EGFR	0.962	ICOSLG	0.754	CCL13	0.441	IL4	0.438
HALLMARK_INTERFERON_GAMMA_RESPONSE	KLRK1	1.4	USP18	0.936	BATF2	0.711	CCL7	0.404	PLA2G4A	0.339

Hallmark pathways are integrated from multiple pathway database resources, among which KEGG is included, the details (e.g. number of pathways in the collection, sizes of the pathways, biological processes described) of Hallmark pathways and KEGG pathways are quite different. The common part of Chapter 5 and this chapter is that the pathways are ranked using the same approach (Individual Pathway Prediction Accuracy). In summary, pancreas-related pathways are identified when training the DIGS2 model using different input datasets. This result gives a new perspective on the stability and reproducibility of the conclusions obtained using DIGS2. Also, it shows that pancreas-related pathways are strongly related to breast cancer. The relationship between pancreatic cancer and breast cancer can be found in the literature. A survey [242] found that BRCA2 mutation appears to be associated with pancreas cancer in the familial breast cancer population. Another study [243] made a similar conclusion using retrospective cohort analysis.

For the other pathways listed in Table 6.6, cholesterol homeostasis is found disrupted in rapidly proliferating breast cancer cells and increasing the proliferation of estrogen receptor (ER)-positive breast cancer cells [244]. The NF-Kb signalling pathway is found to be related to the TNBC by many researchers [228, 245, 246]. These research works show that it is highly related to the Hedgehog Signalling pathway, a well-known pathway related to multiple types of cancer and has been introduced in detail in Chapter 4. Moreover, the NF-kB signalling pathways contribute to deregulating Hedgehog signalling, thereby leading to the acquisition of the mesenchymal phenotype, enhanced growth, and invasion in TNBC. The activation of the IFN-γ pathway plays a key role in immunosuppression and is associated with the depletion and dysfunction of CD8+ T cells in the TNBC [247, 248].

6.5 Integration of PBMC 3k and PBMC 10k

The UMAP projection on the merged pathway activity dataset is shown in Figure 6.9. Both scatter plots use the same UMAP embeddings but use different cell labels. The left plot shows the nine cell types for both datasets (i.e. B_cells, T_cells and Mono/Macro from PMBC 10k and the other six cell types from PBMC 3k). Because the PBMC 10k dataset contains only B cells, T cells and mono/macro cells, the NK cell population in PBMC 3k is removed before merging the two datasets. In principle, the cell annotation for the two datasets should be consistent. However, in this experiment, the cell types in PBMC 10k are annotated coarsely, while in PBMC 3k, the annotations are refined. For example, the T cells in PBMC 10k are annotated with Memory CD4+ T, Naïve CD4+ T and CD8+ T in PBMC 3k.

Overall, the integration result of these two datasets is at an intermediate stage between Figure 6.3a and Figure 6.3b. The negative aspect is that the two datasets are not mixed. The PBMC 10k is clustered into three clusters, and the PBMC 3k is clustered into two clusters. These five clusters can be found in the right-hand side plots. However, the positive aspect is that the distances of clusters for the same cell types are close. For example, the cluster of T cell population from PBMC 10k and the cluster T cell population from PBMC 3k have only a tiny gap between their borders. The situation is similar for the Mono cell populations from these two datasets. The two Mono cell clusters are all located at the bottom right corner of the plot.

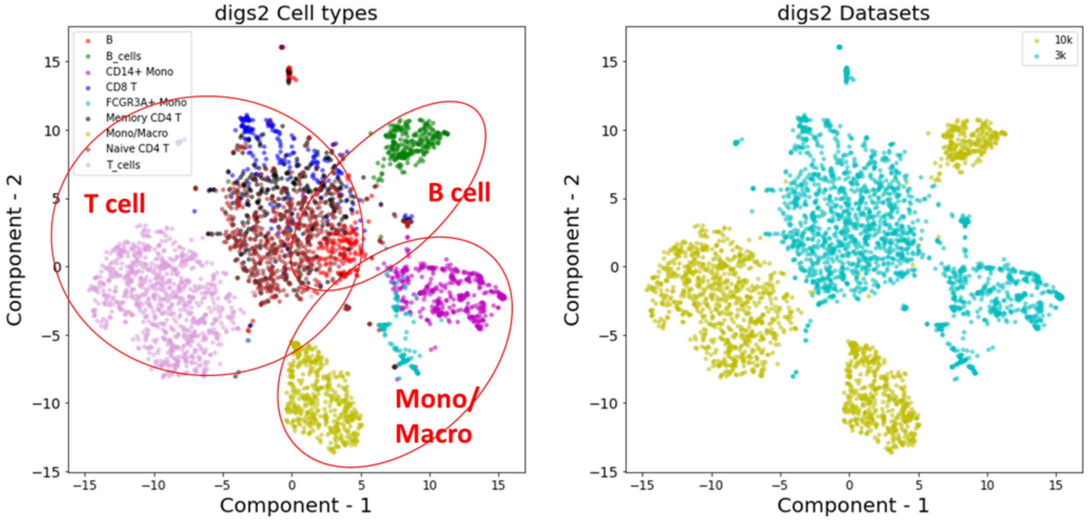


Fig. 6.9 Projection of the integrated PBMC datasets. Both of these two figures are created using the same pathway activity profile of the integrated PBMC datasets. Cells in the left figure is labelled with 7 cell types, and labelled with dataset in the right figure.

In conclusion, the current results do not reach expectations, as the same cell types from the two datasets are not totally clustered together. they still show promising direction for refining the implementation steps. For example, as DIGS2 is a supervised method, the different cell label settings could affect the integration results. Thus, training the PBMC 3k dataset with the same cell labels as PBMC 10k dataset may lead to better results.

6.6 Conclusion and Discussion

This chapter explores the application of the DIGS2 model on scRNA-Seq data. As a newly emerged gene expression profiling technology, scRNA-Seq breaks the limitations of deep sequencing in single cells. However, the sparsity of the scRNA-Seq data is still unavoidable and brings challenges in analysing them. Therefore, this chapter uses a wide range of datasets to test the performance of the DIGS2 model.

Four experiments were implemented on two scRNA-Seq datasets. First, using the pathway activity values to separate the different immune cell types in human peripheral blood mononuclear cells; second, using pathway activity values to separate the cells from different body tissues, i.e. blood and tumour; third, using pathway activity values to separate the four breast cancer subtype and lastly using pathway activity to separate the breast cancer subtype of worst prognosis, TNBC, with the other subtypes. The first experiment was conducted on the PBMC 3k dataset; the other three experiments were implemented on the breast cancer dataset. These experiments aim to provide a comprehensive view for exhibiting the performance of the DIGS2 model on scRNA-Seq datasets by using both the primary PBMC dataset and the more complex breast cancer datasets.

The evaluation in this chapter focuses more on how the cells are clustered with the inferred pathway activity values, as cell clustering and low-dimensional visualisation are crucial steps in analysing single-cell data. Therefore, the method is evaluated using the averaged silhouette score calculated on the cell clusters in the UMAP projection scatter plots. The results showed that DIGS2 outperforms the baseline method GSVA, which is prevalent in analysing scRNA-Seq data. More importantly, the results showed that the DIGS2 model could provide much knowledge for analysing the scRNA-Seq datasets. Compared with the standard analysing pipeline, DIGS2 not only can provide significant genes but also can provide significant pathways. Also, DIGS2 is proven to derive the disease-related pathway and genes, which benefits the research of the complex disease.

Also, this chapter discusses the possibility of using pathway activity values to integrate the different scRNA-Seq datasets. The pathway activity values are produced separately for PBMC 3k and PBMC 10k, and then combined to form integrated pathway activity profiles. The UMAP projection of the integrated pathway activity values showed that, although the datasets are not merged very well, the cells of the same types show the trends of being clustered together. Therefore, it can gain better results after refining the implementation pipeline.

In conclusion, DIGS2 shows its superiority in analysing the scRNA-Seq data and more applications of the pathway activity values of scRNA-Seq data are worth discovering.

Chapter 7

Conclusions and future work

7.1 Overview of thesis

The thesis focused on using optimisation models for pathway activity inference on different kinds of gene profiling data. Chapters 1 and 2 provided the introduction and related background for addressing this problem. Chapters 3 and 4 discussed applying an existing optimisation model on microarray and RNA-Seq datasets. Chapter 5 proposed a new optimisation model and applied it to RNA-Seq data to verify its advancements. Chapter 6 explored using the new proposed model on the newest and most challenging data, scRNA-Seq. A brief overview of each chapter is given below.

Chapter 1 introduced the general topic of research undertaken in this thesis and briefly explained the rationale behind the work. It placed the pathway activity inference problem in the System Biology context, outlining the thesis's key research goals.

Chapter 2 provided a detailed review of the essential background and related work. First, three gene profiling technologies and their properties were introduced. Then an introduction of how biological pathways can be incorporated within the analysis of the profiling data, followed by the definition of pathway activity inference. The functions and advancements of reducing the data dimensionality from gene-level to pathway-level by pathway activity inference methods were described. The overview of the pathway activity inference framework was given in detail, including the profiling data and pathway data collection, pathway activity profile calculation and metrics to evaluate the pathway activity inference method. Finally, a review of pathway activity inference methods was given.

Chapter 3 discussed the problem of classifying the binary phenotypes, i.e. disease and healthy, of ischemic stroke. An existing MILP optimisation model DIGS was applied for inferring the pathway activity values using microarray gene profiling datasets of ischemic stroke. The classification results using six machine learning classifiers showed promising accuracy rates (83.5%) and relatively high AUC values (91.5%) over ten times training and testing data split, which was higher than most current gene-based stroke phenotype prediction methods. Also, the important pathways identified by the DIGS model were proven related to the cause of ischemic stroke and can be seen as the pathway biomarkers of ischemic stroke.

Chapter 4 discussed the problem of classifying the multi-class phenotypes in the four molecular subtypes of colorectal cancer. The DIGS model was applied for inferring the pathway activity values using the advanced high-throughput profiling technology, RNA-Seq gene profiling dataset. In addition to the classification accuracy, this chapter added two more evaluation metrics for evaluating the pathway activity inference method, the robustness against noisy data and survival analysis. The results comprising DIGS with baseline and newly proposed pathway activity inference methods showed that the DIGS model outperformed the other models in all three evaluation metrics. Therefore the adaptation of the DIGS model on RNA-Seq data is illustrated. Moreover, concerning identifying disease-related pathways, this chapter proposed a follow-up approach to calculate the sample classification accuracy for individual pathways based on the outputs of the DIGS model. Therefore the pathways can be ranked according to their classification accuracies. By searching the literature, the top-ranked pathways also proved to be those highly related to colorectal cancer.

Chapter 5 refined the DIGS model to a more robust optimisation-based pathway activity inference method for multi-class disease classification tasks, called DIGS2. DIGS2 is also a MILP optimisation model, which reduces the number of binary variables to decrease computational complexity. Therefore, the DIGS2 model can be solved to global optimality easier than DIGS using the same input data. The DIGS2 model is verified on the breast cancer and colorectal cancer RNA-Seq datasets for binary and multi-class classification tasks. The comparative study showed that DIGS2 improved the computational efficiency compared with DIGS. The number of optimal solutions increased, and the gap values decreased. Also, DIGS2 outperformed the DIGS model and other comparative pathway activity inference methods for the three evaluation metrics. Moreover, the approach of evaluating individual pathways was further extended to build the NaiveDIGS method, which could aggregate the DIGS2 outputs of all pathways and produce the overall classification accuracy. The NaiveDIGS enabled the DIGS2 method to get rid of the outsider machine learning classifiers and directly provide the evaluation for disease phenotype classification.

Chapter 6 explores the application of DIGS2 model on single-cell RNA-Seq datasets. The objectives consisted of three parts. First pathway activity values are used to separate the cell types of healthy human blood cells. Next pathway activity values are used to separate the molecular subtypes of breast cancer cells. The DIGS2 model was compared with a widely used pathway activity inference method applied to scRNA-Seq data. The results showed that DIGS2 performed better on the scRNA-Seq data for clustering and classifying the cells than the comparative method. In the last objective the pathway activity values are used to integrate cells from different datasets. Two human blood cell datasets were merged using the pathway activity value inferred by DIGS2. Although the merged dataset did not reach the best integration status, the results still were promised to be improved with refined cell label settings.

7.2 Research aims revisited

In Chapter 1, five research aims were outlined. These are now revisited in this section in order to ascertain how effectively they have been fulfilled.

1. To build on existing pathway activity inference method to evaluate its ability on analysing bulk gene expression profiling data.

The DIGS method was applied to a microarray dataset in Chapter 3 and two RNA-Seq datasets in Chapter 4 and Chapter 5. Traditionally, the pathway activity inference method is evaluated by machine learning classifiers. The pathway activity profile is passed into the classifier for training and testing. The prediction accuracy is used for assessing how well the pathway activity values can classify sample phenotypes. In this thesis, two more evaluation metrics are designed to evaluate its ability to represent data efficiently. The first is the robustness against noise in gene expression data. As the gene profiling technologies would unavoidably introduce noise into the expression data, pathway activity values are expected to have the ability to eliminate the noise to some extent. The other is how the pathway activity values can predict the survival of patient samples. Compared with using gene-level data to predict the survival probability, pathway-level data incorporates the higher level information, the functions of gene groups. Therefore, the pathway activity values are expected to perform better in survival prediction. The results from these two chapters showed that the DIGS model outperformed the comparative methods for classification accuracy and performed better for robustness and survival analysis. Therefore this research aim was addressed.

2. To build and implement a new pathway activity inference method that has higher solution quality and higher prediction accuracy.

In Chapter 5, a refined MILP optimisation model DIGS2 is proposed. A large number of binary decision variables made the original DIGS model suffer from difficulties in obtaining global optimal solutions. Therefore, in DIGS2 the model was reformulated to reduce the number of binary decision variables. The strategy for reducing the number of binary variables is replacing the pathway activity enclosing constraints with the sample violation distance constraints. To be more specific, the original model needs binary variables to judge whether a sample fills within a class or not, while the new model uses a constant to describe the distance between samples and classes. Therefore the new model can get rid of some binary decision variables. At the same time, DIGS2 retains the core concept of the DIGS model for inference of pathway activity, i.e. calculating the pathway activity as the weighted summation of gene expression values, where the gene weights were optimised to maximise the separation of pathway activity values. Consequently, the new model is expected to perform similarly to the original model in practical usage.

Chapter 5.3.1 compares the computational efficiency between DIGS and DIGS2 models. A great increase can be seen in the number of optimal solutions. Also, DIGS2 model solutions had a much smaller number of Gap values for those feasible solutions. Chapter 5.3.2 compares the performance of these two models in terms of the quality of pathway activities. The results showed that DIGS2 had higher prediction accuracy for binary and multi-class problems. Therefore, the comparisons proved that DIGS2 is a better model with higher solution quality and prediction accuracy.

3. To extend the application of pathway activity inference method from bulk data to single cell data.

Chapter 6 applies the DIGS2 model to single-cell RNA-Seq datasets. As a newly emerging technology, pathway activity inference on scRNA-Seq data has yet to be widely investigated. The sparse nature of single-cell data and large number of cells make analysis methods designed for bulk data only sometimes applicable to single-cell data. Therefore, the flexible nature of the optimization models makes it more suitable for single-cell data than other methods that rely on data quality. The results of Chapter 6 verified this view. The pathway activity calculated by DIGS2 not only showed great improvements for cell clustering and cell label prediction but also got better performance compared with the bulk method that had been applied for scRNA-Seq data analysis (GSVA). Therefore, this research aim was addressed.

4. To evaluate the methodology to show comparability with existing methods from literature.

From Chapter 3 to Chapter 6, the DIGS and DIGS2 models were compared with the most compatible methods listed in Table 2.3. Chapter 3 compared the DIGS model with MEAN, Median and PCA. Chapter 4 compared the DIGS model with MEAN, PCA and Pathifier. Chapter 5 compared two optimisation models with MEAN, PCA, Pathifier and GSVA. Chapter 6 compares the DIGS2 model with GSVA. The other four methods listed in Table 2.3 are not used for comparative study in this thesis because they were only adopted to the binary problem, which did not meet the requirements for solving multi-class classification problems. Among the comparative methods used in this thesis, the MEAN and PCA represented the baseline methods, which are aimed at providing a standard of the performance of pathway activity. GSVA is a well-known and popular method. Pathifier represents the advanced method with complex calculation processes and has been proven the best method from the literature. Therefore, by comparing these methods, DIGS and DIGS2 models showed their comparability with existing methods.

5. To demonstrate the potential of such methods to find meaningful results in biological applications.

One of the most outstanding advantages of the DIGS and DIGS2 models is that they allow in-depth analysis of individual pathways. The DIGS and DIGS2 outputs consist of the pathway activity values and class intervals for each pathway. Given the pathway activity values and class intervals, the samples can be allocated to their nearest class. Consequently, the importance of a single pathway can be assessed by how many samples are allocated into the correct class. Compared with the other methods that rely on statistical tests to evaluate the correlations between pathway activity values with sample labels, DIGS and DIGS2 models provide a highly explainable way to identify highly relevant pathways. Also, as these two optimisation models use gene weights to construct the pathway activity values, the member genes of the pathways can be assessed through their weight values.

In the results and discussion sections of Chapter 3, Chapter 4, Chapter 5, and Chapter 6, the important pathways and genes identified by DIGS or DIGS2 were verified in the literature. Nearly all the high-ranked pathways and gene families had been proven to be related to the disease of interest. Also, the robustness of these identified pathways was shown throughout this thesis. For example, in Chapter 5 and Chapter 6, two different breast cancer datasets (RNA-Seq and scRNA-Seq) and two different pathway collections (KEGG and Hallmark) were input into the DIGS2 model. However, the top-ranked pathways for separating the breast cancer subtypes from these two different datasets are the same pathway (Chapter 6.9). Therefore, the important pathways identified by the DIGS and DIGS2 models are meaningful and reproducible.

7.3 Limitations

The optimisation models used in this thesis focused on dealing with the main concern of the pathway activity inference problem and showed excellent results. However, there are two limitations to these optimisation models.

The first limitation is that they do not consider the topology of pathways. Pathway topology is the role, position, and interaction directions of the pathway genes [109]. From the conclusions of a review [249], for the pathway analysis methods, topology-based methods appear to perform better than non-topology-based methods. This is somewhat expected since the topology-based methods consider the structure of the pathway, which is meant to describe underlying phenomena better than using only the gene expression values.

The second limitation is that the optimisation models cannot indicate the regulation directions of the pathways. Up- or down-regulation is a process by which the availability of molecules involved in the pathway, such as proteins and mRNA, is increased or decreased in the cell [250]. Some popular methods provide an up- or down-regulated call at the pathway level. For example, GSVA compares the rank statistics of pathway member genes with the complement genes to indicate the regulation direction of the pathway [100]. For the optimisation models in this thesis, the gene weight variables can potentially complete this task through their positive or negative signs. However, the signs of gene wights have yet to have specific meanings in current models.

These limitations are not necessary for interpreting the pathway activities but imply the potential improvements for the models built in this thesis. Therefore, it is reasonable to invest time in exploring them further.

7.4 Future work

Future work will involve complementing work in Chapter 6 further. The scRNA-Seq datasets integration process needs more reasonable pathway activity values as inputs. The cell labels of the two datasets should be consistent. Therefore, the first step is re-calculate the pathway activity values of PBMC 3k using the coarse cell type annotations and then repeat the integration process to check the performance.

Future work will also involve the improvement of the methodologies presented in this thesis. There is now a prototype of the new optimisation model for pathway activity inference, named as DIGS3. DIGS or DIGS2 uses the concept of class intervals to limit the pathway activity values of samples to fall within a specific area on the axis to separate the samples of different labels through the objective function. However, the class ranges did not show their power in practical usage. For example, in the usage of DIGS2 in Chapter 6, when the number of classes reached seven, many classes did not have a valid range (i.e. the lower bound equals the upper bound). In particular, some classes have a distribution of pathway activity values that is more prominent than others. However, this is not reflected in its ranges, i.e. the samples of the same class are clustered on the axis but are far away from the class interval.

Therefore, in this new version model, I tried to use another concept, the class anchor, to replace the class intervals. The anchor represents a point on the axis that collects samples from this class. In other words, DIGS3 brings the samples of the same class as close as possible to the anchor point.

The indices, sets and parameters associated with the DIGS3 model are listed below:

Indices:
s cells (s = 1,2,...,S)
m pathway member genes (m = 1,2,...,M)
c class (c = 1,2,...,C)
cs mapping of classes and samples

Parameter:
G_{sm} gene expression values for cell s and gene m

Free Variable:
r_m gene weight
AC_c anchor point of class c
pa_s pathway activity value of sample s

Positive variable:
D_s distance of samples s to its class c

The objective function and the constraints that form the model are:
Objective:

$$Min\ z = \sum_{s} D_s \qquad (7-1)$$

Subject to:

$$pa_s = \sum_{m}^{M} G_{sm} r_m \qquad \forall s \qquad (7-2)$$

$$(AC_c - pa_s)^2 \leq D_s \qquad \forall s\, c_s \qquad (7-3)$$

$$AC_c - AC_k \geq 1 \qquad \forall k < c \qquad (7-4)$$

Because equation (7-3) introduced non-linear constraint into the model, this model is an NLP model. This primary model aims to implement the concept of the class anchor. By testing it on toy data with ten samples of two classes and six genes, the DIGS3 model has gotten better separation for the two phenotypes than the DIGS model. As shown in Figure 7.1, the samples of different label are more clearly separated in DIGS3. Therefore, in this next step, the DIGS3 model needed to be refined and made more reasonable mathematically. In the meantime, it is tested on gene expression datasets to indicate promising potential.

Overall, this thesis demonstrates the flexible and interpretable modelling offered by mathematical optimisation methods in analysing high-throughput data to improve disease classification. A series of DIGS models are implemented and tested on various types of gene expression profiling data. The comparison results of these models showed their superiority of the other pathway activity inference models. Their biological interpretability has also been demonstrated. By analysing the output of the optimised models, this thesis identifies several important genes and pathways associated with disease.

The DIGS series models are keeping improving to deal with the more and more complex up-to-date data and aim to achieve better performance.

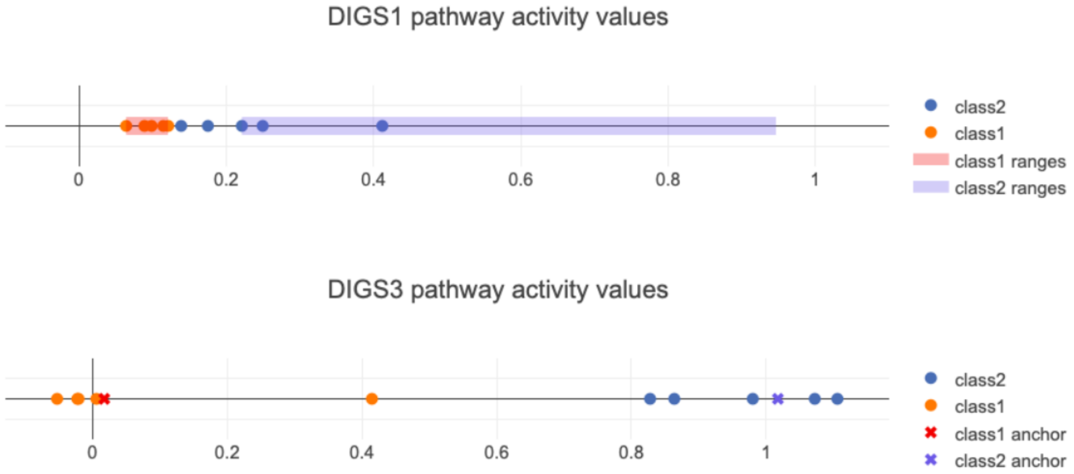


Fig. 7.1 Comparison between DIGS and DIGS3 using toy data.

References

[1] Iman Tavassoly, Joseph Goldfarb, and Ravi Iyengar. Systems biology primer: the basic methods and approaches. *Essays in Biochemistry*, 62(4):487–500, October 2018.

[2] Teri A. Manolio. Genomewide Association Studies and Assessment of the Risk of Disease. *New England Journal of Medicine*, 363(2):166–176, July 2010.

[3] David B. Goldstein, Andrew Allen, Jonathan Keebler, Elliott H. Margulies, Steven Petrou, Slavé Petrovski, and Shamil Sunyaev. Sequencing studies in human genetics: design and interpretation. *Nature Reviews Genetics*, 14(7):460–470, July 2013.

[4] Yunjin Li, Lu Ma, Duojiao Wu, and Geng Chen. Advances in bulk and single-cell multi-omics approaches for systems biology and precision medicine. *Briefings in Bioinformatics*, page bbab024, March 2021.

[5] JG Liao and Khew-Voon Chin. Logistic regression for disease classification using microarray data: model selection in a large p and small n case. *Bioinformatics*, 23(15):1945–1951, 2007.

[6] Harish Bhaskar, David C. Hoyle, and Sameer Singh. Machine learning in bioinformatics: A brief survey and recommendations for practitioners. *Computers in Biology and Medicine*, 36(10):1104–1125, October 2006.

[7] L. Ein-Dor, O. Zuk, and E. Domany. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Sciences*, 103(15):5923–5928, April 2006.

[8] H Paul Williams. *Model building in mathematical programming.* John Wiley & Sons, 2013.

[9] Matthew V. Rockman and Leonid Kruglyak. Genetics of global gene expression. *Nature Reviews Genetics*, 7(11):862–872, November 2006.

[10] Susan D. Thompson, Sampath Prahalad, and Robert Allen Colbert. Integrative Genomics. In *Textbook of Pediatric Rheumatology*, pages 43–53.e3. Elsevier, 2016.

[11] Patrick O. Brown and David Botstein. Exploring the new world of the genome with DNA microarrays. *Nature Genetics*, 21(S1):33–37, January 1999.

[12] Dennise D. Dalma-Weiszhausz, Janet Warrington, Eugene Y. Tanimoto, and C. Garrett Miyada. [1] The Affymetrix GeneChip® Platform: An Overview. In *Methods in Enzymology*, volume 410, pages 3–28. Elsevier, 2006.

[13] Likun Wang, Zhixing Feng, Xi Wang, Xiaowo Wang, and Xuegong Zhang. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, 26(1):136–138, January 2010.

[14] Fangxin Hong, Rainer Breitling, Connor W. McEntee, Ben S. Wittner, Jennifer L. Nemhauser, and Joanne Chory. RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, 22(22):2825–2827, November 2006.

[15] W. Pan. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 18(4):546–554, April 2002.

[16] Roger Bumgarner. Overview of DNA Microarrays: Types, Applications, and Their Future. *Current Protocols in Molecular Biology*, 101(1), January 2013.

[17] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, January 2009.

[18] Fatih Ozsolak and Patrice M. Milos. RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, 12(2):87–98, February 2011.

[19] Bo Li, Victor Ruotti, Ron M. Stewart, James A. Thomson, and Colin N. Dewey. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26(4):493–500, February 2010.

[20] May Ke, Badran Elshenawy, Helen Sheldon, Anjali Arora, and Francesca M Buffa. Single cell RNA-sequencing: A powerful yet still challenging technology to study cellular heterogeneity. *BioEssays*, page 2200084, September 2022.

[21] Xiliang Wang, Yao He, Qiming Zhang, Xianwen Ren, and Zemin Zhang. Direct Comparative Analyses of 10X Genomics Chromium and Smart-seq2. *Genomics, Proteomics & Bioinformatics*, 19(2):253–266, April 2021.

[22] Allon M. Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A. Weitz, and Marc W. Kirschner. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell*, 161(5):1187–1201, May 2015.

[23] Serge Koscielny. Why Most Gene Expression Signatures of Tumors Have Not Been Useful in the Clinic. *Science Translational Medicine*, 2(14), January 2010.

[24] B. van der Vegt, G.H. de Bock, H. Hollema, and J. Wesseling. Microarray methods to identify factors determining breast cancer progression: Potentials, limitations, and challenges. *Critical Reviews in Oncology/Hematology*, 70(1):1–11, April 2009.

[25] Eytan Domany. Using High-Throughput Transcriptomic Data for Prognosis: A Critical Overview and Perspectives. *Cancer Research*, 74(17):4612–4621, September 2014.

[26] Marc J. van de Vijver, Yudong D. He, Laura J. van 't Veer, Hongyue Dai, Augustinus A.M. Hart, Dorien W. Voskuil, George J. Schreiber, Johannes L. Peterse, Chris Roberts, Matthew J. Marton, Mark Parrish, Douwe Atsma, Anke Witteveen, Annuska Glas, Leonie Delahaye, Tony van der Velde, Harry Bartelink, Sjoerd Rodenhuis, Emiel T. Rutgers, Stephen H. Friend, and René Bernards. A Gene-Expression Signature as a Predictor of Survival in Breast Cancer. *New England Journal of Medicine*, 347(25):1999–2009, December 2002.

[27] Yixin Wang, Jan GM Klijn, Yi Zhang, Anieta M Sieuwerts, Maxime P Look, Fei Yang, Dmitri Talantov, Mieke Timmermans, Marion E Meijer-van Gelder, Jack Yu, Tim Jatkoe, Els MJJ Berns, David Atkins, and John A Foekens. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, 365(9460):671–679, February 2005.

[28] Sridhar Ramaswamy, Ken N. Ross, Eric S. Lander, and Todd R. Golub. A molecular signature of metastasis in primary solid tumors. *Nature Genetics*, 33(1):49–54, January 2003.

[29] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171–178, January 2005.

[30] Biological Pathways Fact Sheet, August 2020.

[31] A L Fridman and M A Tainsky. Critical pathways in cellular senescence and immortalization revealed by gene expression profiling. *Oncogene*, 27(46):5975–5987, October 2008.

[32] Zhaohui Feng, Wenwei Hu, Gunaretnam Rajagopal, and Arnold J. Levine. The tumor suppressor p53: Cancer and aging. *Cell Cycle*, 7(7):842–847, April 2008.

[33] A Ben-Ze'ev. Cytoskeletal and adhesion proteins as tumor suppressors. *Current Opinion in Cell Biology*, 9(1):99–108, 1997.

[34] Mathew Loesch. The p38 MAPK stress pathway as a tumor suppressor or more? *Frontiers in Bioscience*, Volume(13):3581, 2008.

[35] Miguel A García-Campos, Jesús Espinal-Enríquez, and Enrique Hernández-Lemus. Pathway analysis: state of the art. *Frontiers in physiology*, 6:383, 2015.

[36] Mike West, Carrie Blanchette, Holly Dressman, Erich Huang, Seiichi Ishida, Rainer Spang, Harry Zuzan, John A. Olson, Jeffrey R. Marks, and Joseph R. Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences*, 98(20):11462–11467, September 2001.

[37] Andrea H. Bild, Guang Yao, Jeffrey T. Chang, Quanli Wang, Anil Potti, Dawn Chasse, Mary-Beth Joshi, David Harpole, Johnathan M. Lancaster, Andrew Berchuck, John A. Olson, Jeffrey R. Marks, Holly K. Dressman, Mike West, and Joseph R. Nevins. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, 439(7074):353–357, January 2006.

[38] Bert Vogelstein and Kenneth W Kinzler. Cancer genes and the pathways they control. *Nature Medicine*, 10(8):789–799, August 2004.

[39] James W. Watters and Christopher J. Roberts. Developing gene expression signatures of pathway deregulation in tumors. *Molecular Cancer Therapeutics*, 5(10):2444–2449, October 2006.

[40] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. 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, October 2005.

[41] Marit Ackermann and Korbinian Strimmer. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10(1):47, December 2009.

[42] Dan Nettleton, Justin Recknor, and James M. Reecy. Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis. *Bioinformatics*, 24(2):192–201, January 2008.

[43] John Tomfohr, Jun Lu, and Thomas B Kepler. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, 6(1):225, December 2005.

[44] J. J. Goeman and P. Buhlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, April 2007.

[45] I. Dinu, J. D. Potter, T. Mueller, Q. Liu, A. J. Adewale, G. S. Jhangri, G. Einecke, K. S. Famulski, P. Halloran, and Y. Yasui. Gene-set analysis and reduction. *Briefings in Bioinformatics*, 10(1):24–34, October 2008.

[46] Qi Liu, Irina Dinu, Adeniyi J Adewale, John D Potter, and Yutaka Yasui. Comparative evaluation of gene-set analysis methods. *BMC Bioinformatics*, 8(1):431, December 2007.

[47] J. J. Goeman, S. A. van de Geer, F. de Kort, and H. C. van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, January 2004.

[48] U. Mansmann and R. Meister. Testing Differential Gene Expression in Functional Groups: Goeman's Global Test versus an ANCOVA Approach. *Methods of Information in Medicine*, 44(03):449–453, 2005.

[49] Irina Dinu, John D Potter, Thomas Mueller, Qi Liu, Adeniyi J Adewale, Gian S Jhangri, Gunilla Einecke, Konrad S Famulski, Philip Halloran, and Yutaka Yasui. Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics*, 8(1):242, December 2007.

[50] W. T. Barry, A. B. Nobel, and F. A. Wright. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, 21(9):1943–1949, May 2005.

[51] Sorin Drăghici, Purvesh Khatri, Rui P. Martins, G.Charles Ostermeier, and Stephen A. Krawetz. Global functional profiling of gene expression This work was funded in part by a Sun Microsystems grant awarded to S.D., NIH Grant HD36512 to S.A.K., a Wayne State University SOM Dean's Post-Doctoral Fellowship, and an NICHD Contraception and Infertility Loan to G.C.O. Support from the WSU MCBI mode is gratefully appreciated. *Genomics*, 81(2):98–104, February 2003.

[52] Sangsoo Lim, Sangseon Lee, Inuk Jung, Sungmin Rhee, and Sun Kim. Comprehensive and critical evaluation of individualized pathway activity measurement tools on pan-cancer data. *Briefings in Bioinformatics*, November 2018.

[53] Zheng Guo, Tianwen Zhang, Xia Li, Qi Wang, Jianzhen Xu, Hui Yu, Jing Zhu, Haiyun Wang, Chenguang Wang, Eric J Topol, Qing Wang, and Shaoqi Rao. Towards precise classification of cancers based on robust gene functional expression profiles. *BMC Bioinformatics*, 6(1):58, 2005.

[54] Emily Clough and Tanya Barrett. The Gene Expression Omnibus Database. In Ewy Mathé and Sean Davis, editors, *Statistical Genomics*, volume 1418, pages 93–110. Springer New York, New York, NY, 2016. Series Title: Methods in Molecular Biology.

[55] Tiago Chedraoui Silva Antonio Colaprico. TCGAbiolinks, 2017.

[56] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[57] Alberto Fernandez, Salvador Garcia, Francisco Herrera, and Nitesh V. Chawla. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, 61:863–905, April 2018.

[58] G. D. Bader. Pathguide: a Pathway Resource List. *Nucleic Acids Research*, 34(90001):D504–D506, January 2006.

[59] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361, January 2017.

[60] D. Croft, G. O'Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, B. Jassal, S. Jupe, I. Kalatskaya, S. Mahajan, B. May, N. Ndegwa, E. Schmidt, V. Shamovsky, C. Yung, E. Birney, H. Hermjakob, P. D'Eustachio, and L. Stein. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research*, 39(Database):D691–D697, January 2011.

[61] Denise N Slenter, Martina Kutmon, Kristina Hanspers, Anders Riutta, Jacob Windsor, Nuno Nunes, Jonathan Mélius, Elisa Cirillo, Susan L Coort, Daniela Digles, Friederike Ehrhart, Pieter Giesbertz, Marianthi Kalafati, Marvin Martens, Ryan Miller, Kozo Nishida, Linda Rieswijk, Andra Waagmeester, Lars M T Eijssen, Chris T Evelo, Alexander R Pico, and Egon L Willighagen. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research*, 46(D1):D661–D667, January 2018.

[62] Carl F. Schaefer, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H. Buetow. PID: the Pathway Interaction Database. *Nucleic Acids Research*, 37(suppl_1):D674–D679, January 2009.

[63] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdottir, P. Tamayo, and J. P. Mesirov. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–1740, June 2011.

[64] E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, N. Schultz, G. D. Bader, and C. Sander. Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research*, 39(Database):D685–D690, January 2011.

[65] Sarah Mubeen, Charles Tapley Hoyt, André Gemünd, Martin Hofmann-Apitius, Holger Fröhlich, and Daniel Domingo-Fernández. The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling. *Frontiers in Genetics*, 10:1203, November 2019.

[66] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems*, 1(6):417–425, December 2015.

[67] Changwon Keum, Jung Hoon Woo, Won Seok Oh, Sue-Nie Park, and Kyoung Tai No. Improving gene expression similarity measurement using pathway-based analytic dimension. *BMC Genomics*, 10(S3):S15, December 2009.

[68] T.-M. Kim, S.-H. Yim, Y.-B. Jeong, Y.-C. Jung, and Y.-J. Chung. PathCluster: a framework for gene set-based hierarchical clustering. *Bioinformatics*, 24(17):1957–1958, September 2008.

[69] Francisco Azuaje, Huiru Zheng, Anyela Camargo, and Haiying Wang. Systems-based biological concordance and predictive reproducibility of gene set discovery methods in cardiovascular disease. *Journal of Biomedical Informatics*, 44(4):637–647, August 2011.

[70] Francisco Azuaje, Yvan Devaux, and Daniel R. Wagner. Integrative Pathway-Centric Modeling of Ventricular Dysfunction after Myocardial Infarction. *PLoS ONE*, 5(3):e9661, March 2010.

[71] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[72] Kun-Huang Chen, Kung-Jeng Wang, Kung-Min Wang, and Melani-Adrian Angelia. Applying particle swarm optimization-based decision tree classifier for cancer classification on gene expression data. *Applied Soft Computing*, 24:773–780, November 2014.

[73] Sandro Sperandei. Understanding logistic regression analysis. *Biochemia Medica*, pages 12–18, 2014.

[74] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[75] Pádraig Cunningham and Sarah Jane Delany. k-Nearest Neighbour Classifiers - A Tutorial. *ACM Computing Surveys*, 54(6):1–25, July 2022.

[76] David M. W. Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. 2020. Publisher: arXiv Version Number: 1.

[77] Margherita Grandini, Enrico Bagli, and Giorgio Visani. Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*, 2020.

[78] Sarah-Jane Schramm, Anna E. Campain, Ricenterd A. Scolyer, Yee Hwa Yang, and Graham J. Mann. Review and Cross-Validation of Gene Expression Signatures and Melanoma Prognosis. *Journal of Investigative Dermatology*, 132(2):274–283, February 2012.

[79] Jorng-Tzong Horng, Li-Cheng Wu, Baw-Juine Liu, Jun-Li Kuo, Wen-Horng Kuo, and Jin-Jian Zhang. An expert system to classify microarray gene expression data using gene selection by decision tree. *Expert Systems with Applications*, 36(5):9072–9081, July 2009.

[80] A.-L. Boulesteix and W. Sauerbrei. Added predictive value of high-throughput molecular data to clinical data and its validation. *Briefings in Bioinformatics*, 12(3):215–229, May 2011.

[81] P. J. Castaldi, I. J. Dahabreh, and J. P. A. Ioannidis. An empirical assessment of validation practices for molecular classifiers. *Briefings in Bioinformatics*, 12(3):189–202, May 2011.

[82] Seungwoo Hwang. Comparison and evaluation of pathway-level aggregation methods of gene expression data. *BMC Genomics*, 13(S7):S26, December 2012.

[83] Masanori Oshi, Tae Hee Kim, Yoshihisa Tokumaru, Li Yan, Ryusei Matsuyama, Itaru Endo, Leonid Cherkassky, and Kazuaki Takabe. Enhanced DNA Repair Pathway is Associated with Cell Proliferation and Worse Survival in Hepatocellular Carcinoma (HCC). *Cancers*, 13(2):323, January 2021.

[84] Anieta M. Sieuwerts, Márcia A. Inda, Marcel Smid, Henk van Ooijen, Anja van de Stolpe, John W. M. Martens, and Wim F. J. Verhaegh. ER and PI3K Pathway Activity in Primary ER Positive Breast Cancer Is Associated with Progression-Free Survival of Metastatic Patients under First-Line Tamoxifen. *Cancers*, 12(4):802, March 2020.

[85] Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma. Gene expression–based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nature Medicine*, 14(8):822–827, August 2008.

[86] Eung-Sirk Lee, Dae-Soon Son, Sung-Hyun Kim, Jinseon Lee, Jisuk Jo, Joungho Han, Heesue Kim, Hyun Joo Lee, Hye Young Choi, Youngja Jung, Miyeon Park, Yu Sung Lim, Kwhanmien Kim, Young Mog Shim, Byung Chul Kim, Kyusang Lee, Nam Huh, Christopher Ko, Kyunghee Park, Jae Won Lee, Yong Soo Choi, and Jhingook Kim. Prediction of Recurrence-Free Survival in Postoperative Non–Small Cell Lung Cancer Patients by Using an Integrated Model of Clinical Information and Gene Expression. *Clinical Cancer Research*, 14(22):7397–7404, November 2008.

[87] Christiana Kartsonaki. Survival analysis. 22(7):263–270.

[88] D. R. Cox. A note on the graphical analysis of survival data. *Biometrika*, 66(1):188–190, 1979.

[89] Ana Fernandez-Teijeiro, Rebecca A. Betensky, Lisa M. Sturla, John Y.H. Kim, Pablo Tamayo, and Scott L. Pomeroy. Combining Gene Expression Profiles and Clinical Parameters for Risk Stratification in Medulloblastomas. *Journal of Clinical Oncology*, 22(6):994–998, March 2004.

[90] Thomas M. Habermann, Sophia S. Wang, Matthew J. Maurer, Lindsay M. Morton, Charles F. Lynch, Stephen M. Ansell, Patricia Hartge, Richard K. Severson, Nathaniel Rothman, Scott Davis, Susan M. Geyer, Wendy Cozen, Stephen J. Chanock, and James R. Cerhan. Host immune gene polymorphisms in combination with clinical and demographic factors predict late survival in diffuse large B-cell lymphoma patients in the pre-rituximab era. *Blood*, 112(7):2694–2702, October 2008.

[91] R. M. Simon, J. Subramanian, M.-C. Li, and S. Menezes. Using cross-validation to evaluate predictive accuracy of survival risk classifiers based on high-dimensional data. *Briefings in Bioinformatics*, 12(3):203–214, May 2011.

[92] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3), September 2008.

[93] Kaci L Pickett, Krithika Suresh, Kristen R Campbell, Scott Davis, and Elizabeth Juarez-Colunga. Random survival forests for dynamic predictions of a time-to-event outcome using a longitudinal biomarker. *BMC Medical Research Methodology*, 21:1–14, 2021.

[94] Sheng Li, Paweł P Łabaj, Paul Zumbo, Peter Sykacek, Wei Shi, Leming Shi, John Phan, Po-Yen Wu, May Wang, Charles Wang, Danielle Thierry-Mieg, Jean

Thierry-Mieg, David P Kreil, and Christopher E Mason. Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nature Biotechnology*, 32(9):888–895, September 2014.

[95] Marcelo P. Segura-Lepe, Hector C. Keun, and Timothy M. D. Ebbels. Predictive modelling using pathway scores: robustness and significance of pathway collections. *BMC Bioinformatics*, 20(1):543, December 2019.

[96] Eunjung Lee, Han-Yu Chuang, Jong-Won Kim, Trey Ideker, and Doheon Lee. Inferring Pathway Activity toward Precise Disease Classification. *PLoS Computational Biology*, 4(11):e1000217, November 2008.

[97] Jeremy A Miller, Chaochao Cai, Peter Langfelder, Daniel H Geschwind, Sunil M Kurian, Daniel R Salomon, and Steve Horvath. Strategies for aggregating gene expression data: The collapseRows R function. *BMC Bioinformatics*, 12(1):322, December 2011.

[98] J. Liu, J. M. Hughes-Oliver, and J. A. Menius. Domain-enhanced analysis of microarray data using GO annotations. *Bioinformatics*, 23(10):1225–1234, May 2007.

[99] E. Edelman, A. Porrello, J. Guinney, B. Balakumaran, A. Bild, P. G. Febbo, and S. Mukherjee. Analysis of sample set enrichment scores: assaying the enrichment of sets of genes for individual samples in genome-wide expression profiles. *Bioinformatics*, 22(14):e108–e116, July 2006.

[100] Sonja Hänzelmann, Robert Castelo, and Justin Guinney. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics*, 14(1):7, December 2013.

[101] B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Routledge, 1 edition, February 2018.

[102] Antonio Canale and David B. Dunson. Bayesian Kernel Mixtures for Counts. *Journal of the American Statistical Association*, 106(496):1528–1539, December 2011.

[103] Michael S. Rooney, Sachet A. Shukla, Catherine J. Wu, Gad Getz, and Nir Hacohen. Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity. *Cell*, 160(1-2):48–61, January 2015.

[104] Junwei Han, Xinrui Shi, Yunpeng Zhang, Yanjun Xu, Ying Jiang, Chunlong Zhang, Li Feng, Haixiu Yang, Desi Shang, Zeguo Sun, Fei Su, Chunquan Li, and Xia Li. ESEA: Discovering the Dysregulated Pathways based on Edge Set Enrichment Analysis. *Scientific Reports*, 5(1):13044, October 2015.

[105] Junjie Su, Byung-Jun Yoon, and Edward R. Dougherty. Accurate and Reliable Cancer Classification Based on Probabilistic Inference of Pathway Activity. *PLoS ONE*, 4(12):e8161, December 2009.

[106] Department of Bioinformatics, Division of Medical Genomics, Medical Research Institute, Tokyo Medical and Dental University 24F M&D Tower Bldg, 1-5-45 Yushima, Bunkyo-ku, Tokyo, Japan, Kaoru Mogushi, and Hiroshi Tanaka. PathAct: a novel method for pathway analysis using gene expression profiles. *Bioinformation*, 9(8):394–400, April 2013.

[107] Nathalie Malo, James A Hanley, Sonia Cerquozzi, Jerry Pelletier, and Robert Nadon. Statistical practice in high-throughput screening data analysis. *Nature Biotechnology*, 24(2):167–175, February 2006.

[108] R. A. Irizarry. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, April 2003.

[109] Frank Emmert-Streib and Matthias Dehmer. Networks for systems biology: conceptual connection of data and function. *IET systems biology*, 5(3):185–207, 2011.

[110] Michael Schubert, Bertram Klinger, Martina Klünemann, Mathew J Garnett, Nils Blüthgen, and Julio Saez-Rodriguez. Perturbation-response genes reveal signaling footprints in cancer gene expression. *bioRxiv*, 2016.

[111] Y. Drier, M. Sheffer, and E. Domany. Pathway-based personalized analysis of cancer. *Proceedings of the National Academy of Sciences*, 110(16):6388–6393, April 2013.

[112] Trevor Hastie and Werner Stuetzle. Principal Curves. *Journal of the American Statistical Association*, 84(406):502–516, 1989. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1989.10478797.

[113] Lingjian Yang, Chrysanthi Ainali, Sophia Tsoka, and Lazaros G. Papageorgiou. Pathway activity inference for multiclass disease classification through a mathematical programming optimisation framework. *BMC Bioinformatics*, 15(1):390, December 2014.

[114] Thomas Brott and Julien Bogousslavsky. Treatment of Acute Ischemic Stroke. *New England Journal of Medicine*, 343(10):710–722, September 2000.

[115] et.al. Feigin. Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet Neurology*, 20(10):795–820, October 2021.

[116] Amy K Saenger and Robert H Christenson. Stroke Biomarkers: Progress and Challenges for Diagnosis, Prognosis, Differentiation, and Treatment. *Clinical Chemistry*, 56(1):21–33, January 2010.

[117] William J Taylor, Annie Wong, Richard J Siegert, and Harry K McNaughton. Effectiveness of a clinical pathway for acute stroke care in a district general hospital: an audit. *BMC Health Services Research*, 6(1):16, December 2006.

[118] Yang Tang, Huichun Xu, Xin Li Du, Lisa Lit, Wynn Walker, Aigang Lu, Ruiqiong Ran, Jeffrey P Gregg, Melinda Reilly, Art Pancioli, Jane C Khoury, Laura R Sauerbeck, Janice A Carrozzella, Judith Spilker, Joseph Clark, Kenneth R Wagner, Edward C Jauch, Dongwoo J Chang, Piero Verro, Joseph P Broderick, and Frank R Sharp. Gene Expression in Blood Changes Rapidly in Neutrophils and Monocytes after Ischemic Stroke in Humans: A Microarray Study. *Journal of Cerebral Blood Flow & Metabolism*, 26(8):1089–1102, August 2006.

[119] Jaime Ramos-Cejudo, María Gutiérrez-Fernández, Berta Rodríguez-Frutos, Mercedes Expósito Alcaide, Fátima Sánchez-Cabo, Ana Dopazo, and Exuperio Díez–Tejedor. Spatial and Temporal Gene Expression Differences in Core and Periinfarct Areas in Experimental Stroke: A Microarray Analysis. *PLoS ONE*, 7(12):e52121, December 2012.

[120] Tiago Krug, João Paulo Gabriel, Ricardo Taipa, Benedita V Fonseca, Sophie Domingues-Montanari, Israel Fernandez-Cadenas, Helena Manso, Liliana O Gouveia, João Sobral, Isabel Albergaria, Gisela Gaspar, Jordi Jiménez-Conde, Raquel Rabionet, José M Ferro, Joan Montaner, Astrid M Vicente, Mário Rui Silva, Ilda Matos, Gabriela Lopes, and Sofia A Oliveira. *TTC7B* Emerges as a Novel Risk Factor for Ischemic Stroke Through the Convergence of Several Genome-Wide Approaches. *Journal of Cerebral Blood Flow & Metabolism*, 32(6):1061–1072, June 2012.

[121] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000.

[122] T. L. Barr, Y. Conley, J. Ding, A. Dillman, S. Warach, A. Singleton, and M. Matarin. Genomic biomarkers and cellular pathways of ischemic stroke by RNA gene expression profiling. *Neurology*, 75(11):1009–1014, September 2010.

[123] Boryana Stamova, Glen C. Jickling, Bradley P. Ander, Xinhua Zhan, DaZhi Liu, Renee Turner, Carolyn Ho, Jane C. Khoury, Cheryl Bushnell, Arthur Pancioli, Edward C. Jauch, Joseph P. Broderick, and Frank R. Sharp. Gene Expression in Peripheral Immune Cells following Cardioembolic Stroke Is Sexually Dimorphic. *PLoS ONE*, 9(7):e102550, July 2014.

[124] Jo Vandesompele, Katleen De Preter, Filip Pattyn, Bruce Poppe, Nadine Van Roy, Anne De Paepe, and Frank Speleman. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biology*, 3(7):research0034.1, 2002.

[125] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, June 2001.

[126] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data - SIGMOD '00*, pages 93–104, Dallas, Texas, United States, 2000. ACM Press.

[127] Michael R. Bussieck and Alex Meeraus. General Algebraic Modeling System (GAMS). In Panos M. Pardalos, Donald W. Hearn, and Josef Kallrath, editors, *Modeling Languages in Mathematical Optimization*, volume 88, pages 137–157. Springer US, Boston, MA, 2004. Series Title: Applied Optimization.

[128] Christian Bliek1ú, Pierre Bonami, and Andrea Lodi. Solving mixed-integer quadratic programming problems with ibm-cplex: a progress report. In *Proceedings of the twenty-sixth RAMP symposium*, pages 16–17, 2014.

[129] Hongyu Diao, Xinxing Li, Sheng Hu, and Yunhui Liu. Gene Expression Profiling Combined with Bioinformatics Analysis Identify Biomarkers for Parkinson Disease. *PLoS ONE*, 7(12):e52319, December 2012.

[130] Gokhan Yilmaz and D. Neil Granger. Cell adhesion molecules and ischemic stroke. *Neurological Research*, 30(8):783–793, October 2008.

[131] Dean Wu, Yuan-Chii G. Lee, Hsing-Cheng Liu, Rey-Yue Yuan, Hung-Yi Chiou, Chia-Hsiu Hung, and Chaur-Jong Hu. Identification of TLR downstream pathways in stroke patients. *Clinical Biochemistry*, 46(12):1058–1064, August 2013.

[132] Konstantinos Theofilatos, Aigli Korfiati, Seferina Mavroudi, Matthew C. Cowperthwaite, and Max Shpak. Discovery of stroke-related blood biomarkers from gene expression network models. *BMC Medical Genomics*, 12(1):118, December 2019.

[133] Cheng Fan, Aleix Prat, Joel S Parker, Yufeng Liu, Lisa A Carey, Melissa A Troester, and Charles M Perou. Building prognostic models for breast cancer patients using clinical variables and hundreds of gene expression signatures. *BMC Medical Genomics*, 4(1):3, December 2011.

[134] Laura J. van 't Veer, Hongyue Dai, Marc J. van de Vijver, Yudong D. He, Augustinus A. M. Hart, Mao Mao, Hans L. Peterse, Karin van der Kooy, Matthew J. Marton, Anke T. Witteveen, George J. Schreiber, Ron M. Kerkhoven, Chris Roberts, Peter S. Linsley, René Bernards, and Stephen H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536, January 2002.

[135] Robin K. Kelley and Alan P. Venook. Prognostic and Predictive Markers in Stage II Colon Cancer: Is There a Role for Gene Expression Profiling? *Clinical Colorectal Cancer*, 10(2):73–80, June 2011.

[136] Laetitia Marisa, Aurélien de Reyniès, Alex Duval, Janick Selves, Marie Pierre Gaub, Laure Vescovo, Marie-Christine Etienne-Grimaldi, Renaud Schiappa, Dominique Guenot, Mira Ayadi, Sylvain Kirzin, Maurice Chazal, Jean-François Fléjou, Daniel Benchimol, Anne Berger, Arnaud Lagarde, Erwan Pencreach, Françoise Piard, Dominique Elias, Yann Parc, Sylviane Olschwang, Gérard Milano, Pierre Laurent-Puig, and Valérie Boige. Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value. *PLoS Medicine*, 10(5):e1001453, May 2013.

[137] Hongtao Liu, Qing Zhang, Qianqian Lou, Xin Zhang, Yunxia Cui, Panpan Wang, Fan Yang, Fan Wu, Jing Wang, Tianli Fan, and Shenglei Li. Differential Analysis of lncRNA, miRNA and mRNA Expression Profiles and the Prognostic Value of lncRNA in Esophageal Cancer. *Pathology & Oncology Research*, 26(2):1029–1039, April 2020.

[138] Eva Budinska, Vlad Popovici, Sabine Tejpar, Giovanni D'Ario, Nicolas Lapique, Katarzyna Otylia Sikora, Antonio Fabio Di Narzo, Pu Yan, John Graeme Hodgson, Scott Weinrich, Fred Bosman, Arnaud Roth, and Mauro Delorenzi. Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *The Journal of Pathology*, 231(1):63–76, September 2013.

[139] Xian-guo Zhou, Xiao-liang Huang, Si-yuan Liang, Shao-mei Tang, Si-kao Wu, Tong-tong Huang, Zeng-nan Mo, and Qiu-yan Wang. Identifying miRNA and gene modules of colon cancer associated with pathological stage by weighted gene co-expression network analysis. *OncoTargets and Therapy*, Volume 11:2815–2830, May 2018.

[140] Zhong Shi, Xiaofu Yu, Meiqin Yuan, Wangxia Lv, Tingting Feng, Rui Bai, and Haijun Zhong. Activation of the PERK-ATF4 pathway promotes chemo-resistance in colon cancer cells. *Scientific Reports*, 9(1):3210, December 2019.

[141] Xiao Yang, ShaoMing Zhu, Li Li, Li Zhang, Shu Xian, Yanqing Wang, and Yanxiang Cheng. Identification of differentially expressed genes and signaling pathways in ovarian cancer by integrated bioinformatics analysis. *OncoTargets and Therapy*, Volume 11:1457–1474, March 2018.

[142] Zhe Liu and Matthew Page. A novel gene and pathway-level subtyping analysis scheme to understand biological mechanisms in complex disease: a case study in rheumatoid arthritis. *Genomics*, 111(3):375–382, May 2019.

[143] Jing Huang. Current developments of targeting the p53 signaling pathway for cancer treatment. *Pharmacology & Therapeutics*, 220:107720, April 2021.

[144] Daniel B. Graham and Ramnik J. Xavier. Pathway paradigms revealed from the genetics of inflammatory bowel disease. *Nature*, 578(7796):527–539, February 2020.

[145] Katrina Ray. A pathway to therapy for HER2+ metastatic biliary tract cancer. *Nature Reviews Gastroenterology & Hepatology*, 18(10):676–676, October 2021.

[146] M. Kanehisa. The KEGG resource for deciphering the genome. *Nucleic Acids Research*, 32(90001):277D–280, January 2004.

[147] Luciano Garofano, Simona Migliozzi, Young Taek Oh, Fulvio D'Angelo, Ryan D. Najac, Aram Ko, Brulinda Frangaj, Francesca Pia Caruso, Kai Yu, Jinzhou Yuan, Wenting Zhao, Anna Luisa Di Stefano, Franck Bielle, Tao Jiang, Peter Sims, Mario L. Suvà, Fuchou Tang, Xiao-Dong Su, Michele Ceccarelli, Marc Sanson, Anna Lasorella, and Antonio Iavarone. Pathway-based classification of glioblastoma uncovers a mitochondrial subtype with therapeutic vulnerabilities. *Nature Cancer*, 2(2):141–156, February 2021.

[148] Sijia Huang, Cameron Yee, Travers Ching, Herbert Yu, and Lana X. Garmire. A Novel Model to Combine Clinical and Pathway-Based Transcriptomic Information for the Prognosis Prediction of Breast Cancer. *PLoS Computational Biology*, 10(9):e1003851, September 2014.

[149] Ke-Qin Liu, Zhi-Ping Liu, Jin-Kao Hao, Luonan Chen, and Xing-Ming Zhao. Identifying dysregulated pathways in cancers from pathway interaction networks. *BMC Bioinformatics*, 13(1):126, December 2012.

[150] Xi Chen and Lily Wang. Integrating Biological Knowledge with Gene Expression Profiles for Survival Prediction of Cancer. *Journal of Computational Biology*, 16(2):265–278, February 2009.

[151] Ruoting Yang, Bernie J Daigle, Linda R Petzold, and Francis J Doyle. Core module biomarker identification with network exploration for breast cancer metastasis. *BMC Bioinformatics*, 13(1):12, December 2012.

[152] Ioulia Karagiannaki, Yannis Pantazis, Ekaterini Chatzaki, and Ioannis Tsamardinos. Pathway Activity Score Learning for Dimensionality Reduction of Gene Expression Data. In Annalisa Appice, Grigorios Tsoumakas, Yannis Manolopoulos, and Stan Matwin, editors, *Discovery Science*, volume 12323, pages 246–261. Springer International Publishing, Cham, 2020. Series Title: Lecture Notes in Computer Science.

[153] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628, July 2008.

[154] Ugrappa Nagalakshmi, Karl Waern, and Michael Snyder. RNA-Seq: A Method for Comprehensive Transcriptome Analysis. *Current Protocols in Molecular Biology*, 89(1), January 2010.

[155] Stephen B. Montgomery, Micha Sammeth, Maria Gutierrez-Arcelus, Radoslaw P. Lach, Catherine Ingle, James Nisbett, Roderic Guigo, and Emmanouil T. Dermitzakis. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, 464(7289):773–777, April 2010.

[156] Marc Beyer, Michael R. Mallmann, Jia Xue, Andrea Staratschek-Jox, Daniela Vorholt, Wolfgang Krebs, Daniel Sommer, Jil Sander, Christina Mertens, Andrea Nino-Castro, Susanne V. Schmidt, and Joachim L. Schultze. High-Resolution Transcriptome of Human Macrophages. *PLoS ONE*, 7(9):e45466, September 2012.

[157] Kai-Oliver Mutz, Alexandra Heilkenbrinker, Maren Lönne, Johanna-Gabriela Walter, and Frank Stahl. Transcriptome analysis using next-generation sequencing. *Current Opinion in Biotechnology*, 24(1):22–30, February 2013.

[158] Shanrong Zhao, Wai-Ping Fung-Leung, Anton Bittner, Karen Ngo, and Xuejun Liu. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PLoS ONE*, 9(1):e78644, January 2014.

[159] Yvette Temate-Tiagueu, Sahar Al Seesi, Meril Mathew, Igor Mandric, Alex Rodriguez, Kayla Bean, Qiong Cheng, Olga Glebova, Ion Măndoiu, Nicole B. Lopanik, and Alexander Zelikovsky. Inferring metabolic pathway activity levels from RNA-Seq data. *BMC Genomics*, 17(S5):542, August 2016.

[160] Christian H. Holland, Jovan Tanevski, Javier Perales-Patón, Jan Gleixner, Manu P. Kumar, Elisabetta Mereu, Brian A. Joughin, Oliver Stegle, Douglas A. Lauffenburger, Holger Heyn, Bence Szalai, and Julio Saez-Rodriguez. Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome Biology*, 21(1):36, December 2020.

[161] Sangseon Lee, Youngjune Park, and Sun Kim. MIDAS: Mining differentially activated subpaths of KEGG pathways from multi-class RNA-seq data. *Methods*, 124:13–24, July 2017.

[162] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Współczesna Onkologia*, 1A:68–77, 2015.

[163] Anguraj Sadanandam, Costas A Lyssiotis, Krisztian Homicsko, Eric A Collisson, William J Gibb, Stephan Wullschleger, Liliane C Gonzalez Ostos, William A Lannon, Carsten Grotzinger, Maguy Del Rio, Benoit Lhermitte, Adam B Olshen, Bertram Wiedenmann, Lewis C Cantley, Joe W Gray, and Douglas Hanahan. A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nature Medicine*, 19(5):619–625, May 2013.

[164] Paul Roepman, Andreas Schlicker, Josep Tabernero, Ian Majewski, Sun Tian, Victor Moreno, Mireille H Snel, Christine M Chresta, Robert Rosenberg, Ulrich Nitsche, Teresa Macarulla, Gabriel Capella, Ramon Salazar, George Orphanides, Lodewyk FA Wessels, Rene Bernards, and Iris M Simon. Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. *International Journal of Cancer*, 134(3):552–562, February 2014.

[165] Felipe De Sousa E Melo, Xin Wang, Marnix Jansen, Evelyn Fessler, Anne Trinh, Laura P M H de Rooij, Joan H de Jong, Onno J de Boer, Ronald van Leersum, Maarten F Bijlsma, Hans Rodermond, Maartje van der Heijden, Carel J M van Noesel, Jurriaan B Tuynman, Evelien Dekker, Florian Markowetz, Jan Paul Medema, and Louis Vermeulen. Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nature Medicine*, 19(5):614–618, May 2013.

[166] Andreas Schlicker, Garry Beran, Christine M Chresta, Gael McWalter, Alison Pritchard, Susie Weston, Sarah Runswick, Sara Davenport, Kerry Heathcote, Denis Alferez Castro, George Orphanides, Tim French, and Lodewyk FA Wessels. Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. *BMC Medical Genomics*, 5(1):66, December 2012.

[167] Justin Guinney, Rodrigo Dienstmann, Xin Wang, Aurélien de Reyniès, Andreas Schlicker, Charlotte Soneson, Laetitia Marisa, Paul Roepman, Gift Nyamundanda, Paolo Angelino, Brian M Bot, Jeffrey S Morris, Iris M Simon, Sarah Gerster, Evelyn Fessler, Felipe De Sousa E Melo, Edoardo Missiaglia, Hena Ramay, David Barras, Krisztian Homicsko, Dipen Maru, Ganiraju C Manyam, Bradley Broom, Valerie Boige, Beatriz Perez-Villamil, Ted Laderas, Ramon Salazar, Joe W Gray, Douglas Hanahan, Josep Tabernero, Rene Bernards, Stephen H Friend, Pierre Laurent-Puig, Jan Paul Medema, Anguraj Sadanandam, Lodewyk Wessels, Mauro Delorenzi, Scott Kopetz, Louis Vermeulen, and Sabine Tejpar. The consensus molecular subtypes of colorectal cancer. *Nature Medicine*, 21(11):1350–1356, November 2015.

[168] Leili Shahriyari. Effect of normalization methods on the performance of supervised learning algorithms applied to HTSeq-FPKM-UQ data sets: 7SK RNA expression as a predictor of survival in patients with colon adenocarcinoma. *Briefings in Bioinformatics*, 20(3):985–994, May 2019.

[169] Tadayoshi Fushiki. Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing*, 21(2):137–146, April 2011.

[170] M. Kanehisa. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, January 2000.

[171] Jaime Lynn Speiser, Michael E. Miller, Janet Tooze, and Edward Ip. A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications*, 134:93–101, November 2019.

[172] Jiangang Hao and Tin Kam Ho. Machine Learning Made Easy: A Review of *Scikit-learn* Package in Python Programming Language. *Journal of Educational and Behavioral Statistics*, 44(3):348–361, June 2019.

[173] P{\"o}lsterl Sebastian. scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020.

[174] Enrico Longato, Martina Vettoretti, and Barbara Di Camillo. A practical perspective on the concordance index for the evaluation and selection of prognostic time-to-event models. *Journal of Biomedical Informatics*, 108:103496, August 2020.

[175] So Yeon Kim, Hyun-Hwan Jeong, Jaesik Kim, Jeong-Hyeon Moon, and Kyung-Ah Sohn. Robust pathway-based multi-omics data integration using directed random

walks for survival prediction in multiple cancer studies. *Biology Direct*, 14(1):8, December 2019.

[176] Zhou Yang, Yusheng Chen, Dejun Wu, Zhijun Min, and Yingjun Quan. Analysis of risk factors for colon cancer progression. *OncoTargets and Therapy*, Volume 12:3991–4000, May 2019.

[177] Mao-lin Wan, Yu Wang, Zhi Zeng, Bo Deng, Bi-sheng Zhu, Ting Cao, Yu-kun Li, Jiao Xiao, Qi Han, and Qing Wu. Colorectal cancer (CRC) as a multifactorial disease and its causal correlations with multiple signaling pathways. *Bioscience Reports*, 40(3):BSR20200265, March 2020.

[178] Yihang Yuan, Ji Chen, Jue Wang, Ming Xu, Yunpeng Zhang, Peng Sun, and Leilei Liang. Development and Clinical Validation of a Novel 4-Gene Prognostic Signature Predicting Survival in Colorectal Cancer. *Frontiers in Oncology*, 10:595, May 2020.

[179] Edward Saehong Oh and Art Petronis. Origins of human disease: the chrono-epigenetic perspective. *Nature Reviews Genetics*, 22(8):533–546, August 2021.

[180] Daniela Mancarella and Christoph Plass. Epigenetic signatures in cancer: proper controls, current challenges and the potential for clinical translation. *Genome Medicine*, 13(1):23, December 2021.

[181] P. W. Franks, E. Melén, M. Friedman, J. Sundström, I. Kockum, L. Klareskog, C. Almqvist, S. E. Bergen, K. Czene, S. Hägg, P. Hall, K. Johnell, A. Malarstig, A. Catrina, H. Hagström, M. Benson, J. Gustav Smith, M. F Gomez, M. Orho-Melander, B. Jacobsson, J. Halfvarson, D. Repsilber, M. Oresic, C. Jern, B. Melin, C. Ohlsson, T. Fall, L. Rönnblom, M. Wadelius, G. Nordmark, A. Johansson, R. Rosenquist, and P. F. Sullivan. Technological readiness and implementation of genomic-driven precision medicine for complex diseases. *Journal of Internal Medicine*, 290(3):602–620, September 2021.

[182] Marie-Claire Wagle, Daniel Kirouac, Christiaan Klijn, Bonnie Liu, Shilpi Mahajan, Melissa Junttila, John Moffat, Mark Merchant, Ling Huw, Matthew Wongchenko, Kwame Okrah, Shrividhya Srinivasan, Zineb Mounir, Teiko Sumiyoshi, Peter M. Haverty, Robert L. Yauch, Yibing Yan, Omar Kabbarah, Garret Hampton, Lukas Amler, Saroja Ramanujan, Mark R. Lackner, and Shih-Min A. Huang. A transcriptional MAPK Pathway Activity Score (MPAS) is a clinically relevant biomarker in multiple cancer types. *npj Precision Oncology*, 2(1):7, December 2018.

[183] Anja van de Stolpe, Laurent Holtzer, Henk van Ooijen, Marcia Alves de Inda, and Wim Verhaegh. Enabling precision medicine by unravelling disease pathophysiology: quantifying signal transduction pathway activity across cell and tissue types. *Scientific Reports*, 9(1):1603, December 2019.

[184] Christian H. Holland, Bence Szalai, and Julio Saez-Rodriguez. Transfer of regulatory knowledge from human to mouse for functional genomics analysis. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1863(6):194431, June 2020.

[185] Yongnan Chen, Konstantinos Theofilatos, Lazaros G. Papageorgiou, and Sophia Tsoka. Identification of Important Biological Pathways for Ischemic Stroke Prediction through a Mathematical Programming Optimisation Model-DIGS. In *Proceedings of the 2020 12th International Conference on Bioinformatics and*

Biomedical Technology, ICBBT 2020, pages 25–31, New York, NY, USA, May 2020. Association for Computing Machinery.

[186] Yan Zhang, Ju Xiang, Liang Tang, Jianming Li, Qingqing Lu, Geng Tian, Bin-Sheng He, and Jialiang Yang. Identifying Breast Cancer-Related Genes Based on a Novel Computational Framework Involving KEGG Pathways and PPI Network Modularity. *Frontiers in Genetics*, 12:596794, August 2021.

[187] Hannah E. Wilson, David A. Stanton, Cortney Montgomery, Aniello M. Infante, Matthew Taylor, Hannah Hazard-Jenkins, Elena N. Pugacheva, and Emidio E. Pistilli. Skeletal muscle reprogramming by breast cancer regardless of treatment history or tumor molecular subtype. *npj Breast Cancer*, 6(1):18, December 2020.

[188] Joel S. Parker, Michael Mullins, Maggie C.U. Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, John F. Quackenbush, Inge J. Stijleman, Juan Palazzo, J.S. Marron, Andrew B. Nobel, Elaine Mardis, Torsten O. Nielsen, Matthew J. Ellis, Charles M. Perou, and Philip S. Bernard. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *Journal of Clinical Oncology*, 27(8):1160–1167, March 2009.

[189] James C Mathews, Saad Nadeem, Arnold J Levine, Maryam Pouryahya, Joseph O Deasy, and Allen Tannenbaum. Robust and interpretable pam50 reclassification exhibits survival advantage for myoepithelial and immune phenotypes. *NPJ Breast Cancer*, 5(1):1–8, 2019.

[190] Paul Black. Dictionary of Algorithms and Data Structures (DADS), 2017. Type: dataset.

[191] Patrick-Simon Welz and S.A. Benitah. Molecular Connections Between Circadian Clocks and Aging. *Journal of Molecular Biology*, 432(12):3661–3679, May 2020.

[192] Supriya Bevinakoppamath, Shobha Chikkavaddaragudi Ramachandra, Anshu Kumar Yadav, Vijaya Basavaraj, Prashant Vishwanath, and Akila Prashant. Understanding the Emerging Link Between Circadian Rhythm, Nrf2 Pathway, and Breast Cancer to Overcome Drug Resistance. *Frontiers in Pharmacology*, 12:719631, January 2022.

[193] Jung-Ae Kim. Peroxisome Metabolism in Cancer. *Cells*, 9(7):1692, July 2020.

[194] Guido Cavaletti, Paola Alberti, and Paola Marmiroli. Chemotherapy-induced peripheral neurotoxicity in the era of pharmacogenomics. *The Lancet Oncology*, 12(12):1151–1161, November 2011.

[195] Yufeng Sun, Wenchao Li, Shiqi Shen, Xuejing Yang, Bing Lu, Xiaojing Zhang, Peng Lu, Yi Shen, and Juling Ji. Loss of alanine-glyoxylate and serine-pyruvate aminotransferase expression accelerated the progression of hepatocellular carcinoma and predicted poor prognosis. *Journal of Translational Medicine*, 17(1):390, December 2019.

[196] P Chen, C Li, X Li, J Li, R Chu, and H Wang. Higher dietary folate intake reduces the breast cancer risk: a systematic review and meta-analysis. *British Journal of Cancer*, 110(9):2327–2338, April 2014.

[197] Priti Tagde, Giriraj T Kulkarni, Dinesh Kumar Mishra, and Prashant Kesharwani. Recent advances in folic acid engineered nanocarriers for treatment of breast cancer. *Journal of Drug Delivery Science and Technology*, 56:101613, April 2020.

[198] Shunda Wang, Jinshou Yang, Lei You, Menghua Dai, and Yupei Zhao. GSTM3 Function and Polymorphism in Cancer: Emerging but Promising. *Cancer Management and Research*, Volume 12:10377–10388, October 2020.

[199] Mary S. Wolff, Julie A. Britton, and Valerie P. Wilson. Environmental risk factors for breast cancer among African-American women. *Cancer*, 97(S1):289–310, January 2003.

[200] Christine B. Ambrosone, Brian F. Coles, Jo L. Freudenheim, and Peter G. Shields. Glutathione-S-transferase (GSTM1) Genetic Polymorphisms Do Not Affect Human Breast Cancer Risk, Regardless of Dietary Antioxidants. *The Journal of Nutrition*, 129(2):565S–568S, February 1999.

[201] B.L Weber and K.L Nathanson. Low penetrance genes associated with increased risk for breast cancer. *European Journal of Cancer*, 36(10):1193–1199, June 2000.

[202] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, Lars J Jensen, and Christian von Mering. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):D607–D613, January 2019.

[203] Xiujuan Lei, Yuchen Zhang, Shi Cheng, Fang-Xiang Wu, and Witold Pedrycz. Topology potential based seed-growth method to identify protein complexes on dynamic PPI data. *Information Sciences*, 425:140–153, January 2018.

[204] Melissa S Cline, Michael Smoot, Ethan Cerami, Allan Kuchinsky, Nerius Landys, Chris Workman, Rowan Christmas, Iliana Avila-Campilo, Michael Creech, Benjamin Gross, Kristina Hanspers, Ruth Isserlin, Ryan Kelley, Sarah Killcoyne, Samad Lotia, Steven Maere, John Morris, Keiichiro Ono, Vuk Pavlovic, Alexander R Pico, Aditya Vailaya, Peng-Liang Wang, Annette Adler, Bruce R Conklin, Leroy Hood, Martin Kuiper, Chris Sander, Ilya Schmulevich, Benno Schwikowski, Guy J Warner, Trey Ideker, and Gary D Bader. Integration of biological networks and gene expression data using Cytoscape. *Nature Protocols*, 2(10):2366–2382, October 2007.

[205] C P S Potter and A L Harris. Diagnostic, prognostic and therapeutic implications of carbonic anhydrases in cancer. *British Journal of Cancer*, 89(1):2–7, July 2003.

[206] Rachel Sutherland, Annette Meeson, and Simon Lowes. Solute transporters and malignancy: establishing the role of uptake transporters in breast cancer and breast cancer metastasis. *Cancer and Metastasis Reviews*, 39(3):919–932, September 2020.

[207] Wojciech Jelski and Maciej Szmitkowski. Alcohol dehydrogenase (ADH) and aldehyde dehydrogenase (ALDH) in the cancer diseases. *Clinica Chimica Acta*, 395(1-2):1–5, September 2008.

[208] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[209] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, January 2020.

[210] Valentine Svensson, Roser Vento-Tormo, and Sarah A Teichmann. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols*, 13(4):599–604, April 2018.

[211] Xinmin Li and Cun-Yu Wang. From bulk, single-cell to spatial RNA sequencing. *International Journal of Oral Science*, 13(1):36, December 2021.

[212] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, Kaiqin Lao, and M Azim Surani. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382, May 2009.

[213] Joep Beumer, Jens Puschhof, Julia Bauzá-Martinez, Adriana Martínez-Silgado, Rasa Elmentaite, Kylie R. James, Alexander Ross, Delilah Hendriks, Benedetta Artegiani, Georg A. Busslinger, Bas Ponsioen, Amanda Andersson-Rolf, Aurelia Saftien, Charelle Boot, Kai Kretzschmar, Maarten H. Geurts, Yotam E. Bar-Ephraim, Cayetano Pleguezuelos-Manzano, Yorick Post, Harry Begthel, Franka van der Linden, Carmen Lopez-Iglesias, Willine J. van de Wetering, Reinier van der Linden, Peter J. Peters, Albert J.R. Heck, Joachim Goedhart, Hugo Snippert, Matthias Zilbauer, Sarah A. Teichmann, Wei Wu, and Hans Clevers. High-Resolution mRNA and Secretome Atlas of Human Enteroendocrine Cells. *Cell*, 181(6):1291–1306.e19, June 2020.

[214] Oliver Stegle, Sarah A. Teichmann, and John C. Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145, March 2015.

[215] Michael Schubert, Bertram Klinger, Martina Klünemann, Anja Sieber, Florian Uhlitz, Sascha Sauer, Mathew J. Garnett, Nils Blüthgen, and Julio Saez-Rodriguez. Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nature Communications*, 9(1):20, December 2018.

[216] Luz Garcia-Alonso, Christian H. Holland, Mahmoud M. Ibrahim, Denes Turei, and Julio Saez-Rodriguez. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Research*, 29(8):1363–1375, August 2019.

[217] Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, Joost van den Oord, Zeynep Kalender Atak, Jasper Wouters, and Stein Aerts. SCENIC: single-cell regulatory network inference and clustering. *Nature Methods*, 14(11):1083–1086, November 2017.

[218] Hongxu Ding, Eugene F. Douglass, Adam M. Sonabend, Angeliki Mela, Sayantan Bose, Christian Gonzalez, Peter D. Canoll, Peter A. Sims, Mariano J. Alvarez, and Andrea Califano. Quantitative assessment of protein activity in orphan tissues and single cells using the metaVIPER algorithm. *Nature Communications*, 9(1):1471, December 2018.

[219] Yaru Zhang, Yunlong Ma, Yukuan Huang, Yan Zhang, Qi Jiang, Meng Zhou, and Jianzhong Su. Benchmarking algorithms for pathway activity transformation of single-cell RNA-seq data. *Computational and Structural Biotechnology Journal*, 18:2953–2961, 2020.

[220] Yan Zhang, Yaru Zhang, Jun Hu, Ji Zhang, Fangjie Guo, Meng Zhou, Guijun Zhang, Fulong Yu, and Jianzhong Su. scTPA: a web tool for single-cell transcriptome analysis of pathway activation signatures. *Bioinformatics*, 36(14):4217–4219, July 2020.

[221] Wenpin Hou and Zhicheng Ji. Unbiased visualization of single-cell genomic data with SCUBI. *Cell Reports Methods*, 2(1):100135, January 2022.

[222] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11), 2008.

[223] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018. Publisher: arXiv Version Number: 3.

[224] Saskia Freytag, Luyi Tian, Ingrid Lönnstedt, Milica Ng, and Melanie Bahlo. Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. *F1000Research*, 7:1297, December 2018.

[225] Elham Azizi, Ambrose J. Carr, George Plitas, Andrew E. Cornish, Catherine Konopacki, Sandhya Prabhakaran, Juozas Nainys, Kenmin Wu, Vaidotas Kiseliovas, Manu Setty, Kristy Choi, Rachel M. Fromme, Phuong Dao, Peter T. McKenney, Ruby C. Wasti, Krishna Kadaveru, Linas Mazutis, Alexander Y. Rudensky, and Dana Pe'er. Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell*, 174(5):1293–1308.e36, August 2018.

[226] Pau Badia-i Mompel, Jesús Vélez Santiago, Jana Braunger, Celina Geiss, Daniel Dimitrov, Sophia Müller-Dott, Petr Taus, Aurelien Dugourd, Christian H Holland, Ricardo O Ramirez Flores, and Julio Saez-Rodriguez. decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinformatics Advances*, 2(1):vbac016, January 2022.

[227] Mihriban Karaayvaz, Simona Cristea, Shawn M. Gillespie, Anoop P. Patel, Ravindra Mylvaganam, Christina C. Luo, Michelle C. Specht, Bradley E. Bernstein, Franziska Michor, and Leif W. Ellisen. Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. *Nature Communications*, 9(1):3588, December 2018.

[228] Behzad Mansoori, Mikkel Green Terp, Ali Mohammadi, Christina Bøg Pedersen, Henrik Jørn Ditzel, Behzad Baradaran, and Morten Frier Gjerstorff. HMGA2 Supports Cancer Hallmarks in Triple-Negative Breast Cancer. *Cancers*, 13(20):5197, October 2021.

[229] Malte D Luecken and Fabian J Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6), June 2019.

[230] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, November 1987.

[231] Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12):1289–1296, December 2019.

[232] Hongxu Ding, Andrew Blair, Ying Yang, and Joshua M. Stuart. Biological process activity transformation of single cell gene expression for cross-species alignment. *Nature Communications*, 10(1):4899, December 2019.

[233] Blue B Lake, Song Chen, Brandon C Sos, Jean Fan, Gwendolyn E Kaeser, Yun C Yung, Thu E Duong, Derek Gao, Jerold Chun, Peter V Kharchenko, and Kun Zhang. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nature Biotechnology*, 36(1):70–80, January 2018.

[234] Arti Parihar, Timothy D. Eubank, and Andrea I. Doseff. Monocytes and Macrophages Regulate Immunity through Dynamic Networks of Survival and Cell Death. *Journal of Innate Immunity*, 2(3):204–215, 2010.

[235] Cornelius Y. Taabazuing, Marian C. Okondo, and Daniel A. Bachovchin. Pyroptosis and Apoptosis Pathways Engage in Bidirectional Crosstalk in Monocytes and Macrophages. *Cell Chemical Biology*, 24(4):507–514.e4, April 2017.

[236] Isabel Barao and William J Murphy. The immunobiology of natural killer cells and bone marrow allograft rejection. *Biology of Blood and Marrow Transplantation*, 9(12):727–741, December 2003.

[237] William H. Kitchens, Shuichiro Uehara, Catharine M. Chase, Robert B. Colvin, Paul S. Russell, and Joren C. Madsen. The Changing Role of Natural Killer Cells in Solid Organ Rejection and Tolerance. *Transplantation*, 81(6):811–817, March 2006.

[238] Cristiana Barbati, Cristiano Alessandri, Marta Vomero, Rosa Vona, Tania Colasanti, Davide Vacirca, Serena Camerini, Marco Crescenzi, Monica Pendolino, Simona Truglia, Fabrizio Conti, Tina Garofalo, Maurizio Sorice, Marina Pierdominici, Guido Valesini, Walter Malorni, and Elena Ortona. Autoantibodies specific to D4GDI modulate Rho GTPase mediated cytoskeleton remodeling and induce autophagy in T lymphocytes. *Journal of Autoimmunity*, 58:78–89, April 2015.

[239] Soizic Daniel, Maria B. Arvelo, Virendra I. Patel, Christopher R. Longo, Gautam Shrikhande, Tala Shukri, Jerome Mahiou, David W. Sun, Christina Mottley, Shane T. Grey, and Christiane Ferran. A20 protects endothelial cells from TNF-, Fas-, and NK-mediated cell death by inhibiting caspase 8 activation. *Blood*, 104(8):2376–2384, October 2004.

[240] K A Sedelies, A Ciccone, C J P Clarke, J Oliaro, V R Sutton, F L Scott, J Silke, O Susanto, D R Green, R W Johnstone, P I Bird, J A Trapani, and N J Waterhouse. Blocking granule-mediated death by primary human NK cells requires both protection of mitochondria and inhibition of caspase activity. *Cell Death & Differentiation*, 15(4):708–717, April 2008.

[241] Colin Adrain, Brona M. Murphy, and Seamus J. Martin. Molecular Ordering of the Caspase Activation Cascade Initiated by the Cytotoxic T Lymphocyte/Natural Killer (CTL/NK) Protease Granzyme B. *Journal of Biological Chemistry*, 280(6):4663–4673, February 2005.

[242] Steinunn Thorlacius, Gudridur Olafsdottir, Laufey Tryggvadottir, Susan Neuhausen, Jon G. Jonasson, Sean V. Tavtigian, Hrafn Tulinius, Helga M. Ögmundsdottir, and Jorunn E. Eyfjörd. A single BRCA2 mutation in male and female breast cancer families from Iceland with varied cancer phenotypes. *Nature Genetics*, 13(1):117–119, May 1996.

[243] Evelina Mocci, Roger L. Milne, Elena Yuste Méndez-Villamil, John L. Hopper, Esther M. John, Irene L. Andrulis, Wendy K. Chung, Mary Daly, Saundra S. Buys, Nuria Malats, and David E. Goldgar. Risk of Pancreatic Cancer in Breast Cancer Families from the Breast Cancer Family Registry. *Cancer Epidemiology, Biomarkers & Prevention*, 22(5):803–811, May 2013.

[244] Erik R. Nelson, Ching-yi Chang, and Donald P. McDonnell. Cholesterol and breast cancer pathophysiology. *Trends in Endocrinology & Metabolism*, 25(12):649–655, December 2014.

[245] Carrie D. House, Valentina Grajales, Michelle Ozaki, Elizabeth Jordan, Helmae Wubneh, Danielle C. Kimble, Jana M. James, Marianne K. Kim, and Christina M. Annunziata. I$\kappa\kappa\varepsilon$ cooperates with either MEK or non-canonical NF-kB driving growth of triple-negative breast cancer cells in different contexts. *BMC Cancer*, 18(1):595, December 2018.

[246] Joyce G. Habib and Joyce A. O'Shaughnessy. The hedgehog pathway in triple-negative breast cancer. *Cancer Medicine*, 5(10):2989–3006, October 2016.

[247] Snahlata Singh, Sushil Kumar, Ratnesh Kumar Srivastava, Ajeya Nandi, Gatha Thacker, Hemma Murali, Sabrina Kim, Mary Baldeon, John Tobias, Mario Andres Blanco, Rizwan Saffie, M. Raza Zaidi, Satrajit Sinha, Luca Busino, Serge Y. Fuchs, and Rumela Chakrabarti. Loss of ELF5–FBXW7 stabilizes IFNGR1 to promote the growth and metastasis of triple-negative breast cancer through interferon-γ signalling. *Nature Cell Biology*, 22(5):591–602, May 2020.

[248] Nami Yamashita, Mark Long, Atsushi Fushimi, Masaaki Yamamoto, Tsuyoshi Hata, Masayuki Hagiwara, Atrayee Bhattacharya, Qiang Hu, Kwok-Kin Wong, Song Liu, and Donald Kufe. MUC1-C integrates activation of the IFN-γ pathway with suppression of the tumor immune microenvironment in triple-negative breast cancer. *Journal for ImmunoTherapy of Cancer*, 9(1):e002115, January 2021.

[249] Tuan-Minh Nguyen, Adib Shafi, Tin Nguyen, and Sorin Draghici. Identifying significantly impacted pathways: a comprehensive review and assessment. *Genome biology*, 20(1):1–15, 2019.

[250] Eva Baxter, Karolina Windloch, Frank Gannon, and Jason S Lee. Epigenetic regulation in cancer progression. *Cell & bioscience*, 4(1):1–11, 2014.

Appendix A

DIGS released COAD relevant pathways

A.1 Ranking for KEGG Pathways for COAD

Table A.1 Top 50 KEGG Pathways of COAD (DIGS)

Pathway	**No. Gene**	**Genes with weights (Ranked High to Low)**				
Primary bile acid biosynthesis	18	ACOT8	HSD17B4	CYP7B1	SCP2	SLC27A5
Other glycan degradation	17	MAN2B1	ENGASE	MANBA	FUCA1	GLB1
Sulfur relay system	7	MOCS3	URM1	CTU2	NFS1	MPST
Glycosaminoglycan biosynthesis keratan sulfate	14	CHST4	ST3GAL3	CHST1	FUT8	CHST6
Maturity onset diabetes of the young	27	RFX6	MNX1	PKLR	HNF4A	GCK
Pantothenate and CoA biosynthesis	22	VNN3	BCAT1	PANK1	VNN1	ENPP3
Non homologous end joining	13	MRE11	XRCC4	FEN1	PRKDC	POLM
2 Oxocarboxylic acid metabolism	19	GPT	BCAT1	BCAT2	ABHD14A-ACY1	IDH3A
One carbon pool by folate	21	FTCD	TYMS	ATIC	ALDH1L2	AMT

Continued on next page

Table A.1 – *Continued from previous page*

Pathway	No. Gene	Genes with weights (Ranked High to Low)				
Glycosaminoglycan biosynthesis chondroitin sulfate dermatan sulfate	20	CHST12	CHST13	DSE	CHSY3	CHST11
Proximal tubule bicarbonate reclamation	25	CA4	CA2	ATP1A3	SLC4A4	SLC9A3
Renin angiotensin system	20	ENPEP	ACE2	ACE	KLK2	ANPEP
Vitamin digestion and absorption	25	SLC19A3	LRAT	SLC23A1	RBP2	CBLIF
Graft versus host disease	38	HLA-DMA	GZMB	PRF1	IFNG	KLRD1
Asthma	30	IL5	HLA-DMA	CCL11	HLA-DPB1	IL10
Phototransduction	28	SAG	GUCA1A	PDE6A	GRK7	CNGB1
Steroid hormone biosynthesis	64	CPN1	UGT2B10	DHRS11	UGT2B7	CYP7B1
Cocaine addiction	50	SLC18A1	DRD2	GRIN2B	BDNF	CREB3L3
Glycosylphosphatidylinositol (GPI) anchor biosynthesis	26	PIGU	PIGH	PIGB	GAB1	PIGC
Mismatch repair	23	RPA4	RFC3	RPA1	POLD3	MLH1
Primary immunodeficiency	37	IGLL1	AIRE	TNFRSF13C	BTK	ORAI1
Hippo signaling pathway multiple species	27	DCHS2	FAT4	FRMD6	RASSF6	TEAD3
Biosynthesis of unsaturated fatty acids	28	BAAT	ELOVL3	HACD4	ELOVL2	HACD1
Terpenoid backbone biosynthesis	21	HMGCS2	FDPS	ICMT	GGPS1	MVD
Type I diabetes mellitus	40	GAD1	HLA-DQA1	TNF	HLA-DRB5	FASLG
Hedgehog signaling pathway	55	GAS1	DHH	SMURF2	HHIP	EVC
Glycosaminoglycan biosynthesis heparan sulfate heparin	24	HS6ST3	GLCE	EXT1	XYLT1	NDST1
Allograft rejection	34	IL5	HLA-DOA	HLA-DMA	HLA-DRB5	CD40LG
Fatty acid elongation	27	ELOVL2	ELOVL3	HACD4	THEM4	ELOVL4
Prolactin signaling pathway	70	MAPK12	HGF	TNFSF11	SHC2	ELF5

Continued on next page

Table A.1 – *Continued from previous page*

Pathway	**No. Gene**	**Genes with weights (Ranked High to Low)**				
Folate biosynthesis	30	ALPG	ALPP	GGH	TH	AR
Fatty acid degradation	44	LBP	ACSL6	CPT1C	ADH4	ACSBG2
Porphyrin and chlorophyll metabolism	43	UGT1A1	ALAS1	UGT1A8	COX10	PPOX
SNARE interactions in vesicular transport	33	STX1B	STX3	VAMP2	STX19	STX16
Estrogen signaling pathway	138	CALML6	HSPA2	KRT16	ADCY6	ADCY4
African trypanosomiasis	41	HP	LAMA3	IFNG	APOA1	IL12A
Cholesterol metabolism	56	APOH	ABCG8	VAPB	STAR	PLTP
Antifolate resistance	34	FOLR1	FOLR2	DHFR	GGH	TYMS
Collecting duct acid secretion	27	ATP6V0A4	CLCNKB	ATP6V0E2	ATP6V1B2	CA2
ABC transporters	45	ABCC9	ABCG8	ABCA12	ABCG2	ABCB11
Ether lipid metabolism	49	PLA2G2F	PLA2G3	PLA2G2D	JMJD7-PLA2G4B	PLA2G2A
Regulation of lipolysis in adipocytes	55	FABP4	ADORA1	PLIN1	AKT3	ADCY2
Protein export	23	IMMP2L	HSPA5	IMMP1L	SEC62	SPCS1
Other types of O glycan biosynthesis	47	POFUT1	COLGALT2	GXYLT2	GALNT14	GALNT4
Fat digestion and absorption	46	MTTP	CEL	ABCG5	FABP2	ABCG8
RNA polymerase	33	POLR2F	POLR3K	POLR2I	POLR2J	POLR3H
Aldosterone regulated sodium reabsorption	39	SGK1	SCNN1G	PDK1	FXYD2	ATP1B1
RIG I like receptor signaling pathway	72	PIN1	MAPK11	CXCL10	RIPK1	RNF125
Yersinia infection	140	CD8B2	ACTR3C	MAPK11	TRAF2	MAPK10
Relaxin signaling pathway	129	ADCY1	MMP1	COL3A1	GNAI2	CREB3L3

A.2 Significant COAD related pathways from literature

[136] analyzed cancer-related signaling pathways from the KEGG for specific deregulation in each cancer subtype signature. They pointed out 19 KEGG pathways that are up- or down- regulated for colorectal cancer using gene set enrichment analysis, among them 6 pathways are given high rank by DIGS. They are Renin-angiotensin

system (rank 12), Mismatch repair (rank 20), Hedgehog signalling pathway (rank 26), Apoptosis (rank 55), ECM-receptor interaction (rank 57), DNA replication (rank 86).

[176] uploaded the DEGenes to the Database for Annotation Visualization and Integrated Discovery (DAVID) online tool for KEGG pathway analysis. They found the Vitamin digestion and absorption is the most relevant pathway for colon cancer progression. This pathway is ranked 13 by DIGS. The forth and fifth important pathways, Fat digestion and absorption and Carbohydrate digestion and absorption are also ranked high in DIGS, which are 45 and 58 respectively.

[177] focus on the relationship between colorectal cancer and signalling pathways. They summarised 9 signalling pathways that are contributing to carcinogenesis. Among them, 2 signalling pathways are given high rank by DIGS, Hippo signalling pathway (rank 22) and Hedgehog signalling pathway (rank 26).

A.3 Pathway maps colored by gene weights

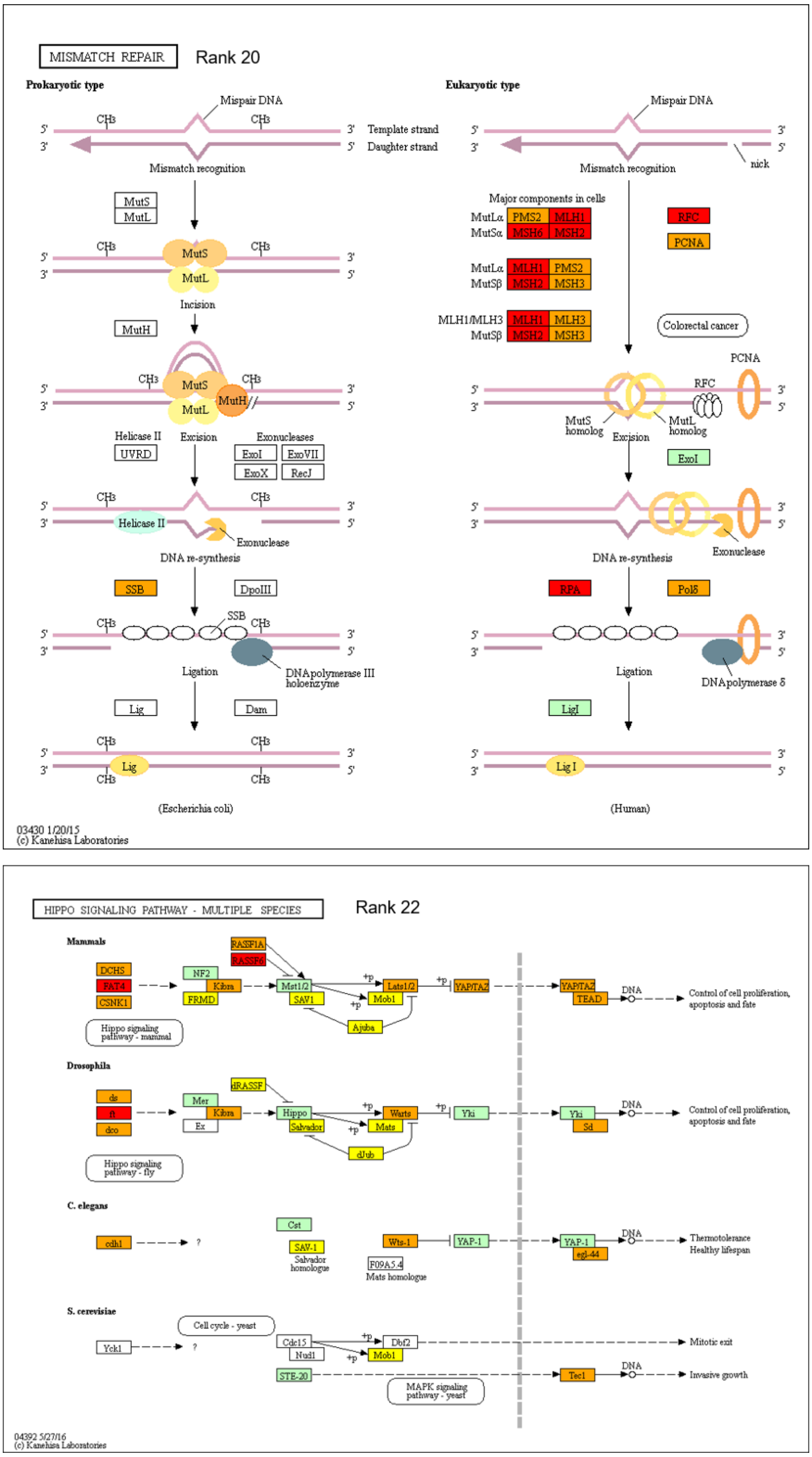
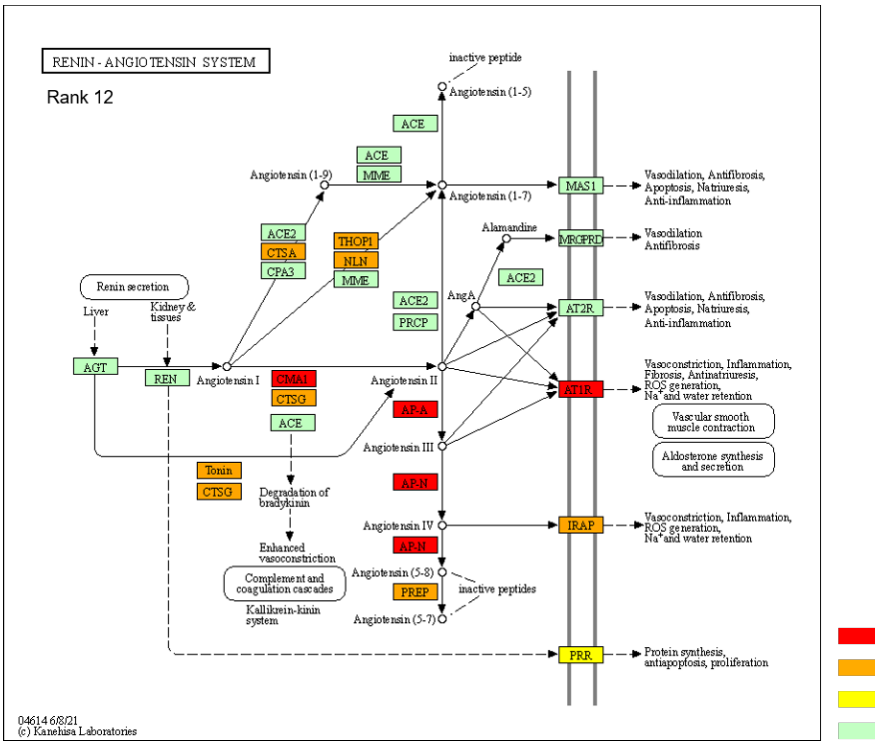


Fig. A.1 KEGG pathway maps with colors for important genes. Genes marked with red, orange and yellow are genes were assigned weights by DIGS, green genes do not have weights.


RENIN - ANGIOTENSIN SYSTEM
Rank 12
inactive peptide
Angiotensin (1-5)
ACE
ACE
MME
Angiotensin (1-9)
Angiotensin (1-7)
MAS1
Vasodilation, Antifibrosis, Apoptosis, Natriuresis, Anti-inflammation
ACE2
CTSA
CPA3
THOP1
NLN
MME
Alamandine
MRGPRD
Vasodilation Antifibrosis
ACE2
Renin secretion
Liver
Kidney & tissues
ACE2
PRCP
AngA
AT2R
Vasodilation, Antifibrosis, Apoptosis, Natriuresis, Anti-inflammation
AGT
REN
Angiotensin I
CMA1
CTSG
ACE
Angiotensin II
AP-A
AT1R
Vasoconstriction, Inflammation, Fibrosis, Antinatriuresis, ROS generation, Na+ and water retention
Vascular smooth muscle contraction
Aldosterone synthesis and secretion
Angiotensin III
Tonin
CTSG
Degradation of bradykinin
Enhanced vasoconstriction
AP-N
Angiotensin IV
IRAP
Vasoconstriction, Inflammation, ROS generation, Na+ and water retention
AP-N
Complement and coagulation cascades
Kallikrein-kinin system
Angiotensin (5-8)
PREP
inactive peptides
Angiotensin (5-7)
PRR
Protein synthesis, antiapoptosis, proliferation
04614 6/8/21
(c) Kanehisa Laboratories

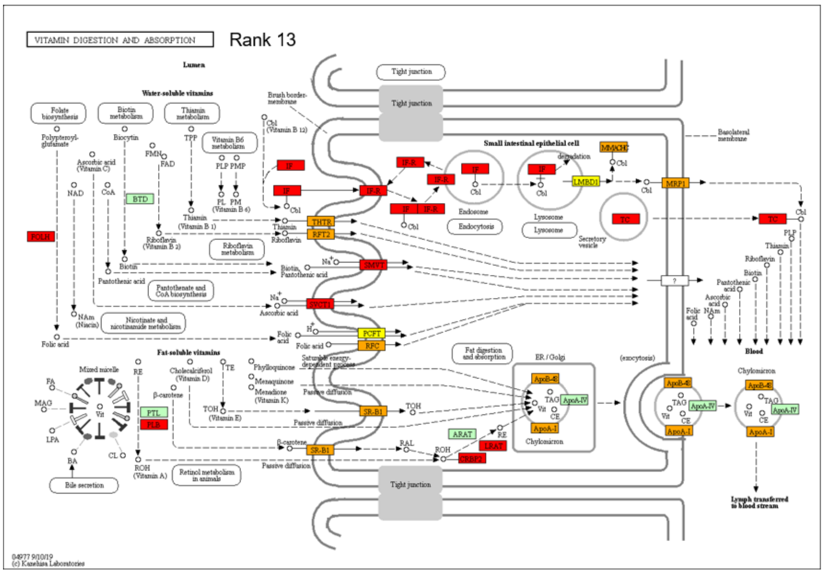


Fig. A.2 KEGG pathway maps with colors for important genes. Genes marked with red, orange and yellow are genes were assigned weights by DIGS, green genes do not have weights.

Appendix B

DIGS2 released COAD and BRCA relevant pathways

Table B.1 Top 50 KEGG Pathways of BRCA (DIGS2)

Pathway	**No. Gene**	**Genes with weights (Ranked High to Low)**				
Pancreatic secretion	104	CA2	CPA1	CHRM3	PRSS2	CPA2
Circadian rhythm	32	RORB	ROR1	PRKAA2	RORA	CUL1
Peroxisome	87	AGXT	HAO2	ACSL6	PEX11A	IDH2
Chemical carcinogenesis	81	GSTM5	PTGS2	GSTA1	GSTA2	CYP1A1
Platinum drug resistance	73	GSTM5	GSTA2	GSTA1	CDKN2A	GSTT2B
Drug metabolism cytochrome P450	70	GSTM5	GSTA1	GSTA2	UGT2B11	FMO2
Folate biosynthesis	30	TPH1	PAH	ALPL	MOCOS	FPGS
Drug metabolism other enzymes	79	GSTM5	GSTA1	GSTA2	XDH	GSTT2B
Cocaine addiction	50	SLC18A2	DRD1	GRIN2A	CREB3L3	SLC18A1
Carbon metabolism	118	AGXT	HAO2	ALDOB	PHGDH	PSAT1
Nicotine addiction	41	CHRNA6	GABRQ	GABRP	CHRNA7	GRIN2A
Complement and coagulation cascades	93	C4BPA	C4BPB	C1QC	C1QB	PLAT
RNA polymerase	33	POLR2F	POLR1F	POLR2D	POLR3G	POLR1B
Glutamatergic synapse	122	GRIK1	PLA2G4A	SLC1A6	ADCY3	PRKCG
Metabolism of xenobiotics by cytochrome P450	77	GSTM5	GSTA1	GSTA2	CYP1A1	UGT2B11
Pathways in cancer	539	GSTM5	GSTA2	NKX3-1	FGF10	F2

Continued on next page

Table B.1 – *Continued from previous page*

Pathway	No. Gene	Genes with weights (Ranked High to Low)				
Herpes simplex virus 1 infection	499	ITGB3	ZNF705E	ZFP57	POU2F3	IL6
Pathways of neurodegeneration multiple diseases	482	ATP2A1	DNAI2	COX7A1	GRIA3	TUBA3E
PPAR signaling pathway	74	UCP1	FABP5	ACSL6	HMGCS2	PLTP
Glycosylphosphatidylinositol (GPI) anchor biosynthesis	26	PIGA	PIGQ	PIGB	PIGT	PIGC
Spinocerebellar ataxia	150	FGF14	RBPJL	PSMD4	ATP2A1	NFYA
Synaptic vesicle cycle	83	SLC18A2	SLC6A2	UNC13C	ATP6V0D2	ATP6V1B2
HIF 1 signaling pathway	106	ALDOB	EGFR	PLCG2	ANGPT1	PRKCG
Amphetamine addiction	69	SLC18A2	GRIA3	DRD1	PRKCG	GRIA1
Steroid hormone biosynthesis	64	CYP1A1	AKR1D1	AKR1C2	UGT2B28	UGT2B7
Other glycan degradation	17	NEU4	FUCA2	MAN2C1	MAN2B2	HEXA
Terpenoid backbone biosynthesis	21	HMGCS2	FDPS	PDSS1	RCE1	IDI1
Primary bile acid biosynthesis	18	BAAT	CYP39A1	AKR1D1	CYP7B1	CYP7A1
Biosynthesis of unsaturated fatty acids	28	BAAT	ELOVL4	SCP2	ELOVL5	HACD2
Collecting duct acid secretion	27	CA2	ATP6V0D2	ATP6V1C2	SLC4A1	ATP6V1B2
Alzheimer disease	375	ATP2A1	WNT7A	IL1A	COX7A1	TUBA3E
Retinol metabolism	68	CYP1A1	RDH12	CYP2W1	CYP27C1	CYP2A7
Mismatch repair	23	MSH2	MSH6	RFC3	POLD1	EXO1
Thyroid hormone synthesis	77	TTN	F2	ATF2	ATP1B3	GPX2
Cholinergic synapse	116	CHRNA6	NRAS	CHRM3	FYN	KCNQ4
Ribosome	154	RPL3L	RPL31	RPL37	RPS3A	RPL14
Biosynthesis of amino acids	75	ALDOB	PHGDH	PSAT1	TKTL1	PAH
Proteasome	49	PSMB4	PSMD4	PSMB5	ADRM1	PSME2
Amyotrophic lateral sclerosis	369	TUBA3E	DNAI2	COX7A1	GRIA1	CASP12
Arginine and proline metabolism	47	CKM	SMO	ODC1	P4HA2	P4HA3

Continued on next page

Table B.1 – *Continued from previous page*

Pathway	**No. Gene**	**Genes with weights (Ranked High to Low)**				
Autoimmune thyroid disease	49	HLA-B	HLA-DRA	HLA-DOB	HLA-F	HLA-DRB1
Sphingolipid metabolism	49	CERS3	UGT8	GLA	ACER1	SPHK1
Lysosome	128	CTSK	ATP6V0D2	LAMP3	ACP5	AP3M1
2 Oxocarboxylic acid metabolism	19	BCAT1	NAGS	ACY1	BCAT2	GOT1
Allograft rejection	34	HLA-DRA	HLA-B	HLA-DRB1	HLA-DOB	HLA-F
Non homologous end joining	13	RAD50	PRKDC	XRCC5	FEN1	MRE11
Huntington disease	312	UCP1	COX7A1	TUBA3E	ITPR1	DNAH9
Oocyte meiosis	128	BUB1	CCNB2	YWHAG	PPP2R1B	ESPL1
Apoptosis multiple species	32	SEPTIN4	BIRC7	NGFR	BIRC5	CASP7

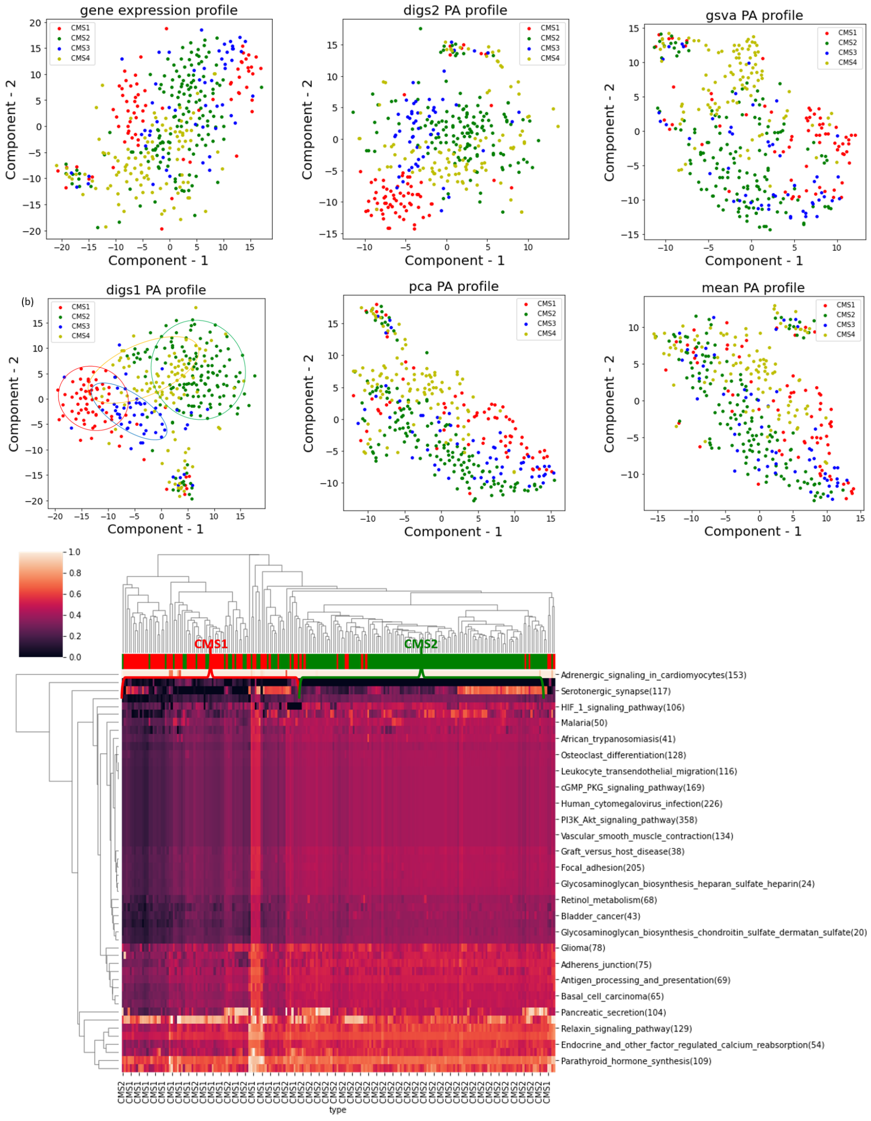


Fig. B.1 Pathway activity dimension reduction plots (TSNE) for six PA inference methods and Hierarchical clustering map of significant pathways (COAD).

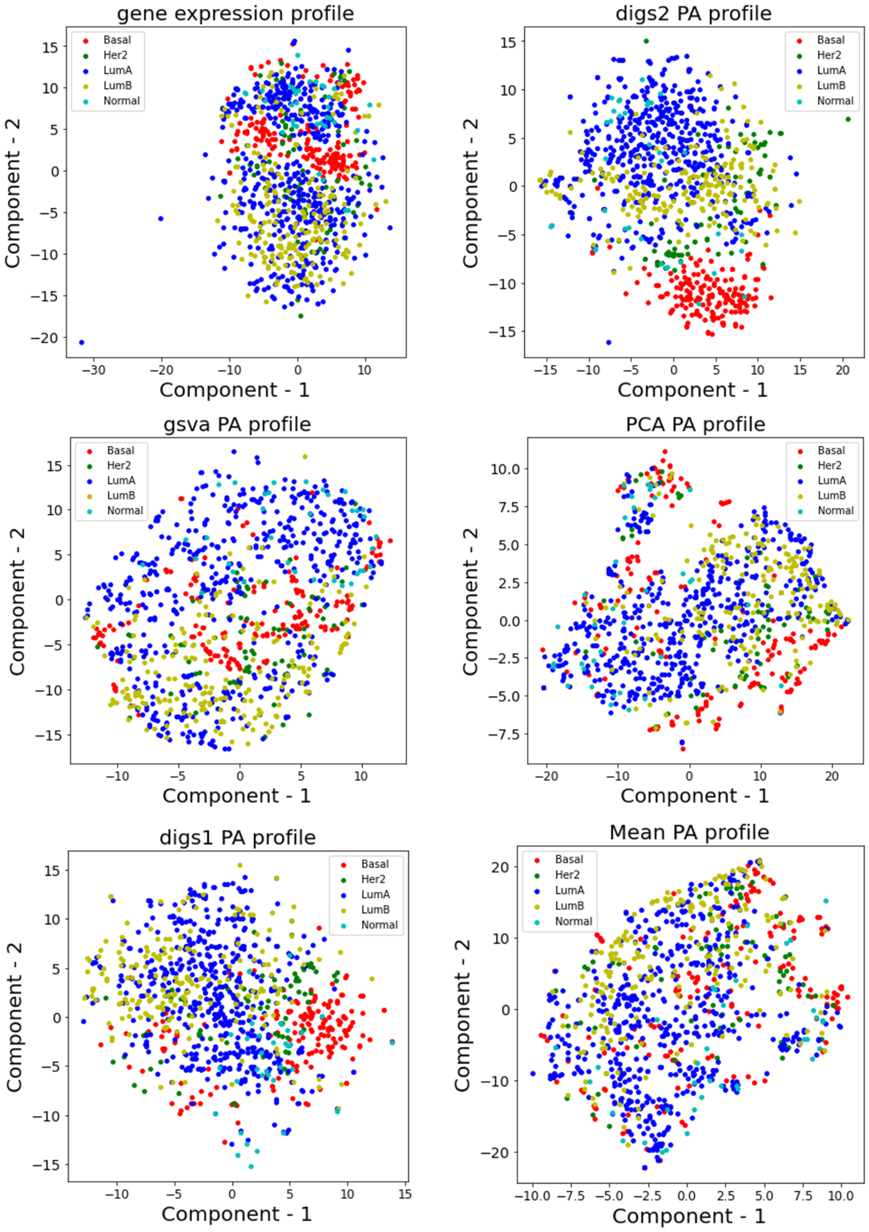


Fig. B.2 Pathway activity dimension reduction plots (TSNE) for six PA inference methods (BRCA).

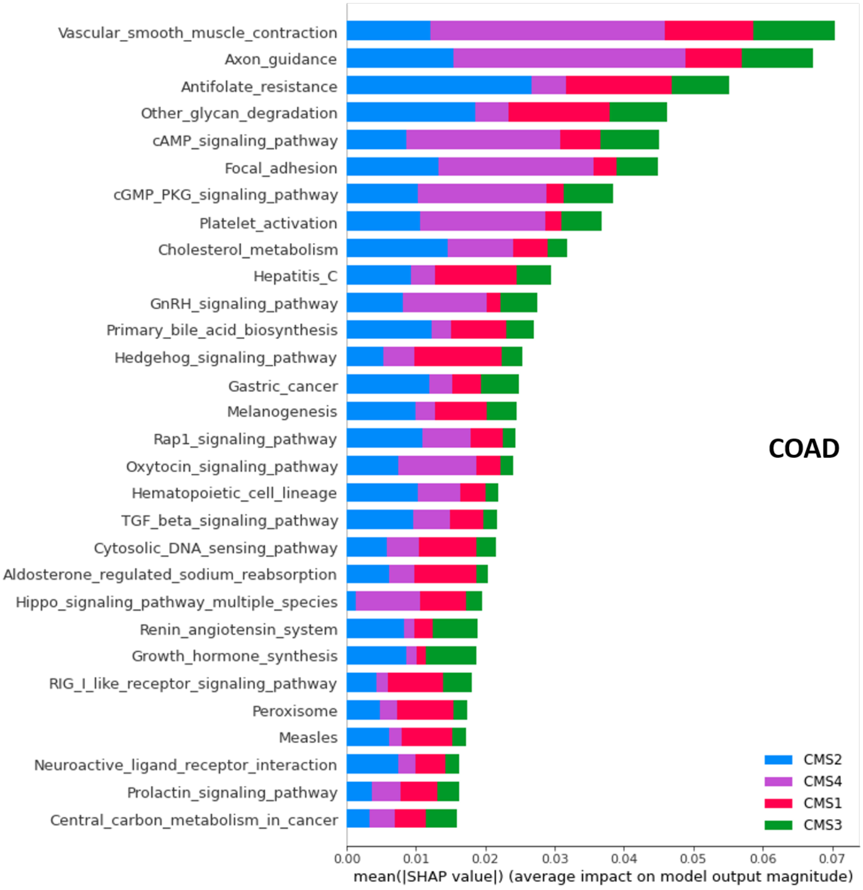


Fig. B.3 Top 30 pathways from SHAP for COAD

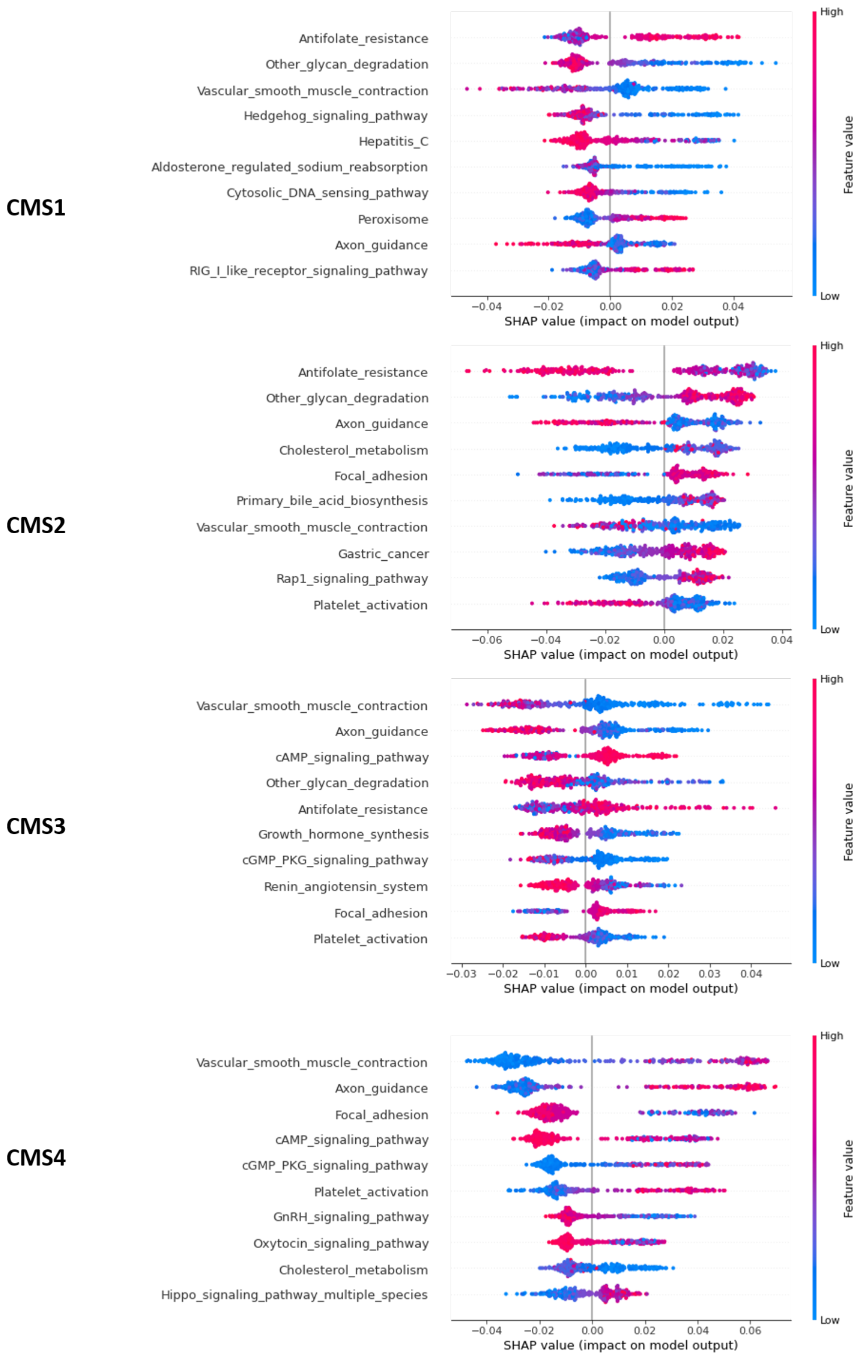


Fig. B.4 Top 10 pathways from SHAP for each subtype of COAD.

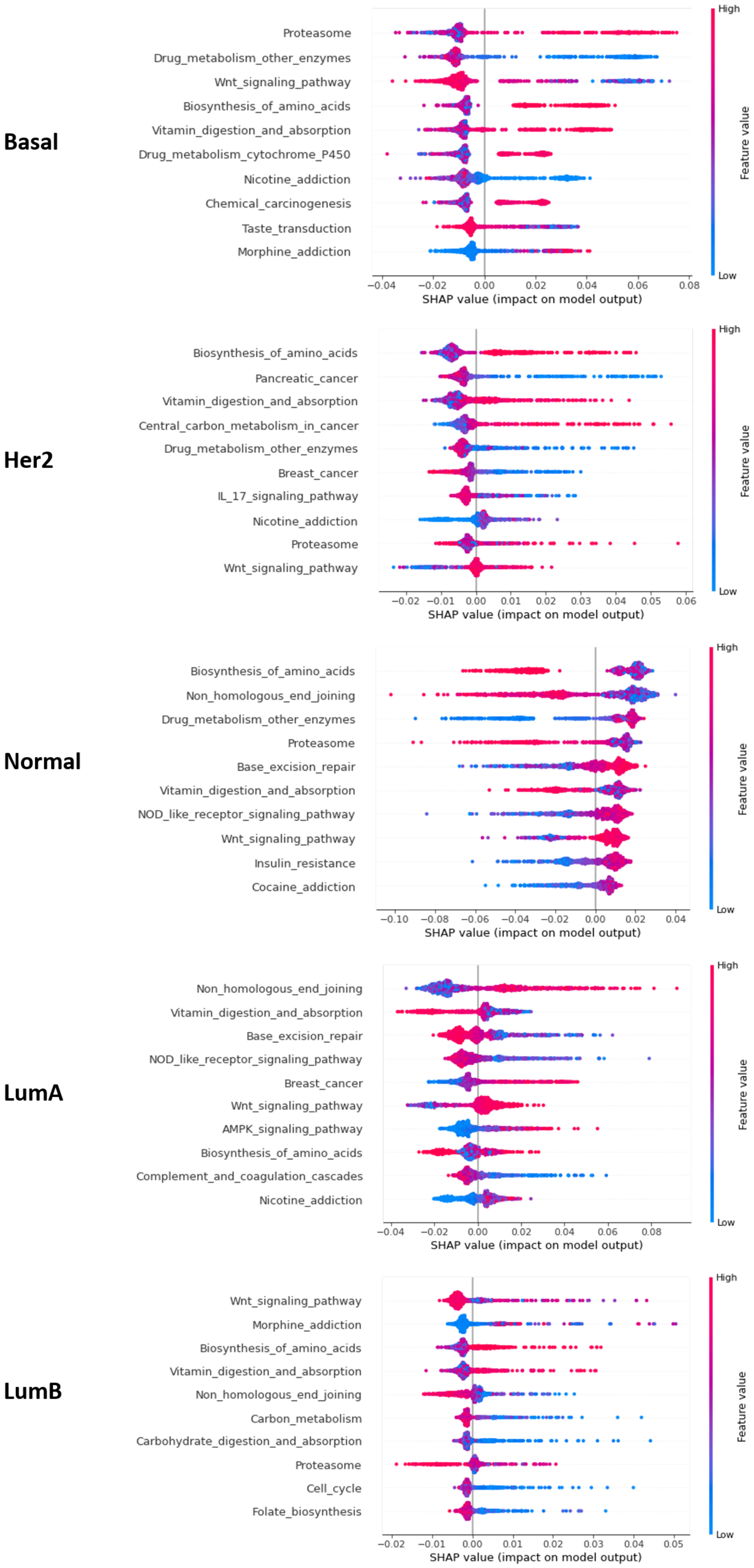


Fig. B.5 Top 10 pathways from SHAP for each subtype of BRCA.