A Systems Biology Approach for Identifying Markers of Chemotherapy Response

by

Kendric Wang

B.Cmp. (Honours), Queen's University, 2009

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate Studies

(Bioinformatics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

April 2012

Abstract

High-throughput gene expression data has been widely used to identify biomarkers for the classification of clinical outcome in cancer studies. In breast cancer, conventional methods have successfully identified molecular markers predictive of disease progression; however, predicting response to chemotherapy has proved more challenging and warrants the development of novel approaches. Recently developed systems biology methods that integrate transcriptomic and proteomic data have shown promising results in various classification problems; therefore, we investigated the use of this approach in predicting response to chemotherapy.

We developed a novel method, called OptDis, which integrates gene expression data with protein-protein interaction networks to efficiently identify subnetwork markers with optimal discrimination between different clinical outcome groups. Application of our method to a public dataset demonstrated three key advantages of using OptDis over previous methods for predicting drug response in breast cancer patients treated with combination chemotherapy. First, subnetwork markers derived from our method provides better classification performance compared with subnetwork and gene marker from existing methods. Second, OptDis subnetwork markers are more reproducible across independent cohorts compared to gene markers and may consequently be more robust against noise and variations in expression data. Third, OptDis subnetwork markers provide insights into mechanisms underlying tumour response to chemotherapy that are missed by conventional methods. Additional analyses using

OptDis showed that the use of prior knowledge from PPI interactions improves marker discovery and subsequent classification performance.

To our knowledge, this is the first study to demonstrate the advantages of applying an integrative network-based approach to the prediction of individual's response to cancer treatment. Markers identified using our method not only improve the classification of outcome, but it also provide novel understandings into the mechanism of drug action. With sufficient validation, this strategy may identify promising clinical markers that can facilitate the effective individualised treatment of cancer patients.

Preface

The development and application of OptDis was a joint effort between Phuong Dao, a PhD candidate in Dr. Sahinalp's lab, and me. The OptDis method was initially designed and implemented by Phuong Dao. My contributions to the development of OptDis comprised optimizing, testing, and further modifying the method for application on the given data. Additionally, I developed a pipeline in R ("bdvTools") that automates the discovery and large-scale validation of biomarkers derived from OptDis and competing subnetwork and gene marker methods (Section 2.3). The pipeline also generates visualizations for comparing marker performance from the different methods.

The OptDis method and its application to predicting chemotherapy response in breast cancer was published [1]: Dao P, Wang K, Collins C, Ester M, Lapuk A, Sahinalp SC: Optimally discriminative subnetwork markers predict response to chemotherapy. *Bioinformatics* 2011, **27**:i205-i213, doi:10.1093/bioinformatics/btr245. Phuong and Kendric contributed equally as first authors in the preparation of this paper. My specific contributions towards the study include:

- Designing the experiments
- Identifying and processing the drug response expression dataset
- Performing cross-dataset classification experiments for markers from different methods
- Performing and interpreting the biological analyses

Chapter 2 and 3, which describes the OptDis method and its application to a drug response dataset, expands on the material published in our paper [1].

Table of Contents

List of Tables

List of Figures

List of Abbreviations

BrCa: breast cancer

BXD: backward cross-dataset validation experiment

Dense: Subnetwork marker method implemented in [2]

FXD: forward cross-dataset validation experiment

GreedyMI: Subnetwork marker method implemented in [3]

MAQC: Microarray Quality Control project

MCC: Matthew's Correlation Coefficient used as the classification performance metric

OptDis: Our published subnetwork marker method described in [1]

pCR: pathological complete response to TFAC treatment

PPI: protein-protein interactions

Gene: classifiers build using gene marker identified by t-test

TFAC: four-drug neoadjuvant chemotherapy regimen provided to breast cancer patients

Acknowledgements

This thesis represents the fruits of my training as a scientist thus far. Undoubtedly, my journey would not have been possible, nor as enriching or pleasurable, without the support from many wonderful people.

First and foremost, I would like to thank my supervisors, Dr. Colin Collins, Dr. Cenk Sahinalp, and Dr. Anna Lapuk for their continual encouragement and guidance. I am grateful to Dr. Collins for providing me with an ideal environment to foster my interests in interdisciplinary and translational research. I am thankful to Dr. Sahinalp for helping me understand the significant role of computational methods in solving complex biological problems. Words alone cannot express my appreciation for the patience and dedication that Dr. Lapuk has shown towards my development as a researcher.

I would like to thank my colleague and close collaborator, Phuong Dao, for his generosity and mentorship. Without him, I truly could not have realised my aspirations for research.

I would like to thank the entire Collins Lab for their advice and for creating such an enjoyable environment to work in. I would like to thank the entire Cenk Lab for their fellowship and for sharing their research interests with me.

I would like to thank Dr. Artem Cherkasov for taking the time to be on my thesis committee, Dr. Martin Ester for taking an interest in my research, as well as Dr. Wyeth Wasserman and Dr. Steven Jones for allowing me the opportunity to expand my scientific horizons in their labs.

I would like to acknowledge the bright and fun students in the Bioinformatics program, many of whom I have gotten to know and befriend over the course of my studies.

Finally, I would like to thank the CIHR Bioinformatics Training Program for funding my research and Sharon Ruschkowski, the Bioinformatics program secretary, for helping me throughout my time in the program.

Dedications

“All our dreams can come true, if we have the courage to pursue them.”
~Walt Disney

I dedicate this thesis to my friends, whom I consider my extended family, for their support and their understanding through our shared trials in life.

I dedicate this thesis to my parents whom have always shown me unconditional love and support. To my father who always tried to steer me to the right path. To my mother who taught me that it is never too late to attain one’s goals.

Chapter 1
Introduction

1.1 Discovery of Cancer Biomarkers using Gene Expression

The advent of high-throughput gene expression assays has facilitated wide use of transcriptome profiling in clinical cancer research. In particular, there has been an increasing focus on identifying expression-based biomarkers predictive of clinical outcome such as cancer progression and response to chemotherapy. While numerous potential biomarkers have been developed using conventional statistical and bioinformatics methods, few of these markers have shown the robustness required for clinical application. Emerging approaches based on systems biology may address the limitations of conventional methods to produce promising markers urgently needed in the management of cancer. The different approaches for biomarker discovery and their respective types of biomarkers are illustrated in Figure 1.1.

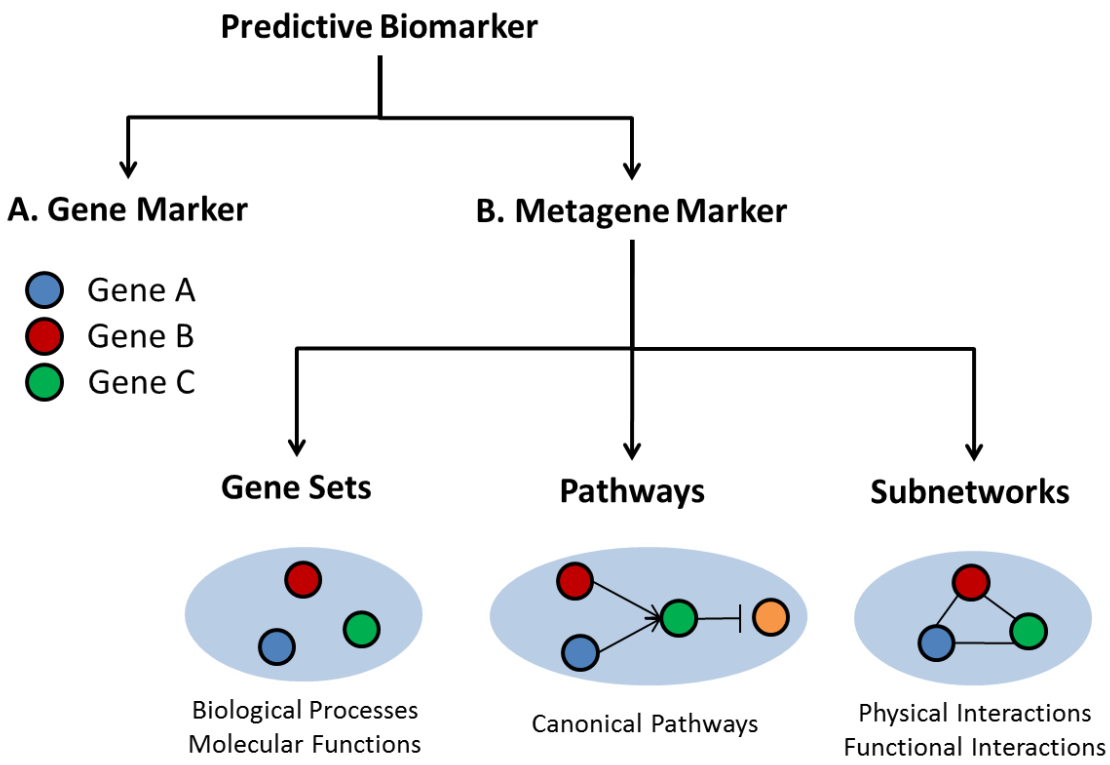


Figure 1.1: Approaches to biomarker discovery and their corresponding types of biomarkers.

1.1.1 Gene Marker Approach and Its Use in Predicting Clinical Outcome

The gene marker approach has been the conventional approach used for identifying gene expression markers predictive of cancer outcome. This approach involves ranking each gene based on its differential expression between different outcome groups and selecting the top ranked genes as predictors [4, 5]. Common ways to measure differential expression between two groups include t-test and mean-based fold-change. In breast cancer (BrCa), use of this type of approach has identified promising single genes implicated in sensitivity to primary chemotherapy such as MAPT [6], as well as multi-gene signatures predictive of metastatic progression [4, 5].

Despite successful findings, this approach of assessing gene importance on an individual basis suffers from multiple weaknesses. First, it may be limited to detecting genes that possess the strongest differential signal, whereas genes with subtle but coordinate changes may be missed [7]. Additionally, for cancer, which in recent years been viewed as a "disease of pathways", detecting rare changes that dysregulate a common pathway will be important for understanding both the etiology and treatment of the disease.

Second, the selection of gene markers in this manner is highly sensitive to noise and variations in gene expression data, resulting in generation of unstable markers which are frequently un-reproducible from additional datasets. Many recent studies have shown that small changes to the sample composition of training data can produce completely different gene signatures, and have suggested that thousands of samples may be needed to generate a stable marker signature [8, 9]. In fact, the well-validated 70-gene and 78-gene prognostic signatures for BrCa developed on different cohorts only had a 3-gene overlap [10]. Reasons suggested for this gene signature instability include small sample size, high correlative nature between genes, cellular heterogeneity within tissue samples, and genetic heterogeneity between patients. It is believed that a more stable (or reproducible) gene signature may provide more robust predictive performance [10], which is required for clinical applications.

Last but not least, gene markers may provide relatively limited insight into the biological mechanisms underlying cancer phenotypes. As a consequence of the first two weaknesses, the top markers from the ranked gene list may comprise a combination of

co-expressed genes related to disparate functions and uninformative false positives, while missing genes that are coordinated regulated such as part of protein complexes or signaling pathways. Finding coherent understanding from such a noisy set of genes through manual literature searches or even use of bioinformatics tools for function or pathway enrichment may be challenging.

1.1.2 Metagene Marker Approach

Although conventional gene marker methods assumes that genes act independently, it has become increasingly evident that genes carry out functions in a coordinated and modular manner [11, 12]. Motivated by these observations, there has been increasing focus on developing metagene marker methods, which aim to address the shortcomings of conventional methods by ranking the importance of a gene within the context of a group of functionally-related genes. In contrast to the gene marker approach, which identifies each marker as a single gene, the metagene marker approach identifies each marker as a group of genes aggregated into a single feature called a metagene. Prior knowledge of gene-gene relationships is used to guide the marker discovery process in this approach.

Earlier metagene methods used known or predicted gene sets such functional GO annotations as surrogates for metagene markers [13–15]. Subsequent studies extended these methods to curated biological pathways (i.e. signaling, metabolic) in hopes of describing specific mechanisms associated with the phenotype [16, 17]. However,

curated pathways model known interactions across many conditions, whereas only portions of the pathway may be active under a specific condition. Therefore, more recent methods attempted to infer pathway activity within a context based on a subset of genes in the pathway [18, 19].

These studies have shown that metagene markers can be as or more accurate than gene markers in predicting phenotype. This supports the view that modular markers are more robust to noise and variations in expression data because they are less sensitive to expression value changes in a single gene. Additionally, these metagene markers offered improved interpretability of the molecular mechanisms underlying clinical outcome at a modular level.

Despite the benefits of using metagene markers over gene markers, metagene marker methods based on *a priori* defined groups still suffer from a number of limitations. First, these methods are restricted to the fraction of genes that have been assigned into categories. Of those genes that have been annotated, there is only a partial understanding of their functions. Second, due to the pleiotropic nature of genes, many curated gene sets overlap and produce redundant markers. Third, curated gene classifications reflect possible functions across many experimental conditions, so methods based on these categories are agnostic to experimental context. Last and most important, use of these defined gene sets does not allow the *de novo* discovery of novel mechanisms associated with a phenotype.

Table 1.1: Characteristics of different types of metagene markers.

Marker	Prior Knowledge	Context-Sensitive	Novel Mechanism
Gene Set	✔		
Pathway	✔	✔	
Subnetwork	✔	✔	✔

1.1.3 Subnetwork Marker Approach

Recently, various groups have aimed to identify *de novo* metagene markers associated with phenotype by integrating gene expression data with prior knowledge from gene networks. In this type of approach, each metagene marker is called subnetwork marker (or "subnetwork"), and its activity is calculated as the aggregate expression of component genes in the subnetwork. The fundamental element of this approach is a search algorithm that identifies subnetworks most differentially active between groups.

Chuang *et al.* [3] published a seminal method by integrating gene expression data with protein-protein interactions (PPI) networks and using a heuristic greedy algorithm to identify subnetwork markers predictive of breast cancer progression. Chowdhury *et al.* [20] incorporated an improved search algorithm based on a branch and bound algorithm to predict colon cancer metastasis with high confidence. More recently, Su et al. [21] proposed a method to discover subnetworks by identifying paths containing genes that are both differentially expressed and co-expressed and greedily combining these paths.

In addition to PPI networks, functional association networks such as STRING [22] can also be used for development of subnetwork markers [2, 23]. Such networks integrate gene-gene relationships from multiple different sources including physical interactions from high-throughput experiments, co-occurrences from literature mining, and co-expression network constructed from microarray experiments. Methods based on functional networks extract dense subnetworks, which contain many more edges than expected and suggest participation in the same biological process or belong to the same protein complex.

These studies have demonstrated many advantages with using subnetwork markers. First, subnetwork markers improve classification performance over gene markers. Second, subnetwork markers derived on different cohorts have greater overlap in genes and therefore a greater degree of stability. As mentioned earlier, it is believed that a more robust gene signature may lead to more generalizable predictive performance. Last, subnetwork modules provide greater biological utility by offering *de novo* hypothesis about the mechanistic cause of phenotype.

Despite these promising results, existing methods for identifying subnetwork marker also have significant limitations. The network-based methods introduced by Chuang et al. [3], Fortney et al. [23], and Su *et al.* [21] are heuristic and thus do not guarantee the optimality of the solution for marker discovery. In other words, these algorithms are not guaranteed to find subnetworks with maximal discrimination power. An optimal solution would presumably provide a better predictive performance. The branch and bound [20] or exhaustive enumeration search algorithms [2] can yield an optimal

solution under some fixed set of parameters; however, their worst-case running time can be super-polynomial (and hence intractable). Therefore, there is a need to design efficient algorithms to retrieve the subnetwork markers that optimally distinguish samples from different classes.

1.2 Predicting Response to Chemotherapy in Breast Cancer

Currently, chemotherapies for treating cancer are selected for each cancer patient based on clinicopathologic features (such as tumour stage and size) with little consideration for the genetic heterogeneity between tumours that can affect response to a therapy [24]. Consequently, a significant fraction of the cancer patient population receives little to no benefit from ineffective treatment. For the treatment of breast cancer, several standard chemotherapy regimens are available, so there is a further question of which regimen should be provided to each patient [25].

The arrival of high-throughput gene expression technologies has motivated numerous clinical studies aimed at discovering molecular biomarkers capable of predicting patient response to chemotherapy prior to treatment [5, 26–28]. In breast cancer, many groups focused on predicting tumour response to neoadjuvant (preoperative) chemotherapy because response can be directly monitored following treatment. Across these studies, the discovery of predictive gene expression markers has been dominated by the use of a conventional gene marker approach (described in Section 1.1.1), where a list of top differentially expressed genes is used to classify

outcome. Although some of these predictive gene signatures have shown promising results in a limited number of patients, they have failed to achieve similar performance in additional validation studies [25, 29]. Thus far, none of these predictive signatures have demonstrated sufficient discriminative accuracy for clinical use [25].

Two recent studies have shown the promising use of metagene markers in the classification of chemotherapy response [30, 31]. In these studies, metagene markers were constructed from sets of co-expressed genes with related biological processes such as mitotic assembly, ceramide metabolism, and stromal biology. Given the success of these metagene markers, we believe that use of knowledge from PPI networks to guide marker discovery would provide further advantages such as improvements in classification accuracy and *de novo* hypotheses regarding the mechanisms of differential drug response.

1.3 Thesis Overview

Conventional gene markers methods have yet to yield robust markers predictive of chemotherapeutic outcome. Given the promising benefits of metagene subnetwork markers in recent studies, the aim of this thesis was to investigate the use this type of approach for predicting chemotherapy response. We examined this problem in three phases. In the first phase, we developed a novel method to identify subnetwork markers for predicting response, which addresses the limitations of previous methods. The details of our method, OptDis, are provided and discussed in Chapter 2. In the

second phase, we evaluated the benefits of OptDis subnetwork markers against gene markers derived from conventional means and other subnetwork markers derived from existing methods. In particular, we assessed the markers based on predictive performance (Section 3.3), reproducibility across cohorts (Section 3.4), and biological insight (Section 3.5). In the third and final phase, we investigated different factors that have contributed to performance improvements for the OptDis method (Section 3.6). We conclude the study by discussing the limitations and future applications of OptDis (Chapter 4).

Chapter 2
Methods

2.1 Overview

This chapter describes OptDis, a novel method that we developed to identify optimally discriminative subnetwork markers. In Section 2.2.1, we present the three main steps in our OptDis strategy. In Sections 2.2.2-2.2.4, we describe the details and mathematical formulations of the novel components in the OptDis method such as the search algorithm. In Section 2.3, we outline a pipeline that was developed to facilitate biomarker discovery and large-scale validation using OptDis and other methods.

2.2 OptDis Method

2.2.1 General Strategy to Identify Subnetwork Markers

The general strategy to identify subnetwork markers comprises three major steps: (1) data integration, (2) search for optimal subnetworks, and (3) marker selection. These steps are illustrated in Figure 2.1.

In the first step, the gene expression profile and PPI data is integrated by overlaying each gene in the expression profile onto its corresponding protein in the PPI network. In this way, a relationship, or an edge, is assigned between each pair of genes if a known physical connection exists between their corresponding proteins. Only genes with

corresponding proteins in the network are used to identify subnetwork markers in the subsequent steps.

In the second step, a search algorithm is employed to identify subnetworks with activities that best correlate with the phenotype (such as response to treatment). A subnetwork is set of connected genes extracted from entire PPI network, and the activity of a subnetwork in any sample is calculated as the aggregate expression level of constituent genes in that subnetwork in that sample. This aggregation essentially collapses many gene features into one subnetwork feature that captures the discriminatory potential of multiple gene markers in a single metagene marker. For example, if gene A discriminates drug-response in one set of patients and gene B discriminates drug-response in a second set of patients, then the aggregate activity of these two genes can potentially discriminate response in both sets of patients. We chose to aggregate expression values by taking the mean expression, which has been commonly used by existing methods. Other ways to aggregate gene expressions have been explored by Su [19].

To identify candidate subnetwork markers, the search algorithm scans the PPI network for subnetworks with maximal discrimination scores, where the discrimination score is a measure of the association between subnetwork activity and the phenotype. If a simple greedy search algorithm is used, it would search in the following manner.

To find the optimal subnetwork that includes a specific gene (“seed gene”), the algorithm starts by including that gene in the subnetwork. Then, it iteratively adds neighbouring genes from the PPI network into the subnetwork if they improve the

subnetwork's discrimination score. If no additional gene can be added to improve the discrimination score, then that subnetwork is considered the optimally discriminative subnetwork containing the seed gene and is subsequently added to the list of candidate subnetwork markers. The search algorithm is applied repeatedly, using each node in the PPI network as the seed gene, in order to return the list of all candidate optimal subnetwork markers.

For the OptDis method, we implemented a novel search algorithm that efficiently returns subnetworks with globally maximal discrimination scores, which improves on the locally maximal solutions returned by heuristic algorithms. Furthermore, our method also uses a superior distance-based function to calculate the discrimination score, compared to statistical scoring functions utilised by existing methods. The motivation and mathematical formulations for these two novel contributions are described in-depth in the subsequent two sections (2.2.2-2.2.3).

In the third and final step, all the candidate subnetwork markers returned by the search algorithm are ranked based on their discrimination scores and the top *x* subnetworks are selected as predictors of outcome. The activity levels of the selected subnetwork markers are used to train a classifier for predicting on new samples.

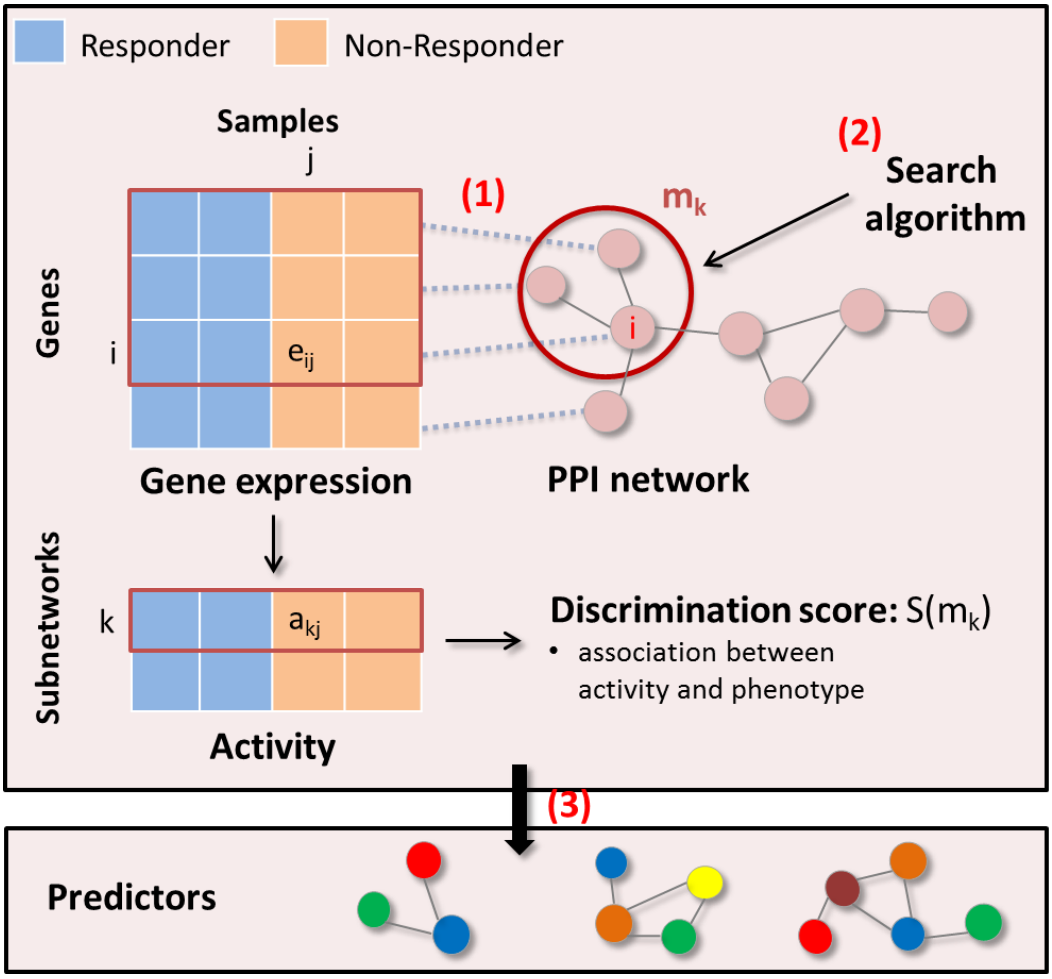


Figure 2.1: Three steps in the general strategy for identifying subnetwork marker. In step (1), gene expression and PPI network data is integrated. In step (2), a subnetwork search algorithm is used to identify candidate subnetworks at each possible seed gene i. In step (3), top candidate subnetworks are ranked by discrimination score and selected as markers of response.

2.2.2 A Distance-based Function for Calculating Discrimination Score

Discrimination score quantifies the ability of a marker to discriminate samples of one class from samples of the other class (i.e. sensitive vs. resistant to therapy). This score is calculated using a defined mathematical function based on either expression levels for gene markers or activity levels for subnetwork markers. For the OptDis

method, we developed a distance-based function that calculates the markers' discrimination score as the difference between its interclass distance and intra-class distance. Interclass distance is defined as the average L_1 (Manhattan) distance between each sample from one class and all samples from the other class. Intra-class distance is defined as the average L_1 distance between all samples of the same class. Intuitively, the activity levels of markers with high discrimination score should maximize the separation between samples from different classes and minimize the distances between samples from the same class. This property is illustrated in Figure 2.2.

We chose to calculate discrimination score using a distance-based function rather than a statistical function such as t-test or information gain because it does not make prior assumptions about the data. T-test assumes that the markers' activity level for samples within the each class are normally-distributed; however, this assumption may be challenged if there are too few samples in a class or if the samples originate from multiple populations (i.e. cancer subtypes). The drawback of information gain is that it requires discretizing the activity level in order to calculate a probability. Based on this reason, we expected a distance-based function to calculate more accurate discrimination scores compared to statistical-based functions.

The search algorithm in the following section uses the distance-based function to look for subnetworks with optimal distance score by simultaneously maximizing the interclass distance and minimizing intra-class distance of samples in the dataset.

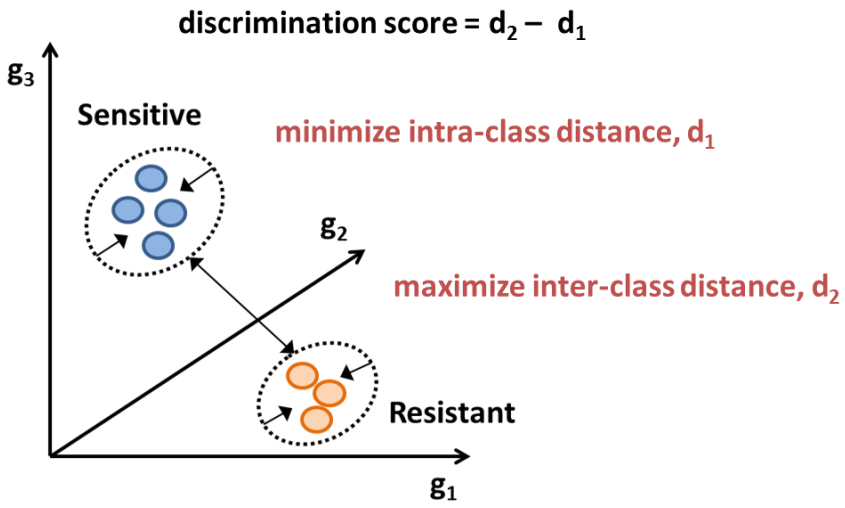


Figure 2.2: Illustration of distance-based discrimination score. Subnetworks with high discrimination scores should maximize inter-class distance and minimize intra-class distance.

2.2.3 An Efficient Randomized Search Algorithm to Identify Optimally Discriminative Subnetworks

First, we formalize the task of searching for subnetworks with optimal discrimination scores as the Optimally Discriminating k-Subnetwork (ODkS) problem and assess the complexity of this problem. Then, we provide a randomized algorithm to solve the problem for any given error probability.

2.2.3.1 Problem Definition

Without loss of generality, we formulate the problem for two classes, but note that it is easy to extend our approach for more than two classes. Let A and A' denote the expression matrices for positive and negative samples, respectively. For each gene g_i,

let A_i and A'_i, respectively, denote the expression profiles of gene g_i in positive class and negative class. For expression matrix A and A', let $A_i(j)$ and $A'_i(j)$ denote the expression of g_i in sample j. Given n genes, let a and a' denote the number of samples in positive class and negative class, respectively. We denote the PPI network by $G=(V,E)$, where $|V|=n$ and $|E|=m$.

In Equation 1, we define the discrimination score function for gene g_i as the difference between the interclass distance (average distance between samples from different classes) and the intra-class distances (average distance between samples from the same class), under L_1 distance. A coefficient c is introduced to weigh the relative contributions of the intra-class distance for the positive and negative class. By default, c is set to 0.5, representing an equal contribution from both classes.

$$w(g_i) = \sum_{j=1}^{a}\sum_{j'=1}^{a'}\frac{|A_i(j) - A'_i(j')|}{aa'} - \left[c\sum_{j=1}^{a}\sum_{j'=1}^{a}\frac{|A_i(j) - A_i(j')|}{aa} + (1-c)\sum_{j=1}^{a'}\sum_{j'=1}^{a'}\frac{|A'_i(j) - A'_i(j')|}{a'a'} \right] \quad \textbf{(1)}$$

In Equation (2), we extend the discrimination score function to subnetwork S by summing the scores from the component genes:

$$w(S) = \sum_{\forall i: g_i \in S} \left(\sum_{j=1}^{a} \sum_{j'=1}^{a'} \frac{|A_i(j) - A_i'(j')|}{aa'} - \left[c \sum_{j=1}^{a} \sum_{j'=1}^{a} \frac{|A_i(j) - A_i(j')|}{aa} + (1-c) \sum_{j=1}^{a'} \sum_{j'=1}^{a'} \frac{|A_i'(j) - A_i'(j')|}{a'a'} \right] \right) \quad \textbf{(2)}$$

For brevity, this can be re-written in terms of the gene score function:

$$w(S) = \sum_{\forall i: g_i \in S} w(g_i) \quad \textbf{(3)}$$

Now, the ODkS problem can be defined as to finding the connected subnetwork from network *G* containing at most *k* genes, S_{OPT} ($|S_{OPT}| \leq k$), such that S_{OPT} distinguishes samples from different classes 'optimally'. From here on end, we denote S_{OPT} as the optimally discriminative subnetwork, where $w(S_{OPT})$ is the maximum among the $w(S)$ for any connected subnetwork *S*. From Equation (3), we can see that identifying S_{OPT} is equivalent to finding the connected subnetwork for which the total score of the vertices (genes) is maximized.

2.2.3.2 Problem Complexity

A variant of the ODkS problem called the "Connected k-Subgraph problem" has been proved to be NP-hard (Hochbaum and Pathria, unpublished). In that problem, the scores of vertices are restricted to either 0 or 1. By reduction from the Connected *k*-Subgraph problem, we proved that the ODkS problem is also NP-hard, even when there

is one sample in each class. The details of this proof are provided in our published paper [1].

2.2.3.2 Randomized algorithm

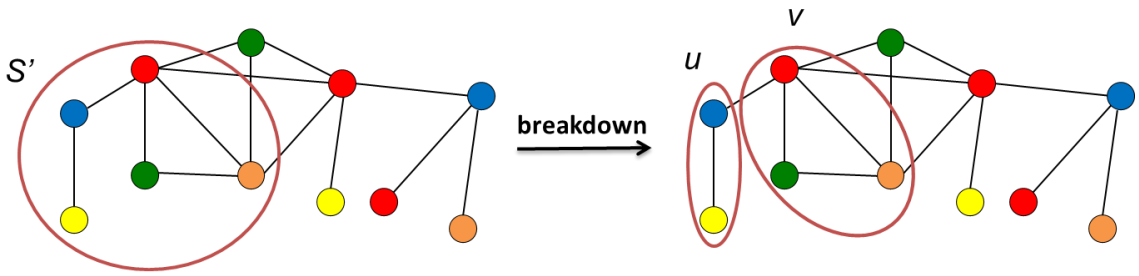
We provide a randomized algorithm to solve the ODkS problem for any given error probability by combining the color-coding technique [32] with dynamic programming. Color coding is an algorithmic technique that was first introduced by Alon *et al.* [32] to detect a simple path or a cycle of length *k* in a given graph. The algorithm consists of a predefined number of iterations. In each iteration, there are two main steps: (1) assign each vertex uniformly at random with one of *k* colors and (2) detect whether there is a 'colorful' path or cycle of length *k* in the given graph. A path or cycle is colorful if no two vertices in the path or cycle have the same color.

The idea behind the algorithm is the clever use of colors to reduce the number of paths that need to consider in the detecting step. In the naive algorithm, it is necessary to keep track of every vertices visited so far, which uses $O(n^k)$ time and space. However, in the color-coding algorithm, it is only necessary to store all possible sets of vertices of distinct colors, which uses $O(n2^k)$ time and space. Color-coding has been successfully applied to many applications including retrieving network motifs and comparing PPI networks of different species [33, 34].

Similar to color-coding technique, our algorithm consists of a predefined number of *i* iterations, where each iteration comprises two main steps (Figure 2.3):

1. Assign each vertex in the network with one of k colors, randomly and uniformly.
2. Identify the colorful connected subnetwork S'_{OPT} ($|S'_{OPT}| \leq k$) with the maximum discriminative score $w(S'_{OPT})$. S_{OPT} is the optimally discriminative connected subnetwork, whereas S'_{OPT} is the colorful optimally discriminative subnetwork in each iteration.

In the second step, an efficient dynamic programming approach is used to retrieve the S'_{OPT}. Dynamic programming solves a complex problem by breaking it into simpler subproblems that can be solved and combined to obtain the overall solution. Since many of the subproblems are the same, dynamic programming only solves each subproblem once, reducing the total number of computations needed. In our application, the complex problem of finding S'_{OPT}, the colorful path with the maximal discrimination score, can be broken down into a simpler problem of finding two smaller connected and non-overlapping colorful paths whose combined discrimination score is maximal. In fact, the general case of finding any colorful path can be broken down into subproblems in this way. In the base case, the simplest colorful path includes only a single vertex. Therefore, the problem of finding S'_{OPT} involves repeatedly combining colorful paths, starting from the base case. The mathematical details for this dynamic programming approach is provided in our published paper [1].



$$score(S') = score(u) + score(v)$$

Figure 2.3: OptDis subnetwork search algorithm. In each iteration of the algorithm, all nodes in the network is colored in order to identify the maximally discriminative colorful subgraph *S'*. Based on the property that finding *S'* is the same as finding the two connected and non-overlapping colorful subgraphs *u* and *v*, use of dynamic programming allows *S'* to be efficiently found.

After repeating the above two-step process for *i* iterations, the algorithm returns the colorful path S'_{OPT} with maximal score across all the iterations. Returning this S'_{OPT} is equivalent to finding S_{OPT}. Given a defined number of iterations of this algorithm, S_{OPT} can be retrieved with a success probability of $1 - \delta$, where δ is the given error probability. The mathematical calculations showing how determining how many iterations need to be run for any given error probability δ is provided in our published paper [1].

Our randomised algorithm using color-coding technique and dynamic programming takes polynomial time to return optimally discriminative subnetworks of size *k* with a fixed probability of error, so long as $k=O(\log n)$ and where *n* is the number of nodes in the network. This means that our algorithm can find optimal solutions for smaller subnetworks in efficient run-time on the relatively sparse PPI networks. The running

time complexity of this randomized algorithm is calculated and explained in our published paper [1].

2.3.4 Rank Subnetworks and Select Markers

The algorithm described in previous section will return the optimally discriminative subnetworks of each size between a min k_0 size and max k size for each vertex in the network. This will produce at most $k*n$ subnetworks, which make up the list of candidate subnetwork markers. To rank subnetworks, each subnetwork is normalized by its size. This normalisation accounts for the bias of subnetwork comprising more constituent genes inherently having greater discrimination scores.

For each subnetwork S, its metagene activity is first calculated as the aggregate expression profiles of genes in S:

$$A_s(j) = \frac{1}{|S|}\sum_{g_i \in S} A_i(j) \; when \; 1 \le j \le a$$
$$A'_s(j') = \frac{1}{|S|}\sum_{g_i \in S} A'_i(j') \; when \; 1 \le j' \le a' \qquad \textbf{(4)}$$

Then, its normalized discriminative score is calculated similarly to the discriminative score for a gene g_i in Equation (1).

Once candidate subnetworks are ranked by their normalized discriminative score, the top x subnetworks are selected as the markers of response. Starting from the subnetworks with greatest normalized discriminatory score, and going down to those with the smallest score, a subnetwork is added to the list of predictors if at least some

fraction of its constituent genes, *f*, is new compared to the genes in all the subnetworks added to the list so far. This criterion limits the maximum overlap between the selected subnetwork markers. By default, we set the fraction overlap, *f* to 0.5. Subnetworks are iteratively added until there are *x* predictors.

2.3 Pipeline for Biomarker Discovery and Validation

We developed 'bdvtools', an in-house software pipeline in R, to support our demand for automated, large-scale biomarker discovery and validation experiments. The design of this pipeline is shown in Figure 2.4. The pipeline includes six functional modules: (1) data pre-processing, (2) feature selection, (3) prediction model building, (4) marker validation, (5) performance computation, and (6) performance visualisation. In each functional module, while many options are readily implemented and available for use, additional options can also be incorporated. For example, probe-gene mapping, in data pre-processing step, can be performed for any Affymetrix gene expression array, but it can easily be extended to non-Affymetrix arrays if the platform annotation file is provided.

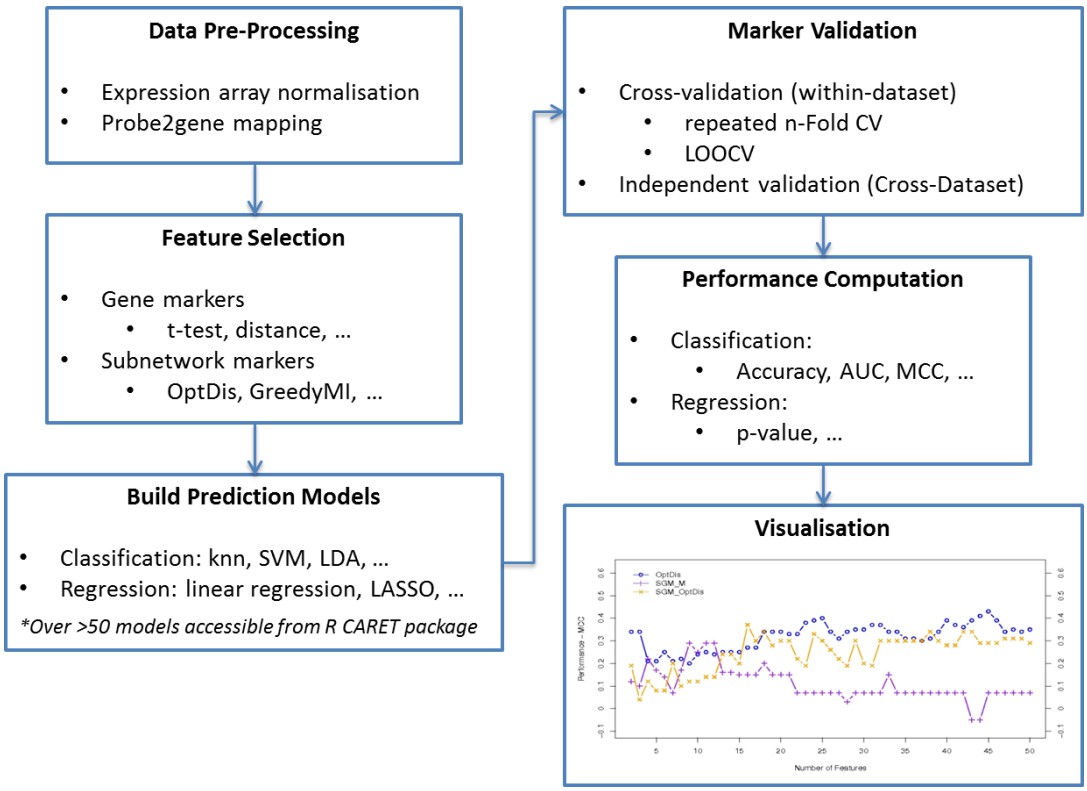


Figure 2.4: Design of the bdvTools pipeline. Each box represents a functional module.

Chapter 3
Results & Discussion

3.1 Overview

This chapter describes the application of the OptDis method on a published gene expression dataset and its evaluation compared to traditional gene marker methods and existing subnetwork marker methods. In Section 3.3, we compared the classification performance of markers derived from different methods and determined that OptDis subnetwork markers offer the best generalizable performance. In Section 3.4, we examined the reproducibility of genes derived from OptDis subnetwork markers against gene markers and found that the OptDis markers produced on different cohorts comprise more recurrent genes. In Section 3.5, we investigated the biological insights offered by the various markers, and found that OptDis subnetwork markers highlighted mechanisms of resistance to chemotherapy missed by gene markers. In Section 3.6, we investigated various factors that contributed to the improvement in performance from using OptDis markers.

3.2 Datasets

We evaluated the OptDis method on a human breast cancer dataset contributed by the University of Texas M.D. Anderson Cancer Center (MDACC, Houston, TX, USA) and

published in the MAQC-II study [35]. The gene expression data was retrieved from NCBI Gene Expression Omnibus (GEO) with accession number GSE20194. Gene expression profiles of 230 Stage I–III breast cancers were generated from fine-needle aspiration specimens of newly diagnosed breast cancers before patients received 6 months of neoadjuvant chemotherapy comprising paclitaxel, 5-fluorouracil, doxorubicin and cyclophosphamide (TFAC) followed by surgical resection of the cancer. Following treatment, patients were categorized into a positive response group if they exhibited pathological complete response, which is described as having no residual invasive cancer in the breast or lymph nodes. Otherwise, they were categorized into a negative response group. RNA extraction and gene expression profiling were performed in multiple batches using Affymetrix U133A microarrays. This dataset was split into two different cohorts according to the time of collection. One cohort consists of 130 samples while the other one consists of 100 samples. The expression profiles were normalized with Robust-chip Median Average (RMA) algorithm [36] and adjusted for batch effect using ComBat [37]. Probes were summarized into their corresponding genes by selecting the probe most differentially expressed between response groups based on t-test. Prior to model generation, the expression value for each gene was mean-centered between the two cohorts.

We retrieved the human PPI data from the Human Protein Reference Database (HPRD) version April 2010 [38]. By including binary interactions and considering each protein complex as a clique of proteins, we obtained 46,370 protein interactions involving 9,617 proteins.

3.3 Classification Performance of Markers

3.3.1 Workflow for Assessing Performance

We evaluated the classification performance of markers in two cross-dataset experiments using the workflow shown in Figure 3.1. For the forward cross-dataset (FXD) analysis, the 130 patient cohort was treated as the training set used for deriving markers of response and building classifiers, and the 100 patient cohort was treated as the independent validation set used for assessing classifier performance. For the complementary backward cross-dataset (BXD) analysis, the cohorts used for training and validation were swapped.

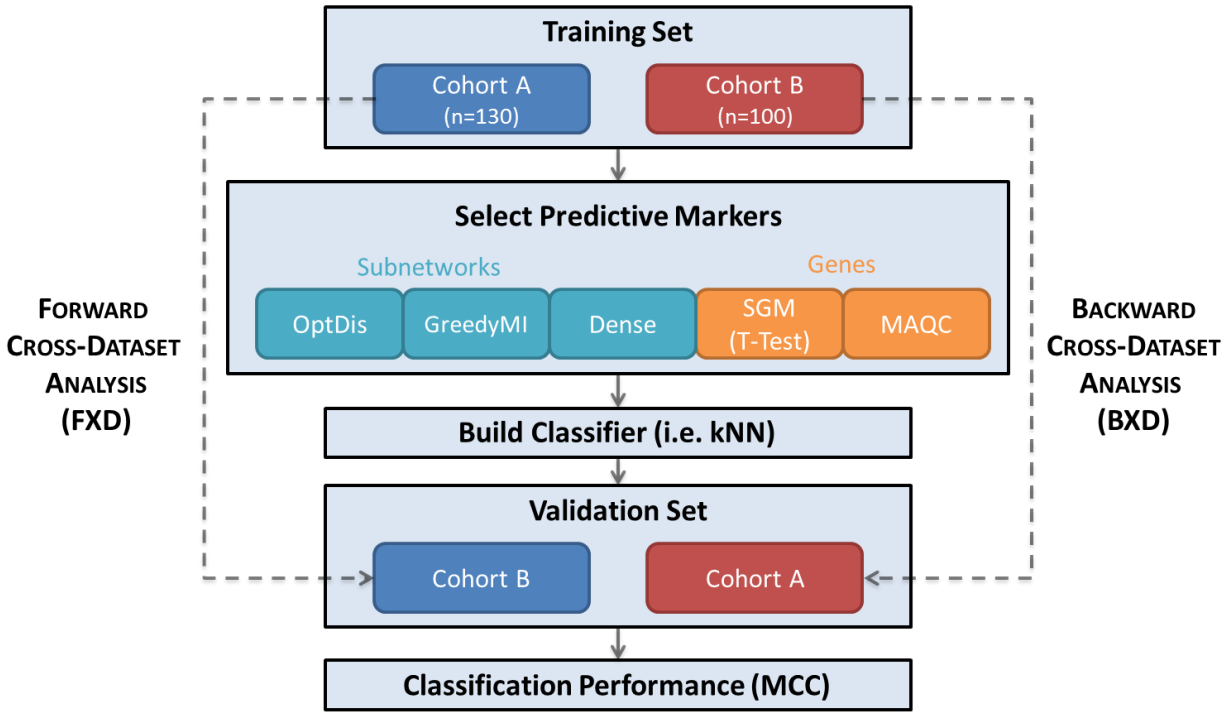


Figure 3.1: Workflow for assessing marker performance in cross-dataset experiments.

3.3.2 Classifier Details

We built classifiers using a k-nearest neighbour (kNN) classification model (k=3, using L_1 distance) and trained it with the top ranked gene or subnetwork markers derived from different methods (Figure 3.1). The top gene markers were derived from a two-sided t-test (assuming unpaired samples and unequal variance) and selected based on t-statistic. For this method (denoted as Gene), only genes with corresponding proteins in the PPI network were considered. The top OptDis subnetwork markers were selected based on the ranking criteria discussed in Section 2.2.3.2. We ran OptDis with error probability δ=0.001 and subnetwork size k_0=4 and k=7 for all experiments. We also compared against subnetwork markers derived from other published methods: (1) GreedyMI, which uses a heuristical search based on mutual information [3] and (2) Dense, which extracts dense subnetworks from the STRING functional network [2]. The density threshold to extract all dense subnetworks is set at 0.7, as implemented in [2]. The top subnetwork markers derived from GreedyMI and Dense were selected based on their mutual information scores.

3.3.3 Performance Metric

Since there is an imbalanced ratio between the number of samples in positive and negative class in this dataset, accuracy was not considered to be an appropriate measure for assessing classification performance. Instead, we utilized Matthews Coefficient Correlation (MCC) to compare the performance between different predictive

models [39]. MCC can be interpreted as the Pearson correlation between the predicted and known class labels for binary classified samples. It is calculated as follows:

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. If any of the four sums in the denominator is zero, then the denominator is set to one and the MCC becomes zero. Intuitively, MCC can be interpreted as follows: 1 is a perfect prediction, −1 is an inverse prediction, and 0 is a completely random prediction. We chose to use MCC over area under ROC curve (AUC) to facilitate comparison to classifiers reported in the MAQC-II study [35].

3.3.4 Classification Results

Using cross-dataset validation experiments, we evaluated the performance of classifiers constructed from subnetwork markers identified by OptDis. In each experiment, classifier performance was calculated for the top 1 to top 50 ranked markers. For comparison, we derived markers of response using competing methods and evaluated their classifier performance across the same range of top markers. The classification performances for these different markers in the FXD and BXD experiments are shown in Figure 3.2. The *x*-axis indicates the number of top markers used in the classifier, where each marker is either a gene marker or a subnetwork marker, and the

y-axis indicates the classifier performance in terms of MCC. In the FXD analysis, OptDis tends to achieve higher MCC than competing methods, starting at 20 markers. In the BXD analysis, OptDis performs better starting at 10 markers.

The average classification performance across the range of top 50 markers for each method is summarized in Figure 3.3. We can see that classifiers based on OptDis markers show statistically higher MCC than other markers in the FXD and BXD analysis, as well as overall in the cross-dataset validation experiments. Subnetwork markers identified by GreedyMI and Dense also predict better than gene markers identified by t-test, but not with statistical significance. To confirm that the improvement in performance was not dependent on classifier selection, we performed the same analysis using linear discriminant analysis (LDA) classification model. The average classification performance of LDA classifiers using a range of top 50 markers for each method is summarized in Figure 3.4.

We also compared against the classification performance of gene markers from the MAQC –II study (MAQC), where MAQC performance was calculated as the mean MCC of the best classifier reported by each of the 36 groups in the study. While OptDis still maintains higher overall MCC in cross-dataset experiments, the gene and subnetwork markers (Gene, GreedyMI, and Dense) show worse performance compared to the MAQC gene markers. This suggests that incorporating different data pre-processing steps and classification models, considered by MAQC groups, may improve prediction accuracy.

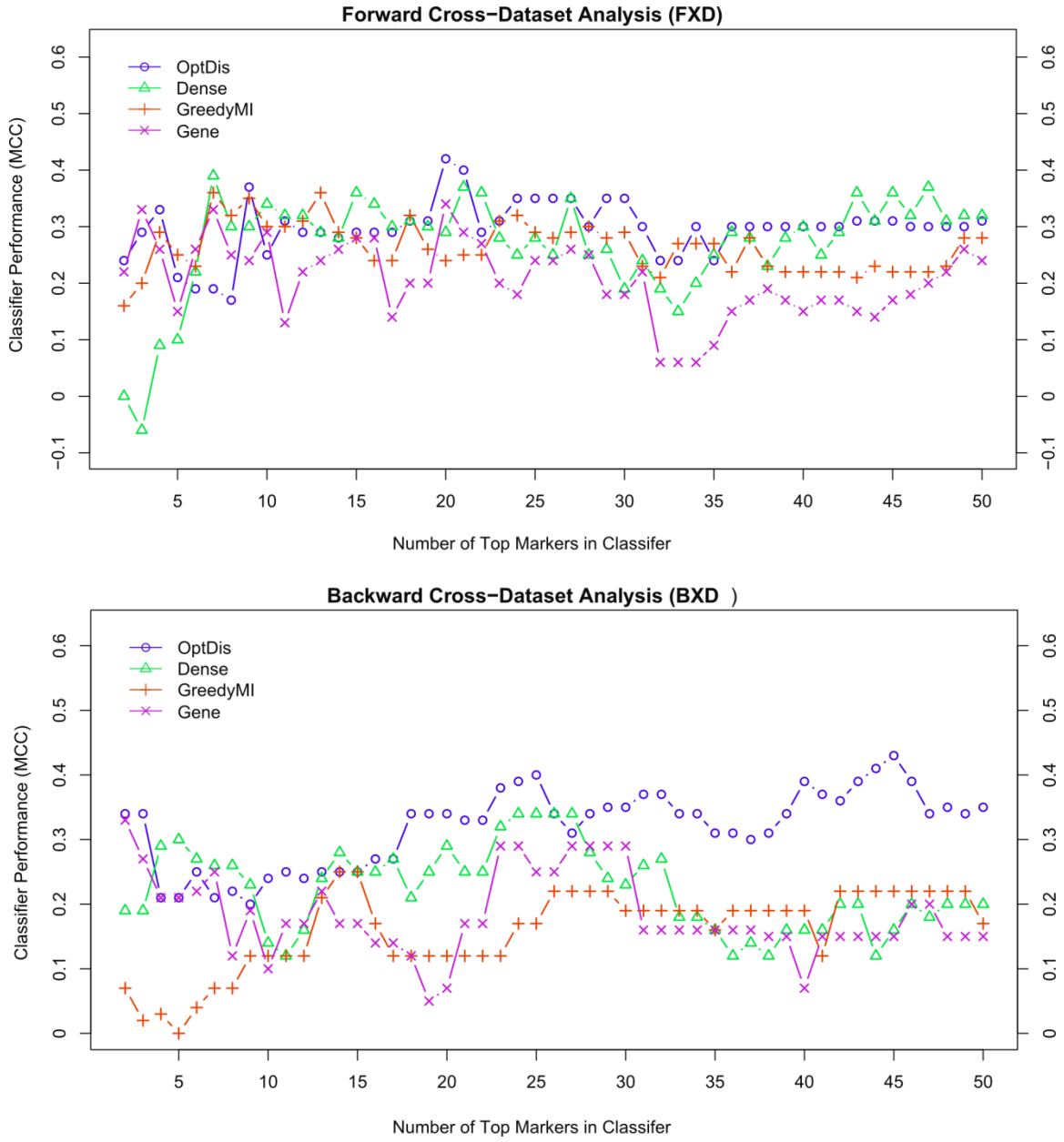


Figure 3.2: Cross-dataset performance for classifiers built using the top 1 to 50 markers derived from different methods. The *x*-axis indicates how many top markers were used in each classifier, where each marker can either a gene marker identified by t-test (Gene) or a subnetwork marker identified by GreedyMI [3], Dense [2], and our OptDis method [1].

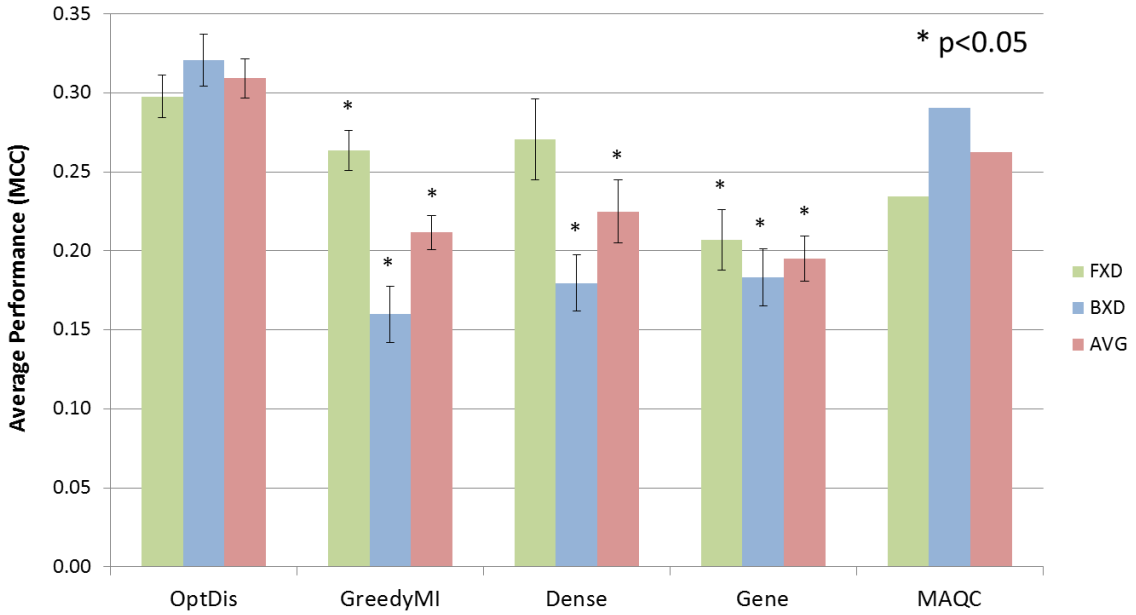


Figure 3.3: Average performance of kNN classifiers built using the top 50 gene or subnetwork markers from different methods. Gene markers are identified by t-test (Gene) or reported by the MAQC-II study (MAQC) [35]. Subnetwork markers are identified using GreedyMI [3], Dense [2], and our OptDis method [1]. Green and blue bars show the average classification performance in the FXD and BXD analyses respectively. Red bars show the overall performance in cross-dataset validation experiment, which is calculated as the mean of the values in the yellow and blue bars. Error bars shows the 95% confidence interval for the average performance. Asterisk indicates when classification performance of OptDis markers is statistically higher at $p<0.05$.

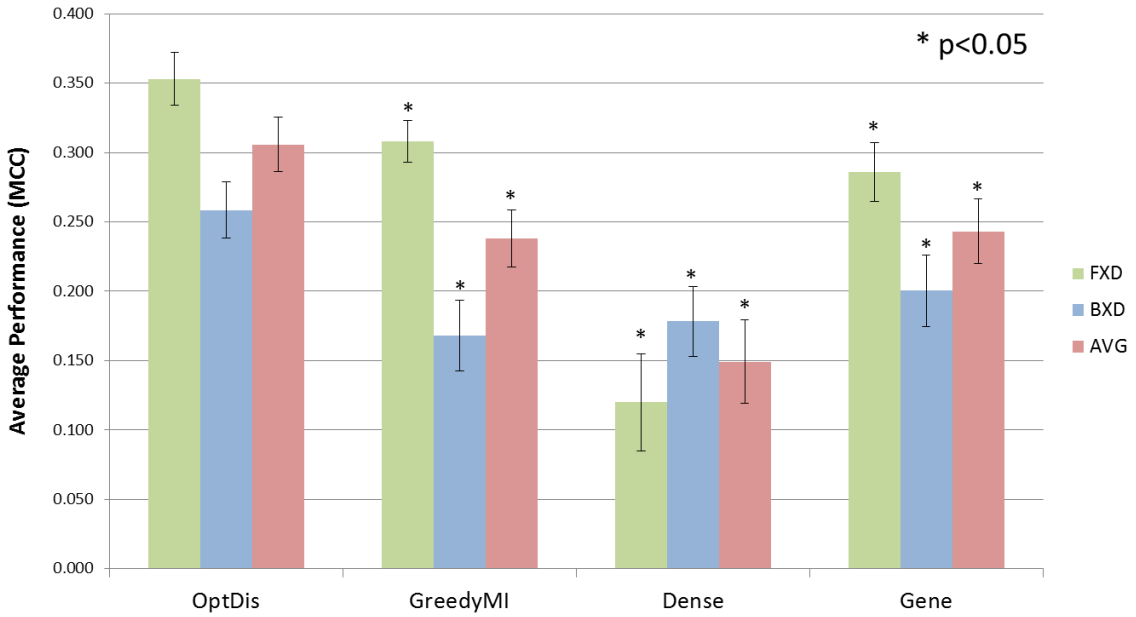


Figure 3.4: Average performance of LDA classifiers built using the top 50 gene or subnetwork markers from different methods. Gene markers are identified by t-test (Gene) or reported by the MAQC-II study (MAQC) [35]. Subnetwork markers are identified using GreedyMI [3], Dense [2], and our OptDis method [1]. Green and blue bars show the average classification performance in the FXD and BXD analyses respectively. Red bars show the overall performance in cross-dataset validation experiment, which is calculated as the mean of the values in the yellow and blue bars. Error bars shows the 95% confidence interval for the average performance. Asterisk indicates when classification performance of OptDis markers is statistically higher at $p<0.05$.

Next, we compared the overall cross-dataset validation performance of the top classifiers from different methods, as shown in Figure 3.5. As a reminder, the overall performance of markers from each method is calculated as the mean MCC of its top classifier from the FXD and BXD analyses. The top OptDis classifier shows similar performance across the two cross-dataset validation experiments and better performance over the top classifiers using gene markers and other subnetwork markers. The top three MAQC classifiers show comparable overall performance to the OptDis

classifier; however, they also suffer a significant difference in MCC between the FXD and BXD analyses. For example, the MAQC_GeneGo classifier has a high overall performance (MCC=0.43), but it also sustains the largest drop in performance between the FXD and BXD analysis (ΔMCC=0.25). The second and third best MAQC classifiers also show similar discrepancy in cross-dataset validation. In contrast, the top OptDis classifier shows consistent MCC when the datasets used for training and validation are swapped, which suggests that our OptDis method is more robust to noise and variations in the training data and may be applicable to different datasets.

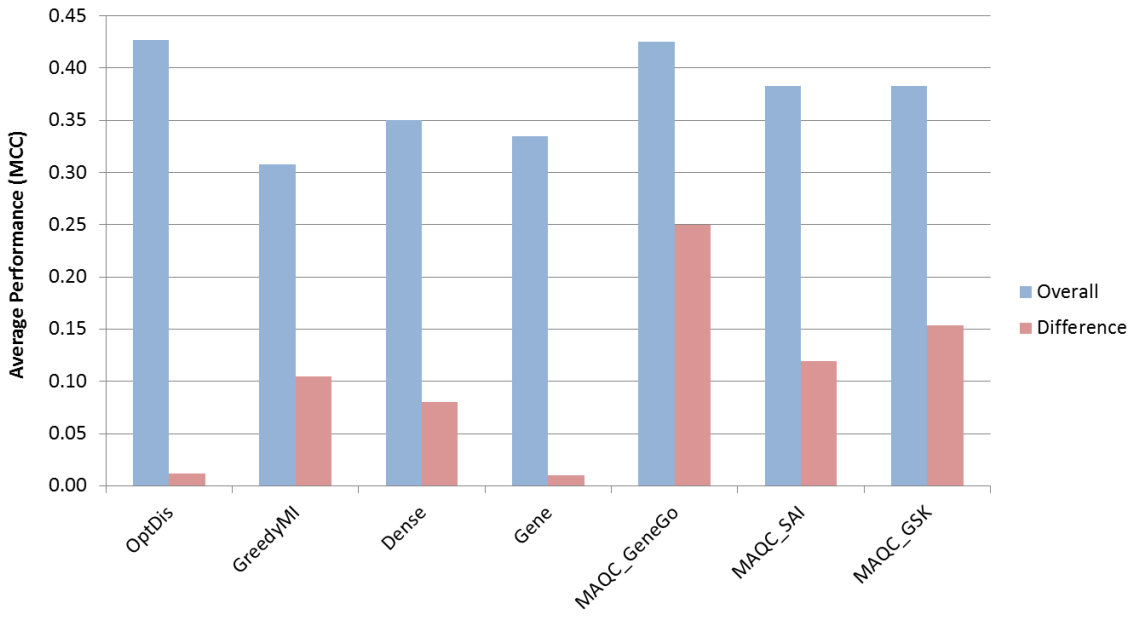


Figure 3.5: Overall cross-dataset performance of the top classifiers built using markers from different methods. Blue bars show overall performance of each classifier, calculated as the mean of the performance that classifier in the FXD and BXD analyses. Red bars show the absolute difference of performance for each classifier between the FXD and BXD analyses. Gene marker classifiers include markers derived by t-test (Gene) and the top 3 classifiers from MAQC-II study (MAQC GeneGo, MAQC SAI, MAQC GSK) [35]. Subnetwork markers are identified using GreedyMI [3], Dense [2], and our OptDis method [1].

Thus far, results from our cross-dataset experiments show that classifiers built using a range of top OptDis subnetwork markers consistently has greater performance compared to classifiers built using the same number of top gene markers (Figure 3.2). While our observation agrees with previous studies [2, 3, 21, 23], this event may simply result from subnetworks using more 'gene information' in each marker – a subnetwork marker contains *k* genes, whereas a gene marker contains 1 gene. Consequently, we investigated whether classifiers built from gene markers could more predict as well as those built from subnetwork markers if the same number of genes is used.

We repeated the earlier cross-dataset validation experiments and compared the classifier performance using the top 1 to top 50 subnetwork markers against the equivalent number of top gene markers identified by t-test (denoted Gene-EQ). The classification results are shown in Figure 3.6. The *x*-axis indicates the number of markers, where each marker is represents one subnetwork marker (containing *k* genes) or *k* gene markers equal to the number of genes in that subnetwork marker. From Figure 3.6, our results still show that OptDis subnetwork markers provides greater performance than the equivalent number of gene markers across the entire range of top markers.

Since a primary motivation for using subnetworks is its potential to identify functionally-related genes with weaker individual discrimination power, we also investigated the benefits of using subnetwork component genes for classification. For this assessment, the component genes taken from the earlier top subnetwork markers were each treated as individual gene markers (denoted as Gene-OptDis), and compared

to equivalent number of gene markers derived from t-test (denoted as Gene-Eq), as well as the subnetwork markers from which they came from (OptDis). These classification results are shown in Figure 3.6.

Interestingly, component gene markers tend to achieve better classification performance than conventional gene markers, but still not as good as subnetwork markers. This suggests two things. First, the OptDis subnetwork method may be better at identifying genes informative to predicting chemotherapy response. Second, summarizing many genes into fewer subnetworks may reduce the number of redundant gene markers, which in turn reduces the negative impact on classifier performance. Although additional experiments are warranted to substantiate these claims, our preliminary results suggest that aggregating multiple genes into a single subnetwork results in better marker for classification and highlights the importance of considering the collective effects of multiple weaker discriminative genes as functional module. Regarding the former claim, additional results in Section 3.6 suggest that the improved classification performance of constituent genes arises from using a distance-based discriminatory score.

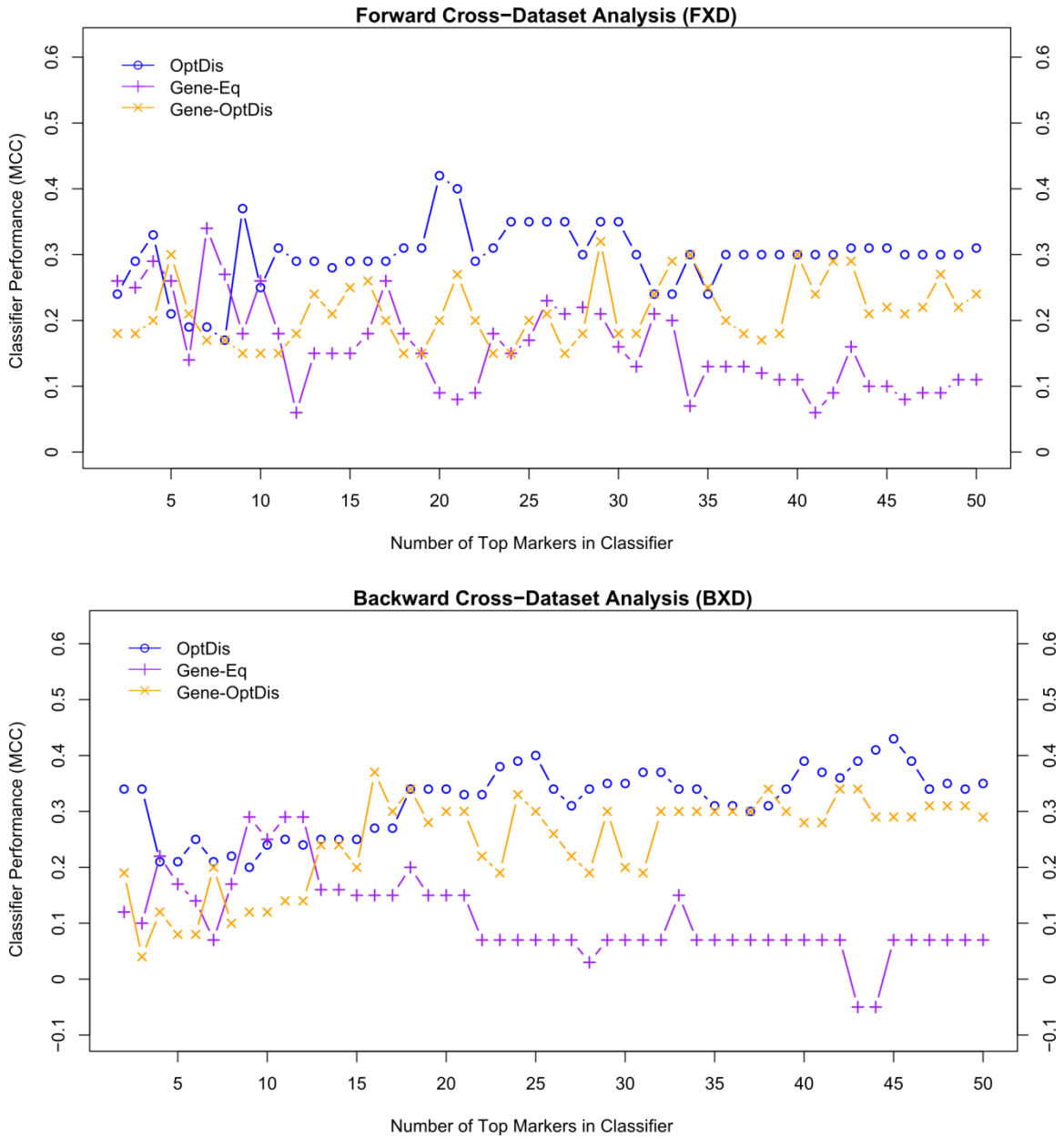


Figure 3.6: Cross-dataset performance for classifiers built using the top markers derived from different methods. The x-axis represents the performance of classifiers built using either the top OptDis subnetwork markers, the component genes from those subnetwork markers treated as gene markers, or the number of gene markers identified by t-test equivalent to the number of genes used in those subnetworks.

3.4 Reproducibility of Markers

We compared the reproducibility of subnetwork markers identified by OptDis against gene markers by deriving top markers from the two different cohorts of breast patients and calculating the number of overlapping genes. Since each subnetwork marker may comprise multiple genes, we compared subnetwork markers to an equivalent number of gene markers equal to the number of genes in those subnetworks (i.e. 1 subnetwork marker = k gene markers). The degree of gene overlap across a range of top markers is shown in Figure 3.7. With ten subnetwork markers, OptDis markers already have 25% reproducibility, which is much higher than the 8% reproducibility for an equivalent number of top gene markers. Although the percentage of overlap for gene markers increases as more genes are considered, it remains consistently lower than the reproducibility of subnetwork markers. The greater reproducibility of OptDis markers may contribute to its more robust performance in cross-dataset validation experiments.

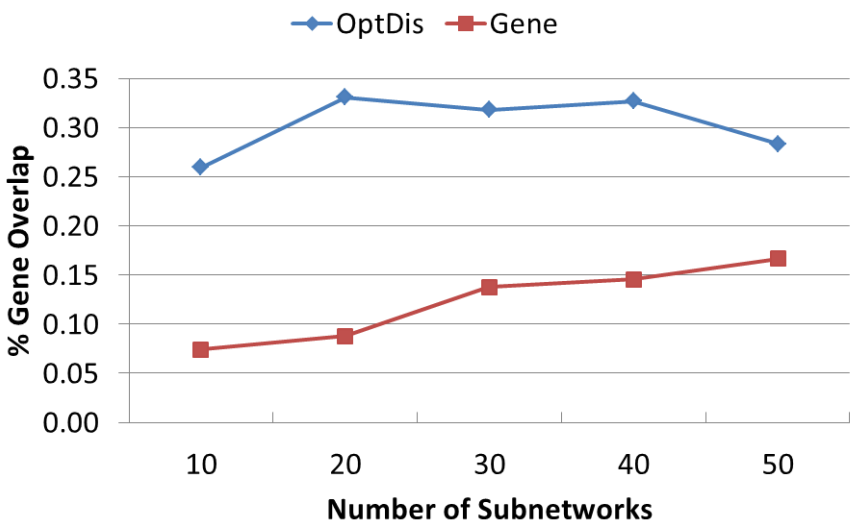


Figure 3.7: Comparison of marker reproducibility. Reproducibility is quantified as the degree of gene overlap between top markers identified from different datasets. This overlap is calculated for 10, 20, 30, 40, and 50 OptDis subnetworks and the equivalent number of genes derived from t-test.

3.5 Insights into Biological Mechanisms of Drug Response

3.5.1 Gene Function Enrichment Analysis

We hypothesized that reproducible genes may be more biologically relevant to the activity of TFAC therapy, so we examined the functions for the set of genes that were common between the two subnetwork signatures derived from the two different cohorts (of 100 and 130 patients). The 39 genes common between the two sets of top 50 subnetwork markers is denoted as O39 genes.

The significantly enriched biological functions for the O39 genes identified using the Ingenuity Pathway Analysis software (IPA; Ingenuity© Systems, www.ingenuity.com) are listed in Table 3.1. About half are implicated in apoptosis, suggesting that changes in strengths of pro-apoptotic and anti-apoptotic signals can induce resistance to

chemotherapy. There are also genes involved in DNA repair, which is expected given many of the anticancer drugs within TFAC therapy induce DNA damage (i.e. cyclophosphamide by cross-linking DNA strands).

Some of the 39 genes have specific functions related to mechanism of individual TFAC drugs. Paclitaxel is a mitotic inhibitor that stabilizes microtubule activity during mitosis and induces cell death. While paclitaxel is known to act on beta-tubulin, some studies [40] have also shown association between the actin and tubulin cytoskeleton in drug response, and suggest that regulation of actin cytoskeleton can induce sensitivity to mitotic-inhibitors. From our O39 list, the EVL, RET and CST3 genes have regulatory roles in the organization and assembly of actin filaments.

Fluouracil's primary anticancer activity blocks DNA replication by suppressing thymidyate synthetase activity and depleting thymidine [41]. *In vitro* studies have shown that AR and IGF2, from our O39 list, can increase incorporation of thymidine, which acts in antagonist to thymidyate synthetase suppression, to allow DNA synthesis through the actions of thymidine kinase [42, 43].

Doxorubicin is an anthracycline antibiotic that intercalates with DNA and causes double-stranded breaks to induce cell apoptosis or disruption in mitosis [44, 45]. SMAD3 from our list has been observed to affect BRCA1-dependent double-stranded DNA break repair in breast cancer cell lines and thus potentially may contribute to differential response to doxorubicin [46].

Table 3.1: Gene function enrichment analysis for the O39 genes. Some of the functions relevant to TFAC chemotherapy response are shown. The enrichment significance for each function is provided as a Benjamini–Hochberg adjusted p-value.

Functions	Gene Symbols	p-value
Apoptosis	AR, EP300, ESR1, GADD45G, IGF2, IGF1R, IGFBP4, IL6ST, MAPK3, MDM2, MED1, NCOA3, PRKACA, RARA, RET, SHC1, SMAD3, SRC, TSC2	1.27E-06
DNA synthesis	AR, ESR1, IGF2, IGFBP4, IL6ST, MDM2, SHC1, SRC	1.74E-06
Actin filament organization	EVL, CST3, RET, SRC, TSC2	7.16E-03
DNA repair	GADD45G, MDM2, RARA, SMAD3	1.89E-02

3.5.2 Pathway Enrichment Analysis

We investigated biological insights into the mechanisms chemotherapy response offered by OptDis subnetwork markers by deriving the top 50 subnetwork markers derived from the combined cohort of 230 patients and using Ingenuity Pathway Analysis software (IPA; Ingenuity© Systems,www.ingenuity.com) to identify significantly enriched pathways. As a baseline for comparison, we also derived two sets of gene markers using t-test from the same dataset: the top 50 gene markers and the top 111 gene markers, which is equivalent to number of genes in the top 50 subnetwork markers. The results are shown in Figure 3.8. Interestingly, several of the top signaling pathways enriched by subnetwork markers were associated with chemotherapy response, whereas no significantly enriched pathways were found for the T50 or T111 gene markers.

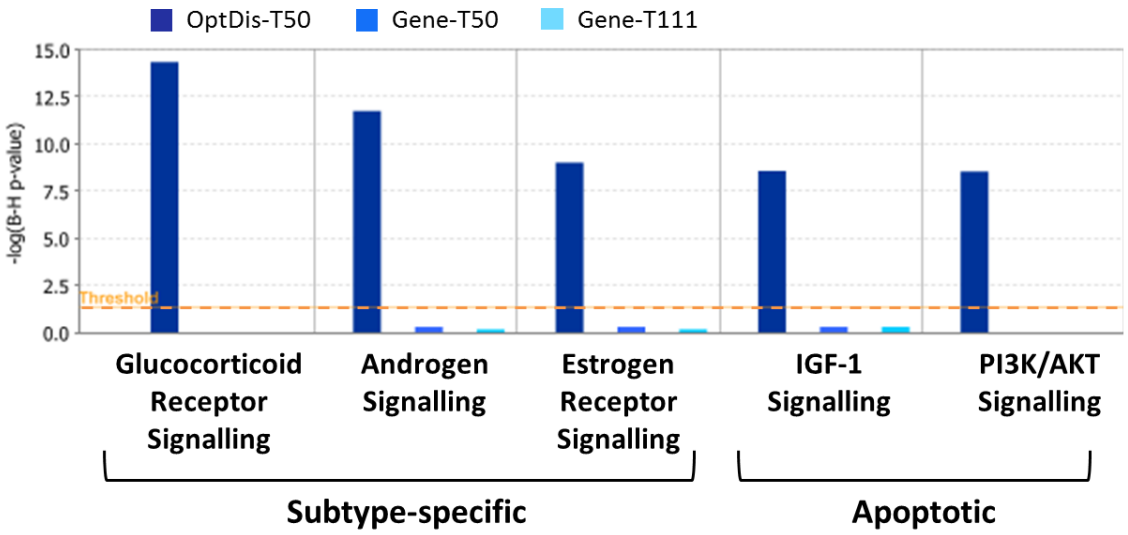


Figure 3.8: Pathway enrichment analysis for the top 50 OptDis subnetwork markers. The top five enriched signaling pathways associated with TFAC response are shown. The enrichment level of the same pathways for the top 50 (middle bar) and top 111 genes (right bar) derived from t-test are also shown. The enrichment level of each pathway is provided in terms of p-value adjusted with Benjamini-Hochberg, where enrichment above 0.05 (dotted line) indicates significance.

A closer examination of these top associated pathways suggests response to TFAC treatment is affected by the cross-talk between tumor subtype specific mechanisms and pathways regulating apoptosis. Chemotherapy response in breast cancer have been observed to be subtype-specific [47], with ER+ tumors exhibiting much higher response rates to taxane-based therapies than ER− tumors [30, 48, 49]. Therefore, it was expected to find that the predictive subnetwork signature was strongly enriched for genes activating the estrogen receptor (ER) signaling pathway. For the same reason, we also observe enrichment for the androgen receptor (AR) signaling pathway. With nearly all ER+ tumors and few ER− tumors showing AR expression [50], it is likely that AR-based subnetworks serve as good markers of TFAC treatment based on their association with

ER status. Experimental studies have shown that expression of ERα selectively inhibits paclitaxel-induced apoptosis through modulation of glucocorticoid receptor activity [51]. Based on the enriched IPA pathways associated with response, we speculate that the differential response between subtypes may be attributed to differential regulation of apoptosis.

Other response-associated pathways may also contribute to differential response to TFAC treatment. For example, signalling of insulin-like growth factor has known functions in cancer proliferation and inhibition of apoptosis, and has been experimentally implicated in chemotherapy resistance [52, 53]. The PI3K/AKT pathway can also increase resistance to taxane-based therapies through downstream anti-apoptotic effectors BCL-2 and BCL-XL [54]. Experiments have shown that tumors with increased phosphorylated BCL-2 expression have increased sensitivity to paclitaxel compared with tumors with reduced expression [55].

3.6 Source of Performance Improvements in OptDis Method

There are four primary factors that may contribute to the improved classification performance of subnetwork markers identified by OptDis:

1. Aggregating gene markers into subnetwork markers
2. Distance-based discrimination score function
3. Use of prior knowledge from PPI networks
4. A search algorithm that retrieves optimally discriminative markers

We investigated the first aspect in Section 3.3.4 and concluded that aggregating genes into subnetwork markers improves classification performance. In this section, we focus on assessing the value of aspects two and three.

We evaluated the worth of the distance-based discrimination score by comparing the average cross-dataset validation performance of the top gene markers identified by the distance score (Gene-Distance) against the top gene markers identified by t-test (Gene-TTest) and the top subnetwork markers from OptDis (OptDis-HPRD). Similar to previous experiments, average performance is calculated average of the 50 classifiers built across the range of top markers. The results are shown in Figure 3.9. Comparing these values, we can see that the gene markers identified by distance-score offer better classification performance than gene markers identified by t-test in both FXD and BXD analyses. This observation supports our motivations for using a distance-based scoring function over a statistical scoring function (described in Section 2.2.2). Surprisingly, the distance-based function appears to contribute to the entire the performance improvements demonstrated by OptDis in the BXD analysis.

We evaluated the value of using prior knowledge, in the form of known protein-protein interactions from PPI networks, to guide marker discovery. To investigate, we re-ran OptDis on randomized networks generated using the Erdős–Rényi model. This model produces a random network by swapping the edges in the network, while maintaining degree distribution of each node. The classification performance of OptDis markers identified using the true PPI network (OptDis-HPRD) against the random network (OptDis-Random) is shown in Figure 3.9.

If interaction knowledge from PPI networks is useful, then the performance of OptDis should decrease when it runs on the random network. In the FXD analysis, we observe a significant drop in performance between OptDis-HPRD and OptDis-Random; however in the BXD analysis, the performance of OptDis-Random stays the same as OptDis-HPRD. Results from the FXD analysis suggest that knowledge of protein-protein interactions improves marker discovery and classification performance. To explain the BXD analysis, we note that the performance of subnetwork markers using the random networks appears to exactly coincide with the performance of gene markers using the distance-based scoring function in both directions of analyses. This suggests that when OptDis does not find edge information from the random networks to help in marker discovery, it simply relies on the discrimination of individual genes in the network. We confirm this view by observing that most of the OptDis-Random subnetworks comprise only one to two genes.

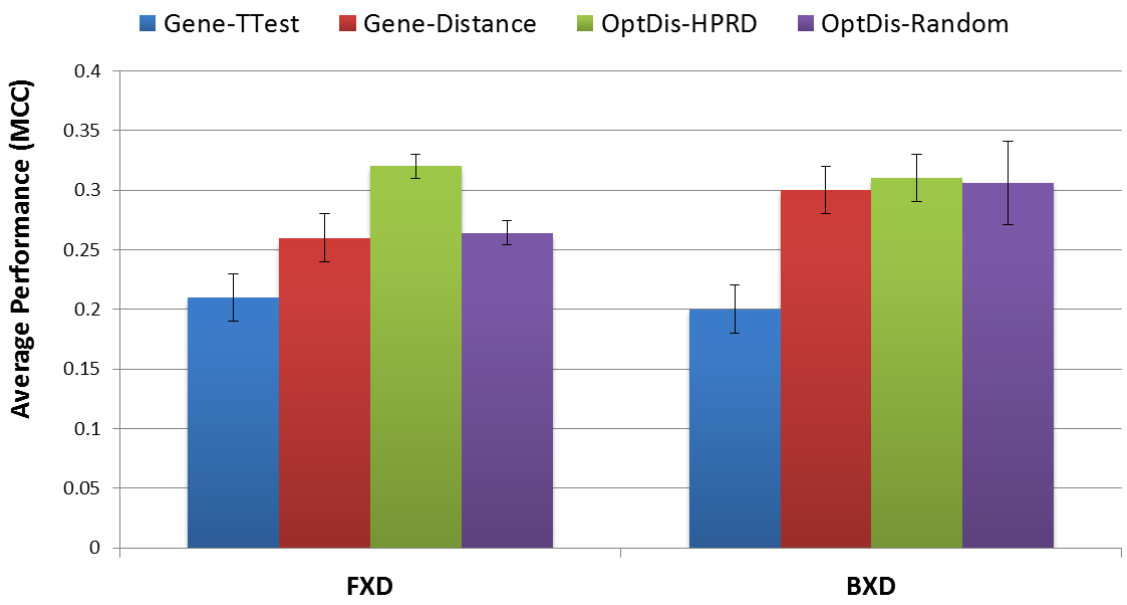


Figure 3.9: Classification performance for different components in OptDis. Bar plots show performance for gene markers identified using t-test (Gene-TTest) and our distance-based scoring function (Gene-Distance), and subnetwork markers identified by OptDis and extracted from the original PPI network (OptDis-HPRD) and the random PPI network (OptDis-Random).The performance OptDis on the random network is reported as the average of its performance across 20 randomly networks generated using Erdős-Rényi model. Performance is reported as the mean MCC across the range of top 1-50 markers.

We did not empirically validate that optimally discriminative subnetworks retrieved by our search algorithm classify better than the subnetworks found by a heuristical search algorithm; however, if a marker can better discriminate different groups, then they should also achieve better classification performance.

Chapter 4
Conclusions

The main contributions of this thesis are the development and application of a network-based approach (OptDis) to efficiently identify subnetwork markers optimally discriminative of clinical outcome. We demonstrate the advantages of our method in the challenging task of predicting patient response to chemotherapy. Our results indicate that subnetwork markers identified by OptDis provide improved performance in classifying response outcome compared to gene markers and subnetwork markers identified by other methods. In addition, subnetwork markers offer the significant benefits of greater reproducibility across cohorts and insights into the molecular mechanisms underlying differential chemotherapy response. We also provide direct evidence supporting the benefit of using protein interactions knowledge from PPI networks in improving marker discovery and classification performance.

4.1 Limitations & Future Directions

While our OptDis method has provided promising results in this study, there are still some aspects about it that can be improved. One aspect is that OptDis infers the activity of subnetworks by taking the mean expression of component genes. While mean-based aggregation is both simple to implement and commonly used in other subnetwork marker methods, this approach restricts our algorithm to finding only subnetworks

whose component genes are under the same direction of regulation: genes must be up-regulated or all down-regulated in group A versus group B. If two functionally-related genes are regulated in opposite directions, then calculating the mean would neutralize these opposing signals. Since signaling pathways in real biological systems comprise genes under a mixture of activating and inhibiting regulation, it is critical to capture this property in the method. One way to address this shortcoming would be to fit a regression function using the component gene expressions and to infer subnetwork activity from the fitted function.

A second aspect is that some of the top ranked subnetwork markers derived by OptDis may have correlated activities, which can lower the classifier's predictive performance. Two subnetworks can be correlated if many of component genes from the first subnetwork have correlated expressions with many of the component genes in the second subnetwork. Instead of removing the correlated subnetwork markers, which may provide informative biology knowledge, this issue could be addressed by merging the correlated markers into one feature.

A final aspect is that the prior knowledge provided by current human PPI networks (and used by OptDis) is not tissue-specific, whereas there is accumulating evidence to support the existence of tissue-specific sets of protein interactions [56]. Results from Section 3.6 show that the performance of subnetwork markers identified by OptDis can depend on the quality of the network, so a corollary from this observation is that more accurate subnetwork markers may be identified using true tissue-specific interactions. One way to infer tissue-specific interactions from the PPI network would be to assign a

confidence weight to each interaction in the PPI network based on the co-expression levels between those interacting genes across many gene expression datasets from the same tissue [57]. The OptDis search algorithm can be modified to consider the interactions weights when it looks for optimally discriminative subnetworks.

Before OptDis subnetwork markers can be considered for clinical use, the utility of these markers will need to be further assessed by comparing their classification performance against the predictive accuracy of traditional clinical parameters or existing clinical predictive tools (i.e. nomograms) used for selecting chemotherapy treatment [58]. In addition to predicting response outcome, the OptDis method can also be applied to classification problems in cancer research such as predicting low versus high cancer prognosis groups or discriminating between benign and malignant tumours.

Bibliography

1. Dao P, Wang K, Collins C, Ester M, Lapuk A, Sahinalp SC: **Optimally discriminative subnetwork markers predict response to chemotherapy.** *Bioinformatics* 2011, **27**:i205-i213.

2. Dao P, Colak R, Salari R, Moser F, Davicioni E, Schönhuth A, Ester M: **Inferring cancer subnetwork markers using density-constrained biclustering.** *Bioinformatics* 2010, **26**:i625-31.

3. Chuang H-Y, Lee E, Liu Y-T, Lee D, Ideker T: **Network-based classification of breast cancer metastasis.** *Molecular Systems Biology* 2007, **3**:140.

4. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH: **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature* 2002, **415**:530-6.

5. Hess KR, Anderson K, Symmans WF, Valero V, Ibrahim N, Mejia J a, Booser D, Theriault RL, Buzdar AU, Dempsey PJ, Rouzier R, Sneige N, Ross JS, Vidaurre T, Gómez HL, Hortobagyi GN, Pusztai L: **Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer.** *Journal of Clinical Oncology* 2006, **24**:4236-44.

6. Rouzier R, Rajan R, Wagner P, Hess KR, Gold DL, Stec J, Ayers M, Ross JS, Zhang P, Buchholz T a, Kuerer H, Green M, Arun B, Hortobagyi GN, Symmans WF, Pusztai L: **Microtubule-associated protein tau: a marker of paclitaxel sensitivity in breast cancer.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**:8315-20.

7. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**:15545-50.

8. Michiels S, Koscielny S, Hill C: **Prediction of cancer outcome with microarrays: a multiple random validation strategy.** *Lancet* 2005, **365**:488-92.

9. Ein-Dor L, Zuk O, Domany E: **Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer.** *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**:5923-8.

10. Ein-Dor L, Kela I, Getz G, Givol D, Domany E: **Outcome signature genes in breast cancer: is there a unique set?** *Bioinformatics* 2005, **21**:171-8.

11. Rives AW, Galitski T: **Modular organization of cellular networks.** *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**:1128-33.

12. Barabási A-L, Oltvai ZN: **Network biology: understanding the cell's functional organization.** *Nature Reviews Genetics* 2004, **5**:101-13.

13. Guo Z, Zhang T, Li X, Wang Q, Xu J, Yu H, Zhu J, Wang H, Wang C, Topol EJ, Wang Q, Rao S: **Towards precise classification of cancers based on robust gene functional expression profiles.** *BMC Bioinformatics* 2005, **6**:58.

14. Svensson JP, Stalpers LJ a, Esveldt-van Lange REE, Franken N a P, Haveman J, Klein B, Turesson I, Vrieling H, Giphart-Gassler M: **Analysis of gene expression using gene sets discriminates cancer patients with and without late radiation toxicity.** *PLoS Medicine* 2006, **3**:e422.

15. Abraham G, Kowalczyk A, Loi S, Haviv I, Zobel J: **Prediction of breast cancer prognosis using gene set statistics provides signature stability and biological context.** *BMC Bioinformatics* 2010, **11**:277.

16. Tian L, Greenberg S a, Kong SW, Altschuler J, Kohane IS, Park PJ: **Discovering statistically significant pathways in expression profiling studies.** *Proceedings of the National Academy of Sciences* 2005, **102**:13544-9.

17. Rapaport F, Zinovyev A, Dutreix M, Barillot E, Vert J-P: **Classification of microarray data using gene networks.** *BMC Bioinformatics* 2007, **8**:35.

18. Lee E, Chuang H-Y, Kim J-W, Ideker T, Lee D: **Inferring pathway activity toward precise disease classification.** *PLoS Computational Biology* 2008, **4**:e1000217.

19. Su J, Yoon B-J, Dougherty ER: **Accurate and Reliable Cancer Classification Based on Probabilistic Inference of Pathway Activity**. *PloS One* 2009, **4**:e8161.

20. Chowdhury SA, Nibbe RK, Chance MR, Koyutürk M: **Subnetwork state functions define dysregulated subnetworks in cancer.** *Journal of Computational Biology* 2011, **18**:263-81.

21. Su J, Yoon B-J, Dougherty ER: **Identification of diagnostic subnetwork markers for cancer in human protein-protein interaction network.** *BMC Bioinformatics* 2010, **11 Suppl 6**:S8.

22. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, von Mering C: **The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored.** *Nucleic Acids Research* 2011, **39**:D561-8.

23. Fortney K, Kotlyar M, Jurisica I: **Inferring the functions of longevity genes with modular subnetwork biomarkers of Caenorhabditis elegans aging.** *Genome biology* 2010, **11**:R13.

24. Weigelt B, Pusztai L, Ashworth A, Reis-Filho JS: **Challenges translating breast cancer gene signatures into the clinic.** *Nature Reviews Clinical Oncology* 2011, **9**:58-64.

25. Fumagalli D, Desmedt C, Ignatiadis M, Loi S, Piccart M, Sotiriou C: **Gene profiling assay and application: the predictive role in primary therapy.** *JNCI Monographs* 2011, **2011**:124-7.

26. Chang JC, Wooten EC, Tsimelzon A, Hilsenbeck SG, Gutierrez MC, Elledge R, Mohsin S, Osborne CK, Chamness GC, Allred DC, O'Connell P: **Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer.** *The Lancet* 2003, **362**:362-9.

27. Cleator S, Tsimelzon A, Ashworth A, Dowsett M, Dexter T, Powles T, Hilsenbeck S, Wong H, Osborne CK, O'Connell P, Chang JC: **Gene expression patterns for doxorubicin (Adriamycin) and cyclophosphamide (cytoxan) (AC) response and resistance.** *Breast Cancer Research and Treatment* 2006, **95**:229-33.

28. Naoi Y, Kishi K, Tanei T, Tsunashima R, Tominaga N, Baba Y, Kim SJ, Taguchi T, Tamaki Y, Noguchi S: **Prediction of pathologic complete response to sequential paclitaxel and 5-fluorouracil/epirubicin/cyclophosphamide therapy using a 70-gene classifier for breast cancers.** *Cancer* 2011, **117**:3682-90.

29. Tabchy A, Valero V, Vidaurre T, Lluch A, Gomez H, Martin M, Qi Y, Barajas-Figueroa LJ, Souchon E, Coutant C, Doimi FD, Ibrahim NK, Gong Y, Hortobagyi GN, Hess KR, Symmans WF, Pusztai L: **Evaluation of a 30-gene paclitaxel, fluorouracil, doxorubicin, and cyclophosphamide chemotherapy response predictor in a multicenter randomized trial in breast cancer.** *Clinical Cancer Research* 2010, **16**:5351-61.

30. Farmer P, Bonnefoi H, Anderle P, Cameron D, Wirapati P, Wirapati P, Becette V, André S, Piccart M, Campone M, Brain E, Macgrogan G, Petit T, Jassem J, Bibeau F, Blot E, Bogaerts J, Aguet M, Bergh J, Iggo R, Delorenzi M: **A stroma-related gene signature**

predicts resistance to neoadjuvant chemotherapy in breast cancer. *Nature Medicine* 2009, **15**:68-74.

31. Juul N, Szallasi Z, Eklund AC, Li Q, Burrell R a, Gerlinger M, Valero V, Andreopoulou E, Esteva FJ, Symmans WF, Desmedt C, Haibe-Kains B, Sotiriou C, Pusztai L, Swanton C: **Assessment of an RNA interference screen-derived mitotic and ceramide pathway metagene as a predictor of response to neoadjuvant paclitaxel for primary triple-negative breast cancer: a retrospective analysis of five clinical trials.** *The Lancet Oncology* 2010, **11**:358-65.

32. Alon N, Yuster R, Zwick U: **Color-coding**. *Journal of the ACM* 1995, **42**:844-856.

33. Alon N, Dao P, Hajirasouliha I, Hormozdiari F, Sahinalp SC: **Biomolecular network motif counting and discovery by color coding.** *Bioinformatics* 2008, **24**:i241-9.

34. Dao P, Hormozdiari F: **Quantifying systemic evolutionary changes by color coding confidence-scored ppi networks**. *In 9th International Workshop on Algorithms in Bioinformatics* 2009:37-48.

35. Shi L, Campbell G, Jones WD, et al.: **The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models.** *Nature Biotechnology* 2010, **28**:827-38.

36. Irizarry R a, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics* 2003, **4**:249-64.

37. Johnson WE, Li C, Rabinovic A: **Adjusting batch effects in microarray expression data using empirical Bayes methods.** *Biostatistics* 2007, **8**:118-27.

38. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A: **Human Protein Reference Database--2009 update.** *Nucleic Acids Research* 2009, **37**:D767-72.

39. Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H: **Assessing the accuracy of prediction algorithms for classification: an overview**. *Bioinformatics* 2000, **16**:412-424.

40. Kavallaris M: **Microtubules and resistance to tubulin-binding agents.** *Nature Reviews Cancer* 2010, **10**:194-204.

41. Longley DB, Harkin DP, Johnston PG: **5-fluorouracil: mechanisms of action and clinical strategies.** *Nature Reviews Cancer* 2003, **3**:330-8.

42. Pedram A, Razandi M, Sainson RC a, Kim JK, Hughes CC, Levin ER: **A conserved mechanism for steroid receptor translocation to the plasma membrane.** *The Journal of biological chemistry* 2007, **282**:22278-88.

43. Yang CQ, Zhan X, Hu X, Kondepudi A, Perdue JF: **The expression and characterization of human recombinant proinsulin-like growth factor II and a mutant that is defective in the O-glycosylation of its E domain.** *Endocrinology* 1996, **137**:2766-73.

44. Minotti G, Menna P, Salvatorelli E, Cairo G, Gianni L: **Anthracyclines: molecular advances and pharmacologic developments in antitumor activity and cardiotoxicity.** *Pharmacological Reviews* 2004, **56**:185-229.

45. Munro AF, Cameron DA, Bartlett JMS: **Targeting anthracyclines in early breast cancer: new candidate predictive biomarkers emerge.** *Oncogene* 2010, **29**:5231-40.

46. Dubrovska A, Kanamoto T, Lomnytska M, Heldin C-H, Volodko N, Souchelnytskyi S: **TGFbeta1/Smad3 counteracts BRCA1-dependent repair of DNA damage.** *Oncogene* 2005, **24**:2289-97.

47. Sørlie T, Wang Y, Xiao C, Johnsen H, Naume B, Samaha RR, Børresen-Dale A-L: **Distinct molecular mechanisms underlying clinically relevant subtypes of breast cancer: gene expression analyses across three different platforms.** *BMC genomics* 2006, **7**:127.

48. Liedtke C, Mazouni C, Hess KR, André F, Tordai A, Mejia JA, Symmans WF, Gonzalez-Angulo AM, Hennessy B, Green M, Cristofanilli M, Hortobagyi GN, Pusztai L: **Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer.** *Journal of Clinical Oncology* 2008, **26**:1275-81.

49. Popovici V, Chen W, Gallas BG, Hatzis C, Shi W, Samuelson FW, Nikolsky Y, Tsyganova M, Ishkin A, Nikolskaya T, Hess KR, Valero V, Booser D, Delorenzi M, Hortobagyi GN, Shi L, Symmans WF, Pusztai L: **Effect of training-sample size and classification difficulty on the accuracy of genomic predictors.** *Breast Cancer Research and Treatment* 2010, **12**:R5.

50. Niemeier LA, Dabbs DJ, Beriwal S, Striebel JM, Bhargava R: **Androgen receptor in breast cancer: expression in estrogen receptor-positive tumors and in estrogen receptor-negative tumors with apocrine differentiation.** *Modern Pathology* 2010, **23**:205-12.

51. Sui M, Huang Y, Park BH, Davidson NE, Fan W: **Estrogen receptor alpha mediates breast cancer cell resistance to paclitaxel through inhibition of apoptotic cell death.** *Cancer Research* 2007, **67**:5337-44.

52. Gooch JL, Van Den Berg CL, Yee D: **Insulin-like growth factor (IGF)-I rescues breast cancer cells from chemotherapy-induced cell death--proliferative and anti-apoptotic effects.** *Breast Cancer Research and Treatment* 1999, **56**:1-10.

53. Benini S, Manara MC, Baldini N, Cerisano V, Massimo Serra, Mercuri M, Lollini PL, Nanni P, Picci P, Scotlandi K: **Inhibition of insulin-like growth factor I receptor increases the antitumor activity of doxorubicin and vincristine against Ewing's sarcoma cells.** *Clinical Cancer Research* 2001, **7**:1790-7.

54. McGrogan BT, Gilmartin B, Carney DN, McCann A: **Taxanes, microtubules and chemoresistant breast cancer.** *Biochimica et Biophysica Acta* 2008, **1785**:96-132.

55. Shitashige M, Toi M, Yano T, Shibata M, Matsuo Y, Shibasaki F: **Dissociation of Bax from a Bcl-2/Bax heterodimer triggered by phosphorylation of serine 70 of Bcl-2.** *Journal of Biochemistry* 2001, **130**:741-8.

56. Bossi A, Lehner B: **Tissue specificity and the human protein interaction network.** *Molecular Systems Biology* 2009, **5**:260.

57. Lopes TJS, Schaefer M, Shoemaker J, Matsuoka Y, Fontaine J-F, Neumann G, Andrade-Navarro MA, Kawaoka Y, Kitano H: **Tissue-specific subnetworks and characteristics of publicly available human protein interaction databases.** *Bioinformatics (Oxford, England)* 2011, **27**:2414-21.

58. Boulesteix A-L, Sauerbrei W: **Added predictive value of high-throughput molecular data to clinical data and its validation.** *Briefings in Bioinformatics* 2011, **12**:215-29.

59. Baggerly KA, Coombes KR, Neeley ES: **Run batch effects potentially compromise the usefulness of genomic signatures for ovarian cancer.** *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2008, **26**:1186-7; author reply 1187-8.

60. Nielsen TO, West RB, Linn SC, Alter O, Knowling MA, O'Connell JX, Zhu S, Fero M, Sherlock G, Pollack JR, Brown PO, Botstein D, van de Rijn M: **Molecular characterisation of soft tissue tumours: a gene expression study.** *Lancet* 2002, **359**:1301-7.

61. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA: **Tackling the widespread and critical impact of batch effects in high-throughput data.** *Nature Reviews Genetics* 2010, **11**:733-9.

62. McCall MN, Bolstad BM, Irizarry RA: **Frozen robust multiarray analysis (fRMA).** *Biostatistics* 2010, **11**:242-53.

Appendices

A. Supporting Details for Methods

A.1 Analysis and Removal of Batch Effect

Batch effects are non-biological variations or systematic biases, which are frequently observed in gene expression data. These biases can be caused by many factors including sample preparation procedure, assay run date [59], and platform differences [60]. These effects can be wide-spread and if not accounted for, may even confound true biological differences and provide invalid conclusions [61].

Since samples in the two cohorts in the BrCa dataset used in this study were collected and processed in different years, we looked for the presence of batch effect prior to data analysis. We identified the presence of batch effect in the dataset by principal component analysis (PCA). The results of PCA analysis on the raw gene expression for the pooled cohort samples are shown in Figure A.1. The first principal component (PC) explains over 40% of the total variance in this dataset and showed a large difference in expression levels between samples in the two cohorts. In contrast, the second PC had comparable expression levels between the cohorts as expected.

We first attempted to correct for batch effects using RMA normalization. However, as seen in Figure A.2, after normalization, the first PC still accounted for 20% of total variance and maintained large difference in expression levels between the two cohorts.

This result is not unexpected given that normalization tends to correct for global effects and not subsets of genes that is affected by batch effect [61].

Next, we tried two statistical methods, ComBat and fRMA, developed specifically to correct for batch effects [37, 62]. Figure A.3 shows that ComBat appeared to be successful such the expression of the first PC became comparable between the two cohorts. In contrast, PCA on fRMA adjusted gene expressions provided little change from the RMA normalised gene expressions (Figure A.4)

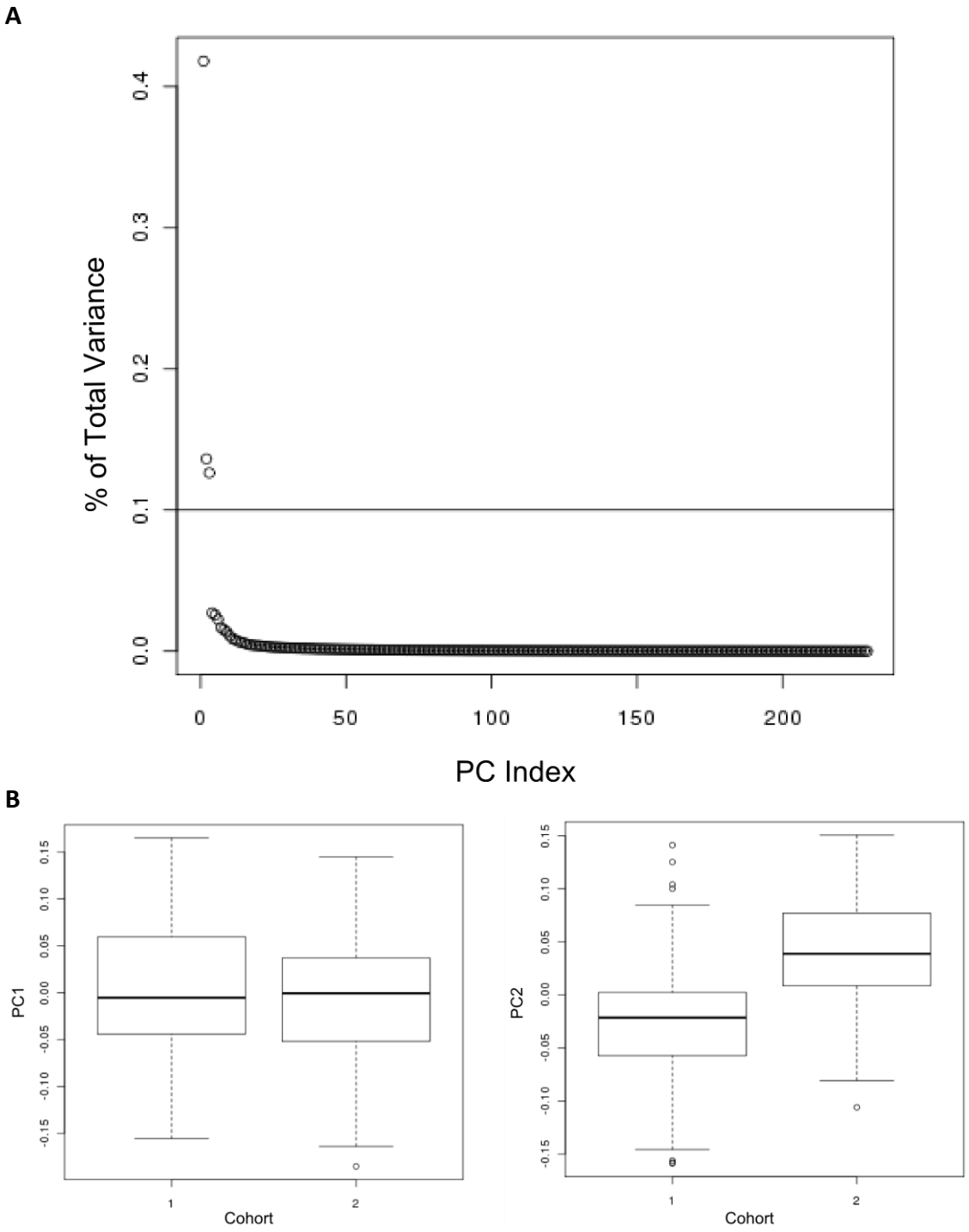


Figure A.1: PCA analysis on raw expression. (A) Fraction of overall variance contributed by the top principal components (PC). **(B)** Boxplots show the expression levels of the first and second PCs in cohort 1 and cohort 2 samples.

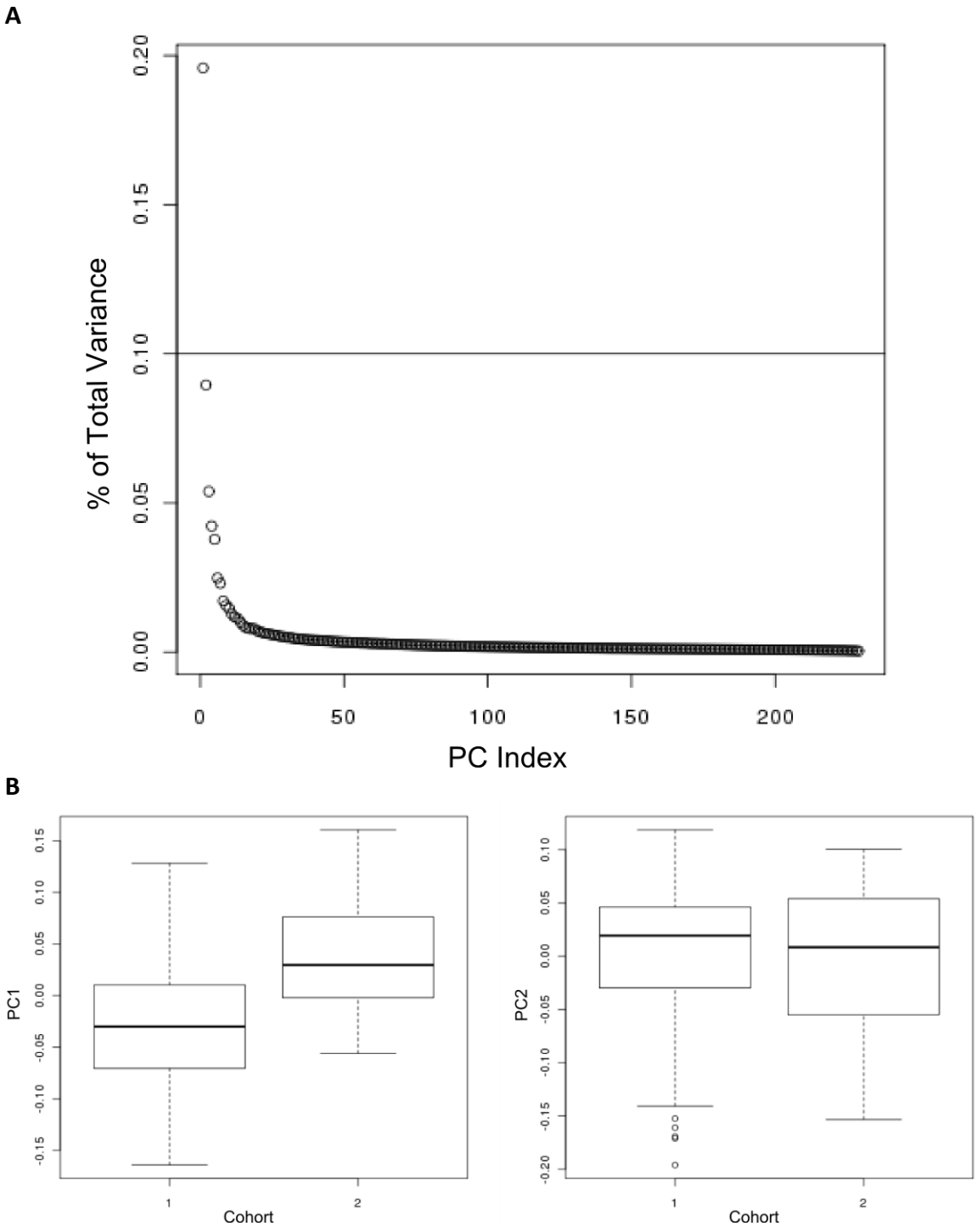


Figure A.2: PCA analysis on RMA normalised expression. (A) Fraction of overall variance contributed by the top principal components (PC). **(B)** Boxplots show the expression levels of the first and second PCs in cohort 1 and cohort 2 samples.

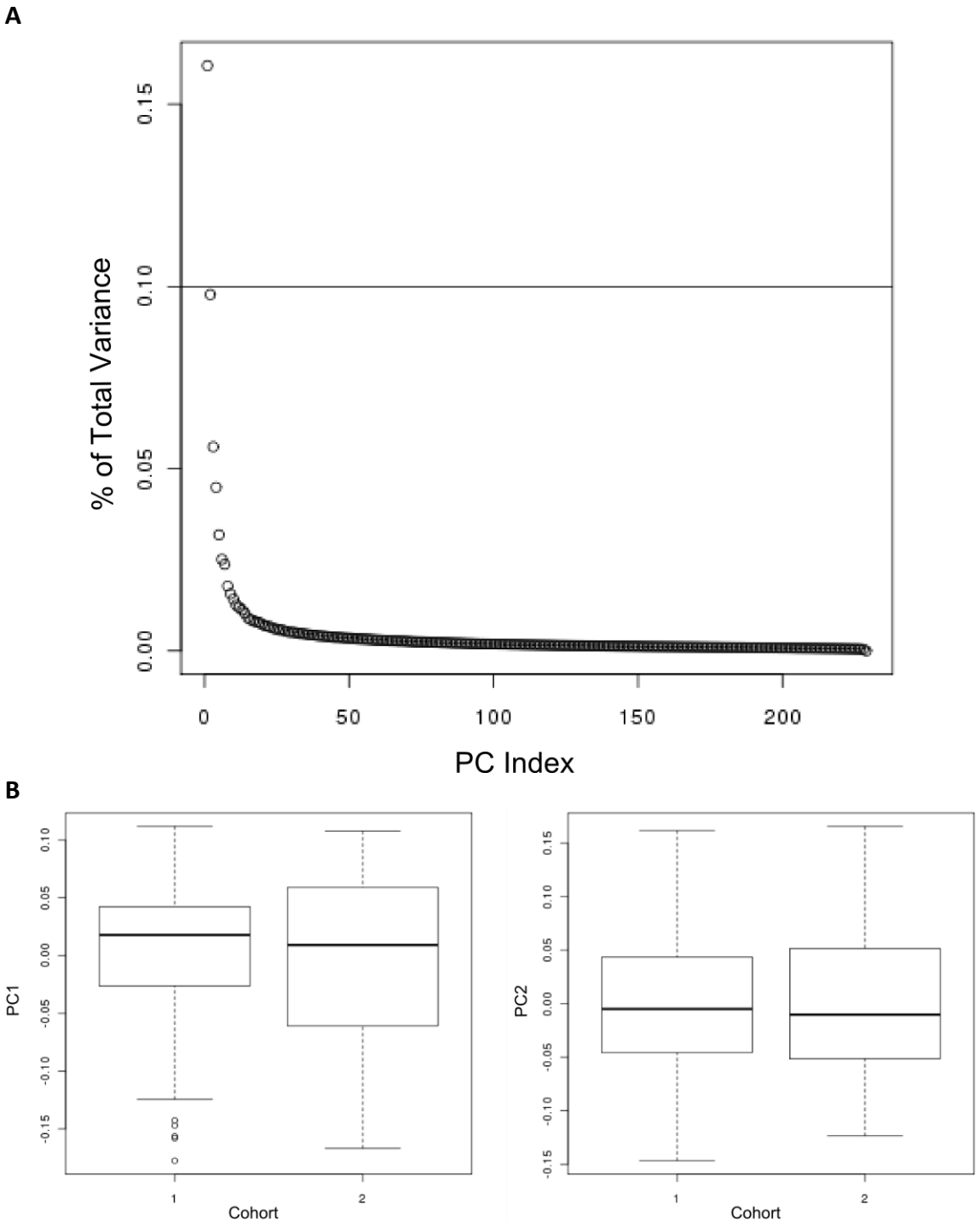
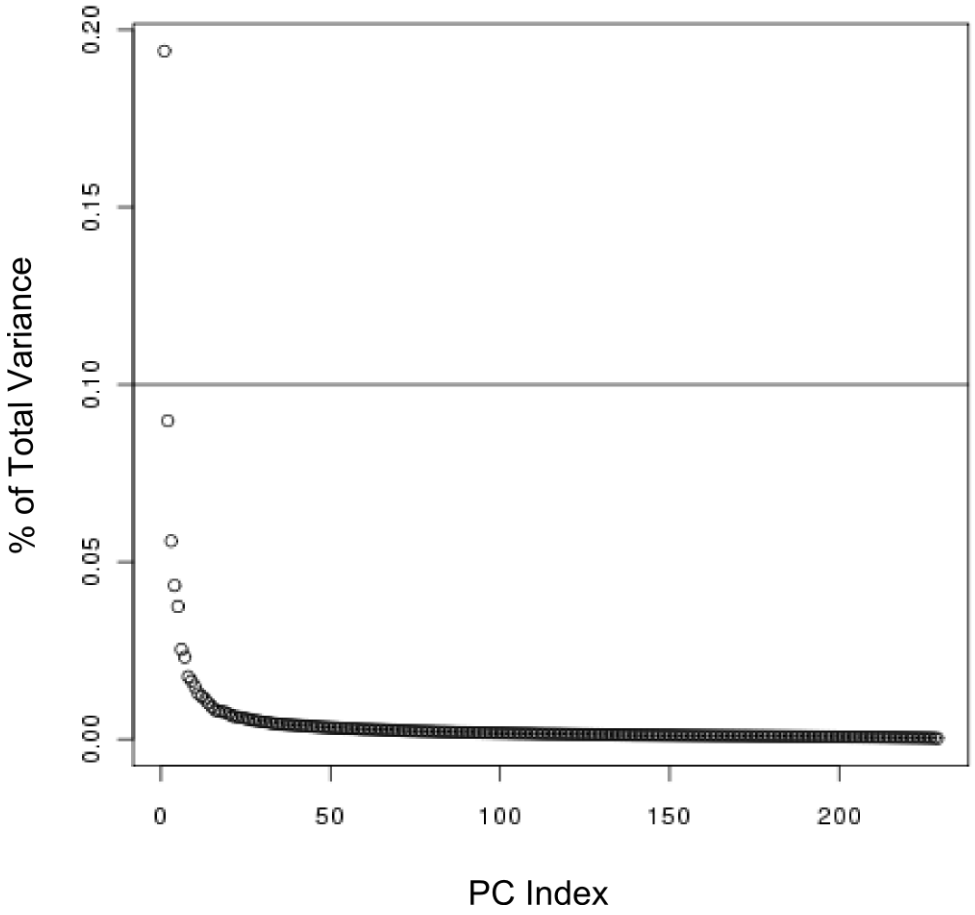


Figure A.3: PCA analysis on RMA normalised expression. (A) Fraction of overall variance contributed by the top principal components (PC). **(B)** Boxplots show the expression levels of the first and second PCs in cohort 1 and cohort 2 samples.

A



B

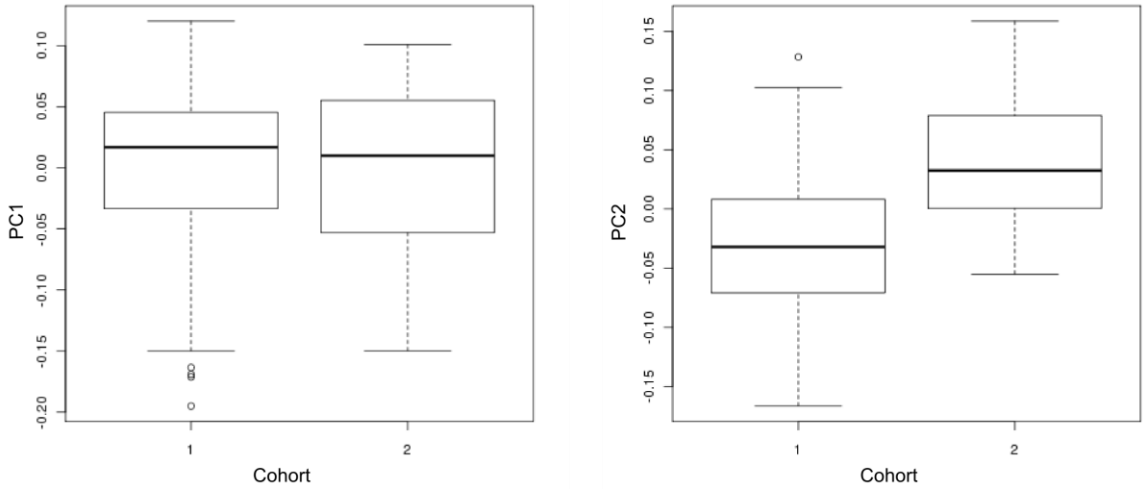


Figure A.4: PCA analysis on batch-corrected expression returned by fRMA. (A) Fraction of overall variance contributed by the top principal components (PC). **(B)** Boxplots show the expression levels of the first and second PCs in cohort 1 and cohort 2 samples.