Computational-Driven Understanding of the Regulatory Mechanisms of the Breast Cancer Transcriptome and Its Implications for Drug Treatment

By

Shujun Huang

A thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

In partial fulfillment of the requirements of the degree of

DOCTOR OF PHILOSOPHY

College of Pharmacy

University of Manitoba

Winnipeg, Manitoba

Abstract

Breast cancer (BC) is the most commonly diagnosed cancer and the major cause of cancer mortality among women worldwide. As a heterogenous disease, BC can be divided into four major molecular subtypes: luminal A, luminal B, HER2-enriched, and triple negative breast cancer (TNBC). The TNBC subtype shows the shortest survival time among the four groups and lacks effective targeted therapeutic strategies. Different BC subtypes display different transcriptome profiles. However, the dysregulated pathways and transcriptional regulators underlying the gene expression profiles of different BC subtypes have yet to be fully elucidated. Current research is investigating the dysregulated genes in BC with the aim to identify potential gene targets for BC while overlooking the fact these genes are often part of a pathway. Moreover, among the multi-omics data, gene expression profiles have been shown to be the most informative data for developing anti-cancer drug response prediction models *in silico*. But these models typically were developed with individual genes. Therefore, this thesis aimed to explore the breast cancer transcriptome with a focus on the TNBC subtype to address three major questions: 1) exploring the regulatory mechanisms driving the unique expression pattern of different BC subtypes; 2) identifying compounds that could affect the expression pattern of the top dysregulated pathways in BC; and 3) developing a drug response prediction model by using BC pathway activity profiles inferred from the transcriptome profiles.

To address the first question, we collected the multi-omics data of BC samples from The Cancer Genome Atlas (TCGA) dataset, including gene expression, DNA methylation, copy number variation (CNV) and microRNA (miRNA) profiles, the transcription factor (TF)-binding data from TRRUST v2.0, and the miRNA-binding data from starBase v3.0. Using these data, the Lasso regression-based integrative analysis identified 25, 20, 15 and 24 key regulators for luminal

A, luminal B, HER2-enriched and TNBC subtypes, respectively. A further look at the TNBC regulators found that many of them are regulating the FOXM1 (i.e., PID_FOXM1_PATHWAY) and PPARA (i.e., BIOCARTA_PPARA_PATHWAY) pathways.

To address the second question, we focused on the FOXM1 and PPARA pathways. Using the Connectivity Map (CMAP) database, which provides drug-induced gene expression changes in MCF7 cell lines, we investigated how different compounds change the activity and expression pattern of the two pathways. Nineteen drugs (such as 5109870, MG-132, MG-262, celastrol, resveratrol, and cephaeline) were identified to decrease the FOXM1 pathway activity scores and reverse the FOXM1 pathway expression pattern while 13 drugs (such as cephaeline, pararosaniline, cycloheximide, monensin, wortmannin, and raloxifene) were identified to increase the PPARA pathway activity scores and reverse the PPARA pathway expression pattern. It may be of interest to validate these compounds experimentally.

To address the third question, we collected the baseline gene expression profiles of 49 BC cell lines along with IC_{50} values of these cell lines to 220 drugs from the Genomics of Drug Sensitivity in Cancer (GDSC) dataset. Using these data, we developed a multiple-layer cell line-drug response network (ML-CDN2) by integrating a one-layer cell line similarity network based on the pathway activity profiles and a three-layer drug similarity network based on three types of drug information. ML-CDN2 demonstrated good prediction performance, with the Pearson correlation coefficient between the observed and predicted IC_{50} values for all cell line-drug pairs of 0.873. Moreover, the ML-CDN2 model could be used to predict the drug response for new BC cell line samples or new BC patient-derived samples.

This thesis demonstrated the transcriptional regulators underlying the transcriptome profiles in different BC subtypes. Moreover, this thesis demonstrated the implications of the BC

transcriptome in drug treatment by identifying the drugs to modulate the two dysregulated pathways in BC and developing the anti-cancer drug response prediction model for BC by incorporating the transcriptome profiles.

Acknowledgements

Here I would like to thank many people who have helped me in various way throughout the four years of my PhD study. First, I would like to thank my advisor Dr. Ted Laskowski for taking me as his student after my past advisor left. His guidance, support, encouragement, and patience throughout my research and the writing of this thesis has been invaluable. I would also like to thank Dr. Pingzhao Hu for supervising me on bioinformatics. His guidance, insights, and expertise has helped me to finish my research and the writing of this thesis. Without the guidance and support from Drs. Ted Lakowski and Pingzhao Hu, this project would not have been continued and this thesis would not have been finished.

I would also like to thank my committee members Drs. Leigh Murphy, Etienne Leygue, and Jim Davie for their advice and instruction on the cancer biology of this research. In every committee meeting, the thesis proposal, and the candidacy examination, they asked the insightful questions to get me to think and evaluate the scientific meaning of my research. Their insights and questions have helped steer both my research and my career directions.

I would like to thank my past advisor Dr. Wayne Xu for bringing the project idea together and helping me to overcome many problems when I was a beginner in the bioinformatics field. I appreciate his knowledge, time, patience, and support during my PhD study in the passing years. Without his guidance and dedication, this project would not have been started.

I'm grateful to all the members of the Lakowski lab (Wenxia Luo and Oluwafolahanmi Ige) and the Hu lab (Mohaiminul Islam, Qian Liu, Yong Won Jin, Shuo Jia, Nikta Feizi, and Rayhan Shikder) for their thought-provoking discussions and comradery. I wish to say particular thanks to Shuo Jia who has volunteered his time to double check my code. I also thank my past

lab member Shavira Narrandes for her perseverance in helping to critique my writing and for her company in the past few years in the lab.

I would like to thank all my family and friends for their support and encouragement as I've pursued this research. To my beloved husband, Sébastien FOCK CHOW THO, I would like to express my sincere thanks for what he has brought to my life. He has always been here by my side and shared every triumph and sacrifice with me throughout this journey.

Dedication

I wish to dedicate this thesis to the patients, who have consented to give out their tissue samples and clinical information.

Publications and Contributions

Portions of Chapter 1 have been published: Huang Shujun, Murphy Leigh, Xu Wayne. Genes and functions from breast cancer signatures. BMC cancer. 2018 Dec;18(1):473. I conducted the data analysis and wrote the manuscript with guidance from Xu W. Murphy L helped write and revise the manuscript.

Chapter 2 has been published: Huang Shujun, Xu Wayne, Hu Pingzhao, Lakowski Ted. Integrative Analysis Reveals Subtype-Specific Regulatory Determinants in Triple Negative Breast Cancer. Cancers. 2019 Apr;11(4):507. I conducted the data analysis and wrote the manuscript with guidance from Hu P and Lakowski TM. Xu W helped the study design.

Chapter 3 has been submitted for publication: Huang Shujun, Xu Wayne, Hu Pingzhao, Lakowski Ted. Identification of Compounds to Modulate the FOXM1 and PPARA Pathway Activities in Breast Cancer. I conducted the data analysis and wrote the manuscript with guidance from Hu P and Lakowski TM. Xu W helped the study design.

Chapter 4 has been submitted for publication: Huang Shujun, Hu Pingzhao, Lakowski Ted. Predicting Drug Response of Breast Cancer Using Multiple-Layer Cell Line-Drug Network Model. I conducted the data analysis and wrote the manuscript with guidance from Hu P and Lakowski TM.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

3'-UTR	Three prime untranslated region
BC	Breast cancer
CCLE	Cancer Cell Line Encyclopedia
CMAP	Connectivity Map
CNV	Copy number variation
CRS	Combined risk score
CSN	Cell line similarity network
CTRP	Cancer Therapeutics Response Portal
DART	Relevance network Topology
DEG	Differentially expressed gene
DFI	Disease-free interval
DM	DNA methylation
DSN	Drug similarity network
DSS	Disease-specific survival
ENCODE	Encyclopedia of DNA Elements
ER	Estrogen receptor
FDR	False discovery rate
FOXM1	Forkhead box protein M1
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
GDSC	Genomics of Drug Sensitivity in Cancer
GEO	Gene Expression Omnibus
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
HER2	Human epidermal growth factor receptor 2
IHC	Immunohistochemistry
KEGG	Kyoto Encyclopedia of Genes and Genomes
KM	Kaplan-Meier
Lasso	Least absolute shrinkage and selection operator
logFC	Log2 fold change

METABRIC	Molecular Taxonomy of Breast Cancer International Consortium
miRNA	MicroRNA
ML-CDN	Multiple-layer cell line-drug response network
mRNA	Messenger RNA
MSigDB	Molecular Signatures Database
OS	Overall survival
PFI	Progression-free interval
PID	Pathway Interaction Database
PPAR	Peroxisome proliferator activator receptor
PPARA	Peroxisome proliferator-activated receptor alpha
PR	Progesterone receptor
RMA	Robust multi-array analysis
RMSE	Root mean squared error
RNA-Seq	RNA-sequencing
RPM	Reads per million miRNAs mapped
RSEM	RNA-Seq by Expectation-Maximization
SMILES	Simplified molecular-input line-entry system
SNF	Similarity Network Fusion
SNP	Single Nucleotide Polymorphism
TCGA	The Cancer Genome Atlas
TF	Transcription factor
TNBC	Triple negative breast cancers
TPM	Transcripts Per Million
UDR	Up-down gene mean expression ratio

Chapter 1: Introduction

In this chapter, parts of section 1.5.1 were previously published. The text incorporated represents the article below that was entirely written by myself:

Huang Shujun, Murphy Leigh, Xu Wayne. Genes and functions from breast cancer signatures. BMC cancer. 2018 Dec;18(1):473.

Breast cancer is the most commonly diagnosed cancer and the primary cause of cancer mortality, accounting for nearly 24.2% of all cancer cases and 15% of cancer-related deaths in women worldwide [1]. According to Canadian Cancer Statistics 2020, BC is the most common cancer and the second leading cause of death from cancer among Canadian women, accounting for 25% of all new cancer cases and 13% of cancer-related deaths [2]. Though typically referred to as a single disease, BC is a heterogeneous disease in histology, progression, therapeutic response and clinical outcome [3]. With the advent of global gene expression (transcriptome) profiling technologies, researchers have been able to divide BC into four major molecular subtypes: luminal A, luminal B, HER2-enriched, and basal-like, with each subtype having a unique gene expression pattern [4]. However, there is still a lack of understanding of the regulatory mechanisms behind the transcriptome in each BC subtype. Knowing more about this would increase understanding of BC heterogeneity, uncover new BC subtype-specific therapeutic targets, and develop better predictors of BC prognosis and treatment response. Thus, this thesis sought to identify differentially expressed pathways in each BC subtype in comparison with normal breast tissue and investigate the regulators that lead to different gene expression patterns in different BC subtypes. Furthermore, we investigated drugs that could modulate the activity of the top differentially

expressed pathways and constructed a network-based model using the pathway activity in terms of gene expression level to predict the anticancer drug response of BCs.

Chapter 1 of this thesis aims at reviewing the literature which has guided my research. Section 1.1 introduces cancer transcriptome, and section 1.2 reviews inter-tumor and intra-tumor heterogeneity in BC as well as the different subtyping schemes for BC. Approaches to profiling the cancer transcriptome are reviewed in section 1.3. The factors that affect gene expression and the methods that model the regulatory mechanisms of gene expression in human cancer with the focus on BC are reviewed in section 1.4. Roles of transcriptome in human cancer precision medicine are summarized in section 1.5, especially focusing on BC. Data sets used through this thesis are introduced in section 1.6. Finally, the hypotheses and objectives of each chapter in this thesis were described in section 1.7.

1.1 Cancer transcriptome and transcriptomics

The human genome is comprised of 3 billion DNA base pairs, which encode approximately 20,000 genes. These genes account for 1~2% of the entire genome and are known as the coding regions. The rest of the whole genome (98~99%) does not encode genes and is referred to as non-coding. However, these non-coding regions influence splicing, expression, chromatin, and genomic organization, among other things. The genome determines what a cell has the potential to do. Genomics is the study of the whole genome of organisms. In the context of cancer, abnormalities that amend the genome (such as by altering DNA sequence, copy number, and chromosome structure) progressively occur and accumulate. For example, germline mutations in *BRCA1*, *BRCA2*, *TP53*, and *CHK2* genes can contribute to BC initiation [5]. Therefore, studying the cancer genome allows researchers to understand the molecular mechanisms that drive cancer.

During transcription, a gene is expressed ultimately resulting in an RNA copy of the relevant coding regions of the gene. There are several types of RNA. The most familiar are mRNAs which encode proteins. But there are also non-coding RNAs such as ribosomal RNAs or miRNAs. All RNA (or sometimes just mRNA) transcripts in a cell or a population of cells are referred to as the transcriptome. In contrast to the genome, which is relatively static across all cells of an organism, the levels of mRNA expression vary, depending on the cell type diversities, cellular growth stages and cell regulatory mechanisms [6]. Thus, transcriptomics, which is the study of the transcriptome of a cell or tissue sample, allows us to understand all the mRNA molecules and all the products of gene transcription in that sample and thus describe the underlying phenotypes in great detail [6].

In gene expression, some RNAs (the mRNAs) are finally translated into proteins. All proteins in a given cell, tissue or biological sample, at a specific developmental or cellular phase, are known as proteome. The proteome is in part responsible for the phenotypes and biological processes of a cell, such as metabolic functions, signalling transduction, regulation of gene transcription. Similar to genomics and transcriptomics, proteomics is the study of the proteome.

Cancer is a complex disease with many causes. Arguably one of the most important cancer mechanisms is aberrant gene expression (transcriptome), which is often the result of gene mutations and epigenetic alterations [7,8]. As the intermediate step between DNAs and proteins, transcriptomics enables researchers to link the cellular phenotypes (proteome) and their molecular mechanisms (genome). When it comes to cancer, studying the transcriptome provides a link between the genetic causes of cancer and their phenotypic consequences. Nowadays, using high-throughput techniques, researchers have been able to profile the transcriptome of individual cancer cells or thousands of cancer samples [9,10]. These data have helped the exploration of the cancer

complexity and heterogeneity and discovery of new biomarkers, signatures or therapeutic strategies for cancer. In this thesis, we focused on the understanding of the regulatory mechanisms and the implications for drug treatment of the BC transcriptome. The following sections review the clinical and cellular properties of different BC subtypes that have been identified according to the transcriptome.

1.2 Breast cancer

1.2.1 Canadian breast cancer statistics

BC is the most frequently diagnosed cancer in Canadian women according to Canadian Cancer Statistics 2020 [2]. In Canada, the lifetime risk of developing BC is about one in eight for women. In 2020, an estimated 27,400 Canadian women will be diagnosed with BC, representing approximately 25% of all newly diagnosed cancer cases in Canadian women in 2020 [2]. Even so, the BC incidence rates in women in Canada have shown an overall small fluctuating decrease since 1991 [2]. Due to the widespread use of early detection programmes and improved treatment, the BC death rates in Canadian women have been decreasing since 1986. The 5-year net survival (estimates for 2012 to 2014) for BC in women in Canada is 88% [2]. In other words, among Canadian women diagnosed with BC, about 88% will survive five or more years after their diagnosis. However, BC still remains the second-most common cause of cancer-related death in Canadian women, just behind lung cancer [2]. In 2020, it is estimated that 5,100 Canadian women will die from BC, accounting for approximately 13% of all cancer deaths in Canadian women in 2020 [2]. BC can also occur in men, but it is very rare. For example, in 2020, it is estimated that 240 men will diagnosed with BC and 55 would die of it in Canada [2]. The 5-year net survival (estimates for 2012 to 2014) for male BC in Canada is 80% [2].

1.2.2 Inter-tumor heterogeneity in breast cancer

BC is typically referred to as a single disease because it originates from the cells of the mammary gland. However, BC is a complex disease with a high degree of inter- and intra-tumor heterogeneity. Inter-tumor heterogeneity refers to differences among tumors from different individuals. In contrast, intra-tumor heterogeneity denotes genetic and phenotypic variability among cancer cells within a single tumor [11]. The heterogeneity in BC has a profound impact on disease progression and therapeutic response, thus making it one of the most important and clinically relevant areas of BC research [12].

Inter-tumor heterogeneity has been recognized through different clinical and histopathological characteristics, biomarker profiles, or the more modern gene expression patterns, among other differences [13,14]. These differences have served as the basis for different BC classification schemes, which are known as BC subtypes [13]. Pathologists have long noted the histological diversity of breast tumors and the morphological heterogeneity is the basis for the histological classification of BC [14]. From a histological point of view, BC can arise from the epithelial cells that line the ducts (ductal carcinoma) or the lobules (lobular carcinoma) [15]. Ductal carcinoma can be ductal carcinoma *in situ* if the cancer cells still stay in the epithelial component of the ducts or invasive ductal carcinoma if the cancer cells have invaded the surrounding tissues [15]. Similarly, lobular carcinoma can also be lobular carcinoma *in situ* which means the tumor is limited to the epithelium of the lobules or invasive lobular carcinoma which indicates the tumor has grown into the stroma [15]. Invasive ductal carcinoma is the most common histological type of BC, accounting for approximately 70~80% of all invasive breast carcinomas [15].

Inter-tumor heterogeneity of BC is arguably best represented by the pathological staging of breast carcinoma, which has a major impact in clinical decision making today [14]. Stage refers

to the extent of a cancer, such as how large the tumor is, what part of the breast has cancer, and if the tumor has spread [16]. Stage is a strong prognostic factor [17]. Staging BC based on physical examination and imaging findings can help physicians to determine the BC prognosis and determine the right treatment options [17]. The most commonly used BC staging system is the Tumor, Node, and Metastasis (TNM) system published by the American Joint Committee on Cancer and the Union for International Cancer Control [18]. The TNM system incorporates the information about tumor size, regional lymph node status and distant metastases to classify a tumor into 5 stages - stage 0 followed by stages 1 to 4 [18]. Generally, the higher the stage is, the more the cancer has spread. Survival varies with each stage of BC. In general, the earlier stage BC is diagnosed and treated, the better the outcome.

Inter-tumor heterogeneity of BC can also be observed in the histological grading of breast carcinomas [14]. There are different grading systems available for BC. The most commonly used is the Nottingham grading system [19], which is a modification of the original Bloom-Richardson breast cancer grading system put forward over 60 years ago [20]. The Nottingham grading system takes three tumor characteristics into consideration to assess the grade of breast tumors: glandular/tubular differentiation (the proportion of cancer cells that are in gland formation), nuclear pleomorphism (the variation of nuclear size and shape between the tumor cells) and mitotic count (how much the tumor cells are dividing or proliferating) [19]. Each of these features is scored from 1-3, and then the individual scores are added together to determine the final grade: grade 1 (low), grade 2 (intermediate) or grade 3 (high) [19]. The grade of breast carcinoma is a robust predictor of survival and reflects the aggressive potential of the tumor, with low-grade cancers generally tending to be less aggressive than high-grade cancers [21]. Determining the grade is

therefore important, because clinicians use this information to guide treatment options for BC patients.

BC inter-tumor heterogeneity can also be characterized by the expression status of well-established molecular biomarkers [14]. Differences in the expression of these biomarkers is known as biomarker heterogeneity [14]. Three biomarkers (estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2)) are recognized by international guidelines as prognostic and predictive factors indispensable for invasive BC therapy decision making [22]. In the clinic, BC subtypes are usually determined by examining protein expression of ER and PR, as well as protein expression of HER2 and/or gene amplification of *HER2*. At diagnosis, the solid tumor tissue samples taken during the biopsy of all invasive breast carcinomas will be routinely tested by immunohistochemistry (IHC) to assess the status of these biomarkers to help treatment decision-making [23]. According to the recommendations by American Society of Clinical Oncology/College of American Pathologist, if any immunohistochemical nuclear staining (irrespective of the signal intensity) is observed in more than 1% of invasive tumor cells, it is considered as hormone receptor-positive (ER-positive and/or PR-positive, i.e., ER+ and/or PR+) [24]. About 80% of breast tumors are ER-positive (ER+) [25] and about 60~70% are PR-positive (PR+) [26]. While approximately 70~80% of breast tumors co-express ER and PR (ER+/PR+), 13% are ER-positive and PR-negative (ER+/PR-) and only a small number (2%) are ER-negative and PR-positive (ER-/PR+) [14]. A tumor is considered as hormone receptor-negative (ER-negative and PR-negative, i.e., ER-/PR-) if less than 1% of cells are stained positively by antibodies against ER and PR [24]. The utility of such IHC ER and PR stratification is that it can be used to make some generalizations. For example, patients with hormone receptor-positive tumors have a more favorable prognosis than those with hormone receptor-negative tumors. In addition, patients with

hormone receptor-positive tumors will likely benefit from hormonal treatment that targets the corresponding hormone receptors, such as tamoxifen and others. The response to hormonal treatment is variable, with the best response of about 60% in ER+/PR+ tumors and lower rates in ER+/PR− and ER−/PR+ tumors [27].

HER2-positive (HER2+) status is characterized by overexpression of HER2 protein assessed by IHC or by amplification of the *HER2* gene detected by an *in situ* hybridization method [22]. According to American Society of Clinical Oncology/College of American Pathologist, the HER2 status assessed by IHC can be positive, negative or equivocal [28]. In many pathology labs, HER2 status is initially assessed by IHC due to its availability and cost-effectiveness. IHC measures HER2 protein expression and the IHC results are scored on a scale from 0 to 3: cases with a score of 0 or 1+ is negative (HER2-); cases with a score of 2+ is equivocal; cases with a score of 3+ is positive (HER2+) [28]. The IHC equivocal cases are subsequently evaluated by an *in situ* hybridization method to detect *HER2* gene amplification to confirm HER2 status [28]. HER2+ cases account for 15~20% of breast tumors [14]. HER2+ tumors have the worst prognosis of all types of invasive BC, but they do respond to anti-HER2 therapy (e.g., trastuzumab and lapatinib) [29]. Breast tumors that do not express ER, PR, and HER2 are called triple-negative breast cancer (TNBC), and are extremely heterogeneous from a histological, genetic, prognostic and treatment response point of view [22,30]. Numerous other biomarkers have been and are being investigated in BC for their potential diagnostic, prognostic, and therapeutic implications. One such marker is Ki-67, which is widely used to assess cell proliferation and predict sensitivty to chemotherapy [22]. Studies have shown that Ki-67 is a prognostic biomarker, with higher Ki-67 expression reflecting higher cancer cell proliferation and diminished prognosis [22]. The Ki-67 expression is assessed using IHC [22]. However, the assessment of Ki-67 expression in BC is

neither standardized nor generally recommended [29]. Thus the cut-off value between high and low expression of Ki-67 varies between laboratories [31]. In the 2013 St Gallen International Breast Cancer Conference, the majority of the Panel voted that a threshold of ≥20% stained nuclei was indicative of high Ki-67 expression [31]. Generally speaking, values <10% stained nuclei are considered low Ki-67 expression and >30% are considered high Ki-67 expression [22].

In the past two decades, inter-tumor heterogeneity of BC has been studied in depth at the molecular level and treatment concepts now take such heterogeneity into consideration. There exist many molecular variations that lead to breast carcinogenesis, and several classification schemes have evolved to categorize breast tumours. The most well-known is the intrinsic classification based on gene expression analysis, which distinguishes four major molecular subtypes of BC with prognostic and therapeutic implications: luminal A, luminal B, HER2-enriched, and basal-like [32–34]. We will discuss the intrinsic and other molecular subtypes of breast tumors with more detail in section 1.2.4, 1.2.5 and 1.2.6.

Several concepts explaining sources of inter-tumor heterogeneity in BC have been proposed, including the differentiation state of the initiating cell (i.e., cell-of-origin of cancer), genetic evolution, cancer cell plasticity, and tumor microenvironment [12]. The cell-of-origin concept states that tumor phenotype depends on the differentiation state of the cell-of-origin and the tumor-initiating genetic and/or epigenetic event(s) [12,13,35]. Accordingly, the luminal and HER2+ tumors may originate from the luminal lineage, whereas basal-like cancers may arise from less differentiated stem cell-like cells [13]. However, the luminal lineage may also serve as the cell-of-origin of basal-like cancers if genetic and/or epigenetic alterations transform cellular phenotypes [13]. The concept of genetic evolution attributes tumor heterogeneity to genetic evolution and proposes that tumorigenesis is a multiple-step evolutionary process during which

genetic and/or epigenetic alterations accumulate in tumor cells and the fittest cells survive [12,14]. Cell plasticity, which enables dynamic and bidirectional cell conversions between cancer stem cells and non-cancer stem cells within a tumor, is also thought to be a source of BC heterogeneity [12]. Tumor microenvironment, which contains fibroblasts, blood vessels, and immune cells, can contribute to tumorigenesis and metastasis and thus plays an important role in tumor heterogeneity[12,36]. For example, physiological interactions between tumor cells and their microenvironment are disrupted since tumor cells are not sensitive to growth inhibitors from their microenvironment [37]. Though these different mechanisms have a crucial impact on the development of BC heterogeneity, no single mechanism can explain all aspects of BC heterogeneity. Thus, several of these mechanisms, including cells-of-origin, genetic evolution during tumor progression and treatment, cancer cell plasticity, and interactions between tumor cells and their microenvironment, are likely to work together to contribute to BC heterogeneity [12].

1.2.3 Intra-tumor heterogeneity in breast cancer

Other than inter-tumor differences, significant heterogeneity may exist among cancer cells within an individual tumor [11,35]. This intra-tumor heterogeneity can be manifested on many levels including histology, as well as genetic and epigenetic aberrations [11,23]. Intra-tumor heterogeneity can be subdivided into spatial and temporal heterogeneity [11]. Spatial heterogeneity represents regional differences within a tumor [11]. Different regions within a given tumor are frequently observed to display different histological characteristics in BC [11]. Also, histologically distinct regions of a tumor are generally thought to harbor different repertoires of genetic and/or epigenetic aberrations, which might contribute to the spatial intra-tumor histological heterogeneity [11]. Given the spatial heterogeneity within a given tumor, a single biopsy might not give an

adequate reflection of the morphological composition of the tumor as a whole and bulk DNA sequencing of the single biopsy is not able to assess the intra-tumor genetic and epigenetic heterogeneity [11]. Temporal heterogeneity reflects genetic/epigenetic and histological changes in the tumor with respect to time [11]. Temporal heterogeneity has been documented between the initial tumor and local or distant recurrences [11], during the progression from *in situ* to a more invasive disease [38], and during the course of therapy [39]. For example, genomic analysis of a primary basal-like breast tumor and its brain metastasis demonstrated 20 shared mutations [40]. However, two *de novo* mutations and a large deletion were observed in the genome of the metastasis but not present in the primary tumor [40]. Tumors evolve over the course of treatment, therefore, it might be misleading to make therapeutic decisions for treatment of recurrent tumors based on the initial scoring of the phenotype at diagnosis.

Intra-tumor heterogeneity also has been observed in the expression of BC biomarkers (ER, PR, HER2), where tumors exhibit differential expression of these biomarkers throughout a tumor [23]. Such variability in biomarker expression within a tumor can cause interpretation problems and discordant results in small biopsies [23]. For example, in a study of examining multiple regions of 96 invasive breast tumors with *HER2* gene amplification detected in whole tissue sections, spatial intra-tumoral heterogeneity of *HER2* gene amplification was observed in 17 (18%) of the 96 cases [41]. Patients with heterogeneous regions of *HER2* gene amplification were found to show shorter disease-free survival than the other cases [41]. Intra-tumoral heterogeneity of biomarker expression assessments has also been recognized between primary tumors and their metastasis [23]. For example, in a study of assessing the discrepancies in biomarker characteristics between primary breast tumors and their metastatic tissues, the discordance rate was 16% for ER IHC staining, 40% for PR IHC staining, and 10% for *HER2* gene amplification [42]. Such

discrepancies between primary BC and metastasis were found to change the therapy for 14 % of patients [42].

Two theories have been proposed to explain the establishment and maintenance of intra-tumor heterogeneity: the clonal evolution/selection model [43] and the cancer stem cell model [44]. Both concepts consider that a single cell which has obtained multiple molecular alterations and acquired indefinite proliferative potential can give rise to tumors [11]. Both concepts assume that the tumor microenvironment has an impact on the tumor evolution [11]. However, there are fundamental differences between the two models. The clonal evolution model suggests that any single cell can give rise to a sequential selection of many diverse subpopulations through continued accumulation of genetic and epigenetic aberrations [11]. Each subpopulation evolves independently and is influenced by its microenvironment, resulting in intra-tumor heterogeneity [11]. In the clonal evolution theory, tumor progression and resistance to therapy would be subject to selective pressure from the microenvironment and treatment [45]. The nature of this selective pressure can also change over time selecting different populations which might occur with treatment changes [45]. Moreover, this kind of heterogeneity includes non-cancerous cells which can support or hinder tumor growth [45]. BC intra-tumor heterogeneity through clonal evolution has been observed in HER2+ tumors, where different patterns of *HER2* amplification have been observed in distinct regions within a single tumor [46]. In the cancer stem cell model, a tumor originates from a cancer stem cell which has self-renewal capacity [11]. Cancer stem cells have capacity to differentiate into non-cancer stem cells and dedifferentiate under certain conditions, which is called plasticity [11]. The cancer stem cell model considers that mutations as well as plasticity of cancer stem cells give rise to different subpopulations and account for intra-tumor

heterogeneity [11]. In the cancer stem cell theory, the cancer cells are hierarchically organized; the cells with high capacity of proliferation and self-renewal are placed in the highest order [11].

1.2.4 The intrinsic subtypes of breast cancers

During the last few decades, molecular classification of BC based on comprehensive gene expression profiling has been extensively explored. The most established is the so-called intrinsic subtype classification, which was first described in 2000 by Perou *et al.* [47] and was later refined by Sørlie *et al.* [33,34]. In 2000, Perou *et al.* identified four different molecular subtypes of BC based on their gene expression patterns [47]. The expression of 8,102 human genes was measured by cDNA arrays in 65 normal or malignant human breast tissue samples from 42 different patients. From these genes, those with significantly greater variation in expression between tumors from different patients than between two paired samples from the same patient were selected, resulting in a total of 496 genes. This subset of genes were termed the intrinsic gene subset as they presumably reflected the phenotypes of individual tumors. Hierarchical clustering based on the expression patterns of these intrinsic genes across all the tumor samples revealed two main clusters (largely separating the samples into clinically ER+ and ER-) with a total of four subclusters. The four subclusters, named intrinsic subtypes, are luminal-like, HER2-enriched, basal-like, and normal-like breast cancers. The luminal-like subtype, enriched for ER+ tumors, was characterized by the relatively high expression of many genes normally expressed by the luminal epithelial cells (hence called luminal-like). The luminal-like characteristics of these tumors were further confirmed by IHC using antibodies against the luminal keratins 8/18. The HER2-enriched subtype was characterized by the overexpression of a subset of genes located in the *HER2* amplicon at chromosome 17q12 including *HER2* and *GRB7* genes. Tumors in this subtype expressed luminal epithelial genes at low levels. The basal-like subtype, enriched for ER- tumors, showed high levels

of expression of many genes characteristic of breast basal epithelial cells. Furthermore, the basal-like tumors can be stained immunohistochemically with antibodies against either basal keratins 5/6 or 17 or both, which corroborated the basal-like characteristics of these tumours. Similar to the HER2-enriched tumors, the basal-like tumors showed low expression of luminal epithelial genes. The normal-like subtype contained tumors that were clustered in the same group as the normal breast samples. The intrinsic gene expression pattern in these normal-like tumors was characterized by high expression of genes characteristic of adipose and basal epithelial cells, and the low expression of genes characteristic of luminal epithelial cells. Due to the limited number (n = 42) of patients in the initial study [47], Sørlie *et al.* reported two follow-up studies including an increased number of patients [33,34]. The first follow-up study included 85 normal or tumor human breast tissue samples from 84 patients and was published in year 2001 [33]. This study refined the previous classification by subdividing the luminal-like tumors into three subgroups (luminal A, B and C) and preserving the basal-like, HER2-enriched and normal-like subtypes, resulting in six intrinsic subtypes. Among the three luminal subtypes, the luminal A tumors demonstrated the highest expression of the luminal epithelial genes (such as *ERα*, GATA binding protein 3, X-box binding protein 1, trefoil factor 3, hepatocyte nuclear factor 3, and estrogen-regulated *LIV-1*). Both the luminal B and C tumors showed low to moderate expression of the luminal epithelial genes. A novel subset of genes with unknown function were highly expressed in the luminal C subtype but not in the other two luminal subtypes, which distinguished the luminal C from the luminal A and B tumors. The percentage of *TP53*-mutated tumors varied among the six intrinsic subtypes. For example, only 13% of the luminal A tumors had mutations in *TP53*, whereas 71% of the HER2-enriched and 82% of the basal-like tumors had *TP53* mutation. Moreover, the six subtypes showed difference in overall survival, with the basal-like and the

HER2-enriched subtypes associated with the shortest survival times. The second follow-up paper from Sørlie *et al.* was published in 2003 [34]. In this study, hierarchical clustering analysis of a total of 115 malignant breast tumors based on their expression patterns of 534 intrinsic genes revealed the five definitive intrinsic subtypes: luminal A, luminal B, HER2-enriched, basal-like, and normal-like. It is noteworthy that the luminal C subtype was not included in the final definition of the intrinsic subtypes.

Since over 500 genes were used in the final definition of the five intrinsic subtypes [34], it posed challenges to the translation of this intrinsic subtype classification into a clinical assay. For this reason, more studies were performed to successively reduce the number of intrinsic genes to 306 [48] and finally to the minimum 50 (PAM50) for defining BC intrinsic subtypes [4]. Similar to the studies of Sørlie *et al.* [34] and Hu *et al.* [48], the PAM50 study used the Single Sample Predictor (SSP) strategy to perform the intrinsic subtype assignment of single samples [4]. The SSP model uses hierarchical clustering to identify the main BC subtypes from intrinsic gene expression data and then build a nearest centroid classifier [4]. For a new single tumor sample, the intrinsic subtype assignment is based on similarities between the given case and the intrinsic subtype centroids [4]. The PAM50 gene set showed high agreement in classification of BC intrinsic subtypes with the larger intrinsic gene sets previously used [32–34,48]. Later, it was successfully commercialized and available in the clinic as the Prosigna® test, which is based on the NanoString nCounter platform (a multiplexed gene expression profiling technology) [49]. The Prosigna® test assigns each tumor sample to the luminal A, luminal B, HER2-enriched, or basal-like subtype [49]. The test includes the PAM50 gene signature in combination with clinical factors (e.g., tumor size) to calculate a numeric score for evaluating the risk of distant relapse for postmenopausal women with hormone receptor-positive, lymph node-negative (i.e., the cancer has

not spread to the lymph nodes) or lymph node-positive (the cancer has spread to the lymph nodes) early-stage BC, with the potential to be informative for assessing the indication of adjuvant chemotherapy [49].

The intrinsic subtype molecular classification of BC has improved the understanding of BC biology while presenting the possibility to refine the prediction of correct BC treatment regimens. Although the intrinsic subtypes are originally defined by gene-expression profiles, the IHC-based surrogate intrinsic subtypes are typically used in clinical practice due to the cost and technical complexities required for gene expression profiling assays [22]. The IHC-based surrogate intrinsic subtypes, determined by histological features as well as IHC measurements of routine biomarkers (ER, PR and HER2) and the proliferation marker Ki-67, are clinically valuable and imply distinct treatment approaches [50,51]. It is noteworthy although the IHC-based surrogate intrinsic subtypes overlap with the PAM50 intrinsic subtypes, some discrepancies exist [52].

Although both the luminal A and B subtypes are enriched for ER+ tumors and have increased expression of genes normally expressed by luminal cells, the luminal B subtype generally has a higher expression of genes involved in cell proliferation and mitosis [53] and shows a worse prognosis than the luminal A subtype [4]. The IHC-based surrogate subtype for luminal A (i.e., luminal A-like) tends to be ER+, PR+, HER2-, low grade, and low Ki-67 [22,51]. These luminal A-like tumors account for 60%~70% of all breast cancers and have a good prognosis [22]. The luminal B tumors can be either HER2- or HER2+ [22]. The IHC-based surrogate subtype for the luminal B/HER- tumors (i.e., luminal B-like) tends to be ER+ but with lower ER/PR expression than the luminal A-like subtype, and displays high grade (histological grade 3) as well as high Ki-67 [51]. The luminal B-like tumors account for 10%~20% of all breast tumors and have an

intermediate prognosis [22]. Clinically speaking, one of the most important questions when dealing with luminal cancers at the early stages is which patients need adjuvant chemotherapy and/or hormonal therapy [22]. The St Gallen International Breast Cancer Conference (2017) Expert Panel has acknowledged that the classification of luminal tumors into luminal A-like and B-like subtypes could be used to inform adjuvant treatment decisions [51]. Specifically, patients diagnosed with luminal A-like disease are likely to benefit from hormonal therapy alone, whereas patients with luminal B-like disease may need chemotherapy in addition to hormonal therapy [51]. The Expert Panel has also agreed that either grade (low/high) or Ki-67 (low/high) could be used to distinguish luminal A-like from luminal B-like tumors [51].

The HER2-enriched subtype is often dominated by high expression of *HER2* amplicon genes [34]. The IHC-based surrogate subtype for HER2-enriched tumors is likely to be ER-, PR-, HER2+, high grade, high Ki-67, and shows an aggressive clinical course [22]. However, the definition of the HER2-enriched subtype is still debated [53]. The tumors classified as HER2 enriched also vary in terms of ER IHC status, copy number aberrations (CNAs), and mutation profiles [54–56]. Furthermore, not all PAM50 HER2-enriched tumors are clinically HER2+ defined by a combination of HER2 protein overexpression and *HER2* gene amplification [57]. Hence, the intrinsic subtypes cannot be used as a predictor with respect to anti-HER2 therapy [53]. In contrast, not all clinically HER2+ tumors fall into the PAM50 HER2-enriched subtype [57]. HER2+ tumors may be included in the HER2-enriched, luminal A, or luminal B subtypes, depending on their IHC status of ER [54]. In the original PAM50 study from Parker *et al.*, HER2+ tumors were mainly found in the HER2-enriched subtype but were also found in the luminal A and B subtypes [4]. Similarly, in the TCGA study, the HER2-enriched subtype captured some but not all clinically HER2+ tumors defined by the American Society of Clinical Oncology/College

of American Pathologist guideline [57]. The TCGA study reported that there existed at least two types of clinically HER2+ tumours: HER2-enriched-mRNA-subtype/HER2+ versus luminal-mRNA-subtype/HER2+ [57]. In addition, in the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) study, which identified 10 integrated cluster subtypes (IntClust1-10) in BC using an integrated analysis of copy number and gene expression, HER2+ tumors dominated the IntClust5 subtype but were also found in other subtypes [58]. In a retrospective study of 1,392 breast tumors defined as HER2+ by IHC and fluorescence *in situ* hybridization, the majority of the tumors (72.1%) were classified as HER2-enriched by PAM50 [59]. The remaining HER2+ tumors were classified as other intrinsic subtypes: the luminal A (9.5%), luminal B (11.4%), and basal-like (7.0%) subtypes [59]. Therefore, the clinically defined HER2+ tumors do not represent a separate subtype but a heterogenous group. Ferrari *et al.* distinguished four different subtypes of clinically defined HER2+ breast tumors based on their gene expression data, with each subtype displaying different somatic mutations, copy number changes or structural variations [54]. Regions and genes of the *HER2* amplicon in HER2+ breast tumors have also been studied. For example, Staaf and colleagues reported that the smallest *HER2* amplicon region was 85.92 kilo base pairs in HER2+ tumors, including six genes (*TCAP*, *PNMT*, *PERLD1*, *HER2*, *C17orf37* and *GRB7*) [60]. Also, HER2+ tumors with higher *HER2* copy numbers were found to exhibit worse prognosis [60]. Sahlberg *et al.* reported that *STARD3*, *GRB7*, *PSMD3* and *PERLD1* together with *HER2* in the *HER2* amplicon were required for the growth and survival of HER2+ breast tumor cells, indicating that targeting these genes at the same time may help improve anti-HER2 therapy efficacy [61].

The basal-like subtype is characterized by expression of genes typical of basal epithelial cells [34]. The basal-like subtype accounts for approximately 15% of all BC cases [62] and is more

common in African Americans, especially young and premenopausal women [63]. Basal-like tumours show aggressive features such as high grade and high proliferation [64]. Patients diagnosed with basal-like tumors have a poor prognosis, and almost 40% of patients experience a relapse within 5 years post-diagnosis [64,65]. The basal-like subtype is generally used in the research setting. Immunohistochemically, basal-like tumors are typically ER-, PR- and HER2- and thus the triple-negative breast cancer (TNBC) phenotype currently is used as a reliable surrogate for the basal-like subtype in the clinical setting [22]. Unlike ER+ tumors and HER2+ tumors, TNBCs generally do not respond to otherwise highly effective therapies such as tamoxifen and aromatase inhibitors for targeting ER or trastuzumab for targeting HER2 [22]. It should be noted, however, there is an incomplete overlap of the intrinsic basal-like subtype and tumors immunohistochemically defined as TNBC. Prat *et al.* reported that 78.6% of TNBC cases had a basal-like gene expression profile, whereas 21.4% did not [66]. Conversely, 68.5% basal-like tumors were identified as TNBCs, whereas 31.5% were not [66].

The normal-like subtype is characterized by showing gene expression pattern similar to that found in normal breast tissue samples [34]. Normal-like breast tumors represent 5-10% of all breast cancers [67] and have an intermediate prognosis between luminal tumors and basal-like/HER2-enriched tumors [33]. There are few studies to characterize this subtype and their clinical implication remains undetermined. Currently, no IHC-based surrogate is available for the normal-like subtype. In fact, though the normal-like subtype is included in the definitive list of the five intrinsic subtypes [34], the PAM50-based Prosigna assay does not have a trained normal-like centroid as a control to assign a tumor to the normal-like subtype [49].

1.2.5 Other molecular subtyping schemes of breast cancers

The intrinsic subtyping scheme is undoubtedly the most well-established molecular classification of BC. As an extension to the intrinsic subtyping scheme, Herschkowitz *et al.* identified a new rare intrinsic subtype, referred to as claudin-low [68]. The claudin-low intrinsic subtype accounts for 7-14% of breast tumors [69]. Claudin-low tumors are characterized by low expression of the claudin genes, and low to absent expression of genes characteristic of the luminal and HER2-enriched subtypes [70]. Claudin-low tumors are enriched for epithelial-to-mesenchymal transition markers, immune response genes and cancer stem cell-like features [70]. Similar to the basal-like subtype, the majority of claudin-low tumors are cases clinically defined as ER-, PR-, and HER2- (i.e., triple negative) [69]. Interestingly, in Lehmann *et al.*'s study which identified six subgroups from TNBC tumors by analyzing gene expression profiles, the claudin-low intrinsic subtype was mostly composed of the mesenchymal and mesenchymal stem-like subgroups [71]. Claudin-low cancers have a poor prognosis and show some chemotherapy sensitivity that is intermediate between basal-like and luminal tumors [70]. Currently, there are no surrogate IHC biomarkers for the claudin-low intrinsic subtype and thus this subtype is not used in the clinical setting yet [22].

In parallel with the discovery of the intrinsic subtypes, Sotiriou and colleagues proposed a similar molecular subtyping scheme for breast cancers based on gene expression profiles [72]. The expression of approximately 8000 probes was measured by cDNA microarray chips for 99 human breast tumor samples, and 706 probes showing high variability across all tumors were selected [72]. By performing unsupervised clustering on the 706 probe elements, the 99 breast cancer specimens were stratified into two main clusters primarily associated with ER status (ER+/ER-) [72]. The predominantly ER+ cluster could be further segregated into three smaller subclusters

based on their luminal characteristics: luminal-like 1, luminal-like 2 and luminal-like 3 [72]. Similarly, the predominantly ER- cluster could also be segregated into three smaller subclasses based on their basal characteristics: Her-2/neu, basal-like 1 and basal-like 2 [72]. Thus, a total of six subtypes were finally proposed by Sotiriou *et al.* [72]. This subtyping scheme was refined using the Subtype Classification Model (SCM) [73–75], which was an alternative molecular classification approach to the SSP method used in the intrinsic subtype assignment studies [4,34,48]. The SSP method performs the classification of single samples into one of the molecular subtypes based on similarities between a given case and molecular subtype centroids [76]. Unlike the SSP method, the SCM model used a mixture of Gaussian distributions based on expression of either three modules (i.e., the ER, HER2 and aurora kinase A (AURKA) modules) or only three key genes (*ER*, *HER2*, and *AURKA*) [73–75]. Three SCM models including different number of genes have been published [73–75]. The first one was SCMOD1 which used 726 genes to represent the ER, HER2, and AURKA modules [73]. The second one was SCMOD2 which used 663 genes to represent the ER, HER2, and AURKA modules [74]. The final one was a simplified SCM (SCMGENE) containing only the *ER*, *HER2*, and *AURKA* genes [75]. Of note, this subtyping scheme has not been translated into a clinical assay.

Based on the gene expression data from 537 samples from the Cartes d'Identité des Tumeurs (CIT) program, Guedj *et al.* proposed a classification scheme called the CIT classification that identified six BC molecular subgroups using a semi-supervised analysis [77]. The six subgroups overlapped with the intrinsic subtypes since the CIT classification also included luminal A, luminal B, basal-like, and normal-like subtypes. But the CIT classification added two additional subtypes (i.e., mApo and luminal C). Noteworthy, the CIT classification scheme did not define an HER2-enriched subgroup. Instead, the HER2-enriched cancers were distributed in the mApo and

luminal C subgroups. The luminal C subgroups, which not only overlapped with the HER2-enriched intrinsic subtype, but also with basal-like, luminal A and B, and normal-like intrinsic subtypes. Importantly, each of these six CIT classification subgroups showed specific copy number alterations and was correlated to specific signaling pathways. These six subgroups varied in terms of tumor grade, metastatic sites, relapse-free survival or chemotherapy sensitivity.

The above molecular BC classification schemes described (the intrinsic schemes [4,32,34,48], the SCMs [73–75], and the CIT classification scheme [77]) were all based on gene expression profiles to stratify patients into transcriptionally distinct subtypes. Jönsson *et al.* proposed a CNA-based classification scheme to distinguish six genomic subtypes from breast tumors: 17q12, basal-complex, luminal-simple, luminal-complex, amplifier, and mixed subtypes [78]. Four of these genomic subtypes (17q12, basal-complex, luminal-simple, luminal-complex) overlapped with the intrinsic subtypes [78]. By intrinsic classification, HER2-enriched tumors were mainly found in the 17q12 subtype, basal-like tumors were mainly found in the basal-complex subtype, while luminal A and luminal B tumors were imperfectly segregated between the luminal-simple and luminal-complex subtypes [78]. The remaining two genomic subtypes (amplifier and mixed) comprised tumors from multiple intrinsic subtypes [78]. The METABRIC study proposed a subtyping scheme by combining gene expression and DNA copy number [58]. The METABRIC study used large BC sample sets for subtype discovery (n = 997) and subtype validation (n = 995) [58]. DNA and RNA isolated from the samples were hybridized to the Affymetrix SNP 6.0 and Illumina HT-12 v3 platforms to determine copy number and gene expression profiling, respectively [58]. The study first identified genes whose expression across breast tumors were driven by recurrent CNAs that were *cis* (within a 3-megabase window surrounding the gene of interest) using an expression quantitative trait loci analysis [58]. These

cis-driven genes were then used to classify the tumors into 10 integrative cluster (IntClust1-10) subtypes with distinct clinical outcomes, which was reproduced in the validation cohort [58]. Among the 10 subgroups, six (IntClust1, 2, 3, 6, 7, and 8) were dominated by the ER+ tumors and roughly corresponded to luminal A and B intrinsic subtypes by PAM50 [58]. But the six subgroups (IntClust1, 2, 3, 6, 7, and 8) have distinct genomic alterations [58]. Three of the 10 subgroups (IntClust4, 5, and 9) captured both ER+ and ER- cases [58]. IntClust4 was characterized by having very few CNAs and was therefore termed the "CNA-devoid" subgroup [58]. IntClust4 also exhibited a strong immune and inflammation signature, and mainly corresponded to luminal A or basal-like subtypes by PAM50 [58]. IntClust5 was dominated by tumors with the high-level amplification on the q-arm of chromosome 17 (17q), centered on the locus including the *HER2* gene [53]. These tumors are categorized by PAM50 as mainly luminal B or HER2-enriched cases [58]. Almost all tumors in IntClust9 exhibited amplifications on the q-arm of chromosome 8 (8q) (including the oncogene *MYC*) [53]. IntClust9 mainly corresponded to luminal A or luminal B subtypes defined by PAM50, but a substantial number of HER2-enriched and basal-like tumors were also found within IntClust9 [53]. Patients in IntClust9 had an intermediate prognosis [53]. IntClust10 was dominated by ER- samples and had high genomic instability with major chromosomal aberrations [53]. The majority of tumors in IntClust 10 were of basal-like subtype by PAM50 [53]. The same research team also proposed an expression-based surrogate scheme for the IntClust classification [79]. The surrogate scheme used the expression data of 612 genes from a training set of 997 breast tumors to estimate 10 centroids for the 10 IntClust subtypes, which were used for IntClust subtype assignment for new samples [79]. When validating the surrogate scheme in an external dataset of 7,544 breast tumors, they observed similar frequencies of the 10

IntClust subtypes to the original METABRIC study and also recapitulated the molecular and survival patterns of the 10 IntClust subtypes confirmed in the original METABRIC study [79].

More recently, Tofigh *et al.* proposed a hybrid subtyping scheme for breast tumors by combining the clinical subtypes and the intrinsic subtypes defined by PAM50 [56]. The clinical subtypes were defined by ER status (ER+ or ER-) as measured by IHC and HER2 status (HER2+ or HER2-) as measured by fluorescence *in situ* hybridization or IHC [56]. The hybrid scheme partitioned ER+ tumors by intrinsic subtypes into five subtypes while partitioned ER- tumors by HER2 status into two subtypes [56]. Thus, the hybrid scheme stratified invasive breast tumor samples into seven hybrid subtypes: ER+/luminal A, ER+/luminal B, ER+/HER2-enriched, ER+/normal-like, ER+/basal-like, ER-/HER2+ and ER-/HER2- [56].

Distinct BC stem cell groups have also been identified [80]. Leucine-rich repeat-containing G protein-coupled receptor 5 (Lgr5) was considered as a candidate marker of actively cycling mammary stem cells [80]. Through gene expression profiling, tetraspanin 8 (*Tspan8*) was found to be the top upregulated gene between Lgr5-positive ($Lgr5^{+}$) versus Lgr5-negative ($Lgr5^{-}$) basal cells. Based on the expression of Lgr5 ($Lgr5^{+}$ or $Lgr5^{-}$) together with Tspan8 (Tspan8-high ($Tspan8^{hi}$) or Tspan8-negative ($Tspan8^{-}$)), Fu *et al.* identified three different mammary stem cell subgroups: $Lgr5^{+}Tspan8^{hi}$, $Lgr5^{+}Tspan8^{-}$, and $Lgr5^{-}Tspan8^{+}$. These MaSC subgroups were mostly quiescent in adults but there were differences in their molecular and epigenetic signatures [80]. Noteworthy, the $Lgr5^{+}Tspan8^{hi}$ subgroup was deeply quiescent, throughout life, but exhibited an expression profile similar to claudin-low tumors.

1.2.6 Molecular subtyping schemes of TNBCs

Although both clinical and biological features distinguish the five intrinsic BC subtypes, there is substantial heterogeneity within each subtype. In particular, basal-like tumors, which are

dominated by the IHC defined TNBCs, share the fewest similarities with the other intrinsic subtypes but have the greatest diversity [53]. Through gene expression profiling, Lehmann *et al*. identified six distinct molecular subtypes of TNBC tumors: basal-like 1 (BL1), basal-like 2 (BL2), immunomodulatory (IM), mesenchymal (M), mesenchymal stem-like (MSL) and luminal androgen receptor (LAR) subtypes [71]. The BL1 subtype was characterized by overexpression of genes enriched in cell cycle and cell division components and pathways [71]. BL1 tumors had a high *TP53* mutation rate (92%) as well as mutations in genes involved in DNA repair mechanisms (such as *BRCA1*, *BRCA2*, and *RB1*) [81]. The BL2 subtype overexpressed genes enriched in growth factor signalling such as the EGF and Wnt/β-catenin pathways, and glycolysis and gluconeogenesis [71]. The IM subtype was characterized by overexpression of genes involved in immune cell signaling (such as B- and T-cell receptor signaling pathways), antigen processing and presentation, and signaling through core immune signal transduction pathways (such as JAK/STAT signaling) [71]. The M and MSL subtypes overexpressed genes encoding regulators of cell motility, invasion and mesenchymal differentiation [71]. The MSL subtype uniquely overexpressed genes associated with epithelial-mesenchymal transition and stemness [71]. The MSL subtype also shared numerous genes involved in immune signaling with the IM subtype [71]. Claudin-low tumors defined by the intrinsic classification were mostly distributed into the M and MSL subtypes [82]. Finally, the LAR subtype was characterized by overexpression of genes involved in mammary luminal differentiation and androgen receptor signalling [71]. LAR tumors displayed increased mutations in *PIK3CA* (55%), *AKT1* (13%) and *CDH1* (13%) genes [81]. In a further study, Lehmann *et al.* refined their TNBC molecular classification from six to four subtypes: BL1 (immune-activated), BL2 (immune-suppressed), M (including most of MSL tumors), and LAR [83].

Similarly, Burstein *et al.* proposed a classification of TNBCs by combining mRNA expression and DNA profiling analyses to identify four distinct subtypes: luminal androgen receptor, mesenchymal, basal-like immunosuppressed and basal-like immune-activated [84]. These overlapped with the four TNBC subtypes proposed by Lehmann *et al.* [83]. Among the four subtypes, basal-like immunosuppressed had the worst prognosis while basal-like immune-activated had the best prognosis [84]. Also, potential therapeutic targets were identified for each subtype. For example, the basal-like immunosuppressed subgroup can be targeted via the SOX transcription factors and the immunoregulatory molecule VTCN1 [84]. Jézéquel *et al.* differentiated TNBC into three subtypes using gene expression profiling [85]. The first subtype (C1) was positive for luminal androgen receptor as measured by IHC [85]. The majority of tumors in C1 were identified as luminal A and B subtypes by PAM50 [85]. The second subtype (C2) was characterized by low immune response and high M2-like macrophages [85]. Almost all tumors in C2 (except one case) were identified as basal-like by PAM50 [85]. The third subtype (C3) was characterized by high immune response and low M2-like macrophages [85]. C3 was mostly composed of basal-like and claudin-low tumors [85].

As stated above, tumor heterogeneity is an important clinical consideration effecting progression, potential treatment and molecular subtyping among other things. This is especially true when considering TNBC which exhibits high molecular heterogeneity. In an effort to study this, Saleh *et al*. identified four microenvironments (stromal axes) which showed different expression of genes in stroma from 57 TNBC samples [30]. The four stromal axes were T-cell, B-cell, epithelial (E), and desmoplasia (D) [30]. Compared to the six TNBC subtypes from Lehmann *et al.*'s study [71], the E-axis was inversely correlated with the LAR subtype, and the D-axis was positively correlated with the MSL subtype [30]. Based on these four stroma axes, Saleh *et al.*

proposed a novel TNBC classification scheme (called STROMA4), in which a score was first assigned along each stromal axis for each patient and then the four axis scores were combined to subtype patients [30]. Unlike the earlier TNBC classifications (such as the one from Lehmann *et al*.) which assigns each patient to exactly one subtype [71,83–85], the STROMA4 scheme subtyped each tumor by scoring it along all four axes [30]. Furthermore, STROMA4 better characterized the microenvironmental (stromal) heterogeneity in TNBC [30]. Even with all these classification schemes, there is no definitive diagnostic method for the classification of TNBC in routine clinical practice.

Being a highly heterogeneous subtype, having the shortest survival time, and lacking effective therapies, makes TNBC particularly interesting to researchers. As a result researchers are focusing on: 1) developing gene expression signatures and biomarkers for TNBC prognosis and treatment response. prediction; and 2) discovering druggable targets for TNBC treatment. For these reasons, this thesis focuses on TNBCs.

1.3 Profiling and differential analysis of cancer transcriptome

This section first summarizes the two major high-throughput technologies that have been widely used to profile cancer transcriptome (subsection 1.3.1) and then discusses the differential analyses that have been applied to the cancer transcriptome in order to identify the dysregulated genes and pathways in cancer (subsection 1.3.2).

1.3.1 Transcriptome profiling technologies

Transcriptome profiling is essential for studying the global gene expression patterns of cancer. The key technique to profile cancer transcriptomes has convincingly switched from microarrays to whole transcriptome sequencing (i.e., RNA-sequencing (RNA-Seq)) in a matter of years [86]. Microarrays are composed of short oligonucleotide probes which are organized onto grid pattern on a surface in such a way that makes deconvolution straightforward [87]. Fluorescently labelled cDNA is generated from mRNA transcripts extracted from samples of interest by reverse transcription and then combined [87]. Next, the cDNA library can bind by hybridization to the probes and the abundance of each mRNA transcript is determined by fluorescence intensity [87]. There are several limitations with this technology [88]. First, they require previously sequenced and annotated genes to generate the oligonucleotide probes for the array [88]. This means that microarrays only allow the detection and quantification of known transcripts and transcript isoforms. Moreover, error in RNA quantitation can affect ratios of expression compared between two treatments [88]. In addition, detection of cDNAs with very low abundance is lacking or underrepresented due to the background noise caused by non-specific hybridization [88]. Some other factors in the microarray experiments, such as cDNA synthesis/labeling, quality of array and hybridization efficiency, can also be sources of error [88].

Nowadays, RNA-Seq, which combines high-throughput sequencing with computational methods, has become the method of choice for profiling transcriptomes [6]. In RNA-Seq, RNA molecules are first extracted from the samples of interest and then reverse-transcribed into cDNA molecules [86]. The cDNA library is then sequenced to produce millions of short reads (i.e., short sequences of base pairs corresponding to all or part of a single cDNA fragment) [86]. The resulting reads can be aligned, either by mapping to a reference genome or by *de novo* assembly [86]. The most common is to map the reads to a known transcriptome or annotated genome [86]. This process can be accomplished using distinct alignment tools (such as TopHat [89], STAR [90] or HISAT [91]) for which a reference genome is required [86]. After the sequence reads have been mapped to the reference genome or transcriptome, the next step is quantification of transcript- or gene-level abundances by assigning the mapped reads to specific transcripts or genes [86]. The abundance of a given mRNA transcript is determined by the number of reads mapped to that transcript [86]. Whereas the abundance of a given gene (i.e., all transcript isoforms produced from that gene) is obtained by either direct read overlap counting of that gene, or transcript-level abundance quantification followed by aggregation to the gene level [86]. The tools commonly used for quantification include RSEM [92], CuffLinks [93], MMSeq [94] and HTSeq [95]. The results of quantification are combined into an expression matrix, where each feature (gene or transcript) is in a row and each sample in a column, where the values are actual read counts or estimated abundances.

RNA-Seq has key advantages over the gene expression microarray approaches [88]. Compared to microarrays, RNA-Seq does not need priori knowledge and thus allows the detection and quantification of both known and novel transcripts and transcript isoforms. Also, RNA-seq can detect transcripts with low abundance and transcripts with high abundance more accurately

relative to microarrays. Moreover, RNA-Seq can detect allele specific variants as well as structural variation (e.g., gene fusion) [88]. Thus RNA-Seq has become an indispensable tool for profiling transcriptional outcomes on a large scale. It is worth mentioning that despite the broad application of bulk RNA-Seq, there are some cases where single-cell resolution is desired, especially when studying intra-tumor heterogeneity [86]. Although bulk RNA-Seq tumor expression data can be computationally deconvoluted to infer the composition of cancer and immune cells, one cannot discover new cell types or perform cell-type-specific analyses with bulk RNA-seq data [96]. Therefore single-cell RNA-Seq may be a more appropriate way to study intra-tumor heterogeneity. Nevertheless, in this thesis, we focused on inter-tumor heterogeneity and used bulk RNA-Seq data (such as those from TCGA) as well as microarray-based gene expression data (such as those from METABRIC) .

1.3.2 Differential analysis of cancer transcriptomes

Differential analysis can identify the gene expression changes of cancer samples with comparison to the patient-matched or unmatched normal tissue samples. The differential approaches can be performed at the gene or gene set level. The former aims to identify the upregulated or downregulated genes (i.e., differentially expressed genes (DEGs)) in the tumor samples relative to the normal tissue samples by accomplishing two steps: 1) calculating the fold change of a given single gene based on the normalized gene expression data between two conditions (tumor/normal); 2) estimating the significance of the differential expression and correct for multiple testing. There are established tools which are computationally efficient and provide comparable results for gene-level differential analysis such as limma-voom [97], edgeR [98] and DESeq2 [99]. All these tools are developed in the R language environment and thus referred to as R packages. It is important to take the experimental design into consideration when deciding which

tool should be used. For example, the limma-voom and DESeq methods tend to have better control of false positives than edgeR while the edgeR method is recommended for experiments with less than 12 samples per condition [100]. Accordingly, in Chapter 2 of this thesis, the limma-voom package was used to identify the DEGs in breast tumor samples compared to the normal tissue samples because of a large sample size and the necessity to have low false positives.

Alternatively, differential analyses could be performed at the gene set level. Unlike the gene-level analysis, which is to test whether the mean expression value for a (single) given gene differs between two conditions (tumor/normal), the gene set-level analysis aims to test whether the expression values for a given pre-defined gene set vary between the tumor and normal groups. The gene sets can be obtained from different sources, such as the sets of genes representing biological pathways, or the sets of genes representing signatures of important biological and clinical states like cancer metastasis and drug resistance. Using the gene set-level analysis has advantages over the gene-level analysis [101,102]. First of all, it reduces the complexity of analyzing thousands of individual genes to just several hundred gene sets. Second, it increases the explanatory power since results from gene set analysis are often easier to interpret than a simple list of DEGs. Gene set-level analyses require two general resources. The first is a knowledge database (*D*) of pathways (gene sets involved in well-known and well-studied biological pathways) or any other functionally-related gene sets. The second is a list *(L*) of genes with the corresponding expression levels in two conditions (i.e., two groups of paired or unpaired samples).

A large number of knowledge databases have been developed to help with the gene set analyses. Some of the most widely and commonly used include Gene Ontology (GO) [103] and Molecular Signatures Database (MSigDB) [104], which were used in this thesis. An important way to organize biological concepts is through ontologies. An ontology formally organizes

biological concepts within a domain in a hierarchical structure and usually consists of a set of terms (or concepts) and the relationships that exist among them. The GO knowledgebase provides three ontologies which systematically describe gene products with respect to three domains: cellular component, molecular function and biological process [103]. Each ontology forms a rooted directed acyclic graph with nodes and edges between nodes. In an ontology graph, each node is a defined GO term representing gene product properties and edges between the nodes are the categorized and defined relationships between the GO terms. In an ontology graph, the child GO terms are more specialized than their parent GO terms. Since genes are known to be involved in multiple biological activities, a GO term may have more than one parent GO term in the same ontology, and sometimes in the other ontologies. A gene annotated with any given GO term is also annotated with all ancestral GO terms, allowing for descriptions of the gene product at varying levels of specialization. Of note, each GO term is also a gene set.

MSigDB is now one of the largest and most widely used gene set databases [104]. Gene sets in MSigDB can be categorized into eight collections, some of which can be further divided into several sub-collections. Some MSigDB gene sets are formed by grouping genes sharing biological characteristics together. For example, the C1 collection (called positional gene sets) groups genes located in the same chromosome or cytogenetic band together as a gene set. MSigDB also generates gene sets according to bioinformatics analysis results. For example, the C3 collection (called motif gene sets) groups genes predicted as potential targets of regulation by the same transcription factors or miRNAs together as a gene set. The C4 collection (called computational gene sets) groups genes sharing a strongly correlated cancer-related gene expression pattern by computational analysis together to form gene sets. The MSigDB also collects gene sets from various online gene set databases. For example, the C5 collection (called Gene Ontology

gene sets) stores the gene sets (i.e., GO terms) derived from the GO database. One sub-collection (called canonical pathways) of the C2 collection (called curated gene sets) curates pathways (gene sets involved in well-known and well-studied biological pathways) from various pathway databases (such as Kyoto Encyclopedia of Genes and Genomes (KEGG) [105], Reactome [106,107], BioCarta [108] and the Pathway Interaction Database (PID) [109]). In addition, MSigDB curates gene sets from the biomedical literature. Over the past several years, various experimental papers based on gene expression data have identified many drug or disease-associated gene signatures correlating with transcriptomics changes induced by genetic (such as RNA interference) and chemical (such as drugs and chemical compounds) perturbations. Some of these signatures are a gene sets composed of two groups of genes, with one group upregulated (i.e., induced) and another group downregulated (i.e., repressed) by the perturbation. One sub-collection (called chemical and genetic perturbations) of the C2 collection curates such gene expression signatures as gene sets. The other commonly used gene set databases are further introduced in section 1.6.1.

Given all the different sources of gene sets, there are typically two approaches to perform gene set-level differential analyses. One is over-representation analysis, which determines whether genes in a given pathway or gene set from the knowledge database *D* are present in the gene list *L* more frequently than expected by chance (over-represented) [102]. A typical over-representation analysis workflow is conducted by the following steps. First, the input gene list *L* is generated by a certain threshold or criteria. As an example, in Chapter 2 of this thesis, we chose genes that are differentially over- or under-expressed in BC samples relative to the adjacent normal breast tissue samples with a false discovery rate (FDR) less than 0.05 and an absolute value of log2 fold change (logFC) not less than 1. Then, for a given gene set in the knowledge database *D*, the number of

genes that are included in both the gene set and the input gene list *L* is counted. Also, for the same gene set, the number of genes that are included in both the gene set and an appropriate background list of genes (e.g., all genes measured) is counted. Next, the statistical significance of the over-representation of the gene set in the input gene list *L* is tested, which are most commonly based on the hypergeometric, chi-square, or binomial distribution. Various established over-representation analysis tools are available, based on either the web (e.g., the L2L tool [110], the Database for Annotation, Visualization and Integrated Discovery [111,112]) or through the R language using packages such as clusterProfiler [113]. Results generated with the same gene set databases using different tools do not differ much from each other because they perform the over-representation analysis with the same statistical tests. However, using the over-representation analysis has some drawbacks. First, the input list of genes are chosen using an arbitrary threshold and potentially informative genes may be excluded. Second, the statistical tests used by the over-representation analysis for identifying significant gene sets only consider the number of genes involved in a gene set while ignoring the measured expression changes between two conditions of these genes. Third, the over-representation analysis treats genes in a given gene set equally. Fourth, the over-representation analysis considers gene sets independent of each other.

An alternative approach, known as Gene Set Enrichment Analysis (GSEA), determines whether genes in a given gene set from the knowledge database *D* are significantly ranked at the extremes (top or bottom) of the list (*L*) of genes which are ranked according to the differential expression between groups (tumor and normal) [114]. The assumption of GSEA is that top-ranking genes in the experiment (which is determined by magnitude of differential expression between two conditions) are more important in terms of biological function. In GSEA, the first step is to obtain the ranked gene list *L* using the original expression matrix. Then, for a given gene set from the

knowledge database *D*, the extent of association (Enrichment Score) of this gene set with the experiment is measured by using a non-parametric running sum statistic (i.e., Kolmogorov-Smirnov statistic) to aggregate statistics for all genes in the gene set. The next step is to compute the statistical significance of the Enrichment Score of the pathway by simulating the distribution of Enrichment Score under the null hypothesis to get the empirical *p*-value. The null hypothesis tested by GSEA is a competitive null hypothesis, which compares the differential expression of the genes in the given gene set against the genes in all the rest of the gene sets and assumes that the genes in the given gene set are not more differently expressed than the genes in all the rest of the gene sets. The final step is correction for multiple testing based on the number of gene sets tested. GSEA can be performed using the Java software GSEA across platforms (Windows, Linux and Mac). Compared to the over-representation analysis, GSEA has a number of benefits [102]. First, GSEA ranks all genes in terms of their expression difference across two groups thus it does not discard any potential informative genes. Second, the Kolmogorov-Smirnov statistic used by GSEA to identify significant gene sets considers the coordinated changes in the expression of genes in the same gene set. Finally, GSEA does not treat genes in a gene set equally, but confers more importance to the top-ranking genes. However, GSEA, similar to the over-representation analysis, also ignores the dependence of gene sets. GSEA was used in Chapter 2 of this thesis to identify pathways or gene sets that are differentially over- or under-expressed in each BC subtype relative to the normal breast tissue samples. Several variations of GSEA, such as GSA [115] and Gorilla [116], have been proposed after the seminal paper of Subramanian *et al.* [114] was published.

According to the above definition of the null hypothesis, Goeman *et al.* reduced all gene set analysis approaches into two categories: competitive and self-contained tests [117]. A

competitive test compares differential expression of a given gene set with the rest of the genes that are not in the gene set [117]. In contrast, a self-contained test compares a given gene set with itself, while ignoring the genes outside the gene set [117]. Most of the proposed approaches for gene set analyses are based on a competitive test, including all the aforementioned methods (such as the overrepresentation test and GSEA). However, there are also several self-contained tests, including the Pathway Level Analysis of Gene Expression (PLAGE) [118] and GlobalTest [119,120]. The GlobalTest can, in a single test, determine if the expression profile of a given pathway (may be any set of genes) is significantly associated with a clinical trait of interest, based on a logistic regression model [119,120].

The majority of the gene set-level differential analysis methods we have described assume two groups of samples (e.g., case/control) and identify overrepresentation or enrichment of gene sets within this context. Other approaches, which allow gene set-level differential analysis for each tumor sample, have been proposed, such as z-score [121], single-sample GSEA (ssGSEA) [122], sample-level enrichment analysis (SLEA) [123], and gene-set variation analysis (GSVA) [124]. GSVA initially uses a non-parametric kernel (Gaussian kernel for microarray and Poisson kernel for RNA-Seq) estimation to calculate gene-level statistics (evaluating whether a gene is high or low expression in individual samples) and then aggregates gene-level statistics into gene set enrichment scores in a similar manner to GSEA. It should be noted that the methods currently available for single sample gene set analysis calculate an enrichment score for a gene set in a single patient compared to all the other patients in the dataset. Hence, the heterogeneity and number of the tumor samples in the dataset can affect the enrichment scores. Recently, Paquet *et al.* proposed an approach named absolute inference of patient signatures (AIPS) [125] based on their previously published absolute intrinsic molecular subtyping (AIMS) methodology [126] which is a PAM50

alternative implementation based on a naïve Bayes classifier of 20 gene pair comparison rules. The AIPS method built a reliable absolute model for each of 1733 different pathways based on a collection of binary rules [125]. Each AIPS model can estimate the activity of a given pathway from the expression profile (either microarray or RNA-seq) of a single BC sample without the need for a large and diverse patient dataset [125].

1.4 Regulatory programs in cancer transcriptomes

This section first introduces different types of factors (subsection 1.4.1) that are involved in the regulation of gene expression and then summarizes some studies (subsection 1.4.2) on modeling regulatory programs in cancer.

1.4.1 Gene regulation mechanisms in cancer

There are a wide range of mechanisms that are used by cells to regulate gene expression (increase or decrease the protein/mRNA production of a specific gene), which is required to establish and maintain normal cell states in humans. The mechanisms include alterations in gene copy numbers, transcription factors, DNA methylation, post-translational modifications to histones, and miRNAs, which could cause the misregulation of gene expression programs and thus lead to cancer.

CNVs are a type of structural variant in the genome that involves changes in the number of copies of certain regions of DNA (loci) [127,128]. These DNA regions are normally larger than one kilobase pairs and may span many different genes. CNVs can arise by deletions (which can lead to gene copy number loss) or duplications (which can lead to gene copy number gain). In principle, a gene has two copies in the diploid genome of normal human cells, with each copy inherited from each parent. But in cancer cells, the number of copies of some genes becomes smaller (gene copy number loss) or greater (gene copy number gain) than two and thus decreases

or increases expression of these otherwise normal genes, respectively. A gain is considered as an amplification when a gene has multiple copies resulting from aberrant DNA duplication [129]. Gene amplification and other CNVs are considered to be detrimental because they alter gene expression patterns and thus lead to an imbalance in cell growth and differentiation [130]. Some CNVs are inherited and known as germline CNVs whereas others that arise spontaneously (or *de novo*), are somatic [131]. The former, which is under negative selective pressure, occurs rarely while the latter plays an important role in cancer development [130]. For example, amplification of some oncogenes (such as *MYCN*, *MYC*, *MYCL1*, *REL*, *AKT2*) accompanied by overexpression can drive cancer development [132]. A classic example of CNV events that contribute to tumorigenesis is the amplification of *HER2* gene and the consequent overexpression of HER2 protein in the clinically defined HER2+ subtype, which accounts for 15-20% of all BCs [46].

Single nucleotide polymorphisms (SNPs) are substitutions of a single nucleotide for another that are typically present in greater than 1% of the population. SNPs are the most common type of genetic variation among humans and most are bi-allelic variants, meaning that only two possible alleles exist in the population [133]. The frequency of the less common allele (minor allele) is called the minor allele frequency and can vary among populations. SNPs can occur in either the coding or non-coding regions throughout the genome. If a SNP occurs in a protein coding region, then this could result in either a synonymous or nonsynonymous change. Synonymous SNPs (also called silent SNPs) do not cause a corresponding change in the amino acid sequence of the resulting proteins because of degeneracy of the genetic code. Nonsynonymous SNPs can be either a missense variant leading to an amino acid substitution or a nonsense variant leading to a premature stop codon Over the past two decades, particularly through genome-wide association studies, researchers have revealed a substantial contribution of SNPs to human disease risk and other

complex traits. For example, approximately 180 common SNPs have been identified as being significantly associated with BC risk, and explaining up to 18% heritability of BC [134]. An easily interpretable mechanism through which SNPs may affect phenotype is the direct alteration of the structure of the coded protein and thus of its functionality such as with nonsynonymous SNPs. Alternatively, SNPs may affect the regulation of gene expression by disrupting specific promoter sequences which bind proteins regulating the expression (e. g., transcription factors). For example, Huo *et al.* reported that 132 SNPs were found to disrupt transcription factor binding and that 97 of them were associated with gene expression in human brain tissues [135]. SNPs in the miRNA binding site of target mRNAs can disrupt miRNA-dependent regulation and thus also alter gene expression in cancer [136]. For example, many SNPs found in primary and precursor miRNA primary sequences can influence the synthesis or stability of these miRNAs, as well as miRNA-mRNA interactions, resulting in changes in the expression of miRNA target genes [136]. Another possibility is the alteration of the structure or sequence of DNA, thereby affecting the functionality of *cis*-regulatory elements such as enhancers, promoters and silencers that ultimately influence gene expression.

Loss of heterozygosity (LOH) is also a type of genetic alteration and is a loss of one of the two alleles (one from each parent) at a heterozygous locus. LOH can occur with copy number losses, which involves the loss of all or part of a chromosome [137]. LOH can also occur without a change in copy number (known as copy number neutral-LOH), which involves the loss of an entire chromosome or a large segment of a chromosome and the duplication of the remaining chromosome or segment [137]. In copy number neutral-LOH, the retained two copies are inherited from only one parent. LOH can originate through direct deletion, deletion due to unbalanced rearrangements, gene conversion (a homologous recombination event), mitotic recombination or

loss of all or part of a chromosome. LOH is a common somatic genetic change in human tumors. When leading to loss of the only remaining normal (wild-type) allele of a tumor suppressor gene, LOH can contribute to cancer development. LOH in this case represents the second hit in the two-hit hypothesis proposed by Knudson in 1971 [138,139], which is an important mechanism for the complete inactivation of tumour suppressor genes. According to this hypothesis, tumor suppressor genes are inactivated by an initial mutation of one normal allele (the first hit) followed by loss of the other normal allele as a result of LOH (the second hit) [138,139]. In hereditary cancer syndromes, the first hit is inherited via the germ cells and the second (i.e., LOH) occurs in somatic cells. In the nonhereditary cancers, both hits are in somatic cells. LOH has been explored in many cancer-related genes and in different types of cancers. In BC, the most important LOH were seen in cancer-related genes such as *BRCA1* and *BRCA2*, *TP53*, *RB1*, *CHK2*, *PTEN*, *CDH1*, *SMAD2* and *SMAD4* [140]. For example, *BRCA1* and *BRCA2* are two canonical tumor suppressor genes that play an important role in DNA double strand break repair via homologous recombination [141]. *BRCA* locus-specific LOH (i.e., loss of the non-mutated allele at the *BRCA1*/*2* locus) is frequently observed and leads to high levels of genomic instability in women with germline *BRCA1/2* mutation-associated breast tumors as well as ovarian cancers [142]. An analysis of TCGA data revealed that 3,780 out of 10,925 human genes (34.6 %) were significantly under-expressed in ovarian cancer samples with LOH compared to those without LOH [137]. For a long time, the Array Comparative Genomic Hybridization and Single Nucleotide Polymorphism genotyping arrays have been used to detect genome-wide CNV or LOH, with the latter being more sensitive than the former due to the improved resolution and the ability to detect SNP alleles [143]. As next-generation sequencing is widely used due to the rapid decrease in price and increase in accuracy, CNV and LOH studies are increasingly using sequencing [144–146].

DNA methylation is an epigenetic modification of DNA that plays an important regulatory role in transcriptional silencing of genes [147]. In the human genome, promoter regions of many genes contain CpG islands, which are genomic regions that contain a high frequency of CpG sites where a cytosine nucleotide is followed by a guanine nucleotide [148]. DNA methylation can occur on CpG sites by the addition of a methyl group to the C5 position of cytosines in CpG dinucleotides to form 5-methylcytosine and this reaction is catalyzed by DNA methyl transferases [149]. Methylation of multiple sites within CpG islands is called hypermethylation and if this happens in a promoter it can ultimately result in transcriptional silencing of the gene [150–152]. In normal cells, only a small number of promoter CpG islands are hypermethylated [147]. However, some cancers can result from, aberrant promoter CpG island hypermethylation leading to inappropriate gene silencing, especially tumor suppressors [150,153]. For example, more than 100 genes have been reported to be transcriptionally silenced by aberrant CpG island hypermethylation in BC [102,154–157]. In particular, CpG hypermethylation suppresses expression of the tumor suppressor cyclin-dependent kinase inhibitor *CDKN2A* (also known as *p16*) and is involved in BC [158], gastric cancer [159], and other types of cancers [160]. CpG island methylation can be prevented by inhibition of DNA methyltransferases [161]. There are also enzymes (such as the ten eleven translocation (TET) group of enzymes) that can remove CpG methylation [162]. This suggests that it is possible to design new drugs to take advantage of the reversible property of the DNA methylation processes [163]. For example, many researchers are trying to turn silenced genes (such as tumor suppressors) back on in a cancer cell and thereby to help re-establish normal growth patterns of the cell [163]. In order to study DNA methylation, bisulfite sequencing-based approaches such as the Illumina

arrays (HumanMethylation450 and HumanMethylation850 arrays) have been used to detect global DNA methylation in the whole genome [162,164].

miRNAs are a family of small (about 22 nucleotides in length) non-coding RNA molecules that negatively regulate gene expression post-transcriptionally [165,166]. As of 2019, there are up to 1,917 annotated miRNA precursor genes in the human genome, which could be further processed to generate 2,654 mature miRNA sequences [167]. miRNAs can bind to the three prime untranslated regions (3'-UTRs) of the target mRNAs that they regulate [165] and either repress the translation or cause the degradation of the target mRNA molecules [165]. miRNA-mediated post-transcriptional regulation is a general regulatory mechanism underlying many biological processes but it can also be involved in tumorigenesis [168]. In cancer cells, dysregulation (i.e., abnormal expression) of miRNAs relative to normal cells due to genomic or epigenomic alterations in miRNA genes (such as CNVs or DNA methylation) have been reported [169,170]. For example, under-expression of miR-9, miR-34b/c, miR-124a, and miR-148a by DNA hypermethylation was found to be associated with BC [171–173]. Also, overexpressing miRNA-182 was found to repress expression of many tumor suppressors in BC, including *BRCA1* [174], *RECK* [175], *PFN1* [176], *FOXO1* [177], *ZEB1* and *HSF2* [178]. Therefore, miRNAs are being studied as potential targets for cancer therapy [178]. Similar to profiling the mRNA transcriptome, the global miRNA expression profile can be detected using microarrays or RNA-Seq [179].

TFs play a key role in the regulation of gene expression via influencing RNA polymerase activity in a gene-specific manner [180,181]. TFs are proteins that typically have two distinct interaction domains: a DNA-binding domain (such as zinc finger, homeodomain and basic helix-loop-helix motif) that interacts with *cis*-regulatory elements, and an activation/repression domain

that interacts with various cofactors (coactivators or corepressors), other transcription factors and even transcriptional machinery (e.g., RNA pol II) [180,181]. TFs typically regulate gene expression by first binding to the *cis*-regulatory elements of their target genes which then recruit other transcription factors, co-activators and repressors which ultimately influence the recruitment of RNA pol II to target genes [180,181]. TF cofactors are large, multi-subunit protein complexes that contribute to either transcriptional activation (coactivators) or repression (corepressors), but they cannot bind to DNA on their own [180,181]. As key cellular components in controlling gene expression, TFs play a central part in controlling many biological processes in cells. Approximately 2000 genes in the human genome code for different TFs [180]. TFs have been known to become dysregulated causing aberrant gene expression in numerous cancer types [182]. In tumor cells, genetic alterations (such as amplification, deletion, rearrangement via chromosomal translocation or point mutations) often occur in genes that encode TFs and hence directly alter the activity of these TFs [183]. TF activity can also be altered indirectly by non-coding DNA mutations that affect TF binding [183]. Indeed, Vaquerizas and colleagues reported that 164 out of 1,391 manually curated TFs (~12%) are directly implicated in 277 diseases including cancer [3]. Cancer genes are genes that play a role in driving cancer upon acquiring alterations, including two major types: tumor suppressors and oncogenes [184]. The Network of Cancer Genes 6.0 curated a total of 2,372 cancer genes, including 711 known cancer genes whose driver role in cancer has been experimentally validated and 1,661 potential cancer genes whose somatic alterations suggest they play a role in cancer but require additional experimental support [184]. When comparing the 2,372 cancer genes to a compiled list of 1,988 human TFs, Ravasi *et al.* found that 460 cancer genes are also TFs [185]. Of these 460 cancer genes which are also TFs, 245 are known cancer genes, accounting for ~34% (245/711) of all known cancer genes [185]. These results implicate a large

number of TFs in human cancers. In recent years, study of the role of TF oncogenes in tumorigenesis has expanded considerably and revealed that TFs modulate properties of cancer, including stem cell properties (e.g., self-renewal), unrestricted cell proliferation, differentiation and/or cell death, development of resistance, autoregulatory circuits and immune evasion. Overexpression of the ETS family TFs (i.e., ETS-related gene and ETS translocation variant 1) caused by chromosomal translocation events has been shown to drive prostate cancer [183]. Overexpression of ETS translocation variant 1 is also implicated in gastrointestinal stromal tumors and melanoma [183]. In addition, the runt-related transcription factor RUNX2 was found to drive the epithelial-to-mesenchymal transition in breast and prostate cancers [183]. Therefore, mutated or dysregulated TFs represent an important and unique class of therapeutic targets for cancer treatment. Various approaches to indirectly or directly target TF activity have been attempted including inhibition of regulators of TF expression, modulation of TF activity by blocking TF-cofactor-protein interactions, blockade of TF-DNA binding, and inhibition (or activation) through the binding of a ligand-based molecule in an activation/inhibition pocket [183,184]. Moreover, several new approaches to target TFs have recently emerged, including modulation of the auto-inhibited state of TFs, proteolysis targeting chimaeras, use of cysteine reactive inhibitors, targeting intrinsically disordered regions of TFs, and modulating TF activity through modulating post-translational modifications of TFs [183]. These innovations in drug development may yield drugs with unique mechanisms that could produce new classes of cancer treatments.

1.4.2 Modeling regulatory programs in cancer

High-throughput approaches enable the investigation of a tumor at multiple levels. For example, the TCGA database has characterized mRNA and miRNA transcriptomes as well as DNA methylation and CNV profiles for thousands of patient samples from various cancer types

[10]. In addition, a large number of databases have provided experimentally derived or computationally predicted miRNA-target mRNA pairs [186–189] and TF-target gene pairs [190–196]. With the availability of these data and considering that there are many factors regulating gene expression, there is an urgent need to develop computational models to integrate these factors by effectively combining heterogeneous data from multiple platforms to investigate the regulatory mechanisms underlying abnormal gene expression in cancer. Two main computational approaches, network-based approaches and regression approaches, have been used to do this.

Gene regulatory networks have been used to describe and computationally reconstruct the complex scheme of interactions behind the regulation of gene expression at the transcriptional level. A network is a mathematical construct composed by a set of vertices (i.e., nodes), and a set of edges (i.e., links) that represent the relationship between them (**Figure 1**) [197]. A network is usually represented by a defined graph $\boldsymbol{G = (V, E)}$, where $\boldsymbol{V}$ denotes the set of vertices, and $\boldsymbol{E}$ denotes the set of edges. If $\boldsymbol{I}$ is a set indexing the vertices, the set of edges is a subset of the Cartesian product $\boldsymbol{E \in I \times I}$, with element $(\boldsymbol{v_i}, \boldsymbol{v_j})$ indicating the existence of an edge between vertex $\boldsymbol{v_i}$ and vertex $\boldsymbol{v_j}$. A network can be either directed or undirected. A directed graph contains edges with a direction, which means all the edges are ordered pairs (**Figure 1a**). In other words, $(\boldsymbol{v_i}, \boldsymbol{v_j})$ and $(\boldsymbol{v_j}, \boldsymbol{v_i})$ indicate two distinct edges. In contrast, an undirected network contains edges with no direction, which means all the edges are unordered pairs of vertices (**Figure 1b**). In other words, whenever edge $(\boldsymbol{v_i}, \boldsymbol{v_j})$ exists also edge $(\boldsymbol{v_j}, \boldsymbol{v_i})$ exists as $(\boldsymbol{v_i}, \boldsymbol{v_j})$ and $(\boldsymbol{v_j}, \boldsymbol{v_i})$ yield the same edge. Moreover, in many cases, each edge in a network is assigned a real number (i.e., edge weight) to capture the varying importance of each edge (**Figure 1c**). Weighted networks have weighted edges and can be directed or undirected. In a gene regulatory network, nodes typically represent genes, both coding (producing mRNA) and noncoding (producing for example, miRNAs

and lncRNAs), whereas edges represent various regulatory relationships between genes and can be weighted to describe the strength of the regulatory relationship. Gene regulatory networks tend to contain directed edges. For example, a regulatory node (TF, miRNA, etc.) influences a target node (target gene) but not *vice versa*. While undirected edges simply indicate the presence of a regulatory relationship between any two nodes. Numerous approaches have been used to reconstruct a gene regulatory network. Concerning the dominant underlying data used to infer gene regulatory networks, a 2019 review article by Mercatelli *et al.* distinguished six categories: co-expression, sequence motifs, chromatin immunoprecipitation (ChIP), orthology, literature and protein-protein interaction data [198]. Concerning the statistical and mathematical concepts used for gene regulatory network inference, there are correlation coefficients, information theoretic scores (e.g., mutual information), regression-based methods, gaussian graphical models, Bayesian networks, Boolean networks, and differential equation methods [197,199,200]. As an example, one of the simplest network architectures is the correlation network based on co-expression which assumes that interacting genes should have correlated expression. Given a set of $\boldsymbol{N}$ expression measurements (e.g., different conditions or samples) for $\boldsymbol{G}$ genes, one arranges them into a data matrix $\boldsymbol{D} \in \mathcal{R}^{\boldsymbol{N}\times\boldsymbol{G}}$, with each row being a sample and each column being a gene. Correlations (Pearson, Kendall or Spearman) between every two columns of $\boldsymbol{D}$ can be calculated and yield a $\boldsymbol{G} \times \boldsymbol{G}$ matrix of pairwise gene correlations. A gene regulatory network can be reconstructed upon the correlation matrix, which can be represented by an undirected graph with edges weighted by correlation coefficients. For more discussions about gene regulatory network inference, please consult the reviews of Huynh-Thu *et al*.[197] and Mercatelli *et al* [198]. Most of the methods for gene regulatory network inference have been implemented in software tools that are freely available to the community [197]. For example, the implementation of the weighted correlation

network analysis method proposed by Zhang *et al.*, which generates gene co-expression networks based on pairwise expression profile correlations, is available as a R package and has been widely used by the community [201].

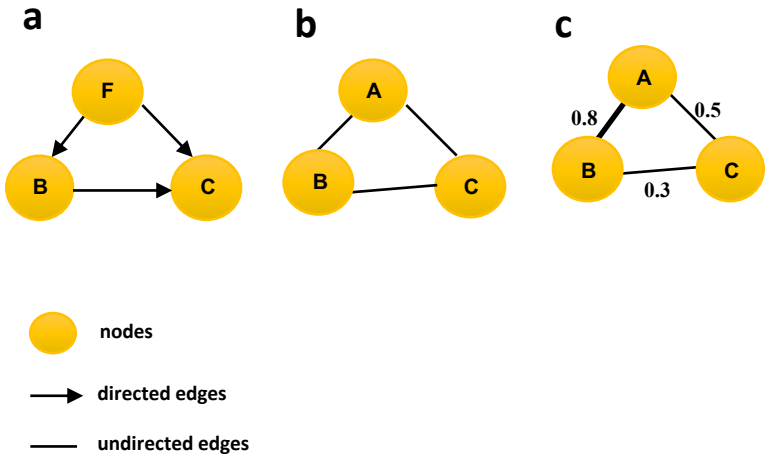


Figure 1.1 Examples of network types
(**a**) Directed network, where the directed relationships between node pairs are usually represented by directed edges (represented by arrows in the graph). (b) Undirected network, where the order of the nodes in a node pair is irrelevant. (**c**) Weighted network, where the weighted edges are represented by lines with different thickness, with the larger the weights the thick the lines.

Linear regression approaches, especially least absolute shrinkage and selection operator (Lasso) regression, have been increasingly used to determine gene regulatory programs. In statistics science, the relationship between a response variable (dependent variable) and one or more explanatory variables (independent variables) can be modeled by linear regression with the assumption that the relationship is linear. Linear regression is referred to as simple linear regression in the case of only one explanatory variable, whereas it is named multiple linear regression when there are many explanatory variables. Of note, in the field of computational biology, the term explanatory variables is often called input variables, features, or predictors while the term response variable is often called outcome variable. If $y \in \mathcal{R}$ is the outcome variable that we would like to predict, and $x \in \mathcal{R}^p$ is the input variables available, which can be written as a p-

dimensional column vector $x = [x_1, \cdots, x_p]^T$ where *T* denotes the transpose of a vector. The linear regression model assumes that the relationship between y and x can be expressed as in Equation 1.

$$y = \beta^T x + э \qquad \textbf{Equation 1}$$

where β is the column vector collecting the parameters (also called the coefficients or weights) of *x* and *э* represents the error term. Since we are assuming there is approximate linear relationship between our input (*x*) and output (*y*) variables, we want to find a regression line that fits the observed data as well as possible. In other words, we want to find β values that reduce the difference between the predicted outcome (y') and the observed outcome (*y*) as much as possible. To this end, the most commonly used method is least squares which minimizes the residual sum of squares between y' and *y*. If we have a dataset $\left\{y^{(i)}, x_1^{(i)}, \cdots, x_p^{(i)}\right\}_{i=1}^{n}$ collecting n observations of corresponding response and explanatory variables, then we can make use of them to find parameter vector β for the best-fitting line. We can get the predicted value $y^{(i)\prime}$ of each $y^{(i)}$ by substituting $\boldsymbol{x}^{(i)} = [x_1^{(i)}, \cdots, x_p^{(i)}]^T$ into **Equation 1**. We can then optimize the least squares function as follows to get the best-fitting parameter vector β for this dataset:

$$\sum_{i=1}^{n}(y^{(i)} - y^{(i)\prime})^2 \qquad \textbf{Equation 2}$$

which is a convex function that can be minimized using the Gradient Descent Algorithm or other optimization techniques.

In the case of datasets with many input variables, such that p is of a high value, regularization techniques are used to prevent the linear regression model from overfitting on the training data. The key idea of regularization is to introduce a constraint on feature coefficients β. Concerning the regularization methods used in linear regression models, there are Ridge regression

[202], Lasso regression [203], and Elastic net regression [204]. Ridge regression, using the L2 regularization method, introduces a L2 penalty term $\lambda \sum_{i=1}^{p} \beta_i^2$ to the previously defined cost function (**Equation 2**) to yield **Equation 3**:

$$\sum_{i=1}^{n}(y^{(i)} - y^{(i)'})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \qquad \textbf{Equation 3}$$

where λ is the regularization parameter which determines how much to penalize the feature coefficients. The sum of squares of all feature coefficients is called the L2 term. Ridge regression forces the coefficients of unimportant features to be small but does not make them zero.

However, in the case where this is a large number of input features (for example, thousands of genes with a few tens to hundreds of samples in an RNA-Seq expression dataset), it is desirable to select a smaller subset of features (genes in the RNA-Seq case) exhibiting the strongest effects on our model's predictability and Lasso regression can be used. This method appeared first in geophysics literature in 1986, and then re-found and became popular in 1996 by Robert Tibshirani. Specifically, Lasso regression, using L1 regularization, adds a L1 penalty term $\lambda \sum_{i=1}^{p} |\beta_i|$ to the previously defined cost function (**Equation 2**) to yield **Equation 4**:

$$\sum_{i=1}^{n}(y^{(i)} - y^{(i)'})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \qquad \textbf{Equation 4}$$

L1 term is the sum of the absolute value of the feature coefficients. Lasso regression does feature selection by assigning insignificant input features with zero coefficients and useful features with a nonzero coefficients. In this way the Lasso method produces a regression model that is simple, interpretable and contains a subset of input features. It's worth mentioning that Elastic net regression is making more progress these days, which incorporates both L1 and L2 regularization methods as below:

$$\sum_{i=1}^{n}(y^{(i)} - y^{(i)'})^2 + \lambda_1 \sum_{j=1}^{p} \beta_j^2 \ + \lambda_2 \sum_{j=1}^{p} |\beta_j| \qquad \textbf{Equation 5}$$

Using a linear regression model to handle continuous data is straightforward. If the data in hand has a categorical variable with n levels, the typical solution is to employ contrast matrix to transform the n-level variable into n new variables, with each new variable having two levels.

In the context of modeling regulatory programs using regression methods, the explanatory variable*s* are derived from databases and could include, but are not limited to, CNV, DNA methylation, TF-binding and miRNA data. These could all influence the response variables which would be expression of particular genes. Questions, such as how can we determine which factors are significant for and how large of a role does each factor play in regulating gene expression, can be answered by using the regression analysis. For example, Setty *et al.* developed a Lasso regularized linear regression model in a glioblastoma multiforme study using TGCA data [205]. The model was fit using changes in mRNA expression in each glioblastoma multiforme sample as the response variable by a linear combination of the explanatory variables including CNV, DNA methylation, miRNA sequence-based predictions and TF-binding sites. The model successfully identified various key miRNAs and TFs as either common or subtype-specific drivers of expression changes. The estimated activities of these key regulators were predictive for glioblastoma multiforme subtypes and patient survival. In a more recent study, Li *et al.* developed a two-stage Lasso regression model to determine key regulators in acute myeloid leukemia [206]. The first stage was to estimate the regulatory activities of TFs and miRNAs in each sample. In the second stage, the sample-specific TF and miRNA activities were then used as inputs to infer specific TF/miRNA-target interactions. In both steps, the TF-target, CNV, DNA methylation, and miRNA expression data were used as explanatory variables while the mRNA expression data was used as response variables. This study identified 18 predominant regulators (miRNAs and TFs). In Chapter 2 of this thesis, the Lasso regression model was used to explore the regulatory

mechanisms in different BC subtypes, therefore, more details about the use of Lasso regression in exploring gene regulatory programs would be discussed there.

1.5 Transcriptomics in precision medicine

In this section, we described how gene expression data have been used in the discovery of prognostic and predictive signatures (subsection 1.5.1) that have been used in the clinic. In subsection 1.5.2, the development of computational drug response prediction models in cancer research are examined.

1.5.1 Multigene prognostic assays for breast cancer

Transcriptome profiles of BC have been extensively studied to either define distinct molecular subgroups of this disease (studies of inter-tumor heterogeneity) or identify multigene expression signatures that could serve as a prognostic (and sometimes predictive) factor [56]. According to National Cancer Institute, a prognostic factor is defined as "a situation or condition, or a characteristic of a patient, that can be used to estimate the chance of recovery from a disease or the chance of the disease recurring (coming back)", while a predictive factor is defined as, "a condition or finding that can be used to help predict whether a person's cancer will respond to a specific treatment". Currently, several multigene expression signatures have been translated into clinical assays, including MammaPrint®, Oncotype DX®, Prosigna®, and Endopredict®.

MammaPrint® is 70-gene DNA assay developed by the company Agendia in the Netherlands. This test was first described in a study published by van 't Veer *et al.* in 2002 [207], which is the first study reporting the identification of a gene expression signature for BC. In this study, the microarray-based expression levels of 70 different genes were able to distinguish two groups from 78 BC patients with different disease outcomes, one group having a poor prognosis

and another group having a good prognosis [207]. Thereafter, this 70-gene signature was validated in a series of 295 consecutive patients with breast carcinomas [208], adapted to a small customized mini-array more suitable for clinical testing [209], and finally approved by the United States Food and Drug Administration (FDA) in 2007 [210]. The test is used for BC patients who are less than 61 years old, with stage I or stage II disease, with tumor size ≤ 5.0 cm and who are lymph node-negative to assess a patients' risk for distant metastasis. MammaPrint® was the first genomic assay approved by the United States FDA and the most widely used genomic assay for BC in Europe.

As the most widely used genomic assay for BC in the United States and Canada, the Oncotype DX® test (Genomic Health, United States) assesses the expression levels of 21 distinct genes using a reverse-transcriptase–polymerase-chain-reaction (RT-PCR) assay in paraffin-embedded tumor tissue and assigns each tested tumor a recurrence score (RS). The RS can estimate a woman's risk of recurrence of early-stage, hormone receptor-positive BC, as well as how likely she is to benefit from chemotherapy after BC surgery. The Oncotype DX® test was first described in the NSABP-B14 study for adjuvant tamoxifen-treated patients with ER+, lymph node-negative BC which was published by Paik *et al.* in 2004 [211]. In this study, 21 prospectively selected genes (five reference genes and 16 cancer-related genes) were used to develop the RS algorithm based on the expression data derived from a RT-PCR assay in paraffin-embedded tumor tissue [211]. The calculated RS, on a scale from 0 to 100, was used to stratify patients with ER+, lymph node-negative BC treated with adjuvant tamoxifen into three groups: the low-risk group ($RS < 18$), the intermediate-risk group ($18 \leq RS < 31$) and the high-risk group ($RS \geq 31$), indicating their risk of distant recurrence at 10 years. The 10-year distant recurrence rate was 6.8% for the low-risk group, 14.3% for the intermediate-risk group, and 30.5%, for the high-risk group [211]. In 2006, the NSABP-B20 study reported that the RS could predict the benefit of chemotherapy in patients

with ER+, lymph node-negative breast tumor [212]. In this study, patients with high-RS tumors ($RS \geq 31$) had a large benefit from chemotherapy while patients with low-RS tumors ($RS < 18$) had the minimal benefit [211]. The assessment of chemotherapy benefit was further assessed in the larger study SWOG 8814 for women with ER+ and lymph node-positive BC in which the RS was prognostic for ER+ and lymph node-positive BC treated with tamoxifen [213]. Also, patients with high RS tumors ($RS \geq 31$) benefited from CAF treatment (i.e., a chemotherapy regimen consisting of cyclophosphamide, Adriamycin, and fluorouracil) while patients with low RS tumors ($RS < 18$) may not benefit from anthracycline-based chemotherapy despite positive lymph nodes [213]. In the more recent TAILORx studies, the RS boundaries were modified to avoid under-treatment producing the low-risk group ($RS < 11$), the intermediate-risk group ($11 \leq RS < 26$) and the high-risk group ($RS \geq 26$) [214,215]. In addition, a 12-gene version Oncotype DX test has been shown to predict the risk of local recurrence in individuals with ductal carcinoma *in situ* [216,217].

The PAM50-based Prosigna® test (NanoString Technologies, United States) was approved by the United States FDA in 2013 [49]. In addition to BC intrinsic subtype assignment, the Prosigna® test can calculate a numeric score, which is called Risk of Recurrence (ROR) score, for a tested sample from the corresponding PAM50 gene signature expression profile in combination with the pathological tumor size as well as a proliferation score (computed from a subset of proliferative genes). The ROR score can be used to evaluate the risk of distant relapse for hormone receptor-positive, lymph node-negative or lymph node-positive early-stage BC in postmenopausal women [49]. A comparative study reported that the Prosigna® ROR score provided more prognostic information in endocrine-treated patients with ER+, lymph node-negative BC than the Oncotype DX® RS (i.e., more patients were scored as high risk and fewer as intermediate risk by

ROR score than by RS) [218]. Of note, both Oncotype DX® and MammaPrint® are performed in centralized (company-owned) laboratories while the Prosigna® test can be performed de-centrally on dedicated instruments.

There exist several other clinical tests available for BC, some of which are already approved by the United States FDA. We will introduce some of them here. The Endopredict® Test (Sividon Diagnostics, Germany) quantifies expression levels of 12 genes (eight cancer-related genes, three RNA reference genes, and one DNA reference gene) using quantitative reverse transcriptase PCR on formalin-fixed, paraffin-embedded tumor samples. The gene expression profile of the 12 genes together with tumor size and lymph node status of a patient is used to calculate a comprehensive risk score, indicating the risk of distant recurrence within 10 years of diagnosis for women with early-stage, ER+ and HER2- BC [219,220]. Although the EndoPredict® Test has currently not been approved by the FDA for marketing in the United States, it has been approved for use in Europe and can be performed in local laboratories on dedicated instruments. The MapQuant DX™ test (Ipsogen, France) is a DNA microarray-based assay that uses formalin-fixed, paraffin-embedded tumor samples to detect the gene expression pattern of the 97-gene Genomic Grade Index signature developed by Sotiriou *et al.* [221]. The test distinguishes low and high-grade breast tumors. The test reclassifies histological grade 2 breast tumors into low or high grade categories with low versus high risk of recurrence [221]. The Breast Cancer Index® (bioTheranostics, United States) is a 7-gene signature that combines the 2-gene ratio index (HOXB13/IL17BR) and the 5-gene molecular grade index [222]. The test is used to assess the risk of late (i.e., 5 to 10 years after diagnosis) recurrence and the potential benefit of a 10-year course of adjuvant endocrine therapy in women with early-stage, lymph node-negative and ER+ BC [223,224]. The test is not currently approved by the FDA for marketing in the United States. So

far, most of the tests we have described are using RT-PCR or microarray to quantify expression profiles of the multigene signatures. However, the tests MammoStrat® and IHC4 are based on the IHC technology and are thus not a true genomic test. The MammoStrat® (Clarient Diagnostic Services, United States) test is a 5-protein (CEACAM5, HTF9C, NDRG1, SLS7A5, and TP53) IHC assay. The test measures expression levels of the five protein markers in formalin-fixed, paraffin-embedded tumor samples to stratify patients with homone receptor-positive tumors receiving endocrine therapy into three groups with different risk of relapse and likelihood of benefit of adjuvant chemotherapy [225]. Application of MammoStrat® in the United States remains limited and it is not approved by the FDA. The IHC4 test assesses expression levels of ER, PR, HER2 and Ki-67 using IHC, which together with clinicopathologic factors are used to calculate a prognostic score for predicting risk of disease recurrence [226]. However, standardization of the IHC assessment to reduce inter-laboratory variability remains challenging, especially with regards to Ki-67. The applicability of the IHC4 test to clinical practice needs further investigation.

Except the several clinically available signatures we have described so far, there exist many other multigene signatures for BC prognosis. Today, over 100 gene signatures have been identified to show prognostic capacity in BC [56]. Although these signatures were derived from similar tumor cohorts for similar purposes and with similar prognostic performance, they had only a small gene overlap [227]. For example, both the 21-gene Oncotype DX® signature and the 70-gene MammaPrint® signature were initially developed for lymph node-negative patients, but were also found to be a good predictor of disease outcome in lymph node-positive patients. However, only one (*SCUBE2*) of the 16 cancer-related genes in the 21-gene Oncotype DX® signature was found to be included in the 70-gene MammaPrint® signature [228]. Only 8-gene overlap (*MELK, CENPA, CCNE2, GMPS, DC13, PRC1, NUSAP1, KNTC2*) was reported between the 70-gene

(MammaPrint®) and the 97-gene Genomic Grade Index (MapQuant DX™) prognostic signatures [228]. While the agreement in prediction with respect to predicting distant metastasis free survival for the individual patient was 88% for the two signatures [227]. This raised a question whether expression pattern of any randomly chosen gene sets is prognostic for BC [229]. Venet *et al.* compared 47 published BC signatures with 1,000 signatures with random constituent genes [229]. The results showed that 28 (60%) of the 47 published signatures did not performed better than random signatures of the same size [229]. Also, more than 90% of random signatures with over 100 constituent genes were prognostic [229]. The large degree of discordance in constituent genes of these signatures may be primarily due to the interrelatedness between the clinicopathological features, molecular tumor subtype, and prognosis of a patient [230]. To explain this phenomenon, Tofigh *et al.* performed a comprehensive and systematic analysis of the performance of 106 published BC prognostic signatures [56]. Regardless of the subtype in which the signature was originally derived, each of the 106 signatures was evaluated in different subtype classification schemes: the clinical classification scheme based on ER and HER2 IHC status, the intrinsic classification scheme by the PAM50 gene set, and the hybrid subtyping schemes stratifying each intrinsic subtype by ER or HER2 IHC status [56]. The study found that the prognostic capacity of signatures were confounded by ER and HER2 status, as stratification of a cohort by ER or HER2 status can reduce misprediction of outcome for those signatures [56]. Thus, the assumed ubiquity of prognostic genes has been greatly exaggerated and interpretation of prognostic signatures must account for breast tumor subtypes and other clinicopathological variables [56]. The discordance in the gene expression profiling platforms, training sets, and/or bioinformatics and statistical tools may also contribute to the small gene overlap between signatures [231]. For example, Vliet *et al.* reported that pooling multiple BC datasets can lead to more accurate prediction of outcome and

the limited overlap of signature genes can be attributed to small sample size [232]. The small degree of overlap between the genes of these signatures is also partially explained by the assumption that such similar disease outcome predictions are based on overlapping underlying biological processes, which is supported by several studies. For example, Thomassen *et al.* reported cell cycle and cell proliferation as being the main overlaps in gene ontology annotations of nine existing prognostic signatures [233]. Yu *et al.* also found that five published prognostic gene signatures shared many common pathways (such as cell cycle, DNA replication, and mitosis) [228]. Reyal and van Vliet *et al.* complied the genes from nine published prognostic gene signatures and found that a core set of genes were enriched in DNA replication, cell cycle and mitosis ontology annotations [234]. Recently, we conducted an analysis to compare genes and biological pathways from 33 published multigene signatures. Out of 1,895 unique genes obtained from all these signatures, 9% (173) genes were shared by any two signatures, 2% (41) genes were shared by any three signatures, 0.7% (14) genes were shared by any four signatures, 0.3% (6) genes were found in any five signatures, 0.2% (4) genes were found in any six or more signatures. Over-representation analysis was performed on gene set from each signature to identify the function terms and resulted in a total 988 unique function terms for all the signatures. Comparison of function terms from all signatures showed that 20% (195) terms were shared by any two signatures, 10% (99) terms were shared by any three signatures, 6% (58) terms were shared by any four signatures, 4% (39) terms were shared by any five signatures, 0.9% (9) terms were shared by any six signatures, and 3% (29) terms were shared by any seven or more signatures. Therefore the lack of gene overlap in multi-gene signatures derived from different research studies can be partially explained by our discovery that the signature genes are associated with similar functions and/or pathways despite being distinct individual genes.

1.5.2 Drug response prediction models

One of the key goals of precision medicine is to predict how a cancer patient will respond to a particular chemotherapy or targeted therapy, which could help clinicians prescribe the most effective and least toxic therapeutic strategy. Researchers have been developing anti-cancer drug response prediction models *in silico* to address this problem. However, it is not feasible to screen a large number of drugs on large populations of cancer patients. Thus, there are not enough patient-derived (*in vivo*) drug response data to train computational anti-cancer drug response prediction models. To overcome this limitation, high-throughput drug screening approaches have been performed on cancer cell lines to provide *in vitro* response data of hundreds of cell lines to many drugs, such as the GDSC [235,236] and Cancer Cell Line Encyclopedia (CCLE) [237,238]. Though concerns regarding the substantial inconsistency between the two independent large pharmacogenomic databases in terms of drug sensitivity measurements have been raised [239,240], other research has highlighted the concordance at different analysis levels of GDSC and CCLE for meaningful analyses [241–244]. For example, the joint study by the CCLE and the GDSC investigators demonstrated a considerable consistency of drug sensitivity measurements as well as drug sensitivity prediction biomarkers (such as gene expression, CNVs and mutations) between the CCLE and GDSC data sets [244]. Therefore, multiple resources have been designed to enable discovery of biomarkers predictive of drug response or integrate pharmacogenomic data from many different cell line databases [245–247]. For example, Smirnov and colleagues developed PharmacoDB which is an integrative database providing access to molecular profiles and drug sensitivity measurements of cell lines from different data sets (such as GDSC and CCLE) [245]. PharmacoDB provides access to PharmacoGx, an R package also developed by Smirnov *et al.* [246], for identifying biomarkers by analyzing the relationships between molecular features

and drug response measurements. Furthermore, the cell line pharmacogenomic data sets have been widely used for *in silico* drug response prediction model development.

Typically, models are developed to make sensitivity predictions for multiple single drugs within cell lines from various cancer types (i.e., pan-cancer) and referred to as pan-cancer models [248]. To date, most of these models have been reported to use gene expression data as input, which have been shown to be the most informative data for drug response prediction in cancer research [236,249–251]. For instance, using the GDSC data, Geeleher *et al.* [249] fitted ridge regression models for single drugs using baseline gene expression levels in the pan-cancer cell lines against the *in vitro* drug 50% inhibitory concentration (IC_{50}) estimates. These multiple single-drug models were independently tested on three publicly available datasets from clinical trials in myeloma, BC and non-small-cell lung cancers [249]. They also applied this ridge regression-based model to the gene expression profiles of the TCGA tumors to impute the drug response [252]. Geeleher *et al.* [249,252] showed that their cell line-derived models can be used for the accurate prediction of drug response in patients. Using CCLE, Dong *et al*. [253] proposed a support vector machine classification model to predict drug response of various cancer cell lines using gene expression data as input features. The model was developed for each single drug and validated on GDSC, which achieved good performance for several drugs [253]. Approaches for predicting drug response of cancer cells can be divided into regression (such as elastic net, ridge, and kernelized regression) and classification (such as support vector machines, random forests, and k-nearest neighbors) models. Comparative analysis suggested that performance of these models differs in terms of settings (such as datasets and drugs) [254]. Besides regression and classification methods, network-based methods have been investigated in drug response prediction field. A classic example is the dual-layer network proposed by Zhang and colleagues, which integrated a cell line

similarity network and drug similarity network [255].

Although gene expression data have been most widely used in modeling drug response, combining several omics data types (e.g. gene mutation, CNVs and DNA methylation) has been reported to improve the predictive power [236]. In the Ding *et al.* study [256], the multi-omics cell line data (mutation, CNV, and gene expression data) from GDSC were concatenated then features were extracted from the concatenated data using deep neural network autoencoder. Taking the learned features as the input, an elastic net regression model was built for each drug. In Sharifi-Noghab *et al.*'s study [257], a deep learning-based model was proposed to take gene somatic mutation, CNV, and gene expression data of a particular cell line as input, and predict the response to a given drug as the output. The model was trained on GDSC and further validated on patient-derived tumor xenograft and patient datasets. In addition, the physical, chemical and pharmacological properties of drugs such as chemical structure, aqueous solubility and potency (IC_{50}) also play important roles in drug response prediction. Thus, computational models by combining genomic profiles with information about the drug's chemical structure would improve drug response prediction *in vitro* and *in vivo* [258,259]. Ammad-ud-dinet *et al.* [259] utilized cell line information and drug chemical properties as additional sources of information using selective data integration.

Other models have focused on making predictions for drug response in a specific cancer type are referred to as cancer-specific models [248]. BC-specific models have been investigated. For instance, Daemen *et al.* generated drug sensitivity data for 138 drugs and molecular profiles (pre-treatment measurements of mRNA expression, CNVs, protein expression, promoter DNA methylation, and gene mutation) from 70 BC cell lines [260]. They then used two independent classification approaches, support vector machine and random forest, to identify pre-treatment

molecular features that could distinguish responders from non-responders for each drug within the cell line panel. They further validated the model on two drugs (tamoxifen and valproic acid) in independent TCGA breast tumors. Later, Costello *et al.* [261] launched the most famous work on this problem - the National Cancer Institute-DREAM Drug Sensitivity Prediction Challenge, which obtained data from Deamen *et al.*'s study [260] and was to integrate multiple-omics measurements and predict drug sensitivity in BC cell lines. Among the 44 drug response prediction algorithms compiled from participating researchers, the Bayesian multitask multiple kernel learning method, which integrated four machine learning principles (kernelized regression, multi-view learning, multi-task learning, and Bayesian inference), was the top-performing one. Due to a limited number of BC-specific drug response models, more investigation is needed. In Chapter 4 of this thesis, a network-based model was proposed to predict anti-cancer drug response of BC.

It is important to point out that although the pharmacogenomic datasets derived from cancer cell lines allowed for discovery of biomarkers or development of more complex models to predict drug response, they still suffer from multiple limitations due to the major concern about the reliability of using cancer cell lines as a cancer model [262]. To this end, patient-derived tumor xenografts and organoids, which more accurately recapitulate the heterogeneity and microenvironment of the origin patient tumor than any cell line model, have emerged as reliable preclinical models for studying cancer drug response [263–266]. Recently, Mer and colleagues developed the freely available Xenograft Visualization and Analysis (Xeva) R package for identifying genomic markers (such as gene expression, CNVs, and mutations) of drug response with patient-derived tumor xenograft-based pharmacogenomic data, which would be beneficial to the research community for better understanding associations between molecular changes and drug response [267].

1.6 Data sets

This section introduces the primary data sets used in this thesis, including the gene set and pathways databases (subsection 1.6.1) used for pathway-level analysis, the public gene expression repositories (subsection 1.6.2) where gene expression data of BC samples could be retrieved, the primary breast cancer cohort data (subsection 1.6.3) which are multi-omics data, as well as the breast tumor cell line data (subsection 1.6.4) used for identifying drugs modulating dysregulated pathways or predicting anti-cancer drug response.

1.6.1 Gene set and pathway databases

To perform pathway analysis, researchers have developed a number of databases that group genes in terms of specific properties (such as functions, chromosome locations) into pathways or gene sets. Databases, such as GO [268,269], KEGG [105], Reactome [106,107], Bio Carta [108], Pathway Interaction Database (PID) [109], and MSigDB [104], are among the most well-known ones. GO[268,269] provides structured knowledge of gene/gene product functions using formal ontologies, which usually consist of a set of classes of gene functions (i.e., GO terms) with specific relations that exist among them. GO terms are described based on three biological function domains: the molecular function of the gene, its subcellular location (cellular component), and biological process (a larger biological network in which the gene functions). In the latest GO database, there are about 45,000 terms (29,698 biological processes, 11,147 molecular functions and 4,201 cellular components), which are linked by almost 134,000 relations.

KEGG [105], as one of the first pathway databases, has been widely used. KEGG pathways are curated from experimental knowledge on gene functions. Each pathway is designed to link genes to gene products in the pathway based on their interactions and reactions to form a network. In KEGG, pathways are represented by manually drawn graphical diagrams. KEGG

pathways include several types, such as metabolism, genetic information procession, environmental information processing, cellular processes, organismal systems, human diseases, and drug development.

Reactome [106,107], a database focusing on a single species, *Homo sapiens*, curated pathways by linking human proteins to their molecular functions, such as signal transduction, transport, DNA replication, metabolism, and other cellular processes. In the current Reactome version (v69, released in May 2019), 11,009 proteins and modified forms of proteins encoded by 10,833 genes, 1,857 small molecules, and 196 drugs, which are linked by 12,505 human reactions, are organized into 2,272 pathways.

BioCarta [108] pathways are curated based on how genes interact and involve over 120,000 genes from multiple species. In BiocCarta, pathways are graphical representations of common metabolic pathways, signal transduction pathway, as well as other biochemical pathways.

PID [109] is a curated collection of human signaling and regulatory pathways. They are networks of events (e.g., gene regulation, transport, small-molecule conversion, and protein-protein interactions,) connected by the participant molecules (e.g., small molecules, RNA, proteins and complexes).

MSigDB [104], originally designed for preforming GSEA, is now one of the largest and most widely used gene set database. There are currently 17,810 gene sets in MSigDB (v6.2, updated July 2018), which can be categorized into 8 collections: 1) 326 positional gene sets which are generated by grouping genes located in the same chromosome or cytogenetic band; 2) 4,762 curated gene sets that are collected from various pathway sources (such as KEGG, Reactome, BioCarta, and PID); 3) 836 motif gene sets which are defined by genes sharing *cis*-regulatory motifs in their promoter (TF targets) or 3'-UTR (miRNA targets) sequences; 4) 858 computational

gene sets which include genes grouped by computational analysis of large cancer-oriented gene expression data; 5) 5,917 GO gene sets that are corresponding to the same GO terms in the GO database; 6) 189 oncogenic signatures which are gene sets representing signatures of typically dysregulated cellular pathways in cancer; 7) 4,872 immunologic signatures which are gene sets representing cell states and perturbations within the immune system; 8) 50 hallmark gene sets which are generated by reducing both variation and redundancy of gene sets from the other collections. All gene sets in MSigDB are represented as lists of human gene symbols from the HUGO Gene Nomenclature Committee[270] and have been reviewed, curated, and annotated manually by the MSigDB curation team. In Chapter 2 of this thesis, MSigDB was used for GSEA to identify differentially expressed pathways in BC.

1.6.2 Public gene expression repositories

Gene Expression Omnibus (GEO) [271] at the National Institutes of Biotechnology Information (NCBI) and ArrayExpress [272] at the European Bioinformatics institute are two major public archives of microarray based gene expression data from tens of thousands of studies worldwide. When accepting data submissions from various experiments, both the two repositories adopt the same reporting standards - the Minimal Information about a Microarray Experiment[273], which enable sufficient and unambiguous annotation of experimental data and thus reuse of public gene expression data. As of 10 August 2019, ArrayExpress contains data from approximately 72,290 experiments with more than two million assays while GEO contains data from approximately 116,571 public series (experiments) with about 19,974 platforms. Though the GEO and ArrayExpress archives are not synchronized, there is a big overlap between the two databases given that ArrayExpress imports data from GEO on a weekly basis. More recently, both two databases started to accept submission of sequence-based gene expression experiments.

1.6.3 Primary breast cancer cohort data

High-throughput methodologies have enabled the generation and public availability of multi-omics data for large cancer cohorts, with TCGA [10] and the International Cancer Genome Consortium [274] being the best-known. In this thesis, we used BC data from the TCGA cohort. In TCGA, multi-omics experiments, in which the same sample is characterized by several methods to generate whole exome sequencing data, genomic CNV data, DNA methylation data, microarray or RNA-Seq based mRNA expression data, RNA-Seq based miRNA expression data, protein expression data, and clinical metadata for the sample, have been performed on primary tumor as well as normal samples. TCGA contains multi-omics data for each of over 20,000 tumor and normal samples collected from over 11,000 patients distributed among 33 different tumor types. The TCGA data can be retrieved from many secondary portals, such as the Firehose Broad GDAC (https://gdac.broadinstitute.org/) and UCSC Xena (https://xena.ucsc.edu). The multi-dimensional omics data are available for over 1,000 breast tumor samples and over 100 adjacent normal breast tissue samples.

Another primary breast tumor dataset used in this thesis is from the METABRIC study [58], which is a joint Canada-UK project with the purpose of further classifying breast tumors using molecular signatures for improved precision medicine. Multi-omics data, including CNV profiles, SNP genotypes, and microarray gene expression data, as well as the clinical traits were generated for each of the normal and tumor samples in the METABRIC study. In the original study, a total of 1,986 samples were divided into two subsets: 997 samples in the *Discovery* subset and 989 samples in the *Validation* subset.

1.6.4 Breast tumor cell line data

Four cell line-derived datasets have been used in this thesis, including the Genomics of GDSC [235,236], CCLE [237,238], Cancer Therapeutics Response Portal (CTRP) [275,276], and CMAP [277]. The first three (i.e., GDSC, CCLE, and CTRP) provide the drug response data and the baseline gene expression profiles of different cancer cell lines whereas the last one (i.e., CMAP) provides drug-induced gene expression changes in cancer cell lines.

In the GDSC release 7.0 (ftp://ftp.sanger.ac.uk/pub4/cancerrxgene/releases/release-7.0, March 2018), approximately 1,001 cell lines from various human cancer types have been characterized at multiple levels, including whole-exome sequencing, SNP array CNV profiles, microarray gene expression profiles, DNA methylation, gene fusions, and microsatellite instability. In parallel, these cancer cell lines have been screened with 267 compounds targeting 24 pathways to fit the drug-response curves. Linking the cell line drug sensitivity to the detailed genomic and transcriptional information allows researchers to discover new therapeutic biomarkers. The current release of GDSC (release 8.0, July 2019) has increased the number of compounds to 453 which have been screened across the same 1,001 cancer cell lines. There are over 50 breast tumor cell lines among the 1,001 cell lines.

The initial CCLE study [237] characterized 947 human cancer cell lines comprising 36 tumor types, including the mutations of 1,650 genes, CNVs, and baseline gene expression profiles. 479 out of the 947 cell lines were treated with 24 anti-cancer compounds to generate dose-response curves. About 40 breast cancer cell lines are included in the 947 CCLE cell lines. The current CCLE database [238] has expanded the characterizations of cancer cell lines to include RNA-Seq gene expression data for 1,019 cell lines, whole-exome sequencing data for 326 cell lines, whole-genome sequencing data for 329 cell lines, protein expression data for 899 cell lines,

DNA methylation data for 843 cell lines, miRNA expression data for 954 cell lines, and global histone H3 modification profiling data for 897 cell lines. However, no extra pharmacologic profiles were generated.

The first release of CTRP (v1, http://portals.broadinstitute.org/ctrp.v1) contained the sensitivity of 242 cancer cell lines to an "Informer Set" of 354 small molecules which perturb targets and processes required for maintaining cancer cell growth and also the annotations of the 242 cancer cell lines including microarray gene expression profiles, SNP CNV profiles, and gene mutation data. The subsequent release (v2, http://portals.broadinstitute.org/ctrp.v2.1/) increased the number of cell lines and molecules 860 and 481, respectively. Over 40 BC cell lines are included CTRP v2.

CMAP [277] is the first large-scale project providing microarray-based transcriptome profiles from cultured human cancer cell lines treated with or without bioactive small molecules. The first version of CMAP (build 01) [277] consisted of 564 microarray gene expression profiles generated from applying 164 small molecules to five cell lines (MCF7, human breast epithelial adenocarcinoma cell line; PC3, human prostate epithelial adenocarcinoma cell line; HL60, human promyelocytic cell line; SKMEL5, human malignant melanoma cell line; and ssMCF7, phenol red free MCF7 cell line). The current version of CMAP (build 02, https://portals.broadinstitute.org/cmap) contains more than 7,000 expression profiles by increasing the number of compounds to 1,309, which have been tested in the same five cell lines. Moreover, CMAP provides an expression-pattern matching algorithm by using gene-expression changes upon drug treatment to connect: 1) drugs to drugs for identifying drug similarities; 2) drugs with diseases for identifying potential treatments [278].

1.7 Thesis Roadmap and Chapter Summaries

High-throughput technologies have enabled the characterization of thousands of BC samples at multiple levels and generated large amounts of multi-omics data, especially gene expression data, for individual BC samples. These *in vivo* (such as TCGA and METABRIC) and *in vitro* (such as GDSC, CCLE, and CMAP) gene expression data allow researchers to explore the BC transcriptome. The aim of the whole thesis was to explore the dysregulated pathways and underlying regulatory mechanisms, as well as the implications for drug treatment of the BC transcriptome.

The gene expression profiles differ among BC subtypes, but the underlying regulatory mechanisms have yet to be fully elucidated. Thus, we studied the transcriptional regulators responsible for gene deregulation in each BC subtype in Chapter 2. We hypothesized that the gene expression changes are influenced by CNV, DNA methylation, TF deposition and miRNA-mRNA interaction and the influence of these regulators on gene expression changes could be fit by sample-specific Lasso regression models. The aim of Chapter 2 was to find the key regulators responsible for the observed gene deregulation and cancer pathogenesis in specific BC subtypes.

Chapter 3 aimed to identify the compounds that could modulate the activities of the top dysregulated pathways (i.e., the FOXM1 and the PPARA pathways) in BC. We hypothesized that the signalling activity of a given pathway could be inferred from the expression levels of genes involved in the pathway. Using the CMAP database, we calculated the activities of the two pathways in MCF7 cells receiving a particular treatment as well as in the paired controls in order to identify compounds that can change the pathway activities.

Transcriptomics data have been reported as the most informative data for anti-cancer drug response prediction among the various omics data. But these models typically were developed with

individual genes. Instead, we used the pathway activity profiles to model drug response in Chapter 4. We hypothesized that similar BC cell lines measured by their pathway activity profiles have similar drug response and similar drugs estimated by their structure, targets, and pan-cancer IC_{50} profiles have similar effects on BC cell lines. By integrating a one-layer cell line similarity network based on the pathway activity profiles and a three-layer drug similarity network based on three types of drug information, Chapter 4 aimed to a develop a multiple-layer cell line-drug response network model with good prediction performance (e.g., high Pearson correlation) on *in vitro* data that also has good performance on *in vivo* data.

Chapter 2: Integrative Analysis of Subtype-Specific Regulation of the Transcriptome in Breast Cancer

This chapter is modified from the published paper below. I produced all analyses, plots, tables, and written text in this thesis myself.

Huang Shujun, Xu Wayne, Hu Pingzhao, Lakowski Ted*. Integrative Analysis Reveals Subtype-Specific Regulatory Determinants in Triple Negative Breast Cancer. Cancers. 2019 Apr;11(4):507. (*denotes equal contribution)*

2.1 Introduction

BC is a heterogeneous disease with different intrinsic subtypes displaying different gene expression patterns [4,32–34]. But the regulatory mechanisms underlying subtype-specific expression profiles in BC have yet to be fully elucidated. Changes in gene expression resulting from alterations in miRNA expression, DNA methylation, dysregulation of transcription factors, and DNA copy number have been implicated in breast carcinogenesis [279]. miRNAs regulate target mRNA translation and degradation by binding to the three prime untranslated region (3′-UTR) of mRNA [170]. The dysregulation of miRNA is associated with BC initiation and progression [279]. In cancer, DNA methylation is perturbed, leading to significant changes in gene expression. For example, aberrant promoter CpG island hypermethylation has been reported in multiple genes associated with BC, typically leading to gene silencing [279]. Genomic instability due to CNVs is also associated with altered gene expression in BC [280]. TFs are key cellular components in controlling gene expression and thus play a central part in controlling many biological processes in cells. Aberrant activity of TFs, frequently caused by genetic alterations in

genes that encode the corresponding TFs, are known to be involved in the tumorigenesis of numerous cancer types including BC [182,183,281]. Research on the regulatory mechanisms of gene expression changes that precipitate BC should take into consideration the interaction of the multiple factors enumerated above. High-throughput approaches enable the investigation of a tumor at multiple levels. For example, the TCGA database characterized a given tumor sample on six major platforms to generate multi-omics data for the sample such as the mRNA/miRNA expression, CNV and DNA methylation profiles [10]. The multi-omics data for breast tumor samples in TCGA, together with miRNA-target and TF-target interaction data provided by other databases (such as StarBase v3.0 [189] and TRRUST v2.0 [192]), allows one to systematically investigate the regulatory mechanisms in BC by integrating multiple regulatory factors.

Considering the multiple factors regulating gene expression presents a challenge in terms of combining data from different experiments to generate biologically meaningful and experimentally testable models. Network-based models that analyze individual genes and their interactions as well as upstream regulatory mechanisms are effective approaches to overcome this challenge [197]. For example, Xu *et al.* proposed a weighted similarity network fusion model to utilize the information from miRNA-TF-mRNA regulatory networks to divide the TCGA breast tumor samples into five groups with different expression patterns [282]. The method works as follows: the interactions among the miRNAs, TFs and mRNAs retrieved from various databases were used as inputs to construct a regulatory network where the nodes represent features (miRNAs, TFs or mRNAs) and the edges represent the interactions between the them [282]. The gene (mRNAs, TFs) or miRNA expression data in each sample, together with the ranking of each feature calculated based on the regulatory network information, were used to construct two weighted similarity networks (one based on genes and the other based on miRNA) [282]. The two similarity

networks were finally integrated through a network fusion approach and clustering analysis was performed to stratify patients into different groups that corresponded to different gene expression patterns with different underlying regulatory programs [282]. Such an integrative strategy using regression models to combine data at multiple levels can also provide insight into gene expression dysregulation in cancer. Using the multi-omics data from TCGA, Setty and colleagues developed a sample-specific Lasso regularized linear regression model to identify key regulatory factors in the different subtypes of glioblastoma multiforme [205]. The model was fit to log fold changes in mRNA expression from glioblastoma multiforme tumor samples (relative to normal brain samples) as the response variable using a linear combination of CNV, DNA methylation, miRNA-binding site counts in mRNA-3′UTR, and TF-binding site counts in gene promoter (filtered by DNA hypersensitive regions from the ChIP-Seq data) as covariates [205]. The study developed the model on a tumor-by-tumor basis and identified miRNAs and TFs as either common or subtype-specific drivers of expression changes [205]. The estimated activities of these key regulators could predict glioblastoma multiforme subtypes and patient survival outcomes [205].

Inspired by Setty *et al.*'s study, a similar sample-specific Lasso regression model was proposed in this study to explore the regulatory mechanisms in the different subtypes of BC. In Setty et al.'s study, they treated each of human genes as a training example when training the Lasso model for a specific tumor and thus identified the key regulators (i.e., miRNAs and TFs) accounting for global gene expression changes in glioblastoma multiforme relative to normal brain samples [205]. In the current study, instead of taking all human genes into consideration, the DEGs in breast tumors relative to the adjacent normal breast tissue samples were focused on. Assuming that tumor versus normal differential expression in DEGs (output variable) is influenced by variation in copy number, DNA methylation, TF deposition and miRNA-mRNA interaction (input

variables) in a given sample, a Lasso regularized linear regression model specific for the tumor sample was built. With all built tumor-specific models, a feature scoring scheme [205] was then used to determine the key regulators (in terms of their statistical significance) accounting for gene expression changes in different BC subtypes (luminal A, luminal B, HER2-enriched and TNBC). The follow-up exploration was focused on key regulators identified in the TNBC subtype, including investigating target pathways potentially regulated by these regulators, developing prognostic signatures using these regulators or their target genes, and constructing a regulatory network where the nodes represent these regulators or their target genes and the edges represent the interactions between them.

2.2 Materials and Methods

2.2.1 TCGA multi-omics data collection

The overall Lasso approach used in this chapter is depicted in **Figure 2.1**. The TCGA multi-omics data for the BC cohort were used in this study (**Figure 2.1a and 2.1b**). The level 3 processed data including mRNA/miRNA expression profiles, DNA methylation and CNV data, were retrieved for breast tumor samples from the UCSC Xena browser (http://xena.ucsc.edu/). Only mRNA/miRNA expression data were downloaded for normal samples. Quantified mRNA expression data in the form of RNA-Seq by Expectation-Maximization (RSEM) [92] values were transformed to log2(RSEM+1) to estimate the gene-level transcription. miRNA mature strand expression data by Illumina Hiseq 2000 were also included. For each sample, all isoform expression in reads per million miRNAs mapped (RPM) values for the same miRNA mature strand are added together and the log2(total_RPM+1) transformation was applied to estimate the miRNA expression level. miRNA IDs from miRBase release 22.1 (October 2018,

ftp://mirbase.org/pub/mirbase/22.1) were used. The mRNA and miRNAs with missing values in more than 50% of the tumor and normal samples were removed. For DNA methylation data, the Illumina methylation450 beta values were used as the methylation scores. When multiple scores were available for the same gene, the average of those scores were used. For CNV data, the whole genome microarray data which have been estimated at the gene-level using the Genomic Identification of Significant Targets in Cancer 2.0 (GISTIC 2.0) method [283] were used. In addition, clinical data for BC patients were downloaded from the Firehose TCGA data portal (http://firebrowse.org/). Patients with the IHC status for the three receptors ER, PR, HER2 were categorized into four molecular subtypes: (1) luminal A (LumA: ER+ and/or PR+ and HER2-); (2) luminal B (LumB: ER+ and/or PR+ and HER2+); (3) HER2-enriched (Her2E: ER- and PR- and HER2+); (4) TNBC (ER- and PR- and HER2-).

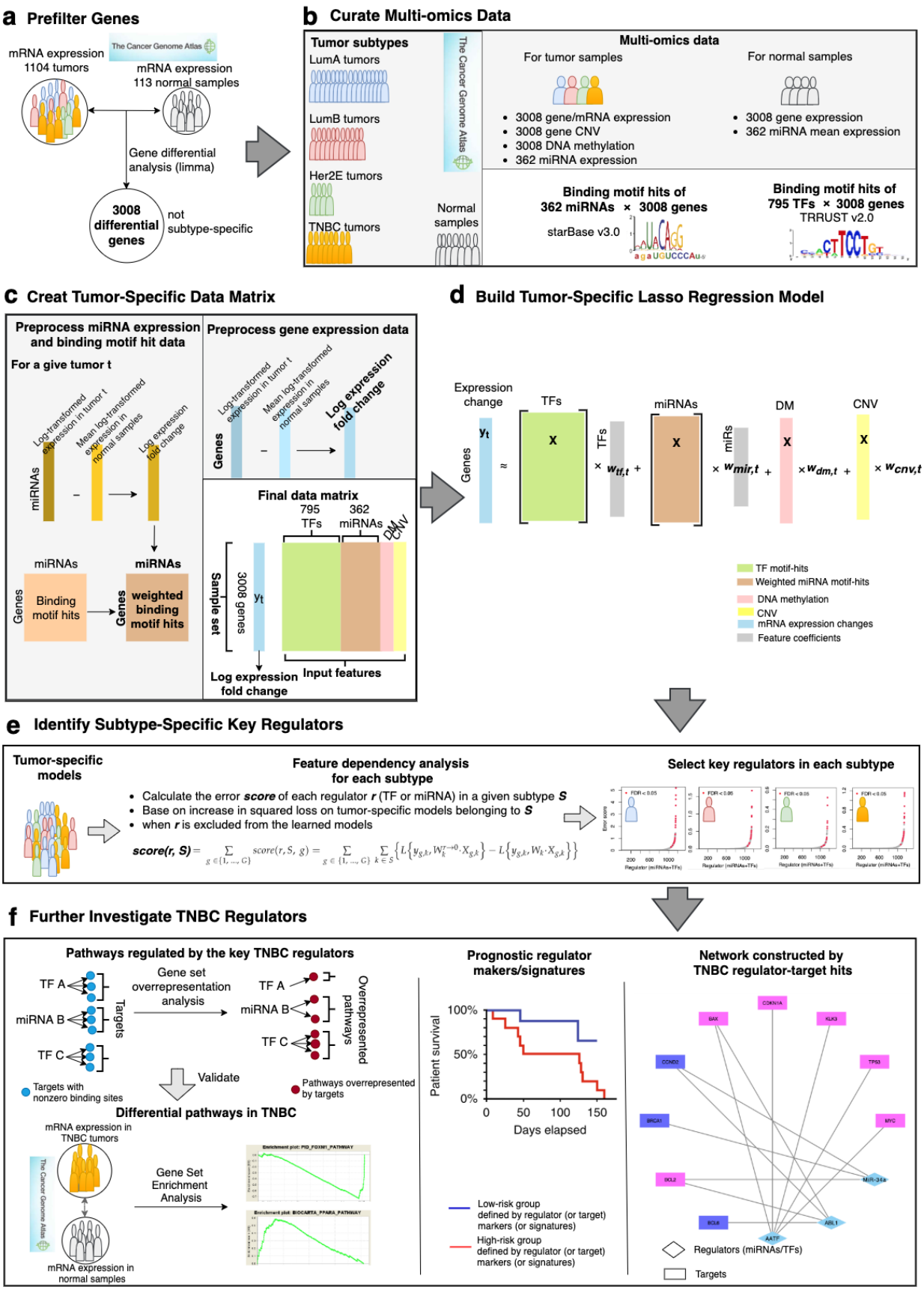

a Prefilter Genes
mRNA expression 1104 tumors
The Cancer Genome Atlas
mRNA expression 113 normal samples
Gene differential analysis (limma)
3008 differential genes
not subtype-specific
b Curate Multi-omics Data
Tumor subtypes
LumA tumors
LumB tumors
Her2E tumors
TNBC tumors
The Cancer Genome Atlas
Normal samples
Multi-omics data
For tumor samples
For normal samples
3008 gene/mRNA expression
3008 gene CNV
3008 DNA methylation
362 miRNA expression
3008 gene expression
362 miRNA mean expression
Binding motif hits of 362 miRNAs × 3008 genes
starBase v3.0
Binding motif hits of 795 TFs × 3008 genes
TRRUST v2.0
c Creat Tumor-Specific Data Matrix
Preprocess miRNA expression and binding motif hit data
For a give tumor t
Log-transformed expression in tumor t
Mean log-transformed expression in normal samples
Log expression fold change
miRNAs
Genes
Binding motif hits
weighted binding motif hits
Preprocess gene expression data
Final data matrix
795 TFs
362 miRNAs
DM
CNV
3008 genes
Sample set
Log expression fold change
Input features
d Build Tumor-Specific Lasso Regression Model
Expression change
TFs
miRNAs
DM
CNV
genes
miRs
TF motif-hits
Weighted miRNA motif-hits
DNA methylation
CNV
mRNA expression changes
Feature coefficients
e Identify Subtype-Specific Key Regulators
Tumor-specific models
Feature dependency analysis for each subtype
Calculate the error score of each regulator r (TF or miRNA) in a given subtype S
Base on increase in squared loss on tumor-specific models belonging to S
when r is excluded from the learned models
Select key regulators in each subtype
f Further Investigate TNBC Regulators
Pathways regulated by the key TNBC regulators
TF A
miRNA B
TF C
Targets
Gene set overrepresentation analysis
Overrepresented pathways
Validate
Targets with nonzero binding sites
Pathways overrepresented by targets
Differential pathways in TNBC
mRNA expression in TNBC tumors
mRNA expression in normal samples
Gene Set Enrichment Analysis
Prognostic regulator makers/signatures
Patient survival
100%
50%
0%
0
50
100
150
Days elapsed
Low-risk group defined by regulator (or target) markers (or signatures)
High-risk group defined by regulator (or target) markers (or signatures)
Network constructed by TNBC regulator-target hits
CDKN1A
BAX
KLK3
CCND1
TP53
BRCA1
MYC
BCL2
MIR-34a
BDNF
ARL1
AATF
Regulators (miRNAs/TFs)
Targets

Figure 2.1 Overview of the study design
(**a**) Gene differential analysis was performed between breast tumors and normal breast tissue samples to prefilter differential genes, resulting in 3008 genes. (**b**) Multi-omics data regarding the 3008 genes were collected from TCGA breast tumor cohort, including the mRNA expression, CNV and DNA methylation profiles. Also, miRNA expression profiles were extracted for TCGA breast tumor cohort. Binding-motif hits between the 3008 genes and 362 miRNAs were extracted from StarBase and binding-motif hits between the 3008 genes and 759 TFs were extracted from TRRUST. (**c**) Data matrix was created for each breast tumor. First, the miRNA expression log fold change was calculated for each tumor relative to the normal breast tissue samples and was further applied to the miRNA-target interaction data, resulting in weighted miRNA-target binding-motif hit data. Second, mRNA expression log fold change was calculated for each tumor relative to the normal breast tissue samples. Finally, multiple types of data were merged together to create two matrices for each tumor: one is a vector matrix containing the 3008 gene expression log fold changes, another one is a 3008 by 1159 matrix with rows being the 3008 genes and columns being different features. (**d**) To infer dysregulated regulatory programs from tumor profiling data, the change in gene expression in a tumor sample is modeled as a linear function of the following input variables: TF binding-motif hits, weighted miRNA binding-motif hits, DNA methylation, CNV. The linear model is trained for each tumor. (**e**) A feature scoring method based on these regression models identified common and subtype-specific regulators. (**f**) The follow-up exploration was performed on TNBC regulators, including investigating potential target pathways TNBC regulators (the left panel), evaluating the clinical relevance of TNBC regulators or their targets (the middle panel), and constructing the regulatory network involving TNBC regulators and their potential target pathways (the right panel).

2.2.2 Target prediction for miRNAs and TFs

As described in **Figure 2.1b**, we collected interaction data for TF-target pairs and miRNA-target pairs as well. For each miRNA-mRNA pair with measured expression in the TCGA data we obtained binding-motif hit information from the StarBase v3.0 database [189]. StarBase v3.0 database contains a large number of publicly available miRNA-target pairs predicted by seven algorithms (PITA, RNA22, miRmap, microT, miRanda, PicTar, TargetScan). In this study, to include as many miRNA-target pairs as possible, those miRNA-target pairs predicted by at least one algorithm were selected.

TF target genes were determined using TRRUST v2.0, where target predictions for human sequence-specific TFs based on motif hits were downloaded [192]. The binding-motif hits for TF-target pairs were used to build the tumor-specific model in this study. We also built the model based on ChIP-Seq derived TF binding data from ReMap v1.2 as a comparison. [194]. ReMap contains a catalog of regulatory regions by compiling the genomic localization of 485 different transcriptional regulators across 346 different human cell lines and tissue types. This database is

composed of 1,066 ChIP-Seq datasets from Encyclopedia of DNA Elements (ENCODE) [284] and 1,763 datasets from Gene Expression Omnibus (GEO) [271] and ArrayExpress [272]. In total, ChIP-Seq data were selected for 125 TFs across 13 diverse untreated BC cell lines. The promoter region for each gene was defined as 2,000 bp upstream to 200 bp downstream from transcription start site based on *homo sapiens* (GRCh38/hg38) annotation. The peaks were then annotated within promoters for each TF using the R package ChIPpeakAnno [285] and the TF binding scores averaged for multiple peaks to represent a single binding site per TF-gene pair.

2.2.3 Prefiltering genes

In this study, DEGs were focused on to develop the tumor-specific Lasso regression model. DEGs were determined using the R package limma-voom [286] with the gene expression profiles for tumor (n = 1,104) and normal breast tissue (n = 113) samples from TCGA (**Figure 2.1a**). A total of 3,008 DEGs with FDR less than 0.05 and an absolute value of log2 fold change not less than 1 and were selected with data available across multiple platforms and for model development.

2.2.4 Developing the tumor-specific Lasso regression model

Preselected multi-omics data regarding the 3008 DEGs for each breast tumor were curated (**Figure 2.1c**). Linear regression models were trained separately for each tumor sample to explain log2 gene expression fold changes (tumor versus normal tissue) using CNV, TF binding site counts in the gene's promoter, and miRNA binding site counts in the gene's 3'UTR as covariates in the models (**Figure 2.1d**). Tumor-specific miRNA expression in fold changes (tumor versus normal tissue), were used to restrict the miRNAs that can be used as explanatory variables. Specifically, suppose there are N tumor samples, G DEGs (g), M miRNAs, and T TFs. For each sample n, we observe the log2 expression fold change in the tumor (relative normal tissue), CNV, DNA methylation value for gene g as well as the log2 expression fold change in the tumor (compared to

normal tissue) for all the M miRNAs (together provided by TCGA). The binding site count for TFs or miRNAs, however, were not available for each sample but rather estimated by sequence-derived methods using TRRUST and starBase, respectively. After preselection and curation, in this study, $N = 391$, $G = 3008$, $M = 362$, and $T = 795$.

To avoid overfitting in the presence of noisy expression data and a large number of explanatory variables, the regularized regression via a Lasso constraint [287] was used to identify a small number of TFs and miRNAs that best explain changes in gene expression on a sample-by-sample basis. The Lasso constraint enforces sparsity in the learned parameters, with the result that most of the regression coefficients are zero. This reduces the number of features included in the model, leading to better prediction accuracy and more interpretable results. The sample-by-sample approach trains a regression model for each tumor sample independently and does not use information about the tumor's assignment to previously defined transcriptomic subtypes. The regularization parameter λ that controls the degree of sparsity in the trained model, was determined by 10-fold cross validation.

The regression coefficient of each regulator (TF, miRNA) establishes the importance of the corresponding regulatory element for the prediction of gene expression changes, while the sign of the coefficient can be interpreted as the predicted direction of regulation. This can be formulated as a regression model with **Equation 2.1**, where y_g is the log2 expression fold change in the tumor (relative normal tissue) for gene g; C_g and D_g are the gene's CNV and DNA methylation values, respectively; $N_{g,TF}$ and $N_{g,\ miR}$ are the number of binding sites for TF or miRNA in the gene's promoter or 3'UTR, respectively; and $W = (W_{CN}, W_{DM}, W_{miR}, W_{TF})$ is the model vector of regression coefficients.

$$y_g = W_{CN}C_g + W_{DM}D_g + \sum_{TF \in \{1,\ldots,T\}} W_{TF}\, N_{g,TF} + \sum_{miR \in \{1,\ldots,M\}} W_{miR} N_{g,miR} \qquad \textbf{Equation 2.1}$$

The performance of models trained with different numbers of features were further compared. The sample-specific models were first built with only miRNAs-binding sites as features. The other features, TFs-binding sites, CNV and DNA methylation, were added to the models one by one. For each modeling method and each sample, the Spearman correlation was computed using 10-fold cross validation on genes that were held-out. Specifically, for each tumor, we performed a 10-fold cross validation by training a model with given features on 90% of the genes and testing its predictions on the remaining 10% of genes that were held-out. After each cross validation run, we obtained the Spearman rank correlation between the predicted and the observed DEG expression changes for the model in the tumor. In this way, the mean cross validation Spearman rank correlation was obtained for each model in each tumor. By contrast, the output gene expression changes were randomized to train sample-specific regression models.

2.2.5 Feature scoring scheme to identify the key regulators in different BC subtypes

After training the tumor-specific model for each tumor, a feature scoring scheme [205] was used to determine the most predominant TFs and miRNAs of gene expression changes in each BC subtype relative to normal samples (**Figure 2.1e**). Specifically, for a given BC subtype *S*, the regression coefficient of each single TF or miRNA regulator *r* was set to zero for all samples belonging to this BC subtype *S* and the change in total squared loss *L* over all genes was computed in these samples as *score(r, S)*. The formula shown in **Equation 3.2**,

$$score(r,S) = \sum_{g \in \{1,\ldots,G\}} score(r,S,g) = \sum_{g \in \{1,\ldots,G\}} \sum_{k \in S} \{L(y_{g,k}, W_k^{r\to 0} \cdot X_{g,k}) - L(y_{g,k}, W_k \cdot X_{g,k})\} \qquad \textbf{Equation 2.2}$$

where *k* indexes the tumor samples and *g* indexes the DEGs, $y_{g,k}$ is the expression change relative to normal tissue of gene *g* in sample *k*, $X_{g,k}$ is the vector of TF and miRNA binding site counts for gene *g* and the CNV and DNA methylation of gene *g* in sample *k*, and *S* ranges over the set of the

four subtypes. L is squared loss and $W_k^{r\to 0}$ denotes the model vector obtained from W_k by setting the coefficient W_k^r to 0. This score measures the degree of influence of the regulator in predicting the changes in gene expression.

To further assess the statistical significance of the feature scores, the randomized data were derived by permuting the motif hits over the genes independently for each TF/miRNA. We carried out the tumor-specific Lasso training procedure on randomized data 5000 times and then computed the resulting random score distribution for each regulator and each subtype. The empirical *p*-value for each regulator and subtype was calculated based on these distributions and adjusted for multiple testing over the TF/miRNA regulators using with the Benjamini-Hochberg method to produce a FDR. Subtype-specific key regulators were finally selected with a FDR < 0.05.

2.2.6 Gene set over-representation/enrichment analysis to identify the potential target pathways of TNBC regulators

A follow-up exploration focused on the above-identified key regulators from the TNBC subtype (i.e., 24 regulators, including 6 TFs and 18 miRNAs) (the left panel in **Figure 2.1f**). A functional enrichment analysis was performed using targets of the identified TNBC key regulators (miRNAs and TFs). Specifically, we extracted the predicted targets having nonzero binding sites for TFs provided by TRRUST and for miRNAs provided by starBase. For each regulator, the target list was intersected with the identified 3008 DEGs, and only targets that were also DEGs were retained as regulators. We then downloaded canonical pathways from MSigDB (c2.cp.v6.2.entrez.gmt). For each target list, we assessed their enrichment for each canonical pathway by the hypergeometric test using the R package clusterProfiler. The maximum gene set size was 500 while the minimum gene set size was 2. The multiple test *p*-value was adjusted usung

the Benjamini-Hochberg method to produce a FDR and we identified pathways related to the selected regulators with enrichment where the FDR < 0.01.

To validate the pathways enriched for targets of the TNBC key regulators, the GSEA analysis was performed on the canonical pathways from MSigDB to determine if the potential target pathways were deferentially expressed in TNBC samples relative to normal samples. Specifically, using the TCGA data, the gene expression values for 114 TNBC tumor and 99 normal breast tissue samples were put through the GSEA software to generate an output of pathways that are upregulated or downregulated in TNBC tumors. We then compared the differential pathways with the pathways over-represented by targets of the TNBC regulators.

2.2.7 Survival outcome association analysis based on TNBC regulators

Next, the prognostic value of the identified TNBC regulators was examined based on their expression levels in TNBC patient cohorts (the middle panel in **Figure 2.1f**). In particular, datasets from the TCGA and the METABRIC databases were used to analyze clinical outcome associations. Curated clinical TCGA data from a study were downloaded from UCSC Xena. The study highlighted four types of carefully curated survival endpoints, and recommended the use of the endpoints of overall survival (OS), disease-specific survival (DSS), progression-free interval (PFI), disease-free interval (DFI) for each TCGA cancer type. For METABRIC, the discovery set was used. The normalized gene expression levels were obtained from the European Genome-Phenome Archive (https://ega-archive.org/dacs/EGAC00001000484) and the curated clinical data were downloaded from a previous study [288]. TNBC cases (127 out of 997 breast cancer cases) with survival data (OS and DSS) and normalized gene expression profiles were selected for the following survival analysis.

Firstly, examination of whether each of the identified 24 TNBC regulators alone could be a prognostic marker for TNBC was done, by performing survival analysis for each regulator in the TNBC and BC cohorts from TCGA. The mean expression level of a given regulator was used as the cutoff to stratify patients into two groups. Kaplan-Meier (KM) survival analysis with different endpoints (OS, DSS, PFI, DFI) was performed, using the R package survcomp to examine the survival differences between the two groups.

Then, using the six TFs from the 24 key regulators and all the 24 key regulators identified in TNBC, the combined risk score (CRS) signatures were built for the six TFs (i.e., 6-TF CRS) and the 24 regulators (i.e., 24-regulator CRS) as in **Equation 2.3**,

$$CRS_s = \sum_{i=1}^{n} coef_i \times x_i \qquad \textbf{Equation 2.3}$$

where n is the number of regulators used to build the CRS model and x_i is the expression level of regulator i in sample s. For each regulator i, the association between its expression and the patient's OS outcome was determined using the univariate Cox proportional hazards model [289] in the 112 TCGA TNBC samples and the coefficient ($coef_i$) was extracted from the Cox model. We first built the CRS model with the six TFs from the 24 key regulators identified in the TNBC subtype. TNBC patients with CRS were binarized into high-risk and low-risk groups using the *xtile* function in the R package statar with a probability parameter set to 0.65. The survival differences between the high- and low-risk groups were assessed by KM curves. We further tested the prognostic power of the 6-TF CRS signature in the METABRIC TNBC data set. In the same way, the 24-regulator CRS signature was built using all 24 key TNBC regulators in the TCGA TNBC data set.

2.2.8 Survival outcome association analysis based on the potential target pathways of TNBC regulators

The prognostic value of the potential target pathways of the identified TNBC regulators was also examined (the middle panel in **Figure 2.1f**). In this study, target genes of six out of the 24 key TNBC regulators were found to be enriched in BIOCARTA_PPARA_PATHWAY (henceforth referred to as PPARA pathway). Target genes of three out of the 24 TNBC key regulators were found to be enriched in KEGG_PPAR_SIGNALING_PATHWAY (henceforth referred to as PPAR pathway), which has similar mechanisms to the PPARA pathway. Also, target genes of two out of the 24 TNBC key regulators were enriched in PID_FOXM1_PATHWAY (henceforth referred to as FOXM1 pathway). Through the GSEA analysis of TNBC versus normal breast tissue, the PPARA pathway was found to be among the top downregulated pathways while the FOXM1 pathway was found to be the top upregulated pathways. Given the importance of the three pathways in TNBC, we examined their clinical outcome relevance. We assumed that genes which are members of the PPARA, PPAR or FOXM1 pathways and are also targets of the key TNBC regulators are important in TNBC tumorigenesis and progression. Thus we tested if these genes could be used to build a signature named Up-Down gene mean expression ratio (UDR) for TNBC prognosis. Specifically, we first collected all gene members from the three pathways from MSigDB (for example, let's call this gene list1). We then intersected these with the 3008 DEGs (which is called gene list2), and further intersected the intersection of gene list1 and list2 with the targets (which is called list3) of the 24 TNBC key regulators. This means that only the overlap among the three gene lists was kept to build the UDR signature. As a result, a total of 51 genes were retained from this process, including17 up- and 34 down-regulated DEGs that are also

members of the PPARA, PPAR or FOXM1 pathways and are also targets of the 24 TNBC key regulators were selected. The UDR signature was calculated using **Equation 2.4**

$$UDR_s = \left(\frac{1}{17}\sum_{i=1}^{17} x_i\right) \Big/ \left(\frac{1}{34}\sum_{j=1}^{34} x_j\right) \qquad \textbf{Equation 2.4}$$

where x_i represents expression level of each of the 17 up-expressed genes and x_j represents expression level of each of the 34 down-expressed genes in a given sample *s*. Therefore, the UDR of patient sample *s* is the ratio of the mean expression level of the 17 up-expressed genes and the mean expression level of the 34 down-expressed genes. The 51-gene UDR signature was tested using the TNBC samples of the TCGA and METABRIC datasets as described in section 2.2.8. Specifically, the UDR signature score was computed for each case in a cohort and the mean of the UDR scores was used as the cutoff to stratify TNBC patients. Cases with a UDR score above the cutoff were assigned to the high-risk group while cases below the cutoff were grouped into low-risk group. We performed KM survival analysis using the R package survcomp [288] to examine the prognostic capability of the UDR signature.

2.2.9 Network construction involving TNBC regulators and their potential target pathways

The regulatory relationships between the TNBC regulators and their target DEGs in the PPARA, PPAR and FOXM1 pathways were further visualized (the right panel in **Figure 2.1f**). To construct the network, the TF/miRNA-gene pairs involving nonzero binding hits between the 24 TNBC regulators and all the identified DEGs were initially compiled. We then extracted the gene sets of the three pathways (the PPARA, PPAR and FOXM1 pathways) from MSigDB [104] and filtered the regulator-gene pairs by gene sets of the three pathways. A matrix was then generated (262 rows by 3 columns), containing 262 interactions between 23 TNBC regulators and 51 DEGs. The first column is the selected TNBC regulators and is used as the source node for the network; the second column is the corresponding targets which are members of at least one of the three

pathways and used as the target node for the network; the third column is the number of hits between the regulator-target pairs and used as the interaction for the network. In the end, the matrix was imported into Cytoscape v3.5.1 [104] to generate the network.

2.3 Results

2.3.1 Curated data sets

To train sample-specific models, data sets from multiple platforms were curated and summarized in **Table 2.1**. The predicted DNA sequence-based TF-binding profile matrix between 795 TFs and 2,492 distinct genes were downloaded from TRRUST. The ChIP-Seq TF-binding site data for 125 TFs were also curated from ReMap. Among the 125 TFs, 95 were derived from luminal A cell lines (with the majority (78) being MCF7), 6 were from luminal B cell lines, 8 were from HER2-enriched cell lines and 16 were from TNBC cell lines. The miRNA-binding profile matrix between 15,168 targets and 600 miRNAs was obtained from starBase. In this study, the tumor-specific Lasso model was based on DEG expression changes instead of global gene expression changes. Gene differential analysis, using the R package limma-voom showed that 4,227 of the 17,675 genes were significantly (FDR < 0.05) differentially expressed between BC and normal tissue samples with the absolute value of logFC not less than 1 and were considered DEGs. Out of the 4,227 DEGs, 3008 have data available at multi-omics levels. The common samples, DEGs and miRNAs for which data were available across different platforms were kept. In the end, there were 391 breast tumor samples with 3,008 DEGs, 795 TFs and 362 miRNAs (**Table 2.1**). For the 3,008 genes, their respective interaction data with the 795 TFs and the 362 miRNAs were also obtained. Among the 391 tumor samples, 239 are luminal A, 74 are luminal B, 21 are HER2-enriched, and 57 are TNBC.

Table 2.1 Multi-dimensional data used to build the tumor-specific Lasso model

Type	Platform/category	Database	All data	Common data
Subtype annotation	IHC	TCGA	694 tumors: 424 LumA, 121 LumB, 37 HER2+, 112 TNBC	391 tumors: 239 LumA, 74 LumB, 21 Her2E, 57 TNBC
mRNA expression [1]	Illumina Hiseq 2000	TCGA	17675 genes × 1104 tumors	3008 genes × 391 tumors
miRNA expression [2]	Illumina Hiseq 2000	TCGA	600 miRNAs × 756 tumors	362 miRNAs × 391 tumors
CNV	Affymetrix SNP 6.0	TCGA	24776 genes × 1080 tumors	3008 genes × 391 tumors
DNA methylation	Illumina Infinium HumanMethylation450	TCGA	26586 genes × 790 tumors	3008 genes × 391 tumors
TF-target	Sequence-based	TRRUST v2.0	2492 genes × 795 TFs	3008 genes × 795 TFs
miRNA-target	Sequence-based	starBase v3.0	15168 targets × 618 miRNAs	3008 genes × 362 miRNAs

[1] 114 normal samples were available for calculating DEG expression change; [2] 76 normal samples are available for calculating miRNA expression change

2.3.2 Construction of the tumor-specific Lasso regression model

The Lasso regularized linear regression approach was applied to the 391 breast tumor profiles and a model was trained for each tumor independently. In the end, $W = (W_{CN}, W_{DM}, W_{miR}, W_{TF})$ was the vector of regression coefficients for each tumor. Combing the 391 tumor-specific models, the following matrices were obtained for further analyses: $\{w\}_{N\times M}$, where N is the 391 tumors/models and M is the 1,159 features (CNV, DNA methylation, 795 TFs and 362 miRNAs).

To justify the model formally, models with different number of features in terms of their abilities to explain DEG log expression fold changes in BC relative to normal breast tissue were compared. **Figure 2.2a** illustrates the 10-fold cross validation results as a distribution of the correlation calculated across all 391 breast tumor samples for each model. Unsurprisingly, the model with all features (red, right most in **Figure 2.2a**) performed the best and significantly better (p-value < 0.0001, Wilcoxon signed rank test) than all the other alternatives, achieving a mean

Spearman correlation of 0.273. Comparing the best model with other models (where some regulatory factors were excluded), it is remarkable to observe that a significant improvement in cross validation performance (p-value < 0.0001, Wilcoxon signed rank test) is attributable to the addition of DNA methylation.

Two alternative models using CNV, DNA methylation and miRNA-binding site data but with different TF-binding site data from TRRUST (the model used in this study) and ReMap, respectively, were further compared. The two models were compared in terms of the Spearman correlation between the predicted and observed gene expression changes via 10-fold cross validation. The second model using ReMap was found to perform significantly better than the first model (p-value < 2.2×10^{-16}, Wilcoxon signed rank test) (**Figure 2.2b**). Therefore, TF-target interaction data derived from ChIP-Seq methods conferred significantly higher explanatory power than the binding motif-based TRRUST database for the DEG expression changes in BC. This is not surprising because ReMap contains experimentally determined TF-binding sites whereas TRRUST predicts potential TF-binding sites. Nevertheless, the decision was made to use the TRRUST model because of the limited ChIP-Seq TF-binding site data that were available during the writing of this manuscript. In addition, the majority of the ChIP-Seq data in ReMap were derived from MCF7 and other luminal cell lines. Therefore, these data do not have representative cell lines from all of the BC subtypes and such overrepresentation of one BC subtype could bias any potential predictive capabilities of our model.

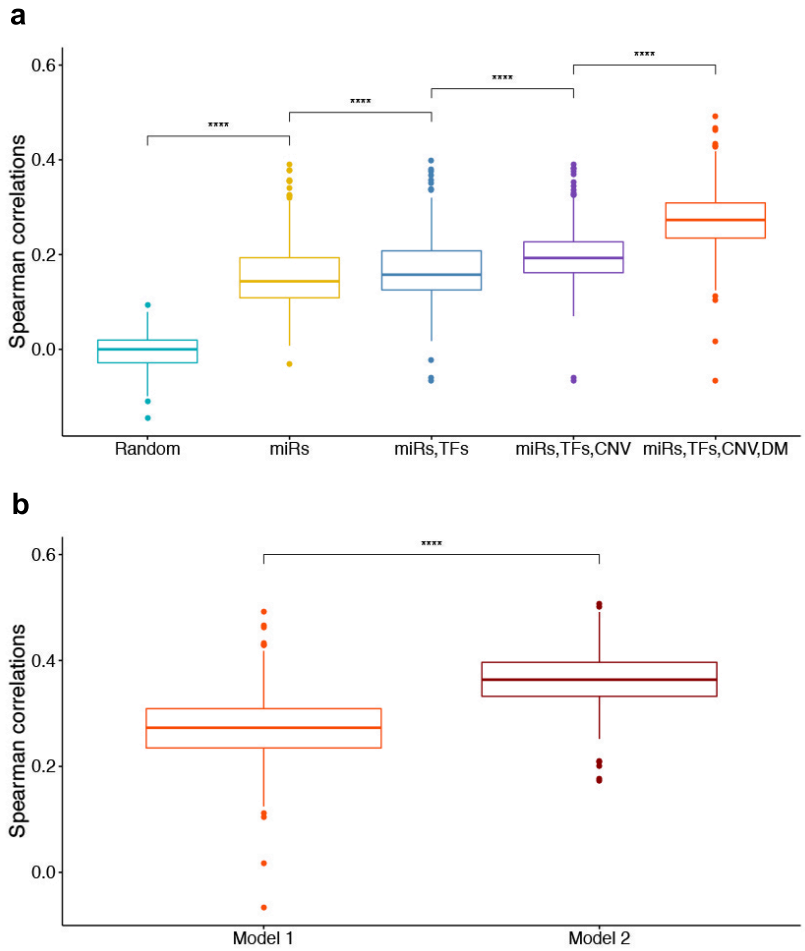


Figure 2.2 Model comparison

(**a**) Boxplot showing Spearman correlations between predicted and actual gene expression changes for all samples in models with different regulators. Using only miRNA-binding sites as features is significantly better than random; adding TF-binding sites significantly improves cross validation performance over using only miRNAs-binding sites; adding CNV significantly improves cross validation performance over using only TF- and miRNA-binding sites, while the model using TF/miRNA-binding sites, CNV and DNA methylation outperforms the others. (**b**) Boxplot shows Spearman correlations for all samples in Model 1 and Model 2. Model 1 was trained using CNV, DNA methylation, miRNA-binding sites and TF-binding sites from TRRUST while Model 2 was trained using CNV, DNA methylation, miRNA-binding sites and TF-binding sites from ReMap. **** p-value < 0.0001 by Wilcoxon signed rank test.

2.3.3 Identification of key regulators in different BC subtypes

To determine the key TF and miRNA regulators in different BC subtypes, a feature scoring scheme was adopted to measure the extent of DEG log expression fold changes in each subtype explained by each regulator (i.e., error score for each regulator in each subtype). **Figure 2.3a-d** shows the results of feature scoring for the four BC subtypes, with miRNAs and TFs sorted based on their error scores in each subtype. Key regulators were selected in each subtype with a threshold corresponding to FDR of 0.05 relative to regression models trained on randomized data. As a result, we identified 25, 20, 15 and 24 key regulators for luminal A, luminal B, HER2-enriched and TNBC subtypes, respectively. A total of 31 unique regulators (including common and subtype-specific regulators) were identified across the four subtypes and summarized in **Table 2.2**. A number of regulators are common for all subtypes, while some regulators are specific to a subtype alone (**Figure 2.3e**). The regulators that are shared among all four subtypes were called common regulators, and they included two TFs and seven miRNAs.

miR-181b-5p was found to be specific to luminal A subtype. A previous study identified miR-181b-5p upregulated more than 2-fold in BC compared to normal adjacent tumor tissues [290]. This study suggested that there is a novel breast cancer progression pathway, involving HMGA1, CBX7 and miR-181b-5p [290]. Three miRNAs, miR-655-3p, miR-654-5p, miR-625-5p, were identified as unique regulators to luminal B subtype. Corroborating this, previous studies showed that circulating miR-654-5p was found to be associated with ER-positive BC [291]. The current study found that miR-9-3p was associated with TNBC and it has also been reported to target β1 integrin to sensitize claudin-low breast tumor cells to MEK inhibition [292]. No significant regulators were found to be unique to the HER2-enriched subtype.

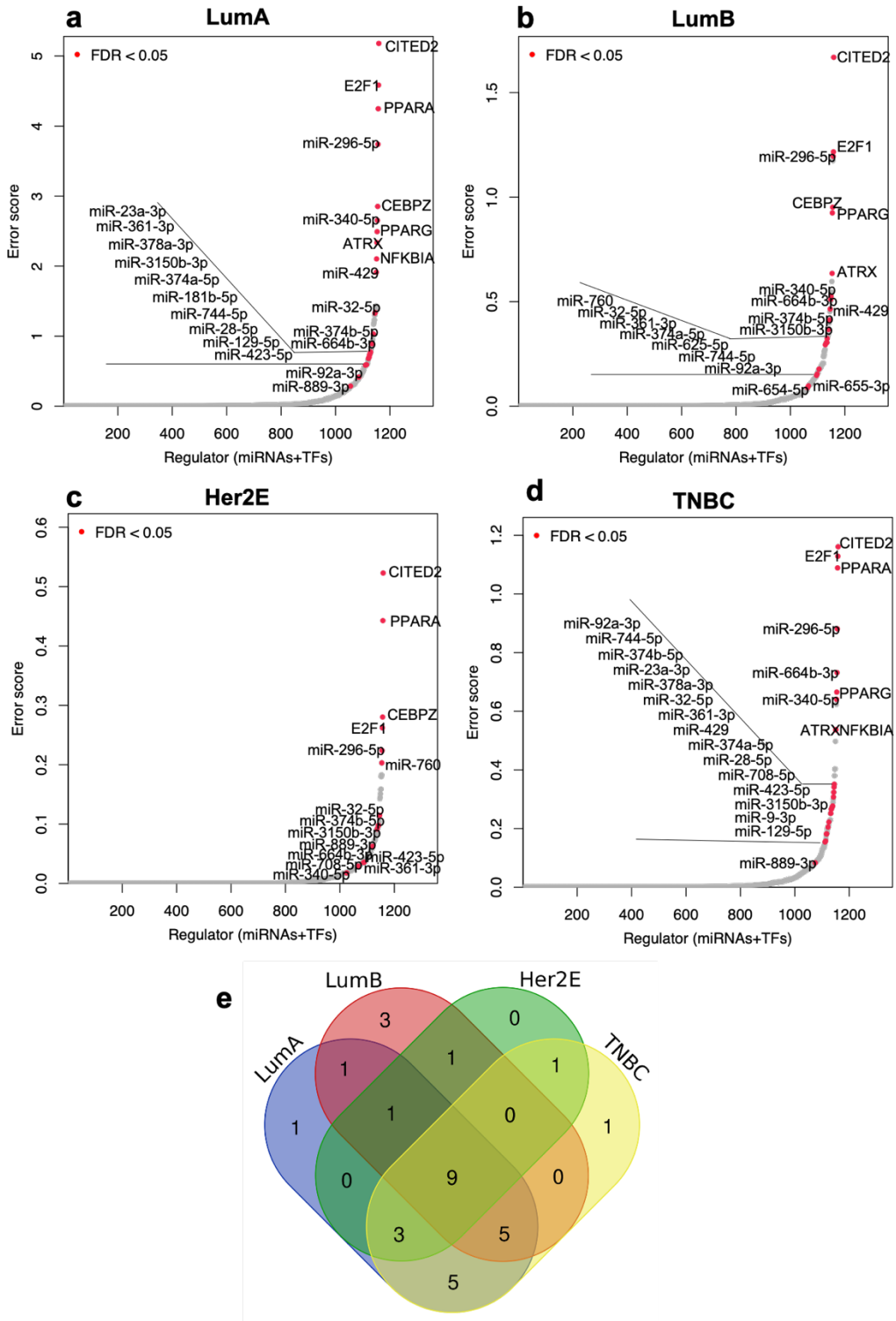


Figure 2.3 Key regulators identified in four BC subtypes by feature scoring

(**a-d**) Regulators are ranked based on error score in LumA (a), LumB (b), Her2E (c) and TNBC (d). The axis x indicates the miRNA/TF regulators, and the axis y indicates error score (i.e., increase in squared error across tumors of a subtype after excluding the regulator from Lasso models). Key regulators for each subtype are identified at an FDR of 5% relative to Lasso models trained on randomized data and are shown as red dots. (**e**) Venn diagram shows the key regulators across LumA, LumB, Her2E and TNBC subtypes.

Table 2.2 The common and subtype-specific regulators identified in four BC subtypes

Regulator	LumA		LumB		HER2E		TNBC	
	Error score	FDR < 0.05	Error score	FDR < 0.05	Error score	FDR < 0.05	Error score	FDR < 0.05
CITED2	5.18	yes	1.67	yes	0.52	yes	1.16	**yes**
E2F1	4.58	yes	1.22	yes	0.26	yes	1.13	**yes**
PPARA	4.25	yes	-	-	0.44	yes	1.09	**yes**
miR-296-5p	3.74	yes	1.2	yes	0.22	yes	0.88	**yes**
CEBPZ	2.85	yes	0.95	yes	0.28	yes	-	no
miR-340-5p	2.65	yes	0.53	yes	0.02	yes	0.64	**yes**
PPARG	2.49	yes	0.92	yes	-	-	0.67	**yes**
ATRX	2.34	yes	0.64	yes	-	-	0.54	**yes**
NFKBIA	2.1	yes	-	-	-	-	0.54	**yes**
miR-429	1.91	yes	0.46	yes	-	-	0.25	no
miR-32-5p	1.34	yes	0.3	yes	0.11	yes	0.27	**yes**
miR-374b-5p	1.33	yes	0.41	yes	0.1	yes	0.32	**yes**
miR-664b-3p	1.03	yes	0.51	yes	0.06	yes	0.73	**yes**
miR-23a-3p	0.89	yes	-	-	-	-	0.31	**yes**
miR-361-3p	0.77	yes	0.3	yes	0.04	yes	0.27	**yes**
miR-378a-3p	0.76	yes	-	-	-	-	0.28	**yes**
miR-3150b-3p	0.74	yes	0.37	yes	0.1	yes	0.18	**yes**
miR-374a-5p	0.73	yes	0.29	yes	-	-	0.25	**yes**
miR-181b-5p	0.7	yes	-	-	-	-	-	no
miR-744-5p	0.7	yes	0.15	yes	-	-	0.34	**yes**
miR-28-5p	0.68	yes	-	-	-	-	0.22	**yes**
miR-129-5p	0.59	yes	-	-	-	-	0.15	**yes**
miR-423-5p	0.59	yes	-	-	0.04	yes	0.18	**yes**
miR-92a-3p	0.42	yes	0.15	yes	-	-	0.35	**yes**
miR-889-3p	0.29	yes	-	-	0.09	yes	0.08	**yes**
miR-760	-	no	0.32	yes	0.2	yes	-	no
miR-625-5p	-	no	0.18	yes	-	-	-	no
miR-708-5p	-	no	-	-	0.03	yes	0.2	**yes**
miR-9-3p	-	no	-	-	-	-	0.16	**yes**
miR-654-5p	-	no	0.09	yes	-	-	-	no
miR-655-3p	-	no	0.1	yes	-	-	-	no

Regulators are miRNAs and TFs identified as key regulators in any of the four BC subtypes
Error score for each regulator in each subtype was calculated from Equation 2.2 for feature scoring.
FDR was calculated to assess the statistical significance of the corresponding error scores. If a regulator has an error score with FDR < 0.05 (i.e., yes) in a subtype, it is identified as a key regulator in this subtype.

2.3.4 Potential target pathways of TNBC regulators

Since there are few effective therapies available for TNBC among all subtypes, it is necessary to investigate the key regulators in this subtype more closely. A total of 24 predominant regulators, consisting of 6 TFs and 18 miRNAs, satisfied a 5% FDR cutoff for the TNBC subtype (see **Table 2.2**). To examine the potential target pathways of these TNBC regulators, a gene set over-representation analysis was performed using their associated DEG targets. Based on the over-representation in the canonical pathways from MSigDB, several interesting pathways related to the TNBC key regulators were discovered (FDR < 0.05, **Table 2.3**). For instance, 11 of the 24 key regulators were engaged in pathways in cancer (KEGG_PATHWAYS_IN_CANCER) whereas seven of the 24 regulators are involved in cell-matrix adhesion pathway (KEGG_FOCAL_ADHESION) (see **Table 2.3**). Additionally, targets of miR-340-5p and E2F1 were enriched in the RB1 pathway (PID_RB_1PATHWAY) and P53 pathway (PID_P53_DOWNSTREAM_PATHWAY). Also, targets of two TNBC key regulators (miR-340-5p and E2F1) were found to be enriched in the FOXM1 pathway (PID_FOXM1_PATHWAY) while targets of six TNBC key regulators (PPARG, PPARA, miR-9-3p, miR-664b-3p, miR-340-5p and miR-129-5p) were found to be enriched in the PPARA pathway (BIOCARTA_PPARA_PATHWAY). In addition, targets of three regulators (PPARG, PPARA and miR-340-5p) are enriched in the PPAR pathway (KEGG_PPAR_SIGNALING_PATHWAY), which has similar mechanisms to the PPARA pathway.

Further GSEA analysis was performed to identify differential pathways in TNBC tumors relative to normal breast tissue samples. Using a FDR of 5% as cut-off, 137 pathways were significantly upregulated in TNBC versus normal breast tissue samples (**Appendix A**). Remarkably, the FOXM1 pathway was found to be the most significantly upregulated pathway in

TNBC (**Figure 2.4a**). The GSEA results for the FOXM1 pathway together with the genes in this pathway are listed in **Appendix B**. No pathways were found to be significantly downregulated in TNBC versus normal breast tissue with the FDR cutoff set to 5%. So the top 20 downregulated pathways were selected in terms of FDR but without adhering to a strict significance level (**Appendix C**). Interestingly, though the PPARA pathway does not show significant downregulation in TNBC versus normal breast tissue (FDR = 0.322), it exhibits the largest normalized enrichment score among the top 20 downregulated pathways (**Figure 2.4b**). The GSEA results for the PPARA pathway together with genes in this pathway are listed in **Appendix D**. The over-representation analysis of the TNBC regulator targets together with the GSEA results support the importance of the PPARA and FOXM1 pathways and their regulators in TNBC.

Table 2.3 Potential target pathways of TNBC regulators

Regulators	Expression [a]	Target # [b]	Enriched Pathways [c]	Ratio [d]	Enrichment FDR [e]
E2F1	Up	48	REACTOME_CELL_CYCLE;	19/241	4.43×10^{-12}
			PID_FOXM1_PATHWAY;	9/40	1.70×10^{-11}
			KEGG_PATHWAYS_IN_CANCER	9/328	2.09×10^{-5}
CITED2	Down	2	REACTOME_EXTRACELLULAR_MATRIX_ORGANIZATION;	2/87	8.03×10^{-4}
			REACTOME_DEGRADATION_OF_THE_EXTRACELLULAR_MATRIX;	2/29	1.74×10^{-4}
			NABA_ECM_REGULATORS	2/238	4.04×10^{-3}
PPARA	Down	20	REACTOME_METABOLISM_OF_LIPIDS_AND_LIPOPROTEINS;	11/478	8.87×10^{-8}
			BIOCARTA_PPARA_PATHWAY;	4/58	2.08×10^{-4}
			KEGG_PATHWAYS_IN_CANCER	4/328	3.76×10^{-2}
PPARG	Down	29	KEGG_PATHWAYS_IN_CANCER;	9/328	5.73×10^{-5}
			REACTOME_HORMONE_SENSITIVE_LIPASE_HSL_MEDIATED_TRIACYLGLYCEROL_HYDROLYSIS;	3/13	5.48×10^{-4}
			REACTOME_METABOLISM_OF_LIPIDS_AND_LIPOPROTEINS;	8/478	3.77×10^{-3}
			BIOCARTA_PPARA_PATHWAY;	3/58	1.92×10^{-2}
			REACTOME_DEGRADATION_OF_THE_EXTRACELLULAR_MATRIX	2/29	4.81×10^{-2}
ATRX	Down	2	BIOCARTA_AHSP_PATHWAY	2/13	1.97×10^{-6}
NFKBIA	Down	3	REACTOME_DEGRADATION_OF_THE_EXTRACELLULAR_MATRIX;	3/29	9.65×10^{-7}
			REACTOME_EXTRACELLULAR_MATRIX_ORGANIZATION;	3/87	1.40×10^{-5}
			KEGG_PATHWAYS_IN_CANCER	2/328	1.37×10^{-2}
miR-374b-5p	Down	589	KEGG_PATHWAYS_IN_CANCER	28/328	4.84×10^{-2}
miR-32-5p	Up	523	KEGG_FOCAL_ADHESION;	19/201	4.45×10^{-2}
			NABA_CORE_MATRISOME	22/275	4.74×10^{-2}
miR-361-3p	-	488	KEGG_FOCAL_ADHESION;	23/201	8.25×10^{-5}
			KEGG_PATHWAYS_IN_CANCER;	27/238	2.41×10^{-3}
			PID_AVB3_INTEGRIN_PATHWAY	11/75	6.61×10^{-3}
miR-340-5p	Up	927	KEGG_PATHWAYS_IN_CANCER;	54/328	5.10×10^{-9}
			PID_AVB3_INTEGRIN_PATHWAY;	19/75	2.64×10^{-5}
			KEGG_FOCAL_ADHESION;	31/201	1.47×10^{-4}
			BIOCARTA_PPARA_PATHWAY;	15/58	1.47×10^{-4}
			REACTOME_EXTRACELLULAR_MATRIX_ORGANIZATION;	13/87	4.14×10^{-2}
			PID_FOXM1_PATHWAY	8/40	4.41×10^{-2}

Table 2.4 Potential target pathways of TNBC regulators (cont.)

Regulators	Expression [a]	Target # [b]	Enriched Pathways [c]	Ratio [d]	Enrichment FDR [e]
miR-664b-3p	Down	431	KEGG_FOCAL_ADHESION;	18/201	9.09×10^{-3}
			BIOCARTA_PPARA_PATHWAY;	9/58	9.09×10^{-3}
			KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION	19/267	4.57×10^{-2}
miR-296-5p	Down	353	KEGG_LEUKOCYTE_TRANSENDOTHELIAL_MIGRATION	12/118	6.14×10^{-3}
miR-423-5p	-	590	KEGG_PATHWAYS_IN_CANCER;	30/328	1.43×10^{-3}
			PID_AVB3_INTEGRIN_PATHWAY;	11/75	1.94×10^{-2}
			REACTOME_GROWTH_HORMONE_RECEPTOR_SIGNALING	6/24	3.31×10^{-2}
miR-92a-3p	Down	524	KEGG_FOCAL_ADHESION;	19/201	4.47×10^{-2}
miR-129-5p	Down	775	KEGG_PATHWAYS_IN_CANCER;	37/328	1.90×10^{-4}
			KEGG_FOCAL_ADHESION;	26/201	5.83×10^{-4}
			REACTOME_EXTRACELLULAR_MATRIX_ORGANIZATION;	13/87	1.41×10^{-2}
			PID_AVB3_INTEGRIN_PATHWAY;	12/75	1.41×10^{-2}
			BIOCARTA_PPARA_PATHWAY	10/58	1.92×10^{-2}
miR-744-5p	-	221	KEGG_PATHWAYS_IN_CANCER;	15/328	1.26×10^{-2}
			KEGG_FOCAL_ADHESION;	12/201	1.20×10^{-2}
			REACTOME_CELL_SURFACE_INTERACTIONS_AT_THE_VASCULAR_WALL	7/91	2.57×10^{-2}
miR-9-3p	-	236	KEGG_PATHWAYS_IN_CANCER;	14/328	2.98×10^{-2}
			BIOCARTA_PPARA_PATHWAY;	6/58	2.98×10^{-2}
			KEGG_MAPK_SIGNALING_PATHWAY	12/267	3.19×10^{-2}

[a] The miRNA/TF expression pattern in 114 TNBC patients when compared to the 99 adjacent normal breast tissue samples (TCGA data set was used);
[b] The number of predicted targets provided by TRRUST (for TFs) or starBase (for miRNAs) overlapped with DEGs;
[c] The enriched canonical pathways from MSigDB in the target DEGs of the corresponding regulator;
[d] The number of the targets involved in a pathway and the total number of genes involved in the pathway;
[e] The Benjamini-Hochberg method adjusted hypergeometric test *p*-value for the corresponding enriched pathway.

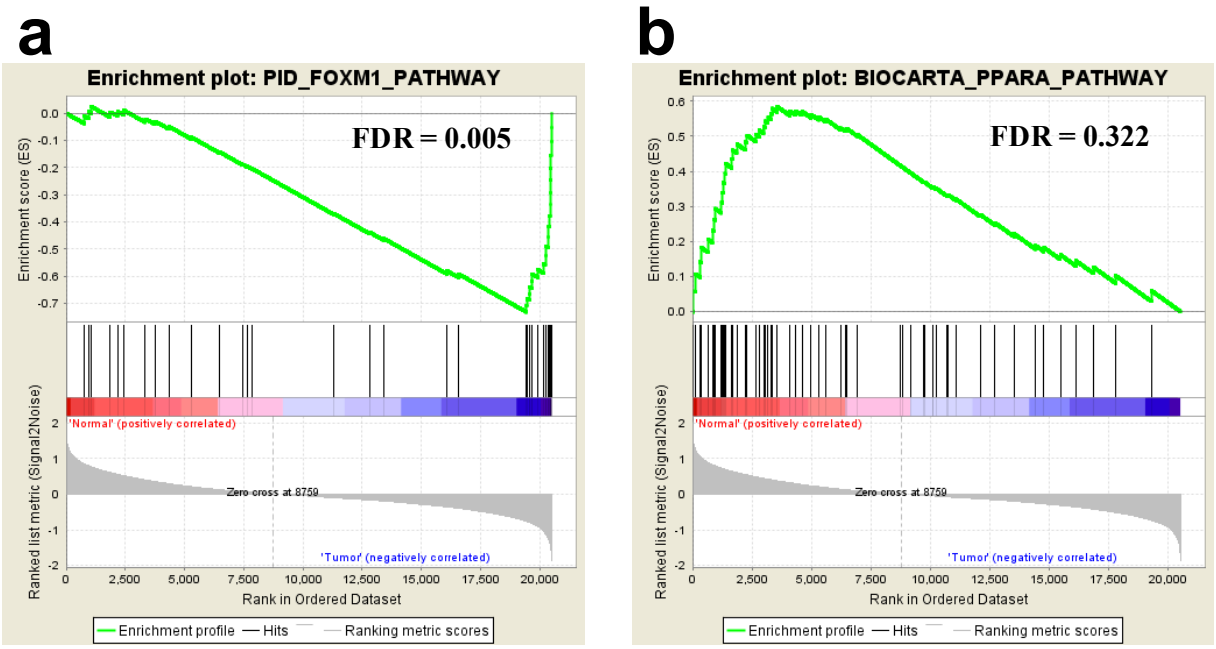


Figure 2.4 The GSEA of mRNA profiling results from TNBC and normal breast tissue samples
(**a**) The GSEA enrichment profile of the upregulated FOXM1 pathway in TNBC versus normal breast tissue. (**b**) The GSEA enrichment profile of the downregulated PPARA pathway in TNBC versus normal breast tissue.

2.3.5 Survival outcome relevance of TNBC regulators

The clinical implication of the expression of the above identified 24 TNBC regulators was examined by evaluating the prognostic value of each regulator. Six regulators were found to show a significant stratification in survival outcomes for TCGA TNBC patients (**Figure 2.5**). For two regulators (ATRX and miR-23a-3p), TNBC patients with high expression showed a significantly worse survival outcome than those with low expression. For the other four regulators (NFKBIA, miR-664b-3p, miR-708-5p, miR-889-3p), TNBC patients with high expression showed a significantly better survival outcome than those with low expression. It was also found that 11 of the 24 TNBC regulators could significantly stratify TCGA BC patients by survival outcome (**Appendix E-1** and **E-2**).

Two CRS signatures were built using different numbers of TNBC regulators. The 6-TF CRS signature could assign TNBC patients into high- and low-risk groups with different survival times. The two groups show a significant difference in OS time in both, the TCGA (p-value = 3.6 $\times 10^{-2}$, log-rank test) (**Figure 2.6a**) and the METABRIC (p-value = 1.3×10^{-2}, log-rank test) (**Figure 2.6b**) data sets. In addition, the CRS with all 24 regulators showed high-risk and low-risk group stratification in terms of both OS (p-value = 5.2×10^{-2}, log-rank test) (**Figure 2.6c**) and DSS (p-value = 4.4×10^{-2}, log-rank test) (**Figure 2.6d**). However, further validation of the 24-regulator CRS signature in METABRIC was not done since there are no expression data available for miRNAs of the 24 regulators.

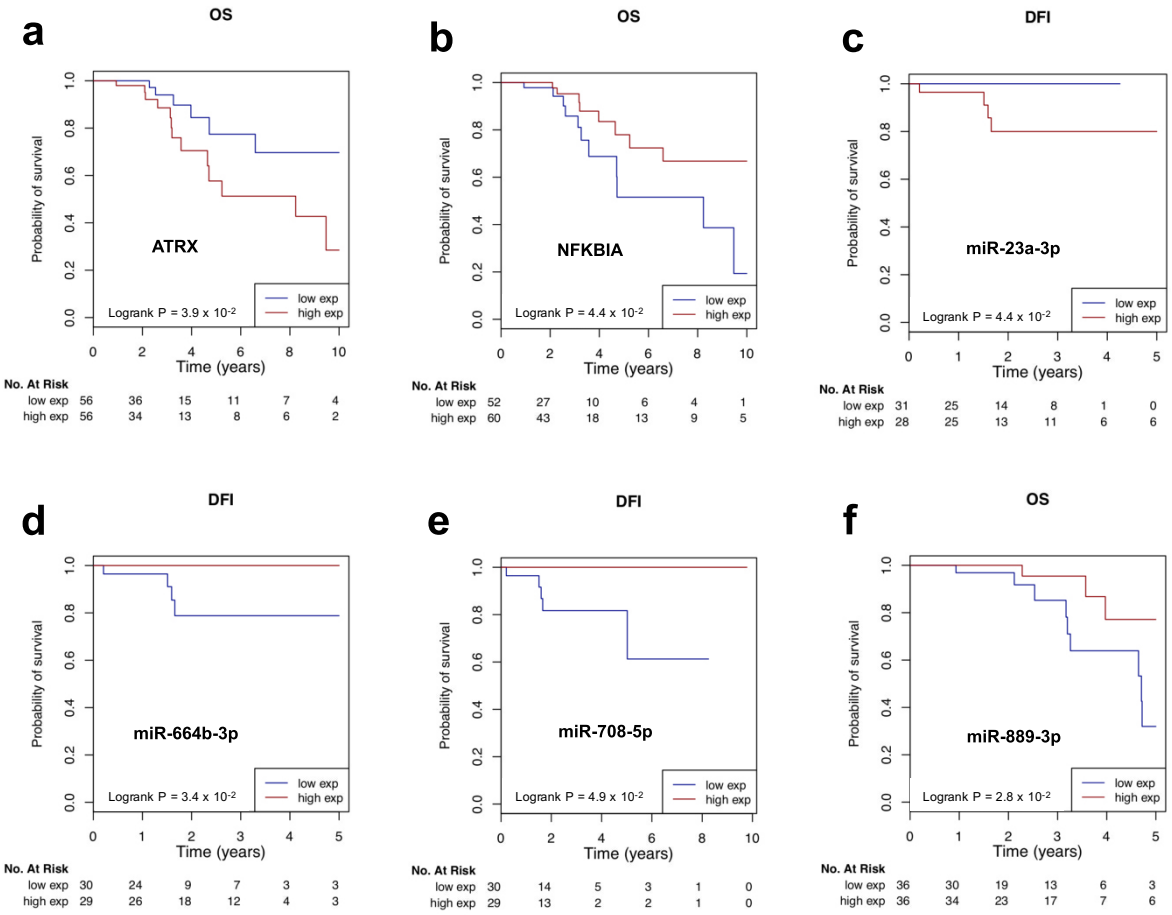


Figure 2.5 KM survival analysis of TNBC regulators in TNBC patients
Based on gene expression levels of a given regulator, TNBC patients from TCGA were divided into over-expressed ("high-exp", red line) and down-expressed ("low-exp", blue line) groups. Survival analysis with different event endpoints (OS, DSS, PFI, DFI) was performed for each regulator. Only significant results are shown here. Eight regulators alone could be a prognostic marker for TNBC patients: (**a**) ATRX, (**b**) NFKBIA, (**c**) miR-23a-3p, (**d**) miR-664b-3p, (**e**) miR-708-5p, and (**f**) miR-889-3p. OS, overall survival; DSS, disease-specific survival; PFI: progression-free interval; DFI, disease-free interval.

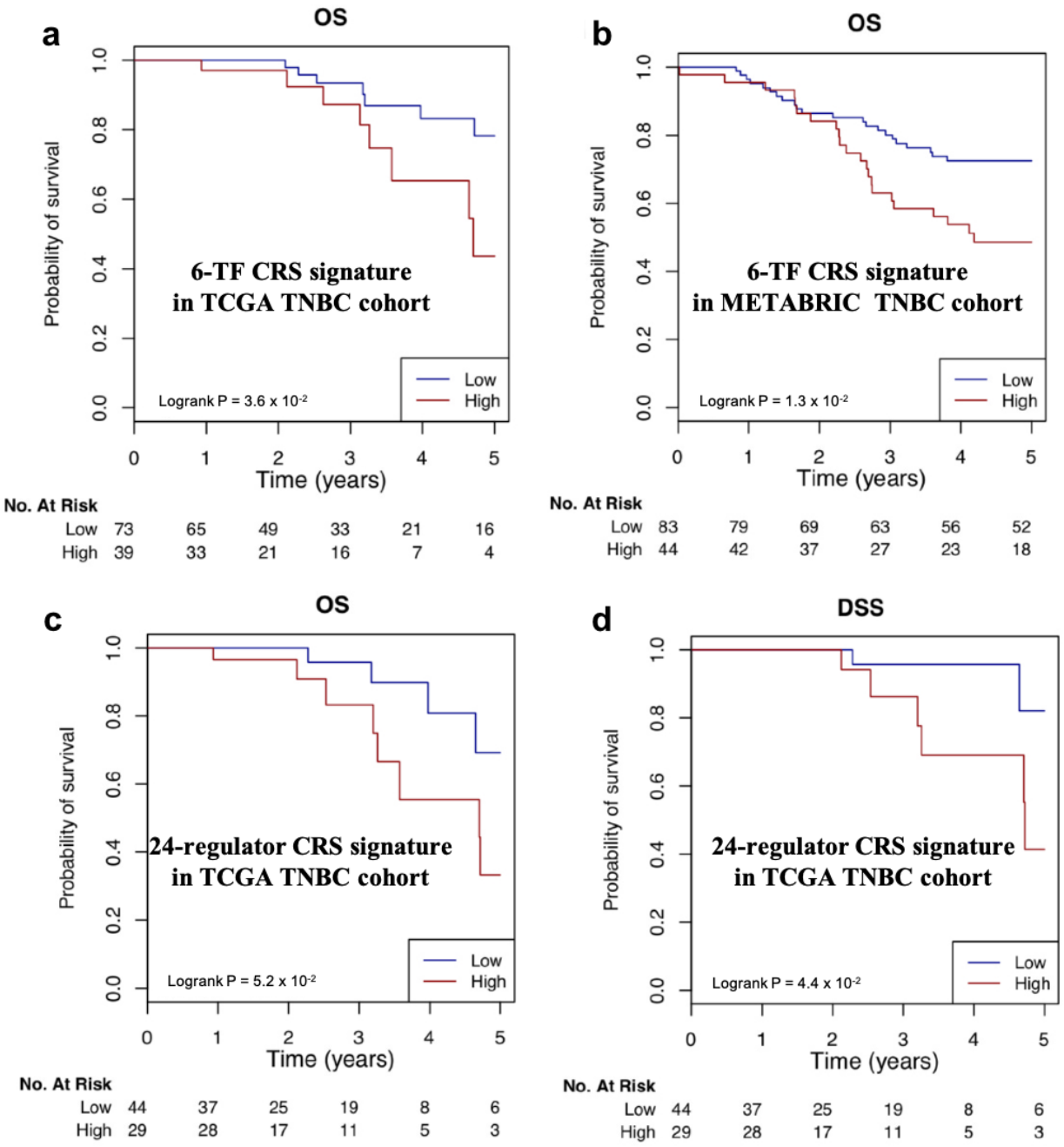


Figure 2.6 KM survival analysis of the CRS signatures, comparing survival for predicted high-risk versus low-risk TNBC patients

TNBC patients were divided into high-risk ("High", red line) and low-risk ("Low", blue line) groups according the 6-TF or 24-regulator CRS signature. Survival fractions as a function of survival time (years) were then plotted for the two groups and the significant separation of the two curves were assessed by log-rank test using the R package survcomp. The 6-TF CRS significantly stratified the TNBC samples from TCGA **(a)** and METABRIC **(b)** into high- and low-risk OS groups. The 24-regulator CRS stratified the TCGA TNBC patients into high- and low-risk groups with respect to OS **(c)** and DSS **(d)**.

2.3.6 Survival outcome relevance of PPARA/PPAR/FOXM1 pathway genes

To determine if TNBC regulator targets involved in the PPARA, PPAR and FOXM1 pathways were clinically relevant, their ability to be used to develop a multigene signature for TNBC prognosis was tested by building a 51-gene UDR signature. The UDR signature stratified the 112 TCGA TNBC samples into high- or low-risk groups with significant accuracy evaluated by both five-year OS (p-value = 5.0×10^{-2}, log-rank test) (**Figure 2.7a**) and DSS (p-value = 4.2×10^{-2}, log-rank test) (**Figure 2.7b**). The UDR signature was further tested using the METABRIC TNBC dataset. The UDR signature significantly stratified the 127 METABRIC TNBC patients into high- and low-risk OS (p-value = 2.1×10^{-2}, log-rank test) (**Figure 2.7c**) and DSS groups (p-value = 4.3×10^{-3}, log-rank test) (**Figure 2.7d**).

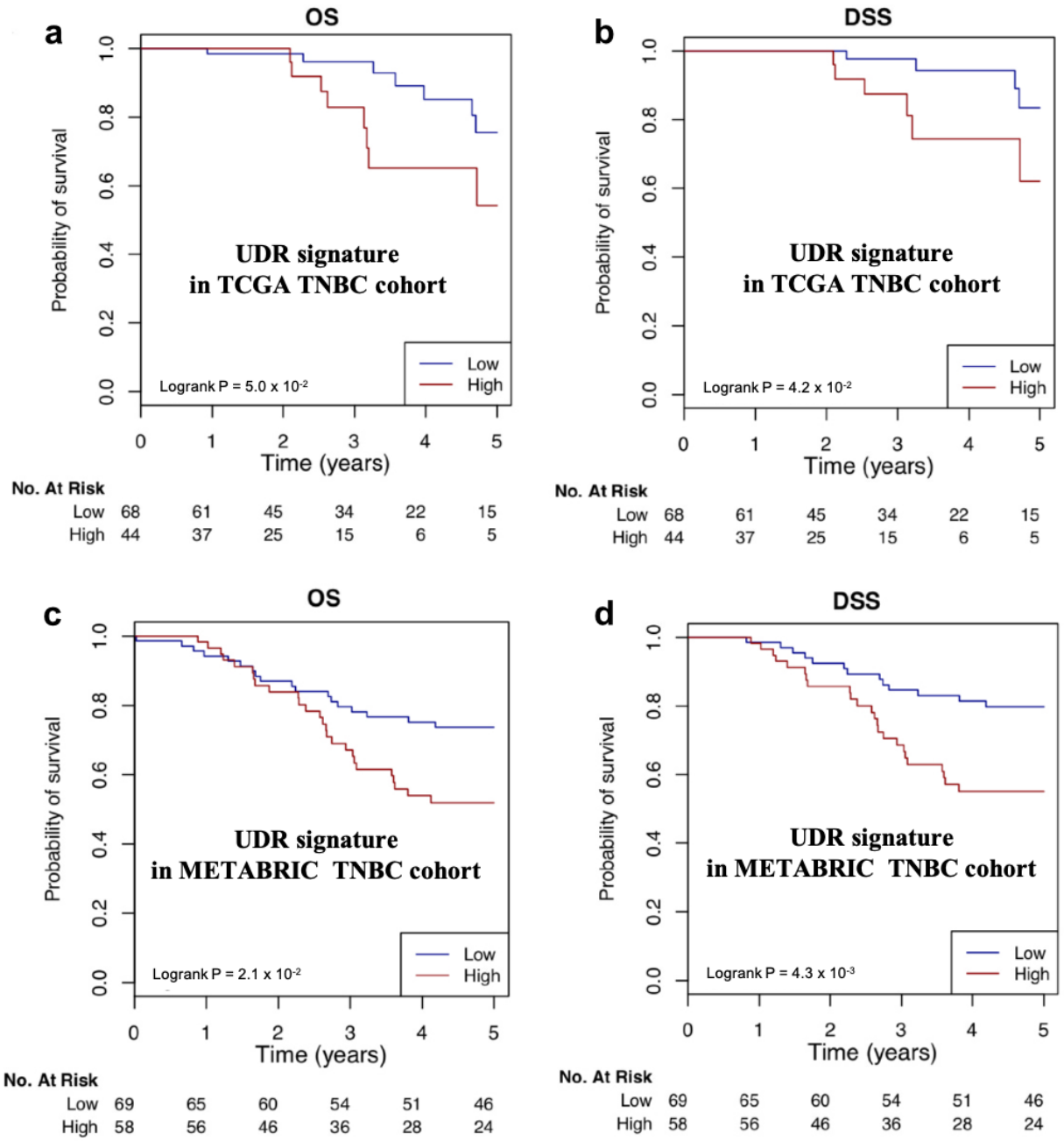


Figure 2.7 KM survival analysis of the UDR signature, comparing survival for predicted high-risk versus lower-risk TNBC patients

TNBC patients were divided into high-risk ("High", red line) and low-risk ("Low", blue line) groups according their UDR signature scores. Survival fractions as a function of survival time (years) were then plotted for the two groups and the significant separation of the two curves were assessed by log-rank test. The UDR signature was tested using TNBC samples of TCGA and METABRIC datasets separately by the R package survcomp. The UDR signature can significantly stratify the 112 TCGA TNBC samples into high- and low-risk groups by both OS (**a**) and DSS (**b**). The UDR signature can also significantly stratify the 127 METABRIC TNBC samples into high- and low-risk groups by both OS (**c**) and DSS (**d**).

2.3.7 Regulatory network involving TNBC regulators and PPARA/PPAR/FOXM1 pathway genes

To visualize the regulatory relationships between the TNBC regulators and their target DEGs involved in the PPARA, PPAR and FOXM1 pathways, the nonzero interactions between 23 TNBC regulators and 51 DEGs were imported into Cytoscape v3.5.1 to construct a network. In **Figure 2.8a**, the resulting regulatory network is densely connected, implying that the TNBC regulators identified in the current study shared many common target DEGs involved in the PPARA, PPAR and FOXM1 pathways. The hierarchical organization of the selected subnetworks involving PPARA and PPARG TFs were examined. A two-layer network architecture was formed with PPARA and PPARG, their targets and the regulators that regulate their expression (**Figure 2.8b**). Specifically, nine miRNAs and the TF E2F1 form the top layer (which were designated as master regulators), while PPARA and PPARG form the bottom layer (regulators that are regulated by the master regulators above it and regulate other gene targets below it). Remarkably, the fact that nine miRNAs regulate PPARA and PPARG and a large cohort of the PPARA, PPAR and FOXM1 pathway-related DEGs either directly via base-pairing or indirectly via PPARA and PPARG provides further support of the important role of these nine miRNAs as master regulators of the PPARA, PPAR and FOXM1 pathways in TNBC.

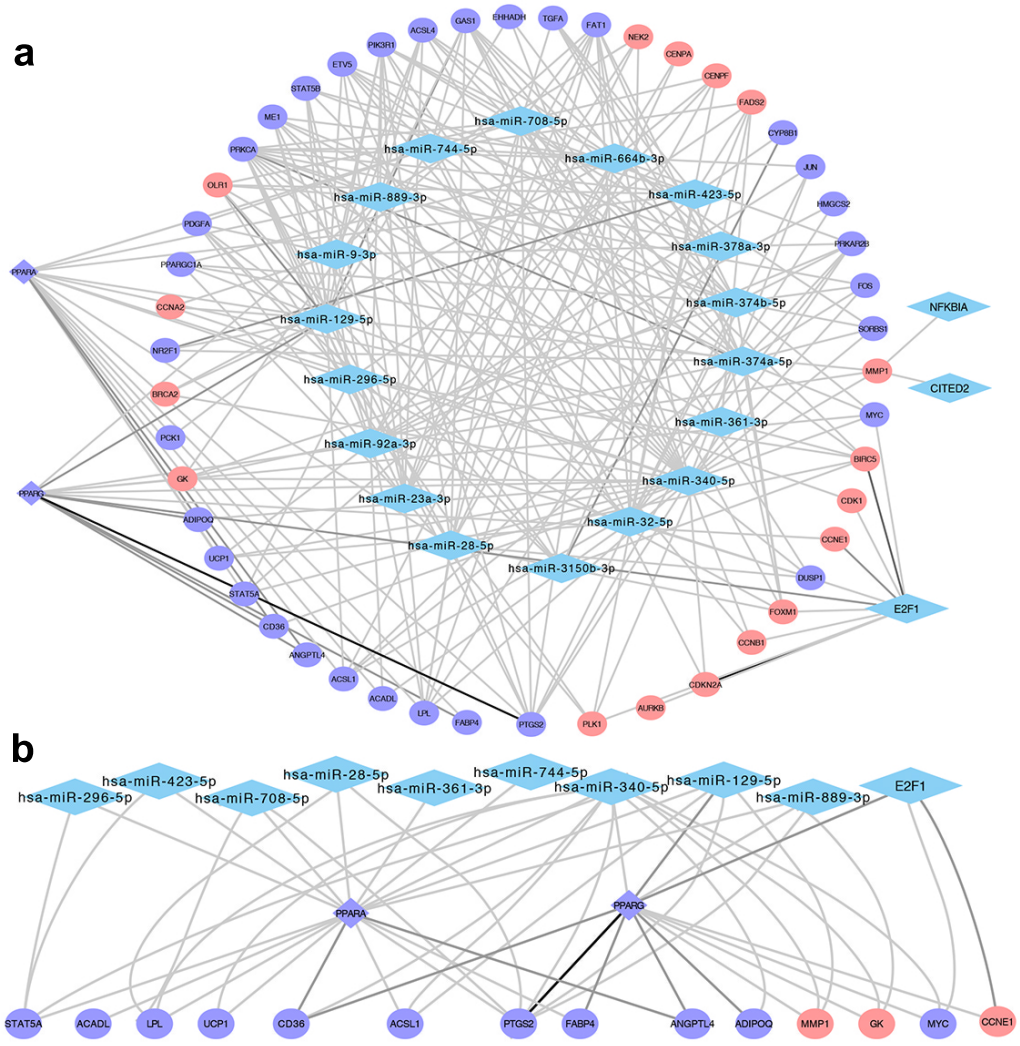


Figure 2.8 TF-target and miRNA-target regulatory networks
(**a**) The network comprises 23 of the 24 key TNBC regulators (one has no targets involved in the PPARA, PPAR, and FOXM1 pathways and thus was not included) (21 blue diamonds and 2 purple diamonds) having nonzero predicted interactions with gene elements (circles) of the PPARA, PPAR, and FOXM1 pathways obtained from MSigDB. The red and purple circles indicate up- and down-expressed genes, respectively. The color of edges ranging from grey to black corresponds to an order of increasing number of interactions. (**b**) A subnetwork contains PPARA and PPARG, their targets and regulators that regulate them. A two-layer hierarchical structure forms with 10 upstream regulators including nine miRNAs and E2F1 on the top layer, and two downstream regulators PPARA and PPARG arranged at the bottom layer.

2.4 Discussion

Instead of building gene-specific models, as done in most other studies[206], this study built a sample-specific model to identify the key common and subtype-specific regulators for DEGs in BC. Remarkably, most of the identified regulators identified have also been shown to be involved in the breast carcinogenesis in the published literature. Two common TFs (E2F1 and CITED2) were identified in each of the four subtypes. E2F1, a member of the E2F transcription factors, plays a crucial role in the control of cell cycle and the action of tumor suppressors[293]. *E2F1* was significantly upregulated in all four subtypes compared to normal samples in the TCGA data set (p-value < 0.05, t-test). CITED2 is a part of the CBP/p300 transcription complex and is known to increase *PPARA* transcription [294] and enhance ER mediated transcription [295]. CITED2 was reported to be a potent prognostic predictor associated with proliferation, migration and chemo-resistance in breast carcinoma [296]. The current study found that *CITED2* was significantly downregulated in BC compared to normal samples in the TCGA data set (p-value < 0.05, t-test).

Among the common seven miRNAs identified for all four BC subtypes, three miRNAs (miR-340-5p, miR-32-5p and miR-3150b-3p) were found to be upregulated (p-value < 0.05, t-test) while three miRNAs (miR-374b-5p, miR-664b-3p and miR-296-5p) were downregulated (p-value < 0.05, t-test) in each of the four BC subtypes versus normal samples. Only miR-361-3p showed no significant difference between each BC subtype versus normal samples. Targets of miR-340-5p include oncogenes such as *KRAS* and *MET* [297]. High levels of miR-374b-5p have been observed to correlate with favorable outcomes in TNBC [298]. miR-296-5p was shown to play a tumor-suppressive role by directly targeting Pin1 in prostate cancer [299].

Among the key regulators identified in the TNBC subtype, PPARA and PPARG are two members of peroxisome proliferator-activated receptors (PPARs) that belong to the nuclear hormone receptor (HR) superfamily which also includes ERs and PRs [300]. PPARs affect the expression of target genes involved in cell proliferation, cell differentiation and immune and inflammation responses [301]. A genetic variant of *PPARA* has been linked to BC risk [302]. NFKBIA inhibits the oncogene NF-κB, which is constitutively activated in ER-negative tumors and TNBCs, mediates adaptive resistance to ionizing radiation and chemotherapy, and is required for epithelial-mesenchymal transition associated with metastatic progression of BC. *NFKBIA* deletions, lead to a haploinsufficient effect on NFKBIA expression and thus the transactivation of several NF-κB target genes with important roles in TNBC carcinogenesis, are effected [303]. miR-423-5p has been observed as upregulated in drug-resistant BC cells from parental MDA-MB-231 cell line [304]. Downregulation of miR-92a-3p is associated with aggressive BC features and increased tumor macrophage infiltration [305]. miR-129-5p was reportedly downregulated in BC cells, in part because of promoter H3K27 methylation and it also regulates epithelial-mesenchymal transition and multi-drug resistance[306]. One of miR-129-5p target genes is *CBX4*, which is up-regulated in BC tissues and thus it promotes cell proliferation[307]. The TNBC-specific regulator miR-9-3p has also been shown to reduce gastric cancer cell invasion[308]. In the MDA-MB-231cell line, miR-9-3p has also been identified as a tumor-suppressor miRNA targeting β1 integrin thereby sensitizing the cell to MEK inhibition [309]. There are seven TNBC miRNA key regulators (miR-378a-5p, miR-708-5p, miR-889-3p, miR-3150b-3p, miR-374a-5p, miR-23a-3p and miR-28-5p), whose targets by gene set over-representation analysis were not enriched in canonical pathways. However, many of them are still involved in breast tumorigenesis. For example, miR-378a-5p is believed to be a switch regulating the Warburg effect in BC [310]. Studies in BC have

shown that miR-708-5p acted as a tumor suppressor through repression of invasion, proliferation, and potentially immune modulation [311]. miR-374a-5p could constitutively activate Wnt/β-catenin signaling to promote BC metastasis and thus may serve as an anti-metastasis therapeutic target in BC [312]. miR-23a-3p was shown to be up-regulated in a variety of human cancers including BC and exert an oncogenic function in human tumorigenesis [313].

When comparing the tumor-specific Lasso models with different number of input features, a significant improvement in cross-validation performance was observed when DNA methylation was added to the models. Accordingly, DNA methylation has recently been studied for early detection and prognosis of BC. Fackler *et al.* found that promoter methylation of four genes (*RASSF1A*, *CCND2*, *TWIST*, *HIN1*) was more frequently detected in breast tumors than in normal breast tissue [314]. In another study, ten hypermethylated genes (i.e., *APC*, *BIN1*, *BMP6*, *BRCA1*, *CST6*, *ESR2*, *GSTP1*, *CDKN2A*, *P21* and *TIMP3*) have been shown to differentiate between BC and normal breast tissues [315]. In the current study, *CCND2*, *TWIST*, *BIN1*, *BMP6* and *TIMP3* were identified as downregulated genes, underscoring the fact that DNA methylation is an important factor for explaining the DEG expression changes in breast tumors versus normal breast tissue.

It is noteworthy that many of the TNBC key regulators were found to regulate the FOXM1 and PPARA pathways, which were identified through GSEA as the top upregulated and the top downregulated pathways in TNBC, respectively. PPARA and PPARG, as two downregulated genes, were also identified as two key regulators in TNBC to regulate the PPARA pathway. In addition, PPARA is member of the PPARA pathway. Thus, a more complete understanding of the roles of PPARA and PPARG in TNBC may aid in determining development of new therapeutic strategies. NEK2, as a member of the FOXM1 pathway, was ranked first among all the genes from

this pathway in the GSEA result and was also identified as an upregulated gene in this study. Moreover, NEK2 was found to be regulated by four of the 24 TN regulators, suggesting its importance in TNBC tumorigenesis.

A limitation of this study is the use of the motif-based TF binding data. Studies have shown that the ChIP-Seq TF binding data conferred significantly higher explanatory power than the motif-based TF binding data for mRNA expression level [206]. The same phenomenon was observed in the current study that experimentally derived ChIP-Seq data from ReMap resulted in a significantly higher Spearman rank correlation than the motif-based TRRUST data (**Figure 2.2b**). Unfortunately, the ChIP-Seq data from ReMap were available for only a subset of the TFs in a limited number of BC cell lines that did not have representatives from each BC subtype. In fact, the majority were from MCF7 cell lines. Therefore, a reasonable alternative in the current study was to use the motif-based TF-target interaction data since it is not representative of any particular BC cell type.

2.5 Conclusions

Using the Lasso regularized linear regression approach, the subtype-specific as well as common TFs and miRNAs in BC were identified. Many of the key regulators identified in TNBC were found to regulate the FOXM1, PPARA and PPAR pathways. Therefore, further exploring these pathways and their regulators potentially could provide novel drug targets that may yield new treatment options for TNBC patients. The DEGs from the three pathways could also be used to build a multigene UDR signature to stratify TNBC further for prognosis. The resulting TNBC groups exhibited different clinical outcomes, which supports the utility of the approach used in this study.

Chapter 3: Identification of Compounds to Modulate the FOXM1 and PPARA Pathway Activities in Breast Cancer

This chapter is an unpublished manuscript to be submitted. All analyses, plots, tables, and written text in the thesis I produced myself.

3.1 Introduction

In Chapter 2, integrative analyses revealed 25, 20, 15 and 24 key TF and miRNA regulators in luminal A, luminal B, HER2-enriched and TNBC subtypes, respectively. Two TFs and seven miRNAs were identified in all four subtypes and thus were referred to as common regulators. Gene set over-representation analysis of targets of the key regulators was performed to investigate pathways potentially regulated by these regulators. miR-340-5p and E2F1 were two common regulators found to be regulating PID_FOXM1_PATHWAY (also referred to as FOXM1 pathway in this thesis). miR-340-5p and another common regulator miR-664b-3p were found to be regulators of BIOCARTA_PPARA_PATHWAY (also referred to as PPARA pathway in this thesis). Moreover, three other regulators (PPARA, PPARG, and miR-129-5p), which were identified in TNBC and together in one or two other subtypes, were found to regulate the PPARA pathway. miR-9-3p, which was identified as a key regulator in TNBC alone, was found to be regulating the PPARA pathway. In addition, the GSEA analysis in Chapter 2 identified the FOXM1 pathway as the top upregulated pathway and the PPARA pathway as the top downregulated pathway in TNBC samples versus normal breast tissue samples.

The FOXM1 pathway is a predefined pathway extracted from the C2 category gene sets in the MSigDB [104] . The FOXM1 pathway is involved in cell cycle control and DNA damage

repair and it ultimately promotes tumor cell proliferation. A total of 40 gene members are engaged in this pathway (**Appendix B**), including tumor suppressors (e.g., *BRCA2*, *CDKN2A*, *CHEK2* and *RB1*), the proto-oncogene family *MYC*, genes encoding cyclins (e.g., *CCNA2*, *CCNB1*, *CCNB2*, *CCND1* and *CCNE1*), genes encoding cyclin-dependent kinases (e.g., *CDK1*, *CDK2* and *CDK4*), *ESR1*, *NEK2* as well as *FOXM1* itself. FOXM1 is one of the most important oncogenic TFs and it is overexpressed in many human cancers [316]. It regulates all hallmarks of cancer, including proliferation, mitosis, epithelial-mesenchymal transition, invasion, and metastasis [317]. Not surprisingly, previously published studies have shown the critical role of FOXM1 in breast tumorigenesis and resistance to chemotherapy. In a study by Yang *et al.* the stable overexpression of FOXM1 was found to promote metastasis of breast cancer cells *in vivo* through stimulating the transcription of Slug, which promotes the epithelial-mesenchymal transition in BC [318]. Xue and colleagues found that the activation of SMAD3/SMAD4 by FOXM1 activated the TGF-β pathway and thus induced invasion and metastasis in BC [319]. The up-regulation of FOXM1 together with XIAP and Survivin induces resistance in breast tumor cells to docetaxel, paclitaxel, and epirubicin. In addition, FOXM1 is overexpressed in 85% of TNBCs [316] and is identified as the key transcriptional driver in the differentially expressed gene signature of TNBC [57]. FOXM1 promotes TNBC proliferation, invasion and progression by directly binding to and thus transcriptionally regulating expression of eEF2K [316]. FOXM1 also plays a role in autophagy by transcriptionally regulating *Beclin-1* and *LC3* genes in TNBCs [320]. Increased expression of the cAMP-response element-binding protein (CBP)/β-catenin/FOXM1 transcriptional complex in TNBC cells *in vivo* was associated with a high proportion of cancer stem cells, high rates of drug resistance and poor survival outcome [321]. In a recent study, Tan *et al.* constructed a gene regulation network in TNBCs and found that FOXM1 was in a key position in the network [322].

They further investigated the FOXM1 function in the TNBC cell line MDA-MB-231 and found that inhibiting FOXM1 can significantly suppress MDA-MB-231 cell tumorigenesis *in vivo* using a mouse xenograft model [322]. Zhang *et al.* found that DEP (dishevelled, EGL-10, pleckstrin) domain-containing (*DEPDC1*) was over-expressed in human TNBCs relative to their paired neighboring non-cancerous tissues using two GEO datasets [323]. Stable DEPDC1 over-expression can facilitate cell proliferation and tumor growth via upregulating FOXM1 in MDA-MB-436 cells and BT549 cells [323]. Taken together, these results suggest that inhibition of FOXM1 function is a potential strategy to treat TNBC.

The PPARA pathway is a predefined pathway extracted from the C2 category gene sets in the MSigDB database [104]. This pathway can induce tumor cell apoptosis and includes 57 gene members (**Appendix D**), such as the tumor suppressors *RB1* and *PIK3R1*, the proto-oncogenes *MYC* and *JUN*, as well as transcription factors *CITED2* and *PPARA*. The stimulation of the PPARA pathway increases the volume and number of peroxisomes which are responsible for, among other things, lipid metabolism and catabolism. Genes in this pathway are regulated by the PPARA transcription factor. Previous studies suggest that PPARA is a tumor suppressor in some cancers, including melanoma [324] and glioblastoma [325]. PPARA also appears to inhibit cell proliferation and tumorigenesis and induces degradation of the proto-oncogene Bcl2 which inhibits apoptosis in developing tumor cells [326]. Moreover, a group of co-expressed genes including *LPL*, *SORBS1*, *PPARG*, *PLIN*, *FABP4*, *AQP7*, *CD36*, and *ADIPOQ* that are involved in the PPARA signaling pathway may also inhibit the pathway and contribute to breast tumor progression [327]. Recently, Saleh *et al.* found that PD-L1 blockade by atezolizumab in the human TNBC cell line MDA-MB-231 downregulated signaling pathways involved in tumor growth, metastasis, and hypoxia, including the PPARA/retinoid receptor a (RXRa) activation pathway [328]. To study the

mechanisms by which adipose tissue in obesity promotes BC progression, Blucher *et al.* treated TNBC cells with adipose tissue conditioned media generated from fat tissue of obese female patients [329]. The adipose tissue treatment changed the expression profiles of TNBC cells resulting in altered expression of many genes regulated by PPAR nuclear receptors [329]. Thus, adipose tissue generated factors that altered PPAR-regulated gene expression and lipid metabolism, and further promoted TNBC progression. These results suggest that the PPAR pathway has potential targets that could be used to develop treatments for TNBCs [329].

Overall, findings from Chapter 2 of this thesis together with the published literature strongly support the roles of the FOXM1 and PPARA pathways in BC (especially TNBC) tumorigenesis. Therefore, identifying compounds that can modulate these two pathways may provide novel therapeutic strategies for BC, and especially TNBC.

In order to assist in identifying novel compounds that might modulate the FOXM1 and PPARA pathways, the CMAP database [330] (see section 1.6.4) was used in this study. CMAP has been widely used to identify novel therapeutic targets for a disease by establishing advantageous connections between the drug treatment and the patient response (phenotypic response) [278,331]. When it is not feasible to directly measure a pathway signaling activity it can be approximated using gene expression. The De-noising Algorithm based on Relevance network Topology (DART) method [332] measures an activity score of a pathway or signature for independent samples in a gene expression matrix while filtering out noise before estimating pathway/signature activity. There are four main steps of the DART algorithm: 1) construct a relevance network where nodes are genes in a given pathway/signature and edges represent the correlations among the genes based on the expression data matrix, 2) evaluate the consistency of the derived relevance network by comparing the gene-gene correlations (i.e., edges) in the network

to the prior information contained in the given pathway/signature, 3) prune the relevance network by removing edges that are not consistent with the prior information contained in the given pathway/signature (this step is called the denoising step), and 4) calculate an activity score for the given pathway/signature based on the pruned relevance network for each sample in the gene expression data matrix. The DART algorithm has been developed into a R package with the same name in which the *DoDART* function can calculate pathway activity scores using the DART algorithm.

The purpose of this study was to investigate the suppression of the FOXM1 pathway and the stimulation of the PPARA pathway on BC cell lines with various drug treatments using the CMAP database. The pathway activity scores of treated and untreated tumor cell lines were first calculated based on the DART algorithm. We then observed and compared the activity score difference in order to identify the drug candidates that could modulate the FOXM1 and PPARA pathway activities. Effective drugs are defined as those that reverse the pathway activities that are activated or inactivated in BC. Therefore, we would expect that an effective drug for the treatment of BC would show FOXM1 pathway inhibition and or PPARA pathway activation. Put another way, the FOXM1 pathway activity score would be decreased, whereas the PPARA pathway activity score would be increased in cells treated with effective drugs. Moreover, we also used the CMAP database tool to identify drugs that could reverse the expression pattern of the two pathways in BC cell lines and compared the drugs identified by the two methods.

3.2 Materials and Methods

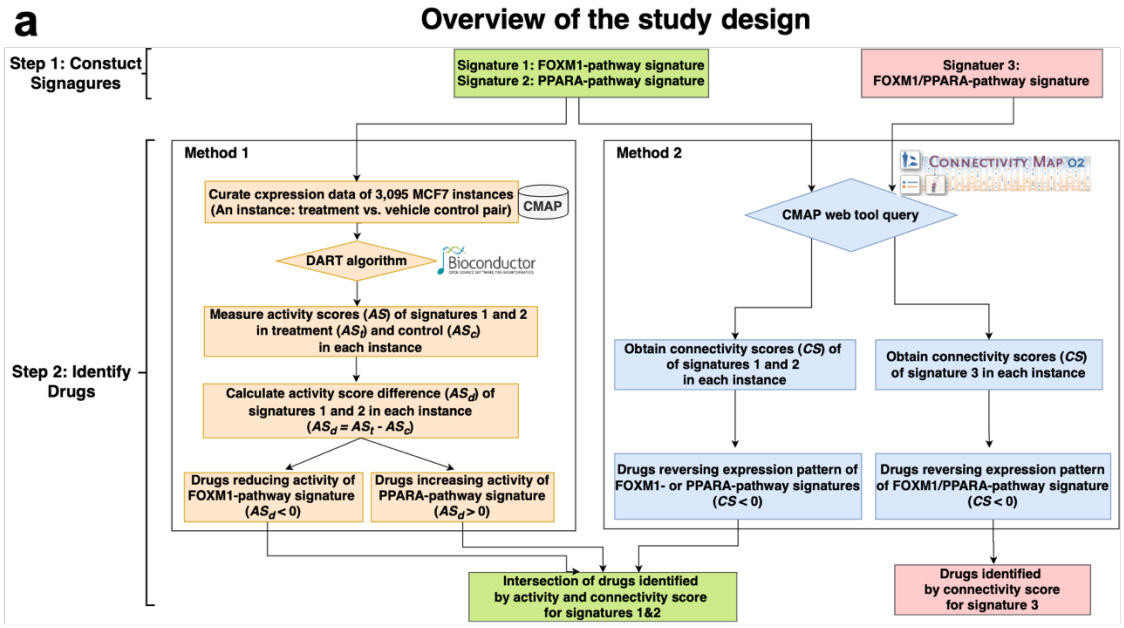
3.2.1 Gene expression data and processing

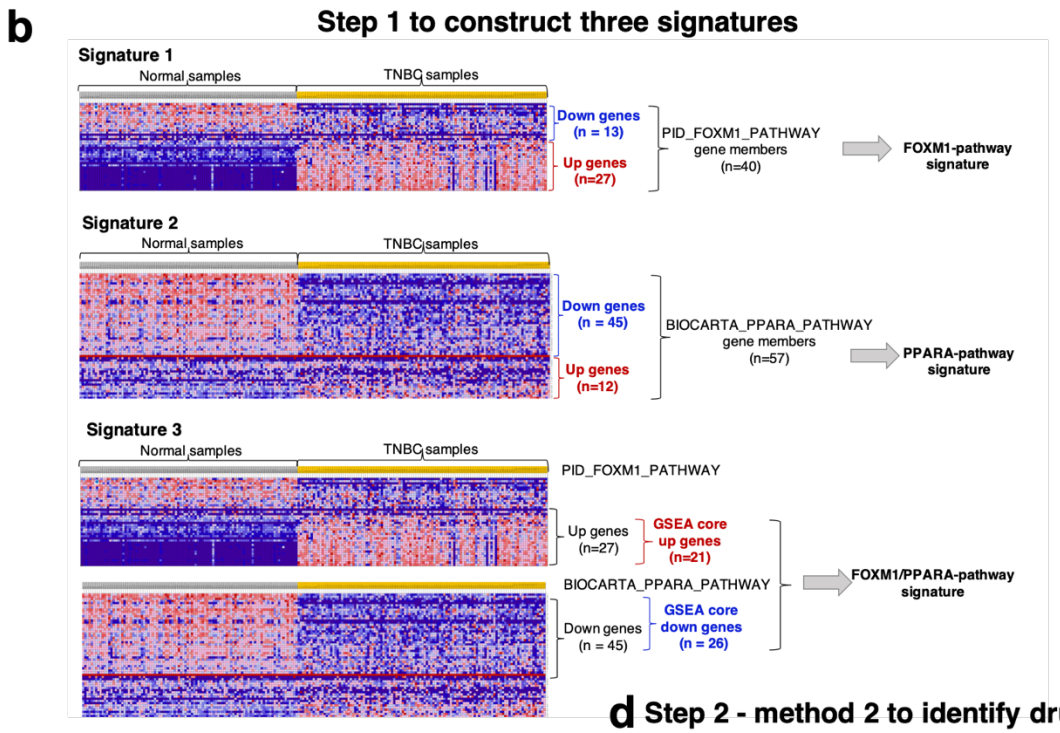
The overall approach is depicted in **Figure 3.1a**. We collected gene expression data from three major resources. The CMAP build 02 database (http://www.broadinstitute.org/cmap), which includes 3,095 gene expression instances (treatment vs. vehicle control pairs) from MCF7 cell lines treated with 1,294 bioactive small chemical molecules at varying concentrations, was used in this study. The raw CEL files were downloaded. Since the CMAP database is based on three different Affymetrix chip types (HG-U133A, HT_HG-U133A and U133AAofAv2), the microarray data were then grouped according to the platforms and the Robust Multichip Average (RMA) [333] method was used to normalize the expression profiles from each chip type. Thirdly, probe IDs were then mapped to gene symbols using the corresponding platform files. If a probe was mapped to multiple or zero genes, the data of this probe were discarded. If multiple probes were mapped to the same gene, the maximum value of the multiple probes was taken as the expression value for the gene. Finally, we kept only genes presenting in all the three chip types and applied the *ComBat* function in the R package sva [334] to remove batch effects. Thus, normalized gene expression data from different platforms could be used for further study.

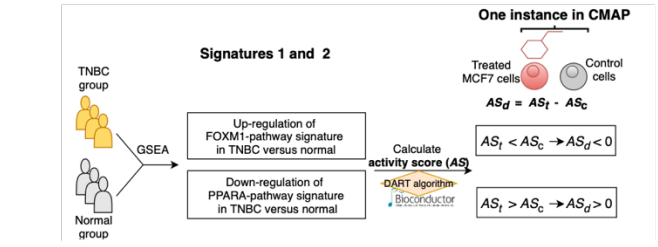
For the TCGA breast tumor cohort, we used the estimated fractions of genes computed by the RSEM method [92] provided by Firehose Broad GDAC (https://gdac.broadinstitute.org), multiplied by 10^6 to obtain Transcripts Per Million (TPM) [92] and log2-transformed. Subtypes were defined for each sample according to the IHC status of ER, PR and HER2 proteins provided in the clinical data.

GSE48213 contains molecular profiles (pre-treatment measurements of mRNA expression, CNV, protein expression, promoter DNA methylation, and gene mutation) from a collection of 84

BC cell lines reported in Daemen *et al.*'s publication [260]. In the current study, baseline expression data in Fragments Per Kilobase of transcript per Million mapped reads (FPKM) values that are available for 56 BC cell line samples were extracted from GSE48213 [260], converted into TPM [335], and log-transformed log2(TPM+1). Cell line subtype information was obtained from Daemen *et al.*'s study [260].


a
Overview of the study design
Step 1: Constuct Signagures
Signature 1: FOXM1-pathway signature
Signature 2: PPARA-pathway signature
Signatuer 3: FOXM1/PPARA-pathway signature
Step 2: Identify Drugs
Method 1
Curate cxpression data of 3,095 MCF7 instances (An instance: treatment vs. vehicle control pair)
CMAP
DART algorithm
Bioconductor
Measure activity scores (AS) of signatures 1 and 2 in treatment (AS_t) and control (AS_c) in each instance
Calculate activity score difference (AS_d) of signatures 1 and 2 in each instance (AS_d = AS_t - AS_c)
Drugs reducing activity of FOXM1-pathway signature (AS_d < 0)
Drugs increasing activity of PPARA-pathway signature (AS_d > 0)
Method 2
CONNECTIVITY MAP 02
CMAP web tool query
Obtain connectivity scores (CS) of of signatures 1 and 2 in each instance
Obtain connectivity scores (CS) of signature 3 in each instance
Drugs reversing expression pattern of FOXM1- or PPARA-pathway signatures (CS < 0)
Drugs reversing expression pattern of FOXM1/PPARA-pathway signature (CS < 0)
Intersection of drugs identified by activity and connectivity score for signatures 1&2
Drugs identified by connectivity score for signature 3


b
Step 1 to construct three signatures
Signature 1
Normal samples
TNBC samples
Down genes (n = 13)
Up genes (n=27)
PID_FOXM1_PATHWAY gene members (n=40)
FOXM1-pathway signature
Signature 2
Normal samples
TNBC samples
Down genes (n = 45)
Up genes (n=12)
BIOCARTA_PPARA_PATHWAY gene members (n=57)
PPARA-pathway signature
Signature 3
Normal samples
TNBC samples
PID_FOXM1_PATHWAY
Up genes (n=27)
GSEA core up genes (n=21)
BIOCARTA_PPARA_PATHWAY
Down genes (n = 45)
GSEA core down genes (n = 26)
FOXM1/PPARA-pathway signature


c
Step 2 - method 1 to identify drugs
Signatures 1 and 2
TNBC group
Normal group
GSEA
Up-regulation of FOXM1-pathway signature in TNBC versus normal
Down-regulation of PPARA-pathway signature in TNBC versus normal
Calculate activity score (AS)
DART algorithm
Bioconductor
One instance in CMAP
Treated MCF7 cells
Control cells
AS_d = AS_t - AS_c
AS_t < AS_c → AS_d < 0
AS_t > AS_c → AS_d > 0

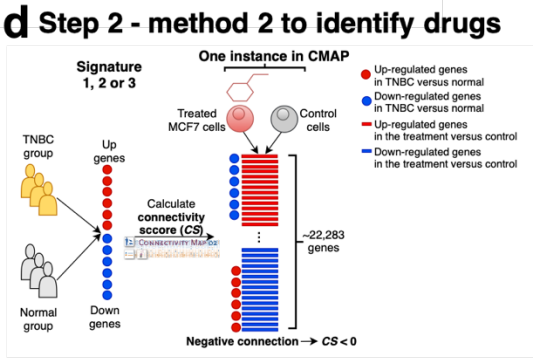

d
Step 2 - method 2 to identify drugs
Signature 1, 2 or 3
One instance in CMAP
Treated MCF7 cells
Control cells
TNBC group
Normal group
Up genes
Down genes
Calculate connectivity sccore (CS)
CONNECTIVITY MAP 02
~22,283 genes
Negative connection → CS < 0
Up-regulated genes in TNBC versus normal
Down-regulated genes in TNBC versus normal
Up-regulated genes in the treatment versus control
Down-regulated genes in the treatment versus control

Figure 3.1 Study design
(**a**) The three major pipelines of this study are depicted. Initially, three signatures were built: two signatures built with each of the FOXM1 and PPARA pathways alone and a third signature built with the two pathways together (**b**). Next, the pathway activity scores of the FOXM1- and PPARA-pathway signatures on genome-wide expression data with drug treatment were computed by the DART algorithm to identify drugs inducing the activity score changes of the two pathways (**c**). In parallel, the connectivity scores for the FOXM1- and PPARA-pathway signatures were computed by the CMAP online tool to identify drugs reversing the expression pattern of the two pathway signatures (**d**). Finally, the intersection of drugs identified from the two pipelines were considered as drugs influencing the activity of the FOXM1 or PPARA pathways. In the third pipeline, the connectivity scores for the FOXM1/PPARA-pathway signature were computed by CMAP online tool to identify drugs reversing expression pattern of the two pathways simultaneously (**d**).

3.2.2 Pathway-based signature construction

Three gene signatures were constructed based on the GSEA results from the two pathways (the FOXM1 and PPARA pathways) in Chapter 2 (see section 2.2.6 and 2.3.4 in Chapter 2) (**Figure 3.1b**). A typical signature was represented by two subsets of genes ("up" tags and "down" tags). The FOXM1-pathway signature was built based on the expression pattern of the 40 gene members in the FOXM1 pathway between TNBC samples and normal breast tissue samples observed in the GSEA results in Chapter 2 (**Appendix A** and **B**). We grouped those genes which are upregulated in TNBC relative to normal breast tissue into "up" while those genes which are downregulated in TNBC compared to normal breast tissue were grouped into "down" (**Figure 3.1b**). Similarly, based on the expression pattern of the 57 gene members form the PPARA pathway between TNBC and normal breast tissue samples observed in the GSEA results in Chapter 2 (**Appendix C** and **D**), the PPARA-pathway signature was built. We grouped those genes which are upregulated in TNBC relative to normal breast tissue into "up" while those which are downregulated in TNBC compared to normal breast tissue were grouped into "down" (**Figure 3.1b**).

Given that the FOXM1 pathway is the top upregulated pathway while the PPARA pathway is the top downregulated pathway in TNBC versus normal breast tissue, the signatures from both pathways were simplified into a single two-pathway signature (FOXM1/PPARA-pathway signature) by selecting genes that are most responsible for upregulation of the FOXM1 pathway and downregulation of the PPARA pathway (**Figure 3.1b**). According to GSEA, for an enriched

gene set, the core members contribute most to the gene set enrichment [114]. Thus, the core members of the FOXM1 pathway which accounted for the upregulation of the FOXM1 pathway in TNBC versus normal (**Appendix B**) were selected and tagged as "up". The core members of the PPARA pathway which accounted for the downregulation of the PPARA pathway in TNBC versus normal (**Appendix D**) were selected and tagged as "down". Thus, the two-pathway signature was represented by the core gene members from the FOXM1 and the PPARA pathways, respectively.

3.2.3 Pathway activity score calculation

To calculate the activity score of the FOXM1- and PPARA-pathway signatures, the *DoDART* function from the DART R package was used. By inputting a gene expression matrix and a predefined signature, the *DoDART* function performed the following steps to calculate pathway activity scores using the DART algorithm: 1) constructing the relevance network where nodes are genes in the signature and edges, represent the correlation among genes based on the expression matrix. The significance of the correlation between a gene pair is defined using the default FDR value of 1.0×10^{-6}, which is stringent and represents a conservative Bonferroni threshold that assumes that a typical signature consists of the order of 100 genes thereby necessitating an estimated 10000 pairwise gene correlations; 2) evaluating the consistency of the gene-gene correlations (i.e., edges) in the relevance network with the prior gene-gene correlations contained in the given signature; 3) filtering out edges inconsistent with the prior information contained in the given signature from the relevance network; and 4) estimating an activity score of the given signature for each individual sample in the gene expression data matrix. A high positive pathway activity score indicates the stimulation of the pathway activity in a sample while high negative score indicates the repression of the pathway activity.

Considering the difference between cancer tissue and cell lines in terms of transcriptional profiles, we asked whether the activity scores of the FOXM1- and PPARA-pathway signatures across different BC subtypes were showing a similar pattern between BC tissue and cell lines. We computed the activity scores of the two signatures for BC cell line samples from the GSE48213 expression matrix and also for BC tissue samples from the TCGA expression matrix. For each signature, we performed the ANOVA test to check the difference of activity score in three BC cell line subtypes in GSE48213 (luminal, HER2 and TNBC). To compare the means of a signature activity score of any two BC cell line subtypes in GSE48213, we performed pairwise t-test between every two subtypes followed by Benjamini-Hochberg correction for multiple testing. For each signature, we also performed the ANOVA test to check the difference of the signature activity score in five BC tissue sample groups in TCGA (luminal A, luminal B, HER2-enriched, TNBC, and normal breast tissue). We performed pairwise t-test between every two BC tissue sample groups followed by Benjamini-Hochberg correction for multiple testing to compare the means of a pathway activity score of any two BC tissue sample groups in TCGA.

To assess the effect of a treatment on the activity of the FOXM1 or PPARA pathways, we applied the *DoDART* function to the CMAP expression profiles and calculated the FOXM1- and PPARA-pathway signature activity scores for each CMAP instance (i.e., the treatment and the paired vehicle control) (**Figure 3.1c**). In the case of multiple controls per treatment, we removed the control with the highest and lowest activity scores as outliers, and then used the mean of the rest as the control. Some compounds with the same dose were exposed to the MCF7 cells for multiple times. In this case, we took the average of the activity scores of the FOXM1 and PPARA pathways. In the end, activity scores of the two pathways for 1,390 MCF7 instances (i.e., treatment vs. vehicle control pairs) were obtained.

3.2.4 Identification of drugs modulating pathway activity scores

Next, we evaluated drug treatments affecting pathway activity in the context of the change of pathway activity score between treated and control cell lines in CMAP (**Figure 3.1c**). For each of the two pathways in a given instance, we defined the activity score difference (AS_d) as the activity score in the treatment (AS_t) minus the activity score in the control (AS_c) (**Equation 3.1**).

$$AS_d = AS_t - AS_c \qquad \textbf{Equation 3.1}$$

The magnitude of AS_d is the degree of change in pathway activity score caused by the corresponding treatment, while the sign of AS_d is the direction of that change. Therefore, a positive sign indicates the treatment increases the pathway activity while a negative sign decreases the pathway activity. We ordered the instances according to increasing difference score for the FOXM1 pathway but decreasing difference score for the PPARA pathway. The top-ranked instances were further investigated. To be more specific, the top-ranked instances for the FOXM1 pathway are those which decrease the pathway activity while for PPARA pathway are those which increase the pathway activity.

3.2.5 Connectivity Map query of drugs modulating pathway expression patterns

In comparison to drug treatments affecting pathway activity in the context of the change of pathway activity score between treated and control cell lines in CMAP, we also evaluated drug treatments in the context of modulating (mimicking or reversing) pathway expression patterns using the CMAP online tool (**Figure 3.1d)**. The three signatures (the FOXM1-pathway signature, the PPARA-pathway signature, and the FOXM1/PPARA-pathway signature) were used as the query signatures to perform the CMAP analysis through the CMAP build 02 web interface. To query the CMAP online tool with a given signature, we first changed the gene list from gene symbols to Affymetrix probe IDs, which are required as input into the CMAP. This probe list was collated into tag sets of "up" or "down" genes and queried against the CMAP database to generate

hits. The similarity between the gene expression profile of the query signature and that of a CMAP instance was measured by the connectivity score, which ranged from -1 to 1. When a query signature receives a high positive score for a treatment, it means with this treatment the up-regulated (i.e., "up") genes in the signature are also up-regulated while the down-regulated (i.e., "down") genes in the signature are also down-regulated. Thus, a high positive connectivity score indicates that the corresponding compounds induced the same changes in expression of the query signature as those caused by BC. When a query signature receives a high negative score for a treatment, it means that the up-regulated genes in the query signature are down-regulated by the treatment while the down-regulated genes in the signature are up-regulated by the treatment (**Figure 3.1d**). Therefore, a high negative connectivity score indicates that the corresponding compounds reversed the expression of the query signature. In the current study, for a query signature, we expected the up-regulated genes to be down-regulated while the down-regulated genes to be up-regulated after a particular treatment. So those treatments returning a large negative connectivity score for the three query signatures are treatments of interest. In other words, we expected to identify the compounds which could reverse the expression patterns of the three signatures.

3.3 Results

3.3.1 Three pathway-based signatures

Gene lists for the three signatures are provided in **Table 3.1**. According to the GSEA results in Chapter 2 (see section 2.2.6 and 2.3.4 in Chapter 2), of the 40 genes in the FOXM1 pathway, 27 were upregulated and 13 downregulated in TNBC compared to normal breast tissue. Thus, the 40-gene FOXM1 pathway signature was represented by two subsets of genes ("up" tags and "down"

tags). In the same way, the 57 genes in the PPARA pathway were divided into two parts, one (12 genes) for upregulation and the other (45 genes) for downregulation. It is noteworthy that the majority of the FOXM1 pathway genes are upregulated while the majority of the PPARA pathway genes are downregulated in TNBC, which is in accordance with the upregulation of the FOXM1 pathway and the downregulation of the PPARA pathway in TNBC identified by the GSEA analysis in Chapter 2.

Given the trend of upregulation of the FOXM1 pathway and downregulation of the PPARA pathway in TNBC, the signatures from both pathways were simplified into a single two-pathway signature by identifying genes that are most responsible for FOXM1 pathway upregulation and PPARA pathway downregulation. Among the 27 upregulated genes from the FOXM1 pathway, 21 were core members that contribute most to the pathway upregulation in the GSEA results in Chapter 2 (**Appendix B**). Among the 45 downregulated genes from the PPARA pathway, 26 were core members that contribute most to the pathway downregulation in the GSEA results in Chapter 2 (**Appendix C**). We used the 21 core gene members from FOXM1 pathway as the “up" genes and the 26 core gene members from the PPARA pathway as the “down” genes to build the FOXM1/PPARA-pathway signature.

Table 3.1 The three pathway-based signatures

Signature	Gene member	Gene number	Tag
FOXM1-pathway signature	*AURKB, BIRC5, BRCA2, CCNA2, CCNB1, CCNB2, CCND1, CCNE1, CDC25B, CDK1, CDK2, CDK4, CDKN2A, CENPA, CENPB, CENPF, CHEK2, CKS1B, ESR1, FOXM1, GSK3A, HIST1H2BA, NEK2, ONECUT1, PLK1, SKP2, XRCC1*	27	up
	CREBBP, EP300, ETV5, FOS, GAS1, LAMA4, MAP2K1, MMP2, MYC, NFATC3 RB1, SP1, TGFA	13	down
PPARA-pathway signature	*APOA1, APOA2, DUT, HSP90AA1, HSPA1A, INS, MED1, MRPL11, NCOR2, NR0B2, RELA, SRA1*	12	up
	ACOX1, CD36, CITED2, CPT1B, CREBBP, DUSP1, EHHADH, EP300, FABP1, FAT1, HSD17B4, JUN, LPL, MAPK1, MAPK3, ME1, MYC, NCOA1, NCOR1, NFKBIA, NOS2, NR1H3, NR2F1, NRIP1, PDGFA, PIK3CA, PIK3CG, PIK3R1, PPARA, PPARGC1A, PRKACB, PRKACG, PRKAR1A, PRKAR1B, PRKAR2A, PRKAR2B, PRKCA, PRKCB, PTGS2, RB1, RXRA, SP1, STAT5A, STAT5B, TNF	45	down
FOXM1/PPARA-pathway signature	*AURKB, BIRC5, BRCA2, CCNA2, CCNB1, CCNB2, CCNE1, CDC25B, CDK1, CDK2 CDK4, CDKN2A, CENPA, CENPF, CHEK2, CKS1B, FOXM1, GSK3A, NEK2, PLK1, SKP2*	21	up
	STAT5B, HSD17B4, PIK3R1, DUSP1, CD36, CITED2, STAT5A, SP1, PRKAR1A, NCOA1, JUN, MAPK3, LPL, PRKAR2B, NCOR1, RB1, RXRA, NR2F1, EHHADH, ACOX1, PRKACB, NRIP1, PDGFA, NR1H3, CREBBP, PRKAR2A	26	down

3.3.2 FOXM1 pathway activity in breast cancer tissue and cell line samples

We calculated the activity score of the FOXM1 pathway for TCGA breast tumor samples, including 424 luminal A, 121 luminal B, 37 HER2-enriched, 112 TNBC, and 112 normal breast tissue samples. There are significant differences among the five groups regarding the FOXM1 pathway activity score (ANOVA, p-value $< 2.2 \times 10^{-16}$) (**Figure 3.2a**). To further check the FOXM1 pathway activity score difference between any two groups, we performed pairwise comparisons between every two groups using t-test followed by Benjamini-Hochberg corrections for multiple testing. We found that the activity score is significantly lower in the normal breast

tissue group than each of the four TCGA BC groups (luminal A, luminal B, HER2-enriched, and TNBC) (**Figure 3.2a**). Among the four TCGA BC subtypes, the TNBC group appears to show a higher FOXM1 activity score than the luminal A, luminal B, and HER2-enriched groups but the results are not significant (**Figure 3.2a**). We note that the GSEA results in Chapter 2 that showed that the FOXM1 pathway was the most significantly upregulated pathway in TNBC (**Appendix A**).

Although there are well documented differences between BC cell lines and BC tissue samples, we expected to see at least some similarity between their pathway activity scores given that BC cell lines are ultimately derived from BC tissue samples. With the gene expression data and subtype annotation from the GSE48213 data set, we obtained data for 14 luminal, 21 HER2-enriched and 10 TNBC cell lines. The GSE48213 data set included two matched normal-like BC lines, on which RNA-Seq was not performed and thus were excluded. We calculated the activity scores of FOXM1 pathway for the GSE48213 BC cell lines. The FOXM1 pathway activity scores showed a significant difference among the three GSE48213 BC cell line subtypes (ANOVA, *p*-value = 0.003) (**Figure 3.2b**). Similar to the results in the four TCGA BC subtypes (**Figure 3.2a**), the TNBC cell line group shows higher FOXM1 activity score than the luminal and HER2-enriched cell line groups (**Figure 3.2b**). The FOXM1 pathway activity score is significantly different between the TNBC and HER2-enriched subtypes but is not significantly different between the TNBC and luminal subtypes. The similarity of the FOXM1 pathway activity score in TCGA breast tumor tissue samples and GSE48213 BC cell lines suggests that the compounds that can modulate the FOXM1 pathway in BC cell lines could also potentially modulate the pathway in breast tumor tissue samples.

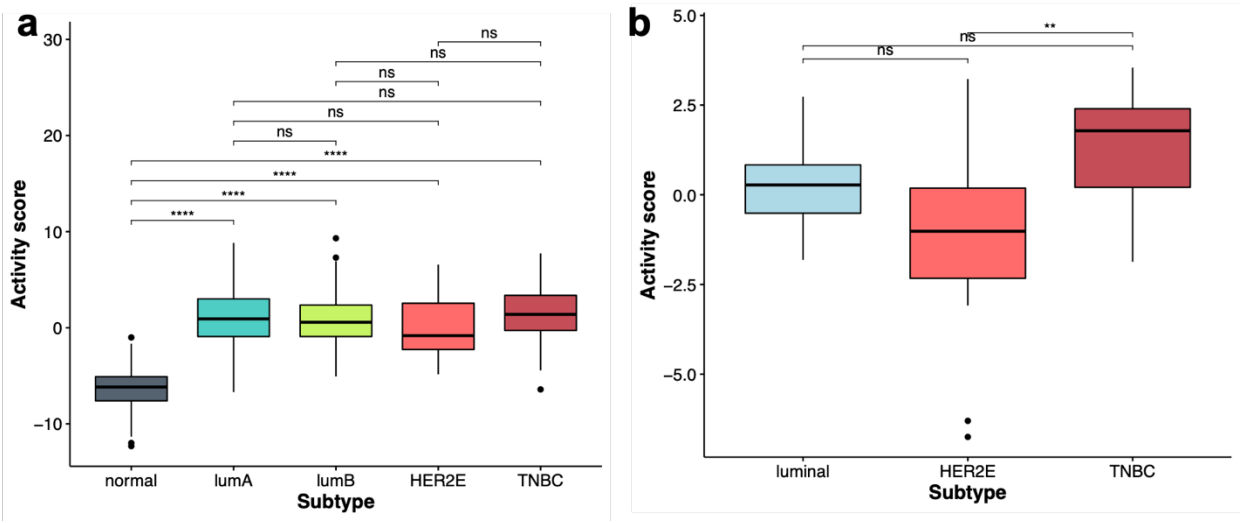


Figure 3.2 The FOXM1 pathway activity score in breast cancer
(**a**) The FOXM1 pathway activity score in TCGA BC tissue samples. (**b**) The FOXM1 pathway activity score in GSE48213 BC cell line samples. The pairwise comparison between every two groups used t-test followed by Benjamini-Hochberg correction. The significance level of adjusted *p-values* is as follows: ns, $p > 0.05$; *, $p \leq 0.05$; **, $p \leq 0.01$; ***, $p \leq 0.001$; ****, $p \leq 0.0001$.

3.3.3 PPARA pathway activity in breast cancer tissue and cell line samples

The activity score of the PPARA pathway was calculated for TCGA breast tumors and BC cell lines from GSE48213. The PPARA pathway activity score is significantly different among the five TCGA BC tissue groups (ANOVA, *p*-value $< 2.2 \times 10^{-16}$) (**Figure 3.3a**). The normal tissue group shows a significantly higher PPARA pathway activity score than each of the four breast cancer tissue groups (**Figure 3.3a**). No significant difference in terms of PPARA pathway activity score was observed between any two of the four TCGA BC subtypes. The activity of PPARA pathway does not show significant difference among the three GSE48213 BC cell line subtypes (ANOVA, *p*-value = 0.15) (**Figure 3.3b**), similar to the results observed in TCGA breast tumor tissue samples. The similarity of the PPARA pathway activity score in TCGA breast tumor tissue samples and GSE48213 BC cell lines suggests that the compounds that can modulate the PPARA pathway in BC cell lines could also potentially modulate the pathway in breast tumor tissue samples.

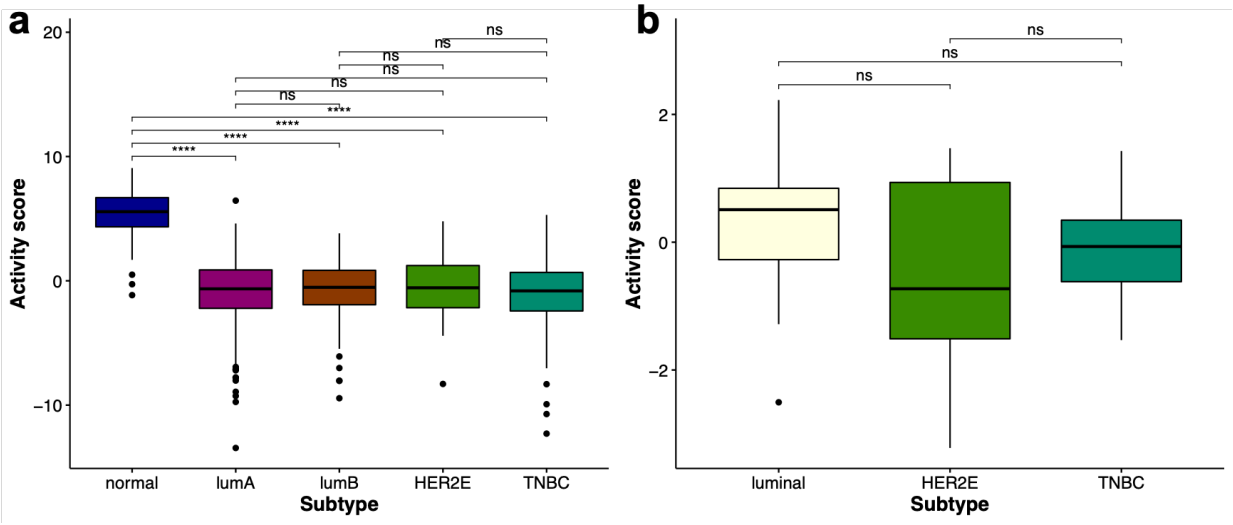


Figure 3.3 The PPARA pathway activity score in breast cancer
(**a**) The PPARA pathway activity score in TCGA BC patients. (**b**) The PPARA pathway activity score in GSE48213 BC cell lines. The pairwise comparison between every two groups used t-test followed by Benjamini-Hochberg correction. The significance level of adjusted *p-values* is as follows: ns, $p > 0.05$; *, $p \leq 0.05$; **, $p \leq 0.01$; ***, $p \leq 0.001$; ****, $p \leq 0.0001$.

3.3.4 Identification of drugs modulating the FOXM1 pathway activity

We evaluated drug treatments in terms of their effect on the FOXM1 pathway activity score. Since the FOXM1 pathway has a high positive activity score (i.e., activated) in BC while a high negative activity score (i.e., inactivated) in normal, an effective drug treatment should reduce the FOXM1 pathway activity score (i.e., suppressing the pathway activity) and thus have a large negative activity score difference (AS_d) between treatment and control. Of the 1,390 instances, 885 instances show a negative difference score for FOXM1 pathway (i.e., the activity score in the treatment is smaller than that in the paired control) (blue bars in **Figure 3.4a**), suggesting that the corresponding treatment could suppress the FOXM1 pathway activity. In comparison, we also evaluated drug treatments in terms of their effect on reversing the FOXM1 pathway expression pattern. Since the majority of gene members in the FOXM1 pathway are over-expressed and the minority are under-expressed in BC. We would expect that an effective drug would reverse the expression pattern by downregulating those over-expressed genes while upregulating those under-

expressed genes. Such effective drugs would show a large negative connectivity score. Out of the 1,390 instances, 1,039 generated a negative connectivity score for the FOXM1 pathway (blue bars in **Figure 3.4b**), suggesting that the corresponding treatment could reverse the FOXM1 pathway expression pattern. Ordering the instances according to increasing difference activity and connectivity scores, respectively and after extracting the top 50 instances from each case we found that 19 drugs decreased the FOXM1 pathway activity scores and reversed the FOXM1 pathway expression pattern simultaneously (**Figure 3.4c**). We listed the 19 drugs with details in **Table 3.2**.

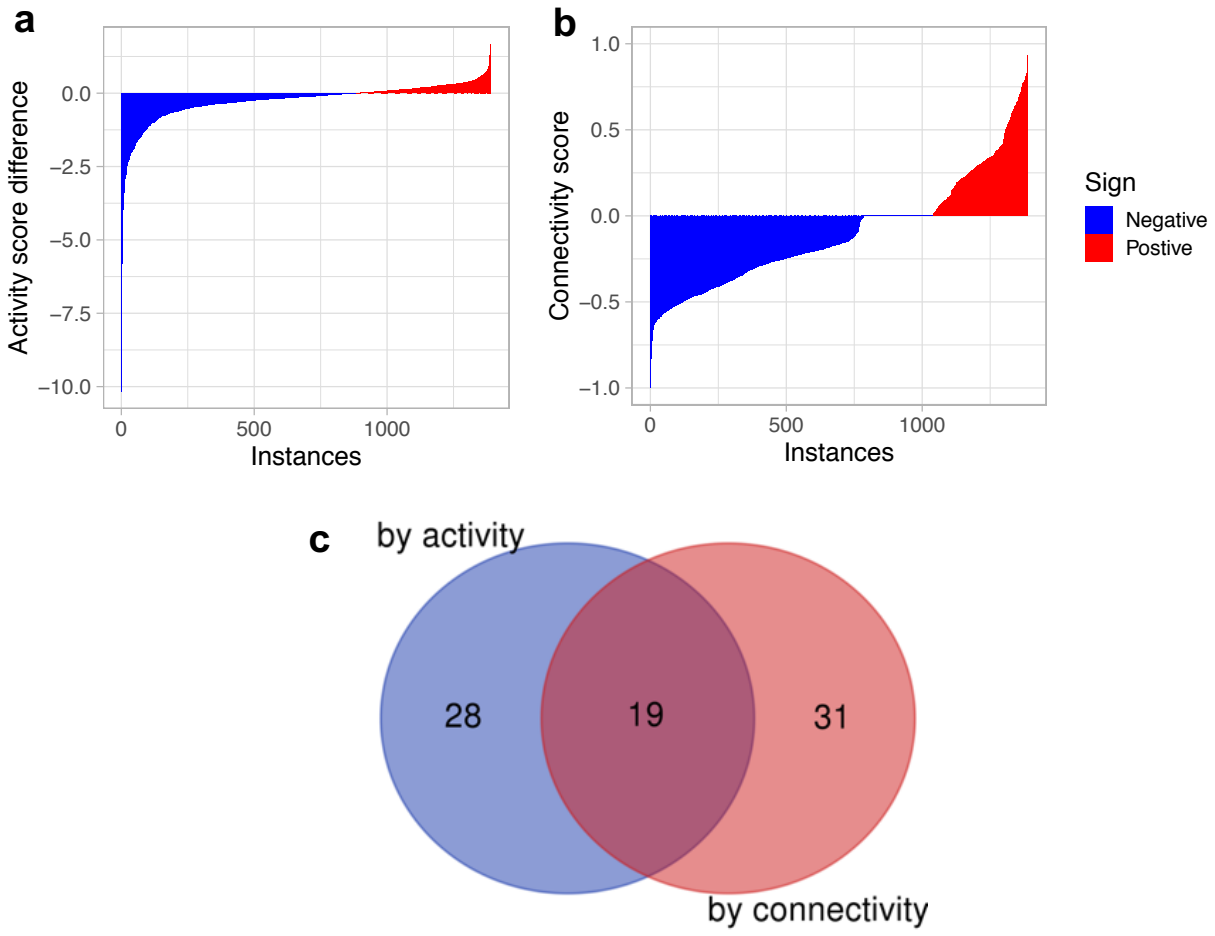


Figure 3.4 Effects of instances on FOXM1 pathway activity and expression pattern
(**a**) The FOXM1 pathway activity score difference of 1,390 CMAP instances. The y-axis indicates the difference between the activity score in the treatment and the paired control for each instance. The x-axis indicates the instances ordered by the corresponding activity score difference increasingly. (**b**) The connectivity score of 1,390 CMAP instances. The y-axis indicates the connectivity score. The x-axis indicates the instances ordered by the corresponding connectivity score increasingly. (**c**) The intersection between the top 50 drugs identified by the activity score difference and the activity score.

Table 3.2 The top effective drugs on FOXM1 pathway

Drug name	Dose (μM)	N [a]	Rank by		Activity score			Connectivity score
			Activity difference (AS_d) [b]	Connectivity score	Treatment (AS_t)	Control (AS_c)	Difference (AS_d)	
5109870	25	1	1	10	-8.57	1.59	-10.16	-0.66
MG-132	21	1	2	1	-4.45	1.97	-6.42	-1
MG-262	0.1	1	4	3	-2.95	1.79	-4.74	-0.83
celastrol	2.5	1	5	7	-2.99	1.59	-4.58	-0.72
resveratrol	50	2	6	8	0.92	5.2	-4.28	-0.7
ciclopirox	15	2	7	37	-2.45	1.53	-3.98	-0.59
pyrvinium	3.4	2	11	2	-1.85	1.56	-3.41	-0.84
emetine	7.2	2	13	14	-1.37	1.71	-3.08	-0.63
15-delta prostaglandin J2	10	8	14	11	-1.18	1.73	-2.91	-0.66
cephaeline	6	3	15	32	-0.95	1.96	-2.91	-0.6
puromycin	7.4	2	16	17	-0.94	1.89	-2.84	-0.62
parthenolide	16.2	2	23	18	-1.43	1.03	-2.46	-0.62
azacitidine	16.4	2	24	49	-0.87	1.53	-2.39	-0.57
cicloheximide	14.2	2	28	45	-0.3	2	-2.29	-0.57
astemizole	8.8	2	29	27	0.04	2.3	-2.26	-0.6
5224221	12	2	36	25	-1.42	0.61	-2.03	-0.6
scriptaid	10	1	42	20	1.32	3.29	-1.97	-0.62
ouabain	5.4	2	43	30	-0.34	1.63	-1.97	-0.6
bepridil	10	2	49	13	0.12	2	-1.88	-0.63

[a] The number of experiments;
[b] Activity difference = activity score in treatment (AS_t) - activity score in control (AS_c).

3.3.5 Identification of drugs modulating the PPARA pathway activity

We evaluated drug treatments in terms of their effect on PPARA pathway activity score. The PPARA pathway shows a high negative activity score (i.e., inactivated) in tumor while at the same time a high positive activity score (i.e., activated) in normal. Thus, an effective drug treatment would be able to increase the PPARA pathway activity score (i.e., stimulating the pathway activity) and thus has a large positive activity score difference (AS_d) between treatment and control. Of the 1,390 instances, 739 instances show a positive score difference for PPARA pathway (i.e., the activity score in the treatment is larger than that in the paired control) (red bars

in **Figure 3.5a**), suggesting that the corresponding treatment could increase the PPARA pathway activity. In comparison, we also evaluated drug treatments in terms of their effect on reversing PPARA pathway expression pattern. Since the majority of gene members in the PPARA pathway are under-expressed and the minority are over-expressed in BC. We would expect an effective drug would reverse the expression pattern by upregulating those under-expressed genes while downregulating those over-expressed genes. Such drugs would show a large negative connectivity score. Out of the 1,390 instances, 739 generated a negative connectivity score for the PPARA pathway (blue bars in **Figure 3.5b**), suggesting that the corresponding treatment could reverse the PPARA pathway expression pattern. Ordering the instances according to increasing difference activity and connectivity scores, respectively and after extracting the top 50 instances from each case, we found that 13 drugs increased the PPARA pathway activity scores and reversed the PPARA pathway expression pattern simultaneously (**Figure 3.5c** and **Table 3.3**).

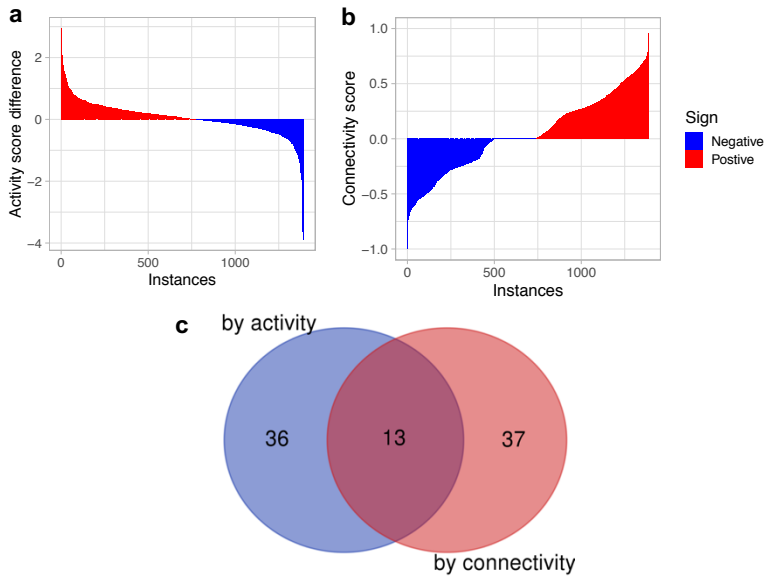


Figure 3.5 Effects of instances on PPARA pathway activity and expression pattern
(**a**) The PPARA pathway activity score difference of 1,390 CMAP instances. The y-axis indicates the difference between the activity score in the treatment and the paired control for each instance. The x-axis indicates the instances ordered by the corresponding activity score difference descendingly. (**b**) The connectivity score of 1,390 CMAP instances. The y-axis indicates the connectivity score. The x-axis indicates the instances ordered by the corresponding connectivity score increasingly. (**c**) The intersection between the top 50 drugs identified by the activity score difference and the activity score.

Table 3.3 The top effective drugs on PPARA pathway

Drug name	Dose (μM)	N [a]	Rank by		Activity score			Connectivity score
			Activity difference (AS_d) [b]	Connectivity score	Treatment (AS_t)	Control (AS_c)	Difference (AS_d)	
anisomycin	15	2	1	15	2.71	-0.26	2.96	-0.66
cephaeline	6	3	2	48	3.01	0.61	2.40	-0.58
pararosaniline	10	1	3	4	2.52	0.30	2.22	-0.73
cicloheximide	14.2	2	6	39	2.05	-0.02	2.08	-0.61
monensin	10.9	1	17	21	0.34	-1.23	1.57	-0.64
wortmannin	1	2	18	44	2.75	1.18	1.56	-0.60
raloxifene	0.1	1	24	47	0.72	-0.66	1.38	-0.58
prednisolone	1	1	25	8	0.65	-0.73	1.38	-0.71
valinomycin	0.1	2	34	29	1.11	-0.09	1.20	-0.62
oligomycin	1	1	37	31	0.51	-0.63	1.15	-0.62
mepacrine	7.8	1	44	14	0.93	-0.10	1.03	-0.66
5186324	2	1	47	20	1.31	0.30	1.01	-0.65
5162773	7	1	48	30	1.31	0.30	1.01	-0.62

[a] The number of experiments;
[b] Activity difference = activity score in treatment (AS_t) - activity score in control (AS_c).

3.3.6 Identification of drugs modulating the FOXM1 and PPARA pathways concurrently

Using the two-pathway signature to query the CMAP database, we examined the individual instances to identify drugs that could reverse the expression pattern of the two pathways at the same time. The top 10 (with the smallest connectivity scores) drugs are listed in **Table 3.4**. To have a further look at effect of these instances on each of the two pathways, we also include the up scores representing the absolute enrichment of the up genes (i.e., the "core" genes which are upregulated in BC and from the FOXM1 pathway) in a given instance and the down scores representing the absolute enrichment of the down genes (i.e., the "core" genes which are downregulated in the PPARA pathway in BC) in a given instance. Both the two types of scores are a value between +1 and -1. A high positive up (down) score indicates that the corresponding drug induced the expression of the up (down) genes. A high negative up (down) score indicates that the corresponding drug repressed the expression of the up (down) genes. In the current study,

we expected the effective drugs to repress the expression of the up genes while inducing the expression of the down genes. Therefore, the most effective drugs were expected to have a high negative up score and a high positive down score. Among the top 10 drugs, MG-262, MG-132, celastrol, ciclopirox, and puromycin, which are also among the top drugs for the FOXM1 pathway, showed a negative correlation with the two-pathway signature, indicating their effect on both the FOXM1 and PPARA pathways. Examination of the up and down scores show that these compounds produce a greater repression of the FOXM1 pathway than their stimulation of the PPARA pathway.

Table 3.4 The top effective drugs on the FOXM1 and PPARA pathways simultaneously

Rank	Drug name	Dose (μM)	N [a]	Up score	Down score	Connectivity score
1	MG-262	0.1	1	-0.55	0.19	-0.91
2	clotrimazole	50	1	-0.32	0.34	-0.82
3	MG-132	21	1	-0.52	0.14	-0.81
4	hycanthone	11	2	-0.44	0.15	-0.73
5	celastrol	3	1	-0.43	0.15	-0.72
6	ciclopirox	15	2	-0.46	0.11	-0.71
7	withaferin A	1	2	-0.39	0.18	-0.70
8	cephaeline	6	3	-0.35	0.19	-0.67
9	pararosaniline	10	1	-0.31	0.23	-0.67
10	puromycin	7	2	-0.34	0.19	-0.66

[a] The number of the MCF7 cell samples treated by the drug at the same dose.

3.4 Discussion

The CMAP analysis has been widely used to identify a novel therapeutic treatment for a disease. The CMAP analysis could be performed by querying a pathway signature against the CMAP database using the gene set enrichment analysis algorithm described by Lamb *et al.* [330]. The generated connectivity score tells the effect of a certain drug treatment in reversing or inducing the expression pattern of the query pathway. This method considers all gene members in a pathway of equal relevance and treats pathways as unstructured lists of genes. The *DART* algorithm evaluates the relevance of the prior information of genes in a given pathway and signature and then estimates pathway activity. In this study, we identified the drug candidates for the FOXM1 and PPARA pathways using both the CMAP query and the DART algorithm. The intersection of drugs identified from each method were considered as candidates with more confidence.

Among the 19 drugs (**Table 3.2**) identified by both the CMAP query and the DART algorithm as repressing the FOXM1 pathway, some have been mentioned in other studies. For example, the ChemBridge compound 5109870 which had the lowest activity score difference (-10.16) and ranked tenth by the connectivity has been found previously to induce HIF-1α-responsive genes, chelate iron and block progression of BC in two distinct mouse models [336]. MG-132 and MG-262 are proteasome inhibitors which have been shown *in vitro* and *in vivo* to have anticancer properties on their own or synergistically with other compounds [337]. In one study, the combination of natural compounds such as gambogic acid with MG-132 or MG-262 showed inhibitory effects on growth of malignant cells and tumors in allograft animal models with no observed systemic toxicity [338]. Celastrol was reported to inhibit breast cancer cell invasion by reducing NF-κB-mediated matrix metalloproteinase-9 expression [339] and also showed anticancer effects on human TNBC potentially by affecting oxidative stress,

apoptosis and the PI3K/Akt pathways [340]. Several studies have suggested that resveratrol may have anti-tumor effects through a variety of mechanisms including but not limited to inhibition and/or activation of histone deacetylases [341] and suppression of the PI3K/Akt signaling pathway [341]. In particular, resveratrol inhibits migration of MDA-MB-435 BC cells via suppression of the PI3K/Akt signaling pathway [341]. In breast cancer, it is reported that the proliferating cell percentage was reduced with resveratrol. In addition, higher doses of resveratrol delayed tumor formation and multiplicity, while the lower dose did not significantly alter these parameters when compared to a control. The work of Chatterjee *et al.* reported that this compound decreased the expression of TGFβ1 and NF-κB and the expression and activity of 5-LOX, as well as cell proliferation, while increasing the faction of cells undergoing apoptosis [342]. Resveratrol also decreased the appearance of single-strand DNA, suggesting that there was less DNA damage [342]. Despite these interesting findings, it should be noted that multiple studies on resveratrol have demonstrated its inconsistent effects on cancer in animals and humans including negative, positive and no effect on cancer outcomes [341,343]. Ciclopirox, which is currently used as a topical antifungal, may also inhibit cell proliferation, inducing cell death, as well as inhibiting angiogenesis and lymph angiogenesis all of which suggest a potential role as an anti-cancer drug. Zhou *et al.* showed that ciclopirox inhibits human breast cancer MDA-MB-231 growth in xenografts [344]. Emetine, which is currently used as an antiemetic and for the treatment of amebiasis, has also demonstrated anticancer properties, by inhibiting both ribosomal and mitochondrial protein synthesis and interfering with the synthesis and activities of DNA and RNA [345].

We found 13 drugs (**Table 3.3**) that could increase the PPARA pathway activity scores by the DART algorithm and reverse the PPARA pathway expression pattern by the CMAP query.

Anisomycin is an antibiotic which is active against protozoa and yeast by inhibiting DNA and protein synthesis [346]. Recently, anisomycin has been found to be active against certain types of cancer, such as ovarian cancer, colon cancer and renal carcinoma [346]. It is also an agonist of p38-mitogen activated-protein kinase and c-Jun N-terminal kinase [346]. Anisomycin has been shown to sensitize glucocorticoid-resistant leukaemia cells to dexamethasone-induced apoptosis through p38-MAPK/JNK [346]. Interestingly, raloxifene is a selective estrogen receptor modulator that produces anti-estrogenic effects in breast and uterine tissues and is used to decrease the risk of BC in post-menopausal women who are at high risk for invasive BC [347,348].

Among the top 10 drugs (**Table 3.4**) that induce the PPARA pathway and repress the FOXM1 pathway concurrently, clotrimazole shows a connectivity score of -0.82, together with similar up score and down score, suggesting its capacity to suppress the FOXM1 pathway while inducing the PPARA pathway. Clotrimazole in a topical antifungal that has been found to preferentially inhibit human BC cell proliferation, viability and glycolysis [349]. Hycanthone and withaferin A were found to reverse the two-pathway signature like clotrimazole. Unlike clotrimazole, these two compounds have a stronger effect on the FOXM1 pathway than on the PPARA pathway. Hycanthone, as one of the thioxanthone derivatives, was found to exhibit antischistosomal, antitumor, and anti-metastatic activities against breast cancer [350]. Withaferin A is a steroidal lactone and negatively regulates breast cancer growth [351]. Pararosaniline, among the top drugs for the PPARA pathway, could also reverse the two-pathway signature. Cephaelin, with a connectivity score of -0.67, demonstrates its effect on both of the two pathways, in accordance with the findings that cephaelin was among the top drugs for either the FOXM1 pathway or the PPARA pathway. MG-262, MG-132, celastrol, ciclopirox, and puromycin, which

are also among the top drugs for the FOXM1 pathway, showed a negative correlation with the two-pathway signature, indicating they could reduce the FOXM1 pathway activity and increase the PPARA pathway activity. Examination of the up and down scores show that these compounds produce a greater repression of the FOXM1 pathway than their stimulation of the PPARA pathway.

Cancer cell lines have been used extensively to screen anti-cancer drug candidates. However, often times the capacity to repress tumor cells *in vitro* is not quantified. In this study, the FOXM1 and PPARA pathways were identified in TCGA breast tumor cells. The FOXM1 pathway has a higher activity score in the four BC subtypes than that in the normal tissue samples, with the highest in TNBC. The PPARA pathway has a higher activity score in the normal breast tissue samples than that in the four BC subtypes. No significant difference was found among the four BC groups for their PPARA pathway activity. With the BC cell lines, we were surprised to see that the distribution of the FOXM1 and PPARA pathway activity scores in breast tumor cell lines is similar to that in TCGA breast tumor patients. The findings suggested that drugs affecting the two pathways in the BC cell line could also potentially affect the two pathways in the BC tumors.

The limitation of this study was that only MCF7 cell line was used to investigate the drug candidates for the two pathways. MCF7 is known to be luminal A breast tumor cell line and therefore cannot fully represent each sub-type of breast cancer, which is a complex and heterogeneous disease. Therefore, the drugs identified for the FOXM1 and PPARA pathways identified in MCF7 cells may not have the same effects in the luminal B, HER2-enriched, and TNBC subtypes. Ideally we would have been able to identify the drugs for the two pathways in different BC cell lines representing different BC subtype cells

3.5 Conclusions

Upregulation of the FOXM1 pathway and downregulation of the PPARA pathway were found in BC cells and in particular TNBC. Therefore, the current study was aimed to identify compounds effective at repressing the FOXM1 pathway activity as well as those inducing the PPARA pathway activity. The former included 5109870, MG-132, MG-262, celastrol, resveratrol, ciclopirox and cephaeline while the latter included anisomycin, cephaeline, pararosaniline cicloheximide and monensin. In addition, the compounds decreasing the FOXM1 pathway activity while increasing the PPARA pathway activity concurrently were identified, including MG-262, MG-132, celastrol, ciclopirox and puromycin.

Chapter 4: Predicting Drug Response of Breast Cancer Using Multiple-Layer Cell Line-Drug Network Model

This chapter is an unpublished paper submitted to BMC Cancer. All analyses, plots, tables, and written text in the thesis I produced myself:

Huang Shujun, Hu Pingzhao, Lakowski Ted*. Predicting Drug Response of Breast Cancer Using Multiple-Layer Cell Line-Drug Network Model. (*denotes equal contribution)*

4.1 Introduction

Predicting the response of a patient to a given drug based on the patient's molecular profiles is one of the key goals of precision medicine in BC. To this end, researchers have been developing computational models to predict anti-cancer drug response of cancer cells based on their molecular profiles (especially gene expression profiles) using cell line-derived (*in vitro*) datasets. Ideally, a drug response prediction model should be first trained using existing patient-derived (*in vivo*) data and then used to predict the response of new patients to a particular drug. However, currently available *in vivo* datasets, such as TCGA [10], do not have enough drug response data to train drug response prediction models whereas *in vitro* datasets, such as the GDSC [235,236] and CCLE [237,238], provide the response data of hundreds of cell lines to many drugs. Therefore, cancer cell line-derived datasets have provided an alternative way for training *in silico* drug response prediction models that can be further used in cancer patients.

A large number of computational models have been developed to predict anti-cancer drug response using cell lines from various cancer types and these models are known as pan-cancer prediction models [248]. Most of these models have been built using gene expression profiles of

cancer cells as input, which have been shown to be the most informative data for drug response prediction in cancer research [236,249–251]. For instance, Dong *et al*. [253] developed a support vector machine model to predict the response of various cancer cell lines to a particular drug. The model was trained on the CCLE dataset and used gene expression data as input features. Other types of molecular profiles, such as gene mutation, CNV, have also been incorporated to develop prediction models to improve the predictive power[236]. For example, Sharifi-Noghab *et al.* [257] proposed a deep learning-based model to take gene somatic mutation, CNV, and gene expression data of a particular cell line as input, and predict the response of the cell line to a given drug as the output. In addition, the physical, chemical, and pharmacological properties of drugs such as chemical structure, aqueous solubility, and potency (IC_{50}) also play important roles in drug response prediction. Thus, computational models by combining genomic profiles with information about the drug's chemical structure would improve drug response prediction *in vitro* and *in vivo* [258,259]. Menden *et al.* [258] developed a neural network model which took mutation, CNV and microsatellite instability data of cell lines together with chemical properties of drugs as inputs to predict drug response in the GDSC dataset. Zhang *et al.* [255] proposed a dual-layer network, which integrated a cell line similarity network based on gene expression profiles and drug similarity network based on drug chemical structures. Very recently, Wei *et al*. [352] proposed a new dual-layer network model, which captures different contributions of all available cell line-drug responses through cell line similarities and drug similarities. The model was applied to CCLE and GDSC independently and demonstrated better performance than some existing studies including the Zhang *et al* study [255]. We note that in Wei *et al.*'s study: 1) the similarity between cell lines was only based on their gene expression profiles while the other omics data types were ignored; 2) the relationships among genes were also overlooked; 3) the similarity between cell

lines in Wei *et al*.'s study was based their chemical structures while the other data types, such as the drug targets, were not taken into consideration.

The models mentioned above, trained on a per-drug and pan-cancer basis, do not take the heterogeneity of cancer types into consideration. Thus, other models have focused on making predictions for drugs for a specific cancer type and are referred to as cancer-specific response prediction models [248]. Some BC-specific models have been developed. The most well-known work is the NCI-DREAM Drug Sensitivity Prediction Challenge [261], which provided drug sensitivity data screened in BC cell lines and along with molecular profiles of BC cell lines to the participants and aimed to predict drug sensitivity in BC cell lines by integrating multiple-omics data. A total of 44 drug response prediction models were submitted to the NCI-DREAM Drug Sensitivity Prediction Challenge and the Bayesian multitask multiple kernel learning method demonstrated the best performance [261]. We note that these approaches take into account the multivariate relationships among genomic features but do not take into account that similar cell lines may have similar therapeutic responses to the same drugs. In addition, these studies did not take into consideration the drugs properties and ignore the fact that drugs with similar properties (such as structure) may have similar therapeutic effect on cell lines.

In this study, motivated by the dual-layer network model in Wei et al.'s study [352], we constructed multiple-layer cell line-drug response network (ML-CDN) models with the focus on BC. In the ML-CDN modeling method, cell line similarity networks (CSNs) were first constructed using either one or all of three types (i.e., layers) of molecular profiles: pathway activity profiles, CNV and mutation profiles. In parallel, drug similarity networks (DSNs) were constructed using either one or all of three types (i.e., layers) of profiles: structures, targets and pan-cancer sensitivity profiles. Next, a CSN and a DSN were connected by linking the cell lines in the first network to

their corresponding (previously tested) drugs in the second network. In the end, ML-CDN was used to predict anti-cancer drug response of BC patients by estimating the similarities between these patients with BC cell lines.

4.2 Materials and Methods

4.2.1 Data resources and preprocessing

The overall approach is depicted in **Figure 4.1**. In this work, we used the GDSC release 7.0 dataset (ftp://ftp.sanger.ac.uk/pub4/cancerrxgene/releases/release-7.0) as a benchmark dataset, which consists of 1,065 cancer cell lines and 266 tested drugs. We downloaded the gene expression, CNV and mutation profiles for 49 BC cell lines. For gene expression data, we used the Robust Multichip Average (RMA) [333] to normalize baseline expression profiles (i.e. coming from untreated samples) for all the BC cell lines. For gene mutation data, the list of genomic somatic variants found in BC cell lines by whole exome sequencing were downloaded and further described as binary features (1 = mutation and 0 = wild type). A gene mutation is annotated if a sequence variation (changes in the protein sequence, e.g. non-synonymous single nucleotide polymorphism) is detected while a gene for which a mutation is not detected in a given cell line is annotated as wild type. CNV data for all genes across all BC cell line samples derived from the Predict Integral Copy Numbers In Cancer analysis [353] of Affymetrix SNP6.0 segmentation data were downloaded and further treated as binary features (1 = amplification or deletion and 0 = wild type). Genes are considered to be amplified when the entire coding sequence is in one contiguous segment as defined by the Predict Integral Copy Numbers In Cancer analysis [353], and have a total copy number of eight or more [235,258,354]. Deletions must occur within a single contiguous segment with copy number zero. The wild type corresponds to a copy number range between 0

and 8, excluding both 0 and 8. Genes annotated as wild-type across all BC cell lines in terms of either CNV and mutation were removed. We also downloaded dose-response curves of the 1,065 cancer cell lines (including the 49 BC cell lines) to 266 drugs, each summarized by its IC_{50} value.

Additionally, cell line gene expression and drug response information from CCLE was used for model validation in this study. From CCLE (https://portals.broadinstitute.org/ccle), we downloaded the gene expression profiles for 40 BC cell lines along with the drug potency measurements evaluated with IC_{50} values for 12 out of the 266 GDSC small molecules. For the RNA-Seq gene expression data, we used expression level computed using the RSEM method [92] provided by CCLE, multiplied by 10^6 to obtain Transcripts Per Million (TPM) [92] and log2-transformed.

Drug chemical structure data were also curated. In order to reduce 3D drug chemical structure data into a 1D usable string variable, we used the simplified molecular-input line-entry system (SMILES). We extracted the canonical SMILES strings for 220 out of the 266 GDSC small molecules from PubChem [355], a database of more than 60 million unique structures. We then used the *parse.smiles* function of the R package rcdk [356] to parse the annotated SMILES strings for existing drugs. Extended connectivity fingerprints (hash-based fingerprints, default length 1,024) across all drugs were subsequently calculated using the *get.fingerprints* function of the R package rcdk. Regarding the drug targets, we extract the interactions between the 220 drugs and 272 target proteins from GDSC.

For the TCGA BC cohort, the patient-specific RNA-Seq gene-level expression data computed by the RSEM algorithm [92] were downloaded from Firehose Broad GDAC (https://gdac.broadinstitute.org), multiplied by 10^6 to obtain TPM [92] and log2-transformed. We

used clinical annotations of the drug response for some patients obtained from a previous study [250]. We also used TCGA BC patients without the drug response in their records.

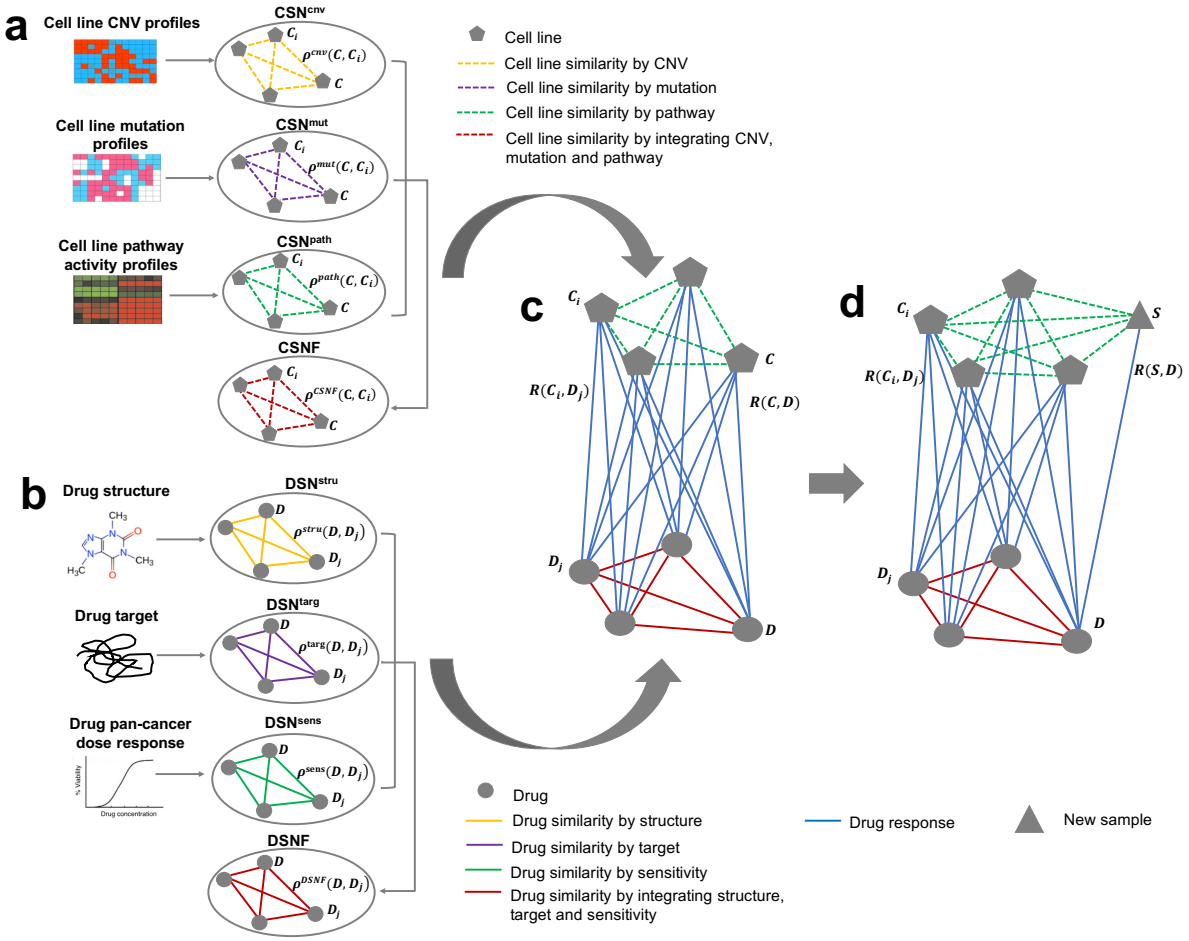


Figure 4.1 Overview of the study design
This study included four major steps: (**a**) four CSNs (CSN^{cnv}, CSN^{mut}, CSN^{path}, CSNF) were constructed separately; (**b**) four DSNs (DSN^{stru}, DSN^{targ}, DSN^{sens}, DSNF) were constructed separately; (**c**) 16 ML-CDNs were constructed by connecting one of the four CDNs to one of the four DSNs, and here only ML-CDN2 was presented which was based CDN^{path} and DSNF; (**d**) ML-CDN2 was employed to predict the response of new breast cancer cell line samples from CCLE or new breast tumor tissue samples from TCGA to the existing drugs.

4.2.2 Pathway activity score calculation

Gene sets of 1,329 canonical pathways, which were curated from various pathway databases (e.g., KEGG, PID, REACTOME, and BIOCARTA), were extracted from the MSigDB website [104]. As described by Wang *et al.* [357], we scored the pathway activities using the gene

expression profiles of BC cell lines or patients from GDSC, CCLE or TCGA datasets. Note that we first standardized gene expressions within each cohort, and then performed pairwise homogenization procedure before scoring pathway activities as described in other studies [249,358] to make the expression measures in different datasets comparable. Briefly, for the three gene expression datasets (GDSC, CCLE and TCGA datasets), we kept only genes presenting in all of the three datasets and applied *ComBat* function from the sva R package[334] to adjust the potential batch effect in the data sets.

4.2.3 Cell line similarity network construction

Multiple molecular profiles of cancer cell lines have been reported to affect drug response prediction [248]. Thus we estimated the cell line similarities in terms of three data types (CNV, mutation, and pathway activity profiles) in this study. We first constructed three CSNs using the three data types separately (**Figure 4.1a**): 1) a CSN using CNV data (CSN^{cnv}), where associations between every two cell lines C and C_i are measured by the Tanimoto correlation ($\rho^{cnv}(C, C_i)$) between their CNV profiles; 2) a CSN using mutation data (CSN^{mut}), which connects every cell line pair C and C_i using the Tanimoto correlation ($\rho^{mut}(C, C_i)$) calculated based on their gene mutation profiles; and 3) a CSN using pathway activity data (CSN^{path}), which connects every two cell lines C and C_i based on their Pearson correlation ($\rho^{path}(C, C_i)$) of pathway activity profiles. The CSNs were generated using GDSC data and all of them are complete graphs of 49 BC cell lines, where interactions between each pair of the cell lines were measured by the correlation coefficients of their respective pathway activity, CNV, and mutation profiles. The three CSNs are a single-layer CSN since each of them was constructed using a single type of molecular data. Next, we used the Similarity Network Fusion (SNF) algorithm [359] to integrate the three CSNs, which included two major steps: 1) an affinity matrix was calculated from each CSN using the

affinityMatrix function as described in the R SNFtool package [359] with default parameters; 2) using the *SNF* function of the SNFtool package, the three affinity matrices were fused into a cell line similarity network fusion (CSNF), which connects every cell line pair C and C_i by the SNF algorithm-derived correlation ($\rho^{CSNF}(C, C_i)$). The CSNF is a three-layer CSN since it was constructed using pathway activity, CNV, and mutation profiles together. Thus, a total of four CSNs were obtained, with CSN^{path}, CSN^{cnv}, and CSN^{mut} using a single data type while CSNF integrating the three data types to measure cell line pair similarities.

4.2.4 Drug similarity network construction

We also estimated drug similarities using three data types (drug structure, target, and pan-cancer response information). Three DSNs, similar to the three CSNs, were first constructed separately (**Figure 4.1b**): 1) a DSN using their structure data (DSN^{stru}), which connects every two drugs D and D_j based on the Tanimoto correlation ($\rho^{stru}(D, D_j)$) of their molecular fingerprint properties; 2) a DSN using their target data (DSN^{targ}), where associations between every two drugs D and D_j are measured by the Jaccard correlation ($\rho^{targ}(D, D_j)$) between their target information; and 3) a drug similarity network using their sensitivity profiles (DSN^{sens}), which connects every drug pair D and D_j using the Pearson correlation ($\rho^{sens}(D, D_j)$) calculated based on their sensitivity profiles with respect to the IC_{50} values across the 1,065 cell lines from all cancer types. The three DSNs are complete graphs of all available drugs, connected with their pairwise correlation between their corresponding structural features, target information, and pan-cancer sensitivity profiles. The three DSNs are a single-layer DSN since each of them was constructed using a single data type. Next, using the SNFtool package[359], we integrated the three DSNs into a drug similarity network fusion (DSNF), which connects every drug pair D and D_j was by the SNF algorithm-derived correlation ($\rho^{DSNF}(D, D_j)$). The DSNF is a three-layer DSN since it was

constructed using drug structure, target, and pan-cancer response information together. Therefore, a total of four DSNs were obtained, with DSN^{stru}, DSN^{targ}, and DSN^{sens} using a single data type while with DSNF integrating the three data types to measure drug pair similarities.

4.2.5 Multiple-layer cell line-drug response network construction

We next integrated a CSN and a DSN together to develop a multiple-layer cell line-drug response network (ML-CDNs), which could be used to make predictions of drug response for BC cell lines (**Figure 4.1c**). In this ML-CDN model, a layer denotes a data type from either the cell lines or the drugs. A ML-CDN connects a CSN and a DSN by linking the cell lines in the first network to their corresponding (previously tested) drugs in the second network. A ML-CDN is a bipartite graph of all cell lines and drugs, labeled with the corresponding response values (IC_{50} values). Note that a ML-CDN is not a complete bipartite graph due to some missing values in the GDSC dataset. With the four CSNs and four DSNs constructed in the above two sections, we obtained a total of 16 ML-CDNs: 1) nine dual-layer CDNs were built using each of the three single-layer CSNs (CSN^{path}, CSN^{cnv} and CSN^{mut}) and each of the three single-layer DSNs (DSN^{stru}, DSN^{targ}, DSN^{sens}); 2) three four-layer CDNs were built using the three-layer CSNF combined with each of the three single-layer DSNs; 3) similarly, another three four-layer CDNs were built using the three-layer DSNF combined with each of the three single-layer CSNs; 4) finally, a six-layer CDN was built using the two integrated similarity networks, CSNF and DSNF. In this study, we referred to the ML-CDN based on CSNF and DSNF as ML-CDN1 and the one based on CSN^{path} and DSNF as ML-CDN2.

For a given cell line-drug pair (C, D), we are able to make a prediction of the response of the cell line C to the drug D using **Equation 4.1**, where Ω is the set of all possible cell line-drug pairs, $\Omega \backslash \{(C, D)\}$ is the set of all other pairs (C_i, D_j) except (C, D), $R(C_i, D_j)$ denotes the observed

response of the pair (C_i, D_j). $\hat{R}(C, D)$ is the predicted response value for the pair (C, D). The product of $w(C, C_i)$ and $w(D, D_j)$ reflects the contribution of $R(C_i, D_j)$ to $\hat{R}(C, D)$.

$$\hat{R}(C,D) = \frac{\sum_{(C_i,D_j)\in\Omega\setminus\{(C,D)\}} w(C,C_i)w(D,D_j)R(C_i,D_j)}{\sum_{(C_i,D_j)\in\Omega\setminus\{(C,D)\}} w(C,C_i)w(D,D_j)} \qquad \textbf{Equation 4.1}$$

where $w(C, C_i)$ is the weight function between cell lines C and C_i and $w(D, D_j)$ is the weight function between drugs D and D_j, which can be calculated as follows:

$$w(C,C_i) = e^{-\frac{[1-\rho(C,C_i)]^2}{2\sigma^2}} \qquad \textbf{Equation 4.2}$$

$$w(D,D_j) = e^{-\frac{[1-\rho(D,D_j)]^2}{2\tau^2}} \qquad \textbf{Equation 4.3}$$

The weight $w(C, C_i)$ between cell lines C and C_i increases with respect to the cell line similarity correlation $\rho(C, C_i)$, and σ measures the decay rate when $\rho(C, C_i)$ decreases. Note that $w(C, C_i)$ is different when $\rho(C, C_i)$ from different CSNs (i.e., $\rho^{cnv}(C, C_i)$ from CSN^{cnv}, $\rho^{mut}(C, C_i)$ from CSN^{mut}, $\rho^{path}(C, C_i)$ from CSN^{path}, and $\rho^{CSNF}(C, C_i)$ from CSNF) is used. We referred to them as $w^{path}(C, C_i)$, $w^{cnv}(C, C_i)$, $w^{mut}(C, C_i)$, and $w^{CSFN}(C, C_i)$. The weight $w(D, D_j)$ between drugs D and D_j increases with respect to the drug similarity correlation $\rho(\text{D}, \text{D}_\text{j})$ and τ measures the decay rate when $\rho(D, D_j)$ decreases. Note that $w(D, D_j)$ is different when $\rho(D, D_j)$ from different DSNs (i.e., $\rho^{\text{stru}}(D, D_j)$ from DSN^{stru}, $\rho^{\text{targ}}(D, D_j)$ from DSN^{targ}, $\rho^{\text{sens}}(D, D_j)$ from DSN^{sens}, $\rho^{\text{DSNF}}(D, D_j)$ from DSNF) is used. We referred to them as $w^{\text{stru}}(D, D_i)$, $w^{\text{targ}}(D, D_i)$, $w^{\text{sens}}(D, D_i)$, and $w^{\text{DSFN}}(D, D_i)$.

The ML-CDN models contain two decay parameters (σ, τ), to be used in the weight function of cell lines and drugs, respectively. The decay parameter pair (σ, τ) was optimized by minimizing the sum of the squared errors for all possible cell line-drug pairs using **Equation 4.1** as the response prediction model. The overall error function (also known as a cost function) is defined in **Equation 4.4**, where $R(C, D)$ is the observed response value of cell line *C* to drug *D*,

and $\hat{R}(C, D)$ is the predicted value of the cell line *C* to the drug *D*. σ and τ are both ranged from 0 to 1 with an increment of 0.01, and the pair (σ, τ) takes all possible combinations.

$$J(\sigma, \tau) = \sum_{(C,D)\in\Omega} (R(C, D) - \hat{R}(C, D))^2 \qquad \textbf{Equation 4.4}$$

To compare the performance of the 16 ML-CDN models, we split all cell line-drug pairs into three folds. Two folds were used as the training set for optimizing the decay parameter pairs (σ, τ) while the remaining fold used as the test set for estimating the prediction performance of the models. The performance was evaluated using Pearson correlation coefficient and root mean squared error (RMSE) between the predicted and the observed drug responses for all drugs. RMSE is the square root of the mean squared error. A higher Pearson correlation coefficient and lower RMSE suggest better prediction performance by a method.

4.2.6 Multiple-layer cell line-drug response network to predict drug response for a new tumor

Although the ML-CDN models can be used to predict either drug response for a new either cell line or tumor tissue sample based on the CSN or cell line response to a new drug based on the DSN, we focused on the former in this study. We therefore expected that the BC cell line-derived ML-CDNs could have good predictive performance in BC patient-derived data. Among the 16 ML-CDN models, ML-CDN2, which was constructed by connecting CSN^{path} and DSNF and demonstrated good prediction performance in terms of Pearson correlation and RMSE (See the Results Section), was chosen for predicting the response of a new BC patient- or cell line-derived sample *S* to a known drug *D* (**Figure 4.1d**). To make a prediction for $R(S, D)$, we first estimated the similarity between the new sample *S* and any existing cell line C_i by calculating the Pearson correlation ($\rho^{path}(S, C_i)$) in terms of their pathway activity profiles and further obtained the weight ($w^{path}(S, C_i)$) between *S* and C_i using **Equation 4.2** with σ optimized from ML-CDN2. In parallel, the similarity between *D* and any existing drugs D_j was measured using $\rho^{DSNF}(D, D_j)$

from DSNF and further weighted using **Equation 4.3** with τ optimized from ML-CDN2 to obtain $w^{DSFN}(D, D_i)$. $R(C_i, D_j)$ is the observed response value of existing cell line C_i to existing drug D_j. Thus, $R(S, D)$ can be predicted by can taking advantage of response data from all existing cell lines C_i based on their weighted similarities with the new sample *S* and all existing drugs D_j based their weighted similarities with *D* as **Equation 4.5**. The predictions of the response to the existing drugs in the benchmark dataset (i.e., GDSC) were made for new BC cell lines from CCLE or new breast tumor tissue samples from TCGA.

$$\hat{R}(S, D) = \frac{\sum_{(C_i, D_j) \in \Omega} w^{path}(S, C_i)\, w^{DSFN}(D, D_j)\, R(C_i, D_j)}{\sum_{(C_i, D_j) \in \Omega} w^{path}(S, C_i)\, w^{DSFN}(D, D_j)} \qquad \textbf{Equation 4.5}$$

4.2.7 Drug response prediction for CCLE cell lines

To validate the performance of ML-CDN2, we employed the ML-CDN2 model, which was trained in the GDSC dataset, to predict the drug response of the same drugs used for BC cell lines from the CCLE dataset using **Equation 4.5**. In this context, we were making predictions of response for new cell lines to existing drugs. First, we measured the similarities between new cell lines from CCLE and the existing cell lines from GDSC using the Pearson correlations of their pathway activity profiles and then obtained the weights between the CCLE and GDSC BC cell lines using **Equation 4.2** with σ optimized from ML-CDN2. The prediction of the drug response could then be made using **Equation 4.5**.

4.2.8 Drug response prediction for TCGA patients

To further study ML-CDN2's performance *in vivo*, we employed the model to predict the drug response of five drugs for patients in the TCGA BC dataset for which drug response was recorded. The five drugs were tested in the GDSC study, including paclitaxel, fluorouracil, tamoxifen, doxorubicin, and docetaxel. For each of the five drugs, the patients were assigned to two groups based on the recorded drug response: Responder (patients showing a "complete

response") and Non-responder (patients showing a "partial response", "progressive disease", or "stable disease"). For these patients, we first calculated their pathway activity scores based on their whole-genome gene expression profiles and then measured the similarity between these BC tumors and the GDSC BC cell lines using the Pearson correlations of their pathway activity profiles. The Pearson correlations were further weighted using **Equation 4.2** with σ optimized from ML-CDN2. In the end, the responses of these TCGA BC patients to the five drugs were made with **Equation 4.5**. Since the IC_{50} values in the GDSC study were measured using cell viability, we expected to see the patients in the Responder group would have a lower predicted IC_{50} value than patients in the Non-responder group.

Using **Equation 4.5**, we also predicted the response of all TCGA BC patients to lapatinib and tamoxifen, which were included in the GDSC study. The former is targeting HER2-overexpressing breast cancers while the latter is targeting ER-positive breast cancers. For lapatinib, we separated the BC patients based on their HER2 overexpression level measured by IHC into four groups: 0, 1+, 2+, 3+, indicating the increasing expression level of HER2. We then compared the predicted IC_{50} values for lapatinib among the four groups. We expected to see that groups with higher HER2 expression level would demonstrate lower predicted IC_{50} values. For tamoxifen, BC patients were divided into two groups (Negative and Positive) based on the IHC status of ER. Then the predicted IC_{50} values for tamoxifen were compared between the two groups. We expected that the predicted IC_{50} values for patients in the ER Positive group would be smaller than that for patients in the ER Negative group.

In addition, we used **Equation 4.5** to predict the drug response of the EGFR and PI3K pathway inhibitors for TCGA BC patients for which there was no drug response recorded. Since these drugs target either the EGFR pathway or the PI3K pathway, we expected the expression level

of the EGFR pathway genes to be strongly correlated with the predicted EGFR inhibitor response while the expression level of the PI3K pathway genes to be strongly correlated with the predicted PI3K inhibitor response. We obtained the gene lists for the EGFR and PI3K pathways from MSigDB[104]. To study the correlation between genes in a pathway and an inhibitor of the pathway, we employed multiple linear regression between the predicted IC_{50} value (response variable) of the inhibitor and the expression levels of the pathway genes (predictors). We obtained the *p*-value for each gene and corrected them for multiple comparison, using Bonferroni correction ($\alpha = 0.05$).

4.3. Results

4.3.1 Data curated

Data used in this study were collected from different databases and are summarized in **Table 4.1**. Gene expression, CNV, and mutation profiles were extracted from the GDSC dataset for 49 common BC cell lines, as well as their response to 220 drugs. Expression profiles of 28 BC cell lines, as well as their response to 13 out of the 220 GDSC drugs were obtained from the CCLE study. Different drugs have different baseline values and ranges of response and in particular IC_{50} values can vary widely based on experimental conditions. Therefore, for each of the two datasets (GDSC and CCLE), we subtracted the mean IC_{50} values of all drugs from each IC_{50} value and then divided each IC_{50} value (mean is already subtracted) by the standard deviation of IC_{50} values of all drugs. This normalization process aimed at making different drugs have the same baseline value and range across all BC cell lines for the GDSC and CCLE studies. Gene expression profiles were collected for 1,100 TCGA BC tissue samples for which responses to five drugs were recorded for 110 patients. Chemical structure information in the form of connectivity fingerprints (hash-based

fingerprints, default length 1,024) and interactions with 272 target proteins for the 220 drugs were extracted from the GDSC dataset using the R package rcdk. In addition, 1,329 canonical gene sets from the MSigDB database were extracted to calculate the pathway activity.

Table 4.1 Data collected from multiple platforms

Dataset	Data type	Platform	Samples
GDSC	Gene expression	Affymetrix Human Genome U219 array	49 cell lines × 14,770 genes
	CNV	Affymetrix SNP6 array	49 cell lines × 3,037 genes
	Mutation	Whole exome sequencing	49 cell lines × 8,849 genes
	Drug response	IC_{50}	49 cell lines × 220 drugs
CCLE	Gene expression	Illumina Hiseq 2000	28 cell lines × 14,770 genes
	Drug response	IC_{50}	28 cell lines × 13 drugs
TCGA	Gene expression	Illumina Hiseq 2000	1,100 tumors × 14,770 genes
	Drug response	RECIST response categories [a]	110 tumors × 5 drugs
Drugs	Chemical structure	rcdk [b]	220 drugs × 1,024 fingerprints
	Target	Curated	220 drugs × 272 targets
MSigDB	Canonical pathways	Curated	1,329 pathway gene sets

[a] Response Evaluation Criteria in Solid Tumors (RECIST), a standard way to categorize treatment response of a cancer patient, including complete response, a partial response, progressive disease, and stable disease.
[b] An R package which can take the SMILES string of a drug as input and output the fingerprints, 1D- and 2D-structres of the drug.

4.3.2 The three cell line data types are predictive of drug response

To measure the similarity of each cell line pair, the correlation $\rho(C,\ C_i)$ between every two cell lines C and C_i were calculated in terms of three data types (the CNV, mutation, and pathway activity profiles) to obtain $\rho^{cnv}(C, C_i)$, $\rho^{mut}(C, C_i)$, and $\rho^{path}(C, C_i)$. The mean correlation of all BC cell line pairs is 88.12% for CNV, 88.10% for mutation and 96.83% for pathway activity. We also calculated the correlation of drug response profiles measured by IC_{50} for each cell line pair. To observe the association between the drug potency and the three types of similarity for the cell line pairs, we divided the similarity level between two cell lines into "low" (minimum <= $\rho(C,\ C_i)$

< 1st quantile), "inter-low" (1st quantile <= $\rho(C,\ C_i)$ < median), "inter-high" (median <= $\rho(C,\ C_i)$ < 3rd quantile) and "high" (3rd quantile <= cc < maximum) for each data type. **Figure 4.2** shows that if two cell lines show similar patterns in terms of the CNV (**Figure 4.2a**), mutation (**Figure 4.2b**), and pathway activity (**Figure 4.2c**) profiles, their responses to certain drugs will be similar, both sensitive or both resistant. These results suggest that cell lines with similar pathway activity, CNV, and mutation profiles exhibit similar drug responses. Therefore, similarity measured by the CNV, mutation, and pathway activity profiles can predict drug response in BC cell lines.

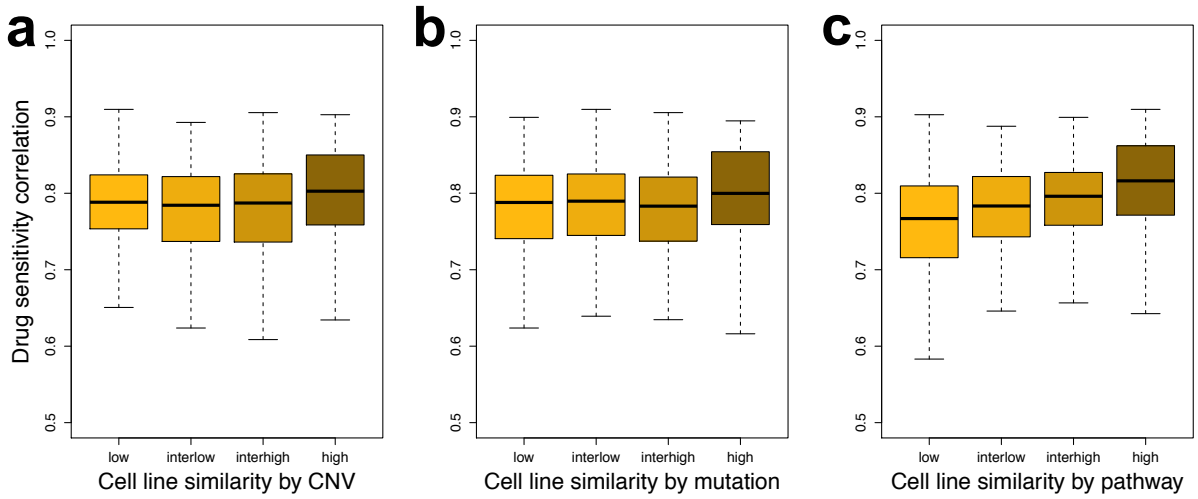


Figure 4.2 Similar cell lines have similar responses
Box plots show that cell lines with similar CNV (**a**), mutation (**b**), and pathway activity (**c**) profiles respond similarly to the same drugs. The x-axis indicates the correlation level between all possible BC cell line pairs based on CNV (**a**), mutation (**b**), and pathway activity (**c**) profiles. The y-axis shows the correlations of their drug response vectors as measured by IC_{50}.

4.3.3 The three drug data types are predictive of drug response

Having noted that similar cell lines exhibit similar drug responses, we next examined whether drugs with similar chemical structures and properties have similar effects on cell lines. We calculated the correlation of IC_{50} values derived from viability in BC cell lines for each drug pair. We extracted the fingerprints to describe the chemical structure and properties and calculated the correlation $\rho^{stru}(D, D_j)$ between a drug pair D and D_j. We divided all drug pairs into four

groups according to their pairwise structural similarities: "low" (minimum <= $\rho^{stru}(D, D_j)$ < 1st quantile), "inter-low" (1st quantile <= $\rho^{stru}(D, D_j)$ < median), "inter-high" (median <= $\rho^{stru}(D, D_j)$ < 3rd quantile) and "high" (3rd quantile <= $\rho^{stru}(D, D_j)$ < maximum). Drug pairs with similar structures have significantly higher drug potency correlations as measured by IC_{50} (**Figure 4.3a**). This suggests that drugs with similar chemical structures exhibit similar inhibitory effects among the BC cell lines tested, which means common patterns can be learned from the fingerprints when predicting drug potency against BC cell lines. We further curated the protein targets of drugs and calculated the coefficient $\rho^{targ}(D, D_j)$ between two drugs D and D_j using these target features. All drug pairs were divided into two groups according to their pairwise target similarities: "low" ($\rho^{targ}(D, D_j)$ <= 0) and "high" ($\rho^{targ}(D, D_j)$ > 0). IC_{50} correlations were significantly higher for drugs with more similar target profiles (**Figure 4.3b**). We also obtained summarized dose-response curves for drugs across 1,065 cell lines from all cancer types from the GDSC dataset. The correlation $\rho^{sens}(D, D_j)$ of pan-cancer IC_{50} profiles was computed for every drug pair D and D_j. All possible drugs pairs were divided into four groups according to their pairwise correlation of pan-cancer IC_{50} profiles: "low" (minimum <= $\rho^{sens}(D, D_j)$ < 1st quantile), "inter-low" (1st quantile <= $\rho^{sens}(D, D_j)$ < median), "inter-high" (median <= $\rho^{sens}(D, D_j)$ < 3rd quantile) and "high" (3rd quantile <= $\rho^{sens}(D, D_j)$ < maximum). Drug pairs with more similar pan-cancer IC_{50} profiles exhibit similar IC_{50} values across the BC cell lines tested (**Figure 4.3c**). These results suggest that integrating the structure, target, and pan-cancer sensitivity data types will improve the drug response prediction.

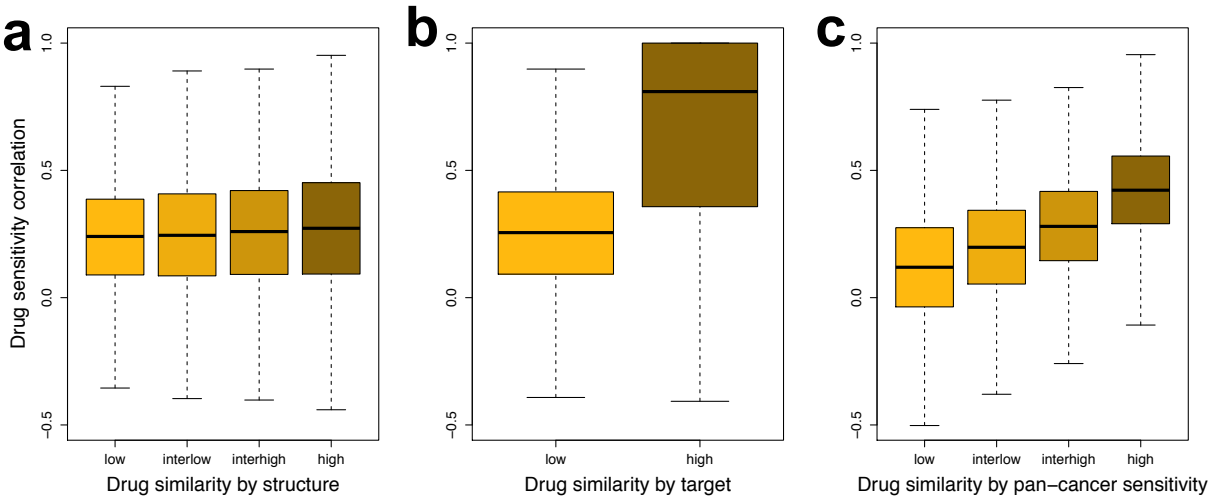


Figure 4.3 Similar drugs have similar responses
Box plots show that drugs with similar structure (**a**), targets (**b**), and pan-cancer cell line sensitivity (**c**) profiles have similar effects on BC cell lines. The X-axis shows the correlation between all possible drug pairs based on structure (**a**), targets (**b**), and pan-cancer cell line sensitivity (**c**) profiles. The Y-axis shows the correlations among their sensitivity vectors tested in BC cell lines as measured by IC_{50}.

4.3.4 Comparison of ML-CDN models

The three types of cell line information enumerated in **Figure 4.2** and the three types of drug information enumerated in **Figure 4.3** are predictive of drug response. So we used each of the four CSNs (CSN^{cnv}, CSN^{mut}, CSN^{path}, and CSNF) and each of the four DSNs (DSN^{stru}, DSN^{targ}, DSN^{sens}, and DSNF) to build a total of 16 ML-CDNs and compared their prediction performance. These models were trained on 2/3 of all cell line-drug pairs and evaluated using the Pearson correlation and RMSE between the observed and predicted drug response values on the remaining pairs. The performance of the 16 ML-CDNs is listed in **Table 4.2** and differs slightly from each other. The optimal model is ML-CDN1 which was constructed based on CSNF and DSNF. Among the three second-best models, we focused on ML-CDN2 which was constructed from CSN^{path} and DSNF. This is because pathway activity profiles derived from transcriptome are the most widely available data for tumor tissue or cell line samples in public databases and DSNF integrated three types of information from drugs.

We optimized the decay parameters for ML-CDN1 and ML-CDN2 models using all cell line-drug pairs. The optimized parameter pair is (σ, τ) = (0.07, 0.06) for ML-CDN1 and (σ, τ) = (0.01, 0.06) for ML-CDN2. The IC_{50} values were predicted for all pairs using the ML-CDN1 and ML-CDN1 models to calculate the Pearson correlation and RMSE. The scatter plots for ML-CDN1 (**Figure 4.4a)** and ML-CDN2 (**Figure 4.4b**) indicate that the good correlations of observed versus predicted responses did not arise from a small number of outliers. We decided to use ML-CDN2 to predict drug response for new cell line-derived or tumor tissue samples because the two models do not differ much in their performance and ML-CDN2 does not use CNV and mutation data of cell lines, which are used in the ML-CDN1 model.

Table 4.2 Performance of the 16 ML-CDN models

CSN	**DSN**	**RMSE**	**R**
CSN^{cnv}	DSN^{stru}	0.513	0.864
CSN^{cnv}	DSN^{targ}	0.562	0.834
CSN^{cnv}	DSN^{sens}	0.504	0.869
CSN^{cnv}	DSNF	0.504	0.869
CSN^{mut}	DSN^{stru}	0.513	0.864
CSN^{mut}	DSN^{targ}	0.562	0.834
CSN^{mut}	DSN^{sens}	0.505	0.869
CSN^{mut}	DSNF	0.504	0.869
CSN^{path}	DSN^{stru}	0.510	0.866
CSN^{path}	DSN^{targ}	0.563	0.833
CSN^{path}	**DSN^{sens}**	**0.496**	**0.873**
CSN^{path}	**DSNF**	**0.497**	**0.873**
CSNF	DSN^{stru}	0.511	0.865
CSNF	DSN^{targ}	0.561	0.834
CSNF	**DSN^{sens}**	**0.496**	**0.873**
CSNF	**DSNF**	**0.493**	**0.875**

R: Pearson correlation coefficient; RMSE: root mean squared error.

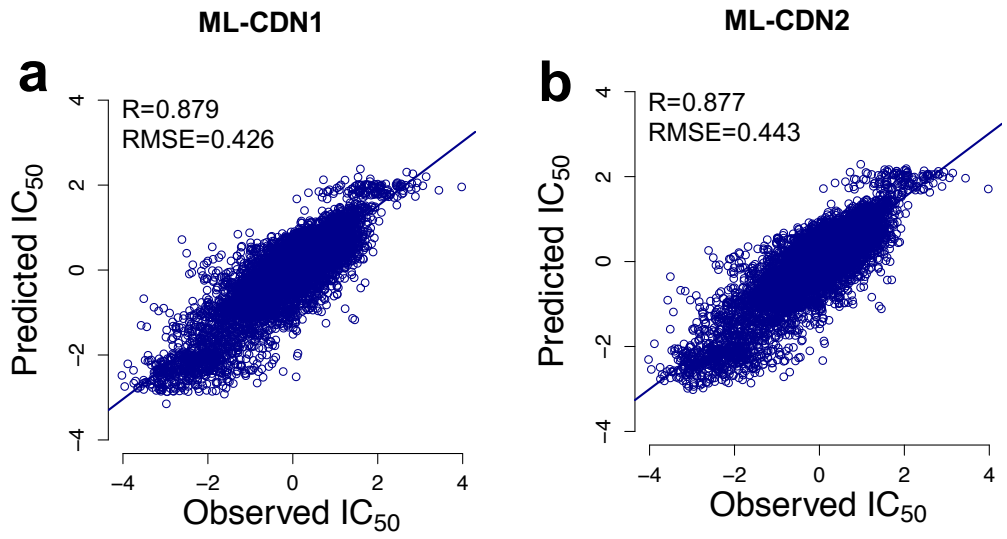


Figure 4.4 Comparison of ML-CDN1 and ML-CDN2
(**a**) Scatter plot of observed and predicted drug responses (IC_{50} values) for all drugs in GDSC using the ML-CDN1 model. (**b**) Scatter plot of observed and predicted drug responses (IC_{50} values) for all drugs in GDSC using the ML-CDN2 model. R: Pearson correlation coefficient; RMSE: root mean squared error.

4.3.5 Comparing ML-CDN2 with other methods

Here we also compared the performance of ML-CDN2 with a dual-layer network proposed by Zhang *et al.* [255]. This method built a weighted model to predict the anticancer drug response using both CSN data and DSN data. The ML-CDN2 model is similar to their method in that both use the CSN and DSN data to predict the drug response of a given cell line. Yet, they constructed the dual-layer integrated cell line-drug network model by combining the predictions from the individual layers. We used the same 220 drugs and 49 BC cell lines from the GDSC study for evaluation in order to make fair comparisons. Following the method of Zhang *et al.*, the CSN was generated using the cell line pairwise Pearson correlations of their gene expression profiles. We extracted the 1-D and 2-D structural features of each drug using PaDEL [360] and calculated the Pearson correlation between each pair of the drugs using these structural features to build the DSN. The performance of the dual-layer model which combined the gene expression-based CSN and PaDEL-based DSN is shown in **Figure 4.5**. The proposed ML-CDN2 model performed better than

Zhang *et al.*'s method. The ML-CDN2 model obtained a correlation between the predicted response and the observed response of 0.873, while Zhang *et al.*'s method had the value of 0.670 (**Figure 4.5**).

We also compared our method with the method of Wei *et al.* [352], which predicted anticancer drug response by capturing different contributions of existing cell line-drug responses through cell line similarities and drug similarities. ML-CDN2 shares something in common with their method in that both methods used the same strategy to integrate the CSN and DSN. Yet, Wei *et al.*'s method measured the cell line similarity by gene expression profiles and the drug similarity by the fingerprint-based chemical structures. In addition, multiple type of data was used for neither cell lines nor drugs in the study of Wei *et al*. We predicted the response values of the 220 drugs for the 49 BC cell lines in GDSC using Wei *et al.*'s method. The ML-CDN2 model performs better than the model of Wei *et al*. (**Figure 4.5**).

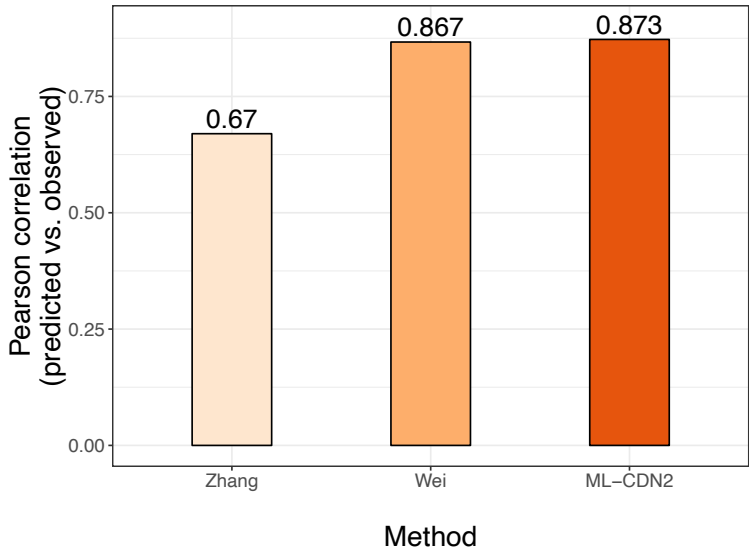


Figure 4.5 Comparison of ML-CDN2 and other network-based models
Bar graph shows the prediction performance of three models, which was estimated based on Pearson correlations (the number on the top of bars) between the predicted and observed IC_{50} values. The first (Zhang) is Zhang *et al.*'s method. The second (Wei) is Wei *et al.*'s method. The third is the ML-CDN2 model in this study using the CSN^{path} and DSNF.

4.3.6 Predicting missing drug responses in GDSC

Out of the possible 49 × 220 BC cell line-drug combinations in the GDSC study, only 81% have corresponding drug response data. With the cell line similarity and drug similarity data, we applied our ML-CDN2 model to predict the missing IC_{50} for these pairs without the responses and compared with those with the responses. We first predicted the missing response of five EGFR inhibitors (afatinib, cetuximab, erlotinib, gefitinib and lapatinib). When grouping the cell lines by their EGFR mutation profiles, it was found that the EGFR-mutant BC cell lines were more sensitive to EGFR inhibitors (**Figure 4.6a**). These predictions were consistent with those in cell lines where response data were observed, and with previously published studies[352,361].

We also predicted the response of three mitogen-activated protein kinase inhibitors (AZD6244, RDEA119 and PD-0325901), for which the response values are missing in a number of BC cell lines. When examining the BRAF mutation status of these cell lines with missing response to the three inhibitors, we found all of them are wild type. **Figure 4.6b** shows the predicted missing IC_{50} values in the BRAF-wildtype cell lines using ML-CDN2 are similar to the response values observed in the BRAF-wildtype cell lines and higher than the existing response values observed in the BRAF-mutant cell lines. Taken together these suggest that our ML-CDN2 model could correctly predict the drug responses of the missing data in the GDSC dataset.

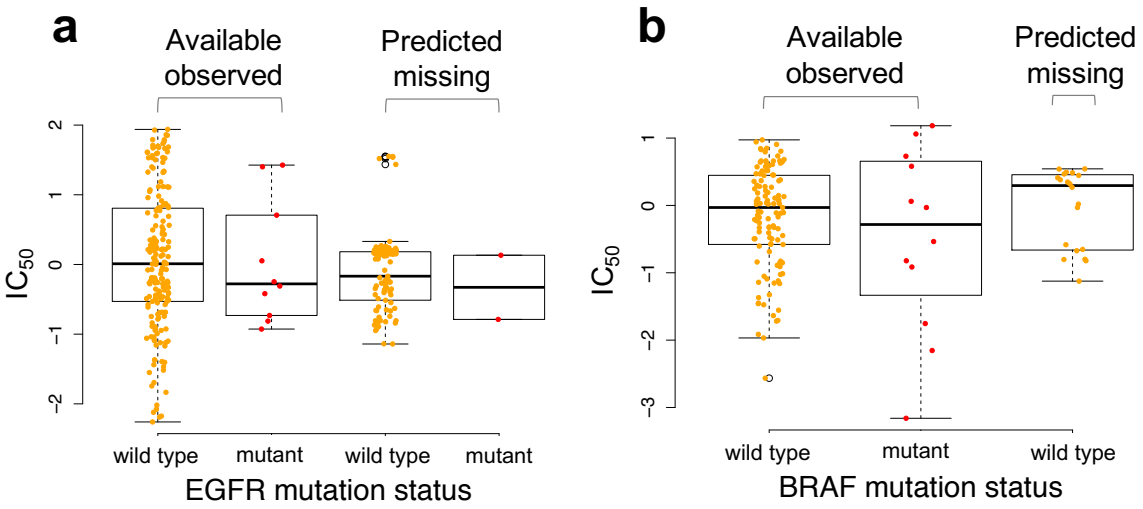


Figure 4.6 Comparison of the predicted missing response values using ML-CDN2 and the existed response values for two types of inhibitors
(**a**) The comparison of predicted and observed activity areas for BRAF mutant and wildtype cell lines for which experimental activity areas were missing from the GDSC dataset for the MEK inhibitors, AZD6244, RDEA119 and PD-0325901. (**b**) The comparison of predicted and observed IC_{50} for EGFR mutant and wildtype cell lines for which experimental IC_{50} values were missing in the GDSC dataset for the EGFR inhibitors, afatinib, cetuximab, erlotinib, gefitinib and lapatinib.

4.3.7 Validating ML-CDN2 in CCLE

We next validated ML-CDN2 in the CCLE dataset, from which the gene expression profiles were extracted for 28 BC cell lines to calculate the pathway activity profiles. The pairwise similarities of the CCLE and GDSC BC cell lines were measured using the Pearson correlations of their pathway activity profiles. We picked the drugs tested in both the CCLE and GDSC datasets, resulting in 13 drugs. Treating each of the 28 cell lines as a new cell line, we employed ML-CDN2 to predict the responses of the new cell line to the 13 drugs. The result is shown in **Figure 4.7**. The Pearson correlation coefficient between the observed and predicted drug responses is 0.718 with a RMSE of 0.783. The results suggest that ML-CDN2 model can be used to predict response values of new BC cell lines to the existing drugs. However, the model did not work well with cell-drug pairs with observed IC_{50} values of 8μM or higher. In CCLE, drugs were tested in eight doses with 8μM being the maximum. Thus, some drugs ended up having an IC_{50} of 8μM or higher in some

cell lines but these are all listed as having an IC_{50} of 8μM. The IC_{50} of 8μM is 0.55 after normalization, hence all IC_{50} values of 8μM or higher from CCLE are represent as the observed normalized value of 0.55 (Figure 4.7). Therefore, some of the predicted IC_{50} values cannot be not linearly correlated with the normalized observed values because of a limitation in the observed IC_{50} data from CCLE.

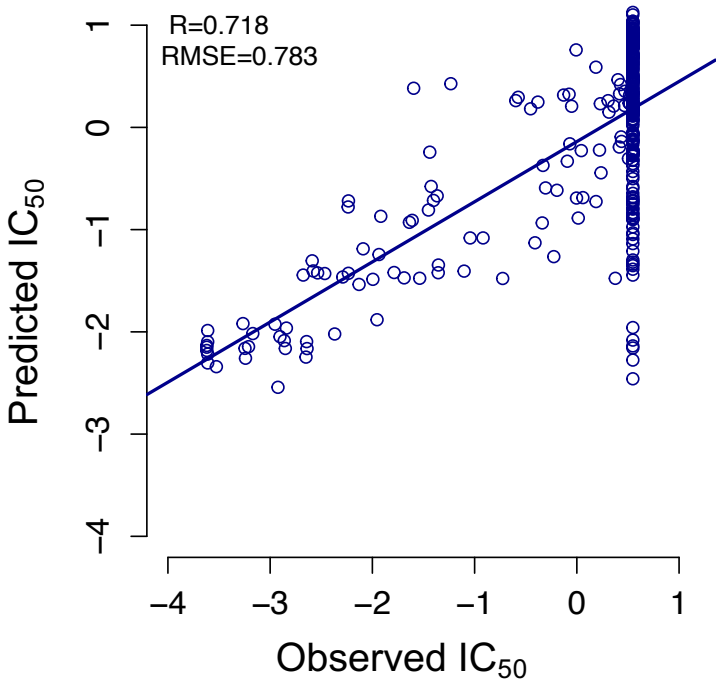


Figure 4.7 Correlation between the predicted and observed drug responses using ML-CDN2 in CCLE. Some cell line-drug pairs have the same normalized observed IC_{50} of 0.55 because the maximum dose in CCLE is 8μM, which cannot correctly represent the true IC_{50} value because it could be higher. However, ML-CDN2 predicted different IC_{50} values for these pairs. So we cannot see a good correlation between the predicted and observed IC_{50} values for the pairs that had an observed IC_{50} value of 8μM or more.

4.3.8 Evaluating ML-CDN2 in TCGA

The TCGA has genomic and transcriptomic information on tumor samples from BC patients, a subset of which have drug response (complete response, partial response, progressive disease, or stable disease) recorded for five drugs that were also tested in the GDSC study:

paclitaxel, fluorouracil, tamoxifen, doxorubicin, and docetaxel. For these breast tumor samples, we calculated the pathway activity scores based on their whole-genome gene expression profiles and further measured the similarity between them and the GDSC BC cell lines. We then applied the ML-CDN2 model to the pathway-based similarity data and predicted the responses of these TCGA breast tumor samples to the five drugs. The predicted IC_{50} values were compared between two groups of tumor samples (Responder and Non-responder) defined according to the recorded drug response. The predicted drug IC_{50} value was lower in the "Responder" patients who were recorded to show a "complete response" to paclitaxel treatment, compared to the patients defined as "Non-responder" who were recorded to show a "partial response", "progressive disease", or "stable disease" to paclitaxel treatment (**Figure 4.8a**, p-value = 0.01 from t-test). Our models did not capture variability in clinical response to the other four drugs. Notably, for tamoxifen, patients in the "Responder" group had a lower median predicted drug IC_{50} value than patients in the "Non-responder" group (**Figure 4.8b**). However, given that there were only five individuals in the Non-responder group, it is not surprising that we did not reach statistical significance. Therefore, a larger clinical cohort will be required to rigorously assess if our models predict the variability in tamoxifen response for BC.

Next, we predicted the response of all TCGA BC samples to lapatinib based on their similarity to the GDSC BC cell lines. Lapatinib is targeting HER2-overexpressing breast cancers and was screened against the GDSC cell lines. When grouping the TCGA BC patients based on their HER2 IHC status, we found that the predicted responses among the four groups are significantly different (**Figure 4.9a**, p-value = 0.01 from ANOVA). Furthermore, the predicted response decreases steadily with HER2 overexpression level, suggesting that drug response is being accurately predicted (**Figure 4.9a**). In addition, we used ML-CDN2 to predict the tamoxifen

response for all TCGA BC samples. In TCGA, ER status was also reported using IHC, as is typical in the clinical setting. Thus, we compared predicted response between the ER-positive and ER-negative groups. Although it should be unsurprising that tamoxifen was predicted to be more sensitive in the ER-positive group these results we not statistically significant (**Figure 4.9b**). These results further suggest that GDSC BC cell line-derived ML-CDN2 could be used to predict the response of BC patients to the existing drugs.

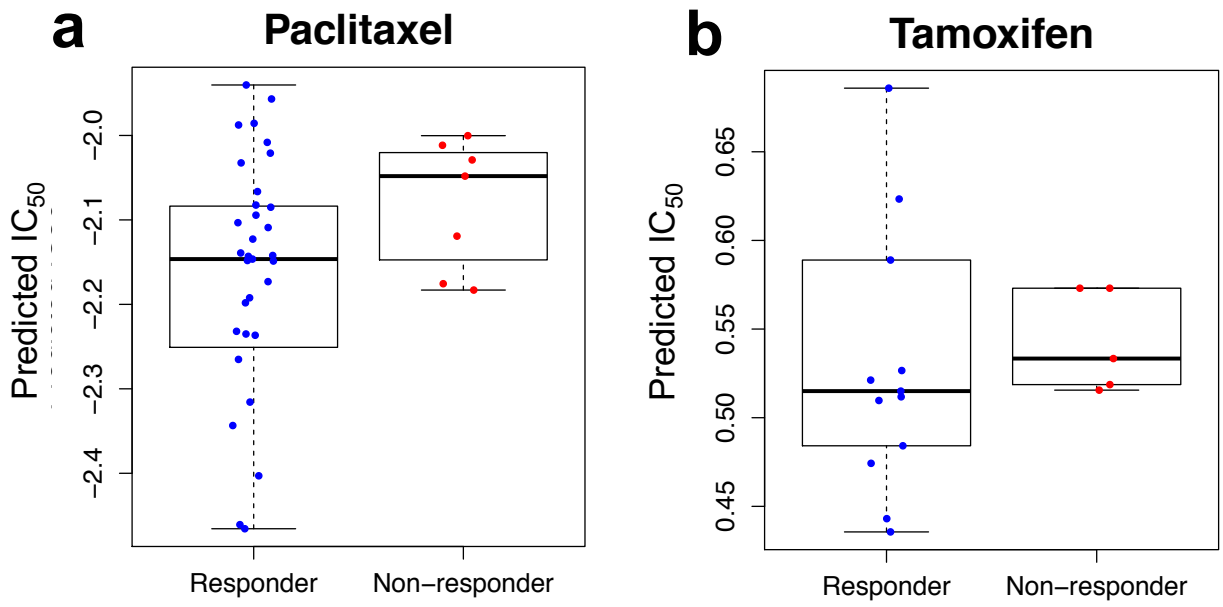


Figure 4.8 Predicted IC_{50} values for TCGA BC samples with recorded drug response
(**a**) Boxplot shows the predicted IC_{50} values for TCGA responders and non-responders to paclitaxel treatment; (**b**) Boxplot shows the predicted IC_{50} values for TCGA responders and non-responders to tamoxifen treatment.

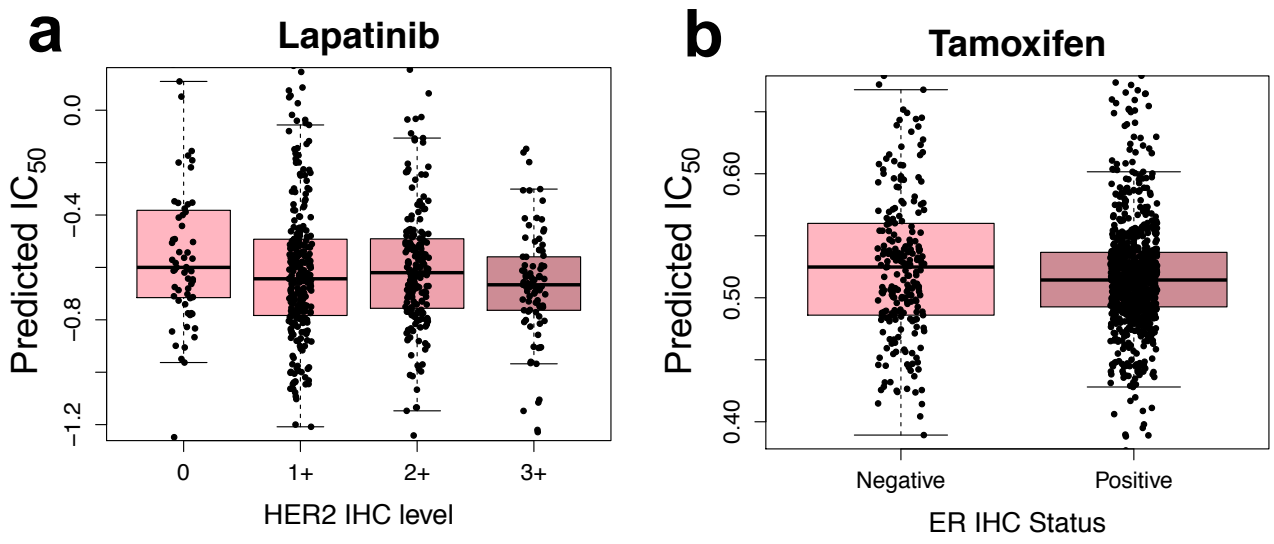


Figure 4.9 Predicted IC_{50} values for TCGA BC samples without recorded drug response
(**a**) Boxplot shows the predicted lapatinib response values in TCGA BC groups with different HER2 overexpression level, which was measured by IHC. (**b**) Boxplot shows the predicted tamoxifen response values in TCGA BC groups with different ER expression status, which was also measured by IHC.

4.3.9 Drug response predictions for TCGA samples have significant associations with expression levels of targeted pathway genes

We applied the ML-CDN2 model to the gene expression data from TCGA BC samples without drug response and predicted the response of seven drugs targeting the EGFR signaling pathway: erlotinib, gefitinib, afatinib, cetuximab, lapatinib, CP724714, and pelitinib. According to the *p*-values derived from multiple linear regression and adjusted by Bonferroni, there are a number of associations between the EGFR pathway genes and the responses of the six drugs predicted by ML-CDN2 (**Table 4.3**). For erlotinib, we found statistically significant associations between the levels of expression of the *ADCY7*, *CDK1*, *FOXO1*, *MAPK1*, and *PAG1* genes and the predicted responses. Lapatinib, exhibited statistically significant associations among the predicted responses and the expression of *AKT3*, *CDK1*, and *ITPR2*. However, no significant associations were observed with gefitinib .

We also employed the ML-CDN2 model to predict the drug response of several PI3K/MTOR signaling pathway inhibitors for TCGA breast tumor samples. Twenty PI3K inhibitors were tested in the GDSC study, and we observed statistically significant associations between the level of pathway gene expression and predicted drug responses for each inhibitor. A total of 120 associations were obtained, and the results for seven inhibitors are listed in **Appendix F**. For temsirolimus, the predicted responses demonstrated statistically significant associations with the expression of *ADRBK1* (p-value = 2.01×10^{-2}), *CXCR4* (p-value = 3.84×10^{-3}), and *NGF* (p-value = 1.92×10^{-2}). For MK-2206, expression of *ADRBK1* (p-value = 2.14×10^{-7}), *PAK4* (p-value = 8.14×10^{-4}), *PPP1CA* (p-value = 1.52×10^{-3}), *PRKAG1*(p-value = 3.32×10^{-2}), and *TRIB3* (p-value = 2.54×10^{-2}) genes were significantly correlated with the predicted response. For dactolisib, the *CXCR4* (p-value = 2.16×10^{-3}) and *FASLG* (p-value = 2.60×10^{-2}) genes were significantly correlated with the predicted response.

Table 4.3 Associations between the expression level of EGFR pathway genes and EGFR inhibitor responses predicted by ML-CDN2

Drug	Gene	Adjusted p-value
Erlotinib	ADCY7	1.51×10^{-3}
	CDK1	2.00×10^{-31}
	FOXO1	4.22×10^{-2}
	MAPK1	1.02×10^{-2}
	PAG1	2.64×10^{-3}
Afatinib	ADAM12	1.64×10^{-5}
	ADCY7	4.63×10^{-4}
	CDK1	8.88×10^{-3}
	EGF	2.94×10^{-2}
Cetuximab	ADCY6	2.67×10^{-2}
	ADRBK1	1.02×10^{-2}
	CDK1	1.59×10^{-19}
	SPRY2	2.18×10^{-4}
Lapatinib	AKT3	8.00×10^{-3}
	CDK1	5.68×10^{-17}
	ITPR2	2.39×10^{-5}
CP724714	CDK1	6.22×10^{-9}

	FOXO1	3.18×10^{-2}
	PDPK1	2.17×10^{-2}
Pelitinib	ADCY9	4.06×10^{-4}
	CDK1	9.07×10^{-5}
	EGFR	2.90×10^{-2}
	ITPR3	3.03×10^{-3}
	PAG1	3.87×10^{-2}
	PIK3R1	1.04×10^{-2}

4.4. Discussion

There are two main steps in the network-based method described here, 1) the construction of cell line and drug similarity networks by using different types of data and 2) developing an effective cell line-drug response model to execute the prediction, and our method improved both steps. For step 1, we integrated multiple types of data from drugs, to construct the drug similarity network fusion. Instead of using gene expression profiles as in other studies [255,352], we used pathways, which contain a set genes in a pathway, to construct the cell line similarity network. For step 2, we used an intuitive weighted model, similar to Wei *et al.*'s method [352], which captured different contributions of existing cell line-drug responses. Thus, our model is more comprehensive and therefore represents an improvement over some existing models [255,352] (**Figure 4.5**).

All the 16 ML-CDNs show a good prediction performance, with the Pearson correlation coefficient greater than 0.8. When the same CSN was used, the ML-CDN model derived from DSN^{targ} shows a smaller Pearson correlation while a greater RMSE than the ML-CDN models derived from DSN^{stru}, DSN^{sens}, and DSNF. These findings suggest that the drug targets are less predictive of drug response in the ML-CDN models compared to the drug structures, pan-cancer IC_{50} profiles, and DSNF. When the same DSN (except DSN^{targ}) was used, the ML-CDN models derived from CSN^{path} and CSNF show a higher Pearson correlation while a lower RMSE than the ML-CDN models derived from CSN^{cnv} and CSN^{mut}, implying the pathway activity profiles and CSNF more informative than the CNV and mutation profiles. The best-preforming one is ML-CDN1 derived from CSNF and DSNF, suggesting that integration of the three types of data from cell lines along with the three types of data from drugs could improve the model.

It is noteworthy that the 16 ML-CDNs do not differ much in their performance. The

Pearson correlation coefficient ranges from 0.833 to 0.875 while the RMSE ranges from 0.493 to 0.563. One possible reason could be that the ML-CDN models are BC-specific models. The idea behind the ML-CDN modeling method is that similar cell lines exhibit similar drug response. These models were trained on 49 BC cell lines, which show high similarity between each other in terms of pathway activity, CNV, or gene mutation profiles. Therefore, any of the three types of profiles can be used by the ML-CDN model to accurately predict drug response.

Recent studies have demonstrated that computational models built on cell line-derived data are applicable to the prediction of drug response for cancer patient samples [249,252,257]. For example, Geeleher *et al.* [249] developed ridge regression models for single drugs using the baseline gene expression data of cancer cell lines as input and the *in vitro* drug IC_{50} values as output. Geeleher *et al*. [252] also applied this ridge regression model to the gene expression profiles of the TCGA tumors to impute the drug response and showed that their cell line-derived models can be used for the accurate prediction of drug response in patients. In our study, the ML-CDN2 model, which was developed on the GDSC dataset, has demonstrated the potential to predict the anti-cancer drug response for breast tumor tissue samples from TCGA. The findings suggest that the high-throughput pharmacogenomics data screened against cancer cell lines could be used for developing computational drug response prediction models, which can further guide precision medicine *in vivo.*

Gene expression data have been extensively used both as a single input and in combination with other omics data for *in silico* drug response prediction [248]. However, most of these models focused on the expression of individual genes. Recent evidence has shown that drug responses were mediated by the coordinate function of a gene set (i.e., pathway) instead of individual genes [362]. In this study, we inferred the pathway activity profiles from the gene expression data. We

estimated the similarity between two GDSC cell lines by calculating the Pearson correlation of their pathway activity profiles, resulting in a mean correlation of all possible cell line pairs of 97.49%, which is higher than the mean (96.83%) of cell line pairwise correlation of the gene expression profiles (**Appendix G**). We also estimated the similarity between every two TCGA BC tumors by computing the Pearson correlation in terms of their genes expression profiles as well as their pathway activity profiles. The mean correlation of all BC tumor pairs is 80.33% for gene expression and 97.49% for pathway activity (**Appendix G**). Moreover, we measured the similarity between the GDSC BC cell lines and the TCGA breast tumor samples. The mean correlation of all possible cell line-tumor pairs is 88.20% for gene expression and 96.57% for pathway activity (**Appendix I**). These finding suggest that pathway activity profiles provide a better way to estimate the similarity of cancer cell line pairs, tumor pairs, as well as cell line-tumor pairs compared to expression profiles of individual genes.

4.5 Conclusions

Instead of using cell lines of various cancer types and making predictions for each drug independently, our model used only BC cell lines to predict responses for all drugs in GDSC. The ML-CDN model (i.e., ML-CDN2) was constructed by connecting a DSN which integrated different data types from drugs to a CSN which was based on the pathway activity profiles of cell lines. We further validated ML-CDN2 in another independent *in vitro* dataset, CCLE and obtained good predictive performance. When applied to the *in vivo* TCGA dataset, the ML-CDN2 model has the potential to predict the drug response for breast tumor tissue samples.

.

Chapter 5: Conclusions and Future Directions

The transcriptome has been extensively studied in cancer research. In this thesis, the regulatory mechanisms underlying the transcriptome profiles of different BC subtypes, as well as the implications of the BC transcriptome for drug treatment were explored. Chapter 5 summarizes the conclusions of each chapter in this thesis and discusses directions for future research.

5.1 Conclusions

The objective of Chapter 2 of this thesis was to explore the regulatory mechanisms underlying transcriptome profiles of each BC subtypes. To achieve this objective, an integrative Lasso regression model for each BC sample was first developed, which took the expression changes of the DEGs in a breast tumor sample relative to the normal breast samples as the response variable while the CNV, DNA methylation, miRNA-binding motif hits and TF-binding motif hits, were used as the explanatory variables. A feature selection procedure to identify the key regulators (miRNAs and TFs) unique to a particular subtype or shared among all subtypes was then performed. A further exploration of the key miRNAs and TFs identified in the TNBC subtype was performed, including identifying the target pathways of the key TNBC regulators and developing CRS signatures for TNBC prognosis using these key regulators. The integrative Lasso regression models identified 25, 20, 15 and 24 key regulators for the luminal A, luminal B, HER2-enriched and TNBC subtypes, respectively. E2F1 and CITED2 were TFs common to all subtypes. miR-374b-5p, miR-32-5p, miR-3150b-3p, miR-361-3p, miR-340-5p, miR-664b-3p and miR-296-5p were common miRNAs to all subtypes. We also found that miR-181b-5p was unique to the luminal A tumors, miR-655-3p, miR-654-5p and miR-625-5p were unique to the luminal B tumors, and miR-9-3p was unique to the TNBCs. When focusing on the TNBC regulators, we found that 6 out of the 24 TNBC regulators (PPARG, PPARA, miR-9-3p, miR-664b-3p, miR-340-5p and miR-

129-5p) regulate the PPARA pathway (i.e., BIOCARTA_PPARA_PATHWAY), while two regulators (miR-340-5p and E2F1) regulate the FOXM1 pathway (i.e., PID_FOXM1_PATHWAY). Moreover, the CRS signatures developed from the TNBC regulators could be used to predict a high-risk and low-risk group stratification with respect to overall survival time. In summary, the results shown in Chapter 2 have suggested the importance of these miRNAs and TFs in the dysregulation of the transcriptome in BC. It has also provided additional evidence of the importance of the FOXM1 and PPARA pathways in BC especially in TNBC.

The objective of Chapter 3 of this thesis was to identify compounds that could modulate the BIOCARTA_PPARA_PATHWAY and the PID_FOXM1_PATHWAY. To achieve this objective, two methods were used. In the first method, data were first curated from the CMAP database, which provides the gene expression profiles of MCF7 cell lines treated with different compounds as well as the paired controls. We then calculated the changes of the FOXM1 and PPARA pathway activities from gene expression profiles under each treatment to identify the drugs that produced a decrease in the FOXM1 pathway activity or an increase in the PPARA pathway activity. In the second method, CMAP database tool was used to identify drugs that could reverse the expression pattern of the two pathways in MCF7. Drugs identified for repressing the FOXM1 pathway or activating the PPARA pathway by the two methods were compared. We identified 19 common drugs (such as 5109870, MG-132, MG-262, celastrol, resveratrol, and cephaeline) that could decrease the FOXM1 pathway activity scores and reverse the FOXM1 pathway expression pattern using the two methods. We also identified 13 common drugs (such as cephaeline, pararosaniline, cycloheximide, monensin, wortmannin, and raloxifene) that could increase the PPARA pathway activity scores and reverse the PPARA pathway expression pattern using the two methods. Cephaelin was found to demonstrate its effect on both of the two pathways.

In summary, the study in Chapter 3 has identified compounds that have the potential to modulate the FOXM1 and PPARA pathway activities in terms of gene expression.

The objective of Chapter 4 of this thesis was to develop anti-cancer drug response prediction models based on the transcriptome profiles of BC. To achieve this objective, data were first curated from the GDSC dataset, which provides the baseline gene expression profiles of 49 BC cell lines along with IC_{50} values of these cell lines to 220 drugs. Secondly, a one-layer cell line similarity network was constructed based on the pathway activity profiles from the transcriptome profiles. Thirdly, a three-layer drug similarity network was constructed by integrating three types of information from the drugs, including the drug structures, targets, and pan-cancer IC_{50} profiles. Fourthly, the cell line similarity network and the drug similarity network were connected to construct a multiple-layer cell line-drug response network (ML-CDN2). ML-CDN2, being the best model, was used to predict drug response in new cell lines or breast tumor tissue samples. We found that the ML-CDN2 demonstrated a good prediction performance, with the Pearson correlation coefficient between the observed and predicted IC_{50} values for all cell line-drug pairs of 0.873. We also found ML-CDN2 showed a good performance when used to predict drug response of new BC cell lines from CCLE, with the Pearson correlation coefficient of 0.718. Moreover, we found that the BC cell line-derived ML-CDN2 model could be applied to the TCGA breast tumor tissue samples for drug response prediction. In summary, the study in Chapter 4 has developed a network-based model using the BC transcriptome profiles and three types of drug data which could be used to predict anti-cancer drug response of BC cell lines as well as BC patient samples.

5.2 Future directions

The regulatory mechanisms underlying the transcriptome profiles in different BC subtypes have been explored *in silico* in Chapter 2. Future research for Chapter 2 may include experimental validation of key miRNAs and TFs identified in each BC subtype. The implications of the BC transcriptome in drug treatment have also been explored *in silico*, including identifying drugs to modulate the two dysregulated pathways in BC in Chapter 3 and developing the anti-cancer drug response prediction model for BC by incorporating the transcriptome profiles in Chapter 4. Future research for Chapter 3 may include investigating the effects of the *in silico* identified compounds on breast tumor cells experimentally. In addition, the study described in Chapter 3 focused on the FOXM1 and PPARA pathways, so future research may try to identify compounds *in silico* that can affect the expression of other dysregulated pathways identified in Chapter 2. Future research for Chapter 4 may extend the ML-CDN2 model to cell lines and tumor samples from various cancer types, which can lead to a larger dataset and may improve the prediction performance.

Bibliography

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68:394–424.
2. Brenner DR, Weir HK, Demers AA, Ellison LF, Louzado C, Shaw A, et al. Projected estimates of cancer in Canada in 2020. Cmaj. 2020;192:E199–205.
3. Dai X, Xiang L, Li T, Bai Z. Cancer Hallmarks, Biomarkers and Breast Cancer Molecular Subtypes. J Cancer. 2016;7:1281–94.
4. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27:1160–7.
5. Kumar B. Breast Cancer Genomics 2014:53–103.
6. Cieslik M, Chinnaiyan AM. Cancer transcriptome profiling at the juncture of clinical translation. Nat Rev Genet. 2018;19:93–109.
7. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz Jr. LA, Kinzler KW, et al. Cancer genome landscapes. Science (80-). 2013;339:1546–58.
8. Stratton M, Campbell P, Futreal A. The cancer genome. Nature. 2009;458:719–24.
9. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin A V, et al. Signatures of mutational processes in human cancer. Nature. 2013;500:415.
10. The Cancer Genome Atlas Research Network, Chang K, Creighton CJ, Davis C, Donehower L, Drummond J, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45:1113–20.
11. Martelotto LG, Ng CKY, Piscuoglio S, Weigelt B, Reis-filho JS. Breast cancer intra-tumor heterogeneity 2014.
12. Koren S, Bentires-Alj M. Breast Tumor Heterogeneity: Source of Fitness, Hurdle for Therapy. Mol Cell. 2015;60:537–46.
13. Polyak K. Heterogeneity in breast cancer. J Clin Invest. 2011;121:3786–8.
14. Turashvili G, Brogi E. Tumor heterogeneity in breast cancer. Front Med. 2017;4.
15. Malhotra GK, Zhao X, Band H, Band V. Histological, molecular and functional subtypes of breast cancers. Cancer Biol Ther. 2010;10:955–60.
16. Brierley JD, Gospodarowicz MK, Wittekind C. TNM classification of malignant tumours. John Wiley & Sons; 2017.

17. Giuliano AE, Connolly JL, Edge SB, Mittendorf EA, Rugo HS, Solin LJ, et al. Breast cancer—major changes in the American Joint Committee on Cancer eighth edition cancer staging manual. CA Cancer J Clin. 2017;67:290–303.
18. Edge SB, Byrd DR, Carducci MA, Compton CC, Fritz AG, Greene FL. AJCC cancer staging manual. vol. 7. Springer New York; 2010.
19. Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. CW Elston & IO Ellis. Histopathology 1991; 19; 403–410: AUTHOR COMMENTARY. Histopathology. 2002;41:151.
20. Bloom HJG, Richardson WW. Histological grading and prognosis in breast cancer: a study of 1409 cases of which 359 have been followed for 15 years. Br J Cancer. 1957;11:359.
21. Davidson TM, Rendi MH, Frederick PD, Onega T, Allison KH, Mercan E, et al. Breast cancer prognostic factors in the digital era: Comparison of Nottingham grade using whole slide images and glass slides. J Pathol Inform. 2019;10.
22. Harbeck N, Penault-Llorca F, Cortes J, Gnant M, Houssami N, Poortmans P, et al. Breast cancer. vol. 5. 2019.
23. Beca F, Polyak K. Intratumor heterogeneity in breast cancer. Nov. Biomarkers Contin. Breast Cancer, Springer; 2016, p. 169–89.
24. Hammond MEH, Hayes DF, Dowsett M, Allred DC, Hagerty KL, Badve S, et al. American society of clinical oncology/college of american pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. J Clin Oncol. 2010;28:2784–95.
25. Harvey JM, Clark GM, Osborne CK, Allred DC. Estrogen receptor status by immunohistochemistry is superior to the ligand-binding assay for predicting response to adjuvant endocrine therapy in breast cancer. J Clin Oncol. 1999;17:1474–81.
26. Bardou V-J, Arpino G, Elledge RM, Osborne CK, Clark GM. Progesterone receptor status significantly improves outcome prediction over estrogen receptor status alone for adjuvant endocrine therapy in two large breast cancer databases. J Clin Oncol. 2003;21:1973–9.
27. Lakhani SR. WHO Classification of Tumours of the Breast. International Agency for Research on Cancer; 2012.
28. Wolff AC, Elizabeth Hale Hammond M, Allison KH, Harvey BE, Mangu PB, Bartlett JMS,

et al. Human epidermal growth factor receptor 2 testing in breast cancer: American society of clinical oncology/ college of American pathologists clinical practice guideline focused update. J Clin Oncol. 2018;36:2105–22.

29. Dean-Colomb W, Esteva FJ. Her2-positive breast cancer: herceptin and beyond. Eur J Cancer. 2008;44:2806–12.
30. Saleh SMI, Bertos N, Gruosso T, Gigoux M, Souleimanova M, Zhao H, et al. Identification of interacting stromal axes in triple-negative breast cancer. Cancer Res. 2017;77:4673–83.
31. Goldhirsch A, Winer EP, Coates AS, Gelber RD, Piccart-Gebhart M, Thürlimann B, et al. Personalizing the treatment of women with early breast cancer: Highlights of the st gallen international expert consensus on the primary therapy of early breast Cancer 2013. Ann Oncol. 2013;24:2206–23.
32. Perou CM, Sørlie T, Eisen MB, Van De Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000;406:747.
33. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci. 2001;98:10869–74.
34. Sørlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci. 2003;100:8418–23.
35. Skibinski A, Kuperwasser C. The origin of breast tumor heterogeneity. Oncogene. 2015;34:5309–16.
36. Runa F, Hamalian S, Meade K, Shisgal P, Gray PC, Kelber JA. Tumor microenvironment heterogeneity: challenges and opportunities. Curr Mol Biol Reports. 2017;3:218–29.
37. Quail DF, Joyce JA. Microenvironmental regulation of tumor progression and metastasis. Nat Med. 2013;19:1423.
38. Kalinsky K, Heguy A, Bhanot UK, Patil S, Moynahan ME. PIK3CA mutations rarely demonstrate genotypic intratumoral heterogeneity and are selected for in breast cancer progression. Breast Cancer Res Treat. 2011;129:635.
39. Balko JM, Giltnane JM, Wang K, Schwarz LJ, Young CD, Cook RS, et al. Molecular profiling of the residual disease of triple-negative breast cancers after neoadjuvant chemotherapy identifies actionable therapeutic targets. Cancer Discov. 2014;4:232–45.

40. Ding LI, Ellis MJ, Li S, Larson DE, Chen K, Wallis JW, et al. Genome remodelling in a basal-like breast cancer metastasis and xenograft. Nature. 2010;464:999–1005.
41. Seol H, Lee HJ, Choi Y, Lee HE, Kim YJ, Kim JH, et al. Intratumoral heterogeneity of HER2 gene amplification in breast cancer: Its clinicopathological significance. Mod Pathol. 2012;25:938–48.
42. Amir E, Miller N, Geddie W, Freedman O, Kassam F, Simmons C, et al. Prospective study evaluating the impact of tissue confirmation of metastatic disease in patients with breast cancer. J Clin Oncol. 2012;30:587–92.
43. Greaves M, Maley CC. Clonal evolution in cancer. Nature. 2012;481:306.
44. Meacham CE, Morrison SJ. Tumour heterogeneity and cancer cell plasticity. Nature. 2013;501:328–37.
45. Prasetyanti PR, Medema JP. Intra-tumor heterogeneity from a cancer stem cell perspective. Mol Cancer. 2017;16:41.
46. Rye IH, Trinh A, Sætersdal AB, Nebdal D, Lingjærde OC, Almendro V, et al. Intratumor heterogeneity defines treatment-resistant HER2+ breast tumors. Mol Oncol. 2018;12:1838–55.
47. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000;406:747–52.
48. Hu Z, Fan C, Oh DS, Marron JS, He X, Qaqish BF, et al. The molecular portraits of breast tumors are conserved across microarray platforms. BMC Genomics. 2006;7:96.
49. Gnant M, Filipits M, Greil R, Stoeger H, Rudas M, Bago-Horvath Z, et al. Predicting distant recurrence in receptor-positive breast cancer patients with limited clinicopathological risk: Using the PAM50 Risk of Recurrence score in 1478 postmenopausal patients of the ABCSG-8 trial treated with adjuvant endocrine therapy alone. Ann Oncol. 2014;25:339–45.
50. Senkus E, Kyriakides S, Ohno S, Penault-Llorca F, Poortmans P, Rutgers E, et al. Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2015;26:v8–30.
51. Curigliano G, Burstein HJ, Winer EP, Gnant M, Dubsky P, Loibl S, et al. De-escalating and escalating treatments for early-stage breast cancer: the St. Gallen International Expert Consensus Conference on the Primary Therapy of Early Breast Cancer 2017. Ann Oncol. 2017;28:1700–12.

52. Lundgren C, Bendahl PO, Borg Å, Ehinger A, Hegardt C, Larsson C, et al. Agreement between molecular subtyping and surrogate subtype classification: a contemporary population-based study of ER-positive/HER2-negative primary breast cancer. Breast Cancer Res Treat. 2019;178:459–67.
53. Russnes HG, Lingjærde OC, Børresen-Dale AL, Caldas C. Breast Cancer Molecular Stratification: From Intrinsic Subtypes to Integrative Clusters. Am J Pathol. 2017;187:2152–62.
54. Ferrari A, Vincent-Salomon A, Pivot X, Sertier AS, Thomas E, Tonon L, et al. A whole-genome sequence and transcriptome perspective on HER2-positive breast cancers. Nat Commun. 2016;7.
55. Carey LA, Berry DA, Cirrincione CT, Barry WT, Pitcher BN, Harris LN, et al. Molecular heterogeneity and response to neoadjuvant human epidermal growth factor receptor 2 targeting in CALGB 40601, a randomized phase III trial of paclitaxel plus trastuzumab with or without lapatinib. J Clin Oncol. 2016;34:542–9.
56. Tofigh A, Suderman M, Paquet ER, Livingstone J, Bertos N, Saleh SM, et al. The prognostic ease and difficulty of invasive breast carcinoma. Cell Rep. 2014;9:129–42.
57. Network CGA, Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490:61–70.
58. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486:346.
59. Perez EA, Ballman K V., Mashadi-Hossein A, Tenner KS, Kachergus JM, Norton N, et al. Intrinsic subtype and therapeutic response among HER2-positive breast tumors from the NCCTG (Alliance) N9831 trial. J Natl Cancer Inst. 2017;109:1–8.
60. Staaf J, Jönsson G, Ringnér M, Vallon-Christersson J, Grabau D, Arason A, et al. High-resolution genomic and expression analyses of copy number alterations in HER2-amplified breast cancer. Breast Cancer Res. 2010;12.
61. Sahlberg KK, Hongisto V, Edgren H, Mäkelä R, Hellström K, Due EU, et al. The HER2 amplicon includes several genes required for the growth and survival of HER2 positive breast cancer cells. Mol Oncol. 2013;7:392–401.
62. Milioli HH, Tishchenko I, Riveros C, Berretta R, Moscato P. Basal-like breast cancer:

molecular profiles, clinical features and survival outcomes. BMC Med Genomics. 2017;10:1–17.

63. Leidy J, Khan A, Kandil D. Basal-like breast cancer: Update on clinicopathologic, immunohistochemical, and molecular features. Arch Pathol Lab Med. 2014;138:37–43.
64. Badve S, Dabbs DJ, Schnitt SJ, Baehner FL, Decker T, Eusebi V, et al. Basal-like and triple-negative breast cancers: A critical review with an emphasis on the implications for pathologists and oncologists. Mod Pathol. 2011;24:157–67.
65. Fulford LG, Reis-Filho JS, Ryder K, Jones C, Gillett CE, Hanby A, et al. Basal-like grade III invasive ductal carcinoma of the breast: Patterns of metastasis and long-term survival. Breast Cancer Res. 2007;9:1–11.
66. Prat A, Adamo B, Cheang MCU, Anders CK, Carey LA, Perou CM. Molecular Characterization of Basal-Like and Non-Basal-Like Triple-Negative Breast Cancer. Oncologist. 2013;18:123–33.
67. Yersal O, Barutca S. Biological subtypes of breast cancer: Prognostic and therapeutic implications. World J Clin Oncol. 2014;5:412–24.
68. Herschkowitz JI, Simin K, Weigman VJ, Mikaelian I, Usary J, Hu Z, et al. Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. Genome Biol. 2007;8:R76.
69. Prat A, Perou CM. Deconstructing the molecular portraits of breast cancer. Mol Oncol. 2011;5:5–23.
70. Prat A, Parker JS, Karginova O, Fan C, Livasy C, Herschkowitz JI, et al. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. Breast Cancer Res. 2010;12:R68.
71. Lehmann BDB, Bauer JAJ, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J Clin Invest. 2011;121:2750–67.
72. Sotiriou C, Neo S-YY, McShane LM, Korn EL, Long PM, Jazaeri A, et al. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. Proc Natl Acad Sci. 2003;100:10393–8.
73. Desmedt C, Haibe-Kains B, Wirapati P, Buyse M, Larsimont D, Bontempi G, et al. Biological processes associated with breast cancer clinical outcome depend on the

molecular subtypes. Clin Cancer Res. 2008;14:5158–65.

74. Wirapati P, Sotiriou C, Kunkel S, Farmer P, Pradervand S, Haibe-Kains B, et al. Meta-analysis of gene expression profiles in breast cancer: Toward a unified understanding of breast cancer subtyping and prognosis signatures. Breast Cancer Res. 2008;10:1–11.
75. Haibe-Kains B, Desmedt C, Loi S, Culhane AC, Bontempi G, Quackenbush J, et al. A three-gene model to robustly identify breast cancer molecular subtypes. J Natl Cancer Inst. 2012;104:311–25.
76. Weigelt B, Mackay A, A'hern R, Natrajan R, Tan DSP, Dowsett M, et al. Breast cancer molecular profiling with single sample predictors: A retrospective analysis. Lancet Oncol. 2010;11:339–49.
77. Guedj M, Marisa L, De Reynies A, Orsetti B, Schiappa R, Bibeau F, et al. A refined molecular taxonomy of breast cancer. Oncogene. 2012;31:1196–206.
78. Jönsson G, Staaf J, Vallon-Christersson J, Ringnér M, Holm K, Hegardt C, et al. Genomic subtypes of breast cancer identified by array-comparative genomic hybridization display distinct molecular and clinical characteristics. Breast Cancer Res. 2010;12.
79. Ali HR, Rueda OM, Chin SF, Curtis C, Dunning MJ, Aparicio SAJR, et al. Genome-driven integrated classification of breast cancer validated in over 7,500 samples. Genome Biol. 2014;15:1–14.
80. Fu NY, Rios AC, Pal B, Law CW, Jamieson P, Liu R, et al. Identification of quiescent and spatially restricted mammary stem cells that are hormone responsive. Nat Cell Biol. 2017;19:164–76.
81. Bareche Y, Venet D, Ignatiadis M, Aftimos P, Piccart M, Rothe F, et al. Unravelling triple-negative breast cancer molecular heterogeneity using an integrative multiomic analysis. Ann Oncol. 2018;29:895–902.
82. Lehmann BD, Pietenpol JA. Clinical implications of molecular heterogeneity in triple negative breast cancer. Breast. 2015;24:S36–40.
83. Lehmann BD, Jovanović B, Chen X, Estrada M V., Johnson KN, Shyr Y, et al. Refinement of Triple-Negative Breast Cancer Molecular Subtypes: Implications for Neoadjuvant Chemotherapy Selection. PLoS One. 2016;11:e0157368.
84. Burstein MD, Tsimelzon A, Poage GM, Covington KR, Contreras A, Fuqua SAW, et al. Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative

breast cancer. Clin Cancer Res. 2015;21:1688–98.

85. Jézéquel P, Loussouarn D, Guérin-Charbonnel C, Campion L, Vanier A, Gouraud W, et al. Gene-expression molecular subtyping of triple-negative breast cancer tumours: importance of immune response. Breast Cancer Res. 2015;17:43.
86. Van den Berge K, Hembach KM, Soneson C, Tiberi S, Clement L, Love MI, et al. RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis. Annu Rev Biomed Data Sci. 2019;2:139–73.
87. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. PLoS Comput Biol. 2017;13:e1005457.
88. Manzoni C, Kia DA, Vandrovcova J, Hardy J, Wood NW, Lewis PA, et al. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. Br Bioinform. 2018;19:286–302.
89. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14:R36.
90. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
91. Kim D, Langmead B, Salzberg SL. HISAT: A fast spliced aligner with low memory requirements. Nat Methods. 2015;12:357–60.
92. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.
93. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7:562–78.
94. Turro E, Su S-Y, Gonçalves Â, Coin LJM, Richardson S, Lewin A. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. Genome Biol. 2011;12:R13.
95. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31:166–9.
96. Racle J, de Jonge K, Baumgaertner P, Speiser DE, Gfeller D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. Elife. 2017;6:e26476.

97. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15:R29.
98. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.
99. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
100. Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, et al. Evaluation of tools for differential gene expression analysis by RNA-seq on a 48 biological replicate experiment. ArXiv Prepr ArXiv150502017. 2015.
101. Creixell P, Reimand J, Haider S, Wu G, Shibata T, Vazquez M, et al. Pathway and network analysis of cancer genomes. Nat Methods. 2015;12:615–21.
102. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8:e1002375.
103. Consortium GO. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 2004;32:D258–61.
104. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27:1739–40.
105. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.
106. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2017;46:D649–55.
107. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, et al. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. 2005;33:D428–32.
108. Nishimura D. BioCarta. Biotech Softw Internet Rep Comput Softw J Sci. 2001;2:117–20.
109. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, et al. PID: the pathway interaction database. Nucleic Acids Res. 2008;37:D674–9.
110. Newman JC, Weiner AM. L2L: a simple tool for discovering the hidden significance in microarray expression data. Genome Biol. 2005;6.
111. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2008;37:1–13.
112. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene

lists using DAVID bioinformatics resources. Nat Protoc. 2009;4:44.

113. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. Omi a J Integr Biol. 2012;16:284–7.
114. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–50.
115. Efron B, Tibshirani R. On testing the significance of sets of genes. Ann Appl Stat. 2007;1:107–29.
116. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 2009;10:1–7.
117. Goeman JJ, Bühlmann P. Analyzing gene expression data in terms of gene sets: Methodological issues. Bioinformatics. 2007;23:980–7.
118. Tomfohr J, Lu J, Kepler TB. Pathway level analysis of gene expression using singular value decomposition. BMC Bioinformatics. 2005;6:1–11.
119. Goeman JJ, Van de Geer S, De Kort F, van Houwellingen HC. A global test for groups fo genes: Testing association with a clinical outcome. Bioinformatics. 2004;20:93–9.
120. Goeman JJ, Oosting J, Cleton-Jansen AM, Anninga JK, van Houwelingen HC. Testing association of a pathway with survival using gene expression data. Bioinformatics. 2005;21:1950–7.
121. Lee E, Chuang HY, Kim JW, Ideker T, Lee D. Inferring pathway activity toward precise disease classification. PLoS Comput Biol. 2008;4.
122. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature. 2009;462:108–12.
123. Gundem G, Lopez-Bigas N. Sample-level enrichment analysis unravels shared stress phenotypes among multiple cancer types. Genome Med. 2012;4.
124. Hänzelmann S, Castelo R, Guinney J. GSVA: Gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics. 2013;14.
125. Paquet ER, Lesurf R, Tofigh A, Dumeaux V, Hallett MT. Detecting gene signature activation in breast cancer in an absolute, single-patient manner. Breast Cancer Res.

2017;19:1–15.
126. Paquet ER, Hallett MT. Absolute assignment of breast cancer intrinsic molecular subtype. J Natl Cancer Inst. 2015;107:1–9.
127. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Genet. 2006;7:85.
128. Shlien A, Malkin D. Copy number variations and cancer. Genome Med. 2009;1:62.
129. Adkison LR. Elsevier's Integrated Review Genetics E-Book: With STUDENT CONSULT Online Access. Elsevier Health Sciences; 2011.
130. Gonçalves E, Fragoulis A, Garcia-Alonso L, Cramer T, Saez-Rodriguez J, Beltrao P. Widespread post-transcriptional attenuation of genomic copy-number variation in cancer. Cell Syst. 2017;5:386–98.
131. Thapar A, Cooper M. Copy number variation: what is it and what has it told us about child psychiatric disorders? J Am Acad Child Adolesc Psychiatry. 2013;52:772–4.
132. Santarius T, Shipley J, Brewer D, Stratton MR, Cooper CS. A census of amplified and overexpressed human cancer genes. Nat Rev Cancer. 2010;10:59.
133. Consortium IH. A haplotype map of the human genome. Nature. 2005;437:1299.
134. Lilyquist J, Ruddy KJ, Vachon CM, Couch FJ. Common genetic variation and breast cancer Risk—Past, present, and future. Cancer Epidemiol Biomarkers Prev. 2018;27:380–94.
135. Huo Y, Li S, Liu J, Li X, Luo XJ. Functional genomics reveal gene regulatory mechanisms underlying schizophrenia risk. Nat Commun. 2019;10.
136. Ryan BM, Robles AI, Harris CC. Genetic variation in microRNA networks: the implications for cancer research. Nat Rev Cancer. 2010;10:389–402.
137. Ryland GL, Doyle MA, Goode D, Boyle SE, Choong DYH, Rowley SM, et al. Loss of heterozygosity: What is it good for? BMC Med Genomics. 2015;8:1–12.
138. Moolgavkar SH, Knudson AG. Mutation and cancer: A model for human carcinogenesis. J Natl Cancer Inst. 1981;66:1037–52.
139. Knudson AG. Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci U S A. 1971;68:820–3.
140. Wang LH, Wu CF, Rajasekaran N, Shin YK. Loss of tumor suppressor gene function in human cancer: An overview. Cell Physiol Biochem. 2019;51:2647–93.
141. Turner N, Tutt A, Ashworth A. Hallmarks of'BRCAness' in sporadic cancers. Nat Rev

Cancer. 2004;4:814–9.

142. Maxwell KN, Wubbenhorst B, Wenz BM, De Sloover D, Pluta J, Emery L, et al. BRCA locus-specific loss of heterozygosity in germline BRCA1 and BRCA2 carriers. Nat Commun. 2017;8:1–11.
143. Fang L, Wang K. Identification of Copy Number Variants from SNP Arrays Using PennCNV. Copy Number Var., Springer; 2018, p. 1–28.
144. Shen W, Szankasi P, Durtschi J, Kelley TW, Xu X. Genome-wide copy number variation detection using NGS: data analysis and interpretation. Tumor Profiling, Springer; 2019, p. 113–24.
145. Abel HJ, Duncavage EJ. Detection of structural DNA variation from next generation sequencing data: A review of informatic approaches. Cancer Genet. 2013;206:432–40.
146. Li W, Xia Y, Wang C, Tang YT, Guo W, Li J, et al. Identifying human genome-wide CNV, LOH and UPD by targeted sequencing of selected regions. PLoS One. 2015;10:1–18.
147. Robertson KD. DNA methylation and human disease. Nat Rev Genet. 2005;6:597.
148. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. Nat Rev Genet. 2013;14:204.
149. Bird A. DNA methylation patterns and epigenetic memory. Genes Dev. 2002;16:6–21.
150. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012;13:484.
151. Domcke S, Bardet AF, Ginno PA, Hartl D, Burger L, Schübeler D. Competition between DNA methylation and transcription factors determines binding of NRF1. Nature. 2015;528:575.
152. Blattler A, Farnham PJ. Cross-talk between site-specific transcription factors and DNA methylation states. J Biol Chem. 2013;288:34287–94.
153. Chatterjee R, Vinson C. CpG methylation recruits sequence specific transcription factors essential for tissue specific gene expression. Biochim Biophys Acta (BBA)-Gene Regul Mech. 2012;1819:763–70.
154. Jovanovic J, Rønneberg JA, Tost J, Kristensen V. The epigenetics of breast cancer. Mol Oncol. 2010;4:242–54.
155. Davalos V, Martinez-Cardus A, Esteller M. The Epigenomic Revolution in Breast Cancer: From Single-Gene to Genome-Wide Next-Generation Approaches. Am J Pathol.

2017;187:2163–74.

156. Basse C, Arock M. The increasing roles of epigenetics in breast cancer: Implications for pathogenicity, biomarkers, prevention and treatment. Int J Cancer. 2015;137:2785–94.

157. Pasculli B, Barbano R, Parrella P. Epigenetics of breast cancer: biology and clinical implication in the era of precision medicine. Semin. Cancer Biol., vol. 51, Elsevier; 2018, p. 22–35.

158. Huang KT, Mikeska T, Li J, Takano EA, Millar EKA, Graham PH, et al. Assessment of DNA methylation profiling and copy number variation as indications of clonal relationship in ipsilateral and contralateral breast cancers to distinguish recurrent breast cancer from a second primary tumour. BMC Cancer. 2015;15:669.

159. He D, Zhang Y, Zhang N, Zhou L, Chen J, Jiang Y, et al. Aberrant gene promoter methylation of p16, FHIT, CRBP1, WWOX, and DLC-1 in Epstein–Barr virus-associated gastric carcinomas. Med Oncol. 2015;32:92.

160. Pfeifer G. Defining driver DNA methylation changes in human cancer. Int J Mol Sci. 2018;19:1166.

161. Zwergel C, Valente S, Mai A. DNA methyltransferases inhibitors from natural sources. Curr Top Med Chem. 2016;16:680–96.

162. Gupta R, Nagarajan A, Wajapeyee N. Advances in genome-wide DNA methylation analysis. Biotechniques. 2010;49:iii–xi.

163. Jones PA, Issa JP, Baylin S. Targeting the cancer epigenome for therapy. Nat Rev Genet. 2016;17:630–41.

164. Wang Z, Wu X, Wang Y. A framework for analyzing DNA methylation data from Illumina Infinium HumanMethylation450 BeadChip. BMC Bioinformatics. 2018;19:115.

165. Ha M, Kim VN. Regulation of microRNA biogenesis. Nat Rev Mol Cell Biol. 2014;15:509.

166. He L, Hannon GJ. MicroRNAs: small RNAs with a big role in gene regulation. Nat Rev Genet. 2004;5:522.

167. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. Nucleic Acids Res. 2018;47:D155–62.

168. Ramassone A, Pagotto S, Veronese A, Visone R. Epigenetics and microRNAs in cancer. Int J Mol Sci. 2018;19:459.

169. Suzuki H, Maruyama R, Yamamoto E, Kai M. DNA methylation and microRNA

dysregulation in cancer. Mol Oncol. 2012;6:567–78.

170. Peng Y, Croce CM. The role of MicroRNAs in human cancer. Signal Transduct Target Ther. 2016;1:15004.

171. Lehmann U, Hasemeier B, Christgen M, Müller M, Römermann D, Länger F, et al. Epigenetic inactivation of microRNA gene hsa-mir-9-1 in human breast cancer. J Pathol A J Pathol Soc Gt Britain Irel. 2008;214:17–24.

172. Hsu P-Y, Deatherage DE, Rodriguez BAT, Liyanarachchi S, Weng Y-I, Zuo T, et al. Xenoestrogen-induced epigenetic repression of microRNA-9-3 in breast epithelial cells. Cancer Res. 2009;69:5936–45.

173. Pronina I V, Loginov VI, Burdennyy AM, Fridman M V, Senchenko VN, Kazubskaya TP, et al. DNA methylation contributes to deregulation of 12 cancer-associated microRNAs and breast cancer progression. Gene. 2017;604:1–8.

174. Moskwa P, Buffa FM, Pan Y, Panchakshari R, Gottipati P, Muschel RJ, et al. miR-182-mediated downregulation of BRCA1 impacts DNA repair and sensitivity to PARP inhibitors. Mol Cell. 2011;41:210–20.

175. Chiang C-H, Hou M-F, Hung W-C. Up-regulation of miR-182 by β-catenin in breast cancer increases tumorigenicity and invasiveness by targeting the matrix metalloproteinase inhibitor RECK. Biochim Biophys Acta (BBA)-General Subj. 2013;1830:3067–76.

176. Liu H, Wang Y, Li X, Zhang Y, Li J, Zheng Y, et al. Expression and regulatory function of miRNA-182 in triple-negative breast cancer cells through its targeting of profilin 1. Tumor Biol. 2013;34:1713–22.

177. Guttilla IK, White BA. Coordinate regulation of FOXO1 by miR-27a, miR-96, and miR-182 in breast cancer cells. J Biol Chem. 2009;284:23204–16.

178. Li P, Sheng C, Huang L, Zhang H, Huang L, Cheng Z, et al. MiR-183/-96/-182 cluster is up-regulated in most breast cancers and increases cell proliferation and migration. Breast Cancer Res. 2014;16:473.

179. Hu Y, Lan W, Miller D. Next-generation sequencing for MicroRNA expression profile. Bioinforma. MicroRNA Res., Springer; 2017, p. 169–77.

180. Bhagwat AS, Vakoc CR. Targeting transcription factors in cancer. Trends in Cancer. 2015;1:53–65.

181. Lee TI, Young RA. Transcriptional regulation and its misregulation in disease. Cell.

2013;152:1237–51.

182. Bradner JE, Hnisz D, Young RA. Transcriptional addiction in cancer. Cell. 2017;168:629–43.

183. Bushweller JH. Targeting transcription factors in cancer — from undruggable to reality. Nat Rev Cancer. 2019;19:611–24.

184. Repana D, Nulsen J, Dressler L, Bortolomeazzi M, Venkata SK, Tourna A, et al. The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. Genome Biol. 2019;20:1–12.

185. Ravasi T, Suzuki H, Cannistraci C V., Katayama S, Bajic VB, Tan K, et al. An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man. Cell. 2010;140:744–52.

186. Hamed M, Spaniol C, Nazarieh M, Helms V. TFmiR: a web server for constructing and analyzing disease-specific transcription factor and miRNA co-regulatory networks. Nucleic Acids Res. 2015;43:W283-8.

187. Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. miRecords: an integrated resource for microRNA–target interactions. Nucleic Acids Res. 2008;37:D105–10.

188. Hsu S-D, Tseng Y-T, Shrestha S, Lin Y-L, Khaleel A, Chou C-H, et al. miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. Nucleic Acids Res. 2014;42:D78–85.

189. Li J-H, Liu S, Zhou H, Qu L-H, Yang J-H. starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Res. 2013;42:D92–7.

190. The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 2004;306:636–40.

191. Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee R, et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res. 2018;46:D260–6.

192. Han H, Cho JW, Lee S, Yun A, Kim H, Bae D, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res. 2018;46:D380–6.

193. Jiang C, Xuan Z, Zhao F, Zhang MQ. TRED: a transcriptional regulatory element database, new entries and other development. Nucleic Acids Res. 2007;35:D137–40.

194. Cheneby J, Gheorghe M, Artufel M, Mathelier A, Ballester B, Chèneby J, et al. ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. Nucleic Acids Res. 2018;46:D267–75.
195. Mei S, Qin Q, Wu Q, Sun H, Zheng R, Zang C, et al. Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. Nucleic Acids Res. 2017;45:D658–62.
196. Han H, Shim H, Shin D, Shim JE, Ko Y, Shin J, et al. TRRUST: a reference database of human transcriptional regulatory interactions. Sci Rep. 2015;5:11432.
197. Huynh-Thu VA, Sanguinetti G. Gene Regulatory Network Inference: An Introductory Survey. Methods Mol Biol. 2019;1883:1–23.
198. Mercatelli D, Scalambra L, Triboli L, Ray F, Giorgi FM. Gene regulatory network inference resources: A practical overview. Biochim Biophys Acta - Gene Regul Mech. 2019:194430.
199. Schlitt T. Approaches to modeling gene regulatory networks: a gentle introduction. Silico Syst. Biol., Springer; 2013, p. 13–35.
200. Hecker M, Lambeck S, Toepfer S, van Someren E, Guthke R. Gene regulatory network inference: Data integration in dynamic models-A review. BioSystems. 2009;96:86–103.
201. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005;4.
202. Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics. 1970;12:55–67.
203. Tishbirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B. 1996;58:267–88.
204. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B (Statistical Methodol. 2005;67:301–20.
205. Setty M, Helmy K, Khan AA, Silber J, Arvey A, Neezen F, et al. Inferring transcriptional and microRNA-mediated regulatory programs in glioblastoma. Mol Syst Biol. 2012;8:605.
206. Li Y, Liang M, Zhang Z. Regression Analysis of Combined Gene Expression Regulation in Acute Myeloid Leukemia. PLoS Comput Biol. 2014;10.
207. Van't Veer LJ, Dai H, Van De Vijver MJ, He YD, Hart AAM, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415:530–6.
208. Van De Vijver MJ, He YD, Van't Veer LJ, Dai H, Hart AAM, Voskuil DW, et al. A gene-

expression signature as a predictor of survival in breast cancer. N Engl J Med. 2002;347:1999–2009.

209. Glas AM, Floore A, Delahaye LJMJ, Witteveen AT, Pover RCF, Bakx N, et al. Converting a breast cancer microarray signature into a high-throughput diagnostic test. BMC Genomics. 2006;7:1–10.

210. Buyse M, Loi S, van't Veer L, Viale G, Delorenzi M, Glas AM, et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J Natl Cancer Inst. 2006;98:1183–92.

211. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med. 2004;351:2817–26.

212. Paik S, Tang G, Shak S, Kim C, Baker J, Kim W, et al. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. J Clin Oncol. 2006;24:3726–34.

213. Albain K, Barlow W, Shak S, Horobagyi G, Livingston R, Yeh I-T, et al. Prognostic and Predictive Value of the 21-Gene Recurrence Score Assay in a Randomized Trial of Chemotherapy for Postmenopausal, Node-Positive, Estrogen Receptor-Positive Breast Cancer. Lancet Oncol. 2010;11:55–65.

214. Sparano JA, Gray RJ, Makower DF, Pritchard KI, Albain KS, Hayes DF, et al. Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. N Engl J Med. 2018;379:111–21.

215. Sparano JA, Gray RJ, Makower DF, Pritchard KI, Albain KS, Hayes DF, et al. Prospective validation of a 21-gene expression assay in breast cancer. N Engl J Med. 2015;373:2005–14.

216. Solin LJ, Gray R, Baehner FL, Butler SM, Hughes LL, Yoshizawa C, et al. A multigene expression assay to predict local recurrence risk for ductal carcinoma in situ of the breast. J Natl Cancer Inst. 2013;105:701–10.

217. Rakovitch E, Nofech-Mozes S, Hanna W, Baehner FL, Saskin R, Butler SM, et al. A population-based validation study of the DCIS Score predicting recurrence risk in individuals treated by breast-conserving surgery alone. Breast Cancer Res Treat. 2015;152:389–98.

218. Dowsett M, Sestak I, Lopez-Knowles E, Sidhu K, Dunbier AK, Cowens JW, et al. Comparison of PAM50 risk of recurrence score with oncotype DX and IHC4 for predicting risk of distant recurrence after endocrine therapy. J Clin Oncol. 2013;31:2783–90.
219. Filipits M, Rudas M, Jakesz R, Dubsky P, Fitzal F, Singer CF, et al. A new molecular predictor of distant recurrence in ER-positive, HER2-negative breast cancer adds independent information to conventional clinical risk factors. Clin Cancer Res. 2011;17:6012–20.
220. Dubsky P, Filipits M, Jakesz R, Rudas M, Rudas M, Greil R, et al. Endopredict improves the prognostic classification derived from common clinical guidelines in ER-positive, HER2-negative early breast cancer. Ann Oncol. 2013;24:640–7.
221. Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, et al. Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst. 2006;98:262–72.
222. Xiao-Jun M, Salunga R, Dahiya S, Wang W, Carney E, Durbecq V, et al. A five-gene molecular grade index and HOXB13.IL17BR are complementary prognostic factors in early stage breast cancer. Clin Cancer Res. 2008;14:2601–8.
223. Jerevall PL, Ma XJ, Li H, Salunga R, Kesty NC, Erlander MG, et al. Prognostic utility of HOXB13: IL17BR and molecular grade index in early-stage breast cancer patients from the Stockholm trial. Br J Cancer. 2011;104:1762–9.
224. Zhang Y, Ph D, Schnabel CA, Schroeder B, Goss PE, Full MD, et al. Prediction of late distant recurrence in estrogen receptor positive breast cancer patients: prospective comparison of the Breast Cancer Index (BCI), Oncotype DX recurrence score, and IHC4 in TransATAC. Lancet Oncol. 2014;14:1067–76.
225. Bartlett JMS, Thomas J, Ross DT, Seitz RS, Ring BZ, Beck RA, et al. Mammostrat®as a tool to stratify breast cancer patients at risk of recurrence during endocrine therapy. Breast Cancer Res. 2010;12:1–11.
226. Cuzick J, Dowsett M, Pineda S, Wale C, Salter J, Quinn E, et al. Prognostic value of a combined estrogen receptor, progesterone receptor, Ki-67, and human epidermal growth factor receptor 2 immunohistochemical score and comparison with the genomic health recurrence score in early breast cancer. J Clin Oncol. 2011;29:4273–8.
227. Haibe-Kains B, Desmedt C, Piette F, Buyse M, Cardoso F, van't Veer L, et al. Comparison

of prognostic gene expression signatures for breast cancer. BMC Genomics. 2008;9:1–9.

228. Yu JX, Sieuwerts AM, Zhang Y, Martens JWM, Smid M, Klijn JGM, et al. Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer. BMC Cancer. 2007;7:182.

229. Venet D, Dumont JE, Detours V. Most random gene expression signatures are significantly associated with breast cancer outcome. PLoS Comput Biol. 2011;7.

230. Iwamoto T, Pusztai L. Predicting prognosis of breast cancer with gene signatures: Are we lost in a sea of data? Genome Med. 2010;2:2–5.

231. Ioannidis JPA, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, et al. Repeatability of published microarray gene expression analyses. Nat Genet. 2009;41:149.

232. van Vliet MH, Fabien R, Horlings HM, van de Vijver MJ, Reinders MJT, Wessels LFA. Pooling breast cancer datasets has a synergetic effect on classification performance and improves signature stability. BMC Genomics. 2008;9:1–22.

233. Thomassen M, Tan Q, Eiriksdottir F, Bak M, Cold S, Kruse TA. Comparison of gene sets for expression profiling: Prediction of metastasis from low-malignant breast cancer. Clin Cancer Res. 2007;13:5355–60.

234. Reyal F, van Vliet MH, Armstrong NJ, Horlings HM, de Visser KE, Kok M, et al. A comprehensive analysis of prognostic signatures reveals the high predictive capacity of the Proliferation, Immune response and RNA splicing modules in breast cancer. Breast Cancer Res. 2008;10:1–15.

235. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature. 2012;483:570–5.

236. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell. 2016;166:740–54.

237. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483:603–7.

238. Ghandi M, Huang FW, Jané-Valbuena J, Kryukov G V, Lo CC, McDonald ER, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature. 2019;569:503.

239. Haibe-Kains B, El-Hachem N, Birkbak NJ, Jin AC, Beck AH, Aerts HJWLWL, et al.

Inconsistency in large pharmacogenomic studies. Nature. 2013;504:389–93.

240. Safikhani Z, Smirnov P, Freeman M, El-Hachem N, She A, Rene Q, et al. Revisiting inconsistency in large pharmacogenomic studies. F1000Res. 2016;5:2333.
241. Geeleher P, Gamazon ER, Seoighe C, Cox NJ, Huang RS. Consistency in large pharmacogenomic studies. Nature. 2016;540:E1–2.
242. Mpindi JP, Yadav B, Östling P, Gautam P, Malani D, Murumägi A, et al. Consistency in drug response profiling. Nature. 2016;540:E5–6.
243. Bouhaddou M, DiStefano MS, Riesel EA, Carrasco E, Holzapfel HY, Jones DC, et al. Drug response consistency in CCLE and CGP. Nature. 2016;540:E9–10.
244. Stransky N, Ghandi M, Kryukov G V., Garraway LA, Lehár J, Liu M, et al. Pharmacogenomic agreement between two cancer cell line data sets. Nature. 2015;528:84–7.
245. Smirnov P, Kofia V, Maru A, Freeman M, Ho C, El-Hachem N, et al. PharmacoDB: An integrative database for mining in vitro anticancer drug screening studies. Nucleic Acids Res. 2018;46:D994–1002.
246. Smirnov P, Safikhani Z, El-Hachem N, Wang D, She A, Olsen C, et al. PharmacoGx: An R package for analysis of large pharmacogenomic datasets. Bioinformatics. 2016;32:1244–6.
247. Luna A, Rajapakse VN, Sousa FG, Gao J, Schultz N, Varma S, et al. rcellminer: exploring molecular profiles and drug response of the NCI-60 cell lines in R. Bioinformatics. 2016;32:1272–4.
248. Azuaje F. Computational models for predicting drug responses in cancer research. Br Bioinform. 2017;18:820–9.
249. Geeleher P, Cox NJ, Huang RS. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. Genome Biol. 2014;15:R47.
250. Ding Z, Zu S, Gu J. Evaluating the molecule-based prediction of clinical drug responses in cancer. Bioinformatics. 2016;32:2891–5.
251. Graim K, Friedl V, Houlahan KE, Stuart JM. PLATYPUS: A Multiple-View Learning Predictive Framework for Cancer Drug Sensitivity Prediction. PSB, World Scientific; 2019, p. 136–47.
252. Geeleher P, Zhang Z, Wang F, Gruener RF, Nath A, Morrison G, et al. Discovering novel pharmacogenomic biomarkers by imputing drug response in cancer patients from large

genomics studies. Genome Res. 2017;27:1743–51.

253. Dong Z, Zhang N, Li C, Wang H, Fang Y, Wang J, et al. Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. BMC Cancer. 2015;15:489.

254. Bayer I, Groth P, Schneckener S. Prediction errors in learning drug response from gene expression data–influence of labeling, sample size, and machine learning algorithm. PLoS One. 2013;8:e70294.

255. Zhang N, Wang H, Fang Y, Wang J, Zheng X, Liu XS. Predicting Anticancer Drug Responses Using a Dual-Layer Integrated Cell Line-Drug Network Model. PLoS Comput Biol. 2015;11:e1004498.

256. Ding MQ, Chen L, Cooper GF, Young JD, Lu X. Precision Oncology beyond Targeted Therapy: Combining Omics Data with Machine Learning Matches the Majority of Cancer Cells to Effective Therapeutics. Mol Cancer Res. 2018;16:269–78.

257. Sharifi-Noghabi H, Zolotareva O, Collins CC, Ester M. MOLI: Multi-Omics Late Integration with deep neural networks for drug response prediction. BioRxiv. 2019:531327.

258. Menden MP, Iorio F, Garnett M, McDermott U, Benes CH, Ballester PJ, et al. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. PLoS One. 2013;8:e61318.

259. Ammad-Ud-Din M, Khan SA, Malani D, Murumagi A, Kallioniemi O, Aittokallio T, et al. Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization. Bioinformatics. 2016;32:i455–63.

260. Daemen A, Griffith OL, Heiser LM, Wang NJ, Enache OM, Sanborn Z, et al. Modeling precision treatment of breast cancer. Genome Biol. 2013;14:R110.

261. Costello JC, Heiser LM, Georgii E, Gonen M, Menden MP, Wang NJ, et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nat Biotechnol. 2014;32:1202–12.

262. Caroli J, Dori M, Bicciato S. Computational Methods for the Integrative Analysis of Genomics and Pharmacological Data. Front Oncol. 2020;10.

263. Hidalgo M, Amant F, Biankin A V, Budinská E, Byrne AT, Caldas C, et al. Patient-derived xenograft models: an emerging platform for translational cancer research. Cancer Discov. 2014;4:998–1013.

264. Meehan TF, Conte N, Goldstein T, Inghirami G, Murakami MA, Brabetz S, et al. PDX-MI: minimal information for patient-derived tumor xenograft models. Cancer Res. 2017;77:e62–6.
265. Conte N, Mason JC, Halmagyi C, Neuhauser S, Mosaku A, Yordanova G, et al. PDX finder: a portal for patient-derived tumor xenograft model discovery. Nucleic Acids Res. 2019;47:D1073–9.
266. Drost J, Clevers H. Organoids in cancer research. Nat Rev Cancer. 2018;18:407–18.
267. Mer AS, Ba-Alawi W, Smirnov P, Wang YX, Brew B, Ortmann J, et al. Integrative pharmacogenomics analysis of patient-derived xenografts. Cancer Res. 2019;79:4539–50.
268. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25:25.
269. Consortium GO. The gene ontology resource: 20 years and still GOing strong. Nucleic Acids Res. 2018;47:D330–8.
270. Braschi B, Denny P, Gray K, Jones T, Seal R, Tweedie S, et al. Genenames. org: the HGNC and VGNC resources in 2019. Nucleic Acids Res. 2018;47:D786–92.
271. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2012;41:D991–5.
272. Athar A, Füllgrabe A, George N, Iqbal H, Huerta L, Ali A, et al. ArrayExpress update–from bulk to single-cell expression data. Nucleic Acids Res. 2018;47:D711–5.
273. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat Genet. 2001;29:365.
274. Consortium ICG. International network of cancer genome projects. Nature. 2010;464:993.
275. Basu A, Bodycombe NE, Cheah JH, Price E V, Liu K, Schaefer GI, et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. Cell. 2013;154:1151–61.
276. Rees MG, Seashore-Ludlow B, Cheah JH, Adams DJ, Price E V, Gill S, et al. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. Nat Chem Biol. 2016;12:109.
277. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease.

Science (80-). 2006;313:1929–35.

278. Musa A, Ghoraie LS, Zhang SD, Glazko G, Yli-Harja O, Dehmer M, et al. A review of connectivity map and computational approaches in pharmacogenomics. Br Bioinform. 2018;19:506–23.

279. Cava C, Bertoli G, Castiglioni I. Integrating genetics and epigenetics in breast cancer: biological insights, experimental, computational methods and therapeutic potential. BMC Syst Biol. 2015;9:62.

280. Chi C, Murphy LC, Hu P. Recurrent copy number alterations in young women with breast cancer. Oncotarget. 2018;9:11541.

281. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: Function, expression and evolution. Nat Rev Genet. 2009;10:252–63.

282. Xu T, Le TD, Liu L, Wang R, Sun B, Li J. Identifying Cancer Subtypes from miRNA-TF-mRNA Regulatory Networks and Expression Data. PLoS One. 2016;11:e0152792.

283. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011;12:R41.

284. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan K-K, Cheng C, et al. Architecture of the human regulatory network derived from ENCODE data. Nature. 2012;489:91.

285. Zhu LJ, Gazin C, Lawson ND, Pagès H, Lin SM, Lapointe DS, et al. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics. 2010;11:237.

286. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47–e47.

287. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B. 1996:267–88.

288. Schröder MS, Culhane AC, Quackenbush J, Haibe-Kains B. survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. Bioinformatics. 2011;27:3206–8.

289. Lin DY, Wei L-J. The robust inference for the Cox proportional hazards model. J Am Stat Assoc. 1989;84:1074–8.

290. Liu J, Shi W, Wu C, Ju J, Jiang J. miR-181b as a key regulator of the oncogenic process and its clinical implications in cancer. Biomed Reports. 2014;2:7–11.
291. Wu X, Somlo G, Yu Y, Palomares MR, Li AX, Zhou W, et al. De novo sequencing of circulating miRNAs identifies novel markers predicting clinical outcome of locally advanced breast cancer. J Transl Med. 2012;10:42.
292. Zawistowski JS, Nakamura K, Parker JS, Granger DA, Golitz BT, Johnson GL. MicroRNA 9-3p targets β1 integrin to sensitize claudin-low breast cancer cells to MEK inhibition. Mol Cell Biol. 2013;33:2260–74.
293. Denechaud P-D, Fajas L, Giralt A. E2F1, a novel regulator of metabolism. Front Endocrinol (Lausanne). 2017;8:311.
294. Tien ES, Davis JW, Heuvel JP Vanden. Identification of the CREB-binding protein/p300-interacting protein CITED2 as a peroxisome proliferator-activated receptor α coregulator. J Biol Chem. 2004;279:24053–63.
295. Lau WM, Doucet M, Huang D, Weber KL, Kominsky SL. CITED2 modulates estrogen receptor transcriptional activity in breast cancer cells. Biochem Biophys Res Commun. 2013;437:261–6.
296. Minemura H, Takagi K, Sato A, Takahashi H, Miki Y, Shibahara Y, et al. CITED 2 in breast carcinoma as a potent prognostic predictor associated with proliferation, migration and chemoresistance. Cancer Sci. 2016;107:1898–908.
297. Unterbruner K, Matthes F, Schilling J, Nalavade R, Weber S, Winter J, et al. MicroRNAs miR-19, miR-340, miR-374 and miR-542 regulate MID1 protein expression. PLoS One. 2018;13:e0190437.
298. Chang JTH, Wang F, Chapin W, Huang RS. Identification of MicroRNAs as breast cancer prognosis markers through the cancer genome atlas. PLoS One. 2016;11:e0168284.
299. Lee K-H, Lin F-C, Hsu T-I, Lin J-T, Guo J-H, Tsai C-H, et al. MicroRNA-296-5p (miR-296-5p) functions as a tumor suppressor in prostate cancer by directly targeting Pin1. Biochim Biophys Acta (BBA)-Molecular Cell Res. 2014;1843:2055–66.
300. Tachibana K, Yamasaki D, Ishimoto K. The role of PPARs in cancer. PPAR Res. 2008;2008.
301. Peters JM, Shah YM, Gonzalez FJ. The role of peroxisome proliferator-activated receptors in carcinogenesis and chemoprevention. Nat Rev Cancer. 2012;12:181.
302. Lianggeng X, Baiwu L, Maoshu B, Jiming L, Youshan L. Impact of Interaction Between

PPAR Alpha and PPAR Gamma on Breast Cancer Risk in the Chinese Han Population. Clin Breast Cancer. 2017;17:336–40.

303. Bredel M, Kim H, Nanda TK, Scholtens DM, Robe PA, Branimir SI, et al. Deletion of the tumor suppressor NFKBIA in triple-negative breast cancer. Int J Radiat Oncol Biol Phys. 2013;87:S98.
304. Zhong S, Chen X, Wang D, Zhang X, Shen H, Yang S, et al. MicroRNA expression profiles of drug-resistance breast cancer cells and their exosomes. Oncotarget. 2016;7:19601.
305. Nilsson S, Möller C, Jirström K, Lee A, Busch S, Lamb R, et al. Downregulation of miR-92a is associated with aggressive breast cancer features and increased tumour macrophage infiltration. PLoS One. 2012;7:e36051.
306. Luan QX, Zhang BG, Li XJ, Guo MY. MiR-129-5p is downregulated in breast cancer cells partly due to promoter H3K27m3 modification and regulates epithelial-mesenchymal transition and multi-drug resistance. Eur Rev Med Pharmacol Sci. 2016;20:4257–65.
307. Meng R, Fang J, Yu Y, Hou LK, Chi JR, Chen AX, et al. miR-129-5p suppresses breast cancer proliferation by targeting CBX4. Neoplasma. 2018;65:572.
308. Meng Q, Xiang L, Fu J, Chu X, Wang C, Yan B. Transcriptome profiling reveals miR-9-3p as a novel tumor suppressor in gastric cancer. Oncotarget. 2017;8:37321.
309. Barbano R, Pasculli B, Rendina M, Fontana A, Fusilli C, Copetti M, et al. Stepwise analysis of MIR9 loci identifies miR-9-5p to be involved in Oestrogen regulated pathways in breast cancer patients. Sci Rep. 2017;7:45283.
310. Krist B, Florczyk U, Pietraszek-Gremplewicz K, Józkowicz A, Dulak J. The role of miR-378a in metabolism, angiogenesis, and muscle biology. Int J Endocrinol. 2015;2015.
311. Monteleone NJ, Lutz CS. miR-708-5p: a microRNA with emerging roles in cancer. Oncotarget. 2017;8:71292.
312. Cai J, Guan H, Fang L, Yang Y, Zhu X, Yuan J, et al. MicroRNA-374a activates Wnt/β-catenin signaling to promote breast cancer metastasis. J Clin Invest. 2013;123.
313. Wang Z, Wei W, Sarkar FH. miR-23a, a critical regulator of “migR” ation and metastasis in colorectal cancer. Cancer Discov. 2012;2:489–91.
314. Fackler MJ, McVeigh M, Mehrotra J, Blum MA, Lange J, Lapides A, et al. Quantitative multiplex methylation-specific PCR assay for the detection of promoter hypermethylation in multiple genes in breast cancer. Cancer Res. 2004;64:4442–52.

315. Nimmrich I, Sieuwerts AM, Meijer-van Gelder ME, Schwope I, Bolt-de Vries J, Harbeck N, et al. DNA hypermethylation of PITX2 is a marker of poor prognosis in untreated lymph node-negative hormone receptor-positive breast cancer patients. Breast Cancer Res Treat. 2008;111:429–37.
316. Hamurcu Z, Ashour A, Kahraman N, Ozpolat B. FOXM1 regulates expression of eukaryotic elongation factor 2 kinase and promotes proliferation, invasion and tumorgenesis of human triple negative breast cancer cells. Oncotarget. 2016;7:16619.
317. O'Regan RM, Nahta R. Targeting forkhead box M1 transcription factor in breast cancer. Biochem Pharmacol. 2018.
318. Yang C, Chen H, Tan G, Gao W, Cheng L, Jiang X, et al. FOXM1 promotes the epithelial to mesenchymal transition by stimulating the transcription of Slug in human breast cancer. Cancer Lett. 2013;340:104–12.
319. Xue J, Lin X, Chiu W-T, Chen Y-H, Yu G, Liu M, et al. Sustained activation of SMAD3/SMAD4 by FOXM1 promotes TGF-β–dependent cancer metastasis. J Clin Invest. 2014;124:564–79.
320. Hamurcu Z, Delibaşı N, Nalbantoglu U, Sener EF, Nurdinov N, Tascı B, et al. FOXM1 plays a role in autophagy by transcriptionally regulating Beclin-1 and LC3 genes in human triple-negative breast cancer cells. J Mol Med. 2019:97:491–508.
321. Ring A, Nguyen C, Smbatyan G, Tripathy D, Yu M, Press M, et al. CBP/B-catenin/FOXM1 is a novel therapeutic target in triple negative breast cancer. Cancers. 2018;10:525.
322. Tan Y, Wang Q, Xie Y, Qiao X, Zhang S, Wang Y, et al. Identification of FOXM1 as a specific marker for triple-negative breast cancer. Int J Oncol. 2019;54:87–97.
323. Zhang L, Du Y, Xu S, Jiang Y, Yuan C, Zhou L, et al. DEPDC1, negatively regulated by miR-26b, facilitates cell proliferation via the up-regulation of FOXM1 expression in TNBC. Cancer Lett. 2019;442:242–51.
324. Grabacka M, Plonka PM, Urbanska K, Reiss K. Peroxisome Proliferator–Activated Receptor α Activation Decreases Metastatic Potential of Melanoma Cells In vitro via Down-Regulation of Akt. Clin Cancer Res. 2006;12:3028–36.
325. Liu D-C, Zang C-B, Liu H-Y, Possinger K, Fan S-G, Elstner E. A novel PPAR alpha/gamma dual agonist inhibits cell growth and induces apoptosis in human glioblastoma T98G cells. Acta Pharmacol Sin. 2004;25:1312–9.

326. Gao J, Liu Q, Xu Y, Gong X, Zhang R, Zhou C, et al. PPARα induces cell apoptosis by destructing Bcl2. Oncotarget. 2015;6:44635.
327. Shi Z, Derow CK, Zhang B. Co-expression module analysis reveals biological processes, genomic gain, and regulatory mechanisms associated with breast cancer progression. BMC Syst Biol. 2010;4:74.
328. Saleh R, Taha RZ, Nair VS, Alajez NM, Elkord E. PD-L1 blockade by atezolizumab downregulates signaling pathways associated with tumor growth, metastasis, and hypoxia in human triple negative breast cancer. Cancers (Basel). 2019;11:1–17.
329. Blücher C, Iberl S, Schwagarus N, Müller S, Liebisch G, Höring M, et al. Secreted Factors from Adipose Tissue Reprogram Tumor Lipid Metabolism and Induce Motility by Modulating PPARα/ANGPTL4 and FAK. Mol Cancer Res. 2020;18:1849–62.
330. Lamb J. The Connectivity Map : a new tool for biomedical research. Nature. 2007;7:54–60.
331. Qu XA, Rajpal DK. Applications of Connectivity Map in drug discovery and development. Drug Discov Today. 2012;17:1289–98.
332. Jiao Y, Lawler K, Patel GS, Purushotham A, Jones AF, Grigoriadis A, et al. DART: Denoising Algorithm based on Relevance network Topology improves molecular pathway activity inference. BMC Bioinformatics. 2011;12:403.
333. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20:307–15.
334. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28:882–3.
335. Pachter L. Models for transcript quantification from RNA-Seq. ArXiv Prepr ArXiv11043889. 2011.
336. Coombs GS, Schmitt AA, Canning CA, Alok A, Low ICC, Banerjee N, et al. Modulation of Wnt/β-catenin signaling and proliferation by a ferrous iron chelator with therapeutic efficacy in genetically engineered mouse models of cancer. Oncogene. 2012;31:213.
337. Rahimi HR, Hasanli E, Jamalifar H. A mini review on new pharmacological and toxicological considerations of protease inhibitors' application in cancer prevention and biological research. Asian J Cell Biol. 2012;7:1–12.
338. Huang H, Chen D, Li S, Li X, Liu N, Lu X, et al. Gambogic acid enhances proteasome

inhibitor-induced anticancer activity. Cancer Lett. 2011;301:221–8.

339. Kim Y, Kang H, Jang SW, Ko J. Celastrol inhibits breast cancer cell invasion via suppression of NF-κB-mediated matrix metalloproteinase-9 expression. Cell Physiol Biochem Int J Exp Cell Physiol Biochem Pharmacol. 2011;28:175–84.
340. Shrivastava S, Jeengar MK, Reddy VS, Reddy GB, Naidu VGM. Anticancer effect of celastrol on human triple negative breast cancer: possible involvement of oxidative stress, mitochondrial dysfunction, apoptosis and PI3K/Akt pathways. Exp Mol Pathol. 2015;98:313–27.
341. Venturelli S, Berger A, Böcker A, Busch C, Weiland T, Noor S, et al. Resveratrol as a pan-HDAC inhibitor alters the acetylation status of jistone proteins in human-derived hepatoblastoma cells. PLoS One. 2013;8:e73097.
342. Chatterjee M, Das S, Janarthan M, Ramachandran HK, Chatterjee M. Role of 5-lipoxygenase in resveratrol mediated suppression of 7, 12-dimethylbenz (α) anthracene-induced mammary carcinogenesis in rats. Eur J Pharmacol. 2011;668:99–106.
343. Singh CK, Ndiaye MA, Ahmad N. Resveratrol and cancer: Challenges for clinical translation. Biochim Biophys Acta (BBA)-Molecular Basis Dis. 2015;1852:1178–85.
344. Zhou H, Shen T, Shang C, Luo Y, Liu L, Yan J, et al. Ciclopirox induces autophagy through reactive oxygen species-mediated activation of JNK signaling pathway. Oncotarget. 2014;5:10140.
345. S Akinboye E, Bakare O. Biological activities of emetine. Open Nat Prod J. 2011;4.
346. Cao C, Yu H, Wu F, Qi H, He J. Antibiotic anisomycin induces cell cycle arrest and apoptosis through inhibiting mitochondrial biogenesis in osteosarcoma. J Bioenerg Biomembr. 2017;49:437–43.
347. Vogel VG. The NSABP study of tamoxifen and raloxifene (STAR) trial. Expert Rev Anticancer Ther. 2009;9:51–60.
348. Dickler MN, Norton L. The MORE Trial: Multiple Outcomes for Raloxifene Evaluation: Breast Cancer as a Secondary End Point: Implications for Prevention. Ann N Y Acad Sci. 2001;949:134–42.
349. Furtado CM, Marcondes MC, Sola-Penna M, De Souza MLS, Zancan P. Clotrimazole preferentially inhibits human breast cancer cell proliferation, viability and glycolysis. PLoS One. 2012;7:e30462.

350. Chen C-L, Chen T-C, Lee C-C, Shih L-C, Lin C-Y, Hsieh Y-Y, et al. Synthesis and evaluation of new 3-substituted-4-chloro-thioxanthone derivatives as potent anti-breast cancer agents. Arab J Chem. 2015.
351. Muniraj N, Siddharth S, Nagalingam A, Walker A, Woo J, Győrffy B, et al. Withaferin A inhibits lysosomal activity to block autophagic flux and induces apoptosis via energetic impairment in breast cancer cells. Carcinogenesis. 2019.
352. Wei D, Liu C, Zheng X, Li Y. Comprehensive anticancer drug response prediction based on a simple cell line-drug complex network model. BMC Bioinformatics. 2019;20:44.
353. Greenman CD, Bignell G, Butler A, Edkins S, Hinton J, Beare D, et al. PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. Biostatistics. 2009;11:164–75.
354. Nguyen L, Dang CC, Ballester PJ. Systematic assessment of multi-gene predictors of pan-cancer cell line sensitivity to drugs exploiting gene expression data. F1000Res. 2016;5.
355. Bolton EE, Wang Y, Thiessen PA, Bryant SH. PubChem: integrated platform of small molecules and biological activities. Annu. Rep. Comput. Chem., vol. 4, Elsevier; 2008, p. 217–41.
356. Yang J, Li A, Li Y, Guo X, Wang M. A novel approach for drug response prediction in cancer cell lines via network representation learning. Bioinformatics. 2018.
357. Wang X, Sun Z, Zimmermann MT, Bugrim A, Kocher JP. Predict drug sensitivity of cancer cells with pathway activity inference. BMC Med Genomics. 2019;12:15.
358. García-Campos MA, Espinal-Enríquez J, Hernández-Lemus E. Pathway analysis: state of the art. Front Physiol. 2015;6:383.
359. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014;11:333–7.
360. Yap CW. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011;32:1466–74.
361. Wang L, Li X, Zhang L, Gao Q. Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. BMC Cancer. 2017;17:513.
362. Shi W, Jiang T, Nuciforo P, Hatzis C, Holmes E, Harbeck N, et al. Pathway level alterations rather than mutations in single genes predict response to HER2-targeted therapies in the neo-ALTTO trial. Ann Oncol. 2017;28:128–35.

Appendices

Appendix A: GSEA identified upregulated pathways in TNBC versus normal breast tissue

This table lists the 137 significantly upregulated pathways in TNBC samples relative to normal breast tissue identified through GSEA with a FDR threshold of 5%.

ID	Gene Set Name	Size [a]	ES [b]	NES [c]	FDR [d]
1	**PID_FOXM1_PATHWAY**	**40**	**-0.732**	**-2.274**	**0.005**
2	REACTOME_CELL_CYCLE_MITOTIC	298	-0.674	-2.167	0.005
3	PID_PLK1_PATHWAY	45	-0.724	-2.176	0.006
4	REACTOME_KINESINS	24	-0.815	-2.177	0.007
5	REACTOME_DNA_REPLICATION	182	-0.749	-2.122	0.008
6	KEGG_CELL_CYCLE	118	-0.648	-2.217	0.008
7	REACTOME_CELL_CYCLE	385	-0.650	-2.189	0.009
8	REACTOME_MITOTIC_M_M_G1_PHASES	162	-0.742	-2.103	0.009
9	REACTOME_MITOTIC_PROMETAPHASE	86	-0.731	-2.079	0.010
10	REACTOME_CELL_CYCLE_CHECKPOINTS	105	-0.732	-2.084	0.010
11	BIOCARTA_G2_PATHWAY	24	-0.761	-2.053	0.011
12	SA_G1_AND_S_PHASES	15	-0.797	-2.059	0.012
13	PID_AURORA_B_PATHWAY	39	-0.698	-2.037	0.013
14	REACTOME_S_PHASE	100	-0.728	-2.032	0.013
15	REACTOME_SYNTHESIS_OF_DNA	84	-0.765	-2.011	0.014
16	REACTOME_MRNA_SPLICING_MINOR_PATHWAY	38	-0.765	-2.013	0.015
17	REACTOME_CHROMOSOME_MAINTENANCE	114	-0.664	-1.998	0.015
18	REACTOME_G2_M_CHECKPOINTS	35	-0.807	-1.971	0.016
19	REACTOME_MITOTIC_G1_G1_S_PHASES	124	-0.676	-1.998	0.016
20	PID_ATR_PATHWAY	39	-0.726	-2.002	0.016
21	REACTOME_MRNA_PROCESSING	146	-0.641	-1.972	0.017
22	REACTOME_FORMATION_OF_TUBULIN_FOLDING_INTERMEDIATES_BY_CCT_TRIC	19	-0.759	-1.933	0.017
23	REACTOME_REGULATION_OF_MRNA_STABILITY_BY_PROTEINS_THAT_BIND_AU_RICH_ELEMENTS	81	-0.639	-1.974	0.017
24	REACTOME_APC_C_CDH1_MEDIATED_DEGRADATION_OF_CDC20_AND_OTHER_APC_C_CDH1_TARGETED_PROTEINS_IN_LATE_MITOSIS_EARLY_G1	64	-0.724	-1.934	0.017
25	REACTOME_PREFOLDIN_MEDIATED_TRANSFER_OF_SUBSTRATE_TO_CCT_TRIC	25	-0.732	-1.948	0.017
26	PID_AURORA_A_PATHWAY	31	-0.602	-1.935	0.017
27	REACTOME_G1_S_TRANSITION	100	-0.694	-1.965	0.017
28	REACTOME_HIV_INFECTION	193	-0.578	-1.937	0.017
29	REACTOME_P53_DEPENDENT_G1_DNA_DAMAGE_RESPONSE	53	-0.725	-1.940	0.017
30	ST_FAS_SIGNALING_PATHWAY	64	-0.516	-1.938	0.017
31	KEGG_PATHOGENIC_ESCHERICHIA_COLI_INFECTION	53	-0.575	-1.975	0.017
32	REACTOME_HOST_INTERACTIONS_OF_HIV_FACTORS	120	-0.629	-1.950	0.017
33	KEGG_PYRIMIDINE_METABOLISM	98	-0.566	-1.959	0.017

Appendix A: GSEA identified upregulated pathways in TNBC versus normal breast tissue (cont.)

ID	Gene Set Name	Size [a]	ES [b]	NES [c]	FDR [d]
34	REACTOME_M_G1_TRANSITION	72	-0.773	-1.986	0.017
35	KEGG_SPLICEOSOME	114	-0.653	-1.952	0.017
36	REACTOME_TRANSCRIPTION_COUPLED_NER_TC_NER	44	-0.628	-1.943	0.018
37	REACTOME_ASSEMBLY_OF_THE_PRE_REPLICATIVE_COMPLEX	57	-0.777	-1.954	0.018
38	REACTOME_MRNA_SPLICING	97	-0.688	-1.981	0.018
39	REACTOME_ORC1_REMOVAL_FROM_CHROMATIN	59	-0.743	-1.940	0.018
40	REACTOME_TRANSCRIPTION	191	-0.584	-1.976	0.018
41	REACTOME_PROCESSING_OF_CAPPED_INTRON_CONTAINING_PRE_MRNA	126	-0.665	-1.955	0.018
42	REACTOME_RNA_POL_II_TRANSCRIPTION	93	-0.592	-1.920	0.018
43	REACTOME_ACTIVATION_OF_ATR_IN_RESPONSE_TO_REPLICATION_STRESS	29	-0.815	-1.911	0.018
44	BIOCARTA_CELLCYCLE_PATHWAY	23	-0.670	-1.920	0.019
45	REACTOME_TRNA_AMINOACYLATION	42	-0.701	-1.921	0.019
46	KEGG_DNA_REPLICATION	36	-0.794	-1.911	0.019
47	REACTOME_APC_C_CDC20_MEDIATED_DEGRADATION_OF_MITOTIC_PROTEINS	65	-0.719	-1.915	0.019
48	REACTOME_CLEAVAGE_OF_GROWING_TRANSCRIPT_IN_THE_TERMINATION_REGION_	34	-0.722	-1.913	0.019
49	REACTOME_REGULATION_OF_MITOTIC_CELL_CYCLE	77	-0.685	-1.902	0.019
50	REACTOME_DNA_REPAIR	105	-0.558	-1.905	0.019
51	REACTOME_TELOMERE_MAINTENANCE	73	-0.693	-1.903	0.020
52	REACTOME_MITOTIC_G2_G2_M_PHASES	74	-0.544	-1.894	0.021
53	REACTOME_P53_INDEPENDENT_G1_S_DNA_DAMAGE_CHECKPOINT	48	-0.741	-1.884	0.021
54	REACTOME_DESTABILIZATION_OF_MRNA_BY_AUF1_HNRNP_D0	50	-0.743	-1.885	0.022
55	REACTOME_ER_PHAGOSOME_PATHWAY	58	-0.744	-1.886	0.022
56	REACTOME_CDT1_ASSOCIATION_WITH_THE_CDC6_ORC_ORIGIN_COMPLEX	48	-0.754	-1.879	0.022
57	REACTOME_DNA_STRAND_ELONGATION	30	-0.847	-1.887	0.022
58	REACTOME_MRNA_3_END_PROCESSING	25	-0.711	-1.880	0.022
59	REACTOME_DEPOSITION_OF_NEW_CENPA_CONTAINING_NUCLEOSOMES_AT_THE_CENTROMERE	58	-0.734	-1.871	0.023
60	REACTOME_REGULATION_OF_ORNITHINE_DECARBOXYLASE_ODC	48	-0.742	-1.872	0.023
61	REACTOME_VIF_MEDIATED_DEGRADATION_OF_APOBEC3G	49	-0.736	-1.867	0.023
62	REACTOME_AUTODEGRADATION_OF_CDH1_BY_CDH1_APC_C	56	-0.698	-1.872	0.023
63	REACTOME_AUTODEGRADATION_OF_THE_E3_UBIQUITIN_LIGASE_COP1	47	-0.724	-1.873	0.023
64	KEGG_HOMOLOGOUS_RECOMBINATION	26	-0.698	-1.863	0.023
65	REACTOME_REGULATION_OF_APOPTOSIS	56	-0.683	-1.860	0.024
66	REACTOME_TRANSPORT_OF_MATURE_TRANSCRIPT_TO_CYTOPLASM	44	-0.702	-1.861	0.024
67	PID_FANCONI_PATHWAY	46	-0.666	-1.855	0.024
68	REACTOME_HIV_LIFE_CYCLE	114	-0.561	-1.854	0.024
69	REACTOME_EXTENSION_OF_TELOMERES	27	-0.808	-1.851	0.024
70	KEGG_P53_SIGNALING_PATHWAY	67	-0.466	-1.853	0.024

Appendix A: GSEA identified upregulated pathways in TNBC versus normal breast tissue (cont.)

ID	Gene Set Name	Size [a]	ES [b]	NES [c]	FDR [d]
71	REACTOME_CDK_MEDIATED_PHOSPHORYLATION_AND_REMOVAL_OF_CDC6	46	-0.748	-1.855	0.025
72	REACTOME_LAGGING_STRAND_SYNTHESIS	19	-0.818	-1.844	0.025
73	REACTOME_PHOSPHORYLATION_OF_THE_APC_C	17	-0.694	-1.842	0.025
74	REACTOME_FANCONI_ANEMIA_PATHWAY	21	-0.724	-1.844	0.025
75	REACTOME_FACTORS_INVOLVED_IN_MEGAKARYOCYTE_DEVELOPMENT_AND_PLATELET_PRODUCTION	125	-0.419	-1.845	0.025
76	KEGG_AMINOACYL_TRNA_BIOSYNTHESIS	41	-0.632	-1.840	0.025
77	REACTOME_SIGNALING_BY_WNT	63	-0.612	-1.840	0.026
78	REACTOME_CROSS_PRESENTATION_OF_SOLUBLE_EXOGENOUS_ANTIGENS_ENDOSOMES	47	-0.729	-1.832	0.027
79	REACTOME_CYCLIN_E_ASSOCIATED_EVENTS_DURING_G1_S_TRANSITION_	62	-0.657	-1.831	0.027
80	REACTOME_BIOSYNTHESIS_OF_THE_N_GLYCAN_PRECURSOR_DOLICHOL_LIPID_LINKED_OLIGOSACCHARIDE_LLO_AND_TRANSFER_TO_A_NASCENT_PROTEIN	28	-0.602	-1.817	0.027
81	REACTOME_APOPTOSIS	143	-0.488	-1.828	0.027
82	KEGG_MISMATCH_REPAIR	23	-0.688	-1.832	0.027
83	REACTOME_ANTIGEN_PROCESSING_CROSS_PRESENTATION	72	-0.669	-1.817	0.027
84	REACTOME_MEIOTIC_RECOMBINATION	79	-0.655	-1.820	0.027
85	REACTOME_SCFSKP2_MEDIATED_DEGRADATION_OF_P27_P21	53	-0.709	-1.821	0.027
86	KEGG_PROTEASOME	44	-0.754	-1.823	0.027
87	REACTOME_LATE_PHASE_OF_HIV_LIFE_CYCLE	101	-0.549	-1.822	0.027
88	PID_MYC_ACTIV_PATHWAY	79	-0.541	-1.824	0.027
89	REACTOME_CYCLIN_A_B1_ASSOCIATED_EVENTS_DURING_G2_M_TRANSITION	15	-0.782	-1.828	0.027
90	REACTOME_METABOLISM_OF_NUCLEOTIDES	72	-0.535	-1.817	0.027
91	REACTOME_INTRINSIC_PATHWAY_FOR_APOPTOSIS	29	-0.550	-1.818	0.028
92	REACTOME_G0_AND_EARLY_G1	23	-0.749	-1.824	0.028
93	KEGG_RNA_POLYMERASE	29	-0.631	-1.808	0.028
94	BIOCARTA_PROTEASOME_PATHWAY	28	-0.787	-1.808	0.028
95	REACTOME_SCF_BETA_TRCP_MEDIATED_DEGRADATION_OF_EMI1	49	-0.694	-1.809	0.029
96	REACTOME_PROTEIN_FOLDING	49	-0.517	-1.809	0.029
97	REACTOME_PROCESSING_OF_CAPPED_INTRONLESS_PRE_MRNA	23	-0.708	-1.804	0.029
98	REACTOME_MEIOSIS	107	-0.568	-1.799	0.030
99	REACTOME_APC_CDC20_MEDIATED_DEGRADATION_OF_NEK2A	21	-0.683	-1.799	0.030
100	REACTOME_DESTABILIZATION_OF_MRNA_BY_KSRP	17	-0.651	-1.797	0.030
101	BIOCARTA_P53_PATHWAY	16	-0.581	-1.792	0.031
102	REACTOME_PROCESSIVE_SYNTHESIS_ON_THE_LAGGING_STRAND	15	-0.798	-1.788	0.032
103	REACTOME_ACTIVATION_OF_THE_PRE_REPLICATIVE_COMPLEX	24	-0.839	-1.790	0.032
104	REACTOME_E2F_MEDIATED_REGULATION_OF_DNA_REPLICATION	27	-0.710	-1.789	0.032
105	REACTOME_ACTIVATION_OF_NF_KAPPAB_IN_B_CELLS	61	-0.612	-1.788	0.032
106	REACTOME_CYTOSOLIC_TRNA_AMINOACYLATION	24	-0.731	-1.786	0.032
107	REACTOME_RNA_POL_III_TRANSCRIPTION_INITIATION_FROM_TYPE_2_PROMOTER	23	-0.616	-1.780	0.034

Appendix A: GSEA identified upregulated pathways in TNBC versus normal breast tissue (cont.)

ID	Gene Set Name	Size [a]	ES [b]	NES [c]	FDR [d]
108	REACTOME_RNA_POL_I_RNA_POL_III_AND_MITOCHONDRIAL_TRANSCRIPTION	114	-0.555	-1.776	0.034
109	REACTOME_APC_C_CDC20_MEDIATED_DEGRADATION_OF_CYCLIN_B	19	-0.666	-1.776	0.035
110	REACTOME_RNA_POL_II_PRE_TRANSCRIPTION_EVENTS	59	-0.526	-1.767	0.037
111	PID_BARD1_PATHWAY	29	-0.676	-1.764	0.038
112	KEGG_BASE_EXCISION_REPAIR	33	-0.617	-1.762	0.038
113	REACTOME_NUCLEOTIDE_EXCISION_REPAIR	49	-0.531	-1.757	0.039
114	REACTOME_METABOLISM_OF_NON_CODING_RNA	47	-0.649	-1.751	0.041
115	REACTOME_SYNTHESIS_AND_INTERCONVERSION_OF_NUCLEOTIDE_DI_AND_TRIPHOSPHATES	19	-0.634	-1.750	0.041
116	REACTOME_DOWNSTREAM_SIGNALING_EVENTS_OF_B_CELL_RECEPTOR_BCR	92	-0.492	-1.748	0.042
117	REACTOME_INHIBITION_OF_THE_PROTEOLYTIC_ACTIVITY_OF_APC_C_REQUIRED_FOR_THE_ONSET_OF_ANAPHASE_BY_MITOTIC_SPINDLE_CHECKPOINT_COMPONENTS	18	-0.666	-1.749	0.042
118	REACTOME_ABORTIVE_ELONGATION_OF_HIV1_TRANSCRIPT_IN_THE_ABSENCE_OF_TAT	23	-0.680	-1.746	0.042
119	BIOCARTA_MPR_PATHWAY	34	-0.520	-1.744	0.042
120	REACTOME_MITOCHONDRIAL_TRNA_AMINOACYLATION	21	-0.659	-1.742	0.043
121	BIOCARTA_ACTINY_PATHWAY	20	-0.622	-1.739	0.044
122	KEGG_OOCYTE_MEIOSIS	112	-0.428	-1.732	0.046
123	REACTOME_ELONGATION_ARREST_AND_RECOVERY	31	-0.598	-1.730	0.046
124	REACTOME_GLUCOSE_TRANSPORT	38	-0.554	-1.724	0.047
125	REACTOME_MRNA_CAPPING	29	-0.566	-1.725	0.047
126	REACTOME_RNA_POL_I_TRANSCRIPTION	81	-0.603	-1.725	0.047
127	PID_ARF_3PATHWAY	19	-0.601	-1.722	0.047
128	REACTOME_POST_CHAPERONIN_TUBULIN_FOLDING_PATHWAY	16	-0.670	-1.722	0.047
129	REACTOME_ASPARAGINE_N_LINKED_GLYCOSYLATION	80	-0.493	-1.727	0.047
130	REACTOME_BASE_EXCISION_REPAIR	19	-0.708	-1.726	0.047
131	REACTOME_MICRORNA_MIRNA_BIOGENESIS	21	-0.611	-1.728	0.047
132	PID_E2F_PATHWAY	70	-0.483	-1.710	0.050
133	REACTOME_INTERACTIONS_OF_VPR_WITH_HOST_CELLULAR_PROTEINS	32	-0.652	-1.710	0.050
134	REACTOME_MEIOTIC_SYNAPSIS	69	-0.562	-1.710	0.050
135	REACTOME_RNA_POL_I_PROMOTER_OPENING	56	-0.694	-1.712	0.050
136	REACTOME_PACKAGING_OF_TELOMERE_ENDS	46	-0.667	-1.711	0.050
137	REACTOME_TRANSPORT_OF_MATURE_MRNA_DERIVED_FROM_AN_INTRONLESS_TRANSCRIPT	33	-0.666	-1.707	0.050

[a] Size is the number of genes included in a gene set.
[b] Enrichment Score (ES) reflects the degree to which a gene set is overrepresented at the bottom of the list of genes ranked according to the normal versus TNBC differential expression.
[c] Normalized enrichment score (NES) is based on the gene set enrichment scores for all dataset permutations and makes GSEA results comparable across gene sets.
[d] False Discovery Rate (FDR) is the estimated probability that a gene set with a given NES represents a false positive finding.

Appendix B: The GSEA results for PID_FOXM1_PATHWAY in TNBC versus normal breast tissue

This table shows the GSEA results for PID_FOXM1_PATHWAY, listing genes in the pathway ordered by their position in the list of genes ranked according to the normal versus TNBC differential expression.

ID	Gene Symbol [a]	Rank in Gene List [b]	Rank Metric Score [c]	Running ES [d]	Core Enrichment [e]
1	FOS	767	0.849	-0.011	No
2	SP1	945	0.799	0.006	No
3	ESR1	1056	0.766	0.025	No
4	RB1	1856	0.610	0.005	No
5	LAMA4	2167	0.558	0.008	No
6	GAS1	2421	0.520	0.012	No
7	CREBBP	3342	0.400	-0.020	No
8	NFATC3	3745	0.355	-0.029	No
9	EP300	4333	0.295	-0.048	No
10	MMP2	5284	0.212	-0.088	No
11	MYC	6487	0.122	-0.143	No
12	CCND1	7464	0.070	-0.188	No
13	MAP2K1	7637	0.061	-0.195	No
14	ETV5	7863	0.048	-0.204	No
15	HIST1H2BA	11300	-0.100	-0.369	No
16	TGFA	12811	-0.162	-0.438	No
17	ONECUT1	13419	-0.191	-0.461	No
18	XRCC1	16090	-0.356	-0.580	No
19	CENPB	16586	-0.395	-0.592	No
20	SKP2	19450	-0.766	-0.708	Yes
21	CDKN2A	19458	-0.768	-0.684	Yes
22	BRCA2	19565	-0.789	-0.664	Yes
23	CDK4	19572	-0.792	-0.639	Yes
24	CDK2	19664	-0.817	-0.618	Yes
25	GSK3A	19679	-0.821	-0.592	Yes
26	CDC25B	19911	-0.891	-0.575	Yes
27	CKS1B	20175	-1.038	-0.555	Yes
28	CCNE1	20246	-1.098	-0.524	Yes
29	CHEK2	20250	-1.100	-0.489	Yes
30	FOXM1	20382	-1.254	-0.456	Yes
31	CCNA2	20385	-1.257	-0.416	Yes
32	BIRC5	20407	-1.297	-0.376	Yes
33	CCNB2	20445	-1.381	-0.334	Yes
34	CENPF	20447	-1.381	-0.291	Yes
35	CDK1	20452	-1.392	-0.247	Yes
36	AURKB	20463	-1.427	-0.202	Yes
37	CENPA	20478	-1.494	-0.155	Yes
38	CCNB1	20492	-1.586	-0.106	Yes
39	PLK1	20496	-1.627	-0.054	Yes
40	NEK2	20499	-1.715	0.000	Yes

[a] Genes included in PID_FOXM1_PATHWAY.
[b] Position of the gene in the ranked list of genes based on the normal versus TNBC differential expression
[c] Score used to position the gene in the ranked list.
[d] The enrichment score at this point in the ranked list of genes.
[e] Genes with a Yes value in this column contribute to the leading-edge subset within the gene set. This is the subset of genes that contributes most to the enrichment result.

Appendix C: GSEA identified downregulated pathways in TNBC versus normal breast tissue

This table lists the top 20 downregulated pathways (in terms of FDR but are not statistically significant) in TNBC relative to normal breast tissue identified through GSEA. The 20 pathways are listed in descending ordered base on normalized enrichment score (NES).

ID	Gene Set Name	Size [a]	ES [b]	NES [c]	FDR [d]
1	**BIOCARTA_PPARA_PATHWAY**	**57**	**0.585**	**1.938**	**0.322**
2	PID_LPA4_PATHWAY	15	0.708	1.830	0.328
3	PID_RXR_VDR_PATHWAY	26	0.582	1.802	0.330
4	REACTOME_PHOSPHOLIPASE_C_MEDIATED_CASCADE	53	0.538	1.795	0.320
5	BIOCARTA_GATA3_PATHWAY	16	0.656	1.783	0.328
6	KEGG_ABC_TRANSPORTERS	44	0.516	1.781	0.307
7	REACTOME_NUCLEAR_SIGNALING_BY_ERBB4	38	0.520	1.767	0.327
8	BIOCARTA_EGF_PATHWAY	31	0.533	1.765	0.310
9	REACTOME_REGULATION_OF_WATER_BALANCE_BY_RENAL_AQUAPORINS	43	0.484	1.759	0.303
10	REACTOME_NUCLEAR_RECEPTOR_TRANSCRIPTION_PATHWAY	47	0.511	1.749	0.310
11	REACTOME_AQUAPORIN_MEDIATED_TRANSPORT	50	0.462	1.741	0.314
12	ST_JNK_MAPK_PATHWAY	40	0.510	1.740	0.301
13	ST_DIFFERENTIATION_PATHWAY_IN_PC12_CELLS	45	0.478	1.736	0.296
14	REACTOME_SIGNALING_BY_BMP	22	0.641	1.731	0.292
15	REACTOME_PI3K_CASCADE	68	0.489	1.722	0.303
16	REACTOME_CGMP_EFFECTS	19	0.614	1.710	0.319
17	KEGG_VASCULAR_SMOOTH_MUSCLE_CONTRACTION	114	0.447	1.690	0.337
18	BIOCARTA_GLEEVEC_PATHWAY	23	0.540	1.686	0.335
19	REACTOME_GROWTH_HORMONE_RECEPTOR_SIGNALING	24	0.525	1.655	0.339
20	REACTOME_NITRIC_OXIDE_STIMULATES_GUANYLATE_CYCLASE	25	0.548	1.645	0.336

[a] Size is the number of genes included in a gene set.
[b] Enrichment Score (ES) reflects the degree to which a gene set is overrepresented at the bottom of the list of genes ranked according to the normal versus TNBC differential expression.
[c] Normalized enrichment score (NES) is based on the gene set enrichment scores for all dataset permutations and makes GSEA results comparable across gene sets.
[d] False Discovery Rate (FDR) is the estimated probability that a gene set with a given NES represents a false positive finding.

Appendix D: The GSEA results for BIOCARTA_PPARA_PATHWAY in TNBC versus normal breast tissue

This table shows the GSEA results for BIOCARTA_PPARA_PATHWAY, listing genes in the pathway ordered by their position in the list of genes ranked according to the normal versus TNBC differential expression.

ID	Gene Symbol [a]	Rank in Gene List [b]	Rank Metric Score [c]	Running ES [d]	Core Enrichment [e]
1	STAT5B	38	1.429	0.058	Yes
2	HSD17B4	106	1.264	0.108	Yes
3	PIK3R1	320	1.055	0.142	Yes
4	DUSP1	376	1.022	0.182	Yes
5	CD36	641	0.893	0.206	Yes
6	CITED2	831	0.829	0.232	Yes
7	STAT5A	884	0.816	0.263	Yes
8	SP1	945	0.799	0.294	Yes
9	PRKAR1A	1218	0.728	0.311	Yes
10	NCOA1	1248	0.721	0.340	Yes
11	JUN	1281	0.716	0.368	Yes
12	MAPK3	1365	0.699	0.394	Yes
13	LPL	1371	0.698	0.423	Yes
14	PRKAR2B	1649	0.644	0.436	Yes
15	NCOR1	1700	0.636	0.460	Yes
16	RB1	1856	0.610	0.478	Yes
17	RXRA	2216	0.552	0.484	Yes
18	NR2F1	2292	0.540	0.503	Yes
19	EHHADH	2650	0.486	0.506	Yes
20	ACOX1	2815	0.465	0.517	Yes
21	PRKACB	2983	0.442	0.528	Yes
22	NRIP1	3037	0.435	0.543	Yes
23	PDGFA	3175	0.420	0.554	Yes
24	NR1H3	3302	0.404	0.565	Yes
25	CREBBP	3342	0.400	0.580	Yes
26	PRKAR2A	3561	0.374	0.585	Yes
27	PPARA	4085	0.320	0.573	No
28	EP300	4333	0.295	0.573	No
29	PIK3CA	4599	0.268	0.571	No
30	ME1	4980	0.236	0.562	No
31	NFKBIA	5308	0.210	0.555	No
32	MED1	5619	0.184	0.548	No
33	MAPK1	6240	0.139	0.523	No

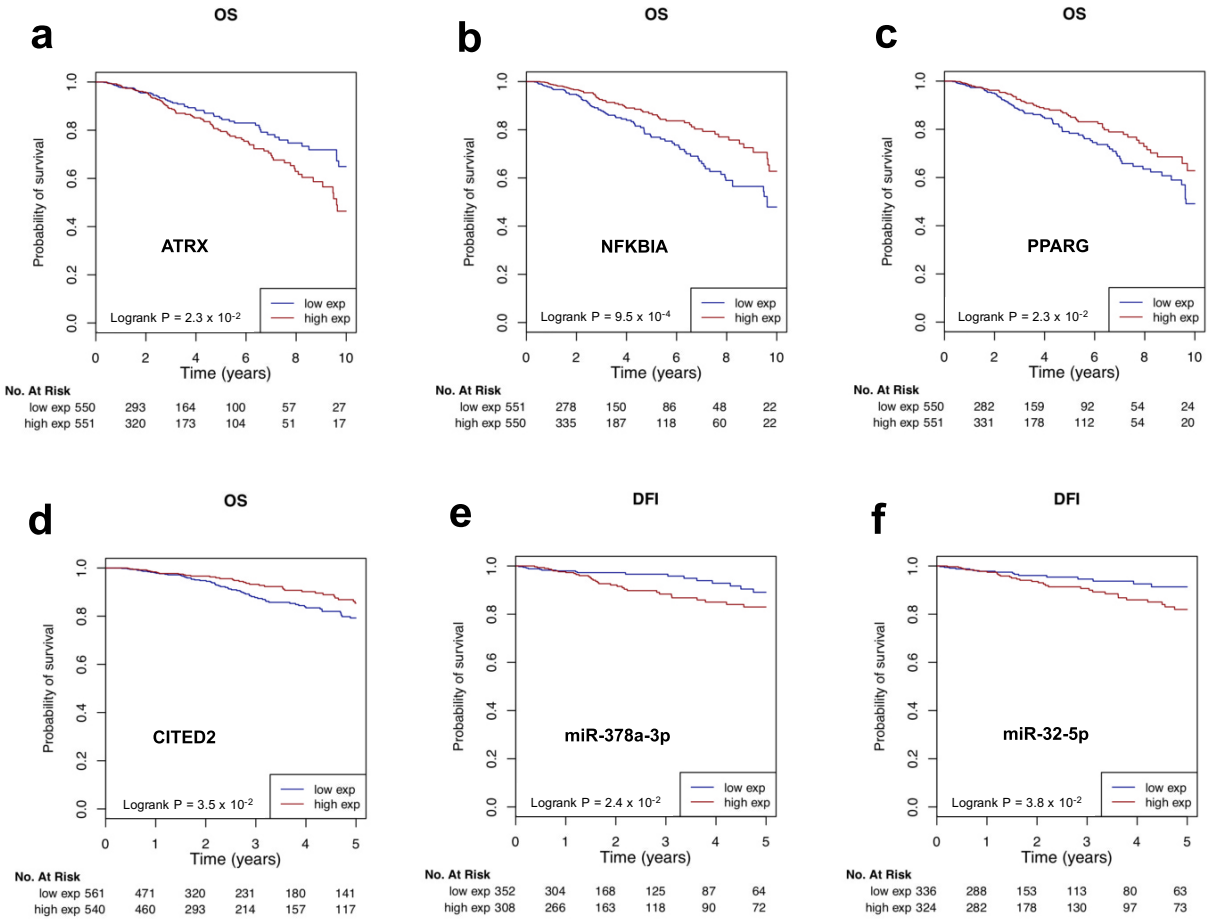
Appendix D: The GSEA results for BIOCARTA_PPARA_PATHWAY in TNBC versus normal breast tissue (cont.)

ID	Gene Symbol [a]	Rank in Gene List [b]	Rank Metric Score [c]	Running ES [d]	Core Enrichment [e]
34	CPT1B	6420	0.127	0.520	No
35	MYC	6487	0.122	0.522	No
36	FAT1	6937	0.100	0.504	No
37	PRKCA	8734	0.002	0.416	No
38	INS	8848	0.000	0.411	No
39	PRKACG	9194	-0.002	0.394	No
40	NR0B2	9716	-0.027	0.369	No
41	HSPA1A	9733	-0.028	0.370	No
42	NCOR2	10087	-0.045	0.354	No
43	PRKAR1B	10119	-0.047	0.355	No
44	PPARGC1A	10246	-0.052	0.351	No
45	FABP1	10696	-0.073	0.332	No
46	PTGS2	10727	-0.075	0.334	No
47	PIK3CG	11072	-0.091	0.321	No
48	NOS2	12092	-0.133	0.276	No
49	APOA2	12707	-0.158	0.253	No
50	PRKCB	13510	-0.195	0.222	No
51	APOA1	14419	-0.247	0.188	No
52	RELA	14770	-0.271	0.182	No
53	SRA1	15458	-0.314	0.162	No
54	DUT	16111	-0.358	0.145	No
55	TNF	16856	-0.419	0.126	No
56	HSP90AA1	17781	-0.512	0.102	No
57	MRPL11	19285	-0.733	0.059	No

[a] Genes included in BIOCARTA_PPARA_PATHWAY
[b] Position of the gene in the ranked list of genes based on the normal versus TNBC differential expression
[c] Score used to position the gene in the ranked list.
[d] The enrichment score at this point in the ranked list of genes.
[e] Genes with a Yes value in this column contribute to the leading-edge subset within the gene set. This is the subset of genes that contributes most to the enrichment result.

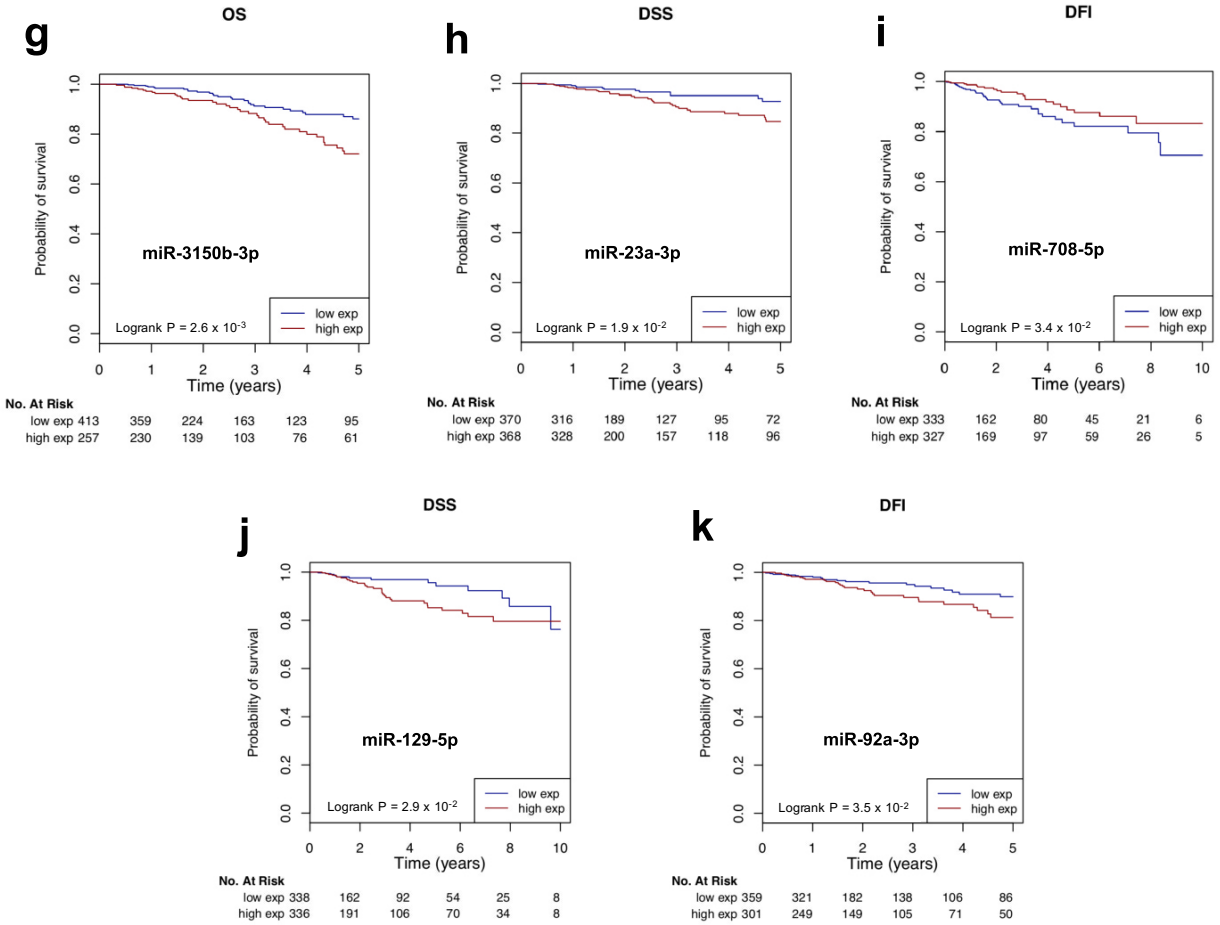
Appendix E-1: Survival analysis of the TNBC regulators in BC cohort

Based on gene expression levels of a given regulator, BC patients from TCGA were divided into over-expressed ("high-exp", red line) and down-expressed ("low-exp", blue line) groups. Survival analysis with different endpoints (OS, DSS, PFI, DFI) was performed for each regulator. Only significant results were shown. 11 regulators alone could be a prognostic marker for BC patients (six were shown here): (a) ATRX, (b) NFKBIA, (c) PPARG, (d) CITED2, (e) miR-378a-3p, and (f) miR-32-5p. OS, overall survival; DSS, disease-specific survival; PFI: progression-free interval; DFI, disease-free interval.



Appendix E-2: Survival analysis of the TNBC regulators in BC cohort

Based on gene expression levels of a given regulator, BC patients from TCGA were divided into over-expressed ("high-exp", red line) and down-expressed ("low-exp", blue line) groups. Survival analysis with different endpoints (OS, DSS, PFI, DFI) was performed for each regulator. Only significant results were shown. 11 regulators alone could be a prognostic marker for BC patients (five were shown here): (g) miR-3150b-3p, (h) miR-23a-3p, (i) miR-708-5p, (j) miR-129-5p, and (k) miR-92a-3p. OS, overall survival; DSS, disease-specific survival; PFI: progression-free interval; DFI, disease-free interval.



Appendix F: Associations between the expression level of PI3K pathway genes and PI3K inhibitor responses predicted by ML-CDN2

This table lists the PI3K pathway genes which show significant association between their expression levels with the predicted IC_{50} values of the PI3K pathway inhibitors.

Drug	Gene	Adjusted *p*-value
Temsirolimus	ADRBK1	2.01E-02
	CXCR4	3.84E-03
	NGF	1.92E-02
MK-2206	ADRBK1	2.14E-07
	PAK4	8.14E-04
	PPP1CA	1.52E-03
	PRKAG1	3.32E-02
	TRIB3	2.54E-02
Dactolisib	CXCR4	2.16E-03
	FASLG	2.60E-02
Pictilisib	ADRBK1	6.62E-06
	DAPP1	4.49E-04
	IL2RG	1.65E-02
	LCK	1.82E-02
	PAK4	2.44E-02
	PPP1CA	6.80E-03
AZD8055	PRKAR2A	2.71E-02
AZD6482	DAPP1	4.57E-03
	TRIB3	3.88E-02
PF-4708671	ADRBK1	3.07E-03
	DUSP3	8.65E-03
	IL2RG	4.92E-05
	LCK	2.13E-02
	PIK3R3	3.98E-03
Pictilisib	ADRBK1	6.62E-06
	DAPP1	4.49E-04
	IL2RG	1.65E-02
	LCK	1.82E-02
	PAK4	2.44E-02
	PPP1CA	6.80E-03
AZD8055	PRKAR2A	2.71E-02
AZD6482	DAPP1	4.57E-03
	TRIB3	3.88E-02
PF-4708671	ADRBK1	3.07E-03
	DUSP3	8.65E-03
	IL2RG	4.92E-05
	LCK	2.13E-02
	PIK3R3	3.98E-03
Pictilisib	ADRBK1	6.62E-06
	DAPP1	4.49E-04
	IL2RG	1.65E-02
	LCK	1.82E-02
	PAK4	2.44E-02
	PPP1CA	6.80E-03
AZD6482	DAPP1	4.57E-03
	TRIB3	3.88E-02

Appendix F: Associations between the expression level of PI3K pathway genes and PI3K inhibitor responses predicted by ML-CDN2 (cont.)

Drug	Gene	Adjusted *p*-value
ZSTK474	ARF1	7.60E-03
	ITPR2	1.76E-02
	NGF	3.70E-02
	PAK4	1.08E-05
	PIK3R3	1.82E-02
	TNFRSF1A	2.12E-04
AS605240	E2F1	4.15E-02
	SFN	1.60E-02
Idelalisib	ADCY2	2.99E-02
	ADRBK1	1.33E-03
	CAB39	5.86E-03
	CDK1	1.51E-02
	DUSP3	2.74E-04
	E2F1	4.44E-02
	PAK4	1.90E-02
	PLCG1	1.66E-03
	PPP1CA	3.47E-02
	PTEN	6.07E-04
	TNFRSF1A	1.72E-04
Omipalisib	AKT1S1	1.88E-02
	ARF1	4.61E-03
	NGF	8.80E-03
	PAK4	1.49E-02
	PIK3R3	9.75E-04
	TNFRSF1A	2.05E-03
PI-103	ARF1	3.27E-02
	PAK4	7.49E-03
	PRKCB	1.90E-02
	PTEN	1.12E-02
	TNFRSF1A	2.39E-06
BX-912	ADCY2	2.05E-04
	CAB39L	3.43E-02
	CDK1	1.48E-11
	CDK2	1.29E-05
	DUSP3	6.59E-05
	E2F1	2.04E-07
	PIN1	4.09E-03
	PLA2G12A	2.48E-04
	PRKAG1	6.38E-05
	TNFRSF1A	6.42E-04
	UBE2D3	2.06E-03

Appendix F: Associations between the expression level of PI3K pathway genes and PI3K inhibitor responses predicted by ML-CDN2 (cont.)

Drug	Gene	Adjusted *p*-value
PIK-93	ADCY2	3.98E-03
	ADRBK1	1.02E-05
	CDK2	7.56E-03
	DUSP3	2.68E-05
	ITPR2	1.38E-02
	NGF	2.10E-03
	PAK4	2.15E-04
	PIK3R3	1.80E-04
	PLCG1	8.51E-03
	PPP1CA	7.62E-05
	TNFRSF1A	3.78E-10
JW-7-52-1	ADRBK1	2.38E-04
	CDK1	2.48E-04
	EGFR	2.64E-02
	EIF4E	3.56E-04
	ITPR2	3.13E-03
	MYD88	1.29E-03
	NOD1	9.46E-04
	PPP1CA	5.90E-04
	RAC1	3.74E-03
	RIT1	6.11E-03
	RPS6KA3	3.55E-04
	SFN	3.99E-03
	STAT2	7.32E-05
	TSC2	2.45E-02
A-443654	ADRBK1	1.71E-03
	CAMK4	2.67E-03
	CDK1	2.43E-02
	EGFR	1.23E-02
	ITPR2	3.22E-03
	RPS6KA3	4.62E-04
	SFN	1.66E-02
	STAT2	3.75E-04
TGX221	CDK1	1.14E-12
	DUSP3	1.28E-03
	E2F1	3.26E-02
	FASLG	1.89E-03
	GNA14	4.40E-04
	MAPK8	7.58E-06
	NOD1	2.53E-02
	PIN1	7.07E-05
	PLA2G12A	2.36E-02
	RPS6KA3	2.00E-02
	TNFRSF1A	7.78E-07
GSK690693	ARF1	1.01E-02
	PAK4	1.63E-02
	PIK3R3	4.98E-04

Appendix G: Correlations of cell line pairs, tumor pairs and cell line-tumor pairs

(a) The pairwise Pearson correlation of the 49 BC cell lines from GDSC was calculated based on their gene expression profiles. (b) The pairwise Pearson correlation of the 1,100 BC samples from TCGA was calculated based their gene expression profiles. (c) The pairwise Pearson correlation between the 49 BC cell lines and the 1,100 BC samples was calculated based their gene expression profiles. (d) The pairwise Pearson correlation of the 49 BC cell lines was calculated based on their pathway activity profiles. (e) The pairwise Pearson correlation of the 1,100 BC samples from TCGA was calculated based their pathway activity profiles. (f) The pairwise Pearson correlation between the 49 BC cell lines and the 1,100 BC samples was calculated based their pathway activity profiles.