DISSERTATION

COMPUTATIONAL APPROACHES TO PREDICT DRUG RESPONSE TO CYTOTOXIC CHEMOTHERAPY

Submitted By

Joshua D. Mannheimer

School of Biomedical Engineering

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2020

Doctorial Committee:

Advisor: Daniel Gustafson
Co-Advisor: Ashok Prasad

Diego Krapf
Douglas Thamm

ABSTRACT

COMPUTATIONAL APPOACHES TO PREDICT DRUG RESPONSE TO CYTOTOXIC CHEMOTHERAPY

Cancer is the second leading cause of death in the United States. Statistically, within a lifetime there is slightly above a one-third chance of developing some form of cancer and a one in five chance of dying from the disease. Thus, it is no hyperbole that the understanding and treatment of cancer is one of the most pressing issues in medical research of the current era. Cytotoxic chemotherapies are a class of anti-cancer drugs that are widely used to treat a number of cancers. While cytotoxic chemotherapies are extremely effective in treating a subset of individuals for some cancers, drug resistance resulting in failure of treatment is a prominent obstacle in many cancer patients. Precision medicine, a novel concept to the 21st century, is the application of disease treatments that are specifically tailored to an individual and the specific attributes of their disease. In oncology, precision medicine particularly refers to the use of gene expression and other biological factors to inform an individual's treatment. Because cancer and its response to treatment result from many complex biological interactions, computational methods have become an essential tool to identify the molecular signatures that are the basis for precision treatment. In this thesis, a systematic analysis of the computational approaches is performed to gain insight necessary for the development of novel computational approaches in precision medicine in cancer.

Statistical learning models are a class of computational modeling methods that identify and extrapolate complex patterns from large amounts of data. Specifically, this involves applying statistical learning approaches on *in vitro* data from cell lines and patient tumor data to predict drug response, particularly for cytotoxic chemotherapies, with an emphasis on understanding the fundamental modeling principles and data attributes driving model performance. The first chapter serves as an introduction to chemotherapy and the advancements that have driven computational approaches to precision applications in cancer. The second chapter serves as a technical introduction to statistical learning models and approaches. In the third chapter a systematic assessment of linear and non-linear modeling approaches are applied to *in vitro* cell lines panel including the National Cancer Institute's 60 cancer cell lines (NCI60) and cell lines of Genomics of Drug Sensitivity in Cancer (GDSC) to predict drug response in several cytotoxic chemotherapies. With in-depth analysis it is shown that the relationship between tumor tissue histotype and drug response is the major driver of model performance and can be maintained in as little as 250 random genes. The fourth chapter utilizes statistical models to explore the influence of drug induced gene perturbations on drug response models in comparison with basal gene expression. The findings indicate that drug induced changes in gene expression are superior predictors of drug response. Second, it is demonstrated that Boolean network representation of gene interactions show distinct topological differences between drug induced changes in gene expression and basal gene expression. Finally, in the fifth chapter, drug induced gene changes demonstrating high levels of connectivity in the previously developed networks are applied to derive a basal gene expression signature to predict response to

combined gemcitabine and cisplatin chemotherapy treatment in patients with bladder cancer. These models show that this derived signature performs better than a random cohort of genes and in some situations genes derived directly from basal gene expression.

ACKNOWLEDGEMENTS

I would like to thank many people that have been active participants throughout my journey. First I would like to thank my advisor Dr. Dan Gustafson and co-advisor Dr. Ashok Prasad for their mentorship throughout the course of my Ph.D. education. Additionally, would like to thank Dr. Diego Krapf and Dr. Douglas Thamm for being valued members of my committee. Likewise I would like to thank my colleagues Keagan Collins and Katherine Cronise for making life just a little bit more interesting and being a part of my journey.

Throughout the course of my life there have been several family, friends, and teachers who have played a crucial role in my success. I would like to thank my former teachers Joan Gardner, Holly Baldwin, Stephanie Derringer, and Angelia Rosende for playing a pivotal role early in my education. Furthermore, I give thanks to Geoff George and Jake Quigley for their mentorship through the years. Additionally, my grandmothers, Carol and Nanette have been an endless source of encouragement. My father, Mark, who has encouraged me along the way.

Finally, I would like to give a very special thanks to my brother, Brandon, and my Mom, Kim. Brandon, has always been there sharing both the darkest and happiest of times and will always be my best friend. Finally, my mom has always been my biggest fan and strongest supporter in this roller coaster called life.

DEDICATION

This work is dedicated to my grandfather Rudy Mannheimer (1917-1989) who unfortunately was taken too soon by cancer to see the great men his grandsons have become and will be forever in our memories

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Cancer as a Disease:

In 2017, the most recent data available on mortality rates, cancer was the second leading cause of premature death in the United States accounting for 599,108 deaths, or 21% of total deaths (1). Likewise, the total number of deaths from cancer in 2017 was estimated to be 9.56 million worldwide according to the Global Health Data Exchange (www.ghdx.healthdata.org). In 2020 it is estimated that 1,806,590 cases of a cancer will be diagnosed and another 606,520 people will succumb to the disease in the United States (1). Furthermore, the American Cancer Society reported that 40.14% of males and 38.7% of females will develop cancer in their lifetimes with about a 1 in 5 chance of dying from the disease (2). The burden cancer places on society extends beyond the number of deaths but also includes economic hardship. As healthcare costs have come under scrutiny in the United States, a 2005 article estimated that medical costs, contributes to 60% of American bankruptcies (3). Undoubtedly, this burden is significant among cancer survivors, where a survey from 2013 to 2016 reported that 43.4% of cancer survivors 18 to 49 expressed having financial hardship due to medical care compared to 30.1% among the general population (4). These statistics, among others, illustrate the importance of research to improve cancer treatment with respect to efficacy while also making it more available in terms of cost.

Early references to cancer can be traced back to ancient Egyptian documents placed between 1600 and 1500 BC (5); however, these documents are thought to be reproductions of an original text dating to somewhere between 3000 and 2500 BC.

Hippocrates (460-370 BC), a Greek physician often credited as being the "*father of medicine",* first used the term "*Karkinos"* to describe ulcerating non-healing lumps (5). Throughout the Roman and middle ages the search for the underlying cause of cancer remained one of the most puzzling medical mysteries at the time. Progress in new surgical techniques to remove tumors, anatomical and pathological knowledge, and humanitarian cancer patients care were advanced (5). Early theories about the origin of cancer included the "*black bile theory",* the imbalance of black bile (one of the four essential fluids (humoral) in the body), which was first championed by Hippocrates and was the predominant theory up until the middle ages (6). Other theories would arise including the notion that cancer might be a contagious disease (6); however, the anatomist and surgeon Alfred Armond Louis Marie Velpeau (1795-1867) would suggest that cancer "is merely a secondary product" of an underlying "intimate element" closely resembling the current understanding of cancer as a genetic disease (5).

At the turn of the 21 century, technological advances, started by the discovery of the structure of DNA in 1953 by Franklin, Watson, and Crick (7), had evolved into a number of discoveries about the human genome and the machinery of cell function. This had resulted in a fundamentally different view of cancer as disease, one which Hanahan and Weinberg would describe as "already complex almost beyond measure" (8). In what has become a seminal paper in cancer literature *The Hallmarks of Cancer*, Hanahan and Weinberg would lay out six fundamental principles which, in their opinion, conceptually linked all cancers; self-sustaining growth signals, insensitivity to antigrowth signals, the ability to evade apoptosis, limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis (8). In 2011, the authors would

amend the list with two emerging hallmarks: deregulation of cellular energetics and ability to evade the immune system (9). Roughly, these together form the simplest definition of cancer: A disease that results from an acquired ability of cells to manipulate mechanisms of cell growth, normal cellular and biological regulation processes, and surrounding tissues to sustain the requirements of unlimited cell replication.

Chemotherapy:

Throughout the nineteenth century and early into the 20th century surgical removal of tumors had been the predominant treatment (10). Discoveries of x-rays (11) and subsequent discoveries in radioactive materials (12) would lead to the development of radiation therapy in the 1920's (10). Meanwhile, early in the twentieth century the term "chemotherapy" would be coined by a German chemist Paul Ehrlich (5, 10, 13). Ehrlich would be instrumental in advancing the concept of using chemicals to treat cancer though with little success personally (13). The effects of mustard gas in World War I coupled with aftermath of a World War II fatal disaster that spilled mustard gas into the Bari Harbor, prompted research into the chemical properties of mustard compounds (5, 13). These studies would lead to subsequent experimental studies that demonstrated anti-tumor activity of nitrogen-mustard, a derivative of mustard gas, in implanted lymphoid tumors in mice (5, 13, 14). This finding led to further observations that nitrogen mustards resulted in brief remissions in lymphoma patients; however, skepticism remained as to the efficacy of drugs to treat cancer, especially in the long term (13, 14). As the search for anti-tumor chemical products continued, a new class of drugs, antifolates, emerged (13, 15). Aminopterin, the first antifolate to become clinically available in 1948, had proven to be effective at producing remissions in children with

acute lymphoblastic leukemia (16). Meanwhile, even with rapid technological advancements, surgery and radiotherapy reached their zenith in the 1950's and were only effective in about a third of all cancers (10). This, coupled with new optimism about chemotherapy, would usher in the advancement of chemotherapy in the decades to come. New drugs and drug combinations continued to be discovered and developed throughout the 1950s and 1960s with continued success in treating childhood leukemia and breakthroughs for treatment of Hodgkin's disease in adults (10, 13). As the foundation was set for the successful treatment of some cancers with chemotherapy, evidence emerged that chemotherapy was successful in treating micro-metastases leading to the practice of adjuvant chemotherapy, chemotherapy in addition to surgical or radiological treatment (10, 13) . Advancements in neoadjuvant chemotherapy, chemotherapy applied before surgery or radiation therapy, has resulted in considerable progress in treating some cancers and remains a standard treatment to many cancers currently.

The therapeutic efficacy of chemotherapeutic agents is thought to be derived from mechanisms which interfere with cell replication and proliferation, thereby having a preference for rapidly dividing cells as is the case of cancer (14, 17). Cytotoxic chemotherapies are broadly categorized into alkylating agents, anti-metabolites, cytotoxic antibiotics, topoisomerase inhibitors, and anti-microtubule agents (14, 17, 18). What follows is a brief overview of the general mechanisms, by category, of chemotherapies that have been studied throughout the body of this text and can be found in Table I.

Table 1.1: Select Cytotoxic Chemotherapies and the commonly used throughout the text and organized by class and the cancers they are FDA approved to treat.

Category	Drug	FDA Approved Use
Alkylating Agents • Nitrogen Mustards	Cyclophosphamide[1]	Acute lymphoblastic leukemia, Acute monocytic leukemia, Acute myeloid leukemia, Breast, Chronic granulocytic leukemia, Chronic myelogenous leukemia, Hodgkin Lymphoma, Neuroblastoma Non-Hodgkin lymphoma, Ovarian, Retinoblastoma, Mycosis fungoides (19)
• Platinum Analog	Cisplatin	Bladder, Ovarian, Testicular (20)
Anti-Metabolites • Pyrimidine analogues	Azacytidine	Myelodysplastic Syndromes(21)
	Cytarabine	Acute non-lymphocytic leukemia, Meningeal leukemia, Acute lymphoblastic leukemia, Chronic myelogenous leukemia (22)
	Gemcitabine	Breast, Non-small lung, Ovarian, Pancreatic (23)
	5-Fluorouracil	Breast, Colorectal, Gastric, Pancreatic (injection) (24) Basal Cell Carcinoma (Topical) (25)
• Antifolate	Methotrexate	Acute lymphoblastic leukemia, Breast, Head and Neck, Lung, Mycosis fungoides, Non-Hodgkin lymphoma, Osteosarcoma

[1] Cyclophosphamide is one example of a nitrogen mustard, several other drugs originate from this class such as melphan, bendamustine, chlorambucil etc.. This list represents an abbreviated of example of different cytotoxic drugs and the classes they belong.

Cytotoxic Antibiotics	Bleomycin	Hodgkin lymphoma, Non-Hodgkin lymphoma, Squamous Cell Carcinoma, Testicular (26)
	Mitomycin C	Gastric, Urothelial (27)
• Anthracyclines	Doxorubicin	Acute lymphoblastic leukemia, Acute myeloid leukemia, Breast, Stomach, Hodgkin lymphoma, Neuroblastoma, Non-small cell lung, Soft tissue sarcoma, Thyroid, bladder, Wilms tumor (28)
Topoisomerase Inhibitors		
• Topo II	Etoposide	Small cell lung, Testicular (29)
• Topo I	Topotecan	Cervical, Ovarian, Small cell lung (30)
	Irinotecan[2]	Colorectal (31), Pancreatic (32)
Anti-microtubule Agents		
• Taxols	Docetaxel	Breast, Non-small cell lung, Prostate, Squamous cell carcinoma (head and neck), stomach adenocarcinoma, Gastroesophageal adenocarcinoma (33)
	Paclitaxel	Breast, Non-small cell, Ovarian (34)
• Vinca alkaloids	Vinblastine	Breast, Hodgkin lymphoma, Kaposi sarcoma, Mycosis fungoides, Non-Hodgkin lymphoma, Testicular (35)
Miscellaneous		
Proteasome Inhibitor	Bortezomib	Mantle cell lymphoma, Multiple myeloma (36)
HDAC Inhibitor	Vorinostat	Cutaneous T-cell lymphoma (37)

[2] Irinotecan is the pro-drug for SN-38

Alkylating agents include the first successful chemotherapies, nitrogen mustards that arose from interest in mustard gases used during and after World War I (38). Traditional alkylating agents involve covalent bonding of reactive alkyl groups to carbon rich cellular molecules (38, 39). The general mechanism by which these drugs work is interacting with the DNA base guanine resulting in DNA damage inhibiting DNA replication or repair and thus inducing apoptosis (14). For example, cyclophosphamide goes through a number of metabolic sets resulting to the active alkylation agent phosphoramide mustard which interacts with the N7 position of multiple guanine resulting in intra-strand cross links that inhibit DNA replication (14, 39). Likewise, the drug cisplatin is known as an alkylating-like agent, in that its main mechanism of cytotoxicity is the introduction of intra-strand cross-links due to preferential binding to guanine similar to alkylating agents (14, 39). However, cisplatin achieves cross-linking through interactions of its platinum complex with DNA (39).

Drugs which derive therapeutic properties by interfering with specific metabolites or metabolic enzymes necessary for cellular processes have been termed antimetabolites (14). One of the first antimetabolites, methotrexate, resulted directly from the success of antifolates in pediatric acute leukemia (13, 15) and resulted in the first radiographically confirmed case of tumor regression in 1956 (40). The anti-tumor properties of methotrexate result from inhibition of folate metabolism, which is a necessary precursor for DNA synthesis (41). Likewise, some chemotherapies from this category act as chemical substitutes for molecular components of DNA, interfering with DNA machinery needed for replication. For example, replacement of pyrimidine nucleosides, DNA bases cytosine and thymine, by chemically modified pyrimidine

analogs, result in anti-tumor activity by interfering with DNA replication (42). For example, integration of the deoxycytidine analog cytarabine in place of the nucleoside deoxycytidine into the DNA backbone significantly interferes with the DNA polymerase resulting in chain termination (43). Gemcitabine, another deoxycytidine analog, interferes with DNA elongation; in addition, gemcitabine inhibits ribonucleotide reductase, the enzyme responsible for converting ribonucleotides to deoxyribonucleotide, this results in depletion of natural deoxynucleotides increasing the uptake of gemcitabine as a substrate for elongation (44). Similar drugs, purine analogs, work in a similar fashion, however, instead interfere with the metabolism of the purine DNA bases adenine and guanine.

Cytotoxic antibiotics consist of a group of naturally occurring compounds that have antibiotic properties but additionally have strong anti-tumor activity (45). The mechanisms of cytotoxic chemotherapies vary as well. Included in the cytotoxic antibiotics are the drugs doxorubicin and mitomycin c. Doxorubicin became clinically available in the 1970's (46) and several mechanisms of doxorubicin cytotoxicity had been studied. One of the more simple mechanisms which has been put forth to explain the anti-tumor effects of doxorubicin involves intercalation between DNA base pairs which interfere in DNA replication and RNA transcription (47-49). Additionally it has been proposed that doxorubicin mediates cytotoxicity as a topoisomerase inhibitor resulting in DNA damage and activation of apoptosis in the G1 or G2 phase of the cell cycle (47, 49, 50). The role of the topoisomerase enzymes are discussed in a subsequent paragraph along with additional topoisomerase inhibitors. The oxidization of doxorubicin into an unstable semiquinone metabolite along with the reverse reaction

creates reactive oxygen species that induce damage to several macromolecules including DNA, proteins, and lipid membranes (49). Furthermore, doxorubicin has been shown to create lipid molecules, known as ceramides, which play a role in the activation of the transcription factor CREB3L1, which upon proteolytic cleavage activate genes involved in cell cycle inhibition (51). Doxorubicin is a drug that is widely used clinically to treat a variety of cancers; however, adverse side effects such as cardiotoxicity (47) introduce a need to better understand the many, known and unknown, cytotoxic effects of the drug which could help identify subpopulations of patients who are likely to benefit from treatment.

As mentioned above, mitomycin c (MMC) is another clinically available drug from the class of cytotoxic antibiotics. The primary mechanism of cytotoxicity mediated by MMC is the formation of intermediate reactive species which introduce intra-strand or inter-strand cross links predominantly in the G1 and S phase of the cell cycle (52). In order for MMC to form intra-strand cross links it must be undergo bioreduction into an active alkylating agent that interacts with DNA to form crosslinks (53, 54). The active form of MMC, mitosene, is formed when the quinone group of MMC is reduced resulting in the active alkylating agent which binds to the N2 position in guanine (53, 54). Inter or intra-strand crosslinks are created when a second alkylating group is formed in a reverse Michael elimination of a carbamate group forming a second bond with guanine at either the N2 position or the N7 position (53, 54). Several bioreductive enzymes have been shown to activate mitomycin c including DT diaphorase (55), NADH cytochrome c reductase (55), NADPH cytochrome c reductase (55), cytochrome P450 reductase (56), xanthine oxidase (56), cytochrome b_5 reductase (57), and xanthine dehydrogenase

(58). The availability of the different bioreductive enzymes might play a crucial role in determining cytotoxicity depending on surrounding microtumor environment. For example, it had been shown in CHO cells that an increase in cytochrome P-450 resulted in minimal increase in MMC activation under aerobic conditions; however, significantly greater MMC activation was exhibited in anerobic; suggesting that cytochrome P450 might play a pivotal role in MMC activation in hypoxic tumor environments (59). Likewise, in EMT6 mouse mammary tumors Gustafson et. al. demonstrated that reduction by xanthine dehydrogenase resulted in the preferential formation of the metabolite, 2,7-diaminomitosene, forming guanine N2/N7 crosslinks, in hypoxic and acidic pH conditions leading to an increase in alkylating ability in hypoxic and acidic environments (58). Furthermore, directly related to MMC sensitivity, it was shown that MMC sensitivity in the NCI60 cancer cell line panel (60) correlated with expression of DT diaphorase, the major reductive enzyme in oxygenated environments (61). Mitomycin C sensitivity tends to favor hypoxic conditions (62-64) suggesting that differences in expression of different bioreductive enzymes might play a role in sensitivity particularly in tumors favoring hypoxic conditions such a breast, uterine cervix, brain, squamous cell carcinoma, brain, and head and neck tumors (65). Doxorubicin and mitomycin c are good examples of how cytotoxic chemotherapies can have multiple mechanisms of action which can influence the sensitivity of a tumor to a certain drugs; however, the availability of biomarkers for cytotoxic chemotherapies are still limited (66).

During DNA replication the two strands of the double helix are pulled a apart by the DNA helicase. This separation results in excess tension in front of the replication

fork eventually making the two strands impossible to separate. In order to relieve this tension super coils began to build up in the DNA strand but eventually the super coiling can no longer mitigate the tension build-up prohibiting the DNA replication machinery form moving forward. The enzymes responsible for mitigating this excess tension are the topoisomerases. Topoisomerases come in two classes; topoisomerase I and topoisomerase II. Topoisomerase I releases tension through a single stranded break within the helix, rotating the other stand through the break and then splicing the break back together. Topoisomerase II induces a double stranded break within a supercoil relaxing the coil and then splices the primary helix back together. Topotecan, an inhibitor of topoisomerase I, binds in the pocket of the enzyme after it induces a strand break, binding the enzyme, the drug, and the DNA together in a combined complex, as this complex approaches the replication fork double strand breaks are introduced resulting initiating apoptosis (67). Alternatively, etoposide, a topoisomerase II inhibitor, stabilizes the cleavage state after a double stranded break is created, preventing the enzyme from repairing the break (68). As these strand breaks become abundant, DNA replication is inhibited, and apoptosis is eventually triggered (68).

Microtubules are essential components of the cell and play a central role in the structure of the cell, acting as molecular “highways” for intra-cellular transport, and are essential in cell division (69). During interphase a portion of the microtubules are disassembled into tubulin subunits and then reassembled to form the mitotic spindle (69). The interference in the assembly and disassembly of microtubules during mitosis is the mechanism of action for anti-microtubule agents. The taxols, which include paclitaxel and docetaxel, work by stabilizing tubulin in its polymeric form thus

decreasing microtubule disassembly, thus decreasing the free tubulin and inhibiting microtubule reorganization that must take place for cell division (70). Alternatively, a class of drugs known as the vinca-alkaloids, represented by vinblastine, binds free tubulin preventing microtubule formation thereby inhibiting spindle formation and thus cell division (70).

One of the biggest challenges to treating cancer with cytotoxic chemotherapy, or any pharmacological therapy, drug resistance. A variety of mechanisms are known, that include everything from decreasing drug concentration in the cell, altering drug targets, and up-regulation of anti-apoptotic regulators (14). A known mechanism of resistance to multiple drugs, including but not limited to doxorubicin, cisplatin, methotrexate, vinblastine, etoposide, and paclitaxel involves increased efflux of the drug out of the cell by upregulated ATP-binding cassette (ABC) drug transporters deceasing cellular concentrations (71, 72). Furthermore, additional mechanisms that decrease drug uptake into the cell can impart resistance, for example, in ovarian cancers doxorubicin is often delivered via encapsulation in a synthetic liposome. Drug resistance is seen in cells that lose the LPP1B cell surface protein which is involved in the transport of liposomes across the cell membrane (14, 73). Another example is seen with methotrexate where decreased expression of the transport protein, reduced folate carrier (RFC), limits cellular uptake (74, 75). Nucleotide excision repair, the removal of cross-linked nucleosides by DNA repair machinery, is observed in cisplatin resistance (76, 77). Furthermore, many chemotherapeutic drugs are prodrugs and require enzymatic activation into their active form. Enzymes that are necessary for activation can become common mechanisms of resistance; for example, gemcitabine requires the addition of

three phosphate groups to form the active metabolite gemcitabine triphosphate after transport into the cell. One of the enzymes responsible for the transformation of gemcitabine to gemcitabine triphosphate is deoxycytidine kinase (dCK) (78). The down regulation of dCK has been observed in gemcitabine resistant pancreatic cells (79). Mutations in drug targets can also contribute to resistance; for example, mutations in tubulin have been shown to decrease vinca-alkaloid binding (80). Furthermore, resistance can be mediated by directly altering the expression of the target such is the case for etoposide where the predominant mechanism of resistance is decreased expression of topoisomerase II (81). Additionally, gene expression changes in signaling pathways that regulate apoptosis or cell cycle can be drivers of resistance. For example, the family of BCL-2 proteins act as an apoptotic activator, overexpression of BH3, a member of the BCL-2 family, correlates with paclitaxel sensitivity in non-small-cell lung cancer (82). Likewise, decreased expression of cell cycle proteins in the MAPK pathway have been associated with doxorubicin resistance in breast cancer (83). Despite the discovery of several new mechanisms of drug resistance, drug resistance continues to be an obstacle to effective cancer treatment (84, 85). Therefore, the development of biomarkers or computational techniques that can identify subpopulations of patients who are likely to be more responsive to certain drugs could play a crucial role in more effective cancer treatment.

Targeted Therapies: Kinase Inhibitors

Cytotoxic chemotherapies have mechanisms that primarily involve interfering with machinery needed in cell replication and division. This toxicity is not specific to cancerous cells but applies to all dividing cells. The theory behind why cytotoxic

chemotherapies have antitumor properties stems from the preference of these drugs to target proliferating cells whereas most cells in the body are not dividing as rapidly as cancerous cells. However, the harsh side-effects that are common with most therapies result from the drugs ability to affect all dividing cells in the body. This is one of the limiting factors in treatment with cytotoxic chemotherapies, both from the standpoint of managing side-effects and planning treatment; a sufficiently long time on chemotherapy would effectively start interfering with normal cell division processes essential for maintenance of different tissues. This would certainly eliminate the cancer but in the process it would kill the patient as well. A better understanding of the genetic and molecular nature of cancer in the 20th century led to the discovery of several genes or mutations necessary for neoplastic growth, oncogenes. This would lead to a more rational drug development approach, by targeting oncogenic pathways the specific mechanisms that allow for neoplastic growth would be inhibited.

The role of kinases in cancer initiation and progression has resulted in a number of small molecule kinase inhibitors, including the first targeted drug imatinib which was shown to inhibit a key tyrosine kinase Bcr-Abl in the formation and proliferation of leukemias (14, 18). There are roughly 538 different protein kinases encoded in the human genome and hundreds have been shown to influence cell transformation, tumor initiation, survival, and proliferation in cancer (86). As a result, 46 kinase inhibitors have been FDA approved as treatments in various cancers (87). Kinase inhibitors are one of the most actively studied in ongoing cancer clinical trials; for example, as of 2018 there were 150 active clinical trials involving kinase inhibitors (86). Most kinase inhibitors block the binding site of ATP, preventing the transfer of a phosphate group from ATP to

a protein substrate. Inhibitor specificity results from high frequency oncogenic mutations specific to certain kinases and cancers (86). Common pathways targeted by kinase inhibitors include Bcr-Abl, VEGF, EFGR, HER1 and HER2 (table 2) (88). Resistance to kinase inhibitors is mediated through two different mechanisms; intrinsic resistance and acquired resistance (89). Intrinsic resistance arises from pre-existing mechanisms that prevent the drug from working; acquired resistance mechanisms develop after initial treatment (89). Similar to multi-drug resistance in cytotoxic chemotherapies, intrinsic mechanisms include increased efflux of the drug out of the cell by increased ABC membrane transporters (90). Acquired resistance to targeted therapies is one of the main challenges in targeted therapy, after an initial positive response recurrence often results from acquired resistance mechanisms (91). For example, acquired resistance can result from a secondary mutation in the binding site of the drug (89). Because there is a large overlap in these signaling pathways, resistance can result through upregulation or activation of redundant or alternative signaling proteins reactivating the signaling pathway or activation of pro-survival signaling thereby disrupting mechanisms of apoptosis (91). Despite the very specific approach of targeted therapies, including kinase inhibitors, factors such as high mutation rates, signal crosstalk and redundancy, and the complexity of signaling in oncogenic pathways can lead to ineffective treatment and a multi-drug approach is required to overcome multiple mechanisms of resistance.

Table 1.2: Select Kinase inhibitors, their targets, and FDA approved uses.(88)

Drug	**Target**	**FDA approved treatments**
Dasatinib	Bcr-Abl	Chronic Myeloid Leukemia, Acute Lymphocytic Leukemia
Erlotinib	EGFR	Non-Small Cell Lung Cancer, Pancreatic Cancer
Lapatinib	HER2	HER2 positive Breast Cancer
Sorafenib	VEGF	Renal Cell Cancer, Hepatocellular Carcinoma
Sunitinib	c-kit, VEGFR PDGFR,	Renal Cell Cancer, Gastrointestinal Stromal Tumor

Cancer: A Genomics View:

Cancer is heterogenous. It is a complex disease and as a result treatment is complex, difficult, and evolving. The idea of precision medicine has arisen as a means to address the heterogenous nature of cancer and leverage specific pathological attributes for diagnosis, prognosis, and treatment (92). Such approaches have given rise to biomarkers, “A characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention.” (93). The discovery and development have been the topic of a great amount of research and resulted in a number of relevant applications in several cancers. What follows is a brief discussion of the development and implementation of genomic biomarkers which have risen over the past 3 decades.

It is estimated that 15.5% of all new cancers in 2020 will be breast cancer and also account for 7% of all cancer related deaths (1). For these reasons a considerable amount of cancer research has been directed to breast cancer and has allowed for some of the biggest breakthroughs in cancer therapy and detection (94). Traditionally, as with all cancers, one of the biggest challenges has been understanding the

relationship between the biological factors contributing to the epidemiology and efficacy of treatment. Historically, histological grading was one of the most prominent methods of predicting outcome. In 1979 Freedman et al. was able to show histological grading based on tubule formation, number of mitotic nuclei, and shape of nuclei was capable of predicting outcome in a cohort of 1759 breast cancer patients (95). Focus has continued on improving the histological grading scale and it continues to be a standard practice today in diagnosis; however, the advent of gene expression profiling has been essential in developing both better diagnosis and treatments (96-98).

Technological advancements in sequencing techniques would usher in the ability to experience biological phenomena on a molecular level including cancer (99). Breast cancer's role in the development of precision medicine is best put by Lukong: "*The molecular classification of breast cancer based on gene expression profiles reported by research groups in the first decade in the 21st century is one of the momentous developments in personalized medicine in recent years." (94).* One of the earliest advancements in molecular cancer treatment in breast cancer was the discovery of the newly sequenced HER2 (human epidermal growth factor 2) gene (100). Soon after discovery it was demonstrated that over expression of HER2 correlated with poor clinical outcomes in female breast cancer (101). The discovery of HER2 would lead to the development of one of the first biomarker based drugs trastuzumab, a multiclonal antibody with high affinity for the extracellular domain of HER2, first granted FDA approval in 1998 for combination use with paclitaxel for HER2-positive metastatic breast cancer (94). In 2006 trastuzumab was approved for neo-adjuvant use on early stage

HER2 positive breast cancers, demonstrating a reduction in recurrence by reportedly 50% (94).

The development of gene array technology in 1995 would launch a new molecular era of cancer research by allowing rapid and consistent measurement of expression from multiple genes simultaneously (102). As this technology progressed, a genetic finger print of breast cancer was starting to emerge, and gene expression would allow for the genetic subtyping of breast cancer (94). Analysis of 43 breast tumors using gene expression for 496 genes showed that hierarchical clustering could subtype the samples into four primary groups: basal like, Erb-B2/HER2 positive, normal breast like, and luminal estrogen receptor (ER) positive based on co-expression patterns (103). Another hierarchal cluster analysis of 78 breast cancer tumor samples of 456 genes also demonstrated the stratification between basal like, ERB-B2 positive, and normal breast like but found that the luminal could be further be discretized into further subtypes, luminal A, luminal B, and luminal C; additionally, they established that basal and ERB-B2 positive cancers were associated with poor prognosis compared with other subtypes (104). Today there are five main molecular subtypes based on genomic analysis: luminal A, luminal B, triple negative basal like, HER2-enriched, and normal like (105). The discovery of these molecular markers have resulted in a number of targeted therapies for breast cancer. In addition to trastuzumab and its derivates for over-expression of HER2, a tyrosine kinase inhibitor, lapatinib, was developed an alternative therapy for HER2 positive breast cancer that is unresponsive to trastuzumab (94). Fulvestrant is an estrogen receptor antagonist used on ER positive breast cancer (106).

Additionally, aromatase inhibitors which interfere with estrogen synthesis, can be effective in preventing relapse in hormone receptor positive breast cancers (107).

As seen in breast cancer, molecular classification has had tangible effects with regards to diagnosis, prognosis, and drug treatments. Thus, there have not only been efforts in breast cancer but many other families of cancers as well. Classification of cancers by histology has relied on morphological appearance; however, histological similarities are rarely indicative of individual patient's response to therapy (108, 109). Golub et al. carried out one of the initial experiments to classify acute leukemias using only gene expression data (108). From a collection of 38 bone marrow samples (27 acute lymphoblastic leukemia (ALL), 11 from acute myeloid leukemia (AML)) and gene expression of 6817 genes, a 50 gene predictor was constructed to discern between the two different classes of leukemia; when applied to an independent collection of leukemia samples the predictor was able to correctly predict the class of 29 of the 34 samples (108).

Clinically significant subtypes of diffuse large B-cell lymphoma (DLBCL) could also be differentiated by gene expression profiles. It was found that there were two distinct gene expression profiles that "reflected variation among tumour proliferation, host response and differentiation state of tumours" (109). Using hierarchical clustering on 2984 genes, DLBCL samples clustered into two distinct groups, those with a similar gene expression profile to germinal center B cells, GCB-like DLBCL, and those with gene expression profiles closer to activated B cells, ABC-like DLBCL (109). Furthermore, it was noted that the clustering was a multi-gene phenomenon, a single gene alone could not differentiate the two subgroups (109). Additionally, the

classification of the DLBCL subgroup had a profound influence in overall survival with GCB-like DLBCL carrying a much higher probability of survival compared to ABC-like DLBCL (109).

Many of the earlier methods of cancer classification utilized univariate correlation and clustering methods to construct gene expression signatures. Alternatively, different approaches utilizing machine learning to classify cancers and construct gene expression signatures were also being explored. Khan et al. utilized artificial neural networks to classify small, round blue cell tumors, including neuroblastoma, rhabdomyosarcoma, non-Hodgkin lymphoma, and Ewing's sarcomas (110). Utilizing an iterative training approach with a two layer artificial neural network (ANN) they isolated a 96 gene signature with 100% accuracy on a test set of 20 tumor samples (111). Similarly, an ANN was trained to distinguish between Barrett's esophagus and esophageal cancer achieving 100% accuracy and a gene signature of 160 genes (112). Several studies applied support vector machines (SVM) to tumor classification as well. Furey et al. used a linear SVM for classification on the same set of AML and ALL tumors as Golub et al. (108), achieving slightly results correctly 30 to 32 tumors correctly, compared to 29 of 35 by Golub, using gene signatures ranging from 25 to 1000 genes (113). Likewise, Su et al. achieved 90% accuracy using a one versus the rest SVM classifier and 110 gene signature to distinguish between tumors from 11 different carcinomas (114).

Predicting Patient Outcome With Gene Expression:

These selected examples demonstrate that gene expression profiling could successfully differentiate between tumor types and subtypes. The ability of gene markers to establish breast cancer subtypes had important consequences in determining treatments and drug development. Additionally, it was shown that multigene markers could differentiate acute leukemias and led to the discovery of two different subtypes in DLBCL that could explain differences in survival. The clinical significance of these results suggested that gene expression might have the ability to predict drug response. Early, demonstrations of the ability of gene expression to predict treatment response was successful in medulloblastoma. Pomeroy et al. took 60 pediatric tumors and applied a K-nearest neighbors (KNN) classifier to classify patients as either responsive to treatment or unresponsive to treatment (115). Using a leave one out cross validation they established an eight gene signature with only a 15% miss-classification rate ($p=0.009$) (115). Additionally, when comparing this classifier to tumor staging, the KNN classifier achieved greater significance ($p=0.002$ vs $p=0.03$) in classification which they claim demonstrates[3] the additional benefit of gene expression based markers (115).

The success of gene expression profiles to delineate clinically meaningful classifications in DLBCL (109) directly resulted in efforts to model patient outcome. Shipp et al. used a cohort of 58 DLBCL patients, 32 with cured disease and 26 with fatal or refractory disease, and used a supervised learning technique to classify samples

[3] This was stated by the authors of the original cited work and is not a viewpoint necessarily taken by authors of this work.

either as cured or those who had fatal/refractory disease after treatment with a regimen of cyclophosphamide, docorubicin, vincristine, and prednisone (CHOP) (116). Using cross validation they showed that a 13 gene model could accurately predict the long term survival of DLBCL patients (116). In a similar study 240 DLBCL patient were split into a preliminary group of a 160 samples and a validation group of 80 samples (117). A 17 gene signature, based on correlation with outcome, was used to construct a Cox proportional-hazards model on the preliminary data, when this model was applied to the validation data they were able to achieve results that were significantly correlated with clinical outcome (117).

Genomic markers in breast cancer gave rise to specific molecular markers that could be used to tailor treatments to specific classes of breast cancer. However, the drugs that have resulted from genomic markers only apply to a subset of patients; additionally, cytotoxic chemotherapy has proven to be effective therapy for women at risk for development of recurrent disease (118, 119). However, estimates of the additional benefit of adjuvant chemotherapy range from a 7-11% increase in 10 year survival for women under age 50 and only 2-3% for women aged 50-69 (120). Gene expression profiling has been shown to predict patients that would benefit from adjuvant chemotherapy and has resulted in clinically utilized technologies. In an analysis of 98 breast cancer tumors in young women a multi-gene marker of 70 genes was constructed using supervised learning to distinguish patients that would develop distant metastases within five years and those who did not without adjuvant chemotherapy intervention; the multi-gene marker resulted in only a 5.3% misclassification rate on an independent validation set of 19 tumors suggesting that the predictor could identify

individuals who might possibly benefit from adjuvant chemotherapy and those who could be successfully treated without it (121). Based on similar work there have been a couple multi-gene platforms developed to help clinicians in making decisions about treatment in breast cancer. For example, *OncotypeDX* (122) is a 21 gene assay which is predictive of 10 year cancer recurrence for patients with early-stage ER positive and lymph node-negative breast cancer after hormone therapy, accounting for approximately 50% of diagnosed female breast cancers (94). Similarly, *MammaPrint* is another FDA approved test, consisting of a 70 gene signature that determines if a patient would benefit from adjuvant chemotherapy for all breast cancers (94, 123).

Application of *in vitro* cell lines to predict Chemosensitivity:

Current drug discovery methods rely on *in vitro* high-throughput screening to generate initial hypothesis about the clinical benefit of a compound (124). With respect to chemotherapy, prior to *in vitro* cell line screens, *in vivo* drug screens were done by implanting tumor cells in mice (13). From 1986 to 1990 the National Cancer Institute (NCI) had developed a panel of 60 immortal tumor cell lines, the NCI60 cell line panel (60). By the early 1990's the advancement of robust drug screening methods would lead to the first drugs, ellipticinium derivatives, to show anticancer effects using the NCI60 cell line panel (60). Cell line screens of anti-tumor compounds on the NCI60 cell line serve as a cornerstone of the NCI Developmental Therapeutics Program and serve as a fundamental tool in drug discovery and development (125). The success of the of the NCI60 drug screening program has inspired other cell line databases, such as Genomics of Drug Sensitivity in Cancer (GDSC) a panel of 987 cell lines screened with

367 different compounds (126) and the Cancer Cell Line Encyclopedia (CCLE) which includes 479 cell lines screened on 27 compounds (127).

One of the difficulties in cancer treatment is that tumors are composed of heterogenous subpopulations of cells, this is the motivation behind multi-drug treatment (14). The effectiveness of omics based models to predict effective personalized therapies relies on the ability to capture sufficient variability in omics based signatures that are explanatory over a range of drug sensitivities (128). *In vitro* cell line models provide an ideal platform where molecular variability is captured over a diverse set of cell lines and easily associated with drug response (129). As a modeling tool, *in vitro* drug screens are a valuable resource for model development and biomarker discovery, representing a large molecular diversity over several compounds. Thus, there has been a concerted effort to accurately model drug response in *in vitro* systems with the expectation that these models may become stepping-stones for developing more effective and diverse precision therapies in a clinical setting.

In vitro drug screens allow rapid model development and validation over hundreds of compounds without the need for large amounts of clinical or *in vivo* data. Staunton et al. took drug screening data from 232 compounds in the NCI60 and developed a classifier between sensitive and resistant cell lines (130). From an initial 6817 genes, a gene signature based on the ability of the gene to discriminate between sensitive and resistant cell lines, was constructed for each drug; classification was made based on weighted comparison of the gene expression signature to the signatures of the sensitive and resistant cell lines (130). Using an independent testing set they claimed that were able to generate statistically significant models for 88 of the

232 compounds (130). Potti et al. used a 50 gene signature and Bayesian binary regression to predict docetaxel response in the NCI60 with 74% accuracy (131). Lee et al. developed a method to extrapolate gene signatures in the NCI60 and applied to human bladder cancer tumors (132) which served as a basis for a recent clinical trial in human bladder cancer (133). Additional cell line data bases have been used in predictive models as well; Barretina et al. constructed predictive models in the CCLE for 24 anti-cancer agents and found several predictive gene-drug associations (127).

Computational methods in biology are rapidly developing as the collection, integration, and storage of large quantities of data become a focus of scientific and medical research. The complexity of the data is going to require creative and novel approaches; however, with vast amounts of tools and data an understanding of the different methodologies is paramount to continue to make progress. In a competitive format the NCI-Dream challenge evaluated 44 drug sensitivity prediction algorithms on a cohort of breast cancer cell lines (128). Given copy number variation, transcript expression values, mutation status, RNA sequencing data, DNA methylation, and reverse phase protein array (RPPA) data for 35 breast cancer cell lines, models were trained to predict drug sensitivity for 28 anti-cancer compounds and evaluated on an independent set of 18 cell lines (128). This was followed by a rigorous independent analysis on modeling approaches and the influence of different data modalities on overall performance (128).

The NCI-DREAM challenge provided several key insights to omics-based modeling. Models ranged from non-linear multitask learning, sparse linear regression, to simply using correlation based prediction; additionally, teams utilized anywhere between

one to all six datasets with some teams including specific known pathway information (128). Thirty four of the forty four models yielded prediction accuracies better than using a random permutation to rank drugs (128). The two top performing models utilized non-linear techniques; however, the third best performing model was strictly correlation based, indicating that model complexity was not necessary to build a top performing model (128). Furthermore, when the influence of different data types on model performance was analyzed, gene expression data was found to contribute most to model performance (128).

Chapter three is largely influenced by the NCI-DREAM and merits further discussion. One of the striking findings was the difference in performance between the best performing model and the third best model despite a stark contrast between model complexity. The best performing model used a multi-view, multi-task Bayesian multi-kernel learner (Bayes-MKL) (128). For simplicity, a kernel can be thought of as a group of functions that measures the similarity between two data points; a more detailed description of kernel functions is presented in chapter 2. The Bayes-MKL approach utilized all six of the given data sets and combined them into multiple data views by combing different data types into feature matrices (128). Each view was used to construct individual non-linear kernel functions then combined by linear weights derived from a Bayesian learning approach. The training was done simultaneously for all 28 drugs, referred to as multi-task learning (128). The third place model utilized gene expression, RNA seq, and RPPA data. Each feature was weighted according to its correlation with drug response and predictions were made using the correlations of the weighted features (128). Interestingly, the performance difference between these two

different strategies was only 2.2% (128). This highlights that a simple statistical method can perform better or comparable to a number of more complex modeling approaches and raises the question if more complex models capture complex data-drug relationships or if these methods are simply a more elaborate method of measuring simple statistical relationships? Either way, this demonstrates the importance of having a better understanding of how data is used to learn specific drug interactions and how they can be leveraged for more robust and accurate prediction of drug response.

The success of Bayes-MKL is representative of current trends in predicting drug response. There has been a focus on how to integrate different types of data for better overall prediction accuracies. This has been facilitated by an a continued interest in multi-kernel learning methods and the recent popularity of deep learning (134, 135). Chang et al. used a convolution neural network (CNN) to combine gene expression data from cancer cell lines and gene mutation status from a cohort of patient tumor samples for 244 drugs; additionally, they applied their model to 1487 approved drugs and were able to identify 37 possible new cancer treatments (136). Similarly, artificial neural networks (ANN) have been used as autoencoders to develop low dimensional representations of high dimensional data sets; for example, Li et al. employed an autoencoder to generate low dimensional representations of gene expression data then combined those representations with chemical structure data to predict drug response in the CCLE and GDSC (137). Furthermore, Ammad-un-din implemented kernelized Bayesian matrix factorization to integrate multiple data views in conjunction with MKL to predict drug response of cell lines in the GDSC (138). Also applying MKL, Cichonska et

al. used cell line data, drug characteristics, and protein level data to predict drug response in the GDSC (139).

Gene Perturbation and Drug Response

The vast majority of the molecular characterization of large-scale *in vitro* cell line databases is limited to a static or basal state. For example, the gene expression data that exist for these cells is limited to cells in generally growth conducive conditions. This is particularly limiting especially when trying to understand the underlying mechanisms of drug response. Drug response, itself, is a dynamic process that involves the interruption of several cellular processes. It is reasonable to associate the cytotoxic effect of a drug with the drugs ability to interrupt crucial cellular processes that cancer cells need to proliferate. This was largely the motivation for the connectivity map, to explore drug mechanisms with disease states, drug influence on drug induced physiological response, and relationships between different drugs based on drug mechanism (140). For example, it was hypothesized that if a disease state was associated with certain genetic differences from a non-diseased state, a drug which induces an opposing genetic response to the diseased state might be able to reverse the disease state on treatment. Similarly, if two different drugs resulted in similar genetic changes after treatment than those drugs might share similar mechanisms (140).

The first version of the connectivity map was released in 2006 by Lamb et al. and consisted of genetic profiles of up to four cancer cell lines perturbed by 164 different small-molecule perturbagens (140). The general concept was to develop multi-gene signatures for disease states and signatures for each drug and find the correlations that could be made between disease states and gene changes induced by drugs (140).

Initial applications of the connectivity map showed several instances of proof of concept. Particularly, they were able to similar signatures for HDAC inhibitors vorinostat and trichostatin; additionally, they were able to reverse dexamethasone resistance in the lymphoid cell line CEM-c1 by introducing treatment with sirolimus which was suggested by profile comparisons in the connectivity map (140). In 2017, the generation of an optimized gene expression assay of 978 genes, the L1000 referring to its 1058 probes, that could infer the expression levels in 81% of the additional gene transcripts generated profiles from the original 164 drugs to 19,881 small molecule drugs among 3 to 77 cell lines (141). Since its inception, the connectivity map has been used to generate hypotheses for new drug treatments (142-144), identify new pathways (145), and identify combination treatments (146, 147). Overall, the connectivity map showed that there was a clear relationship between disease, gene expression, and drug treatment.

Motivation:

The co-expression extrapolation (COXEN) trial in human bladder cancer was recently completed (133). The study was based around the COXEN algorithm which leveraged co-expression patterns between NCI60 cell lines bladder and breast tumors for feature selection and model training to predict drug sensitivity in bladder and breast cancer patients (132). The clinical trial looked at 237 bladder cancer patients which either received neoadjuvant chemotherapy regimens of methotrexate-vinblastine-adriamycin-cisplatin (MVAC) or gemcitabine-cisplatin (GC), COXEN then was used to extrapolate gene profiles to calculate a score that was predictive of tumor downstaging and the best chemotherapy treatment (133). The results of the study stated that the

COXEN score had no discernable significance in predicting successful chemotherapy regimen or downstaging looking specifically within the GC or MVAC treatment groups (133). However, the GC COXEN score for all patients, both patients receiving GC therapy and MVAC therapy, was a significant predictor of downstaging (133). One survey of the literature highlights an important question: why haven't these models that have astounding performance *in situ* not translated to clinically available tools? The statistical epidemiologist, Ioannidis, in a somewhat damning critique titled "*Why Most Published Research Findings are False"* suggest that this is a result of many factors, but two of particular interest he gives are "greater flexibility in designs, definitions, outcomes, and analytical modes" and the "chase of statistical significance" that has become almost a requirement for publication (148). The increase in computational capacity, data volume, and advances in machine learning, particularly deep learning, can capture complex phenomena; however, it is essential to keep in mind that complexity does not necessarily translate to utility.

The purpose of data driven models in oncology is the development of robust and interpretable clinical tools or, alternatively, vehicles that can drive knowledge and discovery acting to bridge the gap between large amounts of hard to interpret data and the biological phenomena underlying those data. *In vitro* models are often the initial step to determining drug cytotoxicity because they offer the most direct way to determine cytotoxicity on a cellular basis; additionally, large amounts of cells and drugs can be characterized relatively quickly and accurately. The simplicity of *in vitro* assays is a necessary proving ground for modeling approaches; if a model cannot predict cytotoxicity at a cellular level it is unlikely that it will be successful in tumors or patients

where factors such as tumor heterogeneity and patient physiology might become important factors. Additionally, in the absence of additional complexities such as the micro-tumor environment, the cytotoxicity can directly be attributed to molecular aspects of the cell. Therefore, modeling *in vitro* chemosensitivity is an essential step to establishing a predictive relationship between genomic features and cytotoxicity that is necessary for more complex models. Furthermore, without additional confounding factors, the association between gene expression and drug response is more clearly defined for biomarker discovery and hypothesis generation which can be leveraged in more complex models in tumors and patients. Therefore, the main focus of this work is to leverage *in vitro* drug response and gene expression data to systematically identify the relationships between modeling practices, gene expression, and drug response to provide a clear and precise foundation for innovations in model development and analysis.

The following chapter serves as a brief technical introduction into statistical and machine learning concepts and practices. The third chapter is largely motivated by the NCI-Dream challenge, providing a more focused view on the tradeoff between linear and non-linear approaches and biomarker selection using gene expression in the GDSC and the NCI60 cell line panels for supervised learning of drug response for cytotoxic chemotherapies. The fourth chapter explores the relationship between drug exposure, changes in gene expression, and drug response using supervised learning and network analysis. The fifth chapter uses genes identified from the network analysis to derive a basal signature and applies them to a drug response model of bladder cancer patients treated with a combination of gemcitabine and cisplatin.

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA: A Cancer Journal for Clinicians. 2020;70(1):7-30.

2. American Cancer Society. Lifetime Risk of Developing or Dying From Cancer 2020 [updated January 13, 2020; cited 2020 June 1]. Available from: https://www.cancer.org/cancer/cancer-basics/lifetime-probability-of-developing-or-dying-from-cancer.html.

3. Himmelstein DU, Warren E, Thorne D, Woolhandler S. Illness And Injury As Contributors To Bankruptcy. Health Affairs. 2005;24(Suppl1):W5-63-W5-73.

4. Zheng Z, Jemal A, Han X, Guy Jr GP, Li C, Davidoff AJ, et al. Medical financial hardship among cancer survivors in the United States. Cancer. 2019;125(10):1737-47.

5. Faguet GB. A brief history of cancer: Age-old milestones underlying our current knowledge database. International Journal of Cancer. 2015;136(9):2022-36.

6. American Cancer Society. Early Theories about Cancer Causes 2014 [updated June 12, 2014; cited 2020 June 16]. Available from: https://www.cancer.org/cancer/cancer-basics/history-of-cancer/cancer-causes-theories-throughout-history.html.

7. Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature. 1953;171(4356):737-8.

8. Hanahan D, Weinberg RA. The Hallmarks of Cancer. Cell. 2000;100(1):57-70.

9. Hanahan D, Weinberg Robert A. Hallmarks of Cancer: The Next Generation. Cell. 2011;144(5):646-74.

10. DeVita VT, Jr., Rosenberg SA. Two hundred years of cancer research. N Engl J Med. 2012;366(23):2207-14.

11. Röntgen WC. On a New Kind of Rays. Science. 1896;3(59):227-31.

12. Sherman AA. Translation of an Historic Paper: On a New, Strongly Radioactive Substance, Contained in Pitchblende: By M. P. Curie, Mme P. Curie and M. G. Bémont; Presented by M. Becquerel. Journal of Nuclear Medicine. 1970;11(6):269-70.

13. DeVita VT, Chu E. A History of Cancer Chemotherapy. Cancer Research. 2008;68(21):8643.

14. Weinberg RA. The Biology of Cancer 2nd ed. New York, NY: Garland Science 2014.

15. Visentin M, Zhao R, Goldman ID. The antifolates. Hematol Oncol Clin North Am. 2012;26(3):629-ix.

16. Farber S, Diamond LK. Temporary remissions in acute leukemia in children produced by folic acid antagonist, 4-aminopteroyl-glutamic acid. N Engl J Med. 1948;238(23):787-93.

17. Nussbaumer S, Bonnabry P, Veuthey J-L, Fleury-Souverain S. Analysis of anticancer drugs: A review. Talanta. 2011;85(5):2265-89.

18. Chabner BA, Roberts TG. Chemotherapy and the war on cancer. Nature Reviews Cancer. 2005;5(1):65-72.

19. National Cancer Institute. Cyclophosphamide 2007 [updated June 7, 2019; cited 2020 June 4]. Available from: https://www.cancer.gov/about-cancer/treatment/drugs/cyclophosphamide.

20. National Cancer Institute. Cisplatin 2007 [updated June 7th 2019; cited 2020 June 4]. Available from: https://www.cancer.gov/about-cancer/treatment/drugs/cisplatin.

21. National Cancer Institute. Azacytidine 2006 [updated December 17, 2018; cited 2020 June 4]. Available from: https://www.cancer.gov/about-cancer/treatment/drugs/azacitidine.

22. National Cancer Institute. Cytarabine 2007 [updated August 20, 2019; cited 2020 June 4]. Available from: https://www.cancer.gov/about-cancer/treatment/drugs/cytarabine.

23. National Cancer Institute. Gemcitabine Hydrochloride 2006 [updated August 16, 2019; cited 2020 June 4]. Available from: https://www.cancer.gov/about-cancer/treatment/drugs/gemcitabinehydrochloride.

24. National Cancer Institute. Fluorouracil Injection 2007 [updated July 5, 2018; cited 2020 June 4]. Available from: https://www.cancer.gov/about-cancer/treatment/drugs/fluorouracil.

25. National Cancer Institute. Flurouracil (Topical) 2016 [updated August 15, 2019; cited 2020 June 4]. Available from: https://www.cancer.gov/about-cancer/treatment/drugs/fluorouracil-topical.

26. American Cancer Society. Bleomycin Sulfate 2009 [updated May 15, 2019; cited 2020 June 4]. Available from: https://www.cancer.gov/about-cancer/treatment/drugs/bleomycin.

27. National Cancer Institute. Mitomycin 2011 [updated May 29, 2020; cited 2020 June 6]. Available from: https://www.cancer.gov/about-cancer/treatment/drugs/mitomycinc.

28. National Cancer Institute. Doxorubicin Hyrdrochloride 2007 [updated April 10, 2020; cited 2020 June 4]. Available from: https://www.cancer.gov/about-cancer/treatment/drugs/doxorubicinhydrochloride.

29. National Cancer Institute. Etoposide 2008 [updated July 19, 2019; cited 2020 June 4]. Available from: https://www.cancer.gov/about-cancer/treatment/drugs/etoposide.

30. National Cancer Institute. Topotecan Hydrochloride 2006 [updated March 9, 018; cited 2020 June 4]. Available from: https://www.cancer.gov/about-cancer/treatment/drugs/topotecanhydrochloride.

31. National Cancer Institute. Irinotecan Hydrochloride 2007 [updated July 25th, 2018; cited 2020 Aug 28th]. Available from: https://www.cancer.gov/about-cancer/treatment/drugs/irinotecanhydrochloride.

32. National Cancer Institute. Irinotecan Hydrochloride Liposome 2015 [updated March 28th, 2019; cited 2020 August 28th]. Available from: https://www.cancer.gov/about-cancer/treatment/drugs/irinotecan-hydrochloride-liposome.

33. National Cancer Institute. Docetaxel 2006 [updated July 1, 2019; cited 2020 June 4]. Available from: https://www.cancer.gov/about-cancer/treatment/drugs/docetaxel.

34. National Cancer Institute. Paclitaxel 2006 [updated March 12, 2020; cited 2020 June 4]. Available from: https://www.cancer.gov/about-cancer/treatment/drugs/paclitaxel.

35. National Cancer Institute. Vinblastine Sulfate 2011 [updated August 10, 2018; cited 2020 June 4]. Available from: https://www.cancer.gov/about-cancer/treatment/drugs/vinblastinesulfate.

36. National Cancer Institute. Bortezomib 2006 [updated March 9, 2018; cited 2020 July 9th, 2020]. Available from: https://www.cancer.gov/about-cancer/treatment/drugs/bortezomib.

37. National Cancer Institute. Vorinostat 2006 [updated March 19th, 2020; cited 2020 July 9th]. Available from: https://www.cancer.gov/about-cancer/treatment/drugs/vorinostat.

38. Colvin M. Alkylating Agents In: Donald W. Kufe REP, Ralph R. Weishselbaum, Robert C. Bast Jr, Test S Gansler, James F Holland, Emil Frei III editor. Holland-Frei Cancer Medicine 6ed. Hamilton (ON): BC Decker; 2003.

39. Bruce A. Chabner DLL. Cancer Chemotherapy and Biotherapy, Principles and Practice 3ed. Philadelphia, PA: Lippincott Williams and Wilkins 2001.

40. National Cancer Institute. Treatment of Solid Tumor Cancer with the Chemotherapy Drug Methotrexate 2014 [cited 2020 Jun e 6]. Available from: https://www.cancer.gov/research/progress/discovery/methotrexate.

41. Richard A. Messman CJA. Antifolates In: Bruce A. Chabner DLL, editor. Cancer Chemotherapy and Biotherapy: Principles and Practice 3rd ed. Philadelphia, PA: Lippincott Wiliams and Wilkins 2001. p. 139-84.

42. Parker WB. Enzymology of purine and pyrimidine antimetabolites used in the treatment of cancer. Chem Rev. 2009;109(7):2880-93.

43. Townsend AJ, Cheng YC. Sequence-specific effects of ara-5-aza-CTP and ara-CTP on DNA synthesis by purified human DNA polymerases in vitro: visualization of chain elongation on a defined template. Mol Pharmacol. 1987;32(3):330-9.

44. Plunkett W, Huang P, Searcy CE, Gandhi V. Gemcitabine: preclinical pharmacology and mechanisms of action. Semin Oncol. 1996;23(5 Suppl 10):3-15.

45. National Institute of Diabetes and Digestive Kidney Diseases. LiverTox: Clinical Research infromation on Drug-Induced Liver Injury. Cytotoxic Antibiotics 2012 [updated December 16, 2013 cited 2020 June 6]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK548321/.

46. Patel AG, Kaufmann SH. How does doxorubicin work? Elife. 2012;1:e00387-e.

47. Tacar O, Sriamornsak P, Dass CR. Doxorubicin: an update on anticancer molecular action, toxicity and novel drug delivery systems. Journal of Pharmacy and Pharmacology. 2013;65(2):157-70.

48. Jawad B, Poudel L, Podgornik R, Steinmetz NF, Ching W-Y. Molecular mechanism and binding free energy of doxorubicin intercalation in DNA. Physical Chemistry Chemical Physics. 2019;21(7):3877-93.

49. Thorn CF, Oshiro C, Marsh S, Hernandez-Boussard T, McLeod H, Klein TE, et al. Doxorubicin pathways: pharmacodynamics and adverse effects. Pharmacogenet Genomics. 2011;21(7):440-6.

50. Marinello J, Delcuratolo M, Capranico G. Anthracyclines as Topoisomerase II Poisons: From Early Studies to New Perspectives. International journal of molecular sciences. 2018;19(11):3480.

51. Denard B, Lee C, Ye J. Doxorubicin blocks proliferation of cancer cells through proteolytic activation of CREB3L1. Elife. 2012;1:e00090.

52. Crooke ST, Bradner WT. Mitomycin C: a review. Cancer Treatment Reviews. 1976;3(3):121-39.

53. Cummings J, Spanswick VJ, Tomasz M, Smyth JF. Enzymology of mitomycin C metabolic activation in tumour tissue: implications for enzyme-directed bioreductive drug development. Biochem Pharmacol. 1998;56(4):405-14.

54. Tomasz M. Mitomycin C: small, fast and deadly (but very selective). Chem Biol. 1995;2(9):575-9.

55. Suresh Kumar G, Lipman R, Cummings J, Tomasz M. Mitomycin C−DNA Adducts Generated by DT-Diaphorase. Revised Mechanism of the Enzymatic Reductive Activation of Mitomycin C. Biochemistry. 1997;36(46):14128-36.

56. Pan SS, Andrews PA, Glover CJ, Bachur NR. Reductive activation of mitomycin C and mitomycin C metabolites catalyzed by NADPH-cytochrome P-450 reductase and xanthine oxidase. Journal of Biological Chemistry. 1984;259(2):959-66.

57. Hodnick WF, Sartorelli AC. Reductive Activation of Mitomycin C by NADH:Cytochrome b₅ Reductase. Cancer Research. 1993;53(20):4907-12.

58. Gustafson DL, Pritsos CA. Bioactivation of Mitomycin C by Xanthine Dehydrogenase From EMT6 Mouse Mammary Carcinoma Tumors. JNCI: Journal of the National Cancer Institute. 1992;84(15):1180-5.

59. Sawamura AO, Aoyama T, Tamakoshi K, Mizuno K, Suganuma N, Kikkawa F, et al. Transfection of human cytochrome P-450 reductase cDNA and its effect on the sensitivity to toxins. Oncology. 1996;53(5):406-11.

60. Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. Nature Reviews Cancer. 2006;6:813.

61. Fitzsimmons SA, Workman P, Grever M, Paull K, Camalier R, Lewis AD. Reductase enzyme expression across the National Cancer Institute Tumor cell line panel: correlation with sensitivity to mitomycin C and EO9. J Natl Cancer Inst. 1996;88(5):259-69.

62. Kennedy KA, Rockwell S, Sartorelli AC. Preferential Activation of Mitomycin C to Cytotoxic Metabolites by Hypoxic Tumor Cells. Cancer Research. 1980;40(7):2356.

63. Strese S, Fryknäs M, Larsson R, Gullbo J. Effects of hypoxia on human cancer cell line chemosensitivity. BMC Cancer. 2013;13:331.

64. Kusumoto T, Maehara Y, Sakaguchi Y, Saku M, Sugimachi K. Hypoxia Enhances the Lethality of Mitomycin C and Carboquone against Human Malignant Tumor Cells in vitro. European Surgical Research. 1989;21(3-4):224-31.

65. Kim Y, Lin Q, Glazer PM, Yun Z. Hypoxic tumor microenvironment and cancer cell differentiation. Curr Mol Med. 2009;9(4):425-34.

66. Marquart J, Chen EY, Prasad V. Estimation of the Percentage of US Patients With Cancer Who Benefit From Genome-Driven Oncology. JAMA Oncol. 2018;4(8):1093-8.

67. Staker BL, Hjerrild K, Feese MD, Behnke CA, Burgin AB, Jr., Stewart L. The mechanism of topoisomerase I poisoning by a camptothecin analog. Proceedings of the National Academy of Sciences of the United States of America. 2002;99(24):15387-92.

68. Hande KR. Etoposide: four decades of development of a topoisomerase II inhibitor. European Journal of Cancer. 1998;34(10):1514-21.

69. Cooper GM. Microtubles 2000 June 7, 2020. In: The Cell: A Molecular Approach [Internet]. Sunderland, MA: Sinauer Associates 2nd. Available from: https://www.ncbi.nlm.nih.gov/books/NBK9932/.

70. Rowinsky EK, Donehower RC. The clinical pharmacology and use of antimicrotubule agents in cancer chemotherapeutics. Pharmacology & Therapeutics. 1991;52(1):35-84.

71. Kara G, Tuncer S, Türk M, Denkbaş EB. Downregulation of ABCE1 via siRNA affects the sensitivity of A549 cells against chemotherapeutic agents. Medical Oncology. 2015;32(4):103.

72. Li W, Zhang H, Assaraf YG, Zhao K, Xu X, Xie J, et al. Overcoming ABC transporter-mediated multidrug resistance: Molecular mechanisms and novel therapeutic drug strategies. Drug Resistance Updates. 2016;27:14-29.

73. Cowin PA, George J, Fereday S, Loehrer E, Van Loo P, Cullinane C, et al. LRP1B deletion in high-grade serous ovarian cancers is associated with acquired chemotherapy resistance to liposomal doxorubicin. Cancer Res. 2012;72(16):4060-73.

74. Assaraf YG. Molecular basis of antifolate resistance. Cancer Metastasis Rev. 2007;26(1):153-81.

75. Gottesman MM. Mechanisms of Cancer Drug Resistance. Annual Review of Medicine. 2002;53(1):615-27.

76. Konishi H, Usui T, Sawada H, Uchino H, Kidani Y. Effects of anticancer platinum compounds on Escherichia coli strains with normal and defective DNA repair capacity. Gann. 1981;72(4):627-30.

77. Fram RJ, Cusick PS, Marinus MG. Studies on mutagenesis and repair induced by platinum analogs. Mutation Research Letters. 1986;173(1):13-8.

78. Mini E, Nobili S, Caciagli B, Landini I, Mazzei T. Cellular pharmacology of gemcitabine. Ann Oncol. 2006;17 Suppl 5:v7-12.

79. OHHASHI S, OHUCHIDA K, MIZUMOTO K, FUJITA H, EGAMI T, YU J, et al. Down-regulation of Deoxycytidine Kinase Enhances Acquired Resistance to Gemcitabine in Pancreatic Cancer. Anticancer Research. 2008;28(4B):2205-12.

80. Cabral FR, Brady RC, Schibler MJ. A mechanism of cellular resistance to drugs that interfere with microtubule assembly. Ann N Y Acad Sci. 1986;466:745-56.

81. Jaffrézou JP, Chen KG, Durán GE, Kühl JS, Sikic BI. Mutation rates and mechanisms of resistance to etoposide determined from fluctuation analysis. J Natl Cancer Inst. 1994;86(15):1152-8.

82. Li R, Moudgil T, Ross HJ, Hu HM. Apoptosis of non-small-cell lung cancer cell lines after paclitaxel treatment involves the BH3-only proapoptotic protein Bim. Cell Death Differ. 2005;12(3):292-303.

83. Smith L, Watson MB, O'Kane SL, Drew PJ, Lind MJ, Cawkwell L. The analysis of doxorubicin resistance in human breast cancer cells using antibody microarrays. Mol Cancer Ther. 2006;5(8):2115-20.

84. Hanane A, Claire L, Caroline A, Fabrice M. Anticancer Drug Metabolism: Chemotherapy Resistance and New Therapeutic Approaches. 2012.

85. Bukowski K, Kciuk M, Kontek R. Mechanisms of Multidrug Resistance in Cancer Chemotherapy. International Journal of Molecular Sciences. 2020;21.

86. Bhullar KS, Lagarón NO, McGowan EM, Parmar I, Jha A, Hubbard BP, et al. Kinase-targeted cancer therapies: progress, challenges and future directions. Molecular Cancer. 2018;17(1):48.

87. Roskoski R. Properties of FDA-approved small molecule protein kinase inhibitors: A 2020 update. Pharmacological Research. 2020;152:104609.

88. Troy AB. Targeted Cancer Therapy: The Next Generation of Cancer Treatment. Current Drug Discovery Technologies. 2015;12(1):3-20.

89. Rosenzweig SA. Acquired Resistance to Drugs Targeting Tyrosine Kinases. Adv Cancer Res. 2018;138:71-98.

90. He M, Wei MJ. Reversing multidrug resistance by tyrosine kinase inhibitors. Chin J Cancer. 2012;31(3):126-33.

91. Neel DS, Bivona TG. Resistance is futile: overcoming resistance to targeted therapies in lung adenocarcinoma. npj Precision Oncology. 2017;1(1):3.

92. Goossens N, Nakagawa S, Sun X, Hoshida Y. Cancer biomarker discovery and validation. Transl Cancer Res. 2015;4(3):256-69.

93. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. Clinical Pharmacology & Therapeutics. 2001;69(3):89-95.

94. Lukong KE. Understanding breast cancer - The long and winding road. BBA Clin. 2017;7:64-77.

95. Freedman LS, Edwards DN, McConnell EM, Downham DY. Histological grade and other prognostic factors in relation to survival of patients with breast cancer. British Journal of Cancer. 1979;40(1):44-55.

96. Rakha EA, Reis-Filho JS, Baehner F, Dabbs DJ, Decker T, Eusebi V, et al. Breast cancer prognostic classification in the molecular era: the role of histological grade. Breast Cancer Research. 2010;12(4):207.

97. Patel C, Sidhu KP, Shah MJ, Patel SM. Role of mitotic counts in the grading and prognosis of the breast cancer. Indian J Pathol Microbiol. 2002;45(3):247-54.

98. Elston CW. The assessment of histological differentiation in breast cancer. Aust N Z J Surg. 1984;54(1):11-5.

99. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016;107(1):1-8.

100. Yamamoto T, Ikawa S, Akiyama T, Semba K, Nomura N, Miyajima N, et al. Similarity of protein encoded by the human c-erb-B-2 gene to epidermal growth factor receptor. Nature. 1986;319(6050):230-4.

101. Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science. 1987;235(4785):177-82.

102. Schena M, Shalon D, Davis RW, Brown PO. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. Science. 1995;270(5235):467.

103. 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.

104. 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 U S A. 2001;98(19):10869-74.

105. Al-Thoubaity FK. Molecular classification of breast cancer: A retrospective cohort study. Ann Med Surg (Lond). 2020;49:44-8.

106. Wakeling AE. Similarities and distinctions in the mode of action of different classes of antioestrogens. Endocr Relat Cancer. 2000;7(1):17-28.

107. Burstein HJ, Prestrud AA, Seidenfeld J, Anderson H, Buchholz TA, Davidson NE, et al. American Society of Clinical Oncology clinical practice guideline: update on adjuvant endocrine therapy for women with hormone receptor-positive breast cancer. J Clin Oncol. 2010;28(23):3784-96.

108. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science. 1999;286(5439):531.

109. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000;403:503.

110. Bishop CM. Pattern Recognition and Machine Learning. New York, New York Springer; 2006.

111. Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine. 2001;7:673.

112. Xu Y, Selaru FM, Yin J, Zou TT, Shustova V, Mori Y, et al. Artificial neural networks and gene filtering distinguish between global gene expression profiles of Barrett's esophagus and esophageal cancer. Cancer Res. 2002;62(12):3493-7.

113. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics. 2000;16(10):906-14.

114. Su AI, Welsh JB, Sapinoso LM, Kern SG, Dimitrov P, Lapp H, et al. Molecular Classification of Human Carcinomas by Use of Gene Expression Signatures. Cancer Research. 2001;61(20):7388.

115. Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature. 2002;415(6870):436-42.

116. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RCT, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature Medicine. 2002;8:68.

117. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, et al. The Use of Molecular Profiling to Predict Survival after Chemotherapy for Diffuse Large-B-Cell Lymphoma. New England Journal of Medicine. 2002;346(25):1937-47.

118. Sachdev JC, Jahanzeb M. Use of Cytotoxic Chemotherapy in Metastatic Breast Cancer: Putting Taxanes in Perspective. Clinical Breast Cancer. 2016;16(2):73-81.

119. Lee SH, Falkenberry SS. Chapter 21 - Conditions of the Female Breast. In: Sokol AI, Sokol ER, editors. General Gynecology. Philadelphia: Mosby; 2007. p. 497-521.

120. Polychemotherapy for early breast cancer: an overview of the randomised trials. The Lancet. 1998;352(9132):930-42.

121. 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.

122. Kaklamani V. A genetic signature can predict prognosis and response to therapy in breast cancer: Oncotype DX. Expert Review of Molecular Diagnostics. 2006;6(6):803-9.

123. Morris SR, Carey LA. Gene expression profiling in breast cancer. Current Opinion in Oncology. 2007;19(6).

124. Blass BE. Chapter 4 - In vitro Screening Systems. In: Blass BE, editor. Basic Principles of Drug Discovery and Development. Boston: Academic Press; 2015. p. 143-202.

125. National Cancer Institute. NCI-60 Human Tumor Cell Lines Screen 2015 [updated August 26, 2015; cited 2020 June 25]. June 25]. Available from: https://dtp.cancer.gov/discovery_development/nci-60/default.htm.

126. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Research. 2013;41(D1):D955-D61.

127. 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.

128. Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nature biotechnology. 2014;32(12):1202-12.

129. Kao J, Salari K, Bocanegra M, Choi YL, Girard L, Gandhi J, et al. Molecular profiling of breast cancer cell lines defines relevant tumor models and provides a resource for cancer gene discovery. PLoS One. 2009;4(7):e6146.

130. Staunton JE, Slonim DK, Coller HA, Tamayo P, Angelo MJ, Park J, et al. Chemosensitivity prediction by transcriptional profiling. Proceedings of the National Academy of Sciences. 2001;98(19):10787.

131. Potti A, Dressman HK, Bild A, Riedel RF, Chan G, Sayer R, et al. Genomic signatures to guide the use of chemotherapeutics. Nature Medicine. 2006;12(11):1294-300.

132. Lee JK, Havaleshko DM, Cho H, Weinstein JN, Kaldjian EP, Karpovich J, et al. A strategy for predicting the chemosensitivity of human cancers and its application to drug discovery. Proceedings of the National Academy of Sciences of the United States of America. 2007;104(32):13086-91.

133. Flaig TW, Tangen CM, Daneshmand S, Alva AS, Lerner SP, Lucia MS, et al. SWOG S1314: A randomized phase II study of co-expression extrapolation (COXEN) with neoadjuvant chemotherapy for localized, muscle-invasive bladder cancer. Journal of Clinical Oncology. 2019;37(15_suppl):4506-.

134. Adam G, Rampášek L, Safikhani Z, Smirnov P, Haibe-Kains B, Goldenberg A. Machine learning approaches to drug response prediction: challenges and recent progress. NPJ Precis Oncol. 2020;4:19.

135. Ali M, Aittokallio T. Machine learning and feature selection for drug response prediction in precision oncology applications. Biophysical Reviews. 2018.

136. Chang Y, Park H, Yang H-J, Lee S, Lee K-Y, Kim TS, et al. Cancer Drug Response Profile scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature. Scientific Reports. 2018;8(1):8857.

137. Li M, Wang Y, Zheng R, Shi X, y l, Wu F, et al. DeepDSC: A Deep Learning Method to Predict Drug Sensitivity of Cancer Cell Lines. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2019:1-.

138. Ammad-ud-din M, Khan SA, Malani D, Murumägi A, Kallioniemi O, Aittokallio T, et al. Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization. Bioinformatics. 2016;32(17):i455-i63.

139. Cichonska A, Pahikkala T, Szedmak S, Julkunen H, Airola A, Heinonen M, et al. Learning with multiple pairwise kernels for drug bioactivity prediction. Bioinformatics. 2018;34(13):i509-i18.

140. 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. 2006;313(5795):1929.

141. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell. 2017;171(6):1437-52.e17.

142. Cheng HW, Liang YH, Kuo YL, Chuu CP, Lin CY, Lee MH, et al. Identification of thioridazine, an antipsychotic drug, as an antiglioblastoma and anticancer stem cell agent using public gene expression data. Cell Death Dis. 2015;6(5):e1753.

143. Yuen T, Iqbal J, Zhu LL, Sun L, Lin A, Zhao H, et al. Disease-drug pairs revealed by computational genomic connectivity mapping on GBA1 deficient, Gaucher disease mice. Biochem Biophys Res Commun. 2012;422(4):573-7.

144. Toscano MG, Navarro-Montero O, Ayllon V, Ramos-Mejia V, Guerrero-Carreno X, Bueno C, et al. SCL/TAL1-mediated transcriptional network enhances megakaryocytic specification of human embryonic stem cells. Mol Ther. 2015;23(1):158-70.

145. Sanda T, Li X, Gutierrez A, Ahn Y, Neuberg DS, O'Neil J, et al. Interconnecting molecular pathways in the pathogenesis and drug sensitivity of T-cell acute lymphoblastic leukemia. Blood. 2010;115(9):1735-45.

146. Lee JH, Kim DG, Bae TJ, Rho K, Kim JT, Lee JJ, et al. CDA: combinatorial drug discovery using transcriptional response modules. PLoS One. 2012;7(8):e42573.

147. Huang L, Li F, Sheng J, Xia X, Ma J, Zhan M, et al. DrugComboRanker: drug combination discovery based on target network analysis. Bioinformatics. 2014;30(12):i228-36.

148. Ioannidis JPA. Why Most Published Research Findings Are False. PLOS Medicine. 2005;2(8):e124.

CHAPTER 2: TECHNICAL REVIEW

The Problem of Learning

Chapters 3, 4, and 5 explore how statistical learning techniques can be used to infer the relationship between gene expression and drug response for *in vitro* cell lines. Tibshirani *et al.* defines statistical learning as the construction of a statistical model *"for predicting, or estimating an output based on one or more inputs"* (1). Thus, in the context of this work the input is gene expression and output is drug response. While a "statistical model" is a generalized term for a number of tools and techniques; central to all statistical models is their dependence on data. Thus, all statistical models leverage statistical patterns to estimate the relationship between input and output variables based on a given set of observations, i.e. data, and applies these relationships to additional inputs with the goal of accurately generating an output.

What follows a general mathematical formulism for the learning problem. First define the input space as $\boldsymbol{X}$ and the output space as $\boldsymbol{Y}$, specifically concerning the models in the following chapters, $\boldsymbol{X}$, is the gene expression profile for all possible cells; likewise, $\boldsymbol{Y}$, is all possible values for drug responses of those cells. Members of $\boldsymbol{X}$ are given as vectors $\boldsymbol{x}_i = \{x_1, \dots, x_D\}$ and for simplicity we will assume $x_d \ni \mathbb{R}$; however, noting that x_d can be any numerical, ordinal, or categorical value. Additionally, the members of $\boldsymbol{Y}$ are given as vectors $\boldsymbol{y}_i = \{y_1, \dots, y_d,\}$, again for simplicity we will assume the members $\boldsymbol{Y}$ are real value scalers such that $\boldsymbol{y}_i = y_i, y_i \ni \mathbb{R}$ while noting that $\boldsymbol{y}_i$ can be of any dimension and y_i can be any numerical, ordinal, or categorical value. Additionally, we will assume that there is some function such that $f: \boldsymbol{x}_i \rightarrow y_i \ \forall \ \boldsymbol{x}_i, y_i$.

Now the objective of learning is to learn the function f given a paired subset $\{X_N, Y_N\}$ where $X_N \supset \boldsymbol{X}, Y_N \supset \boldsymbol{Y}$. However, given that we are given a limited amount of data we are relegated to estimating $\hat{f}$ such that the metric $\left|\hat{f}: \boldsymbol{x_n} - f: \boldsymbol{x}_n\right|$ is minimized for all $\boldsymbol{x}_n \ni X_N$. In order to do this we define a loss function

$$h(\boldsymbol{x}_n) := \begin{cases} 0 \; if \; \left|\hat{f}: \boldsymbol{x_n} - y_n\right| = 0 \\ z > 0 \; if \left|\hat{f}: \boldsymbol{x_n} - y_n\right| \neq 0 \end{cases}$$

then we minimize the sum over all X_N

$$\min_{\hat{f}} E := \sum_{i=1}^{N} h(\boldsymbol{x}_i)$$

Gene Expression Modeling: The Basics

Figure 2.1 shows the general workflow that is followed for model construction. The input is gene expression microarray data for an *in vitro* cell line and the output is the concentration of drug which inhibits the growth of that cell line by 50% (IC_{50}/GI_{50}). Gene expression is measured in microarray experiments; first, mRNA is extracted from the cell line. This is followed by reverse transcription of the mRNA into fluorescently labeled complementary DNA (cDNA) via reverse transcriptase and then transferred to a microarray chip, referred to as DNA hybridization. Each microarray contains multiple sets of tens of thousands small oligonucleotide probes (~20 bases) complementary to conserved regions of the cDNA for each gene. Gene expression is quantified by the number of gene specific probes hybridized to the chip measured by the fluorescent intensity of the gene specific region of the microarray. Drug response is quantified by a cohort of experiments where different concentrations of drug are exposed to fluorescently labeled *in vitro* cell line for a fixed amount of time the growth response is then measured as a fraction of the baseline value. The IC_{50} is typically extrapolated

from a sigmoidal fit to the data and reported with a log scale. For all the work presented in chapters 3,4, and 5 the data was downloaded from publicly available databases.

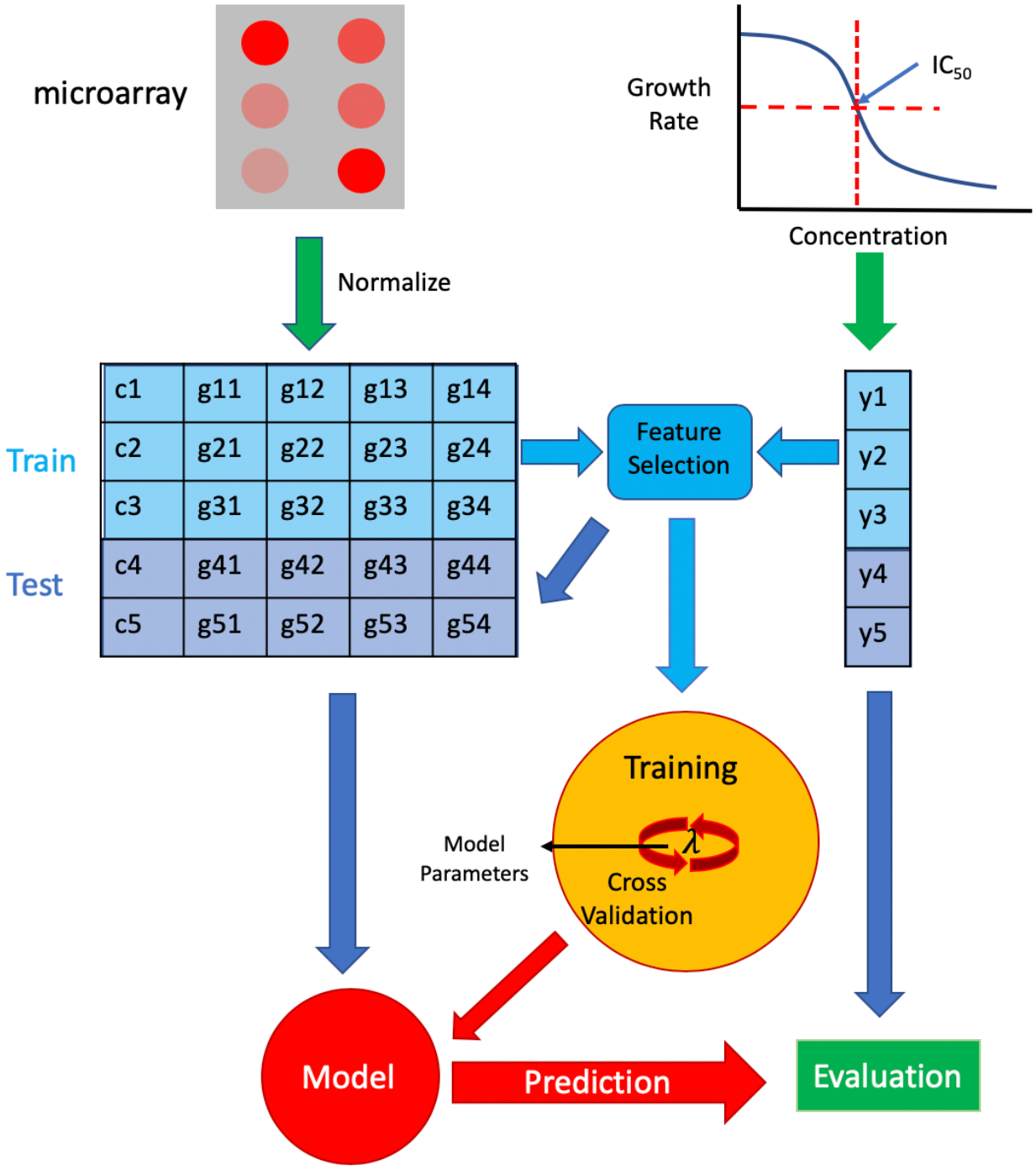


Figure 2.1. Workflow of a general model including data acquisition, feature selection (if applicable), training including parameter optimization (if applicable), and model evaluation.

Each microarray experiment pertains to a single cell line and as discussed, gene expression is quantified based on relative fluorescent intensity which is proportional to the expression of that gene. However, as each microarray is a separate experiment it is subject to random experimental effects; thus, comparison of two different microarray experiments might not be biologically relevant (2). Therefore, it is necessary to adjust the intensity values such meaningful comparisons can be made, referred to as normalization of the data. There are several methods of proposed data normalization techniques (3, 4). Chapters 3, 4, and 5 use a popular method known as robust multi-array-average (RMA) (5) or a variation of RMA called frozen RMA (fRMA) (6). After RMA has been applied gene expression is then reported as a log scaled intensity value which comprises the gene expression matrix (5).

The utility of a model is based on predictive capabilities to a broad range of input data. However, models are constructed on a finite subset of observations which is typically a small subset of all possible observations. In order to estimate the generalizability of the model to a general population the model has to be evaluated on observations independent of the observations the model was constructed. This is accomplished by partitioning the data into training and testing subsets. Again let $\{\boldsymbol{X}, \boldsymbol{Y}\}$ represent the space of all gene expression profiles with the corresponding drug response with the assumption that there is some underlying statistical relationship such that there is a function $f: \boldsymbol{X} \rightarrow \boldsymbol{Y}$. We are limited to a subset of observations $\{X, Y\} \supset \{\boldsymbol{X}, \boldsymbol{Y}\}$ to estimate f. Furthermore, we define the training set to be a subset $\{X_N, Y_N\} \supset \{X, Y\}$ and the testing set to be the complement $\{X_N^c, Y_N^c\}$ of the training set. Generally speaking X is sometimes referred to as the data and Y is sometimes referred

to as the target value. The function, $\hat{f}$, is estimated, or trained, on the training set and then evaluated on the test set to estimate if the model generalizes well to random observation of $\{\boldsymbol{X}, \boldsymbol{Y}\}$. Regression models can be evaluated by a number of statistics including R^2, Pearson correlation, Spearman correlation, root mean squared error (RMSE), or mean absolute difference (MAD). Classification model are typically evaluated by classification accuracy, sensitivity, and specificity. The performance of a model on a test set estimates the predictive accuracy when applied to a random input of $\boldsymbol{X}$.

Often in practice, there is variation in model performance that depends on the individual samples within the testing and training set, this the result of overfitting which will be addressed later. Thus, to quantify the variation due to random sampling the model is trained on several different partitions of the data into training and test sets and the model is evaluated by the average performance on all test sets. This procedure is referred to a cross validation. Additionally, many methods such a regularized regression techniques require a predetermined parameter to be specified; however, it often unknown what parameter will lead to the best performance on the test set. Thus, often cross validation is performed during training on subsets of the training set as the parameter or parameters are systematically changed. The parameter set that performs best overall cross-validation sets is selected then the model is re-trained on all data and then evaluated on the independent test set.

Least Squares Regression

Least squares regression is the standard method that is used to optimize over a linear equation:

$$\hat{f}: \boldsymbol{x}_i = \sum_{j=0}^{M} w_j x_j \quad (2.1)$$

where y_i is the target value for sample i, $\boldsymbol{w}^T = \langle w_0, w_1, \dots, w_M \rangle$ where M is the dimensionality, and $\boldsymbol{x}_i^T = \langle x_0, x_1, \dots, x_M \rangle$ where x_j is the j[th] variable for sample i. Also note that $\boldsymbol{X}^T = \langle \boldsymbol{x}_1, \dots, \boldsymbol{x}_N \rangle$ where N is the total number of samples, additionally $w_0 = b$ and $x_0 = 1$ to make a more compact form of the standard linear equation $y = mx + b$. Furthermore, with respect to the formalism that was presented above, $f: \boldsymbol{x}_i = y_i$ and:

$$h(\boldsymbol{x}_i) = \frac{1}{2}(\boldsymbol{w}^T \boldsymbol{x}_i - y_i)^2 \quad (2.2)$$

therefore least squares regression is defined as the optimization over the following error function:

$$E: \boldsymbol{w} = \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{w}^T \boldsymbol{x}_i - y_i)^2 \quad (2.3)$$

the optimal solution is found by minimizing the error function with respect to $\boldsymbol{w}$:

$$\frac{dE}{d\boldsymbol{w}} = 0 = \sum_{i=1}^{N} \boldsymbol{x}_i^T(\boldsymbol{w}^T \boldsymbol{x}_i - y_i) \quad (2.4)$$

Converting this to matrix form gives the following solution

$$\boldsymbol{w} = \frac{Y\boldsymbol{X}^T}{\boldsymbol{X}^T\boldsymbol{X}} \quad (2.5)$$

where $Y^T = \langle y_1, \dots, y_n \rangle$.

Overfitting:

Generally the goal of learning is to develop a method that is generalizable to all possible values of input data. Overfitting results when a model is very accurate on a training set but does not generalize very well to data outside the training data. Many factors contribute to overfitting such as noise resulting from data acquisition, the number of data points available, and the dimensionality of the data. Omics data are susceptible to all three of these conditions; systematic error and random error will always be present in experimentally collected biological data and the number of data points in a given data set is far exceeded by the number of genes. An additional complication that arises is there is a large co-dependence between genes as several genes may be components in a single biological function, such as a pathway. With respect to dimensionality, generally the training error continuously falls with increasing dimensions and testing falls until it arrives at a minimum additional variables often result in increased error. The effect of dimensionality on model accuracy has been termed “the curse of dimensionality”.

The Curse of Dimensionality:

The total number of human protein-coding genes is not entirely known (7). Current estimates put it somewhere between 20,000 and 25,000 (8). Thus, a genome as a whole is a high dimensional system in comparison to the dimensionality of many other datasets. The phrase “the curse of dimensionality” can be traced back to 1957 and is attributed to Richard Bellman (9). The curse of dimensionality basically refers to the sparsity that arises from finite sampling of data in high dimensional space, particularly when the dimensions greatly exceed the number of data points (10). The

notion of distance is a fundamental concept in machine learning; for instance, in LSQR is can be shown that the optimal solution is the solution which minimizes the Euclidean distance between data points and a point on the best fit line. For example, if we take two random points whose coordinate value on each dimension is a uniform random variable from 0 to 1, increasing the number of dimensions results in the points becoming farther apart (Figure 2.2 A.). Conceptually, the consequences of this can be demonstrated with a simple regression algorithm where we are given a handful of data and target values, a simple way of estimating the value of a new point is just to assign it the same value as the closest point in the data that's been observed. If the majority of the variability in the target value is the result of a small subset of the total variables but the distance measured is going to depend on the total dimensionality, there is a possibility that values that differ substantially in meaningful variables are closer together than points which are closer with respect to their target value. Furthermore, as the dimensions grow to a substantially larger number of points, the points will become effectively equidistant and then noise could have a much bigger effect on the accuracy (Figure 2.2 B-D).

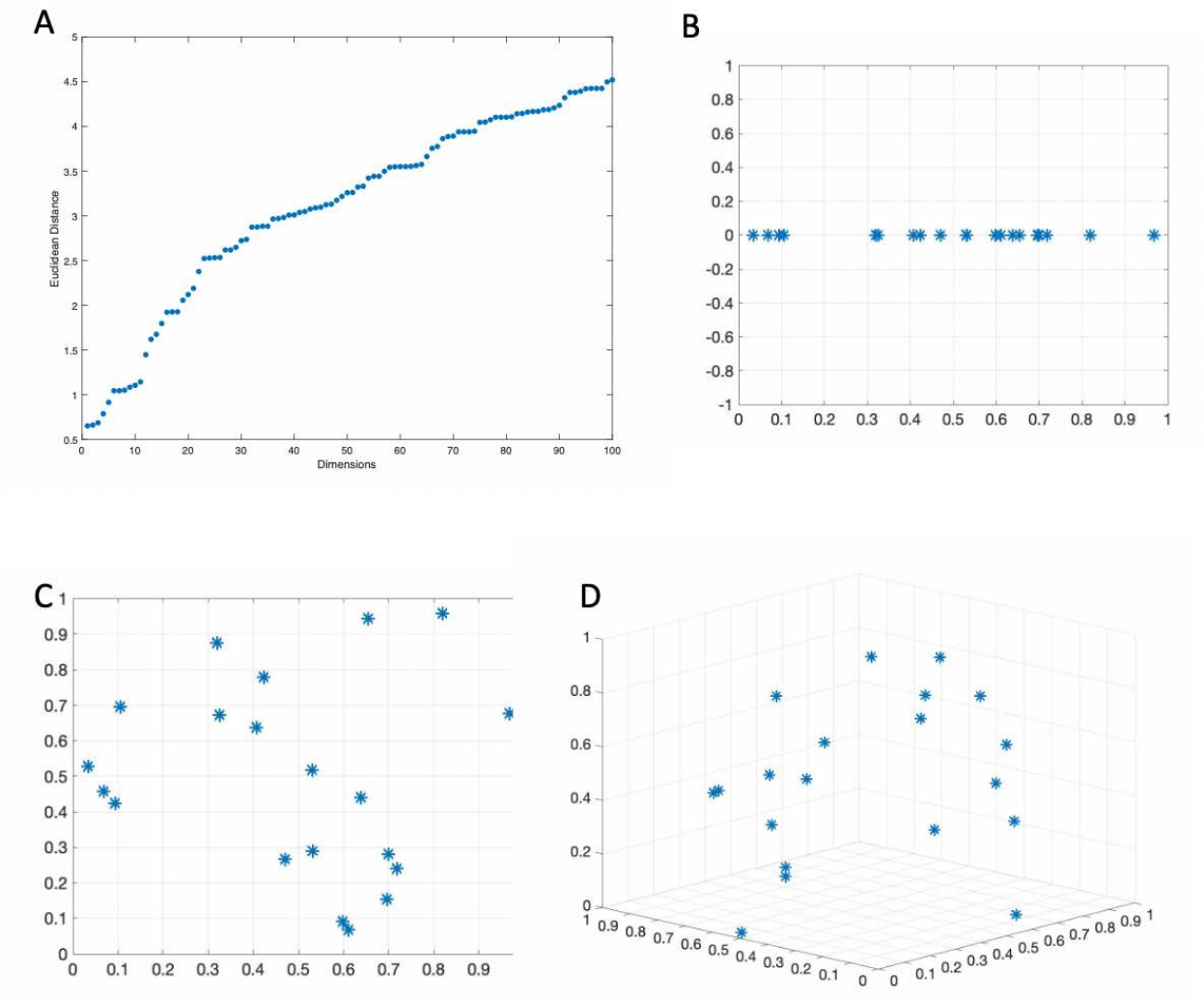


Figure 2.2: A. As the number of the dimensions goes up the distance between two points increase as well. B-D As the number of dimensions increase from 1 (B), 2 (C), 3 (D) the number of points that occupy a d dimensional unit goes from 2 to 0.2 to 0.02

Techniques for Dimensionality Reduction:

There are two ways of mitigating complications that arise from high dimensionality; the first is to increase the number of data points such that the number of data points and the number of dimensions are comparable. However, in omics modeling the number of variables often exceeds the number of data points by several orders of magnitude, thus increasing the data that substantially is not practical. Furthermore, the

curse of dimensionality can be thought of as a combinatorial problem; as dimensions are added the possible locations for a points in space increases exponentially (Figure 2.2 b-d). Thus, by reducing the number of dimensions the possible location of the point decreases dramatically; therefore, the data occupy a much more dense space where the relationships between points can be well defined, which is essential to learning tasks. While dimensionality reduction can be approached from a number of ways, such as various methods of manifold learning, for the purposes of this text the discussion is limited to Principal Components Analysis and feature selection.

Principal Components Analysis:

Principal components analysis (PCA) is a linear dimensionality reduction technique that projects multidimensional data onto a lower dimensional space in such a manner that the variance of the projected data is maximized (10). With respect to genomics modeling where the number of dimensions usually exceeds the number of data points, PCA allows for substantial dimensional reduction while also maintaining the maximum amount of variance. The follow derivation closely follows that of Bishop's (10). First, let each of the N data points be represented by an M dimensional vector, $\boldsymbol{x}_n$. Define another M dimensional vector, $\boldsymbol{u}_1$. Now the objective of PCA is to find the direction of $\boldsymbol{u}_1$ such that when all $\boldsymbol{x}_n$ are projected onto $\boldsymbol{u}_1$ the variance of the projected data is maximized. By definition the variance along $\boldsymbol{u}_1$ is given by

$$\frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{u}_1^T\boldsymbol{x}_n - \boldsymbol{u}_1^T\overline{\boldsymbol{x}})^2 = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{u}_1^T(\boldsymbol{x}_n - \overline{\boldsymbol{x}})(\boldsymbol{x}_n - \overline{\boldsymbol{x}})\boldsymbol{u}_1 = \boldsymbol{u}_1^T\boldsymbol{S}\boldsymbol{u}_1 \quad (2.6)$$

Where $\boldsymbol{S}$ is the covariance matrix of with respect $\boldsymbol{x}_n$ not projected onto $\boldsymbol{u}_1$. Therefore to find the maximum of $\boldsymbol{u}_1^T\boldsymbol{S}\boldsymbol{u}_1$ with respect to $\boldsymbol{u}_1$ the method of Lagrange multipliers is

used to prevent $\|\boldsymbol{u}_1\| \to \infty$. The constraint $\boldsymbol{u}_1^T\boldsymbol{u}_1 = 1$ allows for maximization of the following with Lagrange multiple λ_1

$$\boldsymbol{u}_1^T\boldsymbol{S}\boldsymbol{u}_1 + \lambda_1(1 - \boldsymbol{u}_1^T\boldsymbol{u}_1) \quad (2.7)$$

taking the derivative of this and setting it equal to zero the following solution is obtained

$$\boldsymbol{S}\boldsymbol{u}_1 = \lambda_1\boldsymbol{u}_1 \rightarrow \boldsymbol{u}_1^T\boldsymbol{S}\boldsymbol{u}_1 = \lambda_1$$

Therefore the projection of all $\boldsymbol{x}_n$ onto the vector $\boldsymbol{u}_1$ has the maximum variance of value λ_1. Further principal components, $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_l$ $l \leq N$, can be similarly defined and the process can be repeated, with the additional requirement $\boldsymbol{u}_l$ are orthogonal. The benefit of PCA is that often the majority of variability in high dimensional data can be captured on a lower dimensional subspace. One drawback of PCA is the number of components can be not exceed the number of data points. In high-dimensional problems the number of dimensions is much greater than the number of data points, $\boldsymbol{u}_d \leq \boldsymbol{u}_N$. Therefore, the low dimensional projection of data given by PCA might not be sufficient to capture meaningful variability that is only present in a higher dimensions.

Feature Selection:

PCA can greatly reduce the dimensionality of a dataset. The ability of the principal components to represent the variance of the data in a lower dimensional space can have a dramatic effect on overfitting in a regression of classification task; however, the direction of the principal components are oriented to maximize variance over the entire dataset, not necessarily with respect to a target value in a learning problem. For example, consider spheres of variable volumes and masses in a constant gravitational field all with the same height above the ground. The potential energy of the sphere only depends on its mass if the height is fixed; however, representing each sphere as a two

dimensional data point the, the first principal component would be oriented such that it captures maximum variability with respect to mass and volume. Therefore, if I am trying to learn a function for potential energy given noisy data, and use PCA to reduce the dimension of my data from two to one, there is a possibility of overfitting especially if mass and volume are independent. In order to minimize error on a testing set, it might be best to only use the mass of the spheres for learning. Such a practice is referred to as feature selection in learning applications.

In a model construction, contextual knowledge might be used to select out important features prior to model building; for example, if a model is constructed to estimate the amount of some molecular byproduct in a cell based on gene expression and it is known what enzymes carry out this process, then all genes can be eliminated except for the ones coding for those enzymes. However, with respect to drug response in tumors or cancer cells mechanisms that determine response might not be entirely known, even for targeted agents the mechanisms of resistance might not be fully understood. It is therefore necessary to reduce features based on the data itself. Data based feature selection is broadly categorized into three categories, filter, wrapper, and embedded methods. For the purposes of this review only filter methods are discussed; however, a substantial amount of literature is devoted to feature selection in its entirety both in the context of bioinformatics and other statistical learning applications (11, 12).

The defining feature of filter methods is that features are selected prior to training the model. Often genomic models involve discriminating between two or more states, such as sensitive or resistant in drug response. One technique filters out genes based on the difference in fold change between the two groups (13). Likewise, a two-sample t-

test can be used to determine if a gene belongs to the same distribution in both groups. However, the validity of a t-test relies on the assumption that the random variable follows a normal distribution that has a well-defined mean and variance. In the case of small sample size,s the mean and variance are not well defined and thus a t-test is poorly suited; however, methods using modified t-tests have been implemented with genomic data. For example, the R package limma uses a modified t-statistic that allows for parallel estimation of parameters by “borrowing” information from other genes (14). Additionally, non-parametric methods such as the Wilcoxon rank-sum test have been used to deal with non-normality assumptions (15). Other methods include the use of mutual information to determine relevant features (16).

The above methods are all with respect to discrete classes; however, chapters 3, 4, and 5 are all presented as regression problems and thus require methods that are defined for continuous data. The simplest method is univariate correlation based feature selection (CBF) (17). The process involves calculating the Pearson correlation coefficient between drug IC50 and gene expression for all genes. Genes below a certain correlation cutoff or correlation magnitude to include negative correlations. Alternatively, a level of significance cutoff can be made; however, depending on the level of significance, for a high dimensional system the false discovery rate can be quite high. This can be mitigated by correcting for multiple testing using Bonferroni correction (18), or controlling for the false discovery rate using the Benjamini-Hochberg procedure (19). Again the Pearson correlation assumes normality; however, in a case where the normality assumption does not hold the Spearman rank coefficient offers a non-parametric alternative.

High dimensional systems often involve a large amount of co-linearity. This certainly applies to genomic systems as many genes typically are involved in a cellular process. This can pose challenges when using any kind of filter feature selection. If several genes are highly correlated, they both might meet selection criteria; however, this results in a high level of redundancy between features offering no additional predictive power and might actually be punitive by adding additional noise that can cause overfitting. This suggests that the optimal set of features are the features that can explain the maximum variability in the target variable while minimizing the overlap between features. This is precisely the motivation behind minimum redundancy maximum relevance (MRMR) feature selection (20). The MRMR algorithm is simple: each new feature is added to maximize the following quantity (20):

$$\max_{g_i \in G'} \left\{ f(g_i, y) \Big/ \left[\frac{1}{|G|} \sum_{g_j \in G} |c(g_j, g_i| \right] \right\} \quad (2.8)$$

where $g_j \in G$ which is the set of selected features in this case genes, $g_i \in G'$ is the set of all other genes, $|G|$ is the number of selected genes, $f(g_i, y)$ is a function that measures the similarity between the gene and drug response, y, such as correlation or f-test in the case of categorical targets (i.e. sensitive vs resistant), and $c(g_j, g_j)$ is the Pearson correlation between genes. This is repeated until some criteria is met; for example, a specified cutoff for the number of features. Simply put, the next gene added to the previously selected features should have a maximal relationship with drug response with minimal co-linearity with genes already in the feature set.

Additional Methods of Linear Regression:

Least squares regression is a very effective method to estimate a number of practical phenomena both linear and approximately linear. However, as discussed in high dimensional systems the curse of dimensionality can lead to overfitting. One of the methods to address this issue is to use filter based feature selection; nonetheless, many filter based feature selection methods are univariate ignoring multi-variate interactions that may play important roles. Some filter based methods can be extended to multi-variate interactions; for example, CBF can be applied to sets of variables (17); however, to apply this method to every possible combination of unique sets in high dimensional systems is impractical, especially for anything greater than a two gene interaction. If g is the number of genes and k is the number of covariates in the feature, the number of possible features to be tested is g^k. Furthermore, wrapper and embedded methods can be computationally intensive as well (11). Nonetheless, several modified versions of LSQR have been proposed that can offer substantial improvement based on a generalized linear model.

Principal components analysis is an effective method for conserving the maximum variability of high dimensional data in a lower dimensional subspace. Thus, PCA results in a set of *M* principal components $\boldsymbol{U}_m = (\boldsymbol{u}_1, \dots, \boldsymbol{u}_M)$ with dimension $(l \; x \; M)$ where *M* must be less than or equal to the number of data points. Additionally, let $X = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)$ be the matrix of *N* data points with a dimension of $(l \; x \; N)$. Furthermore, as is the case for genomic data, $l \gg N$. Projecting X onto *M* principal components defines each $\boldsymbol{x}_n$ as a *M* dimensional approximation of each point $\boldsymbol{X}_M = \boldsymbol{U}_M^T \boldsymbol{X}$. Additionally, $\widehat{\boldsymbol{X}_m}$ is the best approximation, i.e. captures the most variability, of the

original data in *M* dimensions. Given that the target value follows a linear function $Y = \boldsymbol{w}^T\boldsymbol{X}$ the value of y as a function in *M* dimensions can be approximated as $Y_M = \boldsymbol{w}^T\boldsymbol{U_M}\boldsymbol{X} = \boldsymbol{w}_M^T\widehat{\boldsymbol{X}_M}$. The principal components regression (PCR) is a LSQR over the lower dimensional data with the following error function

$$E(\boldsymbol{w}_M) = \frac{1}{2}\sum_{n=1}^{N}\left(\boldsymbol{w}_M^T\widehat{\boldsymbol{x}_M}_n - y_n\right)^2 \quad (2.9)$$

The question that remains is how is *M* chosen. Recall that λ_m is the variability along component *m*, additionally $\lambda_1 > \lambda_2 > \cdots > \lambda_M > \cdots > \lambda_N$. One method would be to decide the number of components by the fraction of variability that is explained. The other methods is to systematically add or remove components by the magnitude λ_m and evaluate how many components are needed to perform optimally using cross validation during training. PCR can work exceptionally well if the variability in the data is directly related to the target value. PCR has the added value of not having to eliminate features, as required by feature selection, while finding a lower dimensional representation. However as noted above, the orientations of the components is with respect to the variability in the independent of data itself. Thus, if a large amount of variability in the data is unrelated to the target data, the components might not give the best low dimensional representation with respect to the target variable.

Ridge regression is a popular method used to fit linear functions while also bounding over-fitting. The method adds and additional penalty term to the error function that puts restrictions on the weight vector $\boldsymbol{w}$ referred to as a regularization term. Ridge regression, in particular uses what referred to as the L2 norm of a vector as a penalty function. The L2 norm for a vector is defined as

$$\boldsymbol{w}^T\boldsymbol{w} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{M}} |w_i|^{\boldsymbol{2}} \quad (2.10)$$

The resulting regularized error function is given as

$$E(\boldsymbol{w}) = \frac{\boldsymbol{1}}{\boldsymbol{2}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} (\boldsymbol{w}^T\boldsymbol{x}_i - y_i)^2 + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w} \quad (2.11)$$

Overfitting is usually coupled with larger values for the L2 norm, the penalty term therefore preferences components of $\boldsymbol{w}$ that have the largest overall effect on the accuracy at the expense of some components those which have lesser influence. Intuitively, this prevents overfitting because variables accounting for the "underlying phenomena" should have larger effects on the accuracy than variables that may only contribute to overfitting. Here λ is a Lagrange multiplier that modulates the trade off with higher values being more punitive. Geometrically this can be viewed as restricting the least squares solution to a hypersphere centered at the origin with a fixed radius (Figure 2.3 A).

Generally speaking a LP norm on the vector $\boldsymbol{w}$ is defined as

$$|\boldsymbol{w}|_p = \sum_{i=1}^{M} |w_i|^p \quad (2.12)$$

In addition to a L2 penalty there is a L1 penalty yielding the following error function

$$E(\boldsymbol{w}) = \frac{1}{2} \sum_{i=1}^{N} (\boldsymbol{w}^T\boldsymbol{x}_i - y_i)^2 + \alpha\, |\boldsymbol{w}|_1 \quad (2.13)$$

Where α is a again a Lagrange multiplier. This is known as Lasso regression (21); geometrically the L1 penalty is a hypercube with vertices falling on the axis which drives

select components of $\boldsymbol{w}$ to zero if the optimal solution intersects with a vertex of the hypercube (Figure 2.3 B). Additionally, a regularization penalty has been proposed that combines both L1 and L2 penalties

$$E(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{w}^T\boldsymbol{x}_i - y_i)^2 + \alpha\,|\boldsymbol{w}|_1 + \lambda\boldsymbol{w}^T\boldsymbol{w} \quad (2.14)$$

which is termed elastic net regression and has been shown to outperform both lasso and ridge regression when certain structural dependencies exist within the data (22).

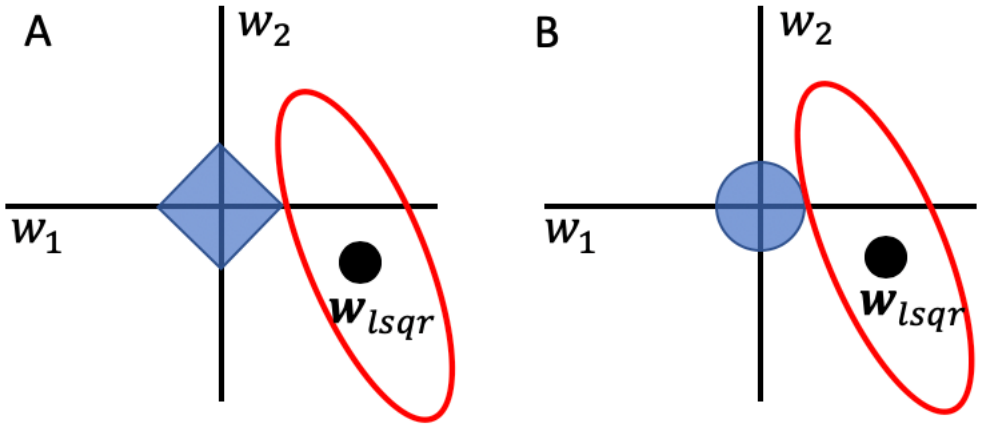


Figure 2.3. A. Example of a L1 norm. B. Example L2 norm. The black dot in the middle of the level curve is the solution is LSQR was performed. The regularized solution is where level curve intersects with the norm. This demonstrates how LASSO can lead to sparse solutions; likewise, ridge regression has the ability to drive many of the coefficients close to zero.

Support Vector Regression:

The challenge in supervised learning problem is that the solution should be generalizable to a broader space of data given that we only have a finite number of samples to train on. An aspect of real world data is that there is some uncertainty or noise associated with the data. Generally, overfitting is the result of the influence of noise during training; thus, minimizing the influence of noise will lead to better

performing models. Dimensionality reduction and regularization are strategies to mitigate the influence noise. Here support vector regression (SVR) is introduced which also aims to decrease the effects of noise. This section is largely influenced by the tutorial provided by Smola and Schölkopf (23) and the text by Bishop (10). Again, let's consider the problem of linear regression, we have been given some data $D = \{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_N, y_N)\}$, $\boldsymbol{x}_i \in \mathbb{R}^d, y \in \mathbb{R}$ and we hypothesize that relationship between $\boldsymbol{x}_i$, and y_i, is a linear function i.e. $f(\boldsymbol{x}_i) = \boldsymbol{w}^T\boldsymbol{x}_i + b$. Given that y_i is some kind of measurement there will be some sort of noise, ε, such that $f(\boldsymbol{x}_i) = y_i \pm \varepsilon$. Thus, the learning problem amounts to finding the best function $f(\boldsymbol{x}_i) = y_i \pm \varepsilon = \boldsymbol{w}^T\boldsymbol{x}_i + b$. As with regularized regression we want to limit the magnitude of $\boldsymbol{w}$. We can define this problem as follows:

$$\min \frac{1}{2}|\boldsymbol{w}|_2 \qquad \text{Subject to } \begin{cases} y_i - \boldsymbol{w}^T\boldsymbol{x}_i - b \leq \varepsilon \\ \boldsymbol{w}^T\boldsymbol{x}_i + b - y_i \leq \varepsilon \end{cases} \tag{2.15}$$

Geometrically, this corresponds to finding a line where all the points fall within a distance ε, which can be viewed as a cylinder with radius ε (referred to as the ε tube) surrounding the line $\boldsymbol{w}^T\boldsymbol{x}_n + b$ (Figure 2.4). However, such line might not exist, given a value ε, and there might be a tradeoff between letting some points lie outside the ε tube and the best fit line. To accomplish this a new variable is introduced, ξ, that allows a tradeoff between the number of points that fall outside the ε tube and stringency of the constraints. This defines the following optimization problem

$$\min\left[\frac{1}{2}|\boldsymbol{w}|_2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)\right]$$

$$\text{Subject to } \begin{cases} y_i - \boldsymbol{w}^T\boldsymbol{x}_i - b \le \varepsilon + \xi_i \\ \boldsymbol{w}^T\boldsymbol{x}_i + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \tag{2.16}$$

Where C is a constant that modulates how far points will be allowed outside the ε tube.

This defines the ε insensitive loss function defined by Vapnik (24)

$$E_\varepsilon = \begin{cases} 0 \; if \; |\boldsymbol{w}^T\boldsymbol{x}_i + b - y_i| \le \varepsilon \\ |\xi| = |\boldsymbol{w}^T\boldsymbol{x}_i + b - y_i| - \varepsilon \; otherwise \end{cases} \tag{2.17}$$

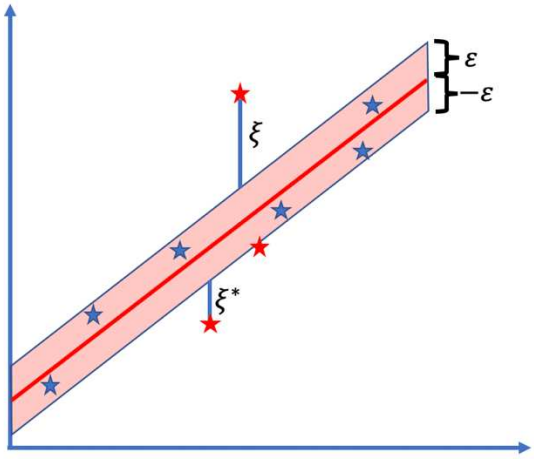


Figure 2.4 Diagram of epsilon insensitive support vector regression

This optimization problem is solved using a Lagrangian with multipliers $(\alpha_i, \alpha_i^*, \eta_i, \eta_i^*) \ge 0$

$$L = \frac{1}{2}|w|_2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)$$

$$-\sum_{i=1}^{N}\alpha_i(\varepsilon + \xi_i - y_i + \boldsymbol{w}^T\boldsymbol{x}_i + b)$$

$$-\sum_{i=1}^{N}\alpha_i^*(\varepsilon + \xi_i^* + y_i - \boldsymbol{w}^T\boldsymbol{x}_i - b) - \sum_{i=1}^{N}(n_i\xi_i + n_i^*\xi_i^*) \tag{2.18}$$

in order to obtain optimality the following conditions must exist

$$\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{x}_i(\alpha_i - \alpha_i^*) = 0 \quad (2.19)$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{N} (\alpha_i^* - \alpha_i) = 0 \quad (2.20)$$

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \eta_i = 0 \quad (2.21)$$

$$\frac{\partial L}{\partial \xi_i^*} = C - \alpha_i^* - \eta_i^* = 0 \quad (2.22)$$

Given these equations $\boldsymbol{w} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{x}_i(\alpha_i - \alpha_i^*)$ and therefore

$$f(x_j) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) \boldsymbol{x}_i^T \boldsymbol{x}_j + b \quad (2.23)$$

Utilizing equations 2.19-2.22 into equation 2.18 defines the dual optimization problem

$$\max[-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle - \varepsilon \sum_{i=1}^{N}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{N} y_i(\alpha_i - \alpha_i^*)] \quad (2.24)$$

$$\text{subject to } 0 \le \alpha_i \le C, 0 \le \alpha_i^* \le C$$

the solution to 2.23 requires the Karush-Kuhn-Tucker (KKT) conditions

$$\alpha_i(\varepsilon + \xi_i + \boldsymbol{w}^T \boldsymbol{x}_i + b - y_i) = 0 \quad (2.25)$$

$$\alpha_i^*(\varepsilon + \xi_i^* - \boldsymbol{w}^T \boldsymbol{x}_i - b + y_i) = 0 \quad (2.26)$$

$$(C - \alpha_i)\xi_i = 0 \quad (2.27)$$

$$(C - \alpha_i^*)\xi_i^* = 0 \quad (2.28)$$

Insight can be gained by quickly observing equations 2.25-2.28. First, the dual variables α_i, α_i^* can only be non-zero if the point lies outside of the ε-tube, thus, these points are the only points used in the calculation of equation 2.23 and are referred to the support

vectors. Second, α_i and α_i^* cannot both have a non-zero value otherwise the point would have to lie simultaneously above and below the ε-tube. Third, $0 \leq \alpha_i \leq C$ and $0 \leq \alpha_i^* \leq C$. There still is the matter of finding b. Based on equation 2.23 it might be tempting to pick a training point, plug it in, an then solve for b; however, for the any training point the equality of 2.23 most likely does not hold but must satisfy the constraints in 2.16. Furthermore, the KKT conditions mandate that for any support vector where $0 < \alpha_i \; or \; \alpha_i^* < C$ by the condition of either 2.27 or 2.28 $\xi_i = 0, \xi_i^* = 0$. Thus, we can pick a point such that $0 < \alpha_i \; or \; \alpha_i^* < C$ and $\xi_i = 0, \xi_i^* = 0$ and either 2.25 or 2.26 must meet the following condition:

$$\begin{cases} \varepsilon + \boldsymbol{w}^T \boldsymbol{x}_\Omega + b - y_\Omega = 0 \; if \; 0 < \alpha_\Omega < C \\ \varepsilon - \boldsymbol{w}^T \boldsymbol{x}_\Omega - b + y_\Omega = 0 \; if \; 0 < \alpha_\Omega^* < C \end{cases}$$

Therefore, lets select a point, $\boldsymbol{x}_\Omega$ for which $0 < \alpha_\Omega < C$ we can solve for b

$$b = y_\Omega - \varepsilon - \boldsymbol{w}^T \boldsymbol{x}_\Omega \quad (2.29)$$

Substituting the dual solution for $\boldsymbol{w} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{x}_i(\alpha_i - \alpha_i^*)$ 2.29 becomes

$$y_\Omega - \varepsilon - \sum_{j=1}^{M} (\alpha_j - \alpha_j^*) \boldsymbol{x}_j^T \boldsymbol{x}_\Omega \quad (2.30)$$

Where M is the number of support vectors Thus subbing this into 2.23 the prediction for a new point $\boldsymbol{x}$ becomes

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) \boldsymbol{x}_i^T \boldsymbol{x} + y_\Omega - \varepsilon - \sum_{j=1}^{M} (\alpha_j - \alpha_j^*) \boldsymbol{x}_j^T \boldsymbol{x}_\Omega \quad (2.31)$$

The Kernel Trick and Kernel Methods

Linear systems represent a very small subset of phenomena in the natural world; however, they are well characterized and readily applied to many problems in science,

engineering, and mathematics. Linear functions represent some of the most basic mathematical functions; nonetheless, a formal mathematical framework is necessary for understanding some of the more complex applications of linearity. As discussed linear regression amounts to finding a function of the form $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$, where $\boldsymbol{x} = \{x_1, \dots, x_d\}$ is a multidimensional set, $\boldsymbol{w} = \{w_1, \dots, w_d\}$ is a d dimensional vector, and b is a constant. The linear framework can be generalized to a much larger subset of functions by using a basis set of functions $\Phi = \{\phi_1, \dots, \phi_l\}$. In the case of a strictly linear function, $\phi(\boldsymbol{x}) = \boldsymbol{x}$; however, Φ, could be a basis of polynomial functions up to degree 2 such that $\Phi = \{\phi_1, \phi_2, \phi_3\} = \{1, \boldsymbol{x}, \boldsymbol{x}^2\}$. Thus, the more general form of a linear function becomes $f(\boldsymbol{x}) = \boldsymbol{w}^T\Phi(\boldsymbol{x}) + b$ giving the following LSQR error function:

$$E: \boldsymbol{w} = \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{w}^T\Phi(\boldsymbol{x}_i) - y_i)^2 \quad (2.31)$$

The use of basis functions allows the extension of the rigorous framework of linear models to a broader class of functions beyond strictly linear functions.

As mentioned earlier the idea of distance plays a central role in regression problems. Geometrically, the solution to least squares problem is the vector, $\boldsymbol{w}$, that minimizes the distance between every data point, $\boldsymbol{x}_n$, and the projection of $\boldsymbol{x}_n$ onto $\boldsymbol{w}$. In Euclidean geometry the projection of a vector $\boldsymbol{x}_n$ onto the vector $\boldsymbol{w}$ is given as the inner product between the two vectors, $\langle \boldsymbol{w}, \boldsymbol{x}_n \rangle = w_1x_{n,1} + w_2x_{n,2} + \cdots + w_mx_{n,m}$. Using the solution of the LSQR problem we found that

$$\boldsymbol{w} = \frac{Y\boldsymbol{X}^T}{\boldsymbol{X}^T\boldsymbol{X}} \quad (2.32)$$

Now considering that we want to make a prediction, y, given a new data point, $\boldsymbol{x}$, after training on a set of N data points, the solution is given by

$$y = \frac{Y\boldsymbol{X}^T}{\boldsymbol{X}^T\boldsymbol{X}}\boldsymbol{x} = \frac{Y}{\boldsymbol{X}^T\boldsymbol{X}}\sum_{i=1}^{N}\langle \boldsymbol{x}_i, \boldsymbol{x}\rangle \quad (2.33)$$

Where $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = x_1y_1 + x_2y_2 + \cdots + x_ny_n$ is the Euclidian inner-product. Additionally, lets define the matrix $\boldsymbol{K} = \boldsymbol{X}^T\boldsymbol{X}$, therefore each elements of $\boldsymbol{K}$ is given by $\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle$ where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are training points. Furthermore, lets define the (1 x N) vector $\boldsymbol{\alpha} = Y\boldsymbol{K}^{-1}$. Therefore the target value, y, for any $\boldsymbol{x}$**,** is given by a linear combinations of inner-products between the new point and each training point.

$$y = \boldsymbol{\alpha}\sum_{i=1}^{N}\langle \boldsymbol{x}_i, \boldsymbol{x}\rangle \quad (2.34)$$

This can further be generalized to any basis set of functions $\Phi = \{\phi_1, \phi_2, \dots, \phi_m\}$

$$y = \boldsymbol{\alpha}\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}}\langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x})\rangle \quad (2.35)$$

Furthermore, we will define a new function called a kernel function $k(\boldsymbol{x}_i, \boldsymbol{x}) = \langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x})\rangle$ the above equation then becomes

$$y = \boldsymbol{\alpha}\sum_{i=1}^{N}k(\boldsymbol{x}_i, \boldsymbol{x}) \quad (2.36)$$

Example: Polynomial Basis Functions

Consider a polynomial with a maximum a degree of two or less for a two dimensional variable $\boldsymbol{x} = \{x_{1,}x_2\}$

$$y = w_0 + w_1x_1 + w_2x_2 + w_3x_1x_2 + w_4x_1^2 + w_5x_2^2 \quad (2.37)$$

This can be defined by the following basis functions

$$\Phi(\boldsymbol{x}) = \{\phi_0, \phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6\} = \{1, x_1, x_2, x_1x_2, x_1^2, x_2^2\}$$

now

$$y = \boldsymbol{w}^T \Phi(\boldsymbol{x})$$

Which is clearly a linear function which can be approximated by LSQR, thus

$$y = \boldsymbol{w}^T \boldsymbol{\Phi}(\boldsymbol{x}) = \frac{Y\boldsymbol{\Phi}^T(\boldsymbol{X})}{\boldsymbol{\Phi}^T(\boldsymbol{X})\boldsymbol{\Phi}(\boldsymbol{X})}\Phi(\boldsymbol{x}) = \frac{Y}{\boldsymbol{\Phi}^T(\boldsymbol{X})\boldsymbol{\Phi}(\boldsymbol{X})}\sum_{i=1}^{N}\langle\Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x})\rangle = \boldsymbol{\alpha}\sum_{i=1}^{N} k(\boldsymbol{x}_i, \boldsymbol{x}) \quad (2.38)$$

Such that

$$k(\boldsymbol{z}, \boldsymbol{x}) = \langle[1, z_1, z_2, z_1z_2, z_1^2, z_2^2], [1, x_1, x_2, x_1x_2, x_1^2, x_2^2]\rangle$$

Where the inner product is the familiar Euclidean dot product as defined above. While, it seems that nothing revelatory has been done the central concept is that we have taken a non-linear two dimensional function and, using a non-linear transformation, $\Phi(\boldsymbol{x})$, transformed it to a linear function in a five dimensional space. Furthermore, every point in this five dimensional space has a defined inner product given by the kernel function. The relationship between kernel functions and linear basis functions is central to allow non-linear function to be represented in a linear space.

In the above example we showed that a kernel function is defined as the inner product of a set of linear basis functions; however, the basis functions were limited to a very small subset of functions, in this case multiplicative products with degree two or less. However, in practice, the basis functions are usually not known or well defined and thus an inner product and its kernel function are not defined as well. Fortunately, the application of Mercer's theorem allows for the generation of orthogonal basis set for any finite set of points. Let $\boldsymbol{X} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$ and let $k(\boldsymbol{x}_i, \boldsymbol{x}_j) \rightarrow \mathbb{R}; \boldsymbol{x}_i, \boldsymbol{x}_j \ni \boldsymbol{X}$ be a continuous function with the following properties:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = k(\boldsymbol{x}_j, \boldsymbol{x}_i)\ and\ \sum_{i=1}^{N}\sum_{j=1}^{N} k(\boldsymbol{x}_i, \boldsymbol{x}_j)c_ic_j \geq 0 \quad (2.39)$$

where $c_i, c_j \ni \mathbb{R}^+$. Then there exists a set of basis functions, Φ, such that

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j) \rangle \; \forall \; \boldsymbol{x}_i, \boldsymbol{x}_j \ni \boldsymbol{X}$$

The first property requires all kernel functions to be symmetric and the second condition requires all kernel functions to be positive semi-definite. In general, any continuous, symmetric, positive semi-definite function defines an inner-product on some set of orthonormal basis functions. This sometimes referred to as the "kernel trick", and it plays a central role in allowing non-linearity in ridge regression and support vector regression.

As stated, any continuous, symmetric, positive semi-definite function can be a kernel; nonetheless, throughout the body of this work two are utilized, linear kernels and radial basis function (RBF) kernels; for a more comprehensive list of kernel functions see Bishop (10). The linear kernel is simply the familiar dot product; let $\boldsymbol{x} = \{x_1, x_2, \dots, x_m\}$ and $\boldsymbol{z} = \{z_1, z_2, \dots, z_m\}$ then the kernel function is

$$k(\boldsymbol{x}, \boldsymbol{z}) = \langle \boldsymbol{x}, \boldsymbol{z} \rangle = x_1 z_1 + x_2 z_2 + \cdots + x_m z_m \quad (2.40)$$

in the case of a linear kernel $\Phi(\boldsymbol{x}) = \{x_1, x_2, \dots, x_m\}$ and only appropriate to approximate linear relationships. One of the more frequently used kernel function is the radial basis function and is defined as follows:

$$k(\boldsymbol{x}, \boldsymbol{z}) = \exp(-\gamma |\boldsymbol{x} - \boldsymbol{z}|_1^2) \quad (2.41)$$

where $|\boldsymbol{x} - \boldsymbol{z}|_1$ is the L1 norm and γ is a constant usually chosen by cross-validation. In the case of a RBF the basis set, Φ, is not so apparent; in fact, in practice Φ is not explicitly known. This is precisely why the kernel trick is so powerful, via Mercer's theorem, for any kernel k there must be a orthonormal basis set Φ such that

$$k(\boldsymbol{x}, \boldsymbol{z}) = \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{z}) \rangle \quad (2.42)$$

Thus, using LSQR as an example

$$y = \boldsymbol{\alpha} \sum_{i=1}^{N} k(\boldsymbol{x}_i, \boldsymbol{x}) \quad (2.43)$$

the value of the kernel function is all that needs to be calculated. However, it is important to remember that the orthonormal basis set is defined by the kernel function via the inner-product; with respect to supervised learning, the basis functions defined by the kernel might not be, and probably are not, the optimal set of basis functions that assigns each data point to a target with minimum loss in generalizability. Nonetheless, if no apparent set of basis functions is well defined for the problem at hand, the kernel trick can define a basis set that can capture a number of non-linear relationships. For example, the RBF kernel captures an infinite set of orthonormal non-linear functions, such that the inner product between two close points is exponentially larger then points that are far apart. Thus, again using LSQR as an example when a target value is estimated for a new data point, the target value is more influenced by points that are close to the new point then those far apart.

The Kernel Trick and Non-Linear Regression:

The kernel trick is not so much as a "trick" but taking advantage of the prevalence of the inner product in learning problems. This was shown with LSQR above; however, the kernel trick is fundamental in generalizing linear regression and classification techniques into non-linear methods. For instance, take the error function for ridge regression replacing $\boldsymbol{x}_i$ with $\Phi(\boldsymbol{x}_i)$

$$E(\boldsymbol{w}) = \frac{\mathbf{1}}{\mathbf{2}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} (\boldsymbol{w}^T \Phi(\boldsymbol{x}_i) - y_i)^2 + \frac{\lambda}{2} \boldsymbol{w}^T \boldsymbol{w} \quad (2.43)$$

Taking the derivative and setting it equal to zero

$$\boldsymbol{w} = -\frac{1}{\lambda}\sum_{i=1}^{N}(\boldsymbol{w}^T\Phi(\boldsymbol{x}_i) - y_i)\Phi(\boldsymbol{x}_i) \quad (2.44)$$

Define

$$\boldsymbol{a} = -\frac{1}{\lambda}\sum_{i=1}^{N}(\boldsymbol{w}^T\Phi(\boldsymbol{x}_i) - y_i) \; = -\frac{1}{\lambda}(\boldsymbol{w}^T\boldsymbol{\Phi}(\boldsymbol{X}) - \boldsymbol{y}) \quad (2.45)$$

Thus

$$\boldsymbol{w} = \boldsymbol{a}\boldsymbol{\Phi}^T(\boldsymbol{X}) \quad (2.46)$$

using equations 2.44, 2.45 and 2.46

$$-\lambda\boldsymbol{a} = (\boldsymbol{a}\boldsymbol{\Phi}^{\boldsymbol{T}}(\boldsymbol{X})\boldsymbol{\Phi}(\boldsymbol{X}) - \boldsymbol{y}) \quad (2.47)$$

Solving equation 2.47 for $\boldsymbol{a}$ and recalling that $\boldsymbol{K} = \boldsymbol{\Phi}^T(\boldsymbol{X})\boldsymbol{\Phi}(\boldsymbol{X})$

$$\boldsymbol{a} = \frac{\boldsymbol{y}}{\boldsymbol{K} + \lambda\boldsymbol{I}_N} \quad (2.48)$$

Where $\boldsymbol{I}_N$ is the (NXN) identity matrix. The form of the equation that minimizes equation 2.1 is $y = \boldsymbol{w}^T\Phi(\boldsymbol{x})$. The using equation 2.46 and 2.48

$$y = \boldsymbol{a}\boldsymbol{\Phi}^T(\boldsymbol{X})\Phi(\boldsymbol{x}) = \frac{k(\cdot,\boldsymbol{x})\boldsymbol{y}}{\boldsymbol{K} + \lambda\boldsymbol{I}_N} \quad (2.49)$$

Where

$$k(\cdot,\boldsymbol{x}) = \boldsymbol{\Phi}^T(\boldsymbol{X})\Phi(\boldsymbol{x}) = \sum_{i=1}^{N}\langle\Phi(\boldsymbol{x}_i),\Phi(\boldsymbol{x})\rangle = \sum_{i=1}^{N}k(\boldsymbol{x}_i,\boldsymbol{x}) \quad (2.50)$$

Applying the kernel trick to ridge regression results in a solution given by equation 2.7 and is simply referred to as kernel ridge regression.

Recall that the solution to the support vector regression problem is given by

$$y(\boldsymbol{x}) = \sum_{i=1}^{N}(a_i - a_i^*)\langle \boldsymbol{x}_i, \boldsymbol{x}\rangle + y' - \epsilon - \sum_{j=1}^{N}\left(a_j - a_j^*\right)\langle \boldsymbol{x}_j, \boldsymbol{x}'\rangle$$

Then generalizes to an arbitrary set of basis functions

$$y(\boldsymbol{x}) = \sum_{i=1}^{N}(a_i - a_i^*)\langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x})\rangle + y' - \varepsilon - \sum_{j=1}^{N}\left(a_j - a_j^*\right)\langle \Phi(\boldsymbol{x}_j), \Phi(\boldsymbol{x}')\rangle \quad (2.51)$$

We can then use the definition of a kernel function

$$y(\boldsymbol{x}) = \sum_{i=1}^{N}(a_i - a_i^*)k(\boldsymbol{x}_i, \boldsymbol{x}) + y_\Omega - \varepsilon - \sum_{j=1}^{N}\left(a_j - a_j^*\right)k\left(\boldsymbol{x}_j, \boldsymbol{x}_\Omega\right) \quad (2.52)$$

For which any continuous, symmetric, positive semi-definite function can be substituted for k. Therefore, as well as in kernel ridge regression, the kernel function allows for an implicit set of orthonormal basis functions which, depending on the kernel, can include a number of non-linear functions.

Artificial Neural Network:

Artificial neural networks (ANN) can be applied to non-linear regression and classification tasks. As the name suggests, ANNs are conceptually inspired by biological neural networks where a single neuron receives electrical input from surrounding neurons, integrates this signal and then based on the magnitude of the integrated signals outputs an appropriate signal to other neurons. In the case of ANN, neurons are represented by nodes, signals are represented by linear functions, and each output is modulated by an activation function. Each ANN is composed of layers input layers feed data into the network, hidden layers perform mathematical operations on input from both input layers, and output layers generate the output from the

accumulation of all other layers. Figure 2.5 shows a simple 4 layer network with two hidden layers.

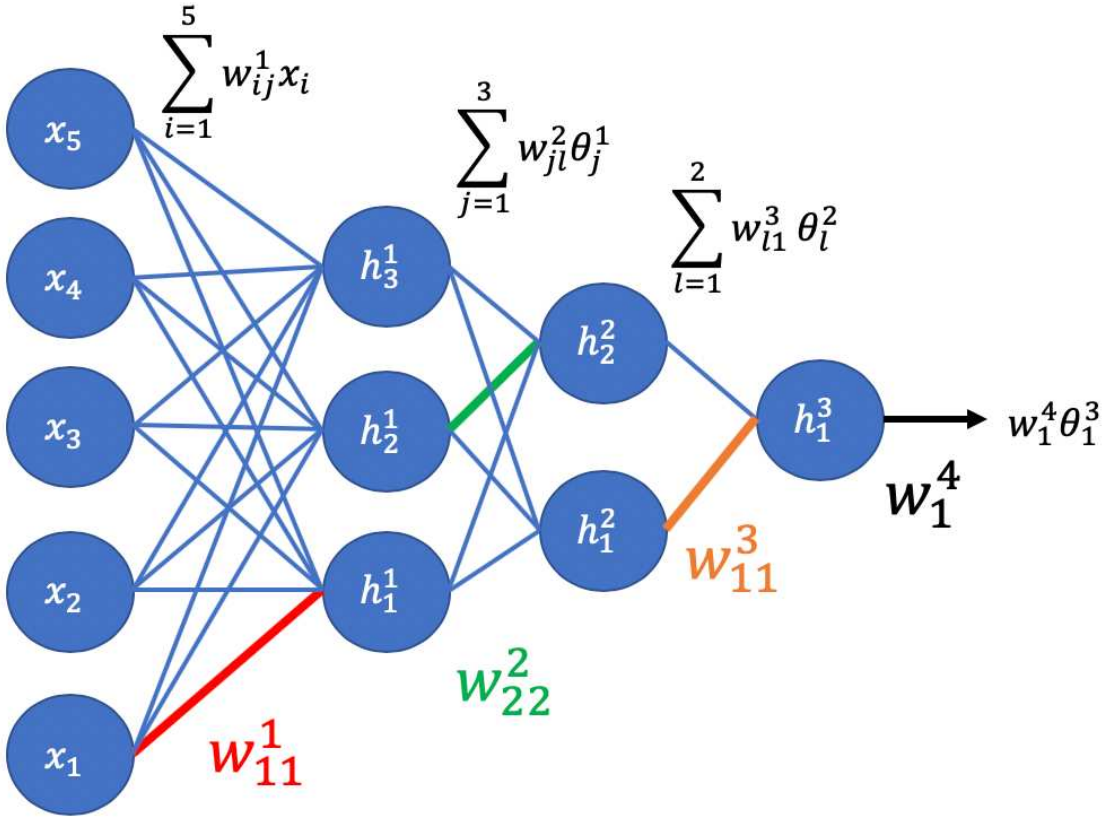


Figure 2.5. Example a 2 hidden layer neural network with a single output node.

For each node in the hidden layer the input is given by a linear combination of basis functions evaluated on the input data.

$$y_j^1 = \sum_{i=1}^{D} w_{ji}^1 \phi_i^1(\boldsymbol{x}_n) \quad (2.53)$$

Let $h_j^1(y_j^1)$ denote the activation function for the j[th] node for the first hidden layer then the output of each node from the first hidden layer is given by

$$y_j^2 = h_j^1\left(\sum_{i=1}^{D} w_{ji}^1 \phi_i^1(\boldsymbol{x}_n)\right) \quad (2.54)$$

Then input to each node of the second hidden layer is given by a linear combination of the outputs of the first hidden layer

$$y_k^2 = \sum_{j=1}^{J} w_{kj}^2\, h_j^1\left(\sum_{i=1}^{D} w_{ji}^1 \phi_i^1(\boldsymbol{x}_n)\right) \quad 2.55$$

following the previous steps the output of the 2nd hidden layer is

$$y_l^3 = \sum_{i}^{l} w_{lk}^3 h_O^2\left(\sum_{j=1}^{J} w_{kj}^2\, h_j^1\left(\sum_{i=1}^{D} w_{ji}^1 \phi_i^1(\boldsymbol{x}_n)\right)\right) \quad (2.56)$$

finally the output node gives the final value

$$f(\boldsymbol{x}_n) = \sum_{k=1}^{1} w_k^4 h_k^3\left(\sum_{i}^{l} w_{lk}^3 h_l^2\left(\sum_{j=1}^{J} w_{jl}^2\, h_j^1\left(\sum_{i=1}^{D} w_{ij}^1 \phi_i^1(\boldsymbol{x}_n)\right)\right)\right) \quad (2.57)$$

For regression the error function for a ANN is given by least squares error function

$$E(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}(f(\boldsymbol{x}_n) - y_n)^2 \quad (2.58)$$

For simplicity let $\phi(\boldsymbol{x}) = \boldsymbol{x}$. The we minimize function 2.58

$$\nabla E(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}(f(\boldsymbol{x}_n) - y_n)\frac{df(\boldsymbol{x}_n)}{d\boldsymbol{w}} \quad (2.59)$$

Thus we are required to calculate $\frac{df(\boldsymbol{x}_i)}{d\boldsymbol{w}}$. At first sight this looks like a daunting task so we will make the following simplifications to equation 2.57

$$\theta_j^1 = h_j^1\left(\sum_{i=1}^{D} w_{ij}^1 x_{i,n}\right)$$

$$\theta_l^2 = h_l^2\left(\sum_{j=1}^{J} w_{jl}^2 \theta_j^1\right)$$

$$\theta_k^3 = h_1^3(\theta_l^2)$$

Given this simplification equation 2.57 becomes

$$f(\boldsymbol{x}_n) = w_1^4 \theta_1^3$$

now lets calculate $\frac{\partial f(\boldsymbol{x}_n)}{\partial w_{11}^1}$

$$\frac{\partial f(\boldsymbol{x}_n)}{\partial w_{11}^1} = \frac{\partial f(\boldsymbol{x_n})}{\theta_1^3}\frac{\partial \theta_1^3}{\partial \theta_l^2}\frac{\partial \theta_l^2}{\partial \theta_1^1}\frac{\theta_1^1}{w_{11}^1} = w_1^4 w_{lk}^3 w_{1l}^2 \frac{\partial h_1^1}{w_{11}^1}$$

Additionally if we calculate $\frac{\partial f(\boldsymbol{x}_n)}{\partial w_{22}^2}$

$$\frac{\partial f(\boldsymbol{x}_n)}{\partial w_{22}^2} = \frac{\partial f(\boldsymbol{x}_n)}{\partial \theta_1^3}\frac{\partial \theta_1^3}{\partial \theta_2^2}\frac{\partial \theta_2^2}{w_{22}^2} = w_1^4 w_{2k}^3 \frac{\partial h_2^2}{w_{22}^2}$$

Thus, for an arbitrary network any derivative for $\boldsymbol{w}$ can be calculated in a similar fashion.

Now we know that a minimum of eq 2.59 is found when the gradient is equal to 0.

However, because of the non-linear nature of ANN often times there are multiple local minima thus there is typically no way to find all minimum to find the global minimum.

Thus, ANN are typically solved using a gradient descent algorithm

$$\boldsymbol{w}^{\tau+1} = \boldsymbol{w}^{\tau} - \eta \nabla E(\boldsymbol{w}^{\tau})$$

Where τ is the iteration and η is a learning rate.

Conclusion:

This chapter has provided a very brief introduction to select concepts and methods in statistical learning and has barely scratched the surface of the vast and developing field. For those looking to dive further into material I suggest starting with the text by Bishop (10). In the following chapters statistical learning is utilized to predict cancer drug response given gene expression either for *in vitro* cell lines or patient tumors. In chapter 3 principal components regression, support vector regression, and artificial neural networks along with methods of feature selection are utilized and evaluated to predict drug response in multiple *in vitro* cell line databases. Chapter 4 utilizes support vector regression to predict *in vitro* drug response using drug perturbed gene expression in NCI60 cell lines. Finally, chapter 5 utilizes a variation of support vector regression called survival support vector machines to predict drug response in bladder cancer patients that have received a combination treatment of the chemotherapy drugs cisplatin and gemcitabine. The potential for application statistical learning in medicine is a rapid and growing field; however, the complexity of biological systems are immense. Thus, an understanding of these applications coupled with biological insight and creativity will be essential in moving forward. The contents of this thesis attempt to explore a small sliver of these applications; however, the vast intrigue of this work is in the questions that are left unanswered rather than the questions that are answered.

REFERENCES

1. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning: with Applications in R: Springer New York; 2013.

2. Quackenbush J. Microarray data normalization and transformation. Nature Genetics. 2002;32(4):496-501.

3. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Research. 2002;30(4):e15-e.

4. Cleveland WS. Robust Locally Weighted Regression and Smoothing Scatterplots. Journal of the American Statistical Association. 1979;74(368):829-36.

5. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4(2):249-64.

6. McCall MN, Bolstad BM, Irizarry RA. Frozen robust multiarray analysis (fRMA). Biostatistics. 2010;11(2):242-53.

7. Salzberg SL. Open questions: How many genes do we have? BMC Biology. 2018;16(1):94.

8. Energy; UDo. About the Human Genome Project 2019 [updated March 26, 2019; cited 2020 June 22]. Available from: https://web.ornl.gov/sci/techresources/Human_Genome/project/index.shtml.

9. Bellman R, Corporation R, Collection KMR. Dynamic Programming: Princeton University Press; 1957.

10. Bishop CM. Pattern Recognition and Machine Learning. New York, New York Springer; 2006.

11. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23(19):2507-17.

12. Liu H, Motoda H. Computational Methods of Feature Selection: CRC Press; 2007.

13. Staunton JE, Slonim DK, Coller HA, Tamayo P, Angelo MJ, Park J, et al. Chemosensitivity prediction by transcriptional profiling. Proceedings of the National Academy of Sciences. 2001;98(19):10787.

14. 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 research. 2015;43(7):e47-e.

15. Thomas JG, Olson JM, Tapscott SJ, Zhao LP. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. Genome Res. 2001;11(7):1227-36.

16. Huang D, Chow TWS. Effective feature selection scheme using mutual information. Neurocomputing. 2005;63:325-43.

17. Hall M. Correlation-based feature selection for machine learning. New Zealand Waikato University; 1999.

18. Haynes W. Bonferroni Correction. In: Dubitzky W, Wolkenhauer O, Cho K-H, Yokota H, editors. Encyclopedia of Systems Biology. New York, NY: Springer New York; 2013. p. 154-.

19. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological). 1995;57(1):289-300.

20. DING C, PENG H. MINIMUM REDUNDANCY FEATURE SELECTION FROM MICROARRAY GENE EXPRESSION DATA. Journal of Bioinformatics and Computational Biology. 2005;03(02):185-205.

21. Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996;58(1):267-88.

22. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005;67(2):301-20.

23. Smola AJ, Schölkopf B. A tutorial on support vector regression. Statistics and Computing. 2004;14(3):199-222.

24. Vapnik V. The Nature of Statistical Learning Theory 2ed. New York: Springer 1995.

CHAPTER 3

A SYSTEMATIC ANALYSIS OF GENOMICS-BASED MODELING APPROACHES FOR PREDICTION OF DRUG RESPONSE TO CYTOTOXIC CHEMOTHERAPIES[4]

Background:

The introduction of cDNA microarrays launched a new era of genomic studies in biological systems (1, 2). This revolutionary new technology allowed researchers to collect vast amounts of data to characterize the genomic landscape fundamental to biological processes. The power of this technology was soon realized to have broad implications in the study of cancer, providing insight into the genomic nature of the disease (3-5). Over the past few decades there has been a concerted community effort to collect both *in vivo and in vitro* data characterizing the molecular blueprints for a variety of cancers (6, 7). This work has spawned countless new insights and has paved the way for a new paradigm of cancer treatment involving precision approaches (8).

The term “Big Data” refers to the collection and storage of large amounts of information for analysis providing insight for a variety of applications (9). The mathematical, statistical, and computational techniques to analyze and extract this information from large sets of complex data encompass the field of statistical learning, having application in science, business, and technology (10). The ability of statistical learning theory to find useful information in large, complex, and often noisy datasets

[4] This work was published in BMC Medical Genomics in June 2019 titled *“A Systematic Analysis of Genomics-Based Modeling Approaches For Prediction of Drug Response to Cytotoxic Chemotherapies”* Authored by J.D. Mannheimer, D. Duval, A. Prasad, and D.L. Gustafson

make it a popular biomedical research area with clear clinical applications (11), including several in cancer diagnostics and treatment (9, 12, 13). A specific area of research has focused on the utilization of statistical learning to predict successful treatment options based on patient and disease specific clinical biomarkers (5, 14-16).

High throughput technologies have allowed researchers to profile the genomics of tumor-derived cell lines and test chemosensitivity to a variety of anti-cancer agents *in vitro*, most notably the National Cancer Institute 60 (NCI60) and the Genomics of Drug Sensitivity in Cancer (GDSC) cell line panels. Several studies have indicated the ability of *in vitro* data to predict patient response in multiple cancers (17-20). Therefore, *in vitro* drug response data offer a simplified format to uncover clinically relevant cancer drug relationships. Thus, models that can accurately capture behavior of *in vitro* experiments are essential to elucidate genomic signatures that can be further applied in more complex clinical models.

To date, one of the most comprehensive analyses of computational methods for predicting drug response with *in vitro* data was a community based challenge sponsored by the Dialogue on Reverse Engineering Assessment and Methods (DREAM) and National Cancer Institutes (NCI) (referred to as the DREAM-NCI challenge) (21). This challenge tasked 44 different research teams to build and train a predictive algorithm given gene expression, DNA methylation, mutation, copy number, protein abundance, and drug response for 35 breast cancer cell lines for 28 different known anti-cancer agents. The methods were then assessed for their ability to predict drug response for the 28 agents on 18 independent breast cancer cell lines. The resulting models highlighted some of the most advanced and cutting edge statistical

learning techniques, with the best model using Bayesian multi-task multiple kernel learning (MKL). However, the third best model differed in performance by only 2.3 percent using only weighted Pearson correlation between feature sets with drug response to make predictions. Overall, the DREAM-NCI challenge demonstrated the ability of statistical learning techniques to capture and predict drug response in *in vitro* environments.

The DREAM-NCI challenge illustrates the balancing act between complexity and simplicity that often presents itself in computational modeling. As “Big Data” takes off, more complex computational techniques will be developed offering new opportunities in precision oncology. However, to fully utilize and develop these techniques a firm understanding of how basic modeling principles influence performance is essential. Biological processes consist of complex dynamic interactions in a high dimensional system. Non-linear methods have the ability to capture complex interactions between players; however, in high dimensional systems these methods have a tendency to incorporate noise, leading to over-fitting. Alternatively, linear methods are more robust to over-fitting but at the cost of potentially missing important non-linear interactions. Furthermore, the high dimensional nature of biological data sets presents challenges in the ability to pinpoint covariates that are most informative to the underlying processes being modeled.

Insights into the molecular nature of cancer has driven a precision approach to cancer pharmacology by capitalizing on specific driver mutations exhibited by certain cancers (22-24). This strategy had been successful in a number of specific instances and continues to be an active area of research and drug development (25, 26).

Cytotoxic chemotherapies were some of the earliest drugs developed for the treatment of cancer and continue to play an important role in cancer therapy (27-30). However, the success of these drugs, as with all therapies, still varies (31). The toxicity associated with these drugs produce substantial side effects and can diminish quality of life for many patients; thus, a precision approach that can identify patients who would benefit could greatly improve the quality and efficacy of treatment. *In vitro* drug assays have become a standard approach to identifying compounds with potential therapeutic benefit (17, 18, 20). Opposed to targeted agents, mutations are poor predictors of efficacy for cytotoxic agents (32) and gene expression signatures have proven to show promise as predictors in cytotoxic agents (33, 34). Therefore, genomic data driven models that can accurately predict chemosensitivity to *in vitro* cell line assays of cytotoxic agents serve as a foundation for improving predictive models in patients.

Here we describe a systematic, pragmatic approach to identify the key components driving model performance when using genomic profiles to predict drug response in cytotoxic agents. While statistical learning offers a vast amount of possible techniques we simplify the approach by breaking down models into two fundamental aspects; the trade-off between linear and non-linear modeling techniques and the influence of feature selection via filter based selection methods. While, our approach is by no means an exhaustive survey of all possible techniques and approaches, our studies illustrate how simple approaches to modeling can offer valuable insight. Mainly we demonstrate that for a given population of cells the association between histotype and drug response is indicative of model performance. The dominance of these traits have important implications when assessing model performance and may prove

instructive in the development of new techniques for modeling drug response across multiple cancers.

Methods:

Preprocessing:

The Genomics for Drug Sensitivity in Cancer (GDSC) is comprised of over 1000 cancer cell lines with response data to 138 anticancer drugs. The available CEL files, containing gene array data using Affymetrix Human Genome U219, array were downloaded at (35). Using the “affy” R package, the CEL files were normalized using Robust Multi-Array Average (RMA) algorithm (36). The data were further corrected for batch variability using COMBAT of the “sva” R package (37). Cells that occurred in duplicate were averaged, resulting in a final gene expression matrix with 968 cell lines and 49386 genomic features. Likewise, for the NCI60, CEL files containing gene array data from Affymetrix Human Genome U133 2-plus array were downloaded from the CellMiner database (38, 39). A total of three CEL files were available for each NCI60 cell line, again the data was normalized using RMA (36) and batch corrected using COMBAT (37). The resulting data were then averaged over the three replicates to give a final gene expression value for each gene and cell. For our analysis in the GDSC we chose 15 cytotoxic drugs Table 3.1. The IC50 data was downloaded from (35). The NCI60 has 61 FDA approved cytotoxic agents (40), the drug response data again downloaded from CellMiner. For the majority of drugs, multiple IC50 measurements were made on multiple cell lines so the final IC50 represents an average over all measurements. For several of the drugs a significant number of the cell lines had the same reported IC50 leading to minimal variability and as such these drugs were

discarded. This left a total of 39 drugs, 14 of the drugs also had data in the GDSC. Several cell lines in multiple drugs in the GDSC reported IC_{50} above the maximum concentration experimentally tested and were not included in any of the models. Given the final number of cell lines as reported in Table 1. 75% of cell lines were randomly chosen and assigned to the training/validation set and the remaining 25% were assigned to the testing set. This was performed six times generating six non-overlapping test-train/validation splits. Likewise, in the NCI60 six random training/validation sets were generated consisting of 75% of the data with the remaining 25% left out for testing. To ensure the presence of each histotype in both testing and training sets, 75% of each histotype was reserved for training and validation with the remaining 25 % in the test set. Prostate cancer cell lines were removed because measurements were limited. Both in the GDSC and NCI60, these generated datasets were used on all models for a given drug.

Table 3.1. Cytotoxic drugs and number of cell lines. 15 cytotoxic agents and the number of cell lines with experimentally determined IC50's for each drug. Training set comprises 75% of the total data while the testing data account for the remaining 25%

Drug	**Abbreviation**	**Number of Cell Lines**
Bleomycin	BLM	632
Bortezomib	BTZ	331
Cisplatin	CIS	146
Cytarabine	CYT	515
Docetaxel	DTX	555
Doxorubicin	DOX	738
Etoposide	ETP	643
Gemcitabine	GEM	583
Methotrexate	MTX	216
Mitomycin C	MMC	759
Paclitaxel	PTX	227
Vinblastine	VBL	719
Vorinostat	VOR	728
SN-38	SN-38	698
5-Fluorouracil	5-FU	409

The choice to limit our analysis to cytotoxic chemotherapies was three-fold; first, as opposed to molecularly targeted therapies, cytotoxic chemotherapies work broadly to inhibit cell proliferation and the mechanisms of action are not dependent on specific driver mutations (22, 23). This has been demonstrated in the NCI60 where mutation status was shown to be a poor predictor of drug response in cytotoxic chemotherapies (32). Second, a study in "The Cancer Genome Atlas" concluded that "the information content content from copy number aberrations, miRNA and methylation is captured at the level of gene expression and protein function" (41). Lastly, several analyses have suggested that gene expression data accounted for the majority of variability in predictive model outcomes (21, 42). By restricting the study to cytotoxic agents, complications that arise from data redundancy could be minimized while also eliminating challenges in integrating different data types. Thus, variability in model performance could directly be attributed to methodological experimental factors.

Model Construction:

Figure 3.1 outlines the basic procedure used to build all models. Feature selection was performed on the training data followed by model training after which the model was validated using the independent test set. Four different regression methods were used for model development including two linear methods; principal components regression (PCR) and support vector regression with a linear kernel function, and two non-linear methods including non-linear support vector regression (NLSVR) and artificial neural networks (ANN). We implemented 3 different feature selection strategies with all four algorithms and an additional seven on our best performing linear model (PCR) and non-linear model (NLSVR). These feature selection methods are

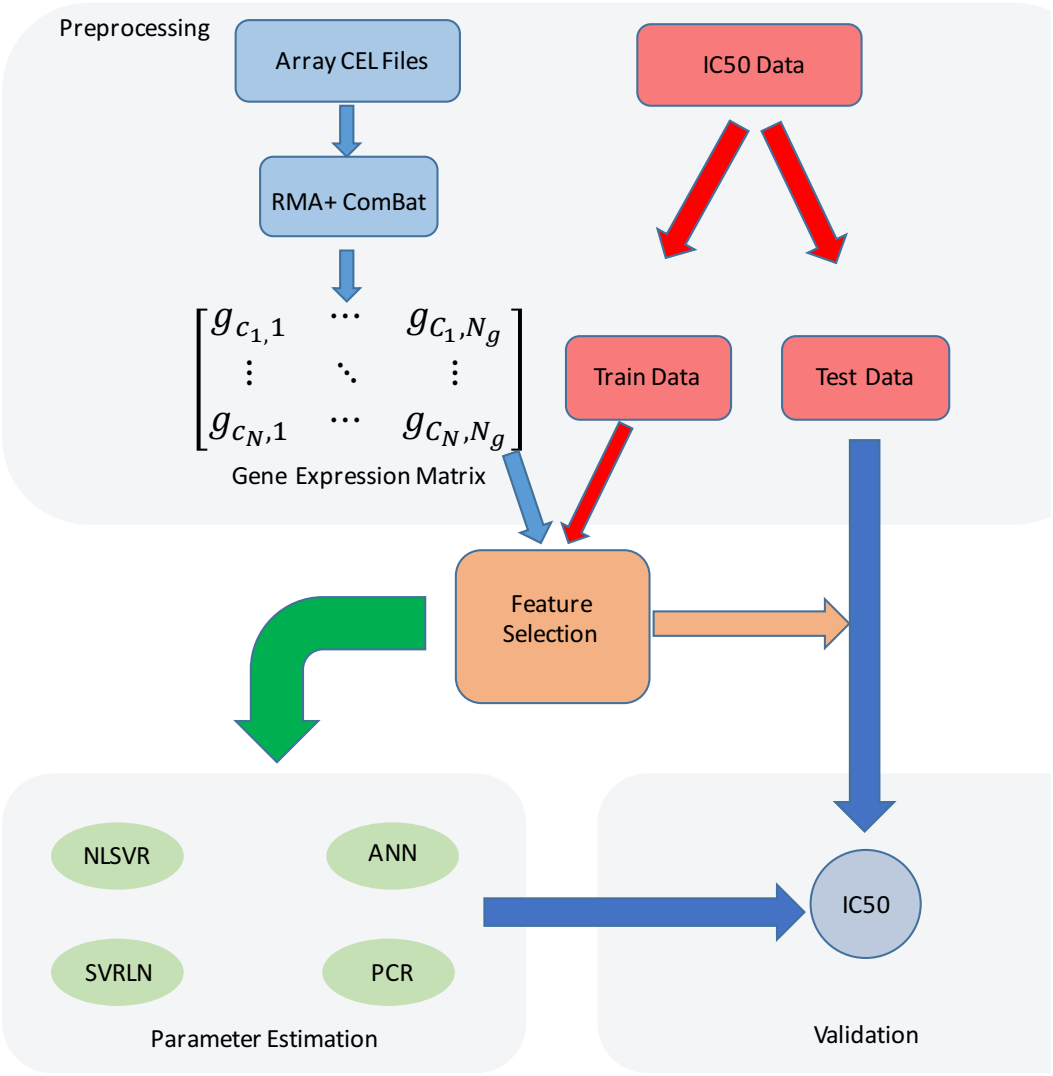


Figure. 3.1 General workflow: The general workflow used to build models.

summarized in Table 3.2. Feature selection was performed in python 2.7 and a generic python 2.7 script was used to read, organize, and write the model output. PCR was implemented in R version 3.2.4 using the PLS package with the number of components chosen by 10-fold Monte Carlo cross validation. Both NLSVR and SVRLN were implemented with scikit-learn version 18.1 (43). For NLSVR parameter optimization was performed on three separate parameters amounting to 210 different three parameter combinations using 10-fold Monte Carlo cross validation. Likewise, SVRLN was

optimized over two parameters for 30 different combinations using 10-fold Monte Carlo cross validation. A single layer ANN with 20 hidden nodes was implemented using the Keras package in python. To combat over-fitting dropout was implemented using 10-fold Monte Carlo cross validation with dropout rates 0,10,25, and 50 percent of total nodes chosen by 10-fold Monte Carlo cross validation. All parameter optimization and model training was performed using only the training data and the independent testing data was used to assess model performance.

Gene expression data is inherently high dimensional, presumably, a given biological response, such as drug response, is influenced by a subset of the total genes. Feature selection provides a means to reduce the number of covariates systematically favoring features that are most relevant to the problem. This often leads to more favorable outcomes by eliminating features that only contribute to noise leading to a more robust signal and a decrease in over-fitting. Filter based feature selection attempt to associate a given feature (gene) to a targeted output (drug response) based on statistical inference. Many such of these algorithms exist for gene expression data (44) and contextually amount to looking at two or more populations (i.e. drug resistant, drug sensitive cells) and determining if a given feature is statistically different between groups. Such methods are often applied to classification problems but can be generalized to continuous responses by looking at populations with distinct responses. However, this method requires reformatting the problem into a binary classification problem and assuming it can be generalized to a continuous response. Alternatively, correlation based feature selection methods (CBF) are more aptly suited to continuous

processes by looking at the statistical relationship between a covariate and target variable based on correlation (45).

To assess the effects of reducing features in our models we used several CBF feature selection methods. First we implemented the non-parametric Spearman correlation using a cutoff of $p<0.05$ to determine as set of differentially expressed probes (DEGs) using the statistics package in scipy 0.17.0. We compared this to a standard method of isolating probes with distinct difference between the 25% of cells with the greatest IC50s (resistant) and the lowest IC50 (sensitive) using the R Limma package (46) with a false discovery rate $q=0.05$. In order to assess the influence of feature selection we performed three control experiments. For the first control (CTR1) we randomly selected a number of probes that corresponded to the the same number of DEGs for a given experiment. The second control (CTR2) consists of all probes that are not selected as DEGs. Lastly, we performed a random control (RCTR) by shuffling the gene array matrix leaving the response vector untouched and then random selecting the same number of probes used in DEGs and CTR1. We addressed multiple testing by using a Bonferroni correction for p cutoff in the spearman correlation. Additionally, we explore a bootstrapping method to decrease false discovery rate (FDR). Lastly we applied a maximum relevance minimum redundancy (MRMR) algorithm (47). All feature selection methods were applied to the training set prior to model fitting. A summary of the different feature selection methods as summarized in Table 3.2.

Table 3.2: Feature selection methods. A summary and definition of the different feature selection methods discussed in the results section. The abbreviations that will be used in the text to refer to these methods are in prentices.

Selection Method	Description
No feature selection (NO FS)	All probes used with a total of 49386 probes.
Differentially Expressed genes (DEGs)	Array probes that have a statistically significant Spearman correlation P <0.05 with drug response
LIMMA	Linear Empirical Bayes with a modified t-statistic as implemented in the LIMMA Bioconductor package in R. Genes were selected by running LIMMA on the top and bottom 25% sensitive and resistance cell lines. A false discovery rate of 5% was chosen as a cutoff.
Bonferroni Correction (BC)	Bonferroni Correction $\rho_{BC} = \frac{\alpha}{m}$ where α is significance level of 0.05 and m is the number of features tested, 49386. $\rho_{BC} = 1.0\ x\ 10^{-6}$
DEG Bootstrap (BS)	Array probes which have a statistically significant Spearman correlation P<0.05 in fifty random subsets containing 75% of the training data
Histotype specific Bootstrap (BS-Hist)	50 subsets of the training data were generated such that each subset contained only one cell from a specific histotype. Probes that have a significant Spearman correlation P<0.05 in 50% of the splits were selected. ** Data not shown, reported in supplementary materials
Maximum Relevance Minimum Redundancy (MRMR)	Maximum Relevance Minimum Redundancy. 1000 Probes are chosen such that they have a maximum correlation with drug response with minimal cross-correlation with other chosen probes.
Control 1 (CTR1)	Probes are randomly selected from all 49836 probes equal to the number of DEGs for each model/trial. For example, bleomycin dataset 1 yielded 5377 DEGs in DEG feature selection thus 5377 probes are selected randomly in control 1 experiments.
Control 2 (CTR2)	The compliment of DEGs. For example, for bleomycin dataset 1 control 2 genes would include 38,009 probes excluded form the 5377 probes selected as DEGs.
Random Control (RCTR)	A number, N, of probes equal to the number of DEGs are randomly selected. This gives N vectors with each entry corresponding to a cell line in the training set. This vector is then shuffled randomly such that the original value is no longer associated with the same cell yielding a feature matrix that is arbitrary.

Histotype Only (HIST)	Each cell line is associated with a 55 dimensional vector where the nth entry is 1 if the cell comes from the corresponding histotype and 0 otherwise. (One hot encoded)

Analysis:

The performance of each model was assessed using the Spearman correlation coefficient between the predicted and measured IC50 values in the testing set using the scipy statistics package version 0.17.0. p values were calculated within the statistics package using a student's t distribution. Additionally, we also calculated a Mean Absolute Difference metric (MAD). The MAD scores were generally reflective of the Spearman correlation; therefore, we have chosen to report the Spearman correlation, as it better highlights particular patterns in the data in the main paper but MAD values for all models can be found in the supplementary materials. K-means clustering was performed using the clustering package in scikit-learn (43). The ability of a given set of genes to assign cells of the same histotype to the same cluster was determined using Clustering Entropy, S_c, (48) which is defined and conceptually illustrated in Figure 3.2. S_c has a minimum value of 0 when histotypes are perfectly clustered together. A theoretical maximum S_c occurs when each cluster contains a uniform distribution of samples from different histotypes, however, since samples are not uniformly distributed across histotypes and each dataset contains a different distribution of histotypes the maximum value was estimated using the random control for each dataset and the values reported are normalized consisting of the S_c of the given dataset divided by the S_c of the random control. Note that by this definition the normalized S_c can be greater than one.

k	p_{1k}	p_{2k}	p_{3k}	s_{Lk}
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0

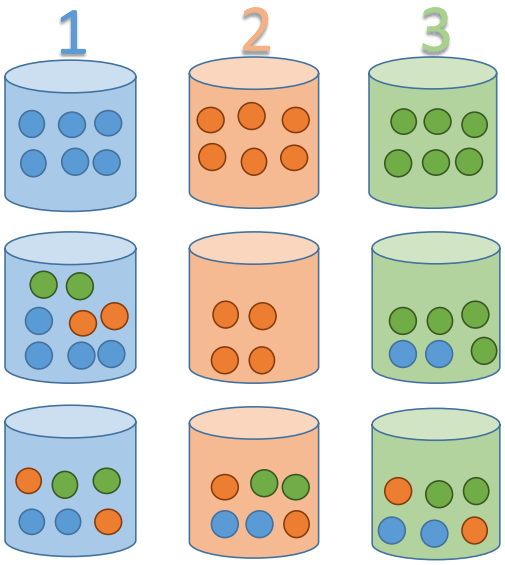
$S_c = 0$

k	p_{1k}	p_{2k}	p_{3k}	s_{Lk}
1	4/8	2/8	2/8	1.5
2	0	1	0	0
3	2/6	0	4/6	0.918

$S_c = 0.971$

k	p_{1k}	p_{2k}	p_{3k}	s_{Lk}
1	2/6	2/6	2/6	1.585
2	2/6	2/6	2/6	1.585
3	2/6	2/6	2/6	1.585

$S_c = 1.585$



$$S_c = \sum_1^K \frac{n_k}{N} s_{Lk} \quad : \quad s_{lk} \sum_1^L -p_{Lk} \log_2 p_{lk}$$

Fig 3.2. Cluster Entropy: Illustration of how cluster entropy, S_c, is calculated. It is a measure of cluster homogeneity, in this case, how many cells of the same histotype are placed in the same cluster.

Results:

Regression models:

Individual Spearman correlations between measured and predicted IC50 values ranged from 0.64 to -0.345 with 51%-84% percent of the models showing significance ($P<0.05$). While NLSVR (0.316-0.331) yielded higher average Spearman correlations than PCR (0.297-0.316) and SVRLN (0.27-0.285), the difference on a per drug basis was minimal (Figure 3A). ANN showed significant drops in performance (0.144-0.266) compared to the other three methods especially when no feature selection was performed, while, the gap narrowed upon the introduction of feature selection,

performance was still substantially less, most notably when compared with NLSVR and PCR (Figure 3A, Table 3.4.).

Correlation based feature selection ($P<0.05$) decreased the number of features by an average of 77% (range 39% to 96%) with the fewest features for cisplatin and most for vorinostat Table 3.3. Model performance was increased for ANN (63%) increasing the average Spearman correlation by 63 percent with only a modest increase for NLSVR (1.5%) and PCR (1.6%). The decrease in features had a minor negative impact on SVRLN (9%) performance. Feature selection by use of the R package Limma was substantially more restrictive than the DEG criteria, leading to an 99% decrease feature number, yielding no features for cisplatin. Despite this substantial decrease in genes, only a 9% average decrease in correlation was observed with similar effects to NLSVR, PCR, and SVRLN (~ 11%) and minimal effects to ANN (1.6%) in comparison to the top performing feature selection method.

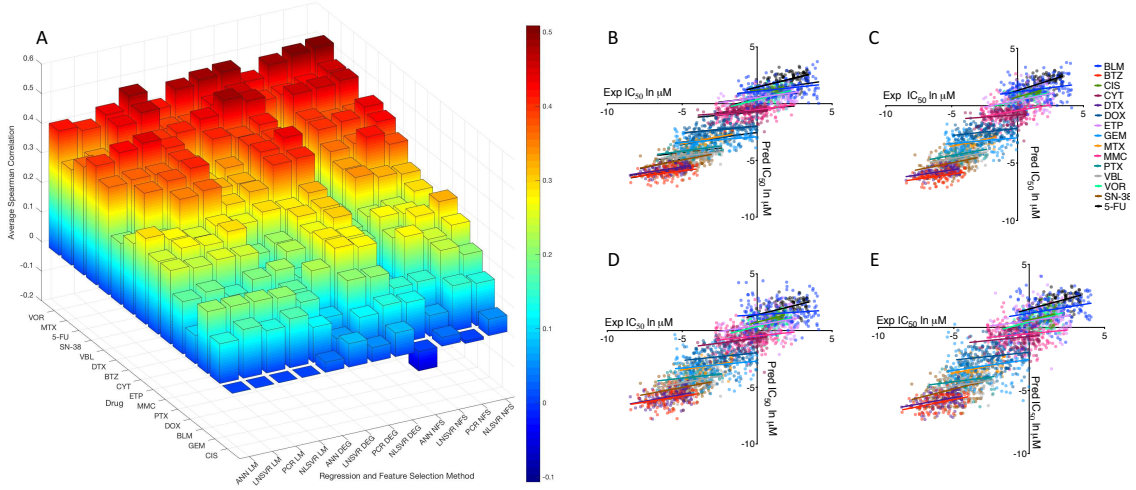


Fig 3.3 Model performance by method and drug: (A) Average spearman correlation coefficients for four different regression methods over three different methods of feature selection. (B-E) Predicted versus Measured IC50 values for each of the fifteen drugs using DEG genes. (B) NLSVR, (C) PCR, (D) SVRLN, (E) ANN

Table 3.3. Model Performance: Average spearman correlations across six different testing sets for all regression and feature selection methods. This data is graphically displayed in Figure 3.3

	NLSVR			PCR			LNSVR			ANN		
	NFS	DEG	LIM	NFS	DEG	LIM	NFS	DEG	LIM	NFS	DEG	LIM
BLM	.207	.202	.202	.239	.208	.208	.151	.1	.209	.147	.17	.21
BTZ	.38	.404	.365	.422	.399	.354	.332	.326	.232	-.009	.299	.24
CIS	.05	.08	N/A	-.009	.047	N/A	.03	.079	N/A	-.066	.034	N/A
CYT	.313	.32	.279	.32	.281	.256	.337	.291	.269	.226	.266	.291
DTX	.422	.44	.408	.367	.409	.382	.357	.319	.359	.185	.318	.207
DOX	.273	.27	.117	.243	.285	.106	.27	.226	.103	.115	.173	.096
ETP	.289	.302	.294	.248	.291	.263	.238	.219	.273	.209	.195	.246
GEM	.143	.139	.166	.153	.117	.143	.07	.063	.165	.131	.119	.134
MTX	.461	.455	.462	.431	.435	.433	.417	.388	.338	.411	.391	.322
MMC	.237	.302	.244	.264	.269	.25	.27	.224	.239	.203	.153	.248
PTX	.32	.27	.198	.287	.282	.159	.233	.170	.191	-.106	.211	.177
VBL	.44	.403	.399	.408	.398	.37	.398	.339	.371	.112	.302	.363
VOR	.509	.495	.486	.5	.487	.439	.484	.471	.404	.445	.42	.42
SN-38	.383	.417	.409	.379	.391	.443	.397	.404	.429	.01	.327	.402
5-FU	.463	.464	.40	.455	.484	.354	.451	.438	.337	.309	.409	.365
AVG	.326	.331	.316	.314	.319	.297	.285	.27	.28	.144	.252	0.266

The results from BC, BS, and MRMR models for NLSVR and PCR in (Figs 4A and 4B). Compared to DEG and NO FS models, all three methods yielded lower average Spearman correlations. BC criteria reduced features by an average of 98.6% with no selected features for cisplatin datasets as well as two datasets for bleomycin and a single dataset for doxorubicin (Table 3.3). The use of BC selected features decreased the overall average Spearman correlation by 11.4% (0.3513 to 0.3112) for NLSVR and 12.2% (0.3406 to 0.2991) across identical datasets using DEG selected features. The most dramatic decreases in performance was seen in bleomycin, cytarabine, doxorubicin, and 5-fluorouracil (Figure 3.4 A and 3.4 B). A small increase in performance was seen for methotrexate in NLSVR models. Despite the decreased

Performance, 80 percent of the models had significant correlations ($P<0.05$) between experimental and predicted IC50 values.

Bootstrap methods resulted in an average 95% decrease of features. Performance decrease was slightly less than that of BC selected genes with an average decrease of in NLSVR (3.4%) and PCR (4.4%) models. A substantial decrease in performance was observed for cisplatin while an increase in performance was seen in etoposide, gemcitabine, and paclitaxel in comparison with DEG models for both NLSVR and PCR (Figure 3.4). Likewise, the modified MRMR algorithm was used to select 1000 features, representing a 98% decrease in features. The drop in performance was similar to that seen with both BC and BS for NLSVR (6.3%) and PCR (6.9%). The general decrease in performance correlated directly with the reduction in the number of genes; however, even a maximum 98.6% decrease in features only resulted in 11.4% drop in performance for NLSVR and 12.2% for PCR. Additional methods of feature selection that attempted to take the histotype into account yielded similar performances (Appendix A).

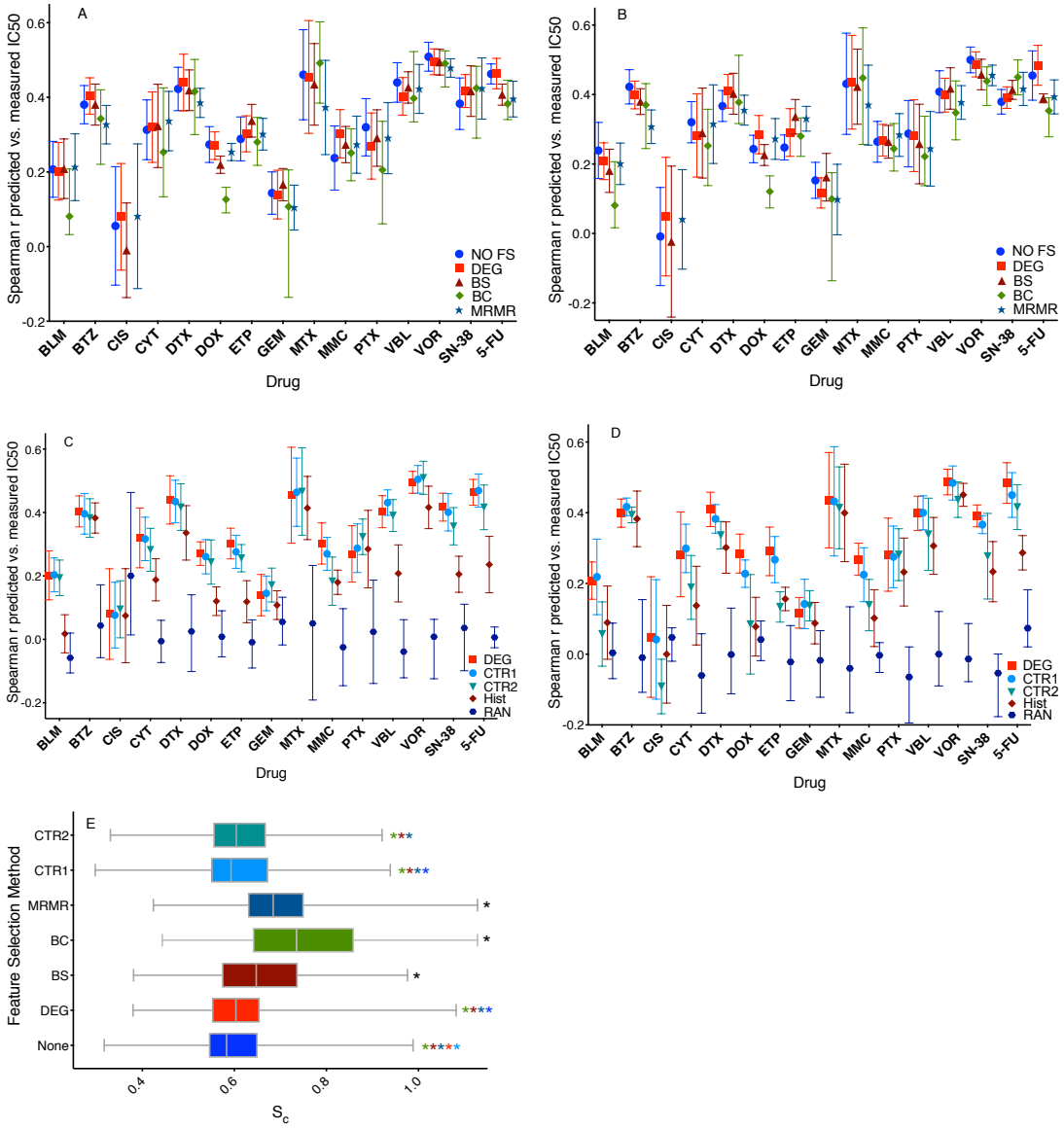


Figure 3.4. Feature selection methods and controls: (A-B) spearman correlation Coefficients for different feature selection methods NLSVR (A), PCR (B). (C-D) spearman correlation coefficients for control models NLSVR (C), PCR (D) the placement of the symbol indicates the mean with the ends representing the range. (E) Cluster Entropy (S_c), indicative of how well cell lines of the same histotype cluster using k-means. comparable S_c as well as little difference in r indicate that histotype recognition drives model performance. S_c is relative to the random control (RCTR) where S_c=1, perfect clustering would have a S_c=0. The asterisks indicate significance ($p<0.05$) between the method and alternative models, indicated by color (black indicates is was significant compared to all other methods), using a non-parametric Wilcoxon match-paired rank

Influence of CBF feature selection:

In order to gain insight into the overall influence of correlation based features we tested several sets of control features on the datasets, designed to address the following questions. First, what was the benefit of using DEGs compared to the same number of randomly selected genes (CTR1)? Second, how influential was the inclusion of correlation based features, thus, what would be the effect of using those genes that had no significant relationship to drug response (CTR2)? Lastly, were these relationships simply an artifact that was introduced during the collection, preprocessing, and normalization of the data, and thus what happens if all causal relationships are removed (Random Control)?

The use of CTR1 genes resulted in a decreased performance <1% (0.331 to 0.329) for NLSVR and a 3.1% (0.319 to 0.309) for PCR in average Spearman correlation. With respect to DEGs in NLSVR, CTR1 genes led to comparable average Spearman correlations for each drug and exceeded DEGs in certain drugs such as methotrexate, paclitaxel, vinblastine, and vorinostat (Figure 3.4 C.). Likewise, for PCR, small increases in average Spearman correlation was seen for bleomycin, bortezomib, cytarabine, and gemcitabine while a minimal decrease for other drugs (Figure 3.4 D.). Surprisingly, removing features with a-priori significant statistical relationships with drug response had little overall negative effects on the average performance of NLSVR models (4.8%) with cisplatin, gemcitabine, paclitaxel, and vorinostat yielding better average performances than the same DEG models (Figure 3.4 C.). However, the performance of CTR2 models in PCR dropped significantly by 26% (0.319 to 0.236) compared to DEG models, however, 64% of the models had significant correlations.

Nonetheless, comparable performances were observed in several drugs including bortezomib, docetaxel, methotrexate, vorinostat, 5-fluorouracil, and gemcitabine while other drugs such as bleomycin, docetaxel, cisplatin, and SN-38 saw dramatic decreases in performance (Fig 3.4 D). Lastly, by randomly assigning expression values to cell-lines (Random Control), there was a significant loss in the predictive ability of the model with average Spearman correlations of 0.0185 for NLSVR and -0.007 for PCR (Figure 3.4 C and 3.4 D.). The loss in predictive capability when the gene-cell line relationship is removed demonstrates that our models are clearly capturing a genomic signature that is indicative of drug response.

Histotype is linked to drug response:

Several of the drugs cell line predictions of the same histotype tended to cluster together as illustrated with vorinostat (Figure 3.5 D.) suggesting that histotype might be predictive of drug response. In order to ascertain if there was an actual differential drug response between histotypes, we performed pairwise F-tests between drug responses categorized by histotype. The number of significant pairwise comparisons ranged from a low of 5.1% for bleomycin (Figure 3.5 A.) to 52.6% for vorinostat (Figure 3.5 B.) with an average of 24.1%. Furthermore, the Spearman correlation between the percentage of significant F-tests and the average Spearman correlation for the two control models was 0.85 and 0.88 for NLSVR and 0.84 and 0.86 for PCR on CTR1 and CTR2 datasets respectively.

To establish the influence of histotype on model performance it had to be shown that histotype could predict drug response, and that any feature selection methods yielded features with the equal ability to distinguish one histotype from another. To

accomplish this a 55 dimensional feature matrix was developed to encode cell line identity to one of the 55 possible histotypes represented in the data using one hot encoding. Then using this feature matrix NLSVR and LSQR were applied to predict drug response. In both NLSVR and LSQR there was a significant drop in the average Spearman correlation, 0.2193 for NLSVR and 0.2218 for LSQR. However, 61-62 % of the models gave significant correlations in both NLSVR and LSQR (Fig 3.4 C and 3.4 D). For several drugs, such as bortezomib, cisplatin, docetaxel, gemcitabine, methotrexate, paclitaxel, and vorinostat gave comparable results. Whereas for others, such as bleomycin, doxorubicin, 5-fluorouracil, and SN-38 substantially lower correlations were obtained. The ability of histotype to predict response is best illustrated between bleomycin (Figure 3.5 A, 3.5 C, and 3.5 E.) and vorinostat (Figure 3.5 B, 3.5 D, and 3.5 F.). Bleomycin has minimal differential drug response between histotypes (Figure 3.5A.) as a result, when given nothing but histotype as input the model will have a tendency to predict the average IC50 values of a given histotype. In the case of bleomycin the average IC50 values of different histotypes do not exhibit a great amount of variability and thus the predictions collapse to the overall average of the data (Figure 3.5 E.). Alternatively, for drugs such as in vorinostat, the histotype average IC50 values exhibit a greater amount of variability (Figure 3.5A.) and as a result this variability is reflected in the predictions (Figure 3.5 F.). Furthermore, the variability of average histotype responses is roughly captured when the features are reduced to only indicators of histotype in drugs such as vorinostat (Figure 3.5 D.) where this is absent in bleomycin (Figure 3.5 C.)

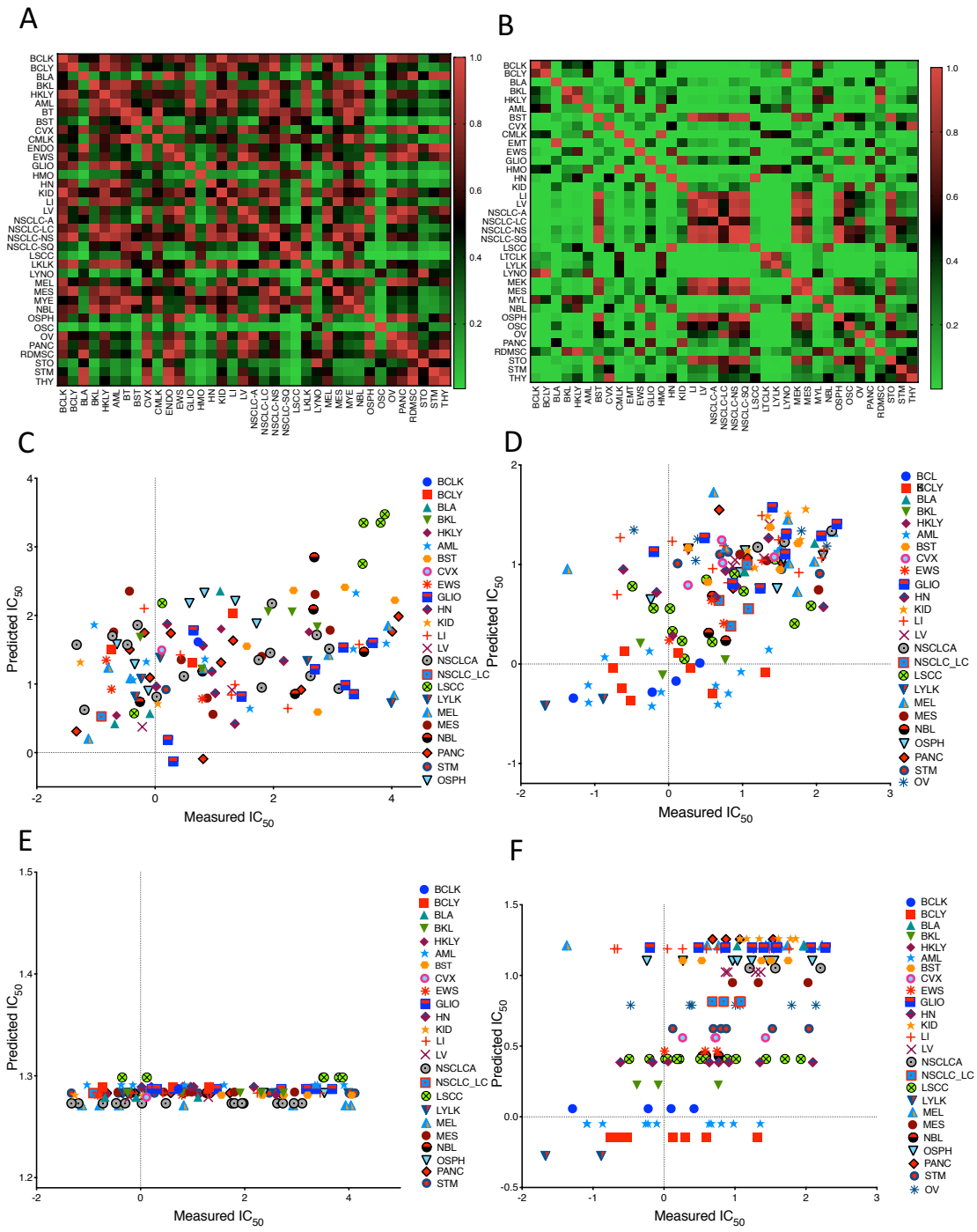


Figure 3.5. Histotype influence on drug response. (A-B) P values for pairwise F-tests between histotype IC50 for Bleomycin (A) and Vorinostat (B). (C-D) Measured vs Predicted IC50 using DEGs for Bleomycin (C) and Vorinostat (D). (E-F) Measured vs. Predicted IC50 for Hist models in Bleomycin (E) and Vorinostat (F). Each symbol color combination indicates a different histotype.

To explore the ability of a given set of features to identify histotype we used k-means clustering to cluster the cells into one of 55 groups and then used cluster entropy, $S_{c,}$ to quantify the consistency of which cells of the same histotype were placed in the same cluster. A pairwise non-parametric Wilcoxen paired T-test showed that there was no significant difference between DEG, CTR1, and CTR2 genes, and while S_c for NO FS was statistically significant it is not apparent if there is a meaningful difference as the average absolute difference was only 8.5% (Figure 3.4 E.). Additionally, while BS, BC, and MRMR had higher average $S_{c,}$ 100% of BS models, 98.8% of MRMR models, and 86.7% of BC models clustered by histotype better than a randomized model. Therefore, the data suggests that the predictive ability of the model is partially dictated by the ability of a set of features to recognize similar histotypes as well as the variability between drug responses between histotypes.

Model performance, number of features, histotype recognition:

To determine how the number of genes affected the performance, if genes statistically linked to drug response became a bigger factor as the number of features decreased, and how both of these affected the ability to cluster cells based on histotype, we constructed models for both NLSVR and PCR using 10, 55, 250, 500 ,1000 randomly selected features from DEGs, CTR1, or CTR2 as well as performing k-means clustering.

As expected, a decrease in overall performance was observed as the number of features decreased. However, the magnitude of performance drop was considerably different depending on the feature selection method. In NLSVR the performance of DEG models dropped by 36%, CTR1 models dropped by 61%, and CTR2 models dropped by

71% (Figure 3.6 A, 3.6 B, 3.6 C, and 3.6 G.). Likewise, PCR models decreased by 35.2%, 51% and 70% in DEG, CTR1 and CTR2 models respectively (Figure 3.6 D, 3.6 E, 3.6 F, and 3.6 H). Furthermore, DEG feature selection in both PCR and NLSVR models are reasonably robust down to 250 features, with NLSVR exhibiting only a 7.5% difference and PCR only 9.1% at 250 features compared to 1000 features (Figure 3.6 A, 3.6 D, 3.6 G, 3.6 H). Likewise, in CTR1 models, only a 7.1% decrease in NLSVR and 11.9% decrease in PCR (Figure 3.6 B, 3.6 E, 3.6 G, and 3.6 H). CTR2 models exhibited a decrease approximately twice as great (14.4% NLSVR and 26% PCR) as that seen with DEGs or CTR1 features (Figure 3.6 E, 3.6 F, 3.6 G, and 3.6 H).

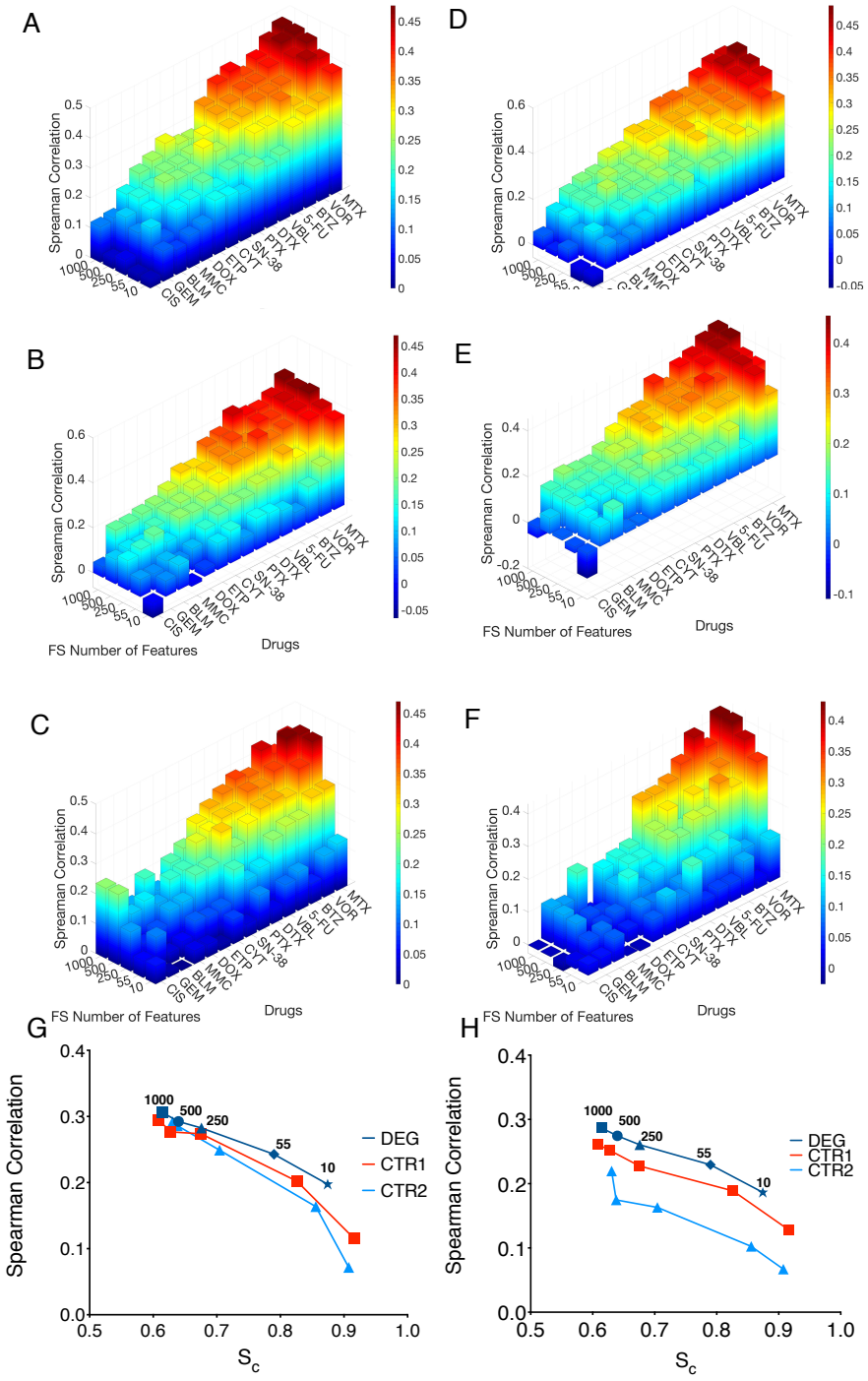


Figure 3.6. Model performance and number of features. Average Spearman correlations for all 15 drugs as a function of features used for NLSVR DEG (A), NLSVR CTR1 (B), NLSVR CTR2 (C), PCR DEG (D), PCR CTR1 (E), PCR CTR2 (F). average S_c vs average Spearman correlation with each symbol representing the number of features used for NLSVR (G) and PCR (H).

For NLSVR models the difference in S_c is minimal down to 500 features for all three feature selection methods and down to 250 genes for DEG and CTR1 features. As a result, there is not a substantial difference in performance down to 500 or less features (Fig 6G). Additionally, for DEG and CTR1 genes the difference in S_c is only about 7% going from 1000 to 250 features (Fig 6G). The drop in S_c reflects a loss in a given set of features to identify histotype and is coupled directly to a loss in performance. Furthermore, substantial performance differences between DEG and CTR1 features began to occur at 55 genes which is also marked by a substantial increase in S_c between DEG and CTR1 genes, and begins happening at 250 CTR2 features (Fig 6G). PCR models exhibit a clear discrepancy between all three feature selection methods; however, as the difference in S_c increases between methods the difference in performance grows as well consistent with the idea that S_c, and therefore, histotype, has a substantial influence on model performance (Fig 6H). Lastly, as the number of features is dropped considerably, DEG features maintain more histotype specificity which suggest that features which are highly correlated to drug response are also highly correlated with histotype.

Comparison with DREAM:

The DREAM-NCI project assessed the performance of each model using a modified concordance index which they called the weighted-probability concordance index (wpc-index). Given the vast diversity of models evaluated we wanted to determine if the models we used were comparable using the wpc-index. The average wpc-index was 0.576 with a range from 0.552 to 0.582 for NLSVR and 0.569 for PCR ranging from 0.552 and 0.58 (Table 5). For Both NLSVR and PCR methods the models with the

highest wpc-index were DEG models while the lowest performing model were the histotype-only models. The Spearman correlation of the wpc-index with average Spearman correlation was 0.9833 (p=0) for NLSVR models and 0.95 (p=0) for PCR models. The top models in the DREAM-NCI paper have wpc-index scores of 0.583, 0.577, and 0.57, and the minimum score was calculated to be 0.485. While, based on wpc-index, we had no models out-perform the top performing model, four NLSVR models had wpc-index scores that would place them in second place, and all, with the exception of the HIST model, scored within the top three. Likewise, PCR, had two models that scored above the second place model and six that placed above the third place model Table 3.3.

The DREAM-NCI project consisted only of a single cancer histotype (breast) and thus, the histoytpe phenomena driving the performance of our models is not a contributing factor in their models. The testing set for the DREAM-NCI project consisted of 18 cells lines, the number of cells of a single histotype did not exceed 20 and was often below 10 for any test split. However, non-small cell lung carcinoma adenocarcinoma (NSCLC-adenocarcinoma) was represented with 10 or greater cells in 10 of the 15 drugs and 43 % of the total testing data sets. Thus, in order to gauge if the models were picking up some cell specific drug response within a histotype we used the WPC index to score DEG, CTR1, CTR2, and No FS models. Compared to wpc scores for our pan-cancer models and several models in the DREAM project the WPC index was smaller ranging from 0.5346 for DEG Models to 0.5084 for CTR2 models (Table 3.3). Considering the variable number of cell lines for each dataset, we assessed the significance by creating a null distribution of 3000 randomly constructed permutations of

the modeled data. DEG and No-FS had wpc scores that significantly differed ($p<0.05$) from what would be expected by random permutation with a wpc value of 0.5. This suggests models which include genes relevant to drug response have some ability to pick up variability in individual histotypes, while genes with no apparent significantly statistical relationship with drug response fail to pick up that variability indicated by having a wpc score consistent with a random permutation of the data.

Modeling the NCI60:

The NCI60 results were highly variable due to the low sample size, but many of the trends that emerged in the GDSC were also evident in the NCI60, mainly, there was not a significant difference in performance between NLSVR and PCR, and minimal difference between selected features (NOFS: 0.4, DEG: 0.403, CTR1: 0.399, CTR2: 0.351) for SVR and (NOFS: 0.412, DEG: 0.406, CTR1: 0.382, CTR2: 0.35) for PCR. One of the more interesting points is that models performance still has a significant relationship with histotype as evidenced by significant correlation histotype models and models constructed with genomic feature with correlation ranging 0.4114 to 0.4576 for NLSVR and 0.2988-0.4547 for PCR (Figure 3.7 A. and B.) This is relationship is even stronger in the GDSC (NLSVR:0.6878-0.7341, PCR: 0.663-0.733) due to the increased number of cell lines and histotypes (Figure 3.7 C. and D.). It is also important to note that the GDSC and NCI60 share 38 cell lines, however, the range is smaller in the drugs we modeled (7-30). For the 14 common drugs with data in the NCI60 and GDSC only 7 had significant correlations in drug response for identical cell lines.

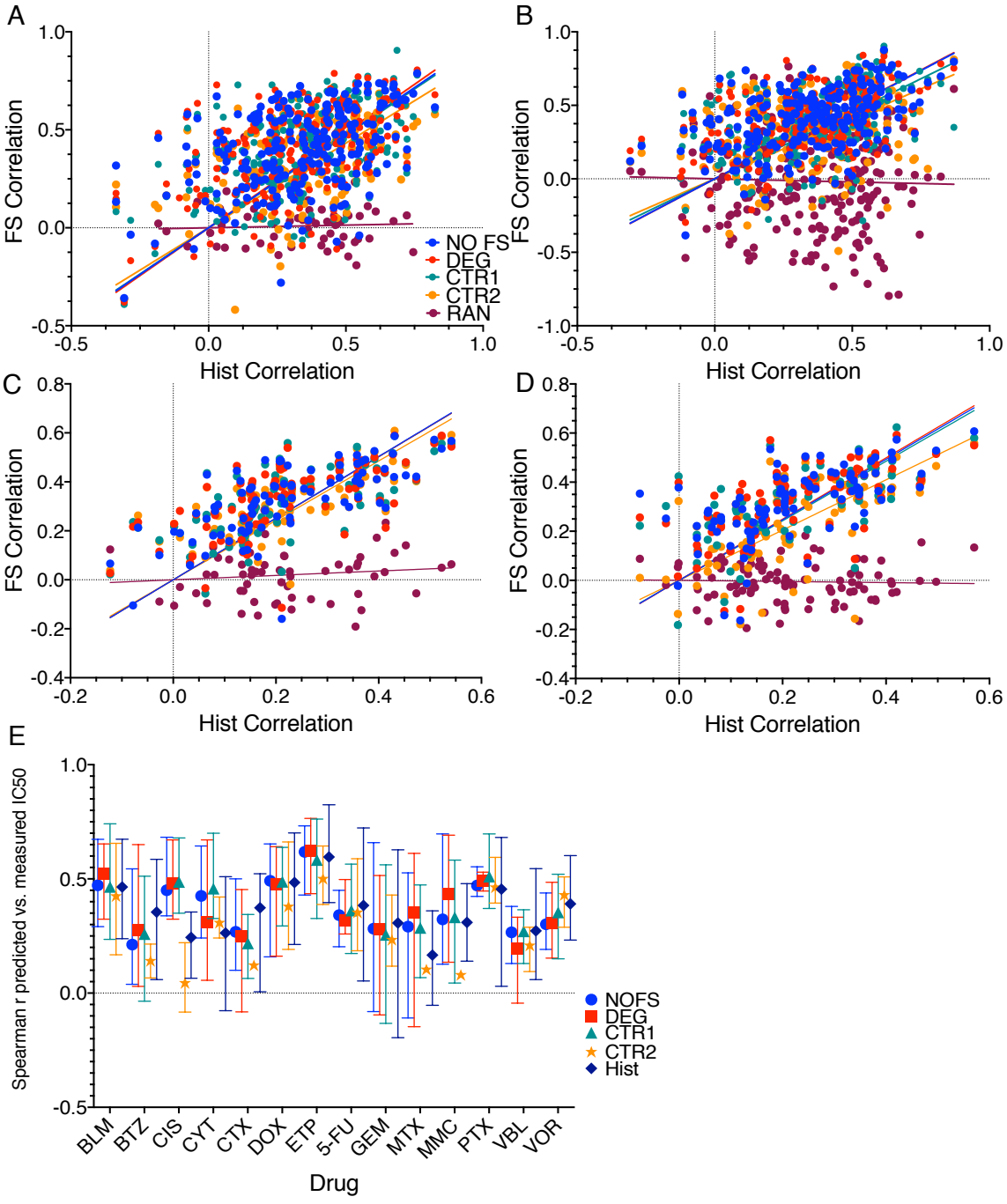


Figure 3.7. NCI60 models vs the GDSC models: (A-D) Spearman correlations for a given feature selection method (FS correlation) verses Spearman correlations for one hot encoded histotype models for NCI60 NLSVR (A), NCI60 PCR (B), GDSC NLSVR (C), GDSC PCR (D). E. NCI60 models for the 14 drugs with data both in the GDSC and NCI60.

Discussion:

The development of molecular tools allows for a unique look into the molecular nature of cancer inspiring a community effort to collect data with the potential for significant clinical impact. Given the complexity and amount of data that is available and continues to be generated, computational approaches are necessary to fully utilize the available information present in the data. As these computational techniques become more advanced, a basic understanding of the factors that influence model performance are essential. We have taken a systematic approach to characterize the influence of basic model complexity in terms of the linearity of the model and basic CBF feature selection methods to predict *in vitro* drug response across a number of cancer cell lines in the GDSC and NCI60. Our results suggest that the complexity of the model and the method of feature selection have marginal effects on the performance of the model with performance largely dictated by the relationship between histotype and drug response.

With the exception of ANN, it is not straightforward to establish which model is superior as it is certainly reasonable that more thorough parameter optimization and different data splits might lead to fractional improvements for one method over others changing the relative rank of each model while having insignificant meaningful quantitative gain. Likewise, there is no substantial gain by eliminating features that do not significantly correlate with drug response. The control experiments (CTR1 and CTR2) as well as the histotype models suggest that histotype has a major influence on predictive capability of model. More rigorous criteria as imposed by Bonferroni correction and bootstrap methods tend to decrease model performance slightly and this is accompanied by a similar decline in selected features ability to cluster cell lines by

histotype as evidenced by a larger S_c. Attempts to remove redundancy in features using MRMR results in higher S_c values suggesting that the histoytpe signal can be somewhat mitigated by removing redundant features but is also accompanied by a decrease in overall performance. Furthermore, even at 500 random features a diffuse histotype signature is maintained which maintains the majority of drug-response information sufficient for producing predictive models.

Our best performing model consisted of Support Vector Regression using a radial basis function. However, the improvement over the best performing linear model, PCR, was only a 3.8% increase in average Spearman correlation. Likewise, the DREAM competition concluded that non-linear models performed slightly better than linear models; however, the performance increase between their top non-linear model and top linear model was only 1.5% with several linear models performing better than many other non-linear models developed (21). Artificial neural networks performed consistently the worst, for comparison, Menden et al. used ANN to build predictive drug response models for 608 cell lines and 111 drugs in the GDSC using genomic and chemical properties of drugs reporting an overall Pearson correlation of 0.85 across all drugs. However, the individual drug correlations ranged roughly from -0.15 to 0.5 similar to the results we achieve (49). The discrepancy in Menden's work between the overall correlation and the individual drug correlations is most likely due to the spread of IC50's across drugs as different drugs have distinct ranges of IC50 values, this can clearly be seen in (Figs 3 B-E) demonstrating how inherent data structure, i.e. different ranges of drug response for different drugs, can introduce artifacts that can potentially affect both the construction and analysis of models. Other modeling strategies in the GDSC that

have focused on targeted agents have produced average spearman correlations slightly higher, (approximately 10%) (50) while models incorporating Bayesian components have led to significantly lower average Spearman correlations (around 50%) (51).

Our results show that neither the linearity of the regression method nor features used have a strong influence on performance with the single most influencing factor being the identity of the drug. Furthermore, this is consistent across multiple data-bases and over many cytotoxic agents despite inconsistencies seen among cell lines shared by the GDSC and NCI60 that could prove a barrier to using the GDSC to train and validate a model and test on the NCI60 or likewise the NCI60 to train and validate testing on the GDSC. This phenomenon results from the tendency that cancers from the same histological background respond similarly to certain drugs. This is reflected in our results; predictive outcomes can be achieved in most drugs simply by identifying the histotype. Consequently, any gene set that has the ability to differentiate histotype also can generate predictive models as demonstrated with our control models. Often the identification of histotype is an essential step into determining specific approaches to successful treatment, clearly not all cancers of the same histotype respond precisely the same to a given drug. The range of responses might have important consequences when it comes to determining effective PKPD parameters for clinical applications. Furthermore, with respect to modeling, this “histotype” effect potentially shields features that have significant predictive capability across all histotypes where the signal to noise ratio is significantly less compared to features that have strong associations with histotype.

Several successful models have been built to classify tumors histologically using genomic profiling (3, 5, 52, 53) demonstrating the ability of statistical learning techniques to learn tissue specific features. Thus, given the differential drug response of cancers with similar histological background the prediction of response loosely defaults to a classification exercise. This simultaneously presents opportunity and challenge. Knowing the histotype, therefore, gives a significant amount of information about the drug response. However, histotype accounts for a large amount of genomic variation as well as variability in drug response. Therefore, feature selection results in the convolution of three possible categories: features that account for variability in histotype having no influence on actual drug response, features that account for variability in histotype and drug response, and lastly features in which variability is exclusively a result of drug response. This is a challenging task, filter methods, such as CBF and mutual information, tend to pick more robust signals associated with drug-histotype interactions. The ability to extract drug response within a histotype and then leverage that information across histotypes, such as ensemble methods, might be a reasonable approach. For example, a filter based feature selection method could be applied on each histotype independently then those features present in all histotypes could be pooled. Additionally, a multiple kernel learning (MKL) method where each individual kernel is applied to a distinct histotype might be an effective way to pool multiple histotype based models into a more generalized pan cancer model. However, number of samples of each histotype in most databases, such as the GDSC, could be a limiting factor for producing robust features in a filter based method or parameter optimization in a MKL setting. Databases such as the NCI60, GDSC, Cancer Therapeutics Response

Portal (CTRP), and Cancer Cell Line Encyclopedia (CCLE) would provide for a broader diversity of data. However, several studies have shown inconsistency in drug response among cell lines derived from the same source (54-56).

Conclusion

The ultimate goal for these types of predictive models is to become a clinical tool that practitioners can utilize to improve the treatment of cancer patients or to inform clinical trials. While the jump from an *in vitro* cancer cell line to a tumor and then eventually a patient is a considerable progression these *in vitro* based experiments certainly add insight to the problem. Previous studies that have leveraged *in vitro* data to inform tumor based predictions have approached drug response as a binary variable, sensitive or resistant (15, 16). However, in such an approach valuable quantitative insight might be lost that could be critical to successful clinical applications. For example, an *in vitro* cell line might exhibit an IC50 that is much lower in comparison to other cell lines, implying sensitivity, but the concentration of drug needed to achieve a comparable exposure in a patient might not be reasonable due to pharmacokinetic or toxicity constraints. Thus, to more effectively use cell line drug exposure the ability to first accurately capture *in vitro* drug response is critical. What our models suggest is that similar pan-cancer cell based models might over emphasize a relationship between histotype and drug response thus could be misleading when applying such techniques to tumor data by effectively only capturing a broad histotype response failing to be applicable to more inter-tumor variability. Therefore, it is paramount that drug-histotype response is considered to improve model performance and utility.

Biological systems are inherently complex, noisy, and high dimensional which makes modeling their behavior a difficult task. Statistical learning allows for the extraction of valuable insights from large sets of data without direct knowledge of the intrinsic mechanisms that are influencing the properties of the system. For this reason, statistical learning provides several tools that are directly applicable to cancer diagnosis and treatment and it has been an active area of cutting edge research in cancer biology, mathematics, statistics, and computer science. Therefore, as the field moves forward it is absolutely imperative to understand how fundamental modeling considerations influence model performance on large complex biological datasets. Systematic approaches with well thought out control experiments are paramount to fully understand the complexities that arise when considering different modeling strategies.

REFERENCES

1. Schena M. Genome analysis with gene expression microarrays. BioEssays. 1996;18(5):427-31.

2. Schena M, Shalon D, Davis RW, Brown PO. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. Science. 1995;270(5235):467.

3. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000;403:503.

4. 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.

5. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science. 1999;286(5439):531.

6. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, et al. A gene expression database for the molecular pharmacology of cancer. Nature Genetics. 2000;24:236.

7. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Research. 2013;41(D1):D955-D61.

8. Collins FS, Varmus H. A New Initiative on Precision Medicine. New England Journal of Medicine. 2015;372(9):793-5.

9. Fessele KL. The Rise of Big Data in Oncology. Seminars in Oncology Nursing. 2018;34(2):168-76.

10. Yaser S. Abu-Mostofa MM-I, Hsuean-Tien Lin. Learning from Data A short Course: AMLbook.com; 2012
.

11. Andreu-Perez J, Poon CCY, Merrifield RD, Wong STC, Yang GZ. Big Data for Health. IEEE Journal of Biomedical and Health Informatics. 2015;19(4):1193-208.

12. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115.

13. Araújo T, Aresta G, Castro E, Rouco J, Aguiar P, Eloy C, et al. Classification of breast cancer histology images using Convolutional Neural Networks. PLOS ONE. 2017;12(6):e0177544.

14. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, et al. The Use of Molecular Profiling to Predict Survival after Chemotherapy for Diffuse Large-B-Cell Lymphoma. New England Journal of Medicine. 2002;346(25):1937-47.

15. Lee JK, Havaleshko DM, Cho H, Weinstein JN, Kaldjian EP, Karpovich J, et al. A strategy for predicting the chemosensitivity of human cancers and its application to drug discovery. Proceedings of the National Academy of Sciences of the United States of America. 2007;104(32):13086-91.

16. Fowles JS, Brown KC, Hess AM, Duval DL, Gustafson DL. Intra- and interspecies gene expression models for predicting drug response in canine osteosarcoma. BMC Bioinformatics. 2016;17:93.

17. Herzog TJ, Krivak TC, Fader AN, Coleman RL. Chemosensitivity testing with ChemoFx and overall survival in primary ovarian cancer. American Journal of Obstetrics and Gynecology. 2010;203(1):68.e1-.e6.

18. Peter S, Elke S, Katja N, Oliver C, Marcel R, Timo S. Prediction of individual response to chemotherapy in patients with acute myeloid leukaemia using the chemosensitivity index Ci. British Journal of Haematology. 2005;128(6):783-91.

19. Wakatsuki T, Irisawa A, Imamura H, Terashima M, Shibukawa G, Takagi T, et al. Complete response of anaplastic pancreatic carcinoma to paclitaxel treatment selected by chemosensitivity testing. International Journal of Clinical Oncology. 2010;15(3):310-3.

20. Williams PD, Cheon S, Havaleshko DM, Jeong H, Cheng F, Theodorescu D, et al. Concordant gene expression signatures predict clinical outcomes of cancer patients undergoing systemic therapy. Cancer research. 2009;69(21):8302-9.

21. Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nature biotechnology. 2014;32(12):1202-12.

22. Weinstein IB. Addiction to Oncogenes--the Achilles Heal of Cancer. Science. 2002;297(5578):63.

23. Weinstein IB, Joe A. Oncogene Addiction. Cancer Research. 2008;68(9):3077.

24. Sawyers C. Targeted cancer therapy. Nature. 2004;432:294.

25. Chapman PB, Hauschild A, Robert C, Haanen JB, Ascierto P, Larkin J, et al. Improved survival with vemurafenib in melanoma with BRAF V600E mutation. The New England journal of medicine. 2011;364(26):2507-16.

26. Afghahi A, Sledge GWJ. Targeted Therapy for Cancer in the Genomic Era. The Cancer Journal. 2015;21(4):294-8.

27. Hellmann MD, Li BT, Chaft JE, Kris MG. Chemotherapy remains an essential element of personalized care for persons with lung cancers. Annals of oncology : official journal of the European Society for Medical Oncology. 2016;27(10):1829-35.

28. Olson JJ, Nayak L, Ormond DR, Wen PY, Kalkanis SN. The role of cytotoxic chemotherapy in the management of progressive glioblastoma. Journal of Neuro-Oncology. 2014;118(3):501-55.

29. Twelves C, Jove M, Gombos A, Awada A. Cytotoxic chemotherapy: Still the mainstay of clinical practice for all subtypes metastatic breast cancer. Critical Reviews in Oncology/Hematology. 2016;100:74-87.

30. Gustavsson B, Carlsson G, Machover D, Petrelli N, Roth A, Schmoll H-J, et al. A Review of the Evolution of Systemic Chemotherapy in the Management of Colorectal Cancer. Clinical Colorectal Cancer. 2015;14(1):1-10.

31. Savage P, Stebbing J, Bower M, Crook T. Why does cytotoxic chemotherapy cure only some cancers? Nature Clinical Practice Oncology. 2008;6:43.

32. Abaan OD, Polley EC, Davis SR, Zhu YJ, Bilke S, Walker RL, et al. The exomes of the NCI-60 panel: a genomic resource for cancer biology and systems pharmacology. Cancer research. 2013;73(14):4372-82.

33. 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.

34. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RCT, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature Medicine. 2002;8:68.

35. Genomics of Drug Sensitivity in Cancer [Available from: https://www.cancerrxgene.org/downloads.

36. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20(3):307-15.

37. Leek JT, Storey JD. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. PLOS Genetics. 2007;3(9):e161.

38. Shankavaram UT, Varma S, Kane D, Sunshine M, Chary KK, Reinhold WC, et al. CellMiner: a relational database and query tool for the NCI-60 cancer cell lines. BMC Genomics. 2009;10(1):277.

39. Reinhold WC, Sunshine M, Liu H, Varma S, Kohn KW, Morris J, et al. CellMiner: A Web-Based Suite of Genomic and Pharmacologic Tools to Explore Transcript and Drug Patterns in the NCI-60 Cell Line Set. Cancer Research. 2012;72(14):3499.

40. Sun J, Wei Q, Zhou Y, Wang J, Liu Q, Xu H. A systematic analysis of FDA-approved anticancer drugs. BMC Systems Biology. 2017;11(5):87.

41. The Cancer Genome Atlas N, Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, et al. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490:61.

42. Jang IS, Neto EC, Guinney J, Friend SH, Margolin AA. SYSTEMATIC ASSESSMENT OF ANALYTICAL METHODS FOR DRUG SENSITIVITY PREDICTION FROM CANCER CELL LINE DATA. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing2014. p. 63-74.

43. Pedregosa F, Ga, #235, Varoquaux l, Gramfort A, Michel V, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825-30.

44. Lazar C, Taminau J, Meganck S, Steenhoff D, Coletta A, Molter C, et al. A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2012;9(4):1106-19.

45. Hall M. Correlation-based feature selection for machine learning. New Zealand Waikato University; 1999.

46. 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 research. 2015;43(7):e47-e.

47. DING C, PENG H. MINIMUM REDUNDANCY FEATURE SELECTION FROM MICROARRAY GENE EXPRESSION DATA. Journal of Bioinformatics and Computational Biology. 2005;03(02):185-205.

48. Bock HH, editor Probabilistic Aspects in Cluster Analysis1989; Berlin, Heidelberg: Springer Berlin Heidelberg.

49. 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(4):e61318.

50. Ammad-ud-din M, Khan SA, Wennerberg K, Aittokallio T. Systematic identification of feature combinations for predicting drug response with Bayesian multi-view multi-task linear regression. Bioinformatics. 2017;33(14):i359-i68.

51. Ammad-ud-din M, Khan SA, Malani D, Murumägi A, Kallioniemi O, Aittokallio T, et al. Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization. Bioinformatics. 2016;32(17):i455-i63.

52. Salem H, Attiya G, El-Fishawy N. Classification of human cancer diseases by gene expression profiles. Applied Soft Computing. 2017;50:124-34.

53. Huang S, Cai N, Pacheco PP, Narandes S, Wang Y, Xu W. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. Cancer Genomics & Proteomics. 2018;15(1):41-51.

54. Haibe-Kains B, El-Hachem N, Birkbak NJ, Jin AC, Beck AH, Aerts HJWL, et al. Inconsistency in large pharmacogenomic studies. Nature. 2013;504(7480):389-93.

55. Ben-David U, Siranosian B, Ha G, Tang H, Oren Y, Hinohara K, et al. Genetic and transcriptional evolution alters cancer cell line drug response. Nature. 2018;560(7718):325-30.

56. Mannheimer J, Fowles JS, Shaumberg K, Duval DL, Prasad A, Gustafson DL. Abstract 1522: Predicting drug sensitivity based on gene array data for cytotoxic chemotherapeutic agents. Cancer Research. 2016;76(14 Supplement):1522.

CHAPTER 4: PREDICTING CHEMOSENSITIVITY USING DRUG PERTURBED GENE DYNAMICS

Introduction

A major focus of cancer treatment is the utilization of phenotypic characteristics that can inform data-driven treatment protocols to target specific vulnerabilities of a patient's cancer (1). There has been a substantial amount work to characterize the genomic and mutational landscape of cancer that have resulted in successful interventions in cancers, harboring specific mutations or genomic signatures (2-4). Nonetheless, for the majority of cancers specific genomic prognostic indicators informing treatment have yet to be discovered with current estimates of only ~15% of cancer patients being eligible for genome-informed treatment (5). Cancer is a complex disease that arises from both numerous and diverse biological interactions. Developments in high-throughput drug screening and genomic profiling have laid a solid foundation characterizing the pharmacogenomic landscape of the disease (6, 7). Even so, developing specific experimental protocols *in vitro* or *in vivo* that probe the entirety of this landscape is an infeasible if not impossible task. A major goal of computational and systems biology has been to integrate and leverage the information inherent in available data to foster new insight about complex biological systems (8). Specifically in cancer, statistical, mathematical and computational approaches, are starting to be utilized to uncover complex drug-disease relationships (9) (10). However, this is an inherently complex task. The genome is innately a high dimensional space, has built in redundancy between genes, and gives rise to several complex multivariate interactions

many of which we have little or no knowledge about. Thus, identifying these relationships requires developing tools and approaches for deconvolution and screening of this complex data pool.

Recently, a clinical trial in human bladder cancer was concluded using computational methods to leverage cell line data to predict prognosis for neoadjuvant chemotherapy (11). The origin of these studies has been driven by similar *in silico* models predicting drug response for *in vitro* cell lines (12, 13). One of the most comprehensive evaluations of these methods was conducted as a team-based competition where 44 teams using a variety of different computational approaches competed to predict drug response for 28 therapeutic agents in a panel of 53 breast cancer cell lines (14). The study concluded that computational approaches could predict drug response using omics data particularly with a high regard to genomics data.

Pan-cancer models have also been shown to predict the response of cytotoxic chemotherapies in large cell line databases such as Genomics for Drug Sensitivity (15) and the National Cancer Institute 60 cell database (NCI60) (16). However, one of the central findings in (16) was that the predictive capabilities of these models was largely driven by associations between certain drugs that had stratified drug response based on histotype; drugs for which drug response was mostly independent of histotype tended to perform poorly in the models compared to those that did. The dimensionality inherent in omics data makes the model more susceptible to weaker broader signals making smaller, yet more informative signals, hard to isolate. The processes in a cell are inherently dynamic and adaptive. With respect to cancer drugs, the purpose of a drug is to interfere with the dynamic and adaptive mechanisms that are responsible for disease

pathology. Therefore, it is reasonable to assume that changes in gene expression after drug perturbation would, in part, be reflective of the underlying mechanisms responsible for drug response. The idea that changes in gene expression are linked to drug mechanism has been reflected in the connectivity map (17, 18) which has shown to give relevant pharmacogenomic insights (19, 20). Additionally, there have been studies that have leveraged specific gene dynamics in the p53 pathway to predict drug response with promising results (21). These results suggest that perturbation-based models have the potential to reflect drug response. Furthermore features identified in perturbation-based models may be predictive even when applied to basal gene expression.

The NCI Transcriptional Pharmacodynamic Workbench (22) is a web based tool that allow users to explore the relationship between changes in gene expression, drug response, and drug exposure for 15 different drugs in the NCI60 panel of cell lines. However, this tool only allows a univariate analysis by correlation of gene expression and drug response. To the best of our knowledge, no one has applied multivariate predictive models using this data. We use support vector regression with a radial basis function (SVR-RBF) to build predictive models of drug response for the data available in the NCI Transcription Pharmacodynamic Workbench. Specifically, there is an emphasis on the predictive capabilities of gene expression under different drug treatments. Additionally, the predictive relationships between these datasets are explored using correlation based feature selection (23). Additionally, network-based analysis is utilized to explore the relationships that exist between selected genes for both basal gene expression and drug induced changes in gene expression.

Methods:

Data Acquisition and Pre-Processing

The Affymetrix U133A 2.0 raw expression data from Monks et.al were downloaded from the gene expression omnibus (https://www.ncbi.nih.gov/geo) series number GSE116436 (22). Each drug had CEL files for gene expression for untreated cell lines (basal, 0 nM), cell lines drugged at a low dose of drug (C_{low}), and cell lines treated with a high dose of drug (C_{high}) at 2,6,and 24 hours. Frozen robust multi-array analysis (fRMA)(24) was performed using the fRMA bioconductor (version 3.8.0) package in R (version 3.5.1) for all CEL files corresponding to each individual drug. At 2,6, and 24 hours, the data were split into gene expression matrices for basal, C_{low}, and C_{high}. All gene expression matrices were scaled to a mean of 0 and unit standard deviation. Perturbation gene expression was obtained by subtracting the basal data from the drug treated data (C_{high},C_{low}) yielding matrices of gene differences at the high and low concentration (ΔC_{high}, ΔC_{low}). Throughout the text, (C_{high}, C_{low}) refer to perturbation gene expression and (ΔC_{high}, ΔC_{low}) to perturbed gene expression deltas or simply expression deltas. NCI60 drug response data were obtained from the *CellMiner version 2.2* https://discover.nci.nih.gov/cellminer (25). The natural log of the GI50 was averaged for all measurements attributed to the same single cell line giving a single average LN GI50 for each cell line per drug.

Training and validation sets were generated randomly using 3-fold nested cross validation. In order to generate a robust measure of performance across all gene expression datasets this process was repeated 2 more times giving a total of 9 random training and validation pairs with each cell line being represented at exactly 3 times

during validation. This, amounted to 9 replicates for each gene expression matrix (basal/0 nM, C_{low}, C_{high}, ΔC_{low}, ΔC_{high}) (Figure 4.1) at 2, 6, and 24 hours.

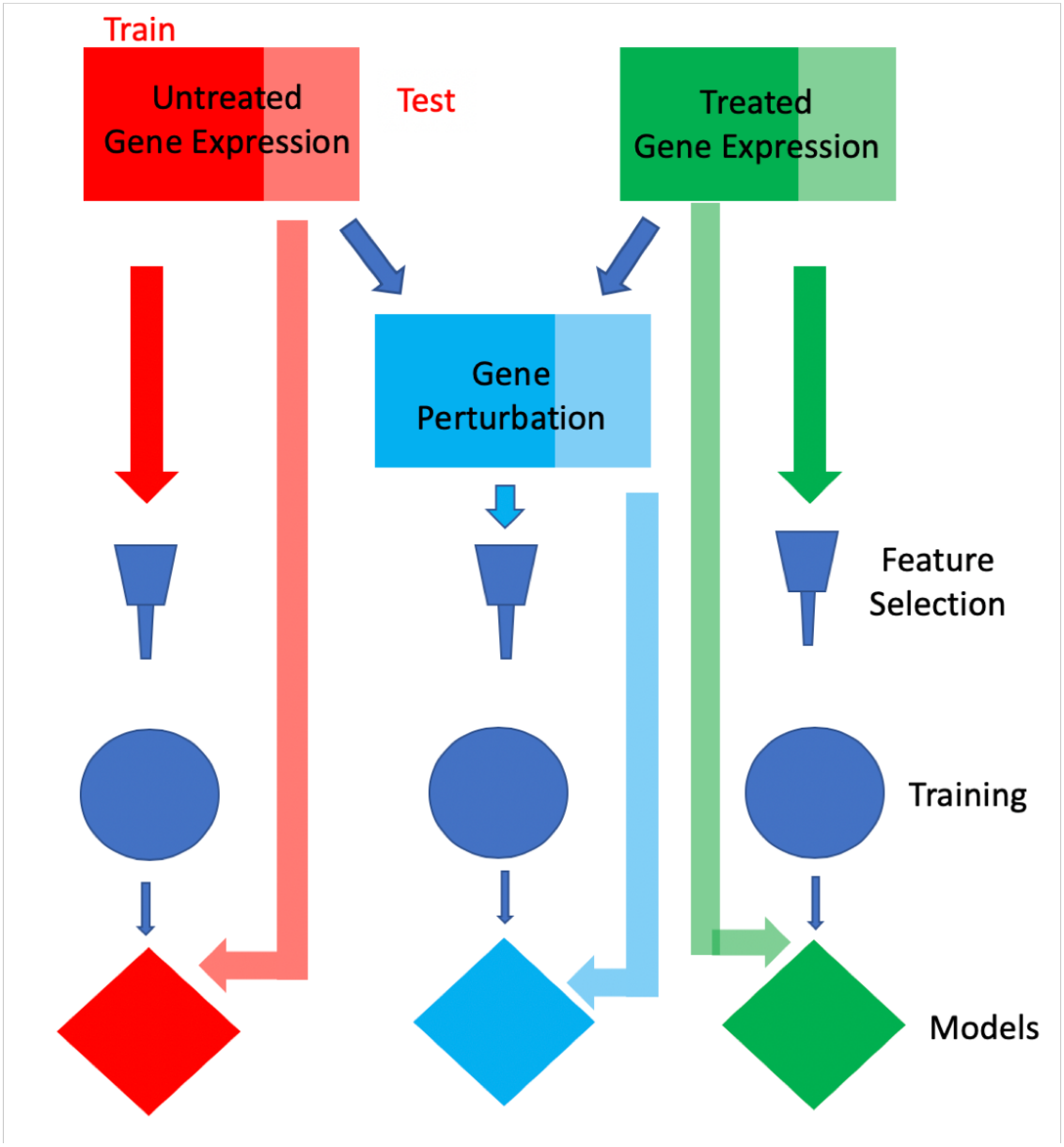


Figure 4.1: Model building outline.

Modeling:

All models were trained using ε-insensitive Support Vector Regression (SVR) using a radial basis kernel function (RBF) from Scikit-learn (26) version 20.3 in python version 3.7.3.

Parameters were optimized using a 10-fold random shuffle cross validation scheme on subsets of the training set. Differentially expressed genes (DEGs) were chosen using correlation based feature selection (CBF) (7) using spearman correlation in scipy version 1.2.1. For the DEGs models, the DEGs chosen are the identity of the genes, however, the gene expression data used in the model will be from a possibly different dataset. For example, if the DEGs are from the C_{high} data set, but the model is being evaluated on ΔC_{low} the gene expression used in the model is from ΔC_{low} but only those genes that were selected from the C_{high} are used. Graphics are generated using Prism Version 8 and to calculate significance, the paired Wilcoxon t test is used in comparing different models.

Topological Network Analysis:

A graph, a mathematical formalism that represents networks, is defined as an ordered set of nodes, V, and the edges, E, that connect nodes $G(V, E)$. The frequency and orientation with which two nodes are connected describe the networks topology. A topological measurement, cliques, are a subset of nodes such that every node in the subset shares an edge with every other node in the subset (Figure 4.5 A.). Additionally, a graphs topology can be described by a connectivity coefficient which is a measurement of the degree to which a node is connected to other nodes (Figure 4.5 A). We use these graph theoretic principles to estimate networks of genes. Graphs, gene

networks, are constructed as follows: using the smallest subset of DEGs from all nine training/validation sets an adjacency matrix was obtained by first constructing a correlation matrix using spearman r, then all matrix values that meet the Bonferroni corrected cutoff p value (α<0.05) are set equal to 1 and all other values are set to 0. The software NetworkX 2.4 (27) was used to generate a undirected graph and calculate cliques and average clustering coefficients.

Results:

Perturbed Gene Expression at 24 hours is a Good Predictor of Drug Response

It can be hypothesized that for each drug there is some timescale when drug induced perturbation is most predictive of drug response. Using the basal and perturbed gene expression at 2, 6, and 24 hours, 135 models (9 models for all 15 drugs), per timepoint, were constructed for each gene expression profile (basal, perturbed, expression deltas) for the 3 different treatment conditions. The best performing models, by average spearman correlation, consisted of gene expression profiles from C_{high} gene expression ($\bar{r} = 0.495$) after 24 hours of treatment while, performance was lowest for ΔC_{high} ($\bar{r} = 0.025$) 2 hours post treatment. Performance was dominated by gene expression profiles 24 hours post treatment (ΔC_{high}, ΔC_{low}, C_{high}, C_{low}) (Figure 4.2). The highest achieved average spearman correlation was achieved for the drug dasatinib ($\bar{r} = 0.848$) using ΔC_{high} gene expression 24 hours post treatment compared to lowest for azacytidine ($\bar{r} = 0.144$) using ΔC_{high} gene expression 6 hours post treatment. The average correlation of the top performing models for each drug drugs was $\bar{r}$= 0.6074 (SD: 0.16) (Table 4.1.).

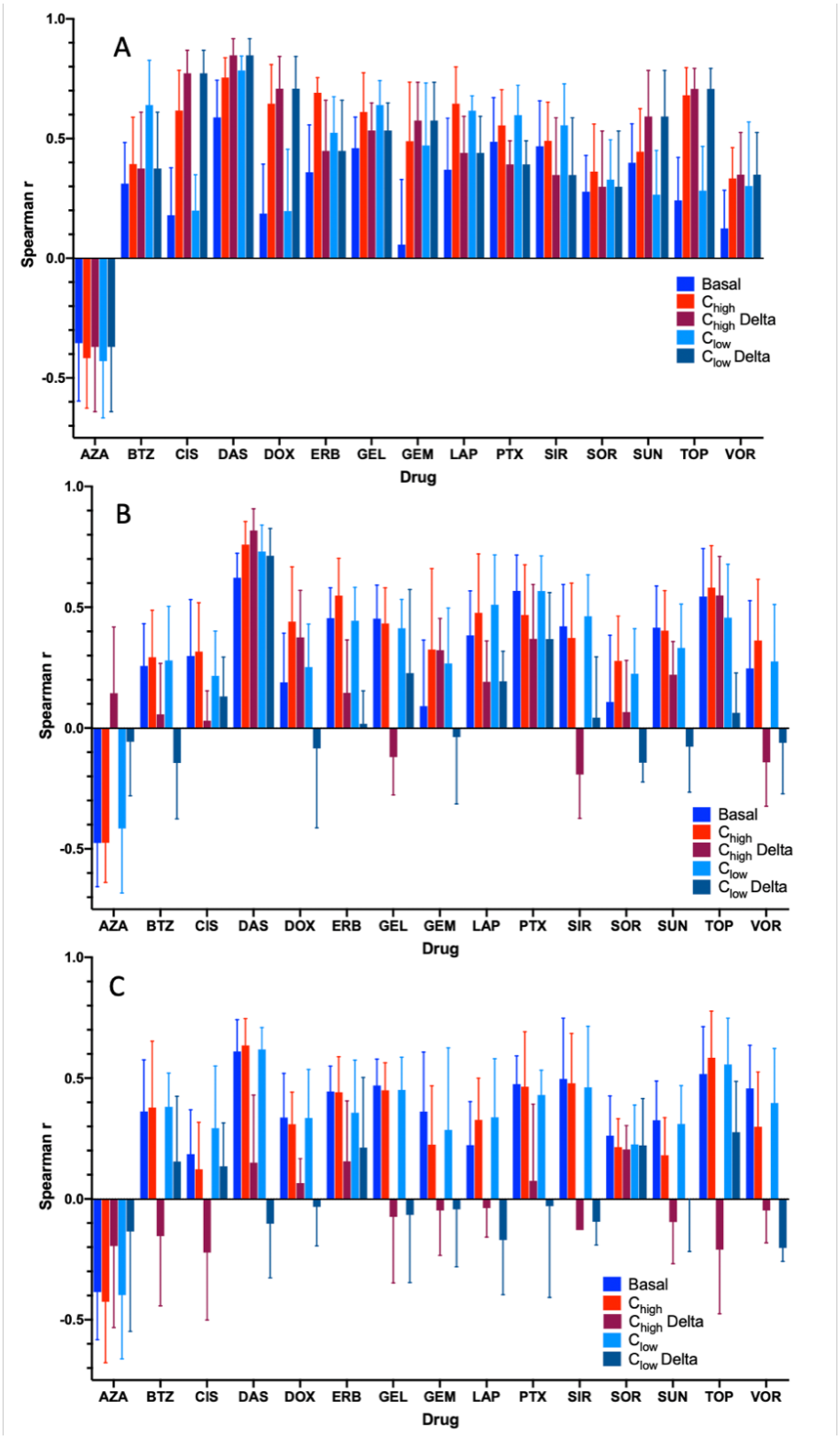


Figure 4.2: Spearman Correlations using different gene expression data for A. 24 hours, B. 6 hours, and C. 2 hours.

Table 4.1. The dataset that gave the best correlation for each drug

Drug	Abbreviation	Dataset	Correlations
Azacytidine	AZA	C_{high} Delta 6hr	0.144
Bortezomib	BTZ	C_{low} 24hr	0.603
Cisplatin	CIS	C_{high} Delta 24hr	0.773
Dasatinib	DAS	C_{high} Delta 24hr	0.848
Doxorubicin	DOX	C_{high} Delta 24hr	0.709
Erlotinib	ERB	C_{high} 24hr	0.692
Geldanamycin	GEL	C_{low} 24hr	0.641
Gemcitabine	GEM	C_{high} Delta 24hr	0.576
Lapatinib	LAP	C_{high} 24hr	0.646
Paclitaxel	PTX	C_{low} 24hr	0.6
Sirolimus	SIR	C_{low} 24hr	0.556
Sorafenib	SOR	C_{low} Delta 24hr	0.566
Sunitinib	SUN	C_{high} Delta 24hr	0.593
Topotecan	TOP	C_{high} Delta 24hr	0.707
Vorinostat	VOR	0nM 2hr	0.457

With respect to each drug and gene expression profile , six drugs were most predictable using ΔC_{high} gene expression at 24 hours post treatment, four drugs using C_{low} gene expression at 24 hours post treatment, two drugs at using C_{high} gene expression at 24 hours post treatment, and a single drug using ΔC_{low} gene expression at 24 hours post treatment (Table 1,Figure 4.3 B). Azacytidine and vorinostat were the

only two drugs that did not have the best performance with 24-hour post treatment gene expression. Azacytidine was most predictive using ΔC_{high} 6 hours post treatment and vorinostat was best predicted by basal gene expression. The interplay between dosage and timing was explored, since some drugs may display more predictive signatures at a low dose and others at a high dose. We found that across all models, gene expression profiles drugged at a high concentration ($C_{high}/\Delta C_{high}$) performed better than similar gene expression drugged with a lower concentration of drug ($C_{low}/\Delta C_{low}$). Both C_{high}/C_{low} (Δr=22%) and $\Delta C_{high}/\Delta C_{low}$ (Δr=0.172%) had large differences in average correlation, however, only C_{high}/C_{low} was significantly different (p_{wc}=0.0005) by Wilcoxon paired t test. Specifically, at 24 hours post treatment C_{high} resulted in models 1.2% better then ΔC_{high} gene expression; however, the difference was not significant (p_{wc}=0.427). Conversely, at the lower concentration ΔC_{low} outperformed C_{low} by 2.88% but not significantly (p_{wc}=0.65). At the 24 hour time point models using basal data gene expression performed significantly lower with respect to C_{high} (42.7%, p_{wc} < 1e-4) and ΔC_{high} (42%, p_{wc}<1e-4). The results were similar at the lower concentration for C_{low} (30% , p_{wc}<1e-4) and ΔC_{low} (32%, p_{wc}<1e-4). With respect to drug exposure (dose x time), C_{low} at 24 hours performs 4.4% better than C_{high} expression at 6 hours despite that drug exposure at C_{high} is greater; however this difference is not significant (p_{wc}=0.1065) (Figure 3.4 A). Additionally, note that C_{high} gene expression at 6 hours post treatment performs better than basal data (18%, p_{wc}=0.0001). C_{low} gene expression at 6 and 2 hours and C_{high} gene expression at 2 hours post treatment are comparable to models using basal gene expression ranging for 3.4% to 9.8% with no significant difference (p_{wc}>0.25).

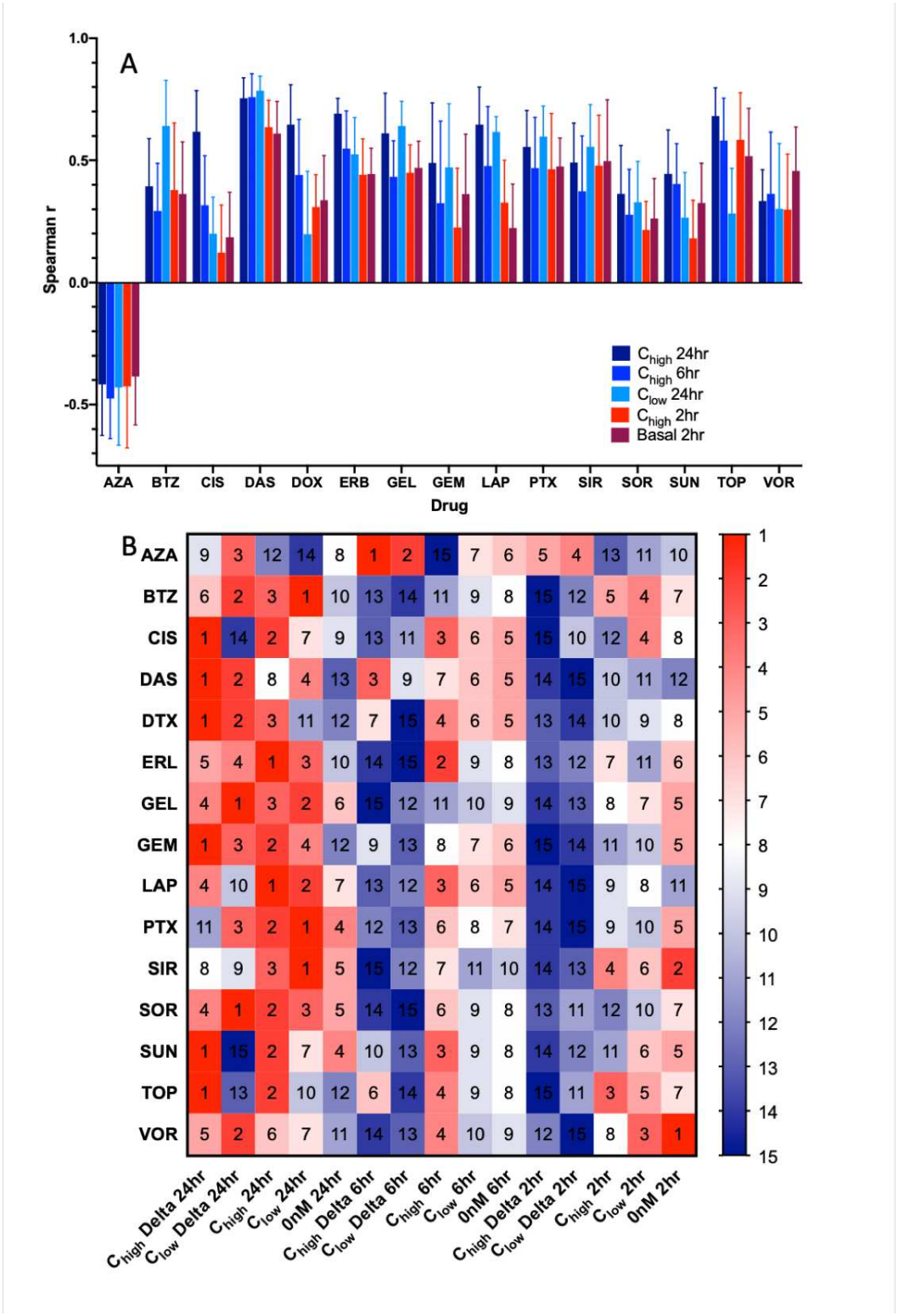


Figure 4.3. Average spearman correlations based on drug exposure. With highest exposure starting on the left working to no drug exposure on the right. Correlations based on exposure with high exposure starting at the left of page. B. Rank of each drug by spearman correlations over different all gene expression profiles.

A smaller set of differentially expressed genes are sufficient to capture drug response

Since each drug has specific modes of action that endow it with cytotoxicity, a smaller set of features, or gene expression signatures, may be as predictive of drug response as the entire ensemble of gene expression. To determine whether a predictive drug response gene expression signature could be found, DEGs were selected for each 24-hour gene expression profile and models based on these DEGs were constructed within each gene expression profile (Figure 4.4 F). DEG gene expression profiles resulted in lower average spearman correlation compared to using all genes (NOFS), with the exception of the ΔC_{high} data (azacytidine was left out in the analysis as it varied greatly between different testing sets within each individual gene expression profile). The increase in performance while using ΔC_{high} DEGs on ΔC_{high} data compared to the entire ΔC_{high} profile (NOFS) was modest (DEG $\bar{r} = 0.5415$, NOFS $\bar{r} = 0.5294$) and not significant (p_{wc}=0.3437). Additionally, when comparing ΔC_{low} gene expression the performance was only slightly less using DEGs ($\bar{r} = 0.4092$) than NOFS ($\bar{r} = 0.4443$) with no significant difference (p_{wc}=0.2516). However, comparing C_{high} DEG models to C_{low} DEG models, the performance of the DEG models was significantly less (C_{high} p<1e-4, C_{low} p=0.0181) with differences in performance of 10% (C_{low}) to 16.2% (C_{high}). The difference between basal DEG models and basal NOFS models data was insignificant (p_{wc}=0.0533) where the DEG model performed about 10% worse ($\bar{r} =$ 0.299/0.331). Comparisons between models using DEGs and NOFS model on a drug by drug basis can be seen in Figure 4.4 F.

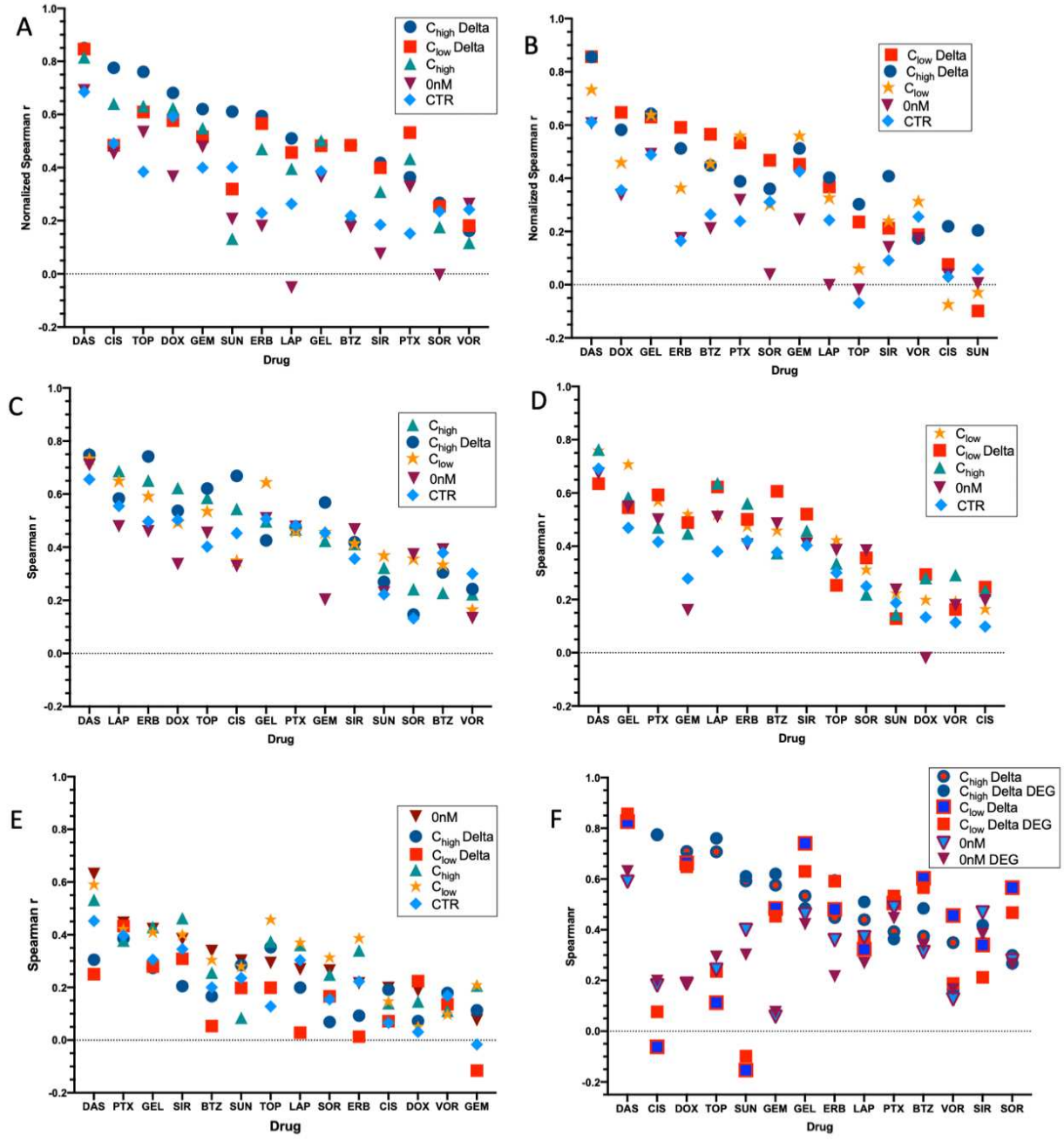


Figure 4.4. The influence on performance using DEGs selected from one gene profile and for models utilizing A. ΔC_{high} , B. ΔC_{low} , C_{high} , D. C_{low} , E. 0 nM. F. Comparison between average spearman correlation where DEGs are selected from the perspective data profiles used in the models.

DEGs selected from different gene expression profiles are not universally predictive when applied across gene expression profiles

The advantage of using perturbation data for feature selection is straight forward; if a gene's expression changes with exposure to drug there is a higher probability that the gene plays a role in the cell line's response to that drug. Thus it is not unreasonable to assume that a gene that has a dynamic response to drug exposure is a good feature to use when modeling. However, often the gene expression data for *in vitro* cell lines and tumor samples after drug exposure is not available. We asked whether it might be possible to use available drug perturbed data from another dataset to select features with a dynamic response, and apply those features to predict response in another dataset. However, it unclear whether a signature derived from drugged gene expression data also reflects drug response under unperturbed conditions. In order to explore this question, we used correlation-based feature selection to select features using one gene expression profile and then applied those selected genes in models utilizing a different gene expression profile (refer to methods).

Differentially expressed genes selected in basal gene expression and applied to ΔC_{high} gene expression resulted in a 46.5% drop in performance ($\bar{r} = 0.542$ to 0.29). For comparison, using 100 randomly selected genes resulted in a smaller drop in performance by only 36%. The difference in performance was minimal (12%) when ΔC_{low} DEGs were applied to ΔC_{high} expression data ($\bar{r} = 0.542$ to 0.48) and was substantially lower when C_{high} DEGs were applied to the same data, resulting in a drop of 22% ($\bar{r} = 0.542$ to 0.429)(Figure 4.4 A). Likewise, with respect to ΔC_{low} data using basal DEGs resulted in a 52% decrease in overall performance ($\bar{r} = 0.41$ to 0.197);

however, contrary to the results for ΔC_{high}, when DEGs from ΔC_{high} were applied to ΔC_{low} gene expression performance increased by 5% ($\bar{r}$ = 0.41 to 0.43) but was not significant (p_{wc}=0.96). The application of C_{low} DEGs to ΔC_{low} gene expression resulted in 16.7% drop ($\bar{r}$ = 0.41 to 0.351, p=0.0028) (Figure 4.4 B).

We also tested the predictive capability of DEGs selected from basal and expression deltas on perturbation gene expression. Similar to what we found for expression deltas, basal DEGs resulted in the greatest drop in performance for both C_{high} 16.7% ($\bar{r}$ =0.476 to 0.396) and C_{low} 15.1% ($\bar{r}$ = 0.4253 to 0.425) (Figure 4.4 C&D). Performance of ΔC_{high} DEGs on C_{high} gene expression resulted in a slight increase in performance by roughly 1.7% ($\bar{r}$ =0.476 to 0.484), while ΔC_{low} DEGs had only a negligible effect when applied to C_{low} gene expression ($\bar{r}$ = 0.4253 to 0.4250) (Figure 4.4 C&D). The application of C_{low} DEGs to C_{high} gene expression resulted in a decrease of roughly 1.7% ($\bar{r}$ = 0.476 to 0.468), a slightly larger, but still models, drop in performance resulted from the use C_{high} DEGs to C_{low} gene expression ($\bar{r}$ = 0.4253 to 0.414)

Finally, DEGs were selected from perturbed gene expression and expression deltas and these DEGs were applied to basal gene expression (Figure 4.4 E). When ΔC_{high} DEGs were applied to basal genes expression, performance substantially decreased by 30% ($\bar{r}$ =0.299 to 0.208) and similarly, using ΔC_{low} DEGs on basal gene expression the performance decreased by 44% ($\bar{r}$ = 0.299 to 0.166). Finally, a random selection of gene expression from 100 random genes resulted in a decreased performance by only 29% ($\bar{r}$ = 0.299 to 0.213). Furthermore, models using basal gene expression increased performance by 6% ($\bar{r}$ = 0.299 to 0.317) when using DEGs from

C_{low} and ΔC_{high} DEGS applied to basal gene expression decreased performance by by 2.7%. These results indicate that changes in gene expression that predict drug response are not predictive features in basal gene expression.

DEGs and Network Topology

One of the fundamental concepts in biology is that cellular systems are an assembly of dynamic interactions forming a network of interacting components, that provide the framework for all functions of the cell. While generally it is well understood that networks involve some kind of connectivity, network models allow for a more rigorous mathematical approach to understand and analyze this general notion of connectivity. Particularly, there has been an interest in applying concepts behind network topology to understand the relationship between genes, disease states, and treatments (21, 28). To explore the relationships between genes under both a basal and perturbed states, DEG networks were constructed using statistically significant correlations between genes. As outlined in methods, this was accomplished by calculating a correlation matrix for both the basal and expression delta gene expression profiles. Then Boolean graphs were constructed by placing edges between genes that had a spearman correlation p-value below a Bonferroni corrected significance level. The topology of the given network was then quantified based on cliques, a subset of nodes which share an edge with every other node in the subset, and the clustering coefficient which measures the connectivity of a subset of nodes all sharing an edge with a single node (Figure 4.5 A).

There was a clear distinction between topological properties of networks formed with expression deltas DEGs compared with basal DEGs. On average networks

constructed from expression DEGs formed 389% more cliques then networks formed with basal DEGs. The number of cliques which exceeded a size of 2 was also much greater using the expression deltas DEGs averaging around 60% similarly compared to 21 % for basal DEGs. Likewise, the average clique length was 314% greater using expression deltas DEGs compared to Basal DEGs (Figure 4.5 B). Additionally, clique participation was much greater in expression delta gene networks, with each node participating in 1.1% of all cliques compared to only 0.3% for basal DEG networks. Lastly, expression deltas DEG networks had an average clustering coefficient that was 2.15x greater than that of the basal DEG networks (Figure 4.5 C). Based on network topological features expression deltas DEGs showed a much greater level of interaction compared to DEGs derived from basal gene expression.

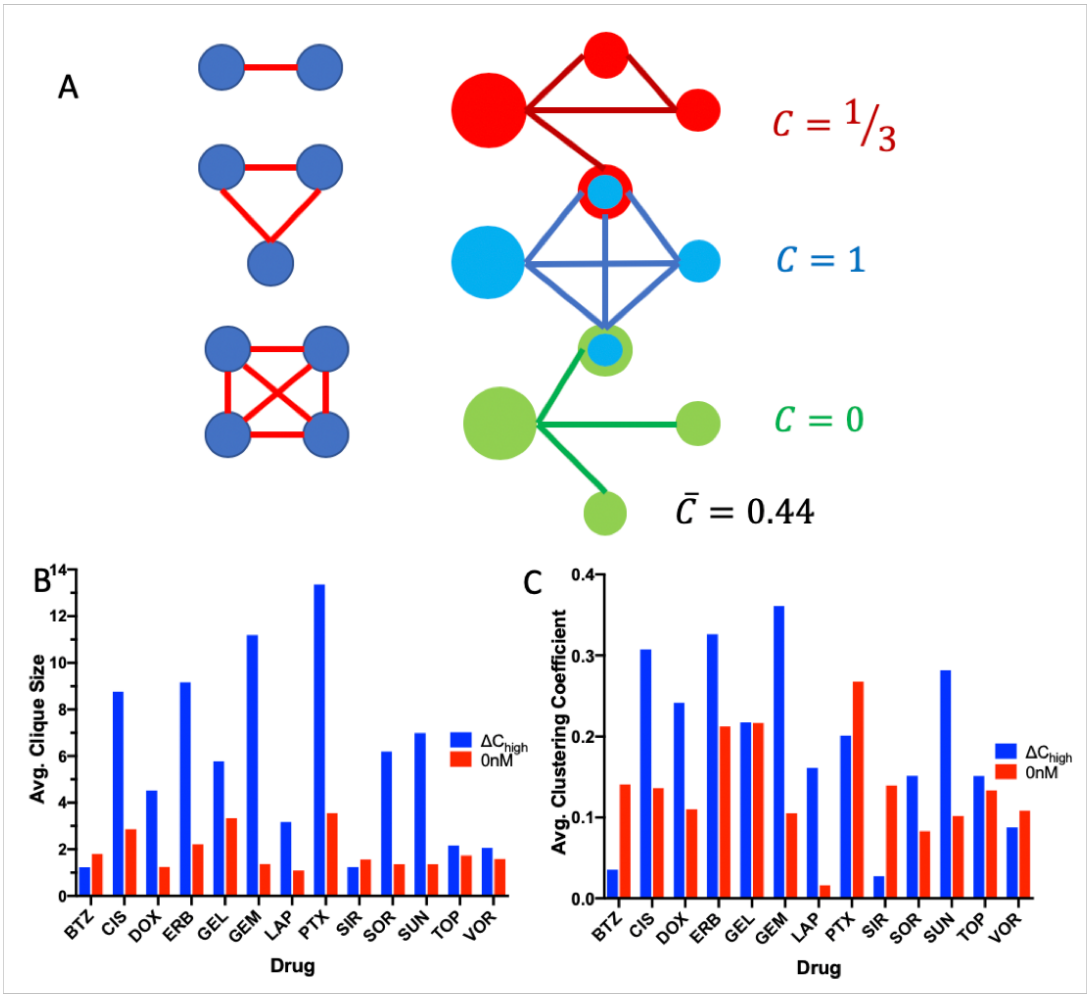


Figure 4.5. A. Illustration of a clique (Right) of size 2 (top), 3 (Middle), 4 (bottom) and Clustering Coefficient for 3 different subgraphs along with the average over all three subgraphs. B. Average clique size, and C. Average clustering coefficient for different drugs for ΔC_{high} and 0n M at 24 hours.

Clique participation is a signature of cancer and drug response association of genes.

In order to explore how clique participation might be associated with cancer association we looked at the 15 genes that participated in the most cliques in both the expression deltas network and the basal gene network. Analysis of expression deltas-based data yielded several genes that had been cited in the literature to have some known association with cancer. This in included tumor suppressor genes, Breast

Cancer Metastasis Suppressor 1 (BMRS1) in bortezomib (29-32) and RNA Binding Protein 5(RBM5) in paclitaxel (33-35). Additionally, Genes that promote proliferation such as MOB Family Member 4 (MOB4, doxorubicin) (36), PBX Homeobox Interacting Protein (PBXIB1, geldanamycin) (37, 38) and Heterogenous Nuclear Ribonucleoprotein A/B (HRNPAB, Sorafenib) (39). Furthermore, apoptosis related genes including Nuclear Receptor Coactivator (NCOA1,cisplatin)(40), TIMELESS interacting protein (TIPIN, lapatinib)(41), and G1 to S Phase Transition 1 (GSPT1, sunitinib)(42). Additionally, cell cycle regulators, Cell Division Cycle 25A (CDC25A, geldanamycin)(43) and Topoisomerase DNA binding Protein (TOPBP1,pacliatxel)(44). Furthermore, some genes linked to drug resistance arose as well such as survivin (BIRC5, erlotinib) which has been found to correlate with paclitaxel resistance (45). These represent a subset of genes that we identified. For all drugs, genes with cancer associated references could be found or genes involved in cell cycle regulation, apoptosis, or translation. A list of these additional genes can be found in Appendix B.

Likewise several genes with maximum clique participation in the basal gene networks had associations with cancer. The gene that participated in the most cliques for bortezomib was ABCE1, a known mediator of drug resistance (46-48). Likewise, for lapatinib ATB binding cassette family F member 1 (ABCF1) also associated with chemoresistance (49). One of the genes picked up for sorafenib was EGFR and among the genes for topotecan was a tumor suppressor genes ST14 (matriptase) (50-52). Additionally, R-Ras 2, a oncogene gene known to be associated with tumorigenesis and metastasis was present for erlotinib (53, 54). Additional genes of interest from basal gene expression networks can be found in Appendix B.

Discussion:

Small molecule gene perturbation studies have become a new focus to understand the relationships between diseases and drugs (17, 18, 22). One of the central roles of this work was to understand how gene dynamics could inform drug response and what roles drug exposure may play. The results suggest, given a limited subset of drugs and cell lines, that the differences in gene expression at 24 hours between untreated and treated cells with a relatively higher high dose of drug is the best predictor of drug response compared to basal gene expression or perturbed responses at earlier time points and lower drug doses. Additionally, the data suggest that elapsed time might play may be a bigger factor than exposure, as models that utilized perturbation gene profiles treated with a low dose at the 24-hour post treatment often outperformed models using high dose gene expression but at an earlier time point. The similarity between the predictive ability of non-drugged gene expression and drugged gene expression at 2 and 6 hours suggest that changes in gene expression are rather minimal at these early time points. One of the questions that might be of interest would be to answer at exactly what time do changes in gene expression become predictive and whether time points greater than 24 hours could possibly be more predictive? Additionally, if the changes in gene expression could be measured with enough temporal resolution it might enable analysis of gene trajectories to see if they are indicative of drug response.

The role of feature selection in omics-based models is somewhat controversial. In certain other data-driven models, feature selection can be critical in eliminating noise, resulting in a better performing model. In genomics two of the most used methods of

feature selection is to use user-defined genes, for example, exploiting all genes known to be in a specific pathway or leveraging statistical inference to select features that can most likely explain the variability in the phenomena, such is the case with CBF or a pairwise t-test like LIMMA (55). With respect to the former, this method inherently gives biological context; however, the problem of predicting drug response is that many of the mechanisms of drug response are not known (56), thus making it difficult to select genes based on prior knowledge. The second approach is very susceptible to noise and it is not clear if there is any advantage over using all features (14, 16). However, the use of all features lacks the specificity to be useful for hypothesis generation. The inferential approach is somewhat of a go between, it does not exclude genes based on any prior bias, but preferences features only by the statistical capability to explain variation with respect to some biological observation such as drug response. The hope is that many of these features have a shared underlying biological context which results in the observed statistics. Drug response has both static and dynamic component; multi drug resistance can result from the overexpression of efflux transporters (56) and down regulation of deoxycytidine kinase is seen in gemcitabine resistance (57). A central question that underlies the DEG experiments is whether statistical inference can capture an underlying relationship between static and dynamic gene expression when it comes to drug response. Based on our results, the drug induced changes in gene expression that best predicted drug response did not have any predictive power in basal gene expression and the same was true of basal gene expression with respect to gene changes. However, this is only in respect to univariate feature selection, it is perfectly reasonable to believe that features exist in a higher dimensional space which might

pertain to an underlying biological phenomena; however, methods to find such features and map them to specific genes are yet to be developed.

The complexities of cancer therapy are immense and thus, especially from a pharmacological view, it is necessary to understand what properties make a cancer susceptible to a certain drug or what mechanisms are responsible for resistance. A recent publication showed that cell proliferation could be maintained without the specific proteins targeted by many therapies; additionally, they also demonstrated that many of these drugs achieved cytotoxicity without the inclusion of the druggable target (58). Targeted therapies are notorious for an initial response followed quickly by developed resistance (59, 60). All in all, gene expression of drug targets in both perturbed and unperturbed data are poor predictors of drug response. However, through network analysis we found that genes with high clique participation are associated with cancer in both changes in gene expression and basal gene expression. Furthermore, the networks constructed from changes in gene proved to be significantly more connected than networks constructed from basal gene expression. This might explain why perturbation is a better predictor of drug response, the redundancies that result from coordinated changes lead to a stronger signal to noise ratio. Alternatively, this might suggest that drug response is a more likely a function of several interacting genes and several different mechanisms and this is reflected better under dynamic changes. This is certainly consistent with the observation that cancer has multiple mechanisms of drug resistance (61, 62).

Conclusion:

There are several roles for computational models in oncology including, but not limited to, patient prognostics and treatment, new treatment development, informing clinical trials, and as a method of hypothesis generation especially when it comes interaction between cellular processes and drug mechanisms. Genetic perturbations would be difficult to leverage in a clinical setting: a prognostic model to aid patient treatment needs to be quick and cost effective. Acquiring gene perturbations of a patient's tumor for multiple drugs would be both difficult and time consuming. *In vitro* drug screens are often the first step in determining the potential of a possible drug candidate and also serve as a platform for hypothesis generation. However, as our data indicate, obtaining predictive models with basal gene expression is difficult and furthermore might not be the best data to determine drug mechanism. However, gene perturbations prove to be better at capturing drug response and they exhibit a high level of connectivity between genomic features. Therefore, modeling *in vitro* gene expression changes could be instrumental to better understand the dynamic mechanisms behind drug response. A better understanding of the underlying gene dynamics induced by a drug would most likely promote insight into the underlying genetic mechanisms of drug response which are essential for developing new treatments and building robust prognostic models at the clinical level. In order to fully take advantage of this strategy it would be essential to generate more perturbation data similar in scope to the genomic and drug profiling that is associated with large cell line panels such as the GDSC, CCLE, and expanded among more drugs in the NCI60. This is a task that funding agencies and foundations that help build these databases should take up without delay

REFERENCES

1. Collins FS, Varmus H. A New Initiative on Precision Medicine. New England Journal of Medicine. 2015;372(9):793-5.

2. Cutter GR, Liu Y. Personalized medicine: The return of the house call? Neurology Clinical practice. 2012;2(4):343-51.

3. Toi M, Iwata H, Yamanaka T, Masuda N, Ohno S, Nakamura S, et al. Clinical significance of the 21-gene signature (Oncotype DX) in hormone receptor-positive early stage primary breast cancer in the Japanese population. Cancer. 2010;116(13):3112-8.

4. Romond EH, Perez EA, Bryant J, Suman VJ, Geyer CE, Davidson NE, et al. Trastuzumab plus Adjuvant Chemotherapy for Operable HER2-Positive Breast Cancer. New England Journal of Medicine. 2005;353(16):1673-84.

5. Marquart J, Chen EY, Prasad V. Estimation of the Percentage of US Patients With Cancer Who Benefit From Genome-Driven Oncology. JAMA Oncol. 2018;4(8):1093-8.

6. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, et al. A gene expression database for the molecular pharmacology of cancer. Nature Genetics. 2000;24:236.

7. Schena M. Genome analysis with gene expression microarrays. BioEssays. 1996;18(5):427-31.

8. Werner HMJ, Mills GB, Ram PT. Cancer Systems Biology: a peek into the future of patient care? Nature Reviews Clinical Oncology. 2014;11:167.

9. Fessele KL. The Rise of Big Data in Oncology. Seminars in Oncology Nursing. 2018;34(2):168-76.

10. Altrock PM, Liu LL, Michor F. The mathematics of cancer: integrating quantitative models. Nature Reviews Cancer. 2015;15(12):730-45.

11. Flaig TW, Tangen CM, Daneshmand S, Alva AS, Lerner SP, Lucia MS, et al. SWOG S1314: A randomized phase II study of co-expression extrapolation (COXEN) with neoadjuvant chemotherapy for localized, muscle-invasive bladder cancer. Journal of Clinical Oncology. 2019;37(15_suppl):4506-.

12. 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(7391):570-5.

13. 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.

14. Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nature biotechnology. 2014;32(12):1202-12.

15. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Research. 2013;41(D1):D955-D61.

16. Mannheimer JD, Duval DL, Prasad A, Gustafson DL. A systematic analysis of genomics-based modeling approaches for prediction of drug response to cytotoxic chemotherapies. BMC Medical Genomics. 2019;12(1):87.

17. 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. 2006;313(5795):1929.

18. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell. 2017;171(6):1437-52.e17.

19. Rho SB, Kim B-R, Kang S. A gene signature-based approach identifies thioridazine as an inhibitor of phosphatidylinositol-3′-kinase (PI3K)/AKT pathway in ovarian cancer cells. Gynecologic Oncology. 2011;120(1):121-7.

20. Sanda T, Li X, Gutierrez A, Ahn Y, Neuberg DS, O'Neil J, et al. Interconnecting molecular pathways in the pathogenesis and drug sensitivity of T-cell acute lymphoblastic leukemia. Blood. 2010;115(9):1735-45.

21. Choi M, Shi J, Zhu Y, Yang R, Cho K-H. Network dynamics-based cancer panel stratification for systemic prediction of anticancer drug response. Nature Communications. 2017;8(1):1940.

22. Monks A, Zhao Y, Hose C, Hamed H, Krushkal J, Fang J, et al. The NCI Transcriptional Pharmacodynamics Workbench: A Tool to Examine Dynamic Expression Profiling of Therapeutic Response in the NCI-60 Cell Line Panel. Cancer Research. 2018;78(24):6807.

23. Hall M. Correlation-based feature selection for machine learning. New Zealand Waikato University; 1999.

24. McCall MN, Bolstad BM, Irizarry RA. Frozen robust multiarray analysis (fRMA). Biostatistics. 2010;11(2):242-53.

25. Mannheimer J, Fowles JS, Shaumberg K, Duval DL, Prasad A, Gustafson DL. Abstract 1522: Predicting drug sensitivity based on gene array data for cytotoxic chemotherapeutic agents. Cancer Research. 2016;76(14 Supplement):1522.

26. Pedregosa F, Ga, #235, Varoquaux l, Gramfort A, Michel V, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825-30.

27. Aric A. Hagber DAS, Pieter J. Swart, editor Exploring netwrok structure, dynamics, and function using NetworkX. 7th Python Science Conference (SciPy2008); 2008 August 2008; Pasadena, CA.

28. Hopkins AL. Network pharmacology: the next paradigm in drug discovery. Nature Chemical Biology. 2008;4:682.

29. Smith PW, Liu Y, Siefert SA, Moskaluk CA, Petroni GR, Jones DR. Breast cancer metastasis suppressor 1 (BRMS1) suppresses metastasis and correlates with improved patient survival in non-small cell lung cancer. Cancer Lett. 2009;276(2):196-203.

30. Kim B, Nam HJ, Pyo KE, Jang MJ, Kim IS, Kim D, et al. Breast cancer metastasis suppressor 1 (BRMS1) is destabilized by the Cul3-SPOP E3 ubiquitin ligase complex. Biochem Biophys Res Commun. 2011;415(4):720-6.

31. Zhao XL, Wang P. [Expression of SATB1 and BRMS1 in ovarian serous adenocarcinoma and its relationship with clinieopathological features]. Sichuan Da Xue Xue Bao Yi Xue Ban. 2011;42(1):82-5, 105.

32. Rivera J, Megias D, Bravo J. Proteomics-based strategy to delineate the molecular mechanisms of the metastasis suppressor gene BRMS1. J Proteome Res. 2007;6(10):4006-18.

33. Rintala-Maki ND, Abrasonis V, Burd M, Sutherland LC. Genetic instability of RBM5/LUCA-15/H37 in MCF-7 breast carcinoma sublines may affect susceptibility to apoptosis. Cell Biochem Funct. 2004;22(5):307-13.

34. Kobayashi T, Ishida J, Musashi M, Ota S, Yoshida T, Shimizu Y, et al. p53 transactivation is involved in the antiproliferative activity of the putative tumor suppressor RBM5. Int J Cancer. 2011;128(2):304-18.

35. Oh JJ, Taschereau EO, Koegel AK, Ginther CL, Rotow JK, Isfahani KZ, et al. RBM5/H37 tumor suppressor, located at the lung cancer hot spot 3p21.3, alters expression of genes involved in metastasis. Lung Cancer. 2010;70(3):253-62.

36. Tang F, Zhang L, Xue G, Hynx D, Wang Y, Cron PD, et al. hMOB3 modulates MST1 apoptotic signaling and supports tumor growth in glioblastoma multiforme. Cancer Res. 2014;74(14):3779-89.

37. Liu L, Huang J, Wang K, Li L, Li Y, Yuan J, et al. Identification of hallmarks of lung adenocarcinoma prognosis using whole genome sequencing. Oncotarget. 2015;6(35):38016-28.

38. van Vuurden DG, Aronica E, Hulleman E, Wedekind LE, Biesmans D, Malekzadeh A, et al. Pre-B-cell leukemia homeobox interacting protein 1 is overexpressed in astrocytoma and promotes tumor cell growth and migration. Neuro Oncol. 2014;16(7):946-59.

39. Zhou ZJ, Dai Z, Zhou SL, Hu ZQ, Chen Q, Zhao YM, et al. HNRNPAB induces epithelial-mesenchymal transition and promotes metastasis of hepatocellular carcinoma by transcriptionally activating SNAIL. Cancer Res. 2014;74(10):2750-62.

40. Wang L, Yu Y, Chow DC, Yan F, Hsu CC, Stossi F, et al. Characterization of a Steroid Receptor Coactivator Small Molecule Stimulator that Overstimulates Cancer Cells and Leads to Cell Stress and Death. Cancer Cell. 2015;28(2):240-52.

41. Baldeyron C, Brisson A, Tesson B, Némati F, Koundrioukoff S, Saliba E, et al. TIPIN depletion leads to apoptosis in breast cancer cells. Mol Oncol. 2015;9(8):1580-98.

42. Lee JA, Park JE, Lee DH, Park SG, Myung PK, Park BC, et al. G1 to S phase transition protein 1 induces apoptosis signal-regulating kinase 1 activation by dissociating 14-3-3 from ASK1. Oncogene. 2008;27(9):1297-305.

43. Li N, Zhong X, Lin X, Guo J, Zou L, Tanyi JL, et al. Lin-28 homologue A (LIN28A) promotes cell cycle progression via regulation of cyclin-dependent kinase 2 (CDK2), cyclin D1 (CCND1), and cell division cycle 25 homolog A (CDC25A) expression in cancer. J Biol Chem. 2012;287(21):17386-97.

44. Yamane K, Chen J, Kinsella TJ. Both DNA topoisomerase II-binding protein 1 and BRCA1 regulate the G2-M cell cycle checkpoint. Cancer Res. 2003;63(12):3049-53.

45. Zaffaroni N, Pennati M, Colella G, Perego P, Supino R, Gatti L, et al. Expression of the anti-apoptotic gene survivin correlates with taxol resistance in human ovarian cancer. Cell Mol Life Sci. 2002;59(8):1406-12.

46. Kara G, Tuncer S, Türk M, Denkbaş EB. Downregulation of ABCE1 via siRNA affects the sensitivity of A549 cells against chemotherapeutic agents. Medical Oncology. 2015;32(4):103.

47. Wang L, Zhang M, Liu D-X. Knock-down of ABCE1 gene induces G1/S arrest in human oral cancer cells. Int J Clin Exp Pathol. 2014;7(9):5495-504.

48. Zheng D, Dai Y, Wang S, Xing X. MicroRNA-299-3p promotes the sensibility of lung cancer to doxorubicin through directly targeting ABCE1. Int J Clin Exp Pathol. 2015;8(9):10072-81.

49. Li X, Li X, Liao D, Wang X, Wu Z, Nie J, et al. Elevated microRNA-23a Expression Enhances the Chemoresistance of Colorectal Cancer Cells with Microsatellite Instability to 5-Fluorouracil by Directly Targeting ABCF1. Curr Protein Pept Sci. 2015;16(4):301-9.

50. Uhland K. Matriptase and its putative role in cancer. Cell Mol Life Sci. 2006;63(24):2968-78.

51. Benaud CM, Oberst M, Dickson RB, Lin CY. Deregulated activation of matriptase in breast cancer cells. Clin Exp Metastasis. 2002;19(7):639-49.

52. Warren M, Twohig M, Pier T, Eickhoff J, Lin CY, Jarrard D, et al. Protein expression of matriptase and its cognate inhibitor HAI-1 in human prostate cancer: a tissue microarray and automated quantitative analysis. Appl Immunohistochem Mol Morphol. 2009;17(1):23-30.

53. Larive RM, Moriggi G, Menacho-Márquez M, Cañamero M, de Álava E, Alarcón B, et al. Contribution of the R-Ras2 GTP-binding protein to primary breast tumorigenesis and late-stage metastatic disease. Nat Commun. 2014;5:3881.

54. Luo H, Hao X, Ge C, Zhao F, Zhu M, Chen T, et al. TC21 promotes cell motility and metastasis by regulating the expression of E-cadherin and N-cadherin in hepatocellular carcinoma. Int J Oncol. 2010;37(4):853-9.

55. 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 research. 2015;43(7):e47-e.

56. Weinberg RA. The Biology of Cancer 2nd ed. New York, NY: Garland Science 2014.

57. OHHASHI S, OHUCHIDA K, MIZUMOTO K, FUJITA H, EGAMI T, YU J, et al. Down-regulation of Deoxycytidine Kinase Enhances Acquired Resistance to Gemcitabine in Pancreatic Cancer. Anticancer Research. 2008;28(4B):2205-12.

58. Lin A, Giuliano CJ, Palladino A, John KM, Abramowicz C, Yuan ML, et al. Off-target toxicity is a common mechanism of action of cancer drugs undergoing clinical trials. Sci Transl Med. 2019;11(509).

59. Ellis LM, Hicklin DJ. Pathways Mediating Resistance to Vascular Endothelial Growth Factor–Targeted Therapy. Clinical Cancer Research. 2008;14(20):6371.

60. Sabnis AJ, Bivona TG. Principles of Resistance to Targeted Cancer Therapy: Lessons from Basic and Translational Cancer Biology. Trends in Molecular Medicine. 2019;25(3):185-97.

61. Gottesman MM. Mechanisms of Cancer Drug Resistance. Annual Review of Medicine. 2002;53(1):615-27.

62. Holohan C, Van Schaeybroeck S, Longley DB, Johnston PG. Cancer drug resistance: an evolving paradigm. Nature Reviews Cancer. 2013;13(10):714-26.

CHAPTER 5: MODELING PATIENT RESPONSE IN BLADDER CANCER TO GEMCITABINE CISPLATIN COMBINATION TREATMENT

Introduction

Chapters 3 and 4 focused specifically on drug response in *in vitro* cell lines. The benefit of leveraging *in vitro* cell lines in statistical learning models is directly related to the simplicity that *in vitro* drug assays provide. Compared to tumor data, drug response in a cell line is relatively easy to interpret, the experimental conditions are absent of higher order complexity such as tumor or patient heterogeneity, and can capture a relatively large amount of genomics/omics variability with the corresponding variability in drug response. These attributes are highly desirable as grounds to construct, validate, and compare modeling approaches as well as *in silico* hypothesis generation, such as identifying relevant biomarkers or identifying possible underlying biological mechanisms of drug response. However, the other appeal of statistical learning is its possible application in diagnosis, prognosis, and treatment in a clinically relevant setting (1, 2). One of the current interests is how to leverage *in vitro* models to inform modeling applications in patient derived data.

As discussed, the high dimensionality of omics data can make a model prone to overfitting and can obscure true biological relevance. Feature selection is one method to eliminate confounding noise and isolate biologically relevant relationships. Because, the relationship between patient tumor data and drug response is not as clearly defined as it is in *in vitro*, strategies to isolate features in *in vitro* data to apply to tumor models have been explored such as COXEN (3). The methodology of COXEN was to extrapolate the

relationship between drug response and co-expression in cell lines to co-expression patterns in tumor data to identify a biologically relevant feature set in tumor drug response models (3). In a similar manner, we wanted to address the question of how cell line based models could inform feature selection in tumor models, particularly with respect to the observation that drug induced gene perturbation was a significantly better predictor of drug response.

It was demonstrated in the previous chapter that differentially expressed genes (DEGs) derived from drug induced changes in gene expression were the best predictors of drug response; however, these genes lacked any predictive power when applied to basal gene expression. Therefore, these genes would only be applicable in a clinical setting if tumor samples could be collected and cultured and gene expression data could be measured before and after drug exposure. This process would be time intensive and costly, making this approach impractical from a clinical perspective. Thus, how can drug induced gene perturbation be used in a clinically relevant manner? A systems view of biology places emphasis on the emergent properties that result out of the interaction of several individual components (4). In this view the emergence of drug response is a result of interactions among genes. Change in gene expression are assumed to be a reaction to drug exposure; however, in a systems view, the reaction of one gene might be indicative of a underlying system of genes. Therefore, changes in one gene might reflect a bigger underlying genetic state that as a whole accounts for the variability in drug response. Thus, in an unperturbed state it is not sufficient for expression of a single gene to predict drug response but the expression of the system as a whole. Therefore, as basal expression of DEGs identified through gene expression

changes are poor predictors of drug response, basal gene expression of the entire system of associated genes with a DEG might be more indicative of drug response. Such an approach would accomplish two tasks, first the identification of a basal gene expression signatures based on observed gene expression changes, and second, the identification of a basal gene expression signature that could be applied to patient derived tumors in a clinical setting. In the following work a method is presented that uses highly connected DEGs discovered by the network analysis of gene perturbations in chapter 4 to identify a basal signature for drug response. Those signatures are then applied to patient derived bladder tumors who received a combination treatment of gemcitabine and cisplatin (GC) to predict patient drug response.

Methods:

Data Acquisition and Preprocessing:

The Laval bladder cancer dataset is a 90 patient cohort of patients that received neo-adjuvant chemotherapy of either gemcitabine and cisplatin or methotrexate, vinblastine, adriamycin, and cisplatin (MVAC) prior to cystectomy ranging anywhere from T2a to T4b. Gene expression from the tumor was measured using Affymetrix U133 2-Plus gene arrays. Survival data included the chemotherapy regimen of each therapy, the survival time in years, censorship status at the survival time given, and the tumor stage of the cystectomy. The censorship status of the patient was one of four possibilities; death from disease uncensored, disease recurrence, death from other causes, and alive at time of measurement. Patients with a status of either death of disease or disease recurrence were considered uncensored and patients with a status of alive of died of other causes were considered censored. Array data was normalized

using frozen robust multiarray analysis (fRMA) (5). Because these models were based on perturbation data from the NCI Transcription Pharmacodynamics Workbench (NCI-TPW) (6), only patients receiving GC treatment were considered. In total there were 36 patient samples, 18 of which were uncensored and 18 censored.

Gene Signature Selection (DEGs):

There are three different sets of gene signatures that are used in models; David derived differentially expressed genes (dDEGs), basal differentially expressed genes (bDEGs), and randomly selected differentially expressed genes (rDEGs). From Chapter 4, gene signatures from drug induced changes in gene expression were identified and network analysis was used to identify genes with maximal clique participation. David Bioinformatics Resources 6.8 (7, 8) was used to identify related genes that had an similarity score greater than 0.35 (refer to appendix C for more information of the similarity score) of the top ten clique participating genes identified in the network analysis (Figure 5.1). This resulted in 203 unique genes represented by 495 U133 2-plus probes for cisplatin and 246 for gemcitabine represented by 514 U133 2-plus probes. Basal DEGs (bDEGs) consisted of all DEGs identified in Chapter 4 by correlation-based feature selection on non-drugged cell lines in the NCI-TPW. Lastly random genes (rDEGs) served as a control where 495 random probes were selected for cisplatin and 514 random probes were selected for gemcitabine.

Modeling:

The test to train ratio was 1/3 testing and 2/3 training. For each test and train set the number of uncensored samples was equal to the number of censored samples thus for testing it was 6 to 6 and 12 to 12 for training. We utilized survival support vector

machines (sSVM) which predicts the relative rank at which events (dod or disease recurrence) occur, taking into account input from censored data (sSVR is expanded on in appendix C of this thesis) (9). Survival support vector machines are implemented using the scikit-survival python module for survival analysis (10). Both linear and radial basis function (rbf) kernels are used. SVMs require two parameters, α and γ, that are typically optimized using cross validation; however, since the dataset is small and contains many censored data points cross validation would not be possible. Therefore a range of the penalty parameter, α, was explored (0.001,0.01,0.1,1) and γ was set to 1 over the number of features. The ability of the model to correctly predict the order of events is characterized by the concordance index (11), which measures the concordance between the rankings predicted and those actually seen. For this score, 1 is a perfect concordance, 0.5 would indicate picking the rank of events at random, and 0 would be ranking the events completely opposite of the correct order. This was repeated 50 times for each different kernel, parameter setting, DEG set, and combination of drugs (GC) and single drugs (i.e. Cisplatin or Gemcitabine). A pairwise t-test was used to determine if there was any significant difference based on the different application of DEGs. This process is outline in Figure 5.1.

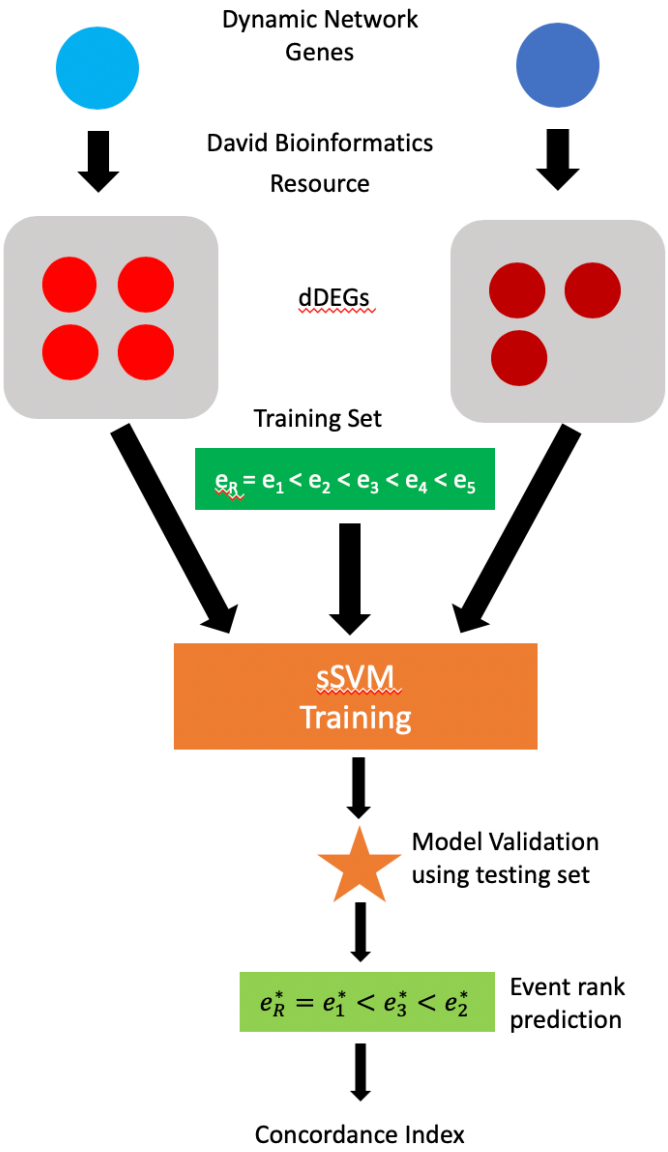


Figure 5.1. General outline of how the modeling workflow

Results:

David Identified DEGs Predict Survival Better Compared to Other DEGs:

With respect to an RBF the range of the concordance scores varied from a minimum of 0.51 using dDEGs for cisplatin to a maximum of 0.649 using dDEGs for Gemcitabine (Table 5.1). Likewise, the best average performance over all levels of α

were the dDEGs gemcitabine (0.641) models and the worst performing models were dDEG models for cisplatin (0.527) (Figure 5.2). For combination (GC) treatment models using dDEGs (0.605) performed better on average than all other DEG models with the exception of dDEG gemcitabine models (Figure 5.2). Furthermore, dDEG GC models performed significantly better than rDEG models for three of the four alpha levels (Figure 5.3). However, compared to bDEGs in GC models, dDEGs only performed significantly better for α=0.001; nonetheless, this was also the best performing dDEG GC model (0.625) in additional to exceeding all other DEG models with the exception of the dDEG Gemcitabine model (0.649) (Table 5.2). Furthermore, the performance comparison between cisplatin and gemcitabine models was completely opposite, dDEG models consistently performed better in gemcitabine while consistently performing the worst in cisplatin for all levels of α (Figure 5.3 and Table 5.1). Additionally, it is of note, that for GC models bDEGs never performed significantly better the rDEGs over all levels of α (Figure 5.3).

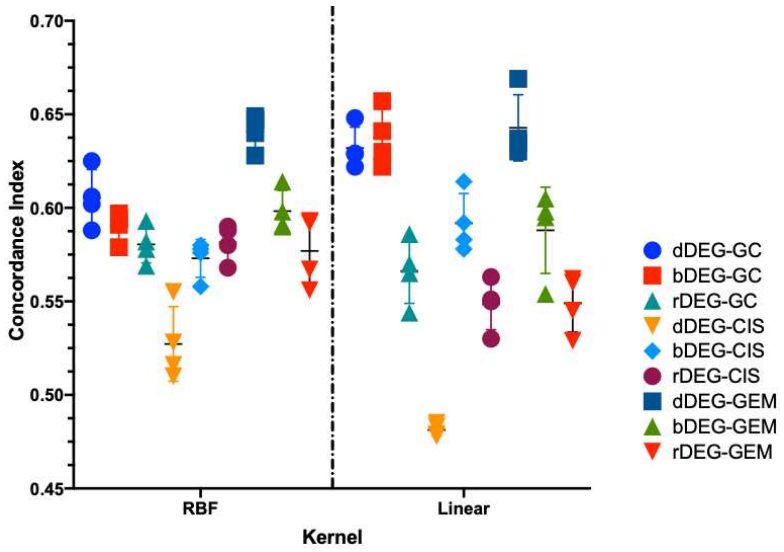


Figure 5.2 . Average Concordance index for each different set of DEGs for all four levels of α

Table 5.1 Minimum, mean, and maximum values for the various models at different levels of α for sSVR models with rbf kernels.

α	0.001	0.01	0.1	1
GC	0.325 0.625 0.842	0.216 0.606 0.8	0.3958 0.602 0.766	0.408 0.588 0.75
Basal GC	0.275 0.591 0.816	0.373 0.597 0.857	0.333 0.579 0.762	0.353 0.594 0.818
R-GC	0.3 0.593 0.812	0.216 0.582 0.756	0.292 0.569 0.786	0.333 0.578 0.784
CIS	0.296 0.555 0.737	0.137 0.528 0.78	0.229 0.51 0.745	0.286 0.516 0.705
Basal CIS	0.25 0.578 0.769	0.392 0.58 0.822	0.364 0.558 0.738	0.326 0.576 0.818
R-CIS	0.325 0.588 0.889	0.177 0.568 0.756	0.395 0.59 0.805	0.353 0.58 0.818
GEM	0.35 0.649 0.868	0.442 0.647 0.844	0.447 0.64 0.81	0.438 0.628 0.804
Basal GEM	0.3 0.591 0.822	0.386 0.59 0.786	0.375 0.598 0.881	0.373 0.614 0.818
R-GEM	0.3 0.5919 0.816	0.275 0.593 0.776	0.25 0.556 0.75	0.326 0.567 0.804

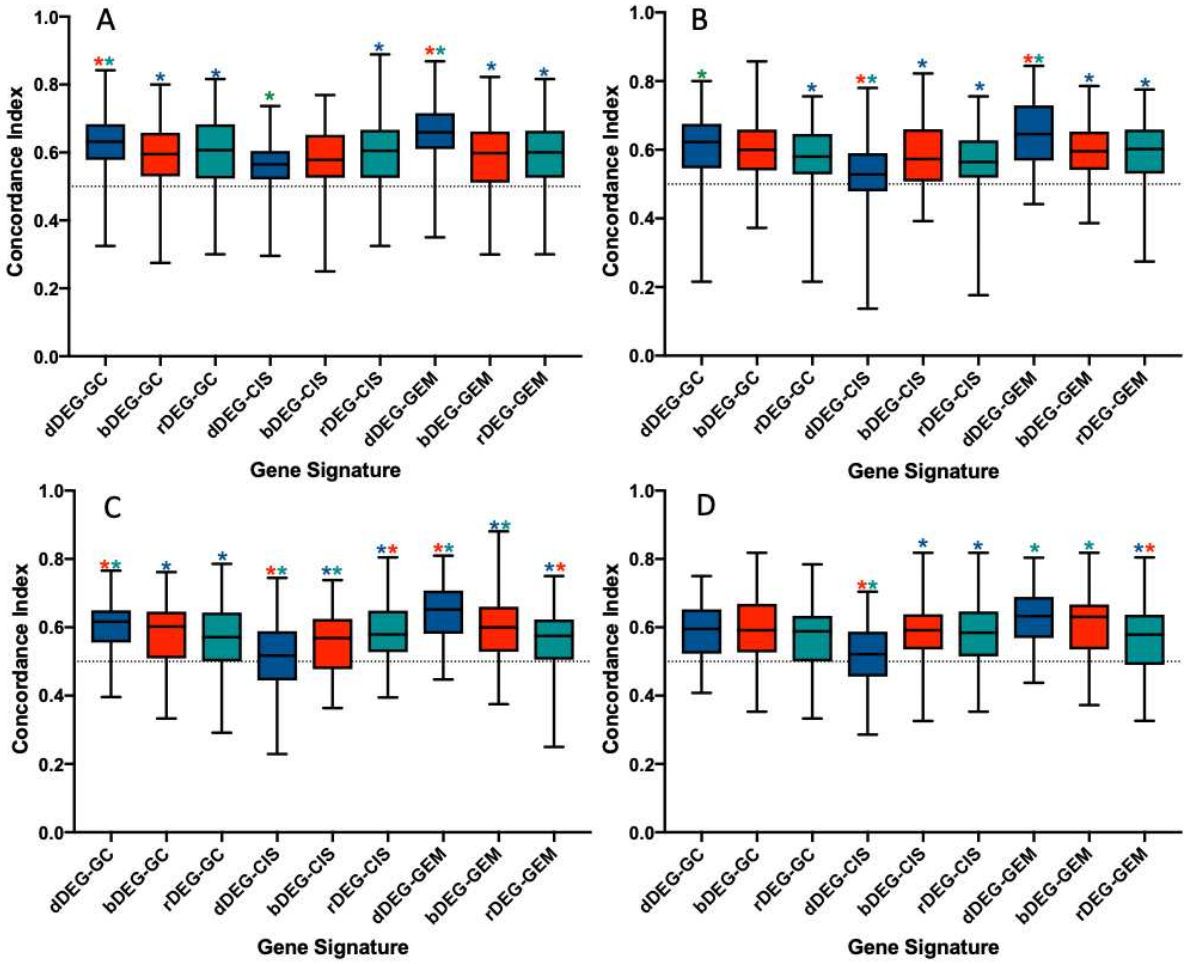


Figure 5.3. Distribution of the concordance scores over 50 random training and testing splits for A. α=0.001 B. α=0.01 C. α=0.1 and D. α=1 using sSVM with a rbf kernel. The stars above each bar indicate a significant difference (p<0.05) by pairwise t-test for the same drugs modeled with a different set of DEGs.

Application of a linear kernel function resulted in the best performing model in gemcitabine using dDEGs at an α level of 1 (0.669) (Table 5.2). For combination models both dDEG (0.632) and bDEG (0.638) models performed about the same (figure 5.2) and performed significantly better than rDEG models at all levels of α (Figure 5.4). Similar to RBF models dDEG gemcitabine models (0.643) had the best average

performance over all and dDEG cisplatin models consistently performed the poorest (Figure 5.2). Furthermore, dDEG gemcitabine models performed significantly better over-all α while the dDEG models for cisplatin performed the worst. With the exception of $\alpha = 0.1$ in gemcitabine bDEG models always outperformed rDEG models (Figure 5.4).

Table 5.2 Minimum, mean, and maximum values for the various models at different levels of α for sSVR models with linear kernels.

α	0.001	0.01	0.1	1
dDEG-GC	0.319 0.622 0.878	0.381 0.629 0.826	0.386 0.629 0.864	0.425 0.648 0.875
bDEG-GC	0.37 0.63 0.881	0.422 0.641 0.837	0.366 0.622 0.86	0.354 0.657 0.881
rDEG -GC	0.333 0.57 0.854	0.317 0.565 0.786	0.261 0.544 0.822	0.225 0.586 0.796
dDEG-CIS	0.192 0.478 0.829	0.186 0.485 0.787	0.13 0.479 0.773	0.184 0.483 0.896
bDEG-CIS	0.275 0.583 0.837	0.4 0.592 0.857	0.304 0.578 0.84	0.26 0.614 0.826
rDEG-CIS	0.196 0.55 0.84	0.34 0.551 0.745	0.326 0.53 0.796	0.271 0.563 0.829
dDEG-GEM	0.326 0.6301 0.824	0.4 0.635 0.809	0.386 0.637 0.886	0.45 0.669 0.857
bDEG-GEM	0.255 0.598 0.842	0.362 0.595 0.838	0.24 0.554 0.881	0.32 0.605 0.864
rDEG-GEM	0.319 0.56 0.857	0.296 0.545 0.811	0.261 0.529 0.784	0.225 0.562 0.8

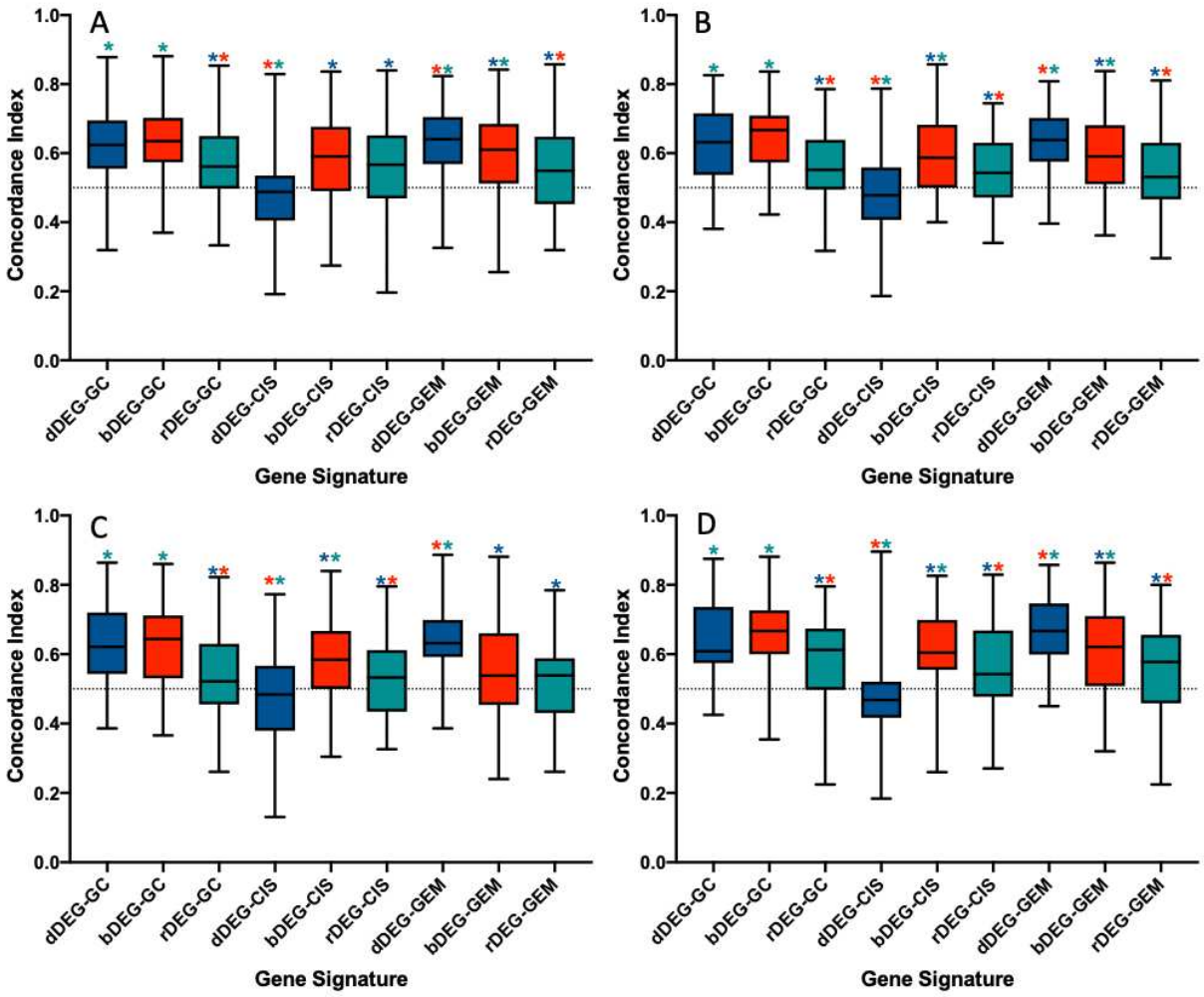


Figure 5.3. Distribution of the concordance scores over 50 random training and testing splits for A. α=0.001 B. α=0.01 C. α=0.1 and D. α=1 using sSVM with a linear kernel. The stars above each bar indicate a significant difference (p<0.05) by pairwise t-test for the same drugs modeled with a different set of DEGs.

Combined DEGs are additive in performance for GC Models

We wanted to see if combining dDEGs and the bDEGs could lead to better models for both a RBF kernel and linear kernel. For models utilizing the RBF kernel models using combined dDEGs and bDEGs did not perform better than any single drug model while the dDEG gemcitabine model continued to perform best (Figure 5.5 A, Table 5.3). However, models using linear kernels performed better in both GC and cisplatin models; additionally, the highest concordance index was achieved using

combined DEGs for GC models at 0.68 and was significantly better than ($p<0.05$) than all single drug models (Figure 5.5 B, Table 5.3).

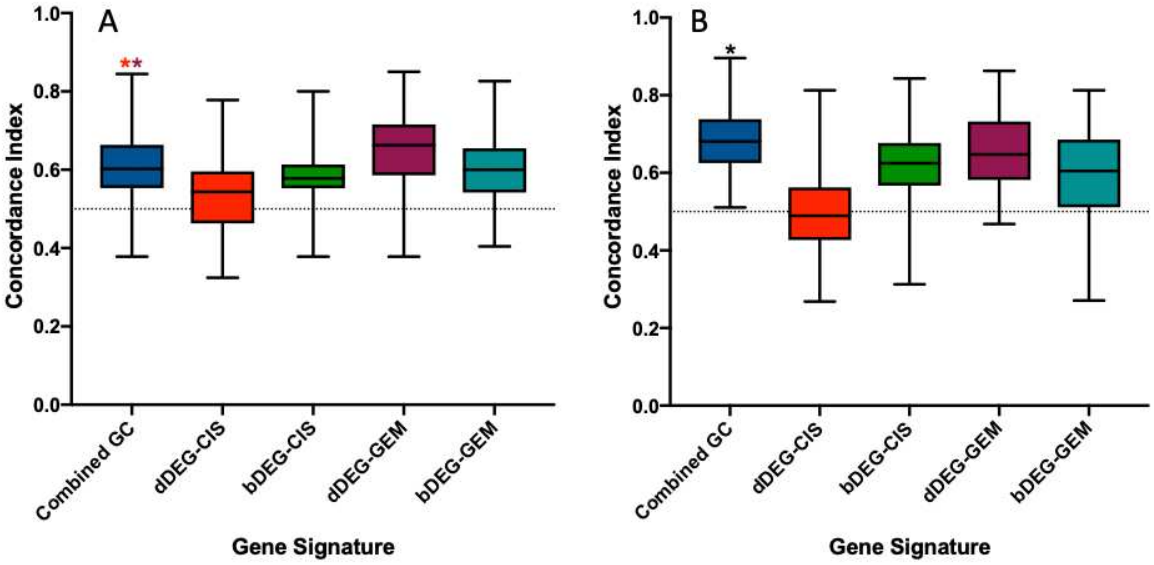


Figure 5.5. Distribution of 50 test train splits using combined dDEGs and bDEGs compared to using a single set of DEGs in a single drug. A. RBF kernel B. Linear Kernel. Stars indicate which DEG sets are significantly different by a paired t tests. The black star in B reflects the the combined DEG set is significantly different from all other DEG sets.

Table 5.3 Minimum, mean, and maximum concordance index over 50 test train splits for the respective set of DEGs and kernels

Kernel	RBF	Linear
dDEG-GC	0.351 0.613 0.822	0.362 0.623 0.875
bDEG GC	0.351 0.588 0.8	0.354 0.658 0.837
Combined GC	0.378 0.604 0.844	0.511 0.68 0.896
dDEG-CIS	0.324 0.532 0.778	0.268 0.505 0.813
bDEG-CIS	0.378 0.587 0.8	0.313 0.623 0.843
Combined CIS	0.324 0.569 0.769	0.383 0.631 0.813
dDEG-GEM	0.378 0.65 0.85	0.468 0.661 0.863
bDEG-GEM	0.404 0.597 0.826	0.271 0.589 0.813
Combined GEM	0.405 0.625 0.822	0.447 0.653 0.875

Discussion:

This work demonstrated three central results. First, drug induced changes in gene expression from *in vitro* cell lines could be used to generate basal signatures to successfully predict patient drug response. Second, this derived signature was better at predicting patient drug response than DEGs derived from basal gene expression of *in vitro* cell lines as well as randomly selected genes. Finally, DEGs selected for each individual drug could be combined to improve patient response to a combination

chemotherapy treatment. The three of these taken together have the potential to have significant impact to building models to predict drug response in patients to combination chemotherapy which is much more clinically common than treatment with a single agent.

Despite these promising results, it is best to remain cautiously optimistic as this is only a pilot study. The results indicate that there is a significant improvement for dDEG models, especially with respect to gemcitabine; however, these experiments would have to be conducted in a larger patient cohort to gain a greater confidence. Likewise, the models were also limited by the fact that half of the data was censored limiting both training and validation of the performance of the model. Ideally, the majority of the patients would be uncensored; however, in the case of cancer this would be a substantially poor treatment as most patients would have died or had reoccurrence of disease. A more concrete endpoint, likely resulting in a more definitive results if dDEGs were indeed better, might be the 5 year survival of patients; nonetheless, the data is always the biggest limiting factor. Additionally, the size and the composition of the dataset did not allow for a reliable cross-validation step to optimize the model parameters, α and γ, where both parameters could be optimized for the feature set independently. However, this would unlikely have a major effect as the different feature sets have approximately the same dimensionality such that the kernel function values, with respect to RBF parameter γ, are similar in the order of magnitude and thus differences results from true differences in the data and is not just an effect of dimensionality. Likewise, with respect to different α levels the lowest performing dDEG gemcitabine model, which seems to be the major driver in dDEG gc models, is greater

than the best models in bDEG and rDEG for gemcitabine. Furthermore, the dDEGs seem to be optimal only for gemcitabine where cisplatin seems to do worse despite the fact that *in vitro* cisplatin models which used drug induced changes in gene expression were among the top performers including better than similar models in gemcitabine. This suggest that perhaps overfitting is occurring in the *in vitro* models and that top performing feature sets in *in vitro* models do not generalize well to tumor data.

One of the difficulties when moving from *in vitro* data to tumor data is the fact that tumor data are vastly more complex. Factors such as tumor heterogeneity and tumor microenvironment most likely play a substantial role in patient drug response. So an interesting question arises can more accurate models be achieved though better feature selection, whether that is based on basal gene expression, extrapolated from drug induced changes in gene expression, or by other methods, or is this a fundamental limitation of gene expression models such that they cannot account for these other factors. If this limitation is the case, what kind of information do we need to produce better models? This is where statistical modeling approaches might be able to play an informative role whereby the solution to produce better models might also be an avenue for better understanding drug response.

Conclusion

In vitro cell lines have served as the bedrock of early drug discovery and development (12, 13). In the modern era drug response can be measured for a large amount of drugs and cell lines in a highly efficient manner making large amounts of data available. Additionally, advances in genomics technologies has allowed genomic characterization of the cell relatively easy and low cost. The wide availability of both

drug response and gene expression for a wide variety of *in vitro* cell lines provides a large and relatively simple platform to build, test, and explore computational models; however, there remains the question of how modeling of these cell lines can be used in a clinically relevant setting. One possibility that has been put forth is using *in vitro* drug response to select a relevant gene set to apply to clinical data. However, previous attempts, such as the COXEN algorithm (3), have seen little success in clinical trials (14). A possible explanation of this is that basal gene expression is not a robust predictor of drug response as was shown in previous chapters (15). However, it was demonstrated in Chapter 4 that signatures from drug induced changes in gene expression were far better as predictors of drug response. Nonetheless, clinical application of these signatures would be impractical at best and how to apply these signature clinically became a central question. This work has shown that a systems based approach, which utilizes drug induced dynamic gene expression as a “Rosetta stone” to uncover a broader underlying biological relationships is a possible application of *in vitro* experimental data to the more practical clinical side. As such, it is important that these continue to be developed and, as the results of chapter four suggests, puts an even greater emphasis on the need to quantify drug induced changes in gene expression for a greater number of cell lines and drugs.

REFERENCES

1. Haendel MA, Chute CG, Robinson PN. Classification, Ontology, and Precision Medicine. New England Journal of Medicine. 2018;379(15):1452-62.

2. König IR, Fuchs O, Hansen G, von Mutius E, Kopp MV. What is precision medicine? European Respiratory Journal. 2017;50(4):1700391.

3. Lee JK, Havaleshko DM, Cho H, Weinstein JN, Kaldjian EP, Karpovich J, et al. A strategy for predicting the chemosensitivity of human cancers and its application to drug discovery. Proceedings of the National Academy of Sciences of the United States of America. 2007;104(32):13086-91.

4. Hillmer RA. Systems Biology for Biologists. PLOS Pathogens. 2015;11(5):e1004786.

5. McCall MN, Bolstad BM, Irizarry RA. Frozen robust multiarray analysis (fRMA). Biostatistics. 2010;11(2):242-53.

6. Monks A, Zhao Y, Hose C, Hamed H, Krushkal J, Fang J, et al. The NCI Transcriptional Pharmacodynamics Workbench: A Tool to Examine Dynamic Expression Profiling of Therapeutic Response in the NCI-60 Cell Line Panel. Cancer Research. 2018;78(24):6807.

7. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1-13.

8. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44-57.

9. Van Belle V, Pelckmans K, Van Huffel S, Suykens JAK. Support vector methods for survival analysis: a comparison between ranking and regression approaches. Artificial Intelligence in Medicine. 2011;53(2):107-18.

10. Pˆlsterl S, Navab N, Katouzian A. An Efficient Training Algorithm for Kernel Survival Support Vector Machines. ArXiv. 2016;abs/1611.07054.

11. Steck H, Balaji K, Cary D-o, Philippe L, Raykar VC. On Ranking in Survival Analysis: Bounds on the Concordance Index. 2008:1209--16.

12. DeVita VT, Chu E. A History of Cancer Chemotherapy. Cancer Research. 2008;68(21):8643.

13. DeVita VT, Jr., Rosenberg SA. Two hundred years of cancer research. N Engl J Med. 2012;366(23):2207-14.

14. Flaig TW, Tangen CM, Daneshmand S, Alva AS, Lerner SP, Lucia MS, et al. SWOG S1314: A randomized phase II study of co-expression extrapolation (COXEN) with neoadjuvant chemotherapy for localized, muscle-invasive bladder cancer. Journal of Clinical Oncology. 2019;37(15_suppl):4506-.

15. Mannheimer JD, Duval DL, Prasad A, Gustafson DL. A systematic analysis of genomics-based modeling approaches for prediction of drug response to cytotoxic chemotherapies. BMC Medical Genomics. 2019;12(1):87.

CHAPTER 6: CONCLUSION

Omics data are playing an increasingly import role in both the understanding of the molecular underpinnings of cancer as well as the development of precision medicine. The large scale and complexity of the data has and will continue to require innovative approaches of computational methods to generate actionable insight for both advancements in treatment and knowledge of cancer as a disease. Despite, successful application of statistical learning models applied to patient tumor data and *in vitro* cell line data the application of these models clinically is still limited. The translation of these computational methods from the research “bench” to a clinical setting requires a comprehensive understanding of the intersection between the biology, computational methodology, and the data. This work has provided a systematic assessment of computational approaches and prediction of *in vitro* drug sensitivity for cytotoxic chemotherapies across a number of pan-cancer cell lines. Additionally, the role of drug exposure and genome wide drug induced gene perturbation has been explored through supervised learning of drug response in *in vitro* drug screens among a pan-cancer databases of cell lines. On the whole, the intention of this work has been to establish a better understanding of computational techniques and the role of gene expression in predictive models of drug response to inform both advancements in modeling and experimental data acquisition practices.

Early on Staunton et al. showed that gene expression could be used to predict chemosensitivity for 80 drugs in the NCI60 cell line panel (1). Much more recently the NCI-DREAM competition established that a variety of modeling approaches applied to

different types of descriptive data could successfully predict drug response in a cohort of breast cancer cell lines (2). Some of the key applications of that study that influenced the work presented in Chapter 3 was the general outperformance of non-linear regression methods compared to linear regression methods, gene expression was most influential in model performance, and the comparative performance of the most complicated method and the simplest method (2). Given these observations, the work presented in Chapter 3 aimed at a similar comparison between linear and non-linear gene expression-based models in pan-cancer cell lines in the GDSC and NCI60 databases specifically for cytotoxic chemotherapies.

The main conclusion of chapter three was that the modeling methodology and feature selection, was much less important than the influence of histotype on drug response. More specifically it was shown that non-linear models only slightly outperformed linear models indicated by the comparative performance between non-linear support vector regression and principal components regression. Furthermore, it was demonstrated that various approaches to filter based feature selection also did not lead to significant improvements in performance; however, on the contrary, models using genes that lacked any significant statistical relationship with drug response performed comparably to models using features selected by correlative strength with drug response. This phenomenon was attributed to the ability of as few as 250 random genes to distinguish between histotype coupled with correlative relationships between drug response and histotype. Therefore, this work suggested that histotype could play a major predictive role in model performance and therefore has to be taken into account when determining the overall generalizability of pan-cancer models. Additionally, feature

selection methods that could differentiate between the relationships strictly resulting from drug histotype interactions and features where the relationship was more specific of drug-gene associations.

The connectivity map explored the basic premise that drug induced gene perturbations might be indicative of a causal relationship between disease states and drug mechanism (3, 4). This concept suggests that drug induced gene dynamics might play a pivotal role in drug response; however, the application of gene dynamics in predictive models of drug response has been limited. Using data generated in the recent development of the NCI Transcriptional Pharmacodynamics Workbench (5) Chapter 4 explored the relationship between drug exposure and drug response. Given gene expression of the NCI60 cell lines at 2 hours, 6 hours, and 24 hours post drug exposure with a high dose of drug, low dose of drug, and no dose for 15 compounds, support vector regression was implemented to predict the log IC_{50}. Particularly, at each time point the predictive capabilities of models built with basal gene expression, drug perturbed gene expression, and the difference between drug perturbed and basal data at both high and low dose of drug were assessed. The results demonstrated that drug perturbed gene expression and differences in perturbed and basal gene expression performed substantially better for both high dose and low dose of drugs at 24 hours. Additionally, at the two and six hour time points little difference was seen between perturbed and basal gene expression while performance using the difference of the two was substantially lower. These results suggest that perturbed gene expression might be much more indicative of drug response encouraging perfroming additional perturbation experiments in additional drugs and cell lines at various timepoints and dosing

concentrations to improve models but more importantly to gain insights into the relationships between gene expression, drug exposure, and drug response.

One of the central questions raised in Chapter 4 was if there was if an underlying relationship between differentially expressed genes in basal gene expression and perturbed gene expression. We approached this question by using differentially expressed gene signatures developed from one gene expression profile as feature inputs for models constructed from different gene profiles. We showed that DEG signatures derived from changes in gene expression, with respect to a treated versus untreated state, had very little predictive capability in models using basal gene expression. Likewise, DEG signatures derived from basal gene expression had diminished predictive capabilities using gene expression profiles from changes in gene expression. This result suggests that statistical dependencies between gene signatures and drug response are not necessarily reflected as statistical dependencies between basal and dynamic gene expression. As a result it might be misleading to use perturbation data as a feature selection method for models in basal data; for example, the drug perturbation signatures contained in the connectivity map do not necessarily translate to predictive signatures of drug response in basal data.

Network theory allows for a rigorous mathematical formalism to represent and analyze complex interactions of genes and proteins (6). The structure of a network, a graphical and mathematical description of the network, is referred as its topology. A theory, proposed by Barabási, suggests that disease states are represented by changes in network topology (7). This concept is naturally extended to pharmacology, as different interactions, network topologies, may give rise to differences in drug response (8). In

chapter 4 it was demonstrated that a network representation of DEG interactions showed differences in topological features when comparing basal DEGs to DEGs derived from the magnitude of change of gene expression after drug treatment (ΔC) . Specifically, it was shown that ΔC DEGs showed a greater level of connectivity than basal DEGs as evidenced by higher clique participation and average clustering coefficient. This suggests that the gene interactions seen in drug induced gene dynamics may have a significant role in determining drug response. Furthermore, genes with high clique participation in both DEGs from basal and ΔC gene profiles had some evidence of playing a role in cancer. Altogether, this shows that gene expression network topology in both basal and drug treated *in vitro* cell lines might be a useful avenue in understanding the complexities that underly variations in drug response.

Finally, Chapter 5 presented a direct extension of how *in vitro* drug induced gene dynamics could be translated to a clinically useful model. We identified a predictive basal gene expression signature based on the analysis of drug induced gene dynamics predictive models and applied those signatures for survival analysis in bladder cancer patients who received neo-adjuvant chemotherapy of a combination treatment with cisplatin and gemcitabine (GC). Particularly, these gene signatures performed significantly better than gene signatures derived from *in vitro* cell line data indicating that drug response information gained from drug induced gene dynamics could be applied to basal gene expression, which might be the only available data, as would be the case in a clinical setting. This demonstrated that *in vitro* drug perturbation experiments are clinically relevant and there could be a great benefit for conducting experiments such as those demonstrated in the NCI Translational Pharmacodynamic Workbench (5)

The Path Forward:

The role of computational biology is ever changing with advancements in computational techniques, improved methods of data acquisition and quantification, and general advancements of knowledge into biological processes and disease. As more data becomes available the use of statistical learning will become more prominent and advanced. With respect to oncology, statistical learning serves a couple of fundamental purposes. Firstly, is an application to precision medicine, developing narrow and focused strategies to provider better diagnosis, prognosis, and treatment of cancer. The second, which is certainly not mutually exclusive from the first, is as a means to explore the complex relationships that dictate drug response for hypothesis generation, drug discovery, and development of novel treatment for cancer as a disease.

Network theory has gained a lot of attention to explain biological phenomena. One of the prevailing arguments for a network based approach is that most biological processes are not the result of a single independent protein/gene but are the result of a number of interactions within multiple proteins/genes. Network medicine is a conceptual framework that frames disease states as topological features in biological networks (7). In particular, for cancer, it was shown how topological differences in protein-protein interaction networks account for breast cancer prognosis (9). The use of network topologies to describe disease has led to the concept of network pharmacology which looks at network topology as a means to identify new candidates as drug targets (10-12). Given this idea that disease, network topology, and drug action are connected it seems natural to extend this idea to understand drug response. For example, are differences in drug response reflected in difference in network topology? This, in some

sense, was the basis behind chapters 4 and 5. In chapter 4 we demonstrated that DEGs selected using drug induced gene dynamics showed high clique participation and connectivity compared to basal DEGs. Furthermore, by using these gene to identify related genes, based on functional annotations in DAVID (13, 14), we were able to identify a basal signature to use in tumor data. This work utilized very basic network metrics and network construction methods and further work should look at enhanced methods of determining network topology as well as measuring network topology. For example, in another DREAM challenge several different teams applied network inference algorithms to reconstruct networks in E. coli and S. cerevisiae (15). Therefore, these methods might be better alternatives to construct gene networks; however, the question also becomes how can network information be used. The first most obvious way is to use network connectivity as a method of finding import features as was done in Chapter 5. Alternatively, could it be possible to learn on networks directly? Convolution neural networks and kernel methods might provide such means.

Kernel methods have become useful tools in computational biology mainly because the flexibility of kernel functions allows for construction of a kernel that is specific to the problem at hand (16). For example, specific string matching kernels have been used for protein classification (17). Based on the definition of a kernel, a continuous function $k(\boldsymbol{x}_i, \boldsymbol{x}_j) \rightarrow \mathbb{R};\ \boldsymbol{x}_i, \boldsymbol{x}_j \ni \boldsymbol{X}$ such that

$$(\boldsymbol{x}_i, \boldsymbol{x}_j) = k(\boldsymbol{x}_j, \boldsymbol{x}_i)\ and\ \sum_{i=1}^{N}\sum_{j=1}^{N} k(\boldsymbol{x}_i, \boldsymbol{x}_j)c_i c_j \geq 0$$

it is possible to define a kernel function $k(G_i, G_j) \rightarrow \mathbb{R}$ on networks G_i and G_j. If the topology of G_i is dependent on its response to a drug the kernel function can be used to

define a scaler value to the differences in topology, (i.e. $k(G_i, G_j) \rightarrow \mathbb{R}$), in turn this allows any kernel method, such as SVM, to be used to predict drug response as a function of network topology. Likewise, convolutional neural networks (CNN) have become one of the most popular and successful methods in image recognition problems (18). Thus, how can the power of convolutional neural networks be applied to learning on biological networks. The interaction between two genes can be described by an adjacency matrix where given two nodes/genes the ith/jth entry in the matrix indicates the interaction between genes i and j. For example, in Figure 6.1 a 5 node network where the width of the edges represents the magnitude of the interaction between nodes is translated into a 25 pixel gray scale image where the shade of each pixel is reflective if the interaction between nodes.

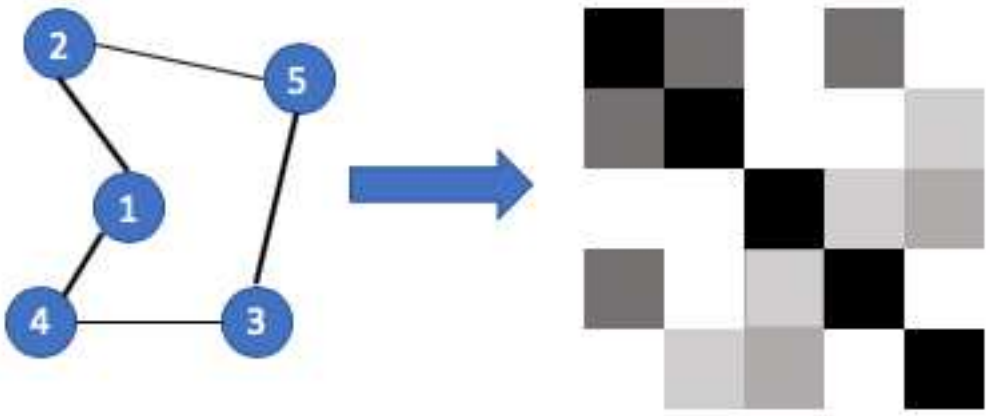


Figure 6.1. The five node network on the right is translated into a 25 pixel image. In the network the width of the edge connecting the two nodes represents the magnitude of the interaction. While in the translated image the shade of the block represents the magnitude of the interaction between nodes.

Therefore, different network topologies and interactions among genes would be reflected by pixilated structures in the image. If certain network topologies and gene interactions gave rise to differences in drug response it would be possible to train a

CNN to distinguish different drug responses based on the network topologies that had been translated into images.

In closing, as the ability to measure the molecular activities of the cell becomes even better large sets of omics data will continue to be collected. Advancements in computational biology will really determine how that data will be used to understand and treat disease. Constant advancements in machine learning, particularly deep learning, will continue to be made and will find medical applications. However, while deep learning has shown impressive capabilities in certain areas it must be remembered that robust models are models that reflect the biology itself and thus for deep learning to be successful it must also reflect the underlying biology. Therefore, looking ahead, finding new ways, such as networks, to impart biological function into complex learning models will be essential to the development of clinically useful tools. While challenging, I am looking forward to see, and hopefully be a part of, the creativity that can drastically alter the way we view and treat cancer.

REFERENCES

1. Staunton JE, Slonim DK, Coller HA, Tamayo P, Angelo MJ, Park J, et al. Chemosensitivity prediction by transcriptional profiling. Proceedings of the National Academy of Sciences. 2001;98(19):10787.

2. Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nature biotechnology. 2014;32(12):1202-12.

3. 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. 2006;313(5795):1929.

4. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell. 2017;171(6):1437-52.e17.

5. Monks A, Zhao Y, Hose C, Hamed H, Krushkal J, Fang J, et al. The NCI Transcriptional Pharmacodynamics Workbench: A Tool to Examine Dynamic Expression Profiling of Therapeutic Response in the NCI-60 Cell Line Panel. Cancer Research. 2018;78(24):6807.

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

7. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nature Reviews Genetics. 2010;12:56.

8. Boezio B, Audouze K, Ducrot P, Taboureau O. Network-based Approaches in Pharmacology. Molecular Informatics. 2017;36(10):1700048.

9. Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, Faria D, et al. Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nature Biotechnology. 2009;27(2):199-204.

10. del Sol A, Balling R, Hood L, Galas D. Diseases as network perturbations. Current Opinion in Biotechnology. 2010;21(4):566-71.

11. Hopkins AL. Network pharmacology: the next paradigm in drug discovery. Nature Chemical Biology. 2008;4:682.

12. Hopkins AL. Network pharmacology. Nature Biotechnology. 2007;25(10):1110-1.

13. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1-13.

14. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44-57.

15. Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, et al. Wisdom of crowds for robust gene network inference. Nat Methods. 2012;9(8):796-804.

16. Kernel Methods in Computational Biology Scholköpf BT, Koji; Vert, Jean-Philippe editor. Cambridge Massachusetts The MIT Press 2004.

17. Christina Leslie RK, Eleazar Esken. Inexact Matching String Kernels for Protein Classification In: Bernard Schölkopf KT, Jean-Philippe Vert, editor. Kernel Methods in Computational Biology Cambridge Massachusetts MIT Press 2004. p. 95-112.

18. Egmont-Petersen M, de Ridder D, Handels H. Image processing with neural networks—a review. Pattern Recognition. 2002;35(10):2279-301.

APPENDIX A: SUPPLEMENTARY MATERIAL FOR CHAPTER 3

Mean Absolute Difference Results:

The Spearman correlation is a relative measure that is particularly useful in describing general trends in the data, however, other measures such as root mean squared error (RMSE) and mean absolute difference (MAD) are better at quantifying the overall accuracy of the model. Here we show that the same quantitative patterns are captured by Spearman correlations are also present in MAD. Average MAD values accompanied with the corresponding average spearman correlations, and the Spearman correlation between MAD values and Spearman correlations for each model are reported in supplementary table 1. Average MAD values ranged from a minimum of 0.881 to 1.391 with a mean value of 0.931 and standard of deviation 0.067. Comparatively, the average Spearman correlations ranged from -0.011 to a maximum value of 0.334 with a mean of 0.253 and a standard of deviation of 0.083. As with Spearman correlation average MAD values gave relative performance as NLSVR>PCR>SVRLN>ANN. The MAD value of a model was strongly negatively correlated to a model Spearman correlation (Figure A.1 C). The correlation between model average MAD scores and average model Spearman correlations also had a strong negative correlation -0.8418 (Figure A.1 B.). MAD scores generally reflect the behavior that drugs with higher correlations generally have lower MAD scores (Figure A.1 B.). However, within each drug there is not a strong correlation between a given datasets correlations and their MAD scores (Figure A.1 B.). This can be explained by the influence of histotype in predictions. For example, for drugs such as vorinostat,

bortezomib, and methotrexate the high correlations are attributed to correctly identifying the association between histotype and drug response, however, for these drugs, within a given histotype the predicted values do not necessarily have a strong correlation with the experimental values. Therefore, within a histotype the MAD score is more a reflection of the variability within each histotype that results as noise around an average IC50 for that histotype. Interestingly, for drugs which show smaller response to histotype, bleomycin, doxorubicin, cytarabine, and mitomycin slight negative correlation can be seen between MAD and Spearman Correlation (Figure A.1 B). Additionally, the range of experimental IC50's can have an effect. As an extreme example Cisplatin, which with respect to Spearman correlation, performs the worst has the lowest MAD values. This results from the fact that the range of IC50's for Cisplatin is much smaller than the other drugs, which might also explain why it is particularly hard to yield good predictions. Therefore, variability of the given dataset plays a role in MAD scores which is minimal with respect to Spearman correlation.

Table A.1: Each models average Spearman Correlation, Average MAD score, and the correlation between the average Spearman correlation and MAD along with the associated P value

	AVG r	AVG MAD	r vs MAD	p
LSQR_HIST_ONLY	0.206600941	0.99311685	-0.789215686	0.00016550
NNet1_DEG	0.268050903	1.001504775	-0.825635314	4.39E-22
NNet1_Limma	0.265734239	1.031461462	-0.684458844	7.06E-13
NNet1_None	0.170836501	1.39071894	-0.519894705	4.02E-07
PCR_BC_DEG	0.299126885	0.892740612	-0.742411924	2.16E-15
PCR_BS_DEG	0.328409628	0.89976564	-0.74097398	7.75E-16
PCR_BS_hist	0.231715541	0.927316692	-0.778454996	2.92E-18
PCR_CTR1	0.327679648	0.89026602	-0.79437076	1.94E-19
PCR_CTR1_10	0.145384873	0.94669059	-0.503735952	1.04E-06
PCR_CTR1_1000	0.282793893	0.907468456	-0.77946745	2.47E-18
PCR_CTR1_250	0.242837144	0.923793142	-0.741905437	6.83E-16
PCR_CTR1_500	0.264982445	0.917180188	-0.738787081	1.04E-15

PCR_CTR1_55	0.204242863	0.935326004	-0.731861901	2.61E-15
PCR_CTR2	0.259641397	0.913025063	-0.777908272	3.19E-18
PCR_CTR2_10	0.069563485	0.958145959	-0.11588539	0.29384029
PCR_CTR2_1000	0.235427127	0.924755282	-0.707765516	5.17E-14
PCR_CTR2_250	0.176621455	0.939921685	-0.608909588	7.98E-10
PCR_CTR2_500	0.187357689	0.93702075	-0.67718943	1.52E-12
PCR_CTR2_55	0.107959474	0.950592666	-0.267712868	0.01381718
PCR_DEG	0.338371044	0.89577949	-0.807168168	1.83E-20
PCR_DEG_10	0.203332115	0.939005307	-0.677533664	1.47E-12
PCR_DEG_1000	0.302904109	0.91008021	-0.804576288	3.00E-20
PCR_DEG_250	0.274990668	0.922032562	-0.732226385	2.49E-15
PCR_DEG_500	0.292483261	0.909676385	-0.761587527	4.08E-17
PCR_DEG_55	0.249617233	0.928768805	-0.734717019	1.79E-15
PCR_DEG_HIST	0.321946531	0.912223626	-0.778758732	2.78E-18
PCR_Limma	0.297227647	0.902728419	-0.722689076	8.44E-15
PCR_MRMR	0.315486982	0.92277003	-0.747777665	3.03E-16
PCR_None	0.336920624	0.886850982	-0.784732206	1.03E-18
PCR_RAN_CTR	-0.01183200	0.967311497	0.054247241	0.62406627
SVRLN_DEG	0.2840744	0.967909486	-0.801761669	5.07E-20
SVRLN_Limma	0.279914537	0.937635103	-0.685896527	6.05E-13
SVRLN_None	0.314662929	0.922962809	-0.79763086	1.08E-19
SVR_BC_DEG	0.311197619	0.890135989	-0.788775971	2.26E-18
SVR_BS_DEG	0.319205813	0.855917246	-0.756	<0.0001
SVR_BS_hist	0.323632754	0.893051569	-0.784266478	1.11E-18
SVR_CTR1	0.347084298	0.881359019	-0.812250683	6.84E-21
SVR_CTR1_10	0.128314394	0.947990492	-0.428895414	4.69E-05
SVR_CTR1_1000	0.31239602	0.896025151	-0.813769363	5.06E-21
SVR_CTR1_250	0.283867927	0.906474413	-0.725442948	5.96E-15
SVR_CTR1_500	0.288746646	0.907812295	-0.759542371	5.54E-17
SVR_CTR1_55	0.210776307	0.927963843	-0.722527083	8.61E-15
SVR_CTR2	0.328040711	0.904836275	-0.749559583	2.35E-16
SVR_CTR2_10	0.072550415	0.95671404	-0.401781918	0.00015177
SVR_CTR2_1000	0.294807627	0.898299147	-0.687366115	2.63E-12
SVR_CTR2_250	0.25805736	0.901661149	-0.760345667	4.38E-16
SVR_CTR2_500	0.28915211	0.924286006	-0.722324592	8.83E-15
SVR_CTR2_55	0.17142137	0.941638918	-0.647949782	2.70E-11
SVR_DEG	0.348704324	0.885688225	-0.79698289	1.21E-19
SVR_DEG_10	0.209667666	0.936247036	-0.66854308	3.68E-12
SVR_DEG_1000	0.319370711	0.896286971	-0.766690291	1.88E-17
SVR_DEG_250	0.299125501	0.904835421	-0.786736863	7.32E-19
SVR_DEG_500	0.310820283	0.896687404	-0.804961021	2.79E-20
SVR_DEG_55	0.258337555	0.919247829	-0.696770274	1.83E-13
SVR_DEG_HIST	0.333098198	0.899440672	-0.809415815	1.19E-20
SVR_HIST_ONLY	0.229659101	0.930236405	-0.784894199	1.00E-18
SVR_Limma	0.316195531	0.895748638	-0.754621849	1.14E-16

SVR_MRMR	0.326862919	0.905420517	-0.76820897	1.49E-17
SVR_None	0.345708834	0.881855077	-0.780074922	2.24E-18
SVR_RAN_CTR	0.008792232	0.967722401	-0.123855092	0.36311230

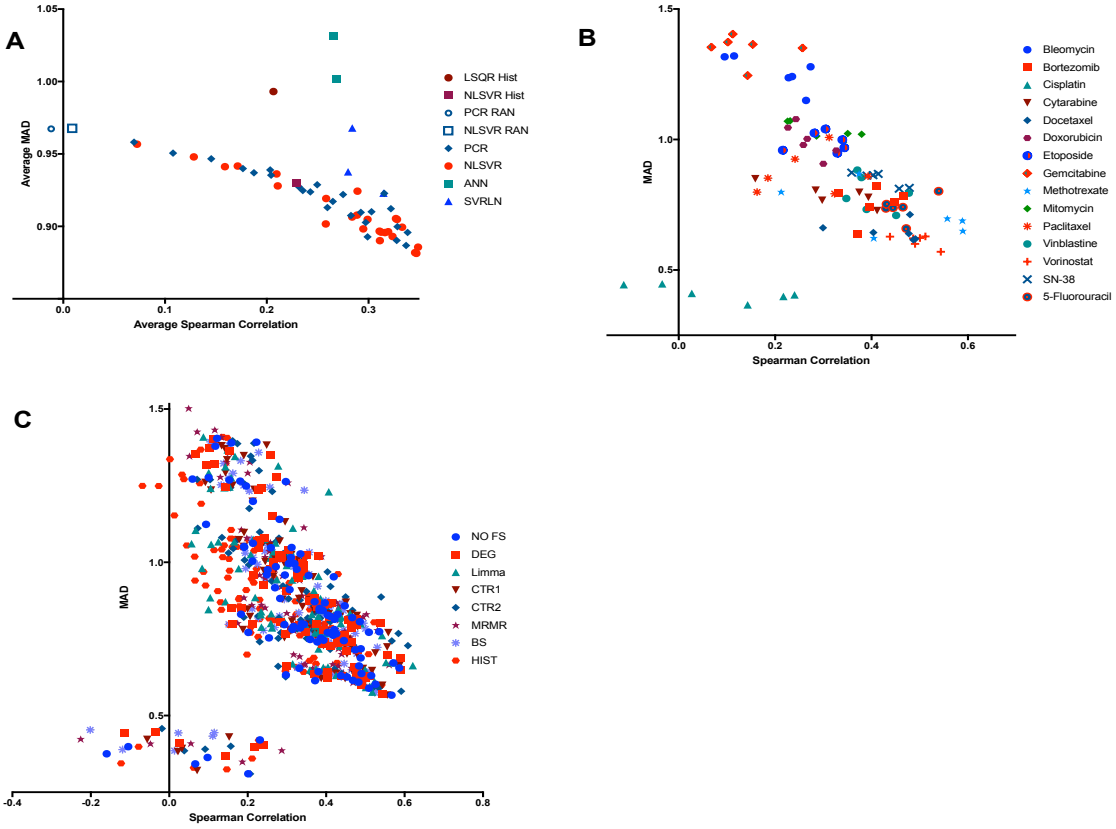


Figure A.1 A. Average Spearman correlations vs Average MAD scores for various models. B. Average Spearman correlations vs average MAD scores for NLSVR DEG datasets. C. Average Spearman correlation vs average MAD scores by feature selection in NLSVR models.

Additional Feature Selection Methods:

The relationship between histotype and drug response dictated both model performance and feature selection. Selected features tended to reflect this relationship, presumably masking features that could account for subtler differences in drug response. Therefore, an approach that could balance variability due to histotype with variability more specific to drug response might allow for better model performance. This suggested that the problem could be addressed either by muting features selected

purely based on histotype or adding additional histotype specific covariates to account for histotype variability while allowing genomic covariates to account for drug variability independent of histotype.

In general, histotypes were not equally represented in each dataset. This becomes problematic when two or more histotypes that are disproportionally represented in the training data have dramatic differences in drug response which results in the selection of genetic features which might show differential expression as a result of histotype rather than drug response. Presumably this effect could be mitigated by using a more uniform representation of histotypes during feature selection. However, by specifically curating datasets such that each histotype had equal representation introduced the possibility that selected features would be biased towards the makeup of the individual choice of dataset. We addressed these issues by constructing 50 subsets of the data containing one sample from a histotype and then taking genes that were significantly correlated ($P<0.05$) in at least half the subsets (Boot Strapped by histotype: “BS Hist”). As expected features were dramatically reduced ranging from 99.99 to 88.65 percent with an average decrease of 98.6 percent. This method resulted in the lowest average Spearman correlation of all feature selection methods, 0.304 for NLSVR (Sup. Figure 2 A) and 0.216 for PCR (Sup. Figure 2 B). S_c also was significantly higher than DEG and no feature selection, consistent with the observation that for a given set of features a decreased ability to discriminate samples by histotype results in a decrease in performance. Analysis of the selected genes showed that genes which met the criteria for selection still showed significant variability according to histotype, thus again these models defaulted to fitting an average histotype IC50.

Alternatively, by adding an additional histotype specific variable we attempted to break the regression into two different groups of terms:

$$\boldsymbol{Y}(h, \boldsymbol{G}) = f(h) + f(\boldsymbol{G})$$

where $\boldsymbol{Y}$ is the drug response, h is the histotype, and $\boldsymbol{G}$ is the gene expression matrix. This was accomplished by combining the feature matrix for the histotype models and the DEG expression values scaled from zero to one to avoid scaling issues. The motivation behind this was to place the variability strictly due to histotype on the histotype term allowing non-redundant variability independent of histotype to play a more active role in predicting drug response. However, the addition of histotype specific variables resulted in an overall decrease in average performance, 0.319 for NLSVR and 0.306 for PCR (Sup figure 2 A and B). The failure of this method further suggests that sources of genomic variability that are small compared to the variability resulting from histotype are effectively treated as noise which both NLSVR and PCR aim to minimize by regularization in NLSVR or the elimination of components in PCR.

REFERENCES

1. Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nature biotechnology. 2014;32(12):1202-12.

2. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23(19):2507-17.

APPENDIX B: SUPPLEMENTARY MATERIAL FOR CHAPTER 4

Combining DEGS:

The best performing models without the inclusion of feature selection utilized 24 hour high dosed drug gene expression (C_{high}) performing marginally better than similar models using perturbed gene expression (1.2%). While, this was not a significant difference a possible explanation for the increase performance is that a signature from this dataset combined aspects from both the basal gene expression and gene perturbations. This presented an opportunity to select DEGS from two different data sets, basal gene expression and gene perturbations , and apply them to the gene expression of a single dataset hopefully capturing a more predictive signature.

The application of DEGs from both the basal data and the perturbed data outperformed models using a single set of DEGs from either the basal data or perturbed data as well as DEGs selected within the C_{high} dataset. The combination of 0nM and ΔC_{high} DEGs resulted in an average spearman correlation of 0.515 (Not including AZA) compared to 0.484 using ΔC_{high} DEGS, 0.396 using basal DEGs, and 0.476 using DEGs selected within (C_{high}) the data; however, with the exception of basal data ($p<0.0001$), there performance of the other DEGs did not prove to be significant by a paired t-test. Nonetheless, for several drugs including bortezomib, doxorubicin, geldanamycin, paclitaxel, sorafenib, and sunitinib proved to outperform using any other DEG set most notably for bortezomib and sunitinib (Figure B.1).

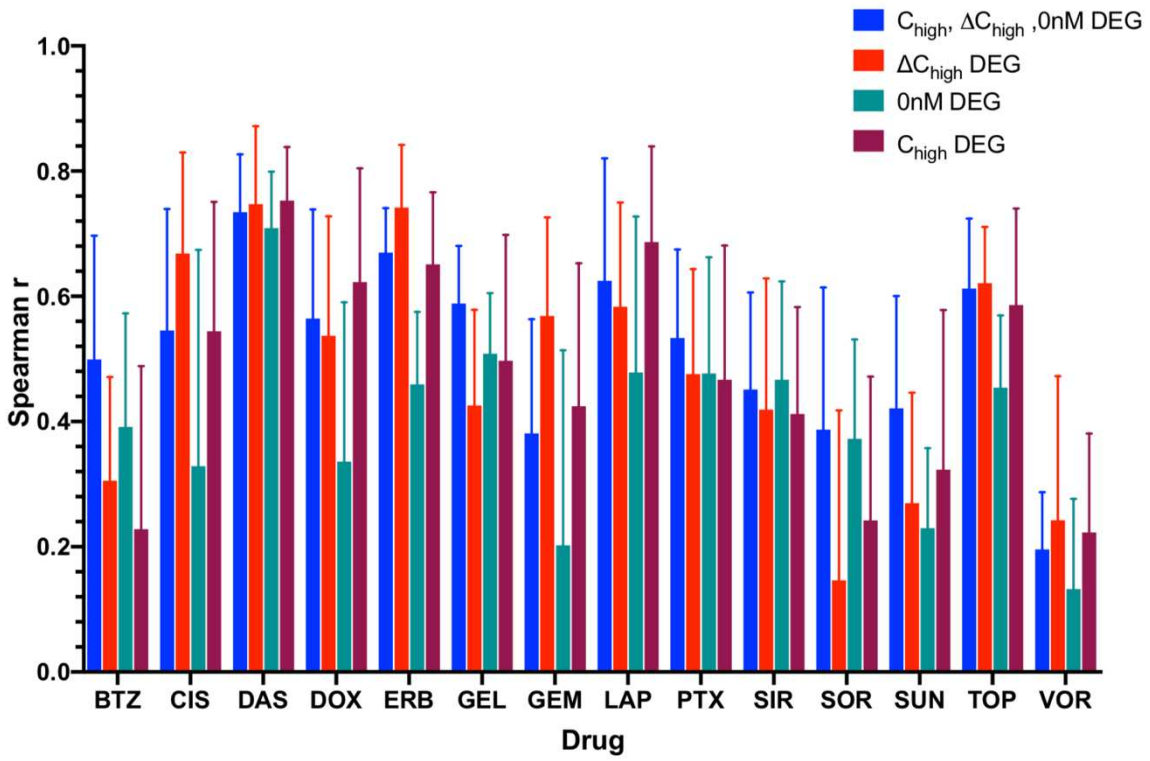


Figure B.1. Models Built with DEGS from ΔC_{high} and 0nM using C_{high} gene expression compared with other DEG Models

Perturbation as Indicators of Drug response:

One of the questions we sought to answer is how indicative is drug induced gene dynamics, specifically without the additional knowledge of drug mechanism or specific routes of resistance. How well can changes in gene expression alone predict drug response? Are predictive genes subject to greater dynamic perturbation? In order to ascertain the influence of we looked at the profile of DEGs, DEGs chosen from 0nM gene expression, and a 100 genes with the greatest magnitude of change in high dosed perturbation data. With the exception of sunitinib and lapatinib (%75,%83) the average magnitude of relative change between models using perturbed DEGs to basal DEGs was on average larger by %111 (SD=0.195) and a maximum of %160. For the 100

genes with the maximum magnitude of change were on average %247 (STDEV 0.45) larger ranging from a minimum of %181 and maximum %326 (Figure B.2 A). When looking at performance despite the fact that drug treatment resulted in a large magnitude of change very few of these genes are predictive of drug response resulting in a correlation of only 0.087 significantly lower than using random genes by %75. As referenced earlier, DEGs from high dosed perturbation performed %46.5 better then DEGs chosen from basal data (Figure B.2 B).

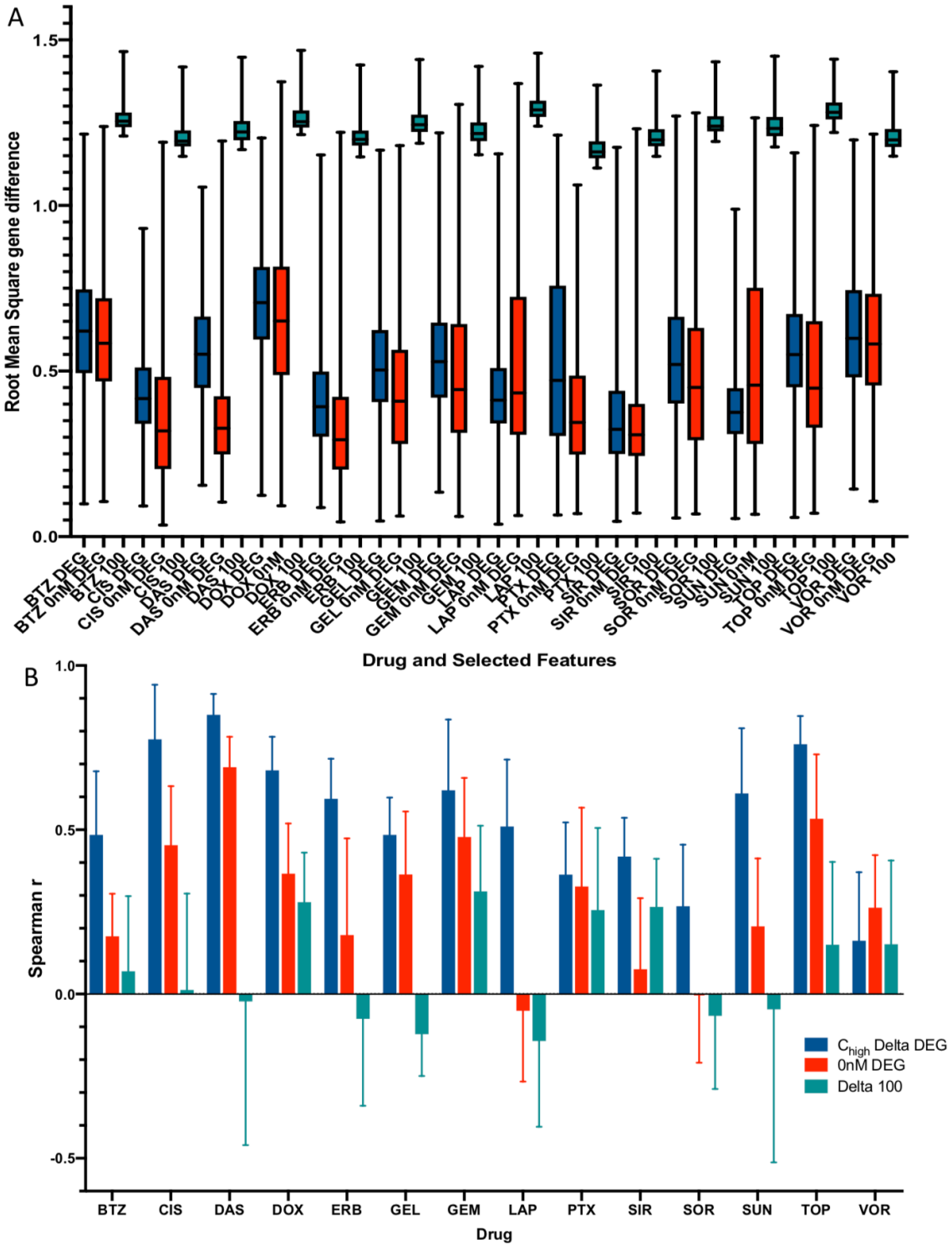


Figure B.2: A. Distribution of the root mean squared magnitude of dynamic changes in gene expression after application of a high dose of drug in different gene sets. B. The performance using the respective DEGs for each drug.

Drug Related Genes and Networks

In order to determine if gene expression of drug related genes at 0nM or ΔC_{high} were predictive the following genes were identified through the iLINCS web tool (1) and STITCH web tool (2) each drug

Table B.1. Gene-drug interactions reported in the iLINCS and Stitch databases

Drug	Genes
Bortezomib	PSMB1, PSMB2, PSMB6, TP53, MAPK8, CASP3, CYCS, JUN
Cisplatin	XIAP, TRAF1,TRAF2,LCK
Dasatinib	ABL1, FYN, LCK, KIT, YES1, EPHA2, LYN, PDGFRB, BCR, SRC, HCK
Doxorubicin	ABCB1, TOP2A, TP53, ABCG2, EFGR, CASP3, AKT1, MYC, ABCC1, ATM
Erlotinib	EFGR, EGF SLK, STAT3, CYP3A4, PTPN9 NOX4, AKT1, GAK,HGF
Geldanamycin	HSP90AA1, HSP90AB1, ERBB2, AKT1, HSP90B1, RAF1, HSPA4, DNAJB1, TRAP1, TP53
Gemcitabine	RRM1
Lapatinib	ERBB2, EGFR, ERBB3, ERBB4, ESR1, AKT1, VEGFA, ABBC10,MCL1, TP53
Paclitaxel	AURKB, CDK1, MAPRE3 TNF, VEGA, MMP2, CYCS, CAMKMT, CDH1, WNTSA, TUBB, NR1I2
Sirolimus	MTOR, FKBP1A
Sorafenib	RET, BRAF, FLT3, KDR, RAF1, FLT1, FLT4, PDGFRA, CSF1R, AXL, RPS6KB1
Sunitinib	PDGFRB, KIT, FLT3, KDR, FLT1, FTL4, PDGFRA, CSF1R, AXL, RPS6KB1
Topotecan	TOP1
Vorinostat	HDAC1,HDAC2,HDAC3,HDAC6, HDAC8, TP53, K2AFX, BCL2L1, HDAC7, HSP90AA1

Compared to DEG models for both ΔC_{high} and 0nM DEGS the performance of models using the genes above was significantly worse (figure B.3).

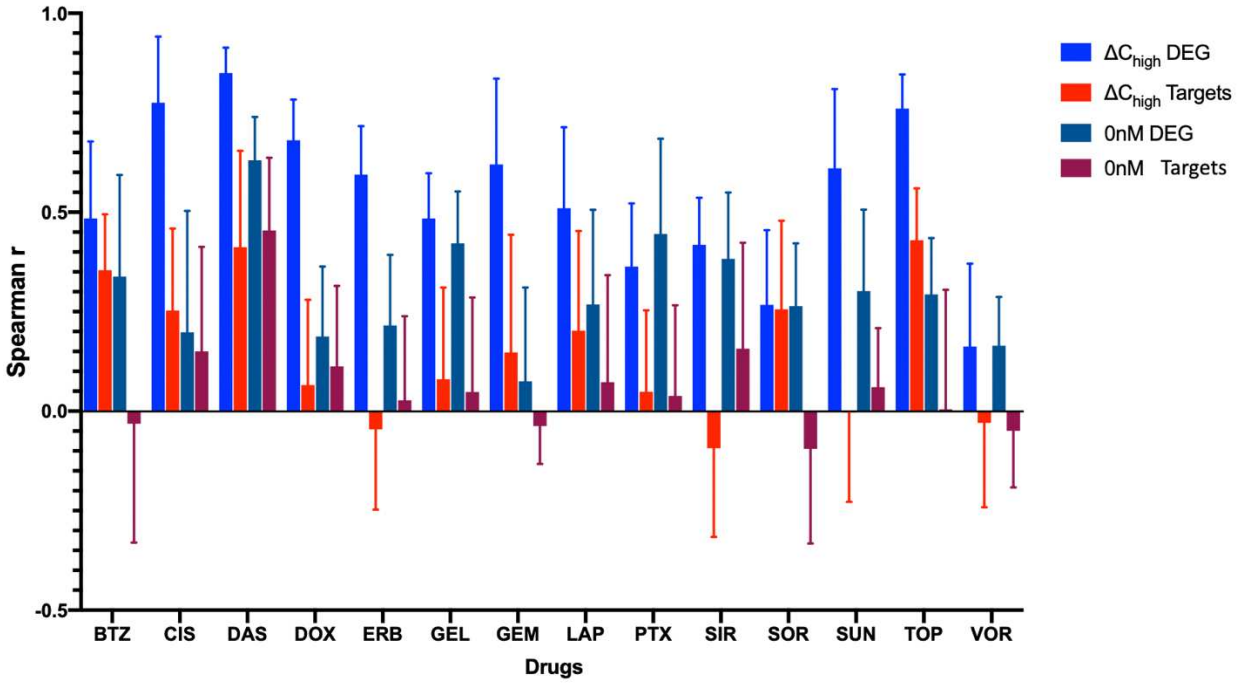


Figure B.3: Models using genes defined in Table S1 compared to Models using DEGs from different gene expression profiles.

Additional Genes From Network Analysis

Table B.2. Additional genes with top clique participation for ΔC_{high} DEGs

Bortezomib	Developmentally regulated GTP binding protein 2 (DRG2)(3) MicroRNA 939 MIR939 (4) Cell division cycle 37 (CDC37) (5, 6) Eukaryotic translation initiation factor 3 subunit G (EIF3G) (7) Squamous cell carcinoma antigen recognized by T-cells 1 (SART1)
Cisplatin	SWI/SNF related, matrix associated, actin dependent regulator of chromatin subfamily c member 1 (SMARCC1) (8, 9) Drosha ribonuclease III (DROSHA) (10, 11) Nuclear receptor coactivator (NCOA1, SRC1) (12) ADAM metallopeptidase domain 10 (ADAM10) (13) Tetratricopeptide repeat domain 28 (TTC28) (14)
Doxorubicin	Heterogenous nuclear ribonucleoprotein A0 (HNRNPA0) (15) Eukaryotic translation initiation factor 4E (EIF4E) (16-18) Cytokine receptor like factor3 (CRLF3, p48.2) (19) MOB family member 4, phocein (MOB4) (20)
Erlotinib	survivin (BIRC5) (21-23) H2A histone family member X(H2AX) (24)

	BUB2 mitotic checkpoint serine/threonine kinase B(BUB1B) (25, 26)
Geldanamycin	Proteasome activator subunit 3 (PSME3) (27, 28) Proliferation-associated 2G4 (PA2G4)(EBP1) (29, 30) RAN binding protein 1(RANBP1)(31, 32) Cell division cycle 25A (CDC25A) (33-35) PBX homeobox interacting protein (PBXIP1,HPIP)(36, 37)
Gemcitabine	Peroxiredoxin 1 (PRDX1) (38, 39) BUD31 homolog (BUD31) (40) DEAD-box helicase 27 (DDX27)(41) Small ubiquitin-like modifier 3 (SUMO3) (42) Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein epsilon (YWHAE, 14-3-3 epsilon) (43, 44)
Lapatinib	Tumor susceptibility 101 (TSG101) (45, 46) Chromosome segregation 1 like (CSE1L) (47, 48) TIMELESS interacting protein (TIPIN) (49) NME/NM23 nucleoside diphosphate kinase1 (NME1) (50-52) GINS complex subunit 2 (GINS2)(53-55)
Paclitaxel	RNA binding motif protein 5(RBM5,LUCA15)(56-58) Topoisomerase DNA II binding protein (TOPBP1) (59-61) BUB3, mitotic checkpoint protein (BUB3)(62) SET domain containing 1B (SETD1B) (63) Histone cluster 4 H4(HIST4H4) (64, 65)
Sirolimus	Ubiquitin associated protein 2 like (UBA2P2L) (66, 67) RE1 Silencing transcription factor (REST) (68, 69)
Sorafenib	Proteasome activator subunit 3(PSME3)(27, 70) Heterogeneous nuclear ribonucleoprotein A/B (HNRNPAB)(71) Chaperonin containing TCP1 subunit 2 (CCT2) (72, 73)
Sunitinib	Eukaryotic translation initiation factor 2 subunit alpha (EIF2S1)(74, 75) Calcyclin binding protein (CACYBP) (76, 77) G1 to S phase transition 1 (GSPT1) (78, 79) DExD-box helicase 39A (DDX39A) (80, 81) Proliferation-associated 2G4 (PA2G4) (30, 82)
Topotecan	RAD23 homolog B. nucleotide excision repair protein (RAD23B) (83) Cullin 4A (CUL4A) (84, 85) BUD31 homolog BUD31 (40) Complement C1q binding protein C1qBP (86, 87) Methionyl aminopeptidase 2 (METAP2) (88-90)
Vorinostat	Eukaryotic translation elongation factor 1 epsilon (AIMP3) (91) NOP16 nucleolar protein (NOP16, HSPC111) (92) Proliferation-associated 2G4 (PA2G4,EBP1) (29, 30, 93, 94)

Table B.3. Genes with top Clique Participation for 0nM DEGs

Bortezomib	Protein phosphatase 2 scaffold subunit Abeta (PPP2R1B) (95, 96) MAD2 mitotic arrest deficient-like MAD2L1 (97-99) Polo like kinase 4 (PLK4,SAK) (100, 101)
Cisplatin	Epithelial cell adhesion molecule (EPCAM) (102-104) Suppression of tumorigenicity 14 (ST14, TADG15) (105) Zinc-Finger protein 165(ZNF165)(106) Serine peptidase inhibitor, Kunitz type 2 (SPINT2)(107, 108) F11 receptor (F11R, JAM-A) (109, 110)
Doxorubicin	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenenase activation protein zeta (YWHAZ, 14-3-3 Zeta) (111-113) Centrosomal protein 57 (CEP57)(114) Fli-1 proto-oncogene, ETS transcription factor (FLI1) (115, 116) Ribosomal protein S4, X-linked(RPS4X) (117, 118)
Erlotinib	Myosin light chain kinase (MYLK) (119, 120) N-myristoyltransferase 2(NMT2) (121) Related RAS viral (r-ras) oncogene homolog 2(RRAS2,TC21)(122-124) NOP14 nucleolar protein (NOP14) (125)
Geldanamycin	GLI pathogenesis related 1(GLIPR1) (126, 127) NUAK family kinase(NUAK1) (128-130) Microtubule associated protein 1B (MAP1B) (131) microRNA 22(MIR22) (132, 133) FOS like 2, AP-1 transcription factor subunit (FOSL2, FRA2) (134, 135)
Gemcitabine	Cadherin 1 (CDH1) (136, 137) CCCTC-binding factor (CTCF) (138, 139) GINS complex subunit 3(GINS3,PSF3) (140)
Lapatinib	RAP1 GTPase activating protein (RAP1GAP) (141-143) Transforming growth factor alpha (TGFA) (144, 145) Lysophosphatidic acid receptor 2 (LPAR2) (146, 147) Tissue factor pathway inhibitor 2 (TFPI2)(148) F11 receptor(F11R,JAM-A) (109, 110)
Paclitaxel	Serine and arginine rich splicing factor 2 (SRSF2) (149, 150) Cell division cycle 25A (CDC25A) (34, 35, 151)
Sirolimus	Nucleolin (NCL) (152) Serine and arginine rich splicing factor 2 (SRSF2) (149, 150) microRNA 1244-1(MIR1244-1) (153) polo like kinase 4 (PLK4,SAK) (154, 155) checkpoint kinase 1(CHEK1) (156, 157)
Sorafenib	Tropomyosin 1 (alpha) (TPM1) (158, 159) CD24 molecule (CD24)(103, 160, 161) E74 like ETS transcription factor 3 (ELF3) (162, 163)

	Caspase recruitment domain family member 10 (CARD10,CARMA3)(164, 165)
Sunitinib	Cell division cycle 25A (CDC25A) (35, 151, 166) Cyclin dependent kinase 8 (CDK8) (167, 168) Squamous cell carcinoma antigen recognized by T-cells 3(SART3)(169, 170) NOP14 nucleolar protein (NOP14)(125) PLAG1 like zinc finger 2(PLAGL2) (171-173)
Topotecan	Cadherin 1 (CDH1)(122, 137) E74 like ETS transcription factor 3 (ELF3)(162, 163, 174) P21 (RAC1) activated kinase 6 (PAK6)(175, 176) Microtubule associated protein 7(MAP7) (177)
Vorinostat	Plasaminogen activator, urokinase (PLAU) (178) Microtubule associated monooxygenase, calponin and LIM domain containing 2 (MICAL2)(179) Transforming growth factor beta2 (TGFB2) (180, 181) Annexin II (ANXA2) (82, 182) Neuropilin 1 (NRP1)(183)

REFERENCES

1. Pilarczyk M, Najafabadi MF, Kouril M, Vasiliauskas J, Niu W, Shamsaei B, et al. Connecting omics signatures of diseases, drugs, and mechanisms of actions with iLINCS. bioRxiv. 2019:826271.

2. Szklarczyk D, Santos A, von Mering C, Jensen LJ, Bork P, Kuhn M. STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. Nucleic Acids Res. 2016;44(D1):D380-4.

3. Xu C, Li H, Zhang L, Jia T, Duan L, Lu C. MicroRNA-1915-3p prevents the apoptosis of lung cancer cells by downregulating DRG2 and PBX2. Mol Med Rep. 2016;13(1):505-12.

4. Ying X, Li-ya Q, Feng Z, Yin W, Ji-hong L. MiR-939 promotes the proliferation of human ovarian cancer cells by repressing APC2 expression. Biomed Pharmacother. 2015;71:64-9.

5. Wu F, Peacock SO, Rao S, Lemmon SK, Burnstein KL. Novel interaction between the co-chaperone Cdc37 and Rho GTPase exchange factor Vav3 promotes androgen receptor activity and prostate cancer growth. J Biol Chem. 2013;288(8):5463-74.

6. Basso AD, Solit DB, Chiosis G, Giri B, Tsichlis P, Rosen N. Akt forms an intracellular complex with heat shock protein 90 (Hsp90) and Cdc37 and is destabilized by inhibitors of Hsp90 function. J Biol Chem. 2002;277(42):39858-66.

7. Kim JT, Lee SJ, Kim BY, Lee CH, Yeom YI, Choe YK, et al. Caspase-mediated cleavage and DNase activity of the translation initiation factor 3, subunit G (eIF3g). FEBS Lett. 2013;587(22):3668-74.

8. Heebøll S, Borre M, Ottosen PD, Andersen CL, Mansilla F, Dyrskjøt L, et al. SMARCC1 expression is upregulated in prostate cancer and positively correlated with tumour recurrence and dedifferentiation. Histol Histopathol. 2008;23(9):1069-76.

9. Andersen CL, Christensen LL, Thorsen K, Schepeler T, Sørensen FB, Verspaget HW, et al. Dysregulation of the transcription factors SOX4, CBFB and SMARCC1 correlates with outcome of colorectal cancer. Br J Cancer. 2009;100(3):511-23.

10. Zhang H, Hou Y, Xu L, Zeng Z, Wen S, Du YE, et al. Cytoplasmic Drosha Is Aberrant in Precancerous Lesions of Gastric Carcinoma and Its Loss Predicts Worse Outcome for Gastric Cancer Patients. Dig Dis Sci. 2016;61(4):1080-90.

11. Zhou J, Cai J, Huang Z, Ding H, Wang J, Jia J, et al. Proteomic identification of target proteins following Drosha knockdown in cervical cancer. Oncol Rep. 2013;30(5):2229-37.

12. Wang L, Yu Y, Chow DC, Yan F, Hsu CC, Stossi F, et al. Characterization of a Steroid Receptor Coactivator Small Molecule Stimulator that Overstimulates Cancer Cells and Leads to Cell Stress and Death. Cancer Cell. 2015;28(2):240-52.

13. Liu S, Zhang W, Liu K, Ji B, Wang G. Silencing ADAM10 inhibits the in vitro and in vivo growth of hepatocellular carcinoma cancer cells. Mol Med Rep. 2015;11(1):597-602.

14. Izumiyama T, Minoshima S, Yoshida T, Shimizu N. A novel big protein TPRBK possessing 25 units of TPR motif is essential for the progress of mitosis and cytokinesis. Gene. 2012;511(2):202-17.

15. Cannell IG, Merrick KA, Morandell S, Zhu CQ, Braun CJ, Grant RA, et al. A Pleiotropic RNA-Binding Protein Controls Distinct Cell Cycle Checkpoints to Drive Resistance of p53-Defective Tumors to Chemotherapy. Cancer Cell. 2015;28(5):623-37.

16. Hoang B, Benavides A, Shi Y, Yang Y, Frost P, Gera J, et al. The PP242 mammalian target of rapamycin (mTOR) inhibitor activates extracellular signal-regulated kinase (ERK) in multiple myeloma cells via a target of rapamycin complex 1 (TORC1)/eukaryotic translation initiation factor 4E (eIF-4E)/RAF pathway and activation is a mechanism of resistance. J Biol Chem. 2012;287(26):21796-805.

17. Wheater MJ, Johnson PW, Blaydes JP. The role of MNK proteins and eIF4E phosphorylation in breast cancer cell proliferation and survival. Cancer Biol Ther. 2010;10(7):728-35.

18. Muta D, Makino K, Nakamura H, Yano S, Kudo M, Kuratsu J. Inhibition of eIF4E phosphorylation reduces cell growth and proliferation in primary central nervous system lymphoma cells. J Neurooncol. 2011;101(1):33-9.

19. Yang F, Xu YP, Li J, Duan SS, Fu YJ, Zhang Y, et al. Cloning and characterization of a novel intracellular protein p48.2 that negatively regulates cell cycle progression. Int J Biochem Cell Biol. 2009;41(11):2240-50.

20. Tang F, Zhang L, Xue G, Hynx D, Wang Y, Cron PD, et al. hMOB3 modulates MST1 apoptotic signaling and supports tumor growth in glioblastoma multiforme. Cancer Res. 2014;74(14):3779-89.

21. Rödel C, Haas J, Groth A, Grabenbauer GG, Sauer R, Rödel F. Spontaneous and radiation-induced apoptosis in colorectal carcinoma cells with different intrinsic radiosensitivities: survivin as a radioresistance factor. Int J Radiat Oncol Biol Phys. 2003;55(5):1341-7.

22. Zaffaroni N, Pennati M, Colella G, Perego P, Supino R, Gatti L, et al. Expression of the anti-apoptotic gene survivin correlates with taxol resistance in human ovarian cancer. Cell Mol Life Sci. 2002;59(8):1406-12.

23. Lee JP, Chang KH, Han JH, Ryu HS. Survivin, a novel anti-apoptosis inhibitor, expression in uterine cervical cancer and relationship with prognostic factors. Int J Gynecol Cancer. 2005;15(1):113-9.

24. Strasberg Rieber M, Viola-Rhenals M, Rieber M. Attenuation of genotoxicity under adhesion-restrictive conditions through modulation of p53, gamma H2AX and nuclear DNA organization. Apoptosis. 2007;12(2):449-58.

25. Ikawa-Yoshida A, Ando K, Oki E, Saeki H, Kumashiro R, Taketani K, et al. Contribution of BubR1 to oxidative stress-induced aneuploidy in p53-deficient cells. Cancer Med. 2013;2(4):447-56.

26. Fragoso MC, Almeida MQ, Mazzuco TL, Mariani BM, Brito LP, Gonçalves TC, et al. Combined expression of BUB1B, DLGAP5, and PINK1 as predictors of poor outcome in adrenocortical tumors: validation in a Brazilian cohort of adult and pediatric patients. Eur J Endocrinol. 2012;166(1):61-7.

27. Li J, Feng X, Sun C, Zeng X, Xie L, Xu H, et al. Associations between proteasomal activator PA28γ and outcome of oral squamous cell carcinoma: Evidence from cohort studies and functional analyses. EBioMedicine. 2015;2(8):851-8.

28. Xu X, Liu D, Ji N, Li T, Li L, Jiang L, et al. A novel transcript variant of proteasome activator 28γ: Identification and function in oral cancer cells. Int J Oncol. 2015;47(1):188-94.

29. Zhang F, Liu Y, Wang Z, Sun X, Yuan J, Wang T, et al. A novel Anxa2-interacting protein Ebp1 inhibits cancer proliferation and invasion by suppressing Anxa2 protein level. Molecular and Cellular Endocrinology. 2015;411:75-85.

30. Zhang Y, Akinmade D, Hamburger AW. Inhibition of heregulin mediated MCF-7 breast cancer cell growth by the ErbB3 binding protein EBP1. Cancer Lett. 2008;265(2):298-306.

31. Amato R, Scumaci D, D'Antona L, Iuliano R, Menniti M, Di Sanzo M, et al. Sgk1 enhances RANBP1 transcript levels and decreases taxol sensitivity in RKO colon carcinoma cells. Oncogene. 2013;32(38):4572-8.

32. Rensen WM, Roscioli E, Tedeschi A, Mangiacasale R, Ciciarello M, Di Gioia SA, et al. RanBP1 downregulation sensitizes cancer cells to taxol in a caspase-3-dependent manner. Oncogene. 2009;28(15):1748-58.

33. Lin TC, Lin PL, Cheng YW, Wu TC, Chou MC, Chen CY, et al. MicroRNA-184 Deregulated by the MicroRNA-21 Promotes Tumor Malignancy and Poor Outcomes in Non-small Cell Lung Cancer via Targeting CDC25A and c-Myc. Ann Surg Oncol. 2015;22 Suppl 3:S1532-9.

34. Li N, Zhong X, Lin X, Guo J, Zou L, Tanyi JL, et al. Lin-28 homologue A (LIN28A) promotes cell cycle progression via regulation of cyclin-dependent kinase 2 (CDK2),

cyclin D1 (CCND1), and cell division cycle 25 homolog A (CDC25A) expression in cancer. J Biol Chem. 2012;287(21):17386-97.

35. Chiu Y-T, Han H-Y, Leung SC-L, Yuen H-F, Chau C-W, Guo Z, et al. CDC25A Functions as a Novel Ar Corepressor in Prostate Cancer Cells. Journal of Molecular Biology. 2009;385(2):446-56.

36. van Vuurden DG, Aronica E, Hulleman E, Wedekind LE, Biesmans D, Malekzadeh A, et al. Pre-B-cell leukemia homeobox interacting protein 1 is overexpressed in astrocytoma and promotes tumor cell growth and migration. Neuro Oncol. 2014;16(7):946-59.

37. Okada S, Irié T, Tanaka J, Yasuhara R, Yamamoto G, Isobe T, et al. Potential role of hematopoietic pre-B-cell leukemia transcription factor-interacting protein in oral carcinogenesis. J Oral Pathol Med. 2015;44(2):115-25.

38. Song IS, Kim SU, Oh NS, Kim J, Yu DY, Huang SM, et al. Peroxiredoxin I contributes to TRAIL resistance through suppression of redox-sensitive caspase activation in human hepatoma cells. Carcinogenesis. 2009;30(7):1106-14.

39. Dey KK, Pal I, Bharti R, Dey G, Kumar BN, Rajput S, et al. Identification of RAB2A and PRDX1 as the potential biomarkers for oral squamous cell carcinoma using mass spectrometry-based comparative proteomic approach. Tumour Biol. 2015;36(12):9829-37.

40. Xu W, Huang H, Yu L, Cao L. Meta-analysis of gene expression profiles indicates genes in spliceosome pathway are up-regulated in hepatocellular carcinoma (HCC). Med Oncol. 2015;32(4):96.

41. Zhou J, Yong WP, Yap CS, Vijayaraghavan A, Sinha RA, Singh BK, et al. An integrative approach identified genes associated with drug response in gastric cancer. Carcinogenesis. 2015;36(4):441-51.

42. Liu J, Sha M, Wang Q, Ma Y, Geng X, Gao Y, et al. Small ubiquitin-related modifier 2/3 interacts with p65 and stabilizes it in the cytoplasm in HBV-associated hepatocellular carcinoma. BMC Cancer. 2015;15:675.

43. Ko BS, Chang TC, Hsu C, Chen YC, Shen TL, Chen SC, et al. Overexpression of 14-3-3ε predicts tumour metastasis and poor survival in hepatocellular carcinoma. Histopathology. 2011;58(5):705-11.

44. Konishi H, Nakagawa T, Harano T, Mizuno K, Saito H, Masuda A, et al. Identification of frequent G(2) checkpoint impairment and a homozygous deletion of 14-3-3epsilon at 17p13.3 in small cell lung cancers. Cancer Res. 2002;62(1):271-6.

45. Zhang Y, Song M, Cui ZS, Li CY, Xue XX, Yu M, et al. Down-regulation of TSG101 by small interfering RNA inhibits the proliferation of breast cancer cells through the MAPK/ERK signal pathway. Histol Histopathol. 2011;26(1):87-94.

46. Gu RJ, Wang SC, Sun G, Zhuang BW, Liu DL. [Expression and significance of tumor susceptibility gene 101 in hepatocellular carcinoma tissues]. Xi Bao Yu Fen Zi Mian Yi Xue Za Zhi. 2012;28(7):738-40.

47. Yuksel UM, Turker I, Dilek G, Dogan L, Gulcelik MA, Oksuzoglu B. Does CSE1L Overexpression Affect Distant Metastasis Development in Breast Cancer? Oncol Res Treat. 2015;38(9):431-4.

48. Shiraki K, Fujikawa K, Sugimoto K, Ito T, Yamanaka T, Suzuki M, et al. Cellular apoptosis susceptibility protein and proliferation in human hepatocellular carcinoma. Int J Mol Med. 2006;18(1):77-81.

49. Baldeyron C, Brisson A, Tesson B, Némati F, Koundrioukoff S, Saliba E, et al. TIPIN depletion leads to apoptosis in breast cancer cells. Mol Oncol. 2015;9(8):1580-98.

50. Tomita M, Ayabe T, Matsuzaki Y, Edagawa M, Maeda M, Shimizu T, et al. Expression of nm23-H1 gene product in esophageal squamous cell carcinoma and its association with vessel invasion and survival. BMC Cancer. 2001;1:3.

51. Galani E, Sgouros J, Petropoulou C, Janinis J, Aravantinos G, Dionysiou-Asteriou D, et al. Correlation of MDR-1, nm23-H1 and H Sema E gene expression with histopathological findings and clinical outcome in ovarian and breast cancer patients. Anticancer Res. 2002;22(4):2275-80.

52. Kushlinskii NE, Delektorskaya VV, Mochal'nikova VV, Sini L, Yurchenko AA, Ryabov AB, et al. Analysis of NM23 protein and components of plasminogen activation system in tumors of patients with stomach cancer with consideration for disease clinical picture and morphology. Bull Exp Biol Med. 2008;146(6):786-90.

53. Rantala JK, Edgren H, Lehtinen L, Wolf M, Kleivi K, Vollan HKM, et al. Integrative functional genomics analysis of sustained polyploidy phenotypes in breast cancer cells identifies an oncogenic profile for GINS2. Neoplasia. 2010;12(11):877-88.

54. Gao Y, Wang S, Liu B, Zhong L. Roles of GINS2 in K562 human chronic myelogenous leukemia and NB4 acute promyelocytic leukemia cells. Int J Mol Med. 2013;31(6):1402-10.

55. Zheng M, Zhou Y, Yang X, Tang J, Wei D, Zhang Y, et al. High GINS2 transcript level predicts poor prognosis and correlates with high histological grade and endocrine therapy resistance through mammary cancer stem cells in breast cancer patients. Breast Cancer Research and Treatment. 2014;148(2):423-36.

56. Rintala-Maki ND, Abrasonis V, Burd M, Sutherland LC. Genetic instability of RBM5/LUCA-15/H37 in MCF-7 breast carcinoma sublines may affect susceptibility to apoptosis. Cell Biochem Funct. 2004;22(5):307-13.

57. Kobayashi T, Ishida J, Musashi M, Ota S, Yoshida T, Shimizu Y, et al. p53 transactivation is involved in the antiproliferative activity of the putative tumor suppressor RBM5. Int J Cancer. 2011;128(2):304-18.

58. Oh JJ, Taschereau EO, Koegel AK, Ginther CL, Rotow JK, Isfahani KZ, et al. RBM5/H37 tumor suppressor, located at the lung cancer hot spot 3p21.3, alters expression of genes involved in metastasis. Lung Cancer. 2010;70(3):253-62.

59. Liu K, Bellam N, Lin HY, Wang B, Stockard CR, Grizzle WE, et al. Regulation of p53 by TopBP1: a potential mechanism for p53 inactivation in cancer. Mol Cell Biol. 2009;29(10):2673-93.

60. Yamane K, Chen J, Kinsella TJ. Both DNA topoisomerase II-binding protein 1 and BRCA1 regulate the G2-M cell cycle checkpoint. Cancer Res. 2003;63(12):3049-53.

61. Forma E, Wójcik-Krowiranda K, Jóźwiak P, Szymczyk A, Bieńkiewicz A, Bryś M, et al. Topoisomerase IIβ binding protein 1 c.*229C>T (rs115160714) gene polymorphism and endometrial cancer risk. Pathol Oncol Res. 2014;20(3):597-602.

62. Yoon YM, Baek KH, Jeong SJ, Shin HJ, Ha GH, Jeon AH, et al. WD repeat-containing mitotic checkpoint proteins act as transcriptional repressors during interphase. FEBS Lett. 2004;575(1-3):23-9.

63. Choi YJ, Oh HR, Choi MR, Gwak M, An CH, Chung YJ, et al. Frameshift mutation of a histone methylation-related gene SETD1B and its regional heterogeneity in gastric and colorectal cancers with high microsatellite instability. Hum Pathol. 2014;45(8):1674-81.

64. Boix-Chornet M, Fraga MF, Villar-Garea A, Caballero R, Espada J, Nuñez A, et al. Release of hypoacetylated and trimethylated histone H4 is an epigenetic marker of early apoptosis. J Biol Chem. 2006;281(19):13540-7.

65. Marquard L, Gjerdrum LM, Christensen IJ, Jensen PB, Sehested M, Ralfkiaer E. Prognostic significance of the therapeutic targets histone deacetylase 1, 2, 6 and acetylated histone H4 in cutaneous T-cell lymphoma. Histopathology. 2008;53(3):267-77.

66. Li D, Huang Y. Knockdown of ubiquitin associated protein 2-like inhibits the growth and migration of prostate cancer cells. Oncol Rep. 2014;32(4):1578-84.

67. Zhao B, Zong G, Xie Y, Li J, Wang H, Bian E. Downregulation of ubiquitin-associated protein 2-like with a short hairpin RNA inhibits human glioma cell growth in vitro. Int J Mol Med. 2015;36(4):1012-8.

68. Zhou Z, Yu L, Kleinerman ES. EWS-FLI-1 regulates the neuronal repressor gene REST, which controls Ewing sarcoma growth and vascular morphology. Cancer. 2014;120(4):579-88.

69. Huang Z, Bao S. Ubiquitination and deubiquitination of REST and its roles in cancers. FEBS Lett. 2012;586(11):1602-5.

70. Chen D, Yang X, Huang L, Chi P. The expression and clinical significance of PA28 γ in colorectal cancer. J Investig Med. 2013;61(8):1192-6.

71. Zhou ZJ, Dai Z, Zhou SL, Hu ZQ, Chen Q, Zhao YM, et al. HNRNPAB induces epithelial-mesenchymal transition and promotes metastasis of hepatocellular carcinoma by transcriptionally activating SNAIL. Cancer Res. 2014;74(10):2750-62.

72. Zou Q, Yang ZL, Yuan Y, Li JH, Liang LF, Zeng GX, et al. Clinicopathological features and CCT2 and PDIA2 expression in gallbladder squamous/adenosquamous carcinoma and gallbladder adenocarcinoma. World J Surg Oncol. 2013;11:143.

73. Guest ST, Kratche ZR, Bollig-Fischer A, Haddad R, Ethier SP. Two members of the TRiC chaperonin complex, CCT2 and TCP1 are essential for survival of breast cancer cells and are linked to driving oncogenes. Exp Cell Res. 2015;332(2):223-35.

74. Rajesh K, Krishnamoorthy J, Kazimierczak U, Tenkerian C, Papadakis AI, Wang S, et al. Phosphorylation of the translation initiation factor eIF2α at serine 51 determines the cell fate decisions of Akt in response to oxidative stress. Cell Death Dis. 2015;6(1):e1591.

75. Tuval-Kochen L, Paglin S, Keshet G, Lerenthal Y, Nakar C, Golani T, et al. Eukaryotic initiation factor 2α--a downstream effector of mammalian target of rapamycin--modulates DNA repair and cancer response to treatment. PLoS One. 2013;8(10):e77260.

76. Fu C, Wan Y, Shi H, Gong Y, Wu Q, Yao Y, et al. Expression and regulation of CacyBP/SIP in chronic lymphocytic leukemia cell balances of cell proliferation with apoptosis. J Cancer Res Clin Oncol. 2016;142(4):741-8.

77. Zhai HH, Meng J, Wang JB, Liu ZX, Li YF, Feng SS. CacyBP/SIP nuclear translocation induced by gastrin promotes gastric cancer cell proliferation. World J Gastroenterol. 2014;20(29):10062-70.

78. Lee JA, Park JE, Lee DH, Park SG, Myung PK, Park BC, et al. G1 to S phase transition protein 1 induces apoptosis signal-regulating kinase 1 activation by dissociating 14-3-3 from ASK1. Oncogene. 2008;27(9):1297-305.

79. Malta-Vacas J, Chauvin C, Gonçalves L, Nazaré A, Carvalho C, Monteiro C, et al. eRF3a/GSPT1 12-GGC allele increases the susceptibility for breast cancer development. Oncol Rep. 2009;21(6):1551-8.

80. Kato M, Wei M, Yamano S, Kakehashi A, Tamada S, Nakatani T, et al. DDX39 acts as a suppressor of invasion for bladder cancer. Cancer Sci. 2012;103(7):1363-9.

81. Sugiura T, Nagano Y, Noguchi Y. DDX39, upregulated in lung squamous cell cancer, displays RNA helicase activities and promotes cancer cell growth. Cancer Biol Ther. 2007;6(6):957-64.

82. Zhang F, Liu Y, Wang Z, Sun X, Yuan J, Wang T, et al. A novel Anxa2-interacting protein Ebp1 inhibits cancer proliferation and invasion by suppressing Anxa2 protein level. Mol Cell Endocrinol. 2015;411:75-85.

83. Linge A, Maurya P, Friedrich K, Baretton GB, Kelly S, Henry M, et al. Identification and functional validation of RAD23B as a potential protein in human breast cancer progression. J Proteome Res. 2014;13(7):3212-22.

84. Yasui K, Arii S, Zhao C, Imoto I, Ueda M, Nagai H, et al. TFDP1, CUL4A, and CDC16 identified as targets for amplification at 13q34 in hepatocellular carcinomas. Hepatology. 2002;35(6):1476-84.

85. Wang Y, Wen M, Kwon Y, Xu Y, Liu Y, Zhang P, et al. CUL4A induces epithelial-mesenchymal transition and promotes cancer metastasis by regulating ZEB1 expression. Cancer Res. 2014;74(2):520-31.

86. McGee AM, Douglas DL, Liang Y, Hyder SM, Baines CP. The mitochondrial protein C1qbp promotes cell proliferation, migration and resistance to cell death. Cell Cycle. 2011;10(23):4119-27.

87. Saha P, Ghosh I, Datta K. Increased hyaluronan levels in HABP1/p32/gC1qR overexpressing HepG2 cells inhibit autophagic vacuolation regulating tumor potency. PLoS One. 2014;9(7):e103208.

88. Tucker LA, Zhang Q, Sheppard GS, Lou P, Jiang F, McKeegan E, et al. Ectopic expression of methionine aminopeptidase-2 causes cell transformation and stimulates proliferation. Oncogene. 2008;27(28):3967-76.

89. Warder SE, Tucker LA, McLoughlin SM, Strelitzer TJ, Meuth JL, Zhang Q, et al. Discovery, Identification, and Characterization of Candidate Pharmacodynamic Markers of Methionine Aminopeptidase-2 Inhibition. Journal of Proteome Research. 2008;7(11):4807-20.

90. Shimizu H, Yamagishi S, Chiba H, Ghazizadeh M. Methionine Aminopeptidase 2 as a Potential Therapeutic Target for Human Non-Small-Cell Lung Cancers. Adv Clin Exp Med. 2016;25(1):117-28.

91. Kim SS, Hur SY, Kim YR, Yoo NJ, Lee SH. Expression of AIMP1, 2 and 3, the scaffolds for the multi-tRNA synthetase complex, is downregulated in gastric and colorectal cancer. Tumori. 2011;97(3):380-5.

92. Butt AJ, Sergio CM, Inman CK, Anderson LR, McNeil CM, Russell AJ, et al. The estrogen and c-Myc target gene HSPC111 is over-expressed in breast cancer and associated with poor patient outcome. Breast Cancer Res. 2008;10(2):R28.

93. He H-c, Ling X-h, Zhu J-g, Fu X, Han Z-d, Liang Y-x, et al. Down-regulation of the ErbB3 binding protein 1 in human bladder cancer promotes tumor progression and cell proliferation. Molecular Biology Reports. 2013;40(5):3799-805.

94. Liu L, Li XD, Chen HY, Cui JS, Xu DY. Significance of Ebp1 and p53 protein expression in cervical cancer. Genet Mol Res. 2015;14(4):11860-6.

95. Tamaki M, Goi T, Hirono Y, Katayama K, Yamaguchi A. PPP2R1B gene alterations inhibit interaction of PP2A-Abeta and PP2A-C proteins in colorectal cancers. Oncol Rep. 2004;11(3):655-9.

96. Zhang Y, Talmon G, Wang J. MicroRNA-587 antagonizes 5-FU-induced apoptosis and confers drug resistance by regulating PPP2R1B expression in colorectal cancer. Cell Death Dis. 2015;6(8):e1845.

97. Avram S, Mernea M, Mihailescu DF, Seiman CD, Seiman DD, Putz MV. Mitotic checkpoint proteins Mad1 and Mad2 - structural and functional relationship with implication in genetic diseases. Curr Comput Aided Drug Des. 2014;10(2):168-81.

98. Du J, Du Q, Zhang Y, Sajdik C, Ruan Y, Tian XX, et al. Expression of cell-cycle regulatory proteins BUBR1, MAD2, Aurora A, cyclin A and cyclin E in invasive ductal breast carcinomas. Histol Histopathol. 2011;26(6):761-8.

99. Choi JW, Kim Y, Lee JH, Kim YS. High expression of spindle assembly checkpoint proteins CDC20 and MAD2 is associated with poor prognosis in urothelial bladder cancer. Virchows Arch. 2013;463(5):681-7.

100. Ledoux AC, Sellier H, Gillies K, Iannetti A, James J, Perkins ND. NFκB regulates expression of Polo-like kinase 4. Cell Cycle. 2013;12(18):3052-62.

101. Sampson PB, Liu Y, Patel NK, Feher M, Forrest B, Li SW, et al. The discovery of Polo-like kinase 4 inhibitors: design and optimization of spiro[cyclopropane-1,3□'[3H]indol]-2'(1'H).ones as orally bioavailable antitumor agents. J Med Chem. 2015;58(1):130-46.

102. Wenqi D, Li W, Shanshan C, Bei C, Yafei Z, Feihu B, et al. EpCAM is overexpressed in gastric cancer and its downregulation suppresses proliferation of gastric cancer. Journal of Cancer Research and Clinical Oncology. 2009;135(9):1277-85.

103. Wang G, Zhang Z, Ren Y. TROP-1/Ep-CAM and CD24 are potential candidates for ovarian cancer therapy. Int J Clin Exp Pathol. 2015;8(5):4705-14.

104. Gao J, Yan Q, Liu S, Yang X. Knockdown of EpCAM enhances the chemosensitivity of breast cancer cells to 5-fluorouracil by downregulating the antiapoptotic factor Bcl-2. PLoS One. 2014;9(7):e102590.

105. Santin AD, Cane S, Bellone S, Bignotti E, Palmieri M, De Las Casas LE, et al. The novel serine protease tumor-associated differentially expressed gene-15 (matriptase/MT-SP1) is highly overexpressed in cervical carcinoma. Cancer. 2003;98(9):1898-904.

106. Singh PK, Srivastava AK, Dalela D, Rath SK, Goel MM, Bhatt ML. Frequent expression of zinc-finger protein ZNF165 in human urinary bladder transitional cell carcinoma. Immunobiology. 2015;220(1):68-73.

107. Dong W, Chen X, Xie J, Sun P, Wu Y. Epigenetic inactivation and tumor suppressor activity of HAI-2/SPINT2 in gastric cancer. Int J Cancer. 2010;127(7):1526-34.

108. Suzuki M, Kobayashi H, Tanaka Y, Hirashima Y, Kanayama N, Takei Y, et al. Suppression of invasion and peritoneal carcinomatosis of ovarian cancer cell line by overexpression of bikunin. Int J Cancer. 2003;104(3):289-302.

109. Ikeo K, Oshima T, Shan J, Matsui H, Tomita T, Fukui H, et al. Junctional adhesion molecule-A promotes proliferation and inhibits apoptosis of gastric cancer. Hepatogastroenterology. 2015;62(138):540-5.

110. Tian Y, Tian Y, Zhang W, Wei F, Yang J, Luo X, et al. Junctional adhesion molecule-A, an epithelial-mesenchymal transition inducer, correlates with metastasis and poor prognosis in human nasopharyngeal cancer. Carcinogenesis. 2015;36(1):41-8.

111. Yang X, Cao W, Zhou J, Zhang W, Zhang X, Lin W, et al. 14-3-3ζ positive expression is associated with a poor prognosis in patients with glioblastoma. Neurosurgery. 2011;68(4):932-8; discussion 8.

112. Rüenauver K, Menon R, Svensson MA, Carlsson J, Vogel W, Andrén O, et al. Prognostic significance of YWHAZ expression in localized prostate cancer. Prostate Cancer Prostatic Dis. 2014;17(4):310-4.

113. Bergamaschi A, Frasor J, Borgen K, Stanculescu A, Johnson P, Rowland K, et al. 14-3-3ζ as a predictor of early time to recurrence and distant metastasis in hormone receptor-positive and -negative breast cancers. Breast Cancer Res Treat. 2013;137(3):689-96.

114. Cuevas R, Korzeniewski N, Tolstov Y, Hohenfellner M, Duensing S. FGF-2 disrupts mitotic stability in prostate cancer through the intracellular trafficking protein CEP57. Cancer Res. 2013;73(4):1400-10.

115. Bilke S, Schwentner R, Yang F, Kauer M, Jug G, Walker RL, et al. Oncogenic ETS fusions deregulate E2F3 target genes in Ewing sarcoma and prostate cancer. Genome Res. 2013;23(11):1797-809.

116. Giorgi C, Boro A, Rechfeld F, Lopez-Garcia LA, Gierisch ME, Schäfer BW, et al. PI3K/AKT signaling modulates transcriptional expression of EWS/FLI1 through specificity protein 1. Oncotarget. 2015;6(30):28895-910.

117. Tsofack SP, Meunier L, Sanchez L, Madore J, Provencher D, Mes-Masson AM, et al. Low expression of the X-linked ribosomal protein S4 in human serous epithelial ovarian cancer is associated with a poor prognosis. BMC Cancer. 2013;13:303.

118. Paquet É R, Hovington H, Brisson H, Lacombe C, Larue H, Têtu B, et al. Low level of the X-linked ribosomal protein S4 in human urothelial carcinomas is associated with a poor prognosis. Biomark Med. 2015;9(3):187-97.

119. Zhou X, Liu Y, You J, Zhang H, Zhang X, Ye L. Myosin light-chain kinase contributes to the proliferation and migration of breast cancer cells through cross-talk with activated ERK1/2. Cancer Lett. 2008;270(2):312-27.

120. Chen L, Su L, Li J, Zheng Y, Yu B, Yu Y, et al. Hypermethylated FAM5C and MYLK in serum as diagnosis and pre-warning markers for gastric cancer. Dis Markers. 2012;32(3):195-202.

121. Ducker CE, Upson JJ, French KJ, Smith CD. Two N-myristoyltransferase isozymes play unique roles in protein myristoylation, proliferation, and apoptosis. Mol Cancer Res. 2005;3(8):463-76.

122. Luo H, Hao X, Ge C, Zhao F, Zhu M, Chen T, et al. TC21 promotes cell motility and metastasis by regulating the expression of E-cadherin and N-cadherin in hepatocellular carcinoma. Int J Oncol. 2010;37(4):853-9.

123. Gutierrez-Erlandsson S, Herrero-Vidal P, Fernandez-Alfara M, Hernandez-Garcia S, Gonzalo-Flores S, Mudarra-Rubio A, et al. R-RAS2 overexpression in tumors of the human central nervous system. Mol Cancer. 2013;12(1):127.

124. Larive RM, Moriggi G, Menacho-Márquez M, Cañamero M, de Álava E, Alarcón B, et al. Contribution of the R-Ras2 GTP-binding protein to primary breast tumorigenesis and late-stage metastatic disease. Nat Commun. 2014;5:3881.

125. Zhou B, Wu Q, Chen G, Zhang TP, Zhao YP. NOP14 promotes proliferation and metastasis of pancreatic cancer cells. Cancer Lett. 2012;322(2):195-203.

126. Sheng X, Bowen N, Wang Z. GLI pathogenesis-related 1 functions as a tumor-suppressor in lung cancer. Mol Cancer. 2016;15:25.

127. Li L, Ren C, Yang G, Fattah EA, Goltsov AA, Kim SM, et al. GLIPR1 suppresses prostate cancer development through targeted oncoprotein destruction. Cancer Res. 2011;71(24):7694-704.

128. Shi L, Zhang B, Sun X, Lu S, Liu Z, Liu Y, et al. MiR-204 inhibits human NSCLC metastasis through suppression of NUAK1. Br J Cancer. 2014;111(12):2316-27.

129. Bell RE, Khaled M, Netanely D, Schubert S, Golan T, Buxbaum A, et al. Transcription factor/microRNA axis blocks melanoma invasion program by miR-211 targeting NUAK1. J Invest Dermatol. 2014;134(2):441-51.

130. Hou X, Liu JE, Liu W, Liu CY, Liu ZY, Sun ZY. A new role of NUAK1: directly phosphorylating p53 and regulating cell proliferation. Oncogene. 2011;30(26):2933-42.

131. Lee SY, Kim JW, Jeong MH, An JH, Jang SM, Song KH, et al. Microtubule-associated protein 1B light chain (MAP1B-LC1) negatively regulates the activity of tumor suppressor p53 in neuroblastoma cells. FEBS Lett. 2008;582(19):2826-32.

132. Guo MM, Hu LH, Wang YQ, Chen P, Huang JG, Lu N, et al. miR-22 is down-regulated in gastric cancer, and its overexpression inhibits cell migration and invasion via targeting transcription factor Sp1. Med Oncol. 2013;30(2):542.

133. Zhang G, Xia S, Tian H, Liu Z, Zhou T. Clinical significance of miR-22 expression in patients with colorectal cancer. Med Oncol. 2012;29(5):3108-12.

134. Wang J, Sun D, Wang Y, Ren F, Pang S, Wang D, et al. FOSL2 positively regulates TGF-β1 signalling in non-small cell lung cancer. PLoS One. 2014;9(11):e112150.

135. Milde-Langosch K, Janke S, Wagner I, Schröder C, Streichert T, Bamberger AM, et al. Role of Fra-2 in breast cancer: influence on tumor cell invasion and motility. Breast Cancer Res Treat. 2008;107(3):337-47.

136. Malaguti C, Rossini GP. Recovery of cellular E-cadherin precedes replenishment of estrogen receptor and estrogen-dependent proliferation of breast cancer cells rescued from a death stimulus. J Cell Physiol. 2002;192(2):171-81.

137. Li YJ, Ji XR. Relationship between expression of E-cadherin-catenin complex and clinicopathologic characteristics of pancreatic cancer. World J Gastroenterol. 2003;9(2):368-72.

138. Rodriguez C, Borgel J, Court F, Cathala G, Forné T, Piette J. CTCF is a DNA methylation-sensitive positive regulator of the INK/ARF locus. Biochem Biophys Res Commun. 2010;392(2):129-34.

139. Docquier F, Farrar D, Arcy V, Chernukhin I, Robinson AF, Loukinov D, et al. Heightened Expression of CTCF in Breast Cancer Cells Is Associated with Resistance to Apoptosis. Cancer Research. 2005;65(12):5112.

140. Tane S, Sakai Y, Hokka D, Okuma H, Ogawa H, Tanaka Y, et al. Significant role of Psf3 expression in non-small-cell lung cancer. Cancer Sci. 2015;106(11):1625-34.

141. Hattori M, Minato N. Rap1 GTPase: functions, regulation, and malignancy. J Biochem. 2003;134(4):479-84.

142. Tsygankova OM, Wang H, Meinkoth JL. Tumor cell migration and invasion are enhanced by depletion of Rap1 GTPase-activating protein (Rap1GAP). J Biol Chem. 2013;288(34):24636-46.

143. Tsygankova OM, Feshchenko E, Klein PS, Meinkoth JL. Thyroid-stimulating hormone/cAMP and glycogen synthase kinase 3beta elicit opposing effects on Rap1GAP stability. J Biol Chem. 2004;279(7):5501-7.

144. Srivastava A, Alexander J, Lomakin I, Dayal Y. Immunohistochemical expression of transforming growth factor alpha and epidermal growth factor receptor in pancreatic endocrine tumors. Hum Pathol. 2001;32(11):1184-9.

145. Rhee J, Han SW, Cha Y, Ham HS, Kim HP, Oh DY, et al. High serum TGF-α predicts poor response to lapatinib and capecitabine in HER2-positive breast cancer. Breast Cancer Res Treat. 2011;125(1):107-14.

146. Wasniewski T, Woclawek-Potocka I, Boruszewska D, Kowalczyk-Zieba I, Sinderewicz E, Grycmacher K. The significance of the altered expression of lysophosphatidic acid receptors, autotaxin and phospholipase A2 as the potential biomarkers in type 1 endometrial cancer biology. Oncol Rep. 2015;34(5):2760-7.

147. Fujita T, Miyamoto S, Onoyama I, Sonoda K, Mekada E, Nakano H. Expression of lysophosphatidic acid receptors and vascular endothelial growth factor mediating lysophosphatidic acid in the development of human ovarian cancer. Cancer Lett. 2003;192(2):161-9.

148. Lai YH, He RY, Chou JL, Chan MW, Li YF, Tai CK. Promoter hypermethylation and silencing of tissue factor pathway inhibitor-2 in oral squamous cell carcinoma. J Transl Med. 2014;12:237.

149. Makishima H, Visconte V, Sakaguchi H, Jankowska AM, Abu Kar S, Jerez A, et al. Mutations in the spliceosome machinery, a novel and ubiquitous pathway in leukemogenesis. Blood. 2012;119(14):3203-10.

150. Edmond V, Merdzhanova G, Gout S, Brambilla E, Gazzeri S, Eymin B. A new function of the splicing factor SRSF2 in the control of E2F1-mediated cell cycle progression in neuroendocrine lung tumors. Cell Cycle. 2013;12(8):1267-78.

151. Nemoto K, Vogt A, Oguri T, Lazo JS. Activation of the Raf-1/MEK/Erk kinase pathway by a novel Cdc25 inhibitor in human prostate cancer cells. Prostate. 2004;58(1):95-102.

152. Mi Y, Thomas SD, Xu X, Casson LK, Miller DM, Bates PJ. Apoptosis in leukemia cells is accompanied by alterations in the levels and localization of nucleolin. J Biol Chem. 2003;278(10):8572-9.

153. Zhang R, Zhang Y, Li H. miR-1244/Myocyte Enhancer Factor 2D Regulatory Loop Contributes to the Growth of Lung Carcinoma. DNA Cell Biol. 2015;34(11):692-700.

154. Li J, Tan M, Li L, Pamarthy D, Lawrence TS, Sun Y. SAK, a new polo-like kinase, is transcriptionally repressed by p53 and induces apoptosis upon RNAi silencing. Neoplasia. 2005;7(4):312-23.

155. Fan G, Sun L, Shan P, Zhang X, Huan J, Zhang X, et al. Loss of KLF14 triggers centrosome amplification and tumorigenesis. Nat Commun. 2015;6:8450.

156. Peng ZG, Yao YB, Yang J, Tang YL, Huang X. Mangiferin induces cell cycle arrest at G2/M phase through ATR-Chk1 pathway in HL-60 leukemia cells. Genet Mol Res. 2015;14(2):4989-5002.

157. Cho SH, Toouli CD, Fujii GH, Crain C, Parry D. Chk1 is essential for tumor cell viability following activation of the replication checkpoint. Cell Cycle. 2005;4(1):131-9.

158. Varga AE, Stourman NV, Zheng Q, Safina AF, Quan L, Li X, et al. Silencing of the Tropomyosin-1 gene by DNA methylation alters tumor suppressor function of TGF-beta. Oncogene. 2005;24(32):5043-52.

159. Wang J, Guan J, Lu Z, Jin J, Cai Y, Wang C, et al. Clinical and tumor significance of tropomyosin-1 expression levels in renal cell carcinoma. Oncol Rep. 2015;33(3):1326-34.

160. Kristiansen G, Denkert C, Schlüns K, Dahl E, Pilarsky C, Hauptmann S. CD24 is expressed in ovarian cancer and is a new independent prognostic marker of patient survival. Am J Pathol. 2002;161(4):1215-21.

161. Zhu J, Nie S, Wu J, Lubman DM. Target proteomic profiling of frozen pancreatic CD24+ adenocarcinoma tissues by immuno-laser capture microdissection and nano-LC-MS/MS. J Proteome Res. 2013;12(6):2791-804.

162. Gajulapalli VNarasihma R, Samanthapudi VSubramanyam K, Pulaganti M, Khumukcham Saratchandra S, Malisetty Vijaya L, Guruprasad L, et al. A transcriptional repressive role for epithelial-specific ETS factor ELF3 on oestrogen receptor alpha in breast cancer cells. Biochemical Journal. 2016;473(8):1047-61.

163. Kohno Y, Okamoto T, Ishibe T, Nagayama S, Shima Y, Nishijo K, et al. Expression of claudin7 is tightly associated with epithelial structures in synovial sarcomas and regulated by an Ets family transcription factor, ELF3. J Biol Chem. 2006;281(50):38941-50.

164. Man X, He J, Kong C, Zhu Y, Zhang Z. Clinical significance and biological roles of CARMA3 in human bladder carcinoma. Tumour Biol. 2014;35(5):4131-6.

165. Xia ZX, Li ZX, Zhang M, Sun LM, Zhang QF, Qiu XS. CARMA3 regulates the invasion, migration, and apoptosis of non-small cell lung cancer cells by activating NF-κB and suppressing the P38 MAPK signaling pathway. Exp Mol Pathol. 2016;100(2):353-60.

166. Mehdipour P, Pirouzpanah S, Sarafnejad A, Atri M, Shahrestani TS, Haidari M. Prognostic implication of CDC25A and cyclin E expression on primary breast cancer patients. Cell Biology International. 2009;33(10):1050-6.

167. Adler AS, McCleland ML, Truong T, Lau S, Modrusan Z, Soukup TM, et al. CDK8 maintains tumor dedifferentiation and embryonic stem cell pluripotency. Cancer Res. 2012;72(8):2129-39.

168. Clark AD, Oldenbroek M, Boyer TG. Mediator kinase module and human tumorigenesis. Crit Rev Biochem Mol Biol. 2015;50(5):393-426.

169. Mohamed ER, Naito M, Terasaki Y, Niu Y, Gohara S, Komatsu N, et al. Capability of SART3(109-118) peptide to induce cytotoxic T lymphocytes from prostate cancer patients with HLA class I-A11, -A31 and -A33 alleles. Int J Oncol. 2009;34(2):529-36.

170. Sasatomi T, Suefuji Y, Matsunaga K, Yamana H, Miyagi Y, Araki Y, et al. Expression of tumor rejection antigens in colorectal carcinomas. Cancer. 2002;94(6):1636-41.

171. Landrette SF, Kuo YH, Hensen K, Barjesteh van Waalwijk van Doorn-Khosrovani S, Perrat PN, Van de Ven WJ, et al. Plag1 and Plagl2 are oncogenes that induce acute myeloid leukemia in cooperation with Cbfb-MYH11. Blood. 2005;105(7):2900-7.

172. Hanks TS, Gauss KA. Pleomorphic adenoma gene-like 2 regulates expression of the p53 family member, p73, and induces cell cycle block and apoptosis in human promonocytic U937 cells. Apoptosis. 2012;17(3):236-47.

173. Liu B, Lu C, Song YX, Gao P, Sun JX, Chen XW, et al. The role of pleomorphic adenoma gene-like 2 in gastrointestinal cancer development, progression, and prognosis. Int J Clin Exp Pathol. 2014;7(6):3089-100.

174. Wang JL, Chen ZF, Chen HM, Wang MY, Kong X, Wang YC, et al. Elf3 drives β-catenin transactivation and associates with poor prognosis in colorectal cancer. Cell Death Dis. 2014;5(5):e1263-e.

175. Liu C, Zhang L, Huang Y, Lu K, Tao T, Chen S, et al. MicroRNA-328 directly targets p21-activated protein kinase 6 inhibiting prostate cancer proliferation and enhancing docetaxel sensitivity. Mol Med Rep. 2015;12(5):7389-95.

176. Liu X, Busby J, John C, Wei J, Yuan X, Lu ML. Direct interaction between AR and PAK6 in androgen-stimulated PAK6 activation. PLoS One. 2013;8(10):e77367.

177. Blum C, Graham A, Yousefzadeh M, Shrout J, Benjamin K, Krishna M, et al. The expression ratio of Map7/B2M is prognostic for survival in patients with stage II colon cancer. Int J Oncol. 2008;33(3):579-84.

178. Biliran H, Jr., Sheng S. Pleiotrophic inhibition of pericellular urokinase-type plasminogen activator system by endogenous tumor suppressive maspin. Cancer Res. 2001;61(24):8676-82.

179. Ashida S, Furihata M, Katagiri T, Tamura K, Anazawa Y, Yoshioka H, et al. Expression of novel molecules, MICAL2-PV (MICAL2 prostate cancer variants), increases with high Gleason score and prostate cancer progression. Clin Cancer Res. 2006;12(9):2767-73.

180. Dallas SL, Zhao S, Cramer SD, Chen Z, Peehl DM, Bonewald LF. Preferential production of latent transforming growth factor beta-2 by primary prostatic epithelial cells and its activation by prostate-specific antigen. J Cell Physiol. 2005;202(2):361-70.

181. Wick W, Platten M, Weller M. Glioma cell invasion: regulation of metalloproteinase activity by TGF-beta. J Neurooncol. 2001;53(2):177-85.

182. Ohno Y, Izumi M, Kawamura T, Nishimura T, Mukai K, Tachibana M. Annexin II represents metastatic potential in clear-cell renal cell carcinoma. Br J Cancer. 2009;101(2):287-94.

183. Stephenson JM, Banerjee S, Saxena NK, Cherian R, Banerjee SK. Neuropilin-1 is differentially expressed in myoepithelial cells and vascular smooth muscle cells in preneoplastic and neoplastic human breast: a possible marker for the progression of breast cancer. Int J Cancer. 2002;101(5):409-14.

APPENDIX C: SUPPPLEMENTARY MATERIAL FOR CHAPTER 5

David Similarity Score:

The method for calculating similarity scores will eb summarized here. For the full explanation of how the similarity score is used consult the DAVID website using the following web address https://david.ncifcrf.gov/helps/linear_search.html#result. Each gene in DAVID is associated with several functional annotations. For example, these functional annotations might be if they involved in cell replication or both nuclear proteins. The similarity score between two genes represent how many functional annotations are shared between two genes. Consider two genes, gene a and gene b the following table describes each gene and its functional annotations

Table C.1. Example of functional annotations of genes.

	Cell Cycle	Nuclear protein	Apoptosis	DNA Synthesis
gene a	1	1	1	0
gene b	1	1	0	0

where 1 indicates the gene has that functional annotation and 0 indicates that the gene lacks that functional annotation. Given the values of Table C.1, the following quantities, defined in Table C.2, can be calculated:

$$O_{ab} = \frac{c_{1,1} + c_{0,0}}{T_{ab}} = \frac{2+1}{4} = 0.75$$

$$A_{ab} = \frac{C_1 * R_1 + C_2 * R_2}{T_{ab}^2} = \frac{3*2+1*2}{16} = 0.5$$

$$K_{ab} = \frac{O_{ab} - A_{ab}}{1 - A_{ab}} = \frac{0.25}{0.5} = 0.5$$

Table C.2: Values used to calculate the Kappa statistic

	gene a			
gene b		1	0	Row Total
	1	2 ($c_{1,1}$)	0 ($c_{1,0}$)	2 (R1)
	0	1 ($c_{0,1}$)	1 ($c_{0,0}$)	2 (R2)
Column Totals		3 (C_1)	1(C_2)	4 (T_{ab})

K_{ab} is referred to the as the Kappa statistic and is the given similarity score (1). The Kappa Statistic ranges from 0 to 1 being complete overlap (i.e. identity) and 0 being no overlap. David defined the following ranges when defining related genes; Very high (0.75-1), High (0.5-0.75), Moderate (0.25-0.5), and Low (0-0.25). The score cutoff used in chapter 5 was 0.35.

Survival Support Vector Machines

Technically given a cohort of patients and time of death after the diagnosis of the disease one could use a regression method to model survival. However, survival of some of the patients might exceed the duration of the study or patients might drop out before the completion of the study, these patients are referred to as censored. In censored cases, such as these, a regression analysis would require those data points be thrown out which could result in a significant loss of data. The other possibility is to transform the problem into a classification task such that survival beyond some time is considered one class and everything before is the other. However, all censored data before that cutoff would have to be eliminated from analysis because there is no

certainty as to which class such samples belong again resulting in data loss. Survival support vector machines salvages this data by using it where appropriate. This problem follow the formulation given in Van Belle *et al. (2).* First the regression problem is turned into a more tractable problem of predicating the rank at which the event (death) happens. Given two patients x_i and x_j associated with observation times y_i and y_j and censorship status δ_i and δ_j which is an indicator function with a value of 1 for uncensored data or a value of 0 for censored data. Define the following function

$$comp(i,j) = \begin{cases} 1 \; if \; \delta_i = 1, \delta_j = 1 \; or \; \delta_i = 1, \delta_j = 0 \; and \; y_i \leq y_j \\ 0 \; otherwise \end{cases}$$

And define the following optimization problem

$$\min_{w} \frac{1}{2} w^T w$$

Subject to

$$w^T(x_i - x_j) \geq 1 \;\; \forall i = 1, \ldots\ldots, n; \forall j: y_i > y_j \; and \; comp(i,j) = 1$$

Again as is the case with regression this might be too stringent of a problem to solve and therefore a slack variable, ε_{ij}, is introduced and the problem becomes

$$\min_{w,\varepsilon} \frac{1}{2} w^T w + \alpha \sum_{i=1}^{N} \sum\nolimits_{j:y_i > y_j} \varepsilon_{ij}$$

Subject to

$$\begin{cases} w^T(x_i - x_j) \geq 1 - \varepsilon_{ij}, \forall i = 1, \ldots, n; \; \forall j: y_i > y_j \; and \; comp(i,j) = 1 \\ \varepsilon_{ij} \geq 0, \forall i = 1, \ldots, n; \; \forall j: y_i > y_j \; and \; comp(i,j) = 1 \end{cases}$$

For a new data point x^* the dual formulation gives the relative rank, u, as

$$u(x^*) = \sum_{i=1}^{N} \sum\nolimits_{j:y_i > y_j, comp(i,j)=1} a_{ij}(x_i - x_j)^T x^*$$

Where a_{ij} are Lagrange multipliers. Therefore $a_{ij} = 0\ if\ y_i < y_j\ or\ comp(i,j) = 0$ similar to how only the Lagrange multipliers for points outside the epsilon tube are used for calculations in support vector regression.

Measuring Performance of Survival Support Vector Machines

The output of the survival SVM demonstrated above is a relative rank where by lower values of $u(x^*)$ indicate a shorter survival time compared to higher values. If the testing data was fully uncensored the performance of the model could be measured by a rank correlation. However, given that the testing set likely contains some censored data, much like as in the SVM case, this needs to be taken into account. First lets define the actual time an event happens as T_i and the relative predicted rank u_i. Given a pair of events (i,j) a concordant pair is defined such that if $T_i > T_j$ then $u_i > u_j$ and discordant when $u_i < u_j$. However, if both T_i and T_j are censored there is no way to tell if they are concordant or discordant and they cannot be used in the calculation. However, if T_j is uncensored and T_i is censored as long as $T_i > T_j$ then the pair is concordant if $u_i > u_j$ and discordant if $u_i < u_j$. The concordance index,c, is given by

$$c = \frac{\#concordant\ pairs}{\#\ concordant\ pairs + \#\ discordent\ pairs}$$

This particular method of calculating the concordance index is known as Harrell's c-index (3).

REFERENCES

1. Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement. 1960;20(1):37-46.

2. Van Belle V, Pelckmans K, Van Huffel S, Suykens JAK. Support vector methods for survival analysis: a comparison between ranking and regression approaches. Artificial Intelligence in Medicine. 2011;53(2):107-18.

3. Harrell FE, Jr., Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the Yield of Medical Tests. JAMA. 1982;247(18):2543-6.