CONNECTING GENOTYPES TO DRUG SENSITIVITIES IN HER2 POSITIVE CANCER CELL LINES

By

Ted Laderas

A DISSERTATION

Presented to the Department of Medical Informatics and Clinical Epidemiology
And the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

March 2014

School of Medicine

Oregon Health & Science University

Certificate of Approval

This is to certify that the PhD Dissertation of

Ted G. Laderas

"Connecting Genotypes to Drug Sensitivities in HER2 Positive Cancer Cell Lines "

Has been approved

_____
Dissertation Advisor – Kemal Sonmez

_____
Committee Member – Laura Heiser

_____
Committee Member – Joe Gray

_____
Committee Member – Wayne Wakeland

Abstract

HER2-positive breast cancer is an aggressive subtype of breast cancer, with patients having significantly lower survival rates compared to other breast cancers. HER2-positive breast cancer shows pathway addiction to multiple mitogenic signaling pathways. One strategy in cancer treatment is targeted drug therapy, which inhibits the function of specific proteins along these signaling pathways. Predicting drug sensitivity using a patient's tumor markers such as somatic mutations and copy number variation can help guide treatment selection towards realization of "precision" medicine.

The motivation for this work is to (1) characterize the degree of cross-phenotype response when a targeted inhibitor is applied and (2) identify potential oncogene collaborations that can drive the decisions of which targeted treatments should be used. These two problems form the basis of my dissertation aims. In this Dissertation, we highlight two approaches: characterizing drug response and resultant system cross-phenotype response using proteomics and a new network based model of oncogenic collaboration using integration of multiple data types (mutation calls, drug sensitivity, and copy number calls) for visualization... In terms of proteomics, we specifically highlight a robust measure of time series that I apply to proteomics data: Area Under the Curve(AUC). AUC is a measure of upregulation/downregulation over time and I show that it is a useful feature for estimating cross-phenotype response. Specifically, we show that the AUC of four proteins (BIM, RB1, ERK, S6) is highly correlated with drug sensitivity. Using AUCs, I then attempted to predict protein cross-phenotype response for four different cancer cell lines (UACC812, BT549, BT20, and MCF7) using a linear modeling approach called Partial Least Squares Path modeling (PLS-PM). PLS-PM modeling highlights proteins which are well characterized by the data (EGFR, HER2, SRC), but also highlights proteins for which we have incomplete knowledge (PDK1, PTEN, MAPK1)

In the network-based approach, we highlight a possible method for integrating unique mutations in a cell line, that of "surrogate mutations". This method may point the way to precision medicine in a patient specific context. We observe that when imposed on a protein-protein interaction network, mutations tend to cluster around certain nodes, and show these mutations are statistically significant under a null model of randomly mutated networks. More importantly, we show that these surrogate mutations are associated with drug sensitivity through the use of Random Forests for classification, with an average error rate across all drugs of 30.9%, comparable to the error rate for expression-based subtype specificity of 29.1%. This suggests that surrogate mutations capture network specific elements that are important to predicting drug sensitivity.

Acknowlegements

A dissertation, especially in this day and age, is not done alone. I would like to thank all of the people and research groups that made this work possible.

First of all, I am grateful to my family, including my husband, Jim Ott, and my mother for their unwavering support and love through this difficult time.

Next, I would like to thank all of the research groups who both generated the data and did much of the preprocessing and analysis and provided valuable feedback. This includes the Gray/Spellman group, especially Paul Spellman, Catie Grasso, Myron Peto, Nicole Nesser, Norene Jeliffe, Katie Crossen, Jim Korkola, Pavana Anur, and Asia Mitchell. I am also grateful for the help from the Gordon Mills group for helping me understand the RPPA data, especially Gordon Mills, Yiling Lu, Rehan Akbani, Zhenlin Ju, Wenbin Liu, and Fan Zhang. Thanks also to Thomas Cokelaer for helping me with the DREAM8 evaluation.

Thank you to the NLM Training program for providing me with financial support and valuable feedback. Thank you to my department, the Department of Medical Informatics and Clinical Epidemiology, including Karen Eden, Bill Hersh, Andrea Ilg, Diane Doctor and Lynne Schwabe. Thank you to all of the other DMICE fellows and students.

Finally, thank you to my committee: Joe Gray, Laura Heiser, Shannon McWeeney, Kemal Sonmez and Wayne Wakeland.

Table of Contents

Table of Figures

Table of Tables

Chapter 1: Introduction and Background

1.1 Introduction

HER2-positive breast cancer is an aggressive subtype of breast cancer, with patients having significantly lower survival rates compared to other breast cancers.[1,2] HER2 positive breast cancer is characterized by tumor cells that overexpress the HER2 cellular receptor on their surface. HER2 is a participant in cellular signaling, the process by which a cell decides to proliferate and differentiate given inputs from surrounding cells, as well as inputs from extracellular sources such as the extracellular matrix. This process, when disrupted, is implicated in cancer formation.[3,4] Disrupted signaling results in overproliferation of cancerous cells. Cancer is thought to arise because aberrant states (or balances) of the signaling feedback loops disrupt the decision making process of the cell.[5]

One strategy in cancer treatment is targeted drug therapy, which inhibits the function of specific proteins along these signaling pathways. Two drugs that specifically target the HER2 receptor are Pertuzumab and Trastuzumab. However, their response in HER2 positive patients is highly variable, with only an estimated 30-50% of patients responding positively to Trastuzumab.[6] It is clear that somatically (non-inherited) acquired mutations in cancer cells drive this variable response to targeted drugs, but little is known about how these mutations alter the balance of signaling systems.[7] Scheduling a targeted therapy is important in that the application of a therapy can place cancerous cells in a different state in terms of proteins and availability of their phosphorylated forms. This state is dependent on not only previous history, but on the genetic background of the cell that predisposes it to a particular state. Thus, understanding how patterns of mutations alter signaling is of vital importance to tailoring drug therapy in cancer patients.

One of the difficulties in relating mutations to cancer sensitivity is that mutations can affect cellular systems in multiple capabilities. Many of these capabilities are described in Weinberg's landmark paper about the Hallmarks of Cancer.[3,4] In this paper, Weinberg describes six capabilities cancer cells must acquire in order to proliferate. The original six hallmarks are 1) sustaining proliferative signaling, 2) evading growth suppressors, 3) resisting cell death, 4) enabling replicative immortality, inducing angiogenesis, and 6) activating invasion and metastasis. Later, the paper was updated to add two additional hallmarks[4] 7) reprogramming of energy metabolism and 8) evading immune response,. Thus, the accumulation of somatic mutations that enable acquisition of these capabilities is critical for cancer cells to develop.

One modality that is especially relevant to drug sensitivity is that of cellular signaling, which is how the cell responds to external inputs such as heat, osmotic stress, or the various intercellular inputs such as growth factors or cytokines. Based on these inputs, the cell may decide to undergo apoptosis (cellular death), or proliferate. In their review, Avraham and Yarden outline many of the challenges in using targeted drugs to influence cellular signaling with feedback, specifically in the EGFR pathway.[5] These challenges include understanding the combinatorial processing of receptors, understanding how mutations alter the steady state behavior of the cells, and how the feedback loops contained in the signaling pathways are altered by both drugs and mutations. They divide signaling into two different temporal domains, *immediate* and *late*, both exhibiting levels of feedback regulation. The immediate group consists of processes such as receptor endocytosis, phosphorylation cascades, microRNAs, and other covalent protein modifications such as ubiquitinylation. The late group consists of transcriptional regulators such as transcription factors, as well as RNA-binding proteins, newly synthesized adaptor proteins (in response to the immediate group), and phosphatases (which dephosphorylate phosphoproteins). Further complicating the matter is

how the two time scales are interlinked. For example, phosphorylation can increase the levels of a transcription factor, which influences the amount of transcript of a protein early in the phosphorylation cascade. Thus understanding how the two time scales interact is of vital importance in understanding how to 'steer' signaling using targeted drugs.

It is clear that somatic mutations affect the state of the cellular system, and the various signaling processes in particular. However, connecting the two is more difficult than simply identifying mutations in signaling pathways. In particular, we will see that a large portion of the mutations to a patient or cell line are unique to that individual or cell.[8,9]

To provide background, in the following sections, we will first discuss subtypes of breast cancer (Section 1.2) and then how drug sensitivity is measured (1.3), and concerns about cell lines (1.4). One of the current hypotheses of drug sensitivity will then be discussed, that of subtype specificity (Section 1.5).

Following that, we will discuss oncogenes and tumor suppressors (Section 1.6), two key genomic features that contribute to cancer, and how they alter cellular systems, such as mitogenic signaling, the cell cycle, and the apoptotic program (Section 1.7). This leads to a discussion of two key questions that are not well resolved in understanding drug sensitivity: 1) What cross-phenotype response exists between cellular systems when drugs are applied? (Section 1.9 and 2) How do oncogenes collaborate in forming drug sensitivity/resistance patterns? (Section 1.10).

1.2 Subtypes of Breast Cancer

It should be noted that there are at least two very different definitions of HER2 positive cancer. Briefly, the definition of HER2 positive cancer affects the characterization of both patients and cell lines because the phenotype can be defined in two different ways. The first is

the clinical definition from the American Society of Clinical Oncology/College of American Pathologists (ASOC-CAP)[10], and the second is the molecular subtype, or PAM50 definition from the National Cancer Institute[11]. The clinical definition defines a number of diagnostic criteria for clinical assays in patients. In this definition, a patient is considered HER2 positive if their immunohistochemical (IHC) staining is at a level of 3 or above (a protein level measure), or their fluorescent in situ hybridization assay shows more than 6 or more copies of HER2 per nucleus (a gene expression level measure).

In contrast, the PAM50 definition relies on gene expression of a set of 50 selected genes to delineate five intrinsic subtypes (LuminalA, LuminalB, Basal-like, normal-like, and HER2 enriched).[11] Essentially, expression levels of tumors derived from these five groups were measured and 50 genes were selected whose expression levels distinguished one subtype from the other using a supervised machine learning method. Thus, the subtypes derived from the PAM50 definition may be phenotypically different than those defined using the ASOC-CAP definition. Importantly, it was shown that these subtypes correlate with survival outcomes.[12] We will also see that these intrinsic subtypes are associated with targeted drug sensitivity.

1.3 Defining Drug Sensitivity

There are many definitions of drug sensitivity. The majority of these are derived from fitting a drug response curve. The drug response curve is measured by applying the drug in a number of increasing doses to cells in a well plate system.[13,14] The dosages may be applied to replicate cells in order to evaluate reproducibility. The plates are then incubated for a fixed period of time and their cellular proliferation is measured via assay. This assay may be a colorimetric assay assessing the amount of apoptosis, or cell death, such as cleaved caspase 3, or it may be assessed through visual inspection of single cells. The amount of cell proliferation is

then plotted against the log concentration of the dosage and sensitivity parameters can be derived by fitting the data to a logistic curve:

$$y = E_{inf} + \left(\frac{E_o - E_{inf}}{1 + \left(\frac{D}{EC_{50}}\right)^{HS}} \right)$$

where y is the response measure (the proliferation value), E_{inf} is the bottom asymptote of the response, E_o is the top asymptote of the response, D is the concentration dosage, HS is a slope coefficient that is equivalent to the Hill Coefficient, and EC_{50} is the concentration at half-maximal effect (Figure 1).[15] An additional metric is E_{max}, or the maximal response of the cells to the drug.

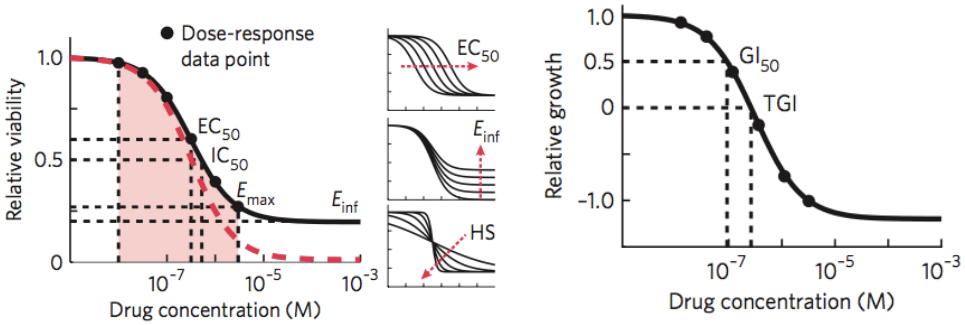


Figure 1. Drug sensitivity parameters such as EC_{50} can be derived from the dosage response curve. Reproduced from Fallahi-Sichiani.[15]

Fallahi-Sichiani et al showed that each of these experimental parameters is highly variable across drugs when measured across cell lines.[15] For example, some drugs such as AKT and PI3K inhibitors, had lower E_{max} than other drugs, whereas drugs like CGC-11144 (a polyamine analog) show high variability in the slope parameter HS. Moreover, each parameter does not capture the same quality of drug sensitivity as they are not highly correlated for cell lines across drugs, suggesting that drug sensitivity may not be completely captured by any single

parameter in this analysis.[15] One example of parameter is the IC50 versus the EC50, which are normally used interchangeably, but their lower correlation suggests that these two metrics are not. However, parameters of dose response curves do seem be associated within drug class.

The sensitivity metric we will use in this manuscript is GI50 (Growth Inhibition to 50 percent). It is defined as the concentration of a drug required to inhibit growth in a cell line to 50% of its original population measured on a Dose-Response Curve (Figure 2).[16] Measuring GI50 requires baseline growth effects of untreated cells within the same time period. It differs from another measure of sensitivity, the IC50, in that it is measured relative to the original population, so that the dependent variable is relative growth, whereas IC50 is measured at the absolute halfway point between the minimum and maximum populations (Figure 2). Thus GI50 actually conflates two effects: 1) the sensitivity of the cell population to growth inhibition and 2) the actual expected growth seen in the cell line.

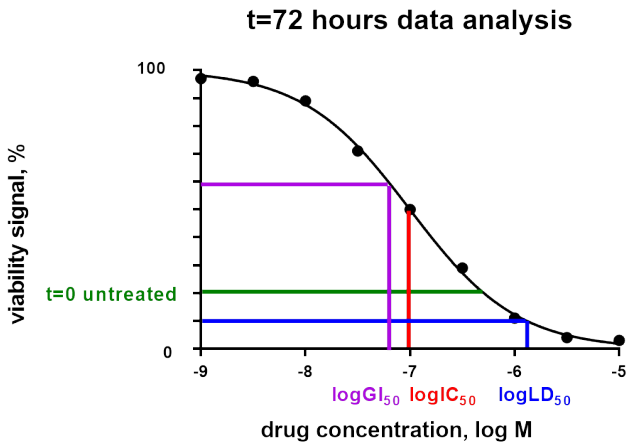


Figure 2. Definition of GI50 (purple). Reproduced from http://www.ntrc.nl/technlogies/oncolinestm

1.4 Cell Lines as a Model System

Cell lines are an important system for testing and understanding the effects of mutations on drug sensitivity. Cell lines are tumor-derived immortalized cells used to study

tumors *in situ*. Neve, et al describe the breast-cancer cell line collection that is used in this study.[17] At the time, this collection consisted of 51 different breast-cancer cell lines, 11 of which are considered HER2 positive. In the Neve paper, it was shown that clustering gene expression data from these 51 cell lines clearly separates this panel into the 4 subtypes (HER2, Luminal (A+B), Normal-like, and Basal) predicted by PAM50. Between these molecular subtypes, focal Copy number analysis shows copy number gains and losses between the luminal and Basal tumors. The HER2 enhanced group in particular was shown to be unstable genomically.[18] However, no analysis of other genomic types such as Methylation or Mutation data was incorporated in this study.

Cell lines by definition are not a complete tumor, nor are completely representative of the tumor *in vivo*. Tumors themselves show high degrees of heterogeneity, and thus one sample may not be representative of the tumor cells (additionally, it has been seen that cell lines themselves are heterogenous). Also, tumors *in vivo* are dependent on the stroma (or supportive tissue), such as vascular tissue and connective tissue, and receive many intercellular signals such as growth factors from this microenvironment.[3] As we will see, we this behavior can be replicated somewhat through introduction of signaling ligands such as EGF. Another complication is that because genomic instability is a feature of some cancer cells, passaging may introduce large genomic changes in the cancer cell lines, resulting in a lack of reproducibility.[19,20] Additionally, selection effects due to passaging may introduce mutations that may not be representative of cancer patients. Thus, caution must be made when interpreting cell line results.

1.5 Drug Sensitivity of Cell Lines

The current state of the art in terms of relating genomic features to drug sensitivity is that different cancer subtypes exhibit differential specificity. That is, intrinsic subtypes are

associated with sensitivity to drugs. Heiser, et al identified subnetworks in a "Superpathway" model associated with drug sensitivity within the panel of cell lines.[21] Three comparisons were done using non-parametric ANOVA: 1) Luminal versus Basal, 2) Luminal versus Basal and Claudin low, and 3) HER2 amplified versus non HER2 amplified. Of the 74 compounds studied, 23 of these had subtype specific responses, meaning that GI50s compared using the ANOVA comparisons were significant for one of the three comparisons.

Within the HER2 amplified subtype, seven compounds were identified as being HER2+ specific: Lapatinib, Gefinitinb, GSK2126458 (PIK3 A/B/D/G), BIBW 2992, GSK2119563 (PIK3CA), and 17-AAG (HSP90AA1). Interestingly, another EGFR inhibitor, Erlotinib, was not called as HER2+ specific, but as basal-specific, having shown significance in comparison 1 (Luminal versus Basal), but not in comparison 3 (HER2 versus non-HER2). Additionally, within the HER2+ subtype, a Beta-Catenin subnetwork was identified as downregulated using the integrative "Superpathway" approach.

There are drugs for which intrinsic molecular subtype is clearly predictive of sensitivity, such as the EGFR/HER2 inhibitors Lapatinib and BIBW2992. Heiser et al also identified subnetworks of genes whose expression is statistically associated with drug sensitivity. However, many of the drugs profiled have sensitivities that are not associated with subtype.

Intrinsic subtype specificity provides much, but not all of the explanation to drug sensitivity in breast cancer. The question is whether we can improve on these features. In order to answer this question, we will need to review current thinking about organizing principles of mutations, notably the classification of driver versus passenger mutations, and the role of rare/unique mutations that are observed in individuals.

1.6 Genomic Alterations and Cancer

Largely, genomic alterations associated with cancer fall into two camps: 1) Oncogenes, which promote tumor growth when altered, and 2) Tumor Suppressor Genes (TSGs), which suppress tumor growth normally, but when altered lose this function.[9] Examples of Oncogenes include PIK3CA (part of the PI3K protein), AKT, Myc, and HER2.[9] Examples of tumor suppressor genes include Rb, which controls a critical transition in the cell cycle, p53, which controls the apoptotic response to cellular damage, and PTEN, which is a negative regulator that dephosphorylates a critical component of PI3K signaling, PIP3.

Oncogenes arise from two main categories of mutations: 1) structural rearrangements, and 2) mutations that lead to elevated expression of the oncoprotein (i.e., the protein product produced from an oncogene). The first category of mutations includes point mutations, and gene fusions. The consequences of point mutations are many, the most common being increasing the activity of the protein. For example, a mutation in Ras at residue 12 increases the ability of the mutant Ras protein to phosphorylate its downstream targets, Ras, Raf, and Ral, resulting in sustained activation of mitogenic (cellular growth) signaling.[9] The point mutation may result in premature truncation of a protein sequence. Truncation of regulatory domains in a protein may make that protein constitutively active, phosphorylating its target in the absence of external mutation, which is observed in a HER2 truncation mutation.[9]

One subtle effect of oncogene mutations is that the affinity of an oncoprotein for its substrate may be changed. This mutation may actually increase the protein's affinity for another substrate, thus potentially rewiring the regulatory network. These effects may be subtle, but have system-wide consequences. One such example is mutations in PIK3CA actually increasing affinity for IRS1.[22] Such mutations are often called "gain of function" or "gain of interaction" mutations.

Gene fusions arise from genomic rearrangements, where portions of chromosomes are translocated, fusing with other chromosomes. One famous example of such a gene fusion product is the BCR/ABL gene fusion, which results from the reciprocal translocation of parts of chromosomes 9 and 22, a translocation called the Philadelphia Chromosome. This translocation occurs in over 95% of Chronic Myelogenous Leukemia (CML) patients and thus the BCR/ABL fusion product made an ideal target for the targeted drug Gleevec.[9]

The second category of oncogene mutations, overexpression, can arise from subtler mechanisms. Overexpressed proteins may result from: 1) copy number gains from genomic rearrangements, 2) a mutation in the promoter sequence that increases transcription factor binding affinity and thus increases expression, 3) promoter/gene fusion resulting from a genomic rearrangement.[9] In the case of HER2 overexpression, the overexpression is clearly from copy number gain. A hyperactive protein may have differing systemic effects than those of having an overabundance of that protein.

One large difference between oncogenes and tumor suppressors is the underlying genetics. Oncogenes are largely heterozygous mutants, indicating that the mutation is dominant.[9] In this case only one copy of a mutation is necessary for the gene to become oncogenic. Tumor suppressor genes on the other hand, require both copies of the gene be disabled in order for the tumor to arise, the so-called "two-hit" model of Tumor Suppressor Genes.[23,24] There are multiple routes to disabling a tumor suppressor gene. The first is through direct mutation of the protein sequence, which can cause a loss of function mutation, or a nonsense mutation, which causes premature truncation of the protein sequence. An alternate route to loss of protein function is through methylation of the gene's promoter sequence, which effectively silences expression of that gene. Methylation of the promoter sequence results in a

number of steps, still unclear to a large degree, that lead of histone deacetylation, resulting in a configuration that blocks transcription. By blocking transcription of an allele of a gene, the function of that gene product is essentially lost.

Still another route to loss of tumor suppressor function is through loss of heterozygosity (LOH), in which a gene replaces an active copy of the gene that is lost with an inactive copy, resulting in two inactive copies.[9] These copies are often lost through chromosome loss that results from the inappropriate segregation of chromosomes during mitosis, resulting in two chromosomes from one parent and none from the other, a phenomenon called uniparental disomy (UPD). UPD is commonly observed in many cancers.[9] Loss of heterozygosity can also occur through gene conversion by DNA repair mechanisms such as mismatch repair.[9]

Finally, one copy of the tumor suppressor gene may be dysfunctional due to a mutation at the germ-line level and inherited from the parents. This inherited gene mutation means the child has a predisposition to cancer, as another loss of function mutation (from LOH, methylation or direct mutation) can lead to cancer.

One crucial difference between oncogenes and tumor suppressor proteins are their druggability. The majority of small molecule drugs are considered to be inhibitors of proteins. Oncogenes are considered to be druggable possibilities because their hyperactivity or overexpression can be possibly inhibited by a targeted inhibitor.[7,9] Tumor Suppressor genes on the other hand, are associated with loss of function and restoring this function is often beyond the scope of small molecule drugs. However, understanding the effect of tumor suppressor genes is critical to understanding the regulatory controls that cancer cells must defeat in order to proliferate.

Oncogenes and tumor suppressor genes may be seen as two sides of the "regulatory coin". One example of this complementary regulation is the proto-oncogene PI3K and the tumor suppressor gene PTEN in the PI3K/AKT pathway. PI3K is responsible for phosphorylating PIP2 to PIP3, which plays a role in the activation of AKT, whereas PTEN dephosphorylates PIP3 to PIP2, essentially acting as a regulator of AKT signaling. It has been observed that either a mutation in PIK3CA, the gene that encodes the catalytic subunit of PI3K, or PTEN is required for tumor cells to proliferate. However, having both is unnecessary. Mutations of this type are called "mutually exclusive."

Some key differences between oncogenes and tumor suppressor genes and the experimental platforms and methods available to detect them are briefly summarized below (Table 1) with further detail regarding detection discussed in Chapter 2.

Table 1. A Comparison of Oncogenes and Tumor Suppressor Genes.

	Oncogenes (structural)	Oncogenes (overexpression)	Tumor Suppressors
Function	Encourages Cell Growth	Encourages Cell Growth	Represses Cell Growth
Mechanism	Mutation increases activity/substrate affinity	Mutation increases expression	Loss of function results in loss of negative feedback control
Genetics	Dominant (only one alteration required)	Dominant	Recessive (requires two alterations to lose function)
Druggability	May be Druggable	May be Druggable	Usually Not (hard to restore function)
Genomic Alterations	Point Mutations, Genomic Translocations (Gene Fusions)	Copy Number Gain, Point Mutations in promoters, Promoter/Gene fusions	Pick two: Inherited familial allele, LOH, methylation, somatic mutation
Technological Platforms	Next Generation Sequencing, RPPA	NGS, Expression Microarray, SNP/Tiling Arrays, RPPA	NGS, Methylation Arrays, RPPA
Methods	Exon Mutation Analysis, Breakpoint Analysis, Structural Rearrangement Analysis, RPPA analysis	Copy Number Analysis, Breakpoint/Structural Analysis, Promoter Mutation analysis	Methylation Analysis, Exon Mutation Analysis, LOH detection

1.7 Systemic Breakdowns in Cancer Tumorigenesis

As we have seen in Weinberg's Hallmarks of Cancer papers, tumors require multiple capabilities in order to proliferate and metastasize throughout the body. Thus tumorigenesis is a process that can in some cases takes decades, with subpopulations of cells within a tissue slowly acquiring these capabilities.[9] These subpopulations gain slight growth advantages over neighboring cells, and thus go through clonal expansion, becoming a larger portion of the tissue. Small subpopulations of these clonal expansions will then acquire new hallmarks, thus selecting for these new subpopulations, resulting in clonal expansion of these subpopulations, until cells that have acquired all necessary capabilities arise.[9,25] We revisit Weinstein's hallmarks of cancer to highlight what cellular systems are affected, as well as examples of oncogenes and tumor suppressor genes in each system (Table 2). We will focus on the roles of three of these systems: Mitogenic (Growth) signaling, the Cell Cycle System, and the Apoptotic system. Other systems,

such as the immune system and inflammatory systems are important as well; however, it is beyond the scope of this work to include them.

Table 2. Hallmarks of cancer and systems they affect along with associated oncogenes and tumor suppressors.

Hallmark/Enabling Characteristic[3,4]	**System Affected**[9]	**Oncogenes/Tumor Suppressors**[9]
Proliferative signaling	Mitogenic (Growth) Signaling	PI3K, Akt, Ras, Myc, Cadherins, Integrins
Evade Growth Suppressors	Cell Cycle	Rb1, CDK inhibitor genes, Cyclins, Myc
Resist Cell Death	Apoptosis	p53, Akt, NFkB, MDM2, Bcl-2, CASP3, CASP8, PTEN, STAT3, ERK, RAF, PI3K
Enable Replicative Immortality	Telomerase	hTERT
Induce angiogenesis	Wound Response	VEGF, MMPs
Activate Invasion/Metastasis	Inflammation	Cadherins, NFKB, beta-Catenin
Genomic Instability	DNA Repair	PCNA, BRCA1, BRCA2
Reprogram Energy Metabolism	Metabolism	Anaerobic Metabolic Pathways
Evade Immune Response	Immune Response	TATA, TSTA, TAP1, FASL
Tumor promoting inflamation	Inflammation	Cadherins, NFKB, beta-Catenin

RTK signaling. In making the decision to proliferate or not, cells are partially dependent on external signals from their neighbors. These extracellular inputs may be growth factors (GFs), such as EGF or HER2, which can be secreted from similar cells or from nearby cellular macrophages, which can be recruited through the inflammatory response.[9] Other inputs include the Cadherins, which govern the strength of intracellular attachment, integrins, which sense components of the extracellular matrix (ECM), and additional signaling pathways. We will concentrate on the receptor tyrosine kinase (RTK) pathways, which govern mitogenic signaling through growth factors.

The main mechanism for the transduction of signal through the proteins in the EGF RTK pathway is through phosphorylation of Tyrosine residues through kinase domains on the

proteins involved in the pathway (Figure 3). One thing to note is that each of the kinases is highly specific, recognizing only a specific tyrosine residue flanked by a three amino acid sequence.[26] This specificity is accomplished through protein domains called SH2 domains, which contain binding sites for both the phosphotyrosine and the flanking 3 amino acid sequence.[26,27]

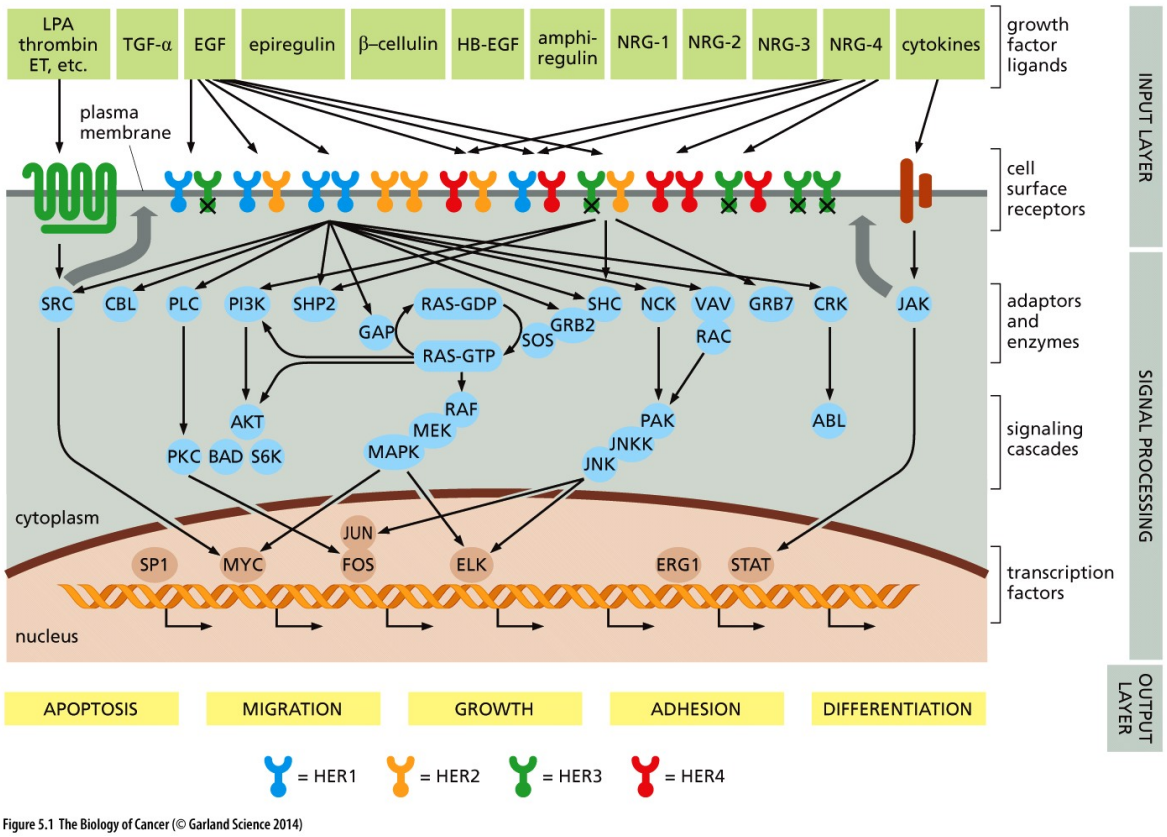


Figure 3. Illustration of the EGF signaling pathway. Reproduced from Weinberg.[9]

As can be seen from Figure 3, extracellular ligands, such as EGF, NRG-1, and HER2 arrive at the membrane, where they bind to the EGFR, HER2, and HER4 receptors on the cellular membrane (HER3 contains a defective receptor domain and thus does not detect ligands, but does participate in signaling). These receptors pair up to produce homodimers, where the two units are the same (such as EGFR/EGFR or HER2/HER2) or heterodimers, where the two units are differing receptors (such as EGFR/HER2, or HER2/HER3).[9] Moasser has suggested that the

HER2/HER3 heterodimer is crucial in HER2 positive signaling.[28,29] When ligands activate the receptor, specific phosphorylation sites are phosphorylated through mechanisms such as transphosphorylation, where kinase portions of the receptors phosphorylate sites on the other receptor in the dimer pair. Other kinases, such as PI3K, and SRC become recruited to activated receptor phosphosites through specific recognition of their SH2 domains to that phosphosite. In turn, these kinases then phosphorylate either other kinases, or substrates such as PIP2, activating these downstream proteins. Still other proteins, such as SHC or GRB2 act as bridges or adaptors between the receptor complex and other proteins, containing pairs of SH2 domains that bridge the receptor to the other protein.[9]

Two signaling cascades that are downstream of the receptors and thus can be activated are the MAPK pathway and PI3K/Akt pathway. Both of these pathways are regulated by the actions of Ras, an oncogene that is frequently mutated in many cancers. The actions of the MAPK cascade directly lead to the activation of two key transcription factors, Elk and Myc, responsible for cell proliferation.[9] The actions of the PI3K/Akt pathway are many, as Akt affects key proteins responsible for protein translation (S6), anti-apoptotic signals (Bad, Caspase-9, FOXO1, MDM2), proliferative signals (GSK-3B, FOXO4, and p21Cip1), and anti-growth signals (TSC2). The PI3K/Akt pathway is one of the most highly mutated of all pathways, with mutations in PIK3CA, the oncogene observed in 18-40% of Breast cancer patients, and loss of function mutations observed in the tumor suppressor PTEN, in 20-33% of breast cancer patients.[9,30]

Downstream of these two cascades are a number of transcription factors that govern the immediate gene response (IEG) and delayed gene response (DEG) of the EGF signaling pathway.[9] These transcription factors include SP1, Jun, Fos, Myc, STAT3, Elk and Erg, each of which are responsible for a specific transcriptional program governing specific decisions such as

growth, apoptosis, migration, adhesion, and differentiation. In the case of the delayed gene response, these expression programs are dependent on the ribosomal synthesis of new transcription factors to start or repress their transcriptional programs.[9]

Additional controls exist for each of these pathways. mTOR is a protein, when forming a complex with with either the Raptor, or Rictor protein, that exerts negative control over the PI3K pathway by controlling available AKT.[9,31,32] In HER2 signaling, Moasser has noted that expression of HER3, a necessary component of the HER2-HER3 receptor heterodimer is negatively regulated when HER2 is inhibited, resulting in increased expression of HER3.[29] Additionally, proteins may be sequestered from the cell surface via endocytosis as an additional form of regulation.

In terms of targeted drug treatments, the EGF system is highly attractive for a number of reasons (Figure 4). A large number of oncogenes occur within this system, suggesting that normal function could be resumed by inhibiting the oncogene products. RTKs are highly specific for their substrate, meaning that their catalytic clefts are very specific for their phosphosite targets, which means that designing highly specific small molecule drug targets are possible.[7] Specificity of these small-molecule drugs is achieved by optimizing their structures such that they bind to multiple residues within their catalytic clefts.

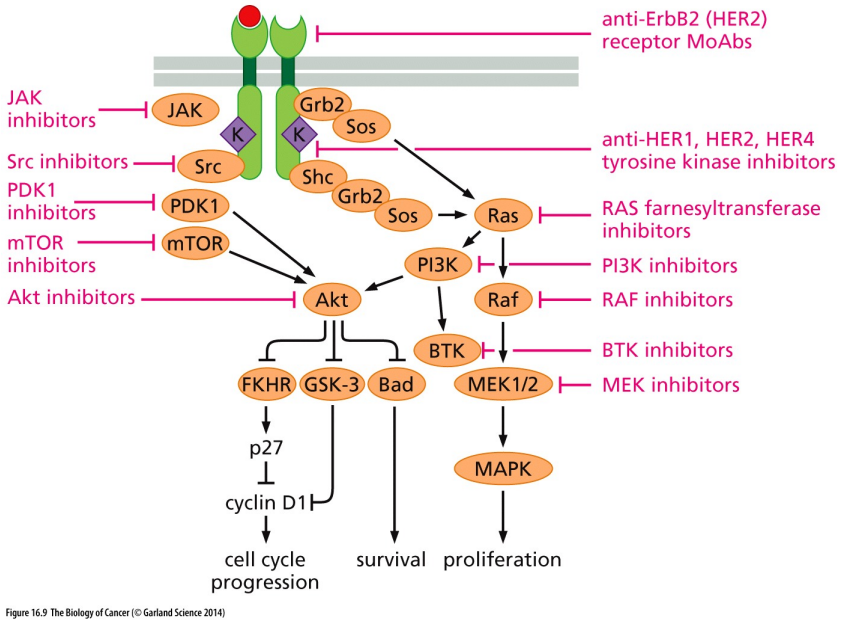


Figure 4. Inhibitors available for the EGF signaling system. Reproduced from Weinberg.[9]

Cell-cycle system. The Cell-cycle system governs the various stages of mitosis, or cell replication and division (Figure 5). Cell cycle activity is governed by the complexing of two complementary protein types: The Cyclins (A, B, D, and E) and the cyclin-dependent kinases (CDKs).[9] Cyclins provide the key timing portion of the cell cycle clock, as they are expressed in cyclic patterns according to the phase of the cell cycle. CDKs are additionally regulated by sensor proteins such as $p21^{Cip1}$, which are known as CDK inhibitors, which inhibit the CDKs/Cyclin complexes and thus the cell cycle in response to physiologic stresses, DNA damage or extracellular signals such as TGF-beta.[9] Other CDK inihibitors, such as $p21^{Kip1}$ and $p27^{Kip1}$ integrate signals from the mitogenic pathways.

In going from a G0, or quiescent state, to a G1, or mitotic state, cells must make a decision to proceed to the replication, or S phase of mitosis. Once replication starts, mitosis must occur, so this decision is a critical one for the cell. In other words, the critical transition in mitosis is the G1 -> S transition, which represents an “all or nothing” decision whether the cell is

going to proceed with DNA replication, and thus cell proliferation. The R-point within the G1 phase marks the irrevocable decision point past which an entire series of cycle-specific transcriptional programs will occur or not, and thus actions such as genome replication, chromosome segregation and cellular division will occur (figure 5). The decision point for this transition is governed by the tumor suppressor gene Rb, whose function is deactivated by hyperphosphorylation, which is initiated by the Cyclin D1-CDK4/6 complexes (which act as an integrator for mitogenic signaling, see below) and then finished by the Cyclin E-CDK2 complex. Deactivation of Rb causes a number of transcription factors called elongation factors (E2Fs), to activate or repress specific transcriptional programs whose expression is required by the G1 -> S transition. If Rb1 function is lost, progression through the cell cycle can continue unabated despite accumulated damage to the cell such as large genomic rearrangements. This is why a large number of cancers have lost Rb1 function. Unfortunately, the cell cycle system is by large undruggable, as it is impossible to restore Rb1 function with small molecule drugs, though CDK inhibitors do exist.[9]

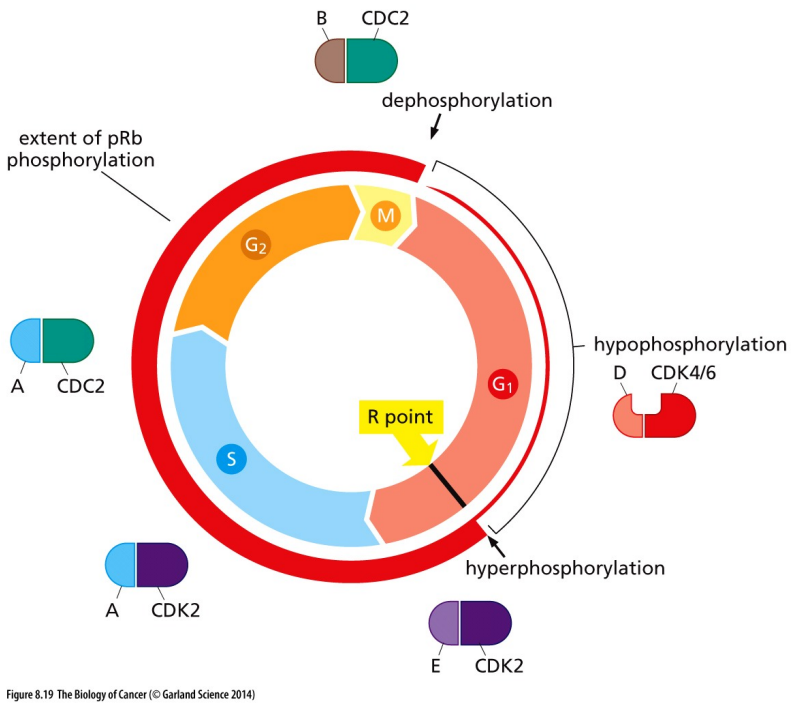


Figure 5. Schematic diagram of Cell Cycle and the R decision point. Reproduced from Weinberg.[9]

Apoptotic system. The apoptotic system governs when a cell should die. There are many reasons for inducing cellular death, and these reasons fall into extrinsic (extra-cellular) causes, such as tissue remodeling or metabolite availability, and intrinsic (within-cell) causes. We will focus on intrinsic causes first. In many cases, cellular processes such as segregation of daughter chromosomes during the G2 phase can fail, meaning that the daughter cells are potentially damaged and unviable, resulting in the activation of the apoptotic response.[9] Large genomic rearrangements (such as observed in tumor cells) and other DNA damage can also trigger the apoptotic response. The apoptotic pathway responds to this irreparable damage in the cell and as a result induces cell death of unviable cells.

The main governor of intrinsic (that is, within-cell) apoptosis is the protein p53, which is a Tumor Suppressor Gene. p53 is a short lived protein that is negatively regulated by MDM2, which ubiquitinates p53, and subsequently makes it a target for protein degradation.[9] p53 acts

as a sensor for various conditions of damage and its function is partially controlled by E2F expression that results from progression through the cell cycle. E2F overexpression is associated with loss of pRb control, and thus loss of cell cycle control.[9]

p53 is a tumor suppressor gene, which means that both copies in a cell must lose function in order to promote tumor development. Loss of function of both copies of p53 through mutation, methylation silencing, or LOH is one of the most frequently observed genomic alterations across all cancer types.[9] By losing p53 function, tumor cells lose the ability to detect cellular damage, and thus can evade the apoptotic response.

Cells may also receive external signals such as from cytokines such as Tumor Necrosis Factor alpha or FAS-ligand, which are then integrated by the extrinsic apoptotic pathway. Depending on the microenvironment of the tumor, surrounding cells may attempt to induce apoptosis through the secretion of these ligands.[9] Evading such external death signals is another capability needed by the tumor.

Downstream of p53 is an additional source of apoptotic signaling, Bcl-2, which integrates other cellular stress signals and can either promote or inhibit apoptosis. Bcl-2 is considered pro-survival whereas p53 is considered pro-apoptotic.[9] Additional inputs include the pro-apoptotic proteins Bim, Puma, tBid, and Bad. The precise balance of these inputs and pro-survival proteins such as BCL-2, BCL-XL, Bcl-W, Mcl-1 and A1 determines whether a cell is programmed to live or die in response to each of these inputs.[9] Downstream of Bcl-2 are a series of Caspases, which begin the process of breaking down the cell, including opening up mitochondrial channels to release the cytochrome-c into the cytoplasm, which assists in breaking down the cell.[9]

There are many redundancies and feedback controls built into the apopotic system, which unfortunately make it a hard system to target.[9] Again, because p53 is a tumor suppressor gene, restoring function is difficult with the current scope of drug therapies.

Cross-phenotype response between systems. One issue with targeted therapy is that there is a lot of *cross-phenotype response* between each of these three systems. As can be observed in Table 2, and Figure 6, a gene itself may participate in multiple systems, meaning that perturbing that gene with a targeted therapy will potentially affect the outputs of each of the systems it participates in. This multi-trait influence of a single gene is also known as pleiotropy. For example, the oncogene Ras has a role in evading apoptosis, mitogenic independence, angiogenesis, and metastasis/invasion.[9]

Another example of cross-phenotype response is Cyclin D1, a protein that actually integrates mitogenic, or growth inputs into the Cell Cycle Response. Cyclin D1 actually serves as a sensor for mitogenic response, as its transcription is controlled by the response mitogenic pathways such as MAPK and the PI3K pathway to external growth inputs such as growth factors.[33] Thus the mitogenic response partially governs the progression of a cell through the cell cycle through Cyclin D1.

Intersystem cross-phenotype response may play a large part in the phenomenon of acquired resistance, where tumor cells develop a resistance to initially effective therapies. Again, small subpopulations of tumor cells may acquire mutations that enable them to bypass signaling pathways that are shutdown by targeted therapies. These mutations again give them an advantage over other tumor cells and thus the tumor will again resume growth through clonal expansion. Additionally, long-term feedback built into these systems can provide compensatory mechanism for increasing levels of the inhibited protein, thus resulting in a

paradoxical response. Thus, if we are rationally to apply a targeted therapy, it is first critical to characterize the cross-phenotype response in the systems that react when a therapy is applied, in order to determine whether whether that therapy should be applied and its role in treatment. Additionally, characterizing system cross-phenotype response can potentially highlight long-term feedback loops that may need to be defeated in order to avoid this acquired resistance. One such route to avoiding acquired resistance could be combination therapies of targeted drugs, in order to induce total shutdown of these systems.

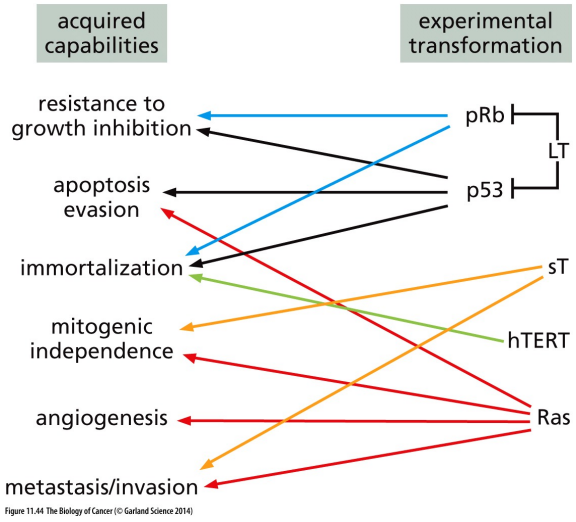


Figure 6. Examples of pleiotropic oncogenes that participate in multiple systems, resulting in cross-phenotype response between these systems. Reproduced from Weinberg.[9]

Oncogene Collaboration. Additionally, there is the phenomenon of oncogene collaboration, in which two or more oncogenes may have a synergistic effect on the system output. The classic example of oncogene collaboration is between Ras and Myc. Some mutations in Myc are insufficient to transform a cell into a tumor like phenotype. However, when combined with a Ras mutation, which affects mitogenic signaling, survival time is greatly reduced.[9] Many examples of "ras-like" (components of mitogenic cascades) and "myc-like" (nuclear proteins) oncogene collaboration exist, such as Pim1 ("Ras-like") and Myc, Notch-1

("ras-like") and E1A ("myc-like").[9] However, oncogene collaboration may have subtler effects that increases the viability of cells through regulation.

Many examples of cross-phenotype response and collaboration can be observed in the systemic diagram of Figure 7. The interactions between four systems, the mitogenic system, invasiveness, growth inhibition and differentiation, and cell survival can be seen to be highly connected in this diagram. However, much of the cross-phenotype response between this systems has not be completely characterized, nor has the true number of collaborations between oncogenes required to promote tumor development been completely characterized.

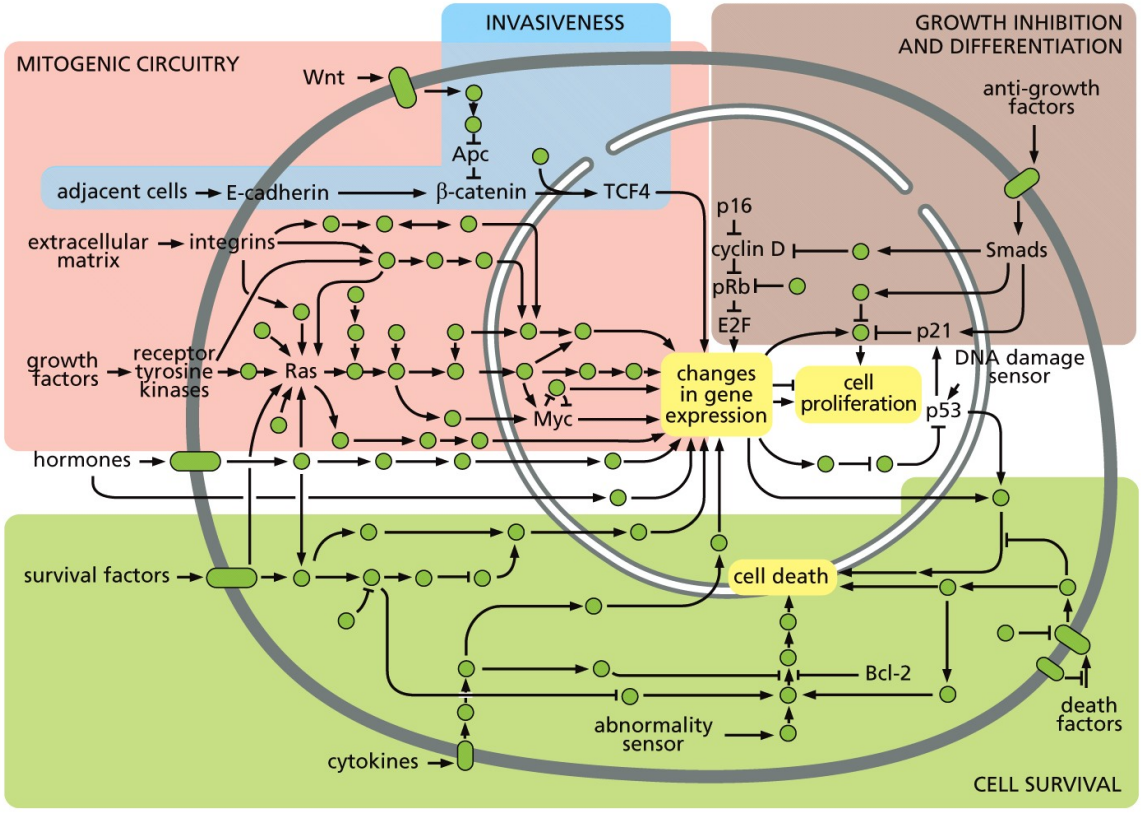


Figure 11.45 The Biology of Cancer (© Garland Science 2014)

Figure 7. Oncogene collaboration within multiple systems that govern extracellular response. Reproduced from Weinberg. [9]

The consequences of cross-phenotype response between cellular systems and oncogene collaboration make targeted therapy much more difficult than simply inhibiting the protein product of an oncogene. Cross-phenotype response means that inhibiting an oncogene product can affects multiple systems, some of which are compensatory. These compensatory effects may arise from negative feedback systems built into the pathways, which can result in a paradoxical response. For example, application of Lapatinib, a HER2/EGFR inhibitor can actually result in the increase in HER2 expression in the long-term.[28] Additionally, cross-phenotype response between signaling systems means that receptors (and thus other extracellular inputs) from other signaling pathways may affect the response to a targeted drug such as Lapatinib.[34]

Thus, *characterizing the degree of cross-phenotype response* when a targeted inhibitor is applied is an important problem. *Finding potential oncogene collaborations* is also important and can drive the decisions of which targeted treatments should be used. These two problems form the basis of my dissertation aims.

1.8 Two Problems involving Drug Sensitivity

There are two problems involving the genomics of cancer and drug sensitivity that I have identified as the focus of this work.

1) What is the actual targeted drug response and how much cross-phenotype response does it induce between systems?
2) What is the contribution of patient-unique mutations to drug sensitivity? Are they organized collaboratively?

These two problems motivate my aims:

Aim 1: Characterize the cross-phenotype response between cellular systems in response to targeted drugs.

Aim 2: Identify potential oncogenic collaborations that can inform the application of targeted drugs.

1.9 Problem 1: Characterizing cellular system cross-phenotype response in response to targeted drugs

In this section, I will discuss the current state of phosphoprotein research and attempts to characterize the phosphoprotein response to targeted drugs such as Lapatinib. Characterizing this response is complicated by many aspects, including 1) Cell Line Specific Responses, 2) Microenvironment, and 3) mechanism of the targeted drug and 4) the scheduling of drug dosage.

1.9.1 Characterizing the cross-phenotype response due to targeted drugs

Using antibodies that target proteins and their phosphorylated states, the dynamic response of proteins that participate in signaling cascades can be identified and quantified. Most importantly, key pathways and specific proteins and their phosphorylated states can be identified that are altered in response to a targeted therapy. This is important in identifying key protein outputs that drive drug response as well as identifying subnetworks of proteins involved in transducing the drug response from its target.

Additionally, it is important to establish the Reference Behavior Pattern (RBP) of the cells before any dynamic modeling can occur. As defined by Sterman, a RBP is a set of time series data that describes the system of interest and describes the development of the problem over time. This RBP serves as a reference during model development, describing the essential aspects of the system behavior that the model needs to capture.[35] Thus, the RBP defines the

scope of the model and acts as a reference during the model verification phase to ensure that the model is behaving correctly, and can act as an explicit check during the model validation process.

Western blotting is the gold standard for measuring protein and phosphoprotein expression levels.[36] However, it is only generally done on the level of 6-12 proteins. However, a new technology called Reverse Phase Protein Arrays (RPPA), allows for the querying of hundreds of proteins. While not on the scale of Gene Expression Microarray technology for the transcriptome, RPPA allows for the simultaneous measurement of protein levels across hundreds of conditions, such as time points, stimuli, and inhibitors. The RPPA data and its implications for characterizing the drug response are discussed in depth in Chapter 2.

10.9.2 DREAM8 Data: A source of cross-phenotype response data

One rich source of data for identifying cross-phenotype response given targeted drugs is the RPPA data provided by the DREAM8 Breast Cancer Network challenge. The DREAM8 challenge was proposed in 2013 with 3 specific challenges: 1) Network Inference, given protein time series data, 2) Time-course prediction, given the same time-series data and 3) Visualization of the protein time-series data.[37]

As we will see in chapter 2, the dataset provided for this challenge is a potentially rich dataset for characterizing protein cross-phenotype response in that it provides protein time-series in response to 4 inhibitor conditions, for 4 phenotypically different cell lines, under a number of different extracellular stimuli, including extracellular ligands such as EGF, FGF, Insulin and serum conditions.[37]

1.10 Problem 2: A New Model of Oncogenic Collaboration

In section 1.7, we have mentioned some simple two-gene models of oncogenic collaboration, notably the "Ras-like" and "Myc-like" type of collaborations. However, motivated by visualizing the mutation data onto protein-protein interaction (PPI) networks, we propose a new type of oncogenic collaboration, that of network-based collaboration. Mutations visualized on PPI networks show a preferential grouping around known oncogenes, such as BRCA1, a protein implicated in breast cancer, and ESR1, the estrogen receptor. This observation shows potential for incorporating many genes we refer to as "passenger" mutations into understanding drug sensitivity. Such passenger mutations are often unique to the individual, but are not considered because they are rare.

1.10.1 Driver versus Passenger Mutations: A False Dichotomy?

The dominant thinking about mutations in cancer is that mutations can be divided into two categories. The first category is Driver mutations, which are higher frequency (in the population studied) mutations that enable one of the capabilities discussed in Weinberg's 'Hallmarks of Cancer' paper and thus confer a selective advantage to the tumor.[3,4] The second category consists of 'passenger mutations', or evolutionarily neutral mutations that occur along with the driver mutations but confer no selective advantage to the tumor. One explanation for these mutations is because many driver mutations occur in proteins that participate in the DNA damage response, such as BRCA1 or XRCC1, and thus these proteins cannot correct mutations introduced by replication errors or externally damaging factors such as UV radiation.[38]

Driver mutations are often found using a frequentist approach. This approach utilizes multiple cancer samples, whether direct biopsies or derived cell lines, frequently occurring mutations are counted. These mutations may be summarized at multiple levels. The most detailed level of summarization is at the level of protein residue, followed by the domain level,

the transcript level, and the gene level, and the pathway level.[39] Each level of summarization provides potential insight into the organizational principles of the mutations.

However, for an individual, the non-synonymous group of mutations contains a large number of mutations unique to that individual. Using frequentist methods, these unique mutations will never be assessed as possible driver mutations. An open question remains whether these are truly passenger mutations, or whether groups of these unique mutations may affect the cell. The question is whether these mutations are truly passengers, or whether there is another organizational concept such as networks that unifies these mutations. I begin to examine the potential for such an organizational concept in Aim 2 of this work.

1.10.2 Towards New Organizational Principles of Mutations

Many sequencing studies in Cancer have used signaling pathways as an organizational principle for mutations.[40–43] However, one weakness of using pathways is that they cover only a small subset of proteins; thus many mutations that are unique to an individual may not map to them.

Protein/Protein interaction (PPI) networks may provide a more general way to organize mutations. These networks are built from both direct experimental data, such as Yeast2Hybrid interactions, or from the scientific literature. Examples of PPI networks include the Human Reference Protein Database (HPRD) and STRING, among others. [44,45] In Aim 2, I focus on a new methodology for integrating mutations using PPI networks.

1.11 Organization of this Dissertation

The remainder of this dissertation is structured around the two aims. In Chapter 2, we will discuss characterizing the drug response and resultant system cross-phenotype response using the RPPA technology. In Chapter 3, we discuss a new model of oncogenic collaboration,

that of network mutations, and the integration of multiple datatypes (mutation calls, drug sensitivity, and copy number calls) required to visualize this collaboration. In Chapter 4, we discuss general issues with the data and subsequent consequences to our approach, and future directions. Finally, in Chapter 5 we discuss conclusions.

Chapter 2: Aim 1 – Assessing the Protein Expression Response and Cross-phenotype response to Targeted Drugs in HER2+ Cells

2.1 Research Question

In this chapter, I will present my attempt to answer the following questions:

What is the phosphoprotein response to a targeted drug such as Lapatinib? What cross-phenotype response exists between cellular systems? Can we predict this cross-phenotype response using simple linear network models?

These questions lead to my Specific Aim, which is:

Aim 1: Characterize the cross-phenotype response between cellular systems in response to targeted drugs.

Within this aim, I have a number of subaims:

Subaim 1-1: ANOVA analysis of a multiple cell line dataset in response to Lapatinib

Subaim 1-2: Characterizing pathway activity and cross-phenotype response using the Area Under the Curve (AUC)

Subaim 1-3 Visualizing RPPA data onto Networks and Pathway Diagrams

Subaim 1-4 Timecourse Prediction of the cross-phenotype response drug response using AUCs

In section 2.3, I discuss the necessary background in order to justify these subaims.

2.2 Introduction

Characterizing the protein response to targeted drugs in cellular systems is an important task, as inhibiting a targeted node can have unforeseen consequences, such as compensatory subsystems. We have discussed that this is an important feature of cellular systems is due to the pleiotropic nature of their genes. Thus characterizing the cross-phenotype response that is the result of targeted therapies is an important step in understanding drug resistance and overcoming this drug resistance through combination therapies.

In this chapter, we outline an approach for characterizing systemic cross-phenotype response response to targeted therapies. The main data we use to characterize the response is time-series proteomic data, in particular data generated using an antibody-based technique called Reverse Phase Protein Arrays (RPPA). We will see that RPPA is analogous to a high-throughput ELISA technique that enables us to query the drug response over a number of conditions (such as stimuli and drugs) and timepoints.

We utilize both statistical techniques and a robust summarization metric for the time series data known as Area Under the Curve (AUC). Using AUCs, we can quantify the dynamic response of the individual proteins in order to gauge pathway activity and thus quantify the degree of cross-phenotype response across these cellular systems. We apply AUCs into two different applications: 1) quantifying cross-phenotype response across cell lines stimulated in Lapatinib, a HER2/EGFR inhibitor, and 2) predicting cross-phenotype response within individual cell lines given combinations of extracellular stimuli and inhibitor pairs in the DREAM8 data.

2.3 Background

In this section, we will first discuss the Reverse Phase Protein Array technology (Section 2.3.1), how the data are generated, and caveats to interpreting the RPPA data. We will see that

the largest drawback to RPPA data as it is used in our datasets, is that it generates relative, and not absolute concentrations, and that expression values between antibodies cannot be directly compared.

We then will discuss current research on drug sensitivity in cell lines, focusing on Lapatinib (Section 2.3.2). We will finally discuss a summarization method for time series data, called Area Under the Curve (AUC) in Section 2.3.3 and its potential for summarizing pathway activity and thus cross-phenotype response between the cellular systems probed by the RPPA data.

2.3.1 Reverse Phase Protein Arrays

Tibes, et al. describe the basic design and validation behind Reverse Phase Protein Array (RPPA) experiments.[46] RPPA is a medium-throughput technique that can be thought of as a highly parallel ELISA (Enzyme Linked Immunosorbent assay)[47], allowing the investigator to interrogate protein and phosphoprotein expression across a wide array of conditions and proteins.[48] These conditions can include time course experiments under differing stimuli and drugs.

Overview of RPPA process. With regard to the RPPA process (Figure 8), cell lysate representing samples of interest is first processed to produce 5 2-fold serial dilutions, producing a series of sample dilutions that span from 0 (no dilution) to 1:16 (16-fold dilution). The reason for the dilution series is two-fold: 1) it ensures that at least some of the dilution values are within the antibody's dynamic range and 2) it aids in the reduction of random error in estimating the relative concentration (see below).[49] Each dilution is then treated with glycerol to a 30% concentration. The reason for the glycerol is to prevent spreading of the sample dilutions when they are spotted onto the slide.[50] These sample dilutions are then plated onto a set of 364 well

plates (called the master plate) that are then placed into a plating machine that then places the samples onto a series of nitrocellulose backed slides along with two different sets of positive controls, a control lysate, and a set of 48 standards (the rationale behind these positive controls will be discussed below). About 30 ug of lysate derived from either patient tumors or cell lines is required to produce the slides to be probed with the standard antibody list of 250 antibodies. Each antibody slide has a unique identifier that aids in Quality Control.

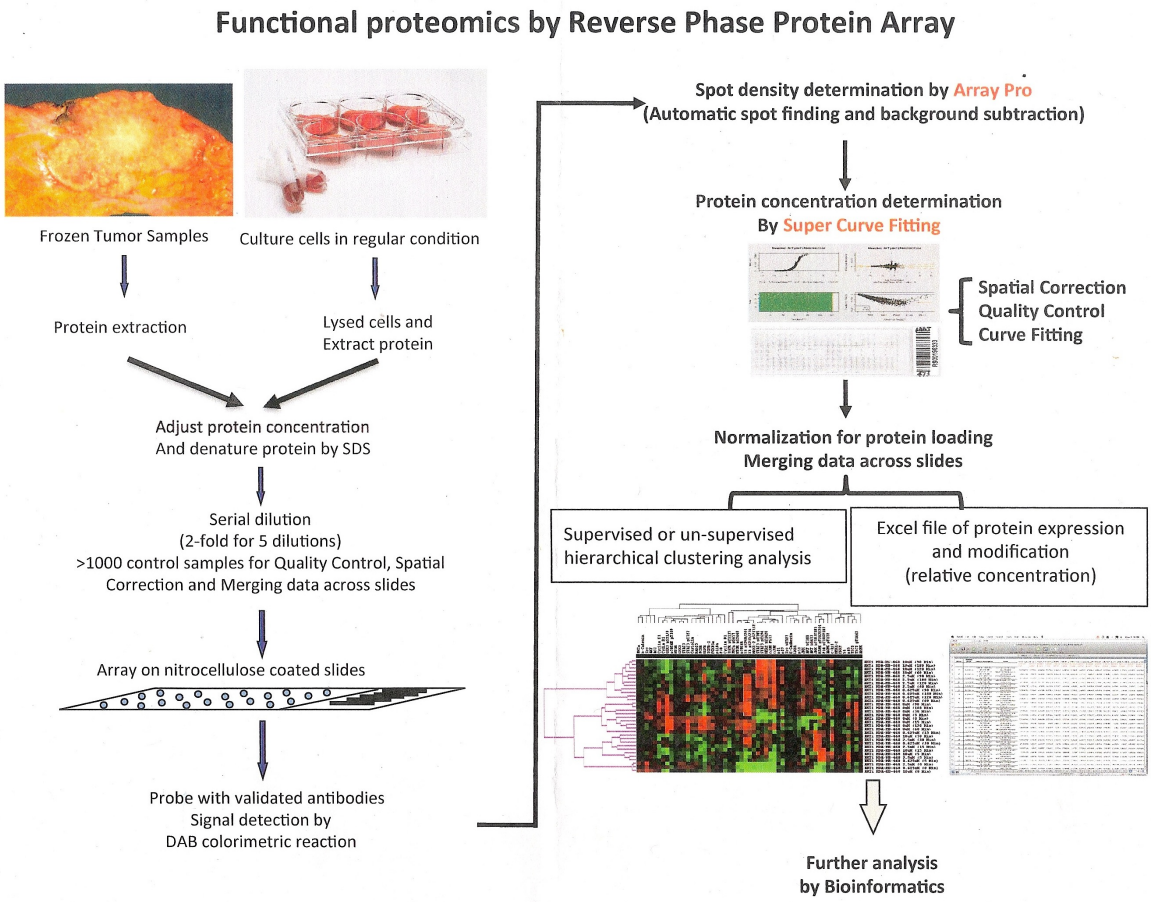


Figure 8. Overview of RPPA process. Reproduced from http://www.mdanderson.org/education-and-research/resources-for-professionals/scientific-resources/core-facilities-and-services/functional-proteomics-rppa-core/rppa-process/index.html

A single slide is then probed with a reporter antibody that is sensitive and specific to the protein or phosphoprotein state of interest. This process is then repeated over all the

antibodies of interest. The reason for spotting samples (reverse-phase) instead of the antibodies (forward-phase) to the slides is that finding optimal incubation conditions for probing multiple antibodies spotted to a slide is difficult.[51] These reporter antibodies are then interrogated with a secondary antibody conjugated to BIotin. The Biotin is then amplified using a Dako Cytomation-catalyzed system, and then the amount of amplified Biotin is measured using a DAB colorimetric assay, which precipitates the biotin and the amount of protein is quantified by the change in color observed at the spot.[52] The amplification step is done because it reduces the amount of lysate required on a slide.[50,53] This is one large difference between RPPA as it is used in our datasets and the analogous Gene expression microarray experiments for the transcriptome – the quantitation is not fluorescence but color (Fluoresence can also be used in RPPA, as well as alternative visualization systems such as quantum dots, however the drawback to both of these alternative quantitation systems is that much more protein lysate is required because these systems have no amplification step[53]). The slides are then scanned and each sample spot is quantified and locally background corrected using image analysis software called Array-Pro Analyzer. Array-Pro Analyzer produces sample spot intensities and also attempts to correct for local background issues.[54] The sample spot intensities are then further processed to remove unwanted variation using an R-based package called SuperCurve, which will be discussed below.

Slide Design. Because the slide design contains a number of controls, it is important to discuss the slide layout (Figure 9). Remember that each sample on the slide is represented by a dilution series, and thus is represented as 5 spots on the array. These five spots are placed within an 11x11 grid.[50] These grids are arranged in 4 rows by 12 columns, allowing for a total of 1056 samples represented by 5280 dilution spots on the array with 528 cell lysate controls, for a total of 5808 spots on the array. Each 11x11 grid allows for 22 samples to be placed two in a row, occupying 11 rows by 10 columns. The last column of each grid is occupied by two dilution

series of control lysates and is called a vertical control. This control lysate is derived from 32 cell lines under a wide variety of stimulation conditions. The goal of this lysate is to produce some signal for all of the antibodies probed. The vertical controls are used to correct for spatial variability and also used in the QC process.[49]

The other set of positive controls consist of 48 standard samples that are spotted as dilution series across the array. These samples are derived from a wide range of cell lysates that are again meant to express a wide variety of proteins across a wide range of stimulation and conditions. These cell lysates include MDA-231 cells treated with insulin, MDA-468 cells treated with/without EGF, Jurkat cells treated with/without alpha-FAS (in order to represent the cellular death response), phosphatase treated lysate, and lysates from 30 other cell lines.[53] These standard samples are used to normalize across sample batches (discussed below).

In order to reduce spatial effects, samples for a particular experiment are spotted so they are in close proximity to each other on the slide. Though complete randomization of sample position would be desirable, it is currently difficult to do so given the current dilution process and spotting procedure.

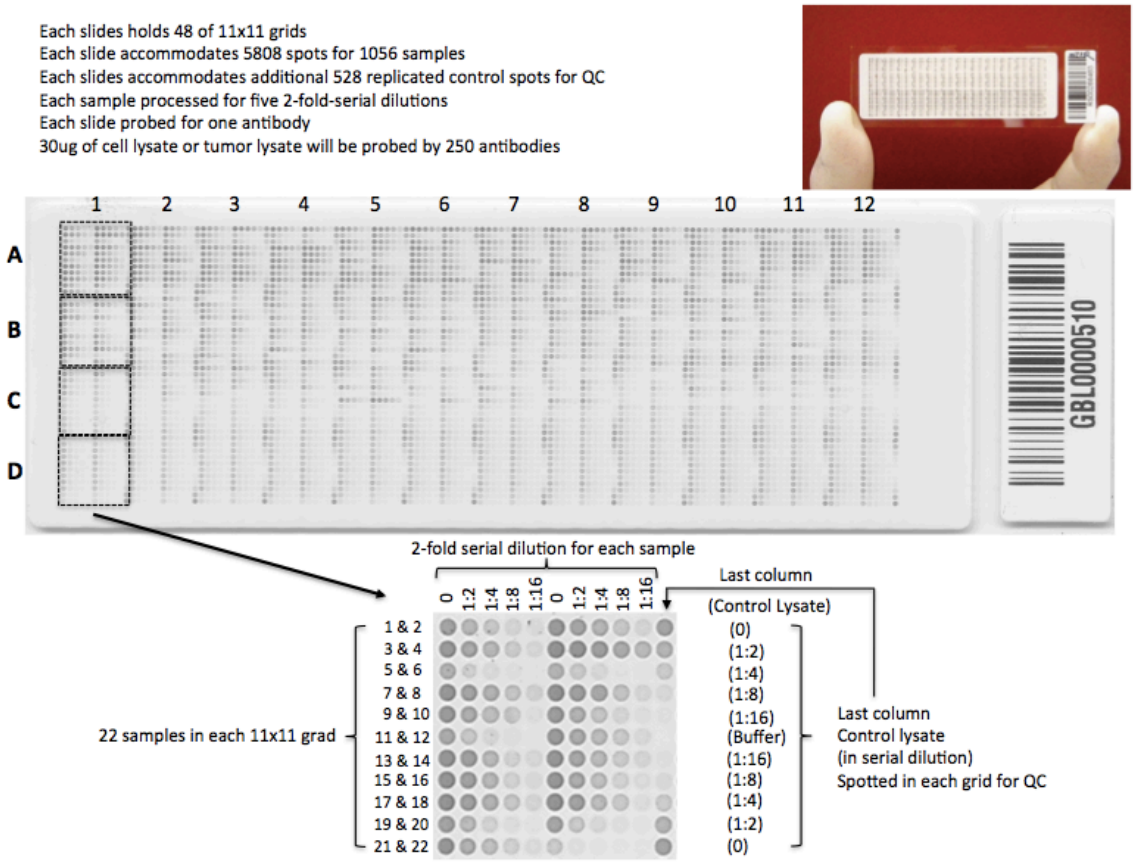


Figure 9. RPPA slide design. The top panel shows the overall slide design. The bottom panel shows the anatomy of a single grid location on the slide. Reproduced from Lu.[50]

QC process. All slides are subjected to a rigorous QC process. The first process is to visually examine both the net expression values and the local background values that are measured using the ArrayPro software. Examination of these values visually can expose experimental issues such as possible smudges on the array. Additionally, a probability measure is derived through the linear combination of 4 quality control factors. These factors are 1) the variability of the positive controls, 2) the linearity of the log dilution line, 3) the observed dynamic range across all samples in a slide, and 4) the deviation in linearity across the 96 horizontal controls, as they should be flat across 96 controls of the same dilution.[55]

SuperCurve. Thus, the color amount from the secondary body provides an estimate for the expression level of protein or its phosphorylated state. The relative concentration for each dilution series is quantified by fitting the color value of all samples on an antibody slide to a calibration curve, or "SuperCurve".[56–58] It is called a SuperCurve because all spots on the antibody slide are used to do this fitting, rather than just using single dilution series. This calibration curve can be 1) a logistic curve or 2) a nonparametric curve whose only assumption is that the curve be monotonically increasing, which is approximated using a quadratic spline.[56]

The logistic curve was initially chosen due to the nature of the colorimetric probes: at high probe concentrations, quenching is observed, leading to a saturation effect, and at low probe concentrations, background noise obscures the true signal.[56] The logistic model is expressed in the following equation:

$$y_{ij} = \alpha + \beta \frac{2^{\gamma(x_i + EC50_j)}}{1 + 2^{\gamma(x_i + EC50_j)}} + \epsilon_{ij} ,$$

where y_{ij} is the observed expression level at dilution level i and sample j, i is the dilution level (1,…5), j is the sample, and $x_i = i - (1+5)/2$, or the dilution level index in Step I, and $EC50_j$ is the median effective concentration of sample j, which is estimated from the dilution series, which is what we desire.[56] By using nonlinear least squares, the global parameters for the SuperCurve, α, β, and γ as well as the EC50s can be found. The final term, ϵ_{ij}, is a random error term for the ith dilution level and the jth sample that is assumed to be normally distributed with mean 0 and variance σ^2.

The nonparametric model on the other hand, is of the form

$$y_{ij} = g\left(x_i + EC50_j\right) + \epsilon_{ij},$$

where g() is defined as a unknown monotonically increasing function. Both g() and the EC50s are iteratively estimated using quadratic splines.[56] An initial estimate for the x_is and EC50s are estimated using a simple linear model of the expression values y_{ij}. In Step 2, the function g is obtained by regressing y_{ij} on $x_i + EC50_j$ for all the data points. In Step 3, the EC50s, are updated by minimizing an objective function. Steps 2 and 3 are repeated until a convergence criterion in the EC50 values is met.

The nonparametric model is now the preferred model used in the quantitation.[50] The nonparametric model fits a larger set of data than the logistic model can, and does not have the large estimation bias as the logistic model.

Potential sources of variability in RPPA. With RPPA, the data is subject to many levels of variability. The first type is variability within a single antibody slide (refered to as *within-slide*). The second level is variability within a sample batch (that is, a set of antibody slides, refered to as *within-batch*), and the final level is variability between-sample batch (referred to as *between-batch*). We will discuss sources of variation within all three levels and how the variation is handled both experimentally and within the SuperCurve software.

For within-slide variability, the first potential source of variability in RPPA has to do with the *antibodies probed onto the slides*. There is an enormous dynamic range observed across antibodies; some produce large amounts of signal, whereas some antibodies' signal is so low that they are lost in the background.[53] Additionally, the specificity of the antibody matters, as non-specific antibodies that may conjugate to other proteins, producing cross target signal that has nothing to do with the desired protein. Thus all antibodies used in the RPPA process are subjected to a rigorous validation process, in which expression of the antibody on an RPPA slide is correlated to expression of the Western Blot.[59] Only antibodies that are highly correlated with

Western Blot expression (Pearson correlation > 0.7) are then validated for use in the RPPA process. Additionally, the western blot must only have a single band, corresponding to a high specificity for the protein.

More forms of within-slide variability are *noise due to random error* and also variation due to *spatial effects* across the slide. Due to the antibody probing and washing procedure, an antibody may stain unevenly across the slide, resulting in spatial effects where certain areas on the slide may be higher in expression than others.[52] However, this is where the vertical controls can help. Because the vertical controls are spotted across all 48 grids in the slide, they form a positive control surface that is used to estimate spatial effects. Spatial effects can be minimized by scaling all grids such that the vertical control values are equal.[51]

Random error is reduced because five points in the dilution series are used to estimate the concentration. Because the assumption of the noise is that it is Gaussian and centered at zero, the deviation of each of the dilution spots from the SuperCurve is expected to average to zero.[56]

Another source of variability in the data at the single antibody slide level is variation due to background noise. The assumption is that the background noise (after local background subtraction) is the same across the entire array. Background is estimated as the median value across all samples within the slide.[49,51] The median value is then either divided (on the linear scale) or subtracted (on the log scale) from all values in the array.

A large source of variability in the data within an antibody slide is due to *protein loading*, in which the amount of protein spotted can vary from sample to sample. This variability can happen due to the dilution process or simply due to the spotting procedure. Because the amount of loading can vary across samples, protein loading can obscure true expression

differences between samples. This variability is minimized both experimentally and in preprocessing. Experimentally, a preloading procedure is done by adjusting the total protein concentration in all sample lysates to 1 ug/uL.[50] By adjusting all lysate samples to the same total protein concentration, a large portion of this variability can be reduced.[50,53] Additionally, after SuperCurve processing, an additional normalization procedure is done to adjust for protein loading. This procedure relies on utilizing the power across all the antibody slides for a particular sample batch to adjust for protein loading.[49,51] The protein loading of a sample is estimated as its median value across all printed antibody slides. This median value is called a Correction Factor (CF), and the sample values across antibodies are adjusted by dividing by this amount on the linear scale (this amounts to dividing all values in a single row in the protein matrix). CFs that are outside of the experimentally determined range of 0.25-2.5 are flagged as being possible outliers. Because protein loading is an issue, other normalization methods such as NormaCurve[60] have been proposed to adjust for protein loading, background subtraction and spatial variation, though they are not used on either of our datasets.

Finally, because the number of samples in an experiment can be larger than the number of spots on a slide, batch effects can be observed.[49] These effects can be due to protein loading, or variation due to sample processing (dilutions across experiment samples can happen on different days, for instance). The current procedure to adjust for batch effects relies on the 48 known standards that are spotted throughout each array and is called Replicate-Based Normalization (RBN). Because the data used in this study has not been normalized by RBN, we will not discuss it here.

Caveats to SuperCurve approach for RPPA. There is an important caveat to this approach to the RPPA data. SuperCurve measures relative expression, not absolute expression

due to the fitting approach. That means that protein expression can only be readily compared within a protein, not between proteins because of the variation in binding affinity across antibodies. Thus additional information about the absolute abundances of the proteins are needed to complete a model, in the form of either a known protein standard dilution curve, or through MS quantitation of the antibody of one of the standards.

Caution should be taken when a set of samples is known to be biased (that is, unbalanced) in expression for a particular protein, as the median centering procedure could be biased upwards or downwards in estimating background. Such a situation could arise if the number of samples is small, or predominantly one cell type and a single condition is run. Additionally, the protein loading adjustment procedure can be biased if the total number of antibodies probed is small (< 20 antibodies).

The datasets used in our study are specifically timecourse experiments in response to targeted drugs such as Lapatinib. We will now discuss previous approaches to understanding the Lapatinib response.

2.3.2 Previous work with Lapatinib

Lapatinib, a small molecule inhibitor that inhibits two Tyrosine Kinase Receptors: HER2, and EGFR, is the one of the key drugs focused on in this Aim. Lapatinib's mechanism of action is through binding reversibly to the ATP binding site of the Tyrosine Kinase domain of EGFR and HER2.[61] Through this binding, it is thought to prevent phosphorylation and thus results in the silencing of downstream signaling events.

Much of the work in characterizing the Lapatinib response largely focuses on gene expression, as gene expression microarrays are readily available. O'Neill et al used microarrays in their study of six cell lines (BT474, SKBR3, EFM192A, HCC1954, MDAMB453 and MDAMB231)

and identified four genes (RB1CC1, FOXO3A, NR3C1 and ERBB3) whose transcription switched from up-regulated to down-regulated in response to Lapatinib after 12 hours of exposure.[62] Additionally, they noted that the expression of CyclinD1, a cell cycle regulator that controls the transcription of these 4 genes, was correlated with the Lapatinib response.

Hegde et al note a similar response in their comparison of the expression of two Lapatinib sensitive cell lines (BT474 and SKBR3) and two insensitive cell lines (MDA-MB-468 and T47D).[63]

A few studies characterizing the protein response to Lapatinib do exist. Imami et al used a mass spectroscopy approach to quantitate the protein response to Lapatinib within SKBR3 cells.[64] Briefly, the advantage of this approach over antibody-based approaches is that many more proteins can be discovered at the expense of sensitivity. Additionally, because proteins are identified at the peptide level (short protein sequences), there are many cases for which a peptide cannot be unambiguously mapped to a protein. Using isotope-specific labeling, they compared three conditions: EGF, EGF + Lapatinib (1 μM) and EGF + Lapatinib (10 μM) at the time points of 0, 1 minute, 5 minutes and 60 minutes. Of the 4953 phosphopeptides identified, 62 of these phosophopeptides mapped to the EGFR/HER2 pathways, and 21 of these phosphopeptides were differentially expressed between the EGF and EGF + Lapatinib 10 uM conditions. Among their findings they identified 8 EGFR and 13 HER phosphorylation sites that were upregulated and downregulated in response to Lapatinib.

One caveat to many of these studies is that they only characterize the early (0-4 hrs) response to the cells; there is evidence that long-term regulation of transcription is also involved in the Lapatinib response.[65,66] We will find evidence of this long-term response in our dataset

that spans a much longer time period, that of 72 hours, and that much of the cross-phenotype response is found in the later response.

2.3.3 Previous Work: Area Under the Curve Analysis

We utilize time-series response data in order to characterize cellular system cross-phenotype response, specifically the total protein and phosphoprotein response as quantified as RPPA data. There are many methods to characterize time-response data, and thus quantify the protein response, pathway response, and thus the cross-system response. Some of these may focus on quantifying the rate at which a protein responds (derivative-based methods), the final state of the protein (steady state analysis), or the shape of the protein response (trend-based) methods.[67]

We focus on quantifying pathway activity by *quantifying the amount of upregulation and downregulation over time* in the protein observed after the drug is applied. However, due to the lack of replication in our dataset and the possibility of spurious datapoints, there was a need for a robust summarization technique. We have chosen to integrate the Area Under the Curve (AUC) as our summarization metric.

Other researchers have shown the usefulness of using Area Under the Curve (AUCs) as a method for summarizing timecourse data. Di Camillo et al used Area Under the Curve (AUC) as a feature in deciding differential significance for time course expression experiments for which replicates were lacking.[68] They compared this method with two other methods for time course data for deciding on differential expression using synthetic data. The first was a threshold-based method that utilized a model of the experimental error, or noise in the data. The second method they compared was a method that utilized spline fitting to compare the time series profiles. The AUC method outperformed these two methods in terms of precision and recall

when the time series was short and sparse, whereas the spline-based methods outperformed AUC and the threshold-based method for data with extended timepoints.

Other approaches that use AUCs note that summarizing time series data with AUCs tends to make these summaries more robust to fluctuations in individual timepoints.[69] This is especially useful with the RPPA time series data, as replicates are not available. Summarization with AUC is helpful in minimizing these fluctuations, such as sudden spikes in a single timepoint.

We will see that by quantifying pathway activity using AUCs that we can identify and quantify cross-phenotype response across cell lines and correlate the cross-phenotype response with drug sensitivity. Additionally, we will utilize AUCs in order to predict cross-phenotype response within four phenotypically distinct breast cancer cell lines using a linear network method known as Partial Least Squares Path Modeling (PLS-PM).

2.4 Methods

2.4.1 Datasets Used

Two RPPA datasets were used in this analysis, which will be referred to as the Lapatinib dataset and the Stimulus/Inhibitor (or DREAM8) dataset.

The Lapatinib dataset consists of 15 cell lines exposed to two conditions: DMSO (control) and Lapatinib (the drug of interest). Multiple timepoints (0.5, 1, 2, 4, 8, 24 and 72 hours) were measured for 131 proteins and their 48 associated phosphorylated proteins. This dataset was used for subaims 1-1, 1-2, and 1-3.

The Stimulus/Inhibitor data set is described in detail in the DREAM8 challenge website.[37] This dataset was used for subaim 1-4, in order to predict cross-phenotype response using PLS-PM. This dataset was done for 4 cell lines: UACC812 (Luminal, HER2+/PR+), BT20 (Basal, Triple

Negative), BT474 (Luminal Type, PR+), and MCF7 (Luminal-Type, ER+/PR+). Briefly, this data consists of multiple inhibitors over multiple stimuli measured over a period of time (0 – 4 hrs). The inhibitors were first applied to the cells, followed by stimuli two hours later, which was considered as the zero timepoint. The training set consists of four inhibitors: DMSO, an AKT inhibitor (GSK690693), a combination of MEK + AKT inhibitors (GSK690693 + GSK1120212), and an FGFR inhibitor (PD173074). Across the 8 stimuli (PBS, Serum, NRG1, FGF1, HGF, EGF, Insulin, and IGF1), that makes a total of 32 conditions in the DREAM8 training set.

The test set consists of five inhibitors: 1) EGFR inhibitor, 2) EGFR + HER2 inhibitor, 3) MTOR inhibitor, 4) SRC/MEK inhibitor and 5) SRC inhibitor. The actual drug names were not supplied to the contestants. The DREAM8 challenge required all the responses to all 5 of these inhibitors be predicted across the 8 stimuli, 40 timecourses total.

2.4.2 Preprocessing of RPPA datasets

Both RPPA datasets were processed using the SuperCurve approach outlined in section 2.3.1. One important thing to note is that some of the antibodies did not pass the correlation test described in Chapter 1, a test of the linearity of the calibration curve. Wherever possible, we will note these cases (described as F, or failing).

Also, it should be noted that the expression values of one antibody cannot be compared with another. However, multiple conditions can be compared within antibodies, and we take advantage of this by using DMSO as a reference. By subtracting the DMSO trace from the Inhibitor of interest, we focus on the relative effect of the inhibitor, which simplifies the overall dynamics of the data.

2.3.3 Methods Sub Aim 1-1: ANOVA analysis in the Lapatinib Drug Set

The Lapatinib data is unique in that it is consists of multiple cell lines (10 HER2 positive, 5 non-HER2) and their response to DMSO and Lapatinib. Examination of the Lapatinib GI50s for the cell lines separated the cell lines into two groups: High Sensitivity (blue values) and Low Sensitivity (white values) (Figure 10A). Within each group, we are interested in proteins whose expression under Lapatinib treatment is statistically significantly different from the DMSO treatment (Figure 10B). By looking within each group, we can treat cell lines as replicates and gain statistical power for this analysis.

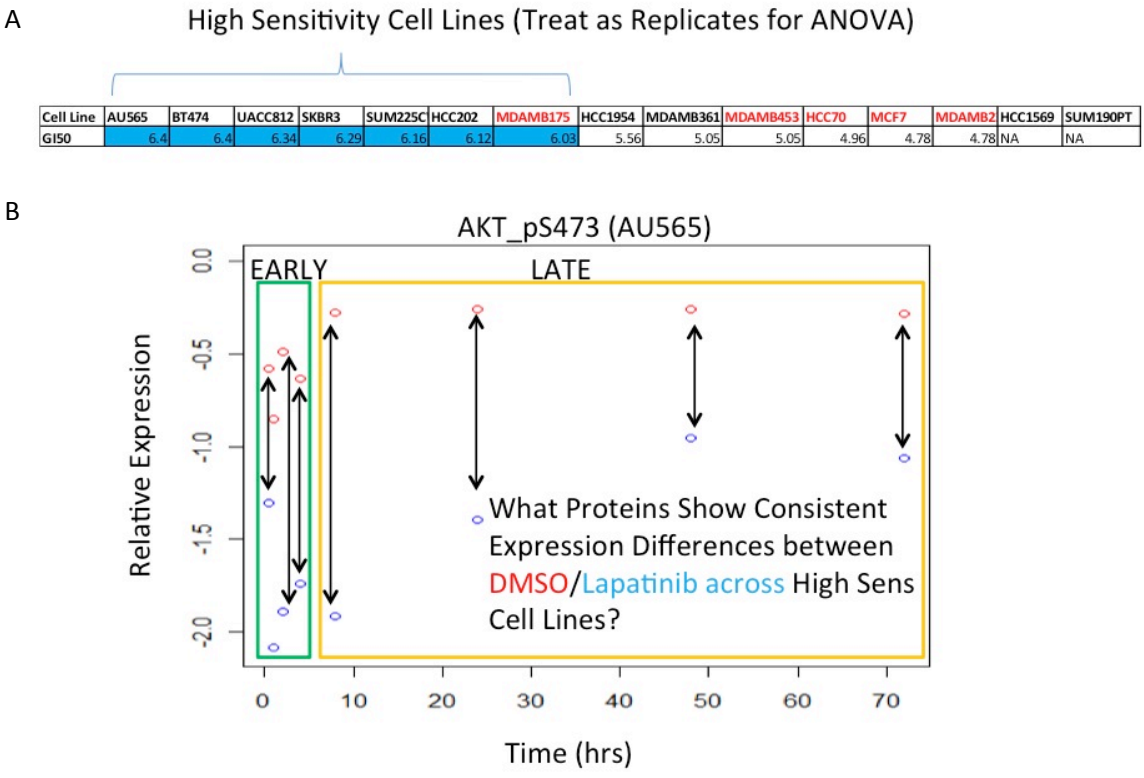


Figure 10. Basic structure for identifying early and late responses in high sensivity HER2 positive cell lines using ANOVA. A) Definition of high sensitivity cell lines (blue) versus low sensitivity cell lines (white) using Lapatinib GI50s. B) Distinction between early and late time series. Time series data was divided into early (green) and late (yellow) timepoints and analyzed separately.

We examined both groups using ANOVA (Analysis of Variance) separately in order to identify groups of proteins that had significant expression differences between the DMSO and

Lapatinib traces. In order to do this, the time series data was divided into early (0 – 8 hrs) and late (8 – 72 hrs) groupings. A separate ANOVA analysis was conducted for each (early and late) grouping.

ANOVA analysis was conducted using the Limma software package for Bioconductor, which takes as inputs a model matrix (which defines the groups to be compared in the ANOVA analysis), and the actual RPPA data.[70] The null hypothesis of ANOVA is that the group means across all groups (DMSO versus Lapatinib) are identical. Because there are two groups, this is equivalent to a t-test comparing means between the DMSO and Lapatinib values. For a list of all packages and their versions, please refer to the session information in the Appendix.

P-values from the ANOVA analysis were adjusted for False Discovery Rate (FDR) using the qvalue package.[71] A q-value cutoff of 0.05 was chosen.

2.4.4 Methods Sub Aim 1-2: Using AUCs as indicators of Drug Sensitivity in the Lapatinib Drug Set

AUCs are a potentially robust method to summarize time series data. By summarizing the data over a time interval, the effect of spurious data points is minimized (Figure 11A).

AUCs were derived by taking the difference between the DMSO trace and the Lapatinib trace for a particular antibody. Thus the AUC represents the relative difference in expression between Lapatinib and DMSO (Figure 11B). Using relative expression helps to emphasize and clarify the true effect of Lapatinib in the system. The actual Lapatinib and DMSO trajectories are somewhat variable across the cell lines, and concentrating on relative expression allows us to somewhat filter out these trajectory differences.

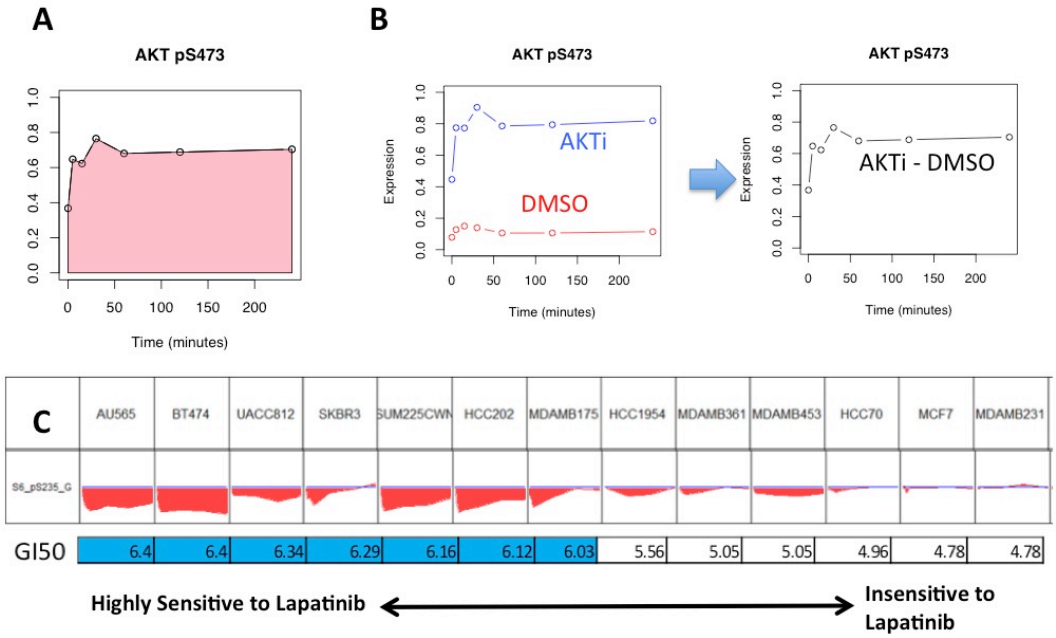


Figure 11. Illustration of AUC in RPPA dataset. A) Integration of upregulation over time in AKT pS473 antibody to produce AUC. B) Subtraction of DMSO trace (red) from AKTi trace to produce trace that shows relative response. C) AUC and drug sensitivity. Visualized AUCs correlate with drug sensitivity, suggesting that the downregulation of S6 phosphoprotein is correlated with GI50.

The Bolstad R package was used to numerically integrate the time series data into AUC values.[72] This package calculates AUCs using Simpson's Rule of integration, which estimates the area under the curve in an interval using a polynomial approximation of the curvature observed in that interval. One weakness in using Simpson's rule to integrate time series data is that if the time intervals used are not are larger than observed spikes in the data, it can potentially underestimate the area. In order to compensate for this underestimation, small intervals were used to estimate the total AUC for a time series. For a list of all packages and their versions, please refer to the session information in the Appendix.

Once the AUCs were calculated for each cell line, AB specific correlations were calculated by calculating Pearson's correlation coefficient between the AB AUCs and the Lapatinib GI50 values across the 13 cell lines for which Lapatinib GI50 data were available

(Figure 11C). We visualized the distribution of correlations across all antibodies in order to identify potentially interesting outputs correlated with drug sensitivity in the tails.

2.4.5 Methods Sub Aim 1-3: Visualization of RPPA Data On Interaction Networks

One method to aid in interpretation of the RPPA data is visualizing the RPPA time courses on a network. Visualizing timecourses on a network is an important intermediary step between the differential abundance analysis and modeling. Possible causal relations between proteins can be examined by comparing the time series between the proteins.

Two network visualization approaches were examined: 1) an undirected network approach, and 2) a pathway based (and thus directed) network approach. Both approaches offer a different view of the data. The directed pathway approach allows us to compare known and well curated interactions to see if the data fits the data, while the undirected approach allows us to compare less curated interactions to see if they are possibly correlated.

For the undirected approach, Protein/Protein Interaction networks were used to visualize the signature of the highly sensitive lines identified in sub Aim 1-1. This network was derived from the Human Protein Reference Database (HPRD), a database of protein-protein interactions in humans. [44,73,74] Antibodies were mapped to nodes (specified by Entrez Gene ID) in this HPRD network by first mapping the AB to HUGO gene symbol and then to Entrez Gene ID.

For the directed approach, two pathways derived from the Kyoto Encyclopedia of Genes and Genomes (KEGG) were used for visualization: the mTOR pathway and the PI3K/AKT pathway.[75] Pathways were loaded from the KGML files using the KEGGgraph package in R. Nodes in the KEGG graphs were mapped to entrez genes, which allowed us to associate proteins with nodes in the network. We used the Rgraphviz package to layout the pathways, ggplot to generate the timecourse plots for individual nodes, and the grid package to place the

timecourse nodes on the graphs. For a list of all packages and their versions, please refer to the session information in the Appendix.

2.4.6 Methods Sub Aim 1-4: Validation of the AUC approach on Stimulus/Inhibitor Dataset

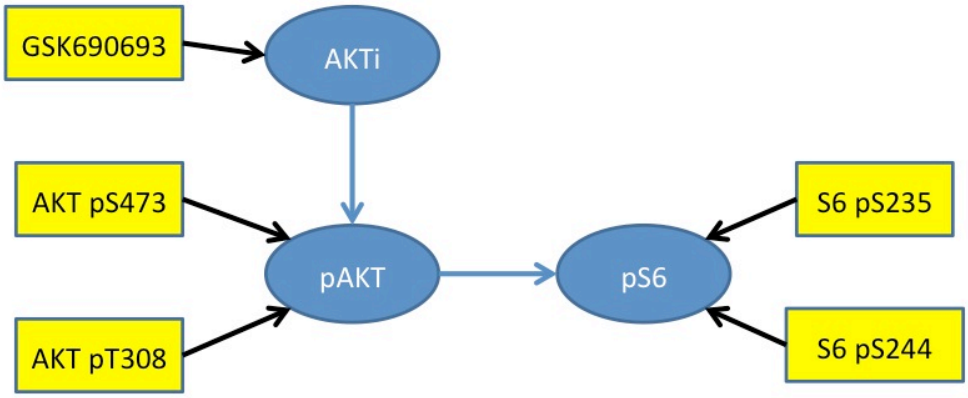
Partial Least Squares Path Modeling (PLS-PM) was used to predict cross-phenotype response (in the form of AUCs) in the Stimulus/Inhibitor dataset. Briefly, PLS-PM is descended from Partial Least Squares approaches, which attempt to summarize multiple variables through the use of weights, and Structural Equation Model approaches, which attempt to detect associations between latent variables in the data.[76,77] PLS-PM is a latent-variable (LV) based approach, meaning that the modeled entities are not measured directly, but through linear combinations of indicator variables. For example, the latent variable "pAKT" may be derived from two different phospho-AKT measurements: pS473 and pT308 (figure 12).

LV approaches are common in the social sciences and marketing, where the latent variables to be modeled might include abstract concepts such as Satisfaction, which are not directly measurable. Thus LV approaches are often called 'soft' in that they do not directly model the observed data, but rather surrogate variables. PLS-PM has also found use in the biosciences in predicting genomic associations with obesity.[78–80] If the Latent Variables are clearly defined and its relationship to the indicator variables is direct, PLS-PM can be a powerful technique. Sorger, et al used a related technique, Partial Least Squares Regression, to relate inputs of a signaling model to known outputs.[81]

Three inputs are required for a PLS-PM model (Figure 12): 1) The inner model, which specifies the linear relationships (blue arrows) between the latent variables (blue ellipses), 2) the data, which consists of physically observed indicator variables (observed data) (yellow boxes

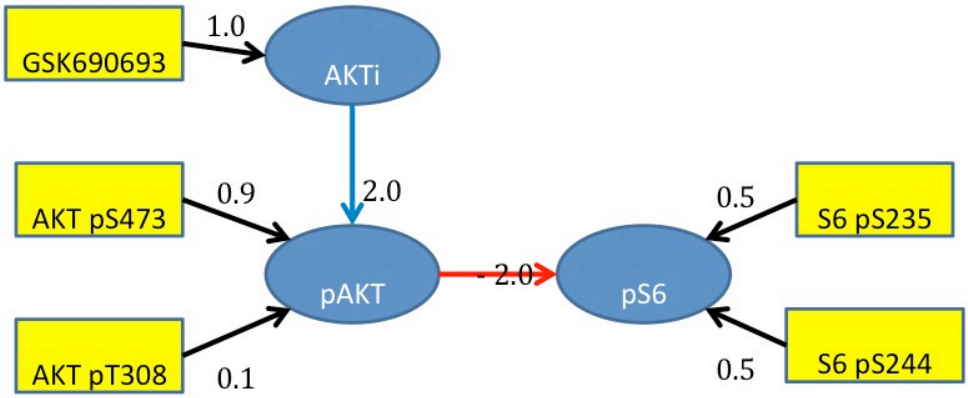
in Figure 12) and 3) a mapping, or outer model, that relates the individual indicator variables to the latent variables. Essentially PLS-PM models ask how strong the linear relationships are between the latent variables described by the inner model.

There are three outputs of a PLS-PM model. The first is the specification of the outer-model weights, or the contributions of the indicator variables to their latent variable. The outer model weights are decided using an iterative hill-climbing procedure that maximizes the correlations between variables and their parents, which are specified in the network. For example, In Figure 13, the weights for the two S6 phosphoproteins are decided by maximizing the correlation between the pS6 and pAKT latent variables. The second is the quantified relations between the latent variables. Each endogenous (that is, variables with parents) latent variable is specified by a linear relationship between its parents. (Figure 13). The last output consists of the values for the Latent Variables themselves. Tenenhaus notes that these values can be of further use in calculating additional non-linear regressions, such as quadratic regression between Latent Variables.



GSK690693	AKT pS473	AKT pT308	S6 pS235	S6 pS244
0	2.1	1.5	0.5	0.1
0	2.3	1.5	0.1	0.3
1	4.1	1.5	-5	-4
1	4.3	1.5	-6	-6

Figure 12. Illustration of Inputs to a Partial Least Squares Path Model (PLS-PM) for genes.



$$pAKT = \beta_{pAKT0} + 2.0 * AKTi$$

$$pS6 = \beta_{pS60} - 2.0 * pAKT$$

Figure 13. Illustration of outputs to a Partial Least Squares Path Model (PLS-PM) for genes.

Estimating the parameters of a PLS-PM model is an iterative process, involving examination of both the weights of the outer model and the strength of the linear relations.

Sanchez describes such an iterative methodology: 1) examining first the outer weights, or loadings and then 2) the R-square values of the inner model, which provide an assessment of the percentage of variance accounted for by the local linear model for that variable.[82] Additionally, by bootstrap sampling the data, a rudimentary measure of significance of each interaction coefficient can be assessed. Based on these values, the modeler may then 1) Add, remove, or transform indicator variables or 2) Add/Remove Latent Variables, or 3) Add or remove interactions in the inner model. This process is repeated until the modeler is satisfied with the fits.[77] Because we only have 24 conditions in the training data (3 inhibitors x 8 stimuli), we could not use the bootstrap approach. Thus, we use the R-square values for the variables in order to assess the model fit to the training data.

Preprocessing the DREAM8 RPPA Data. We subtracted the appropriate DMSO-stimulus condition from each inhibitor/stimulus condition. For example, for the MEKi/AKTi-NRG inhibitor-stimulus condition, we subtracted the DMSO-NRG trace. AUCs were calculated for these relative expression traces using the Bolstad package as in subaim 1-2. The reshape2 and plyr packages were used to summarize the data and calculate these expression differences.

Additionally, the MEKi/AKTi condition was missing for the stimulus NRG1, PBS and IGF conditions for the BT20 cell line. The model was trained without these conditions. The 60 minute timepoint was missing for the BT549 for a number of stimuli for the DMSO condition. The zoo package was used to interpolate these values as part of a time-series [Ref].

Filtering the RPPA data. We assessed activity within the phosphoprotein set by sorting the phosphoproteins by Coefficient of Variation (CV) across all conditions available in the training set. Variability is a common filtering criterion when there is no reference. A large number of phosphoproteins are not active across the multiple conditions in the test data and

thus are not included within the phosphonetwork. The resulting data was transformed into two groups of AUCs: Early: (0-1 hr) and Late (1-4 hrs) using the same Bolstad package used in Aim 1-2. This resulted in two sets of predictions for each cell line: early predictions and late predictions.

Building the Phosphonetwork input. The filtered phosphoproteins are then connected into a network using the HPRD data, resulting in an undirected network. Because PLSPM requires a directed network, each edge was assigned a direction using KEGG pathways if available, or annotated using a reference if not available. We built four distinct networks based on the filtered phosphoproteins for each Cell Line (UACC812, BT20, BT549, and MCF7). Available data was then mapped to these nodes in the network, resulting in the outer map. An additional requirement is the networks needed to be acyclic; i.e., no cycles can be present in the data. We pruned out reactions that would lead to feedback cycles, but these feedback loops could be potentially accommodated using a dynamic version of PLS-PM.[83]

Multicollinearity of indicator variables. One issue that is important to address is the issue of multicollinearity, or correlations between the indicator variables. Multicollinearity can affect the interpretation of the model results by affecting the outer model weights for two collinear variables unequally. For example, the two phosphosites for the pAKT latent variable (pS473 and pT308) might be highly correlated and the weights might be highly skewed towards one of them. This can affect the predictions for model, especially if a resulting weight is very small. Backtransforming from such a small weight can result in a nonsensically large prediction. Thus, it is necessary to assess which phosphosite to include for each model based on phosphosite plots (Figures 24-27).

The plspm package for R was used to do the actual PLS-PM modeling. This package was chosen because it is open-source, and thus the analysis steps are transparent and potentially easy to modify.[84] For a list of all packages and their versions, please refer to the session information in the Appendix.

Generating AUC predictions from PLS-PM models. Because the inner model must be acyclic, the inner model coefficients generated by PLS-PM can be used to predict LV values using a special property of directed acyclic graphs. All directed acyclic graphs can be topologically-sorted by node such that each node is preceded by its parents (Figure 14).[85] If we specify values for the latent nodes that are the exogenous inputs, we can evaluate the values of the other latent nodes by evaluating the network in the order of the topological sort. The predictions for each node are consistent with the local linear models specified by the inner model.

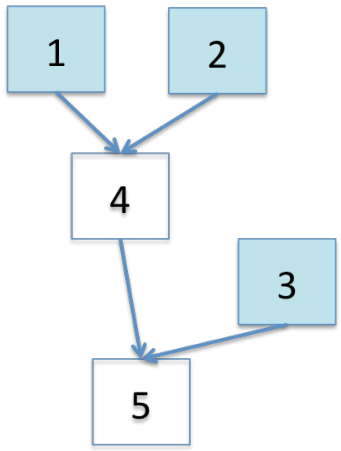


Figure 14. Example of a topologically sorted acyclic network.

In order to predict the values for inhibition, we specify values of the exogenous inputs. Because the inhibitors in the test set were not specified in the model, we model inhibitors as reducing the value of the target node by subtracting 1.5 times the current value of that node.

Backtransforming the AUC predictions. The AUC predictions are relative, not absolute predictions as required by the DREAM8 challenge. The AUCs can be visualized as a box whose area is equal to the AUC over a timeframe of interest. We transform our relative result to an absolute one by dividing the AUC value over the entire interval and adding this value to the DMSO prediction for that particular timepoint. For the 60 minute timepoint, which exists in both AUC calculations, we add an average of the two AUC values.

2.5 Results

2.5.1 Results 1-1 - Identification of a common set of protein responses in HER2+ Cell Lines to Lapatinib

The significant results from the ANOVA analysis can be seen in Tables 3 (Early timepoints) and 4 (Late Timepoints). All antibodies called as significant in the early time series can also be seen in the late time series, with unique antibodies to the late time series highlighted in orange in table 4. An alpha of 0.05 was used to filter both lists.

The early response (Figure 15) is dominated by the phospho-AKT response, which is downregulated in Lapatinib compared to DMSO. This is to be expected, as inhibiting HER2 in this system is expected to have downregulating effects on downstream proteins. An upstream adapter protein, SHC1 is also downregulated in response to Lapatinib. Additional downstream antibodies are downregulated, including the protein translation components S6 and its related kinase S6K. Already within 8 hours we also observe the transcriptional response due to the transcription factors STAT6 and STAT5-alpha being downregulated in response to Lapatinib. Overexpression of STAT5 is associated with Breast Cancer.[86]

Table 3. Early response proteins differentially expressed between DMSO and Lapatinib.

Antibody	GeneName	adj.P.Val	CorrFilter
Akt_pS473	AKT1, AKT2, AKT3	3.67E-09	P
Akt_pT308	AKT1, AKT2, AKT3	3.36E-06	P
mTOR_pS2448	FRAP1	0.001978	F
p70S6K_pT389	RPS6KB1	0.016472	P
S6_pS235_S236	RPS6	2.53E-08	P
S6_pS240_S244	RPS6	0.000113	P
SHC	SHC1	0.01026	M
STAT5-alpha_py594	STAT5A	0.000113	P
STAT6_pY641	STAT6	0.003424	P

In the late response (Table 4), there are additional proteins which are downregulated and upregulated (highlighted in orange). Most of these proteins are downstream of the early

proteins from Table 3, which would support the idea that they are differentially expressed later in the time-series (Figure 16).

Table 4. Late Response Proteins Differentially Expressed between DMSO and Lapatinib.

Antibody	GeneName	adj.P.Val	CorrFilter
Akt_pS473	AKT1, AKT2, AKT3	1.07E-09	P
Akt_pT308	AKT1, AKT2, AKT3	3.54E-08	P
BIM	BCL2L11	0.001189	P
Cyclin_B1	CCNB1	1.04E-05	F
MAPK_pT202_Y204	MAPK1, MAPK3	0.00079	P
mTOR_pS2448	FRAP1 (MTOR)	1.04E-05	F
p70S6K_pT389	RPS6KB1	0.002511	P
Rad51	RAD51	0.040067	P
Rb_pS807_S811	RB1	4.92E-07	P
S6_pS235	RPS6	1.17E-11	P
S6_pS240	RPS6	1.60E-09	P
SHC	SHC1	0.00013	M
STAT5-alpha_py594	STAT5A	2.59E-05	P
STAT6_pY641	STAT6	0.002213	P
YB-1_pS102	YBX1	0.002511	P

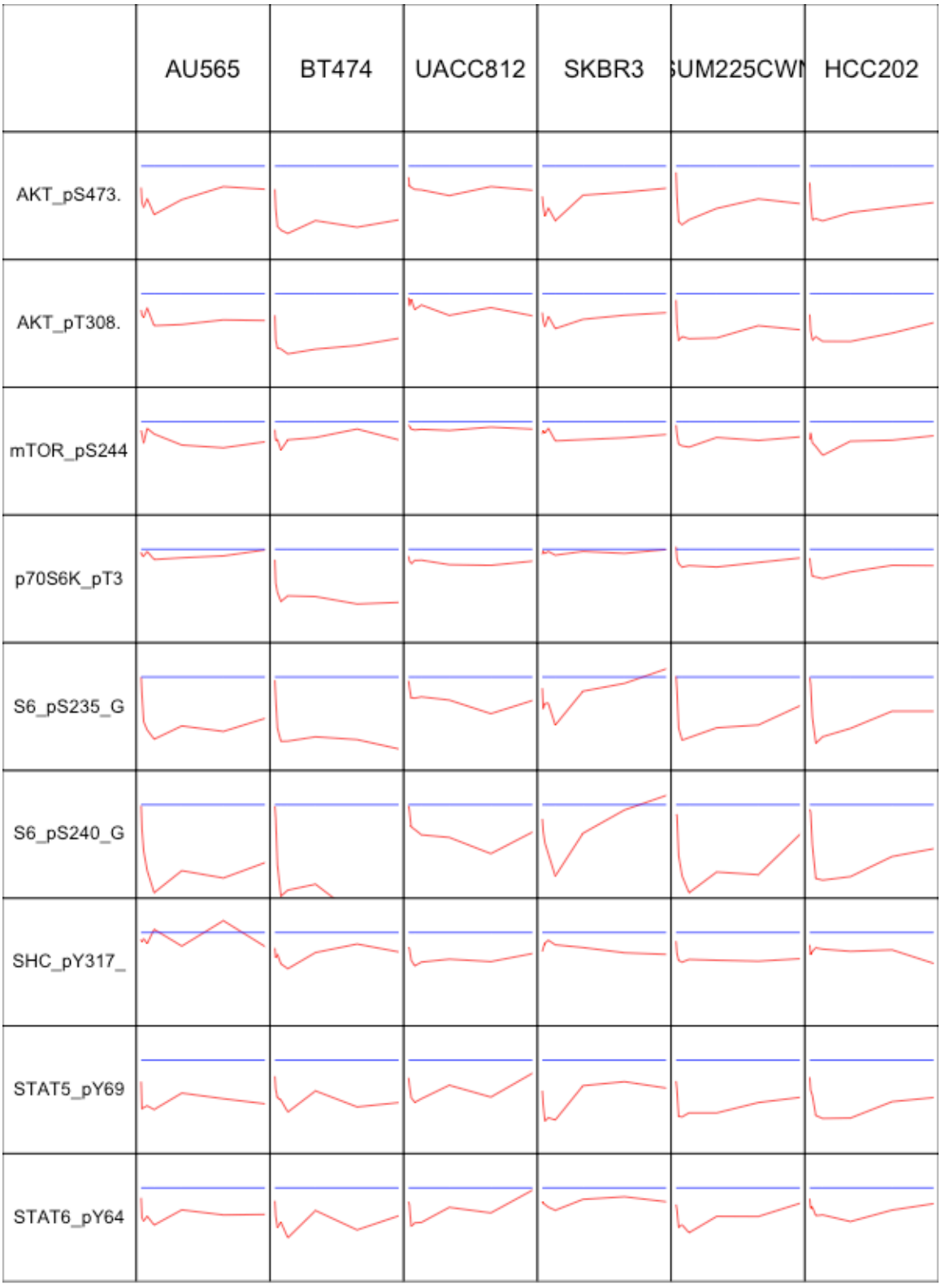


Figure 15. Early Expression Candidates from ANOVA analysis across the sensitive cell lines. The expression scale is identical and ranges from -3 to 1. Time scale (x-axis) is identical across all traces and ranges from 0 – 72 hrs.

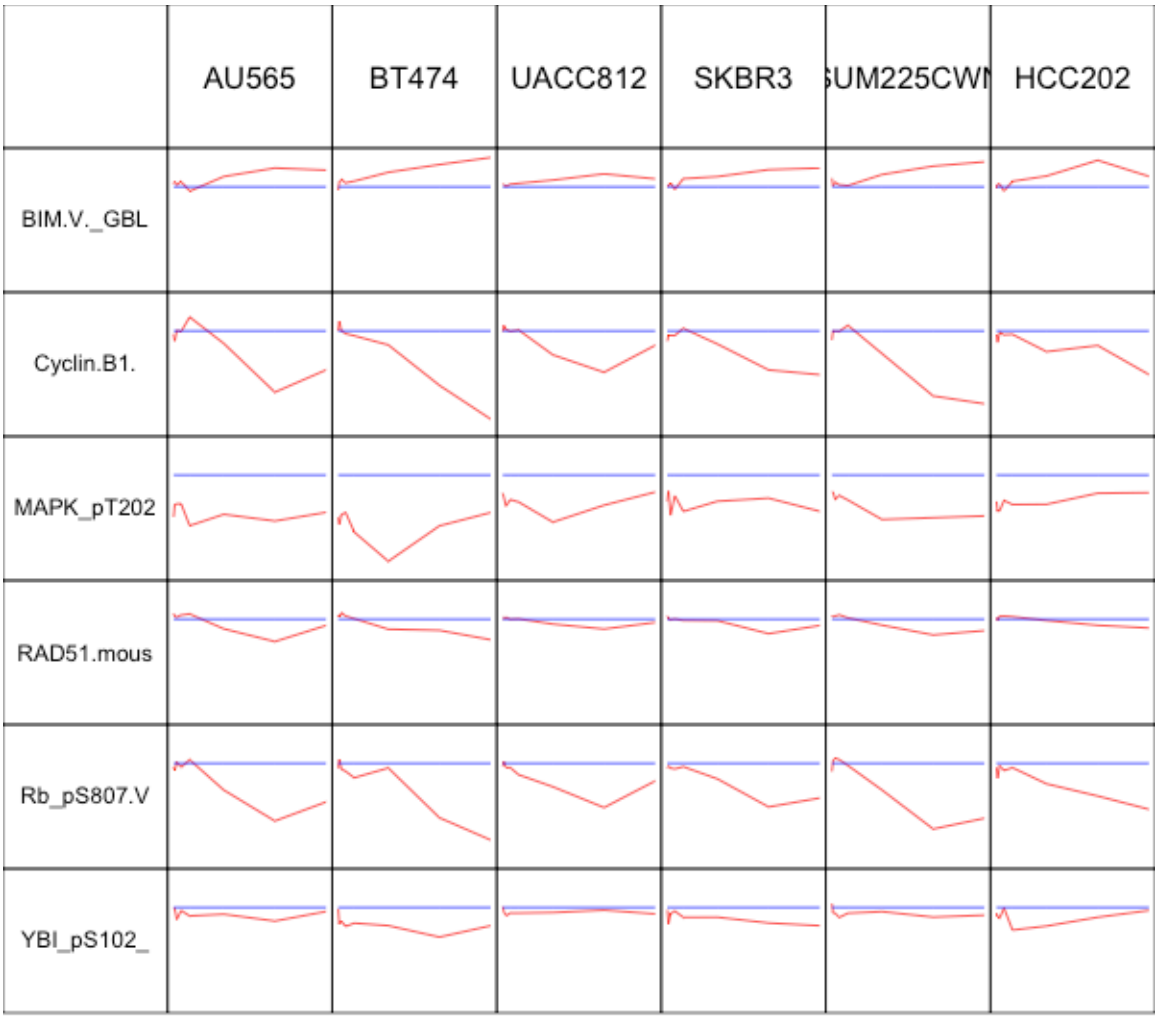


Figure 16. Late expression candidates across the highly sensitive cell lines. Scales are identical to figure 15.

2.5.2 Results 1-2: Identification of key proteins correlated with GI50 response across cell lines

The distribution of correlations for all antibodies with the lapatinib GI50s across the thirteen cell lines appears to be bi-modal (figure 17). We examined the top 5 antibodies that were correlated and top 5 antibodies that were anti-correlated with Lapatinib GI50 (Table 5).

The top anti-correlation candidate, RB1 pS807-811, showed a high amount of anticorrelation (-0.87, Figure 18). Downregulation of this phosphoprotein is highly anti-

correlated with the GI50 response. RB1 is a protein that is involved in regulating the cell cycle response, and thus affects the growth of cancer cells (see section 1.7).

The top correlation candidate, the protein BIM, is highly correlated with GI50 (0.815, Figure 19). Note that this is an antibody for the total protein. Upregulation of this phosphoprotein is associated with apoptosis, another outcome that can affect the growth of cancer cells (see section 1.7).

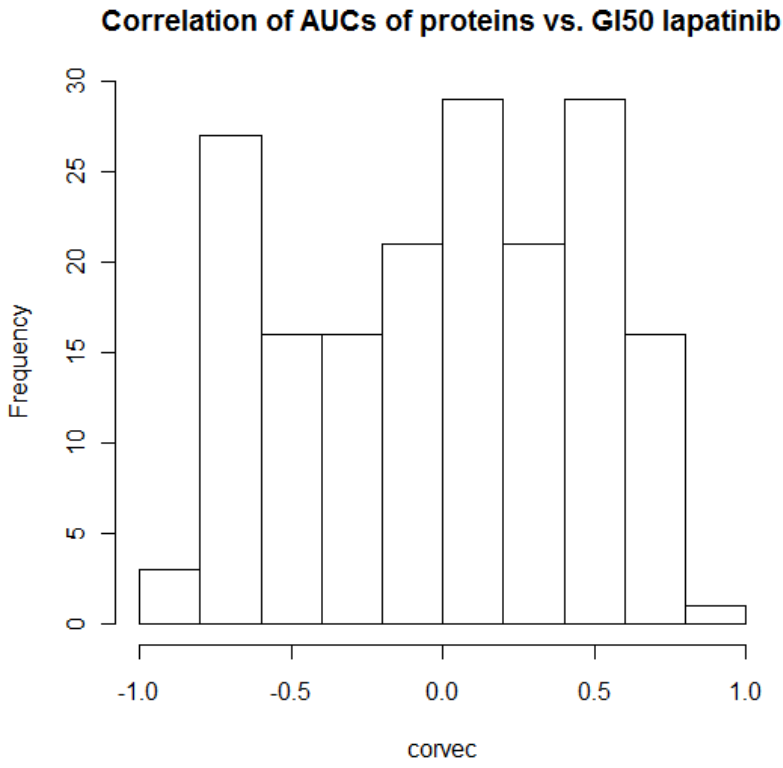


Figure 17. Distribution of Correlations between AUCs and GI50s for all antibodies in RPPA dataset.

Table 5. Top 5 correlated and anti-correlated antibodies with GI50.

Antibody	Correlation	Function
Rb_pS807	-0.871	Cell Cycle
Cyclin.B1	-0.817	Cell Cycle
MAPK_pT202	-0.801	Signal Transduction
S6_pS240	-0.795	Protein Translation
STAT6_pY641	-0.786	Transcription Factor
Caspase.9.Cleaved.Asp315	0.706	Apoptosis
Cyclin.E2	0.713	Cell Cycle
N.Cadherin	0.735	Cell Adhesion
SRC	0.737	Signal Transduction
BIM	0.815	Apoptosis

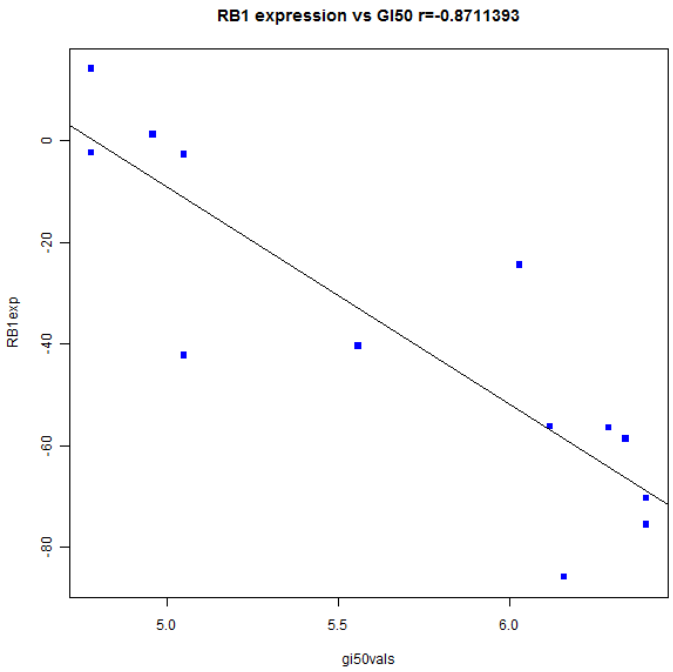


Figure 18. RB1 AUC is highly anti-correlated with GI50.

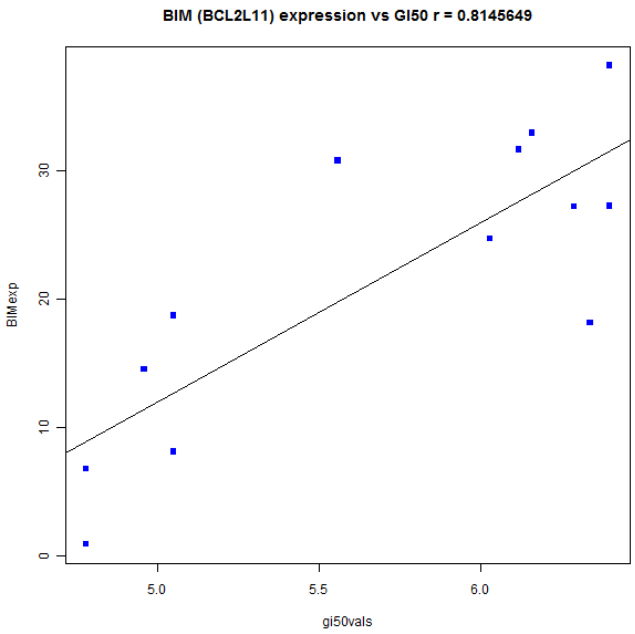


Figure 19. BIM AUC is highly correlated with GI50.

2.5.3 Results 1-3: Visualization of RPPA data onto Network and Pathway Diagrams

We hoped that visualizing the RPPA data on known pathways would highlight temporal interactions between proteins. Visualizing the RPPA data on the MTOR pathway for the UACC812 cell line is useful in assessing whether the putative causal relationships that are suggested by the pathway are actually supported in the data (Figure 20).

Visualizing the core set of proteins identified in Subaim 1-1 using the protein/protein interaction network from HPRD for the UACC812 cell line revealed interesting patters of mutations (Figure 21). Mutations that are directly connected to these RPPA proteins are visualized in Red. Of particular interest are the mutations that seem to cluster around nodes such as SHC1 and RB1. This visualization was the inspiration for the work in Aim 2, where we ask whether mutations cluster around specific nodes.

By visualizing a combination of both the phosphoprotein and mutation/copy number data superimposed on the PI3K/AKT pathway for the SKBR3 cell line, potential nodes were highlighted that possibly affect the response (figure 22). The results of Aim 1-1 are shown as a blue pathway, and possible SKBR3 specific responses are shown as a red pathway.

While not immediately useful, visualizing the RPPA data onto pathways and networks helped to generate further ideas about relating the timecourse and mutation data together. In Aim 2, we will explore one such idea, that of surrogate mutations.

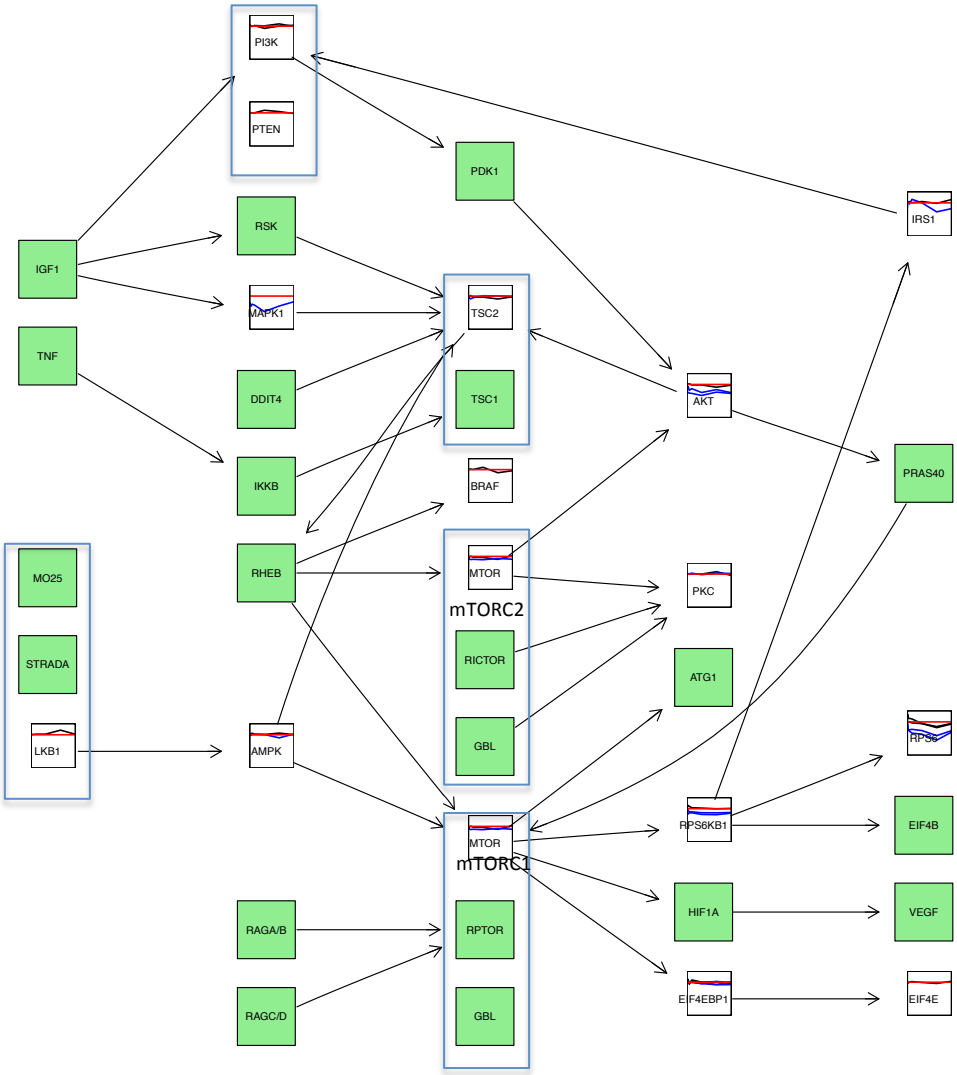


Figure 20. UACC812 RPPA timecourses visualized on KEGG mTOR pathway. Proteins with RPPA data in the pathway are represented with a small graph over the entire time series (0 to 72 hrs), with a red line acting as a reference for the zero point (0 = no difference from the DMSO trace). Complexes (such as MTORC1 and MTORC2 are represented by boxes enclosing multiple proteins), and proteins that have no RPPA data are represented as green nodes. Within the graphical boxes, a blue trace represents a phosphoprotein, while a black trace represents a non-phosphorylated antibody.

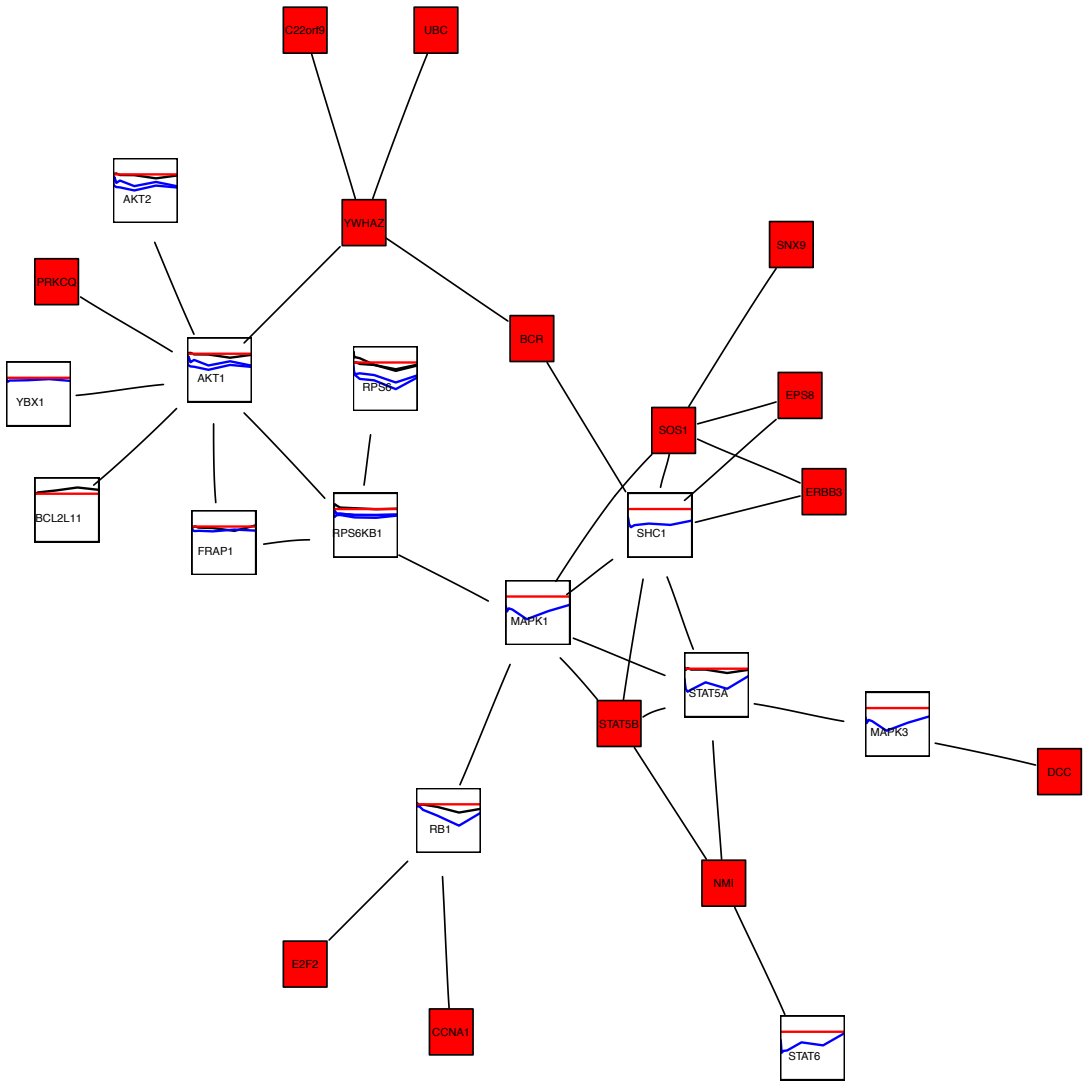


Figure 21. Visualization of UACC812 differentially expressed network (graph nodes) with Mutations (red nodes). Network interactions are derived from the Human Protein Reference Database (HPRD). This visualization was the inspiration for Aim 2, in that we wondered about the role of mutations that surround a node such as SHC1 or RB1 that may act as a "surrogate" mutation.

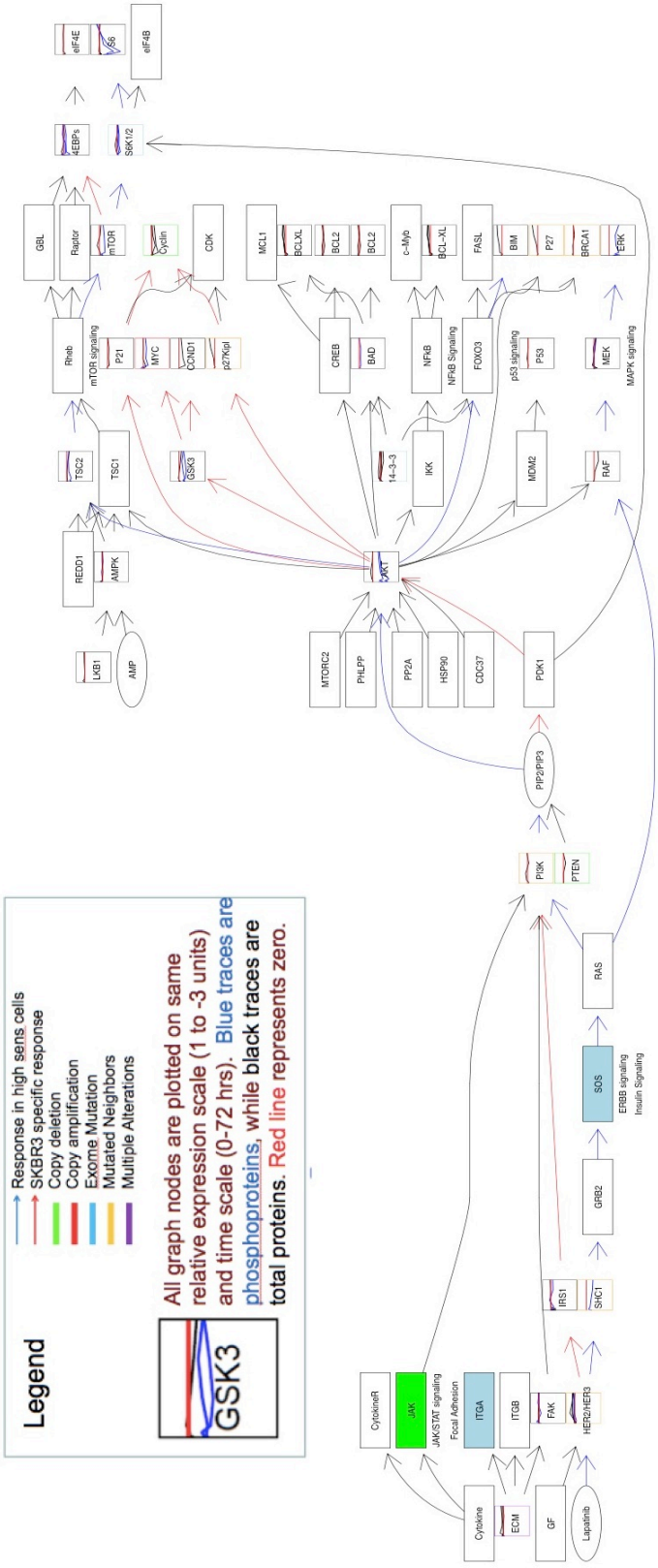


Figure 22. Visualization of SKBR3 timecourses on KEGG PI3K/AKT pathway.

2.5.4 Results 1-4: Can We Predict Protein Cross-phenotype response for Inhibitors in DREAM8 Data?

The sets of active phosphoproteins show variation across cell lines. Figure 23 shows the proteins for the early AUCs of UACC812 filtered by Coefficient of Variation (CV) across all conditions in the training data. A number of these phosphoproteins show very little variation (including EGFR pY1173, RSK pT359, and ACCpS79), and so were not incorporated into the early PLS-PM model. Visualizing the variation of the relative expression AUCs is a valuable filtering technique for excluding nodes in the model, as we do not include nodes whose expression is not notably different than the DMSO condition, simplifying the networks modeled. Here again we see the usefulness of AUC as a filter for cross-phenotype response activity.

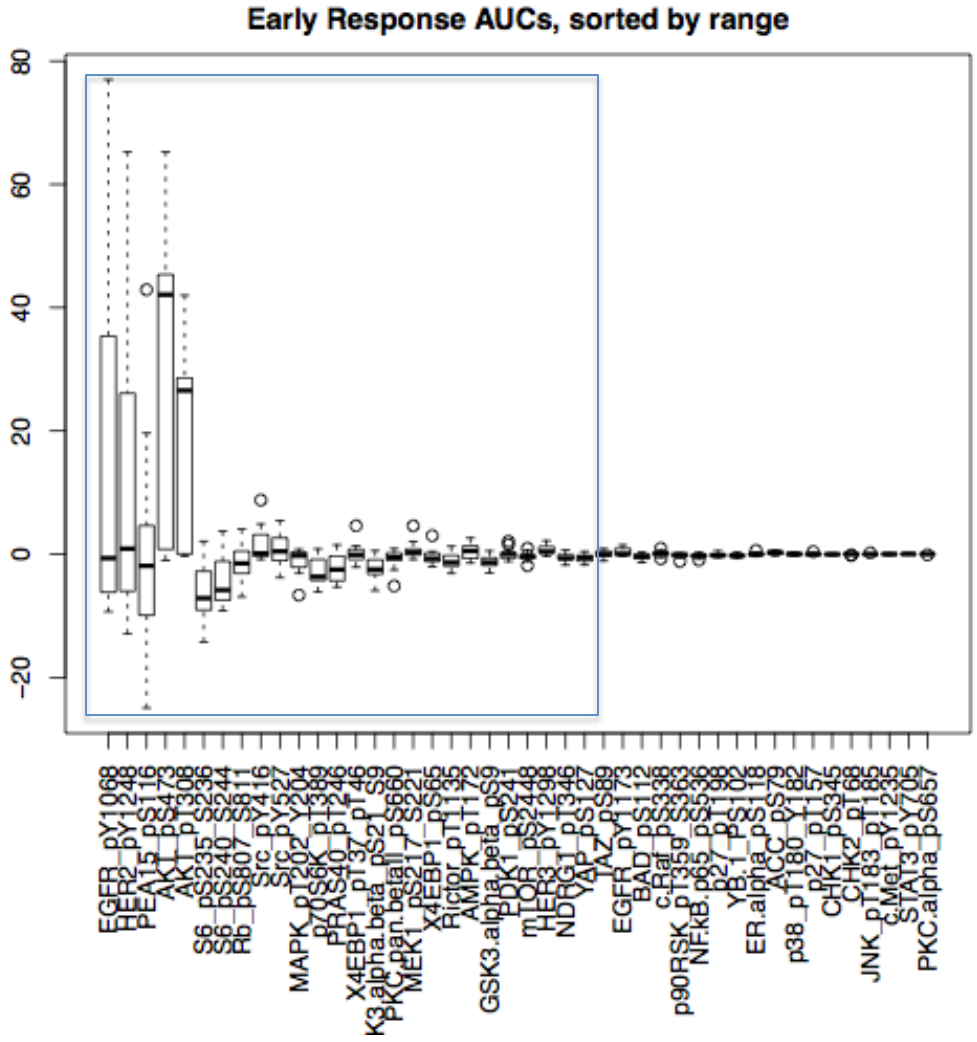


Figure 23. Variability in the Early UACC812 phosphonetwork. AUCs are shown for all conditions in the training set.

Multicollinearity analysis shows differential phosphosite usage. Figures 24-27 show the multicollinearity plots for proteins with multiple phosphosites, plotting AUCs of one phosphosite versus the other. One interesting quality is that the AKT phosphosites pT308 and pS473 are highly correlated in UACC812 (figure 24) and MCF7 (figure 25) cells, but less so in BT549 and BT20 cells (figures 26 and 27). That is, both phosphosites contribute equivalent information to the model. Because of this, we chose to include only AKT pS473 in our model. To calculate the value of AKT pT308, we estimated the linear relationship between AKT pS473

and pT308 in our training data using linear regression and used this linear relationship to transform our AKT pS473 predictions to pT308 values.

Additionally, we noticed differential phosphosite usage for the SRC phosphosites pY527 and pY416. EGFR only appears to be highly expressed in the HER2 cell line UACC812 and BT20.

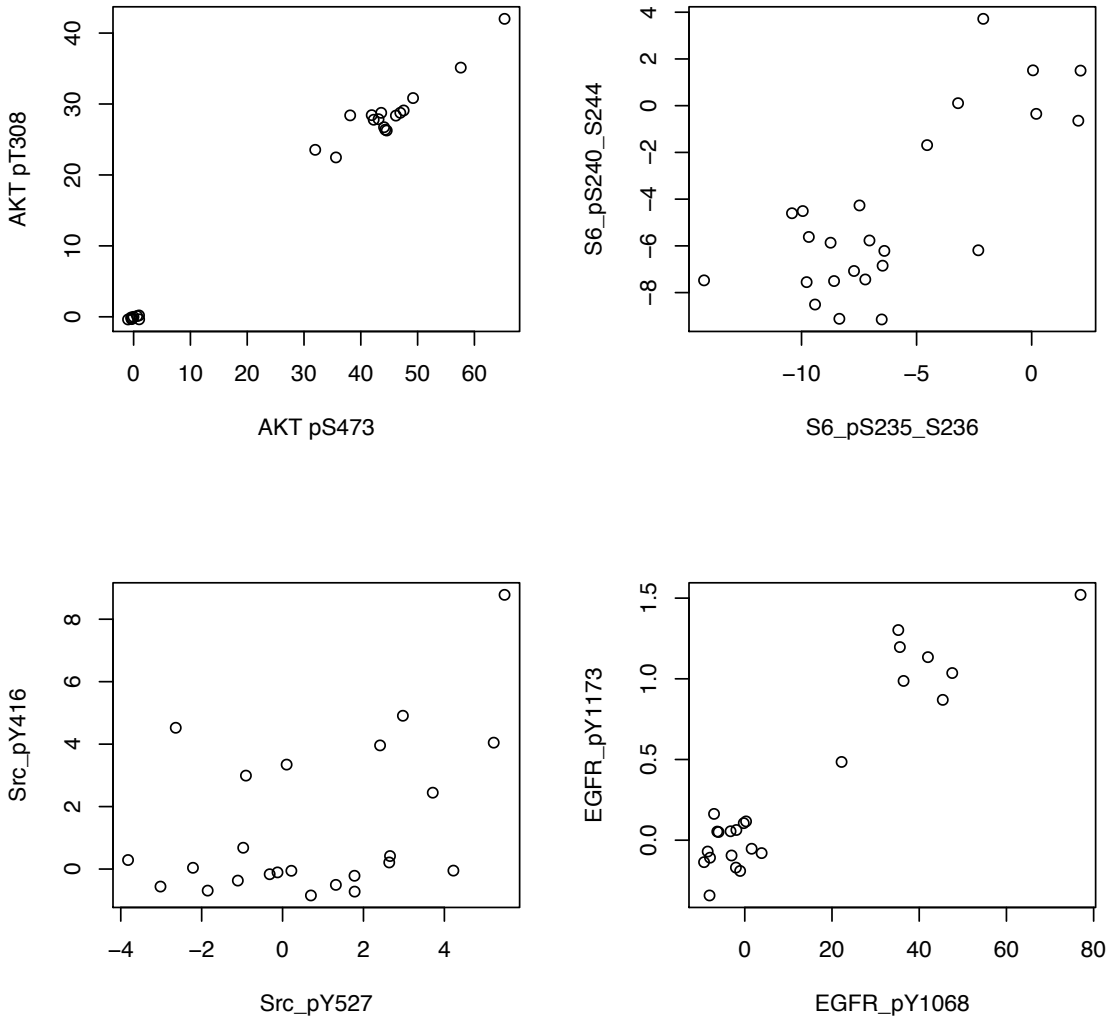


Figure 24. Multicollinearity analysis of multiple phosphosites in UACC812 cell line. In each plot, the AUCs for one phosphosite are plotted against the other phosphosite. Note that Src pY527 shows a continuous response, while Src pY416 appears to show a binary response.

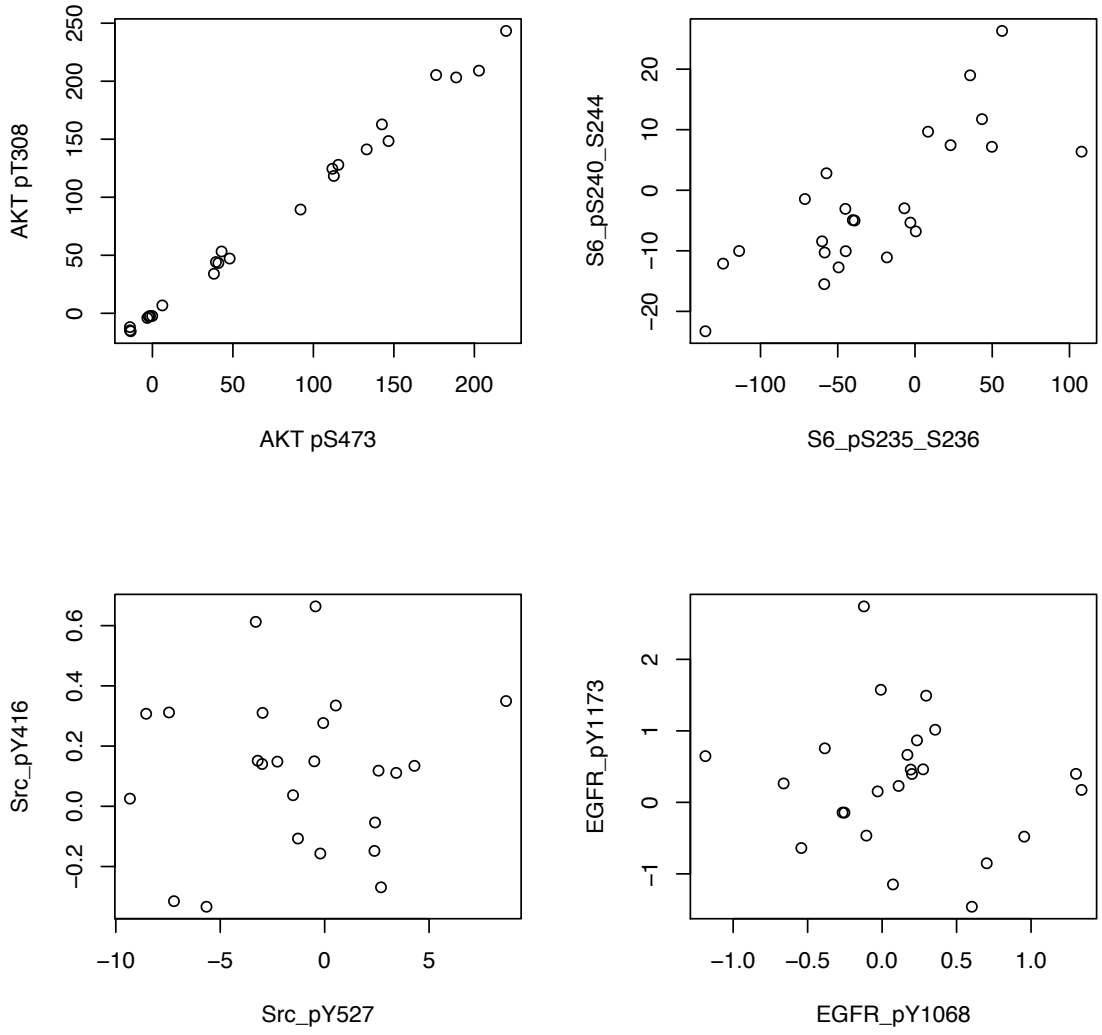


Figure 25. Multicollinearity Analysis of MCF7 cells. Note the possible differential usage of Src pY527, as its AUC expression is higher than Src pY416.

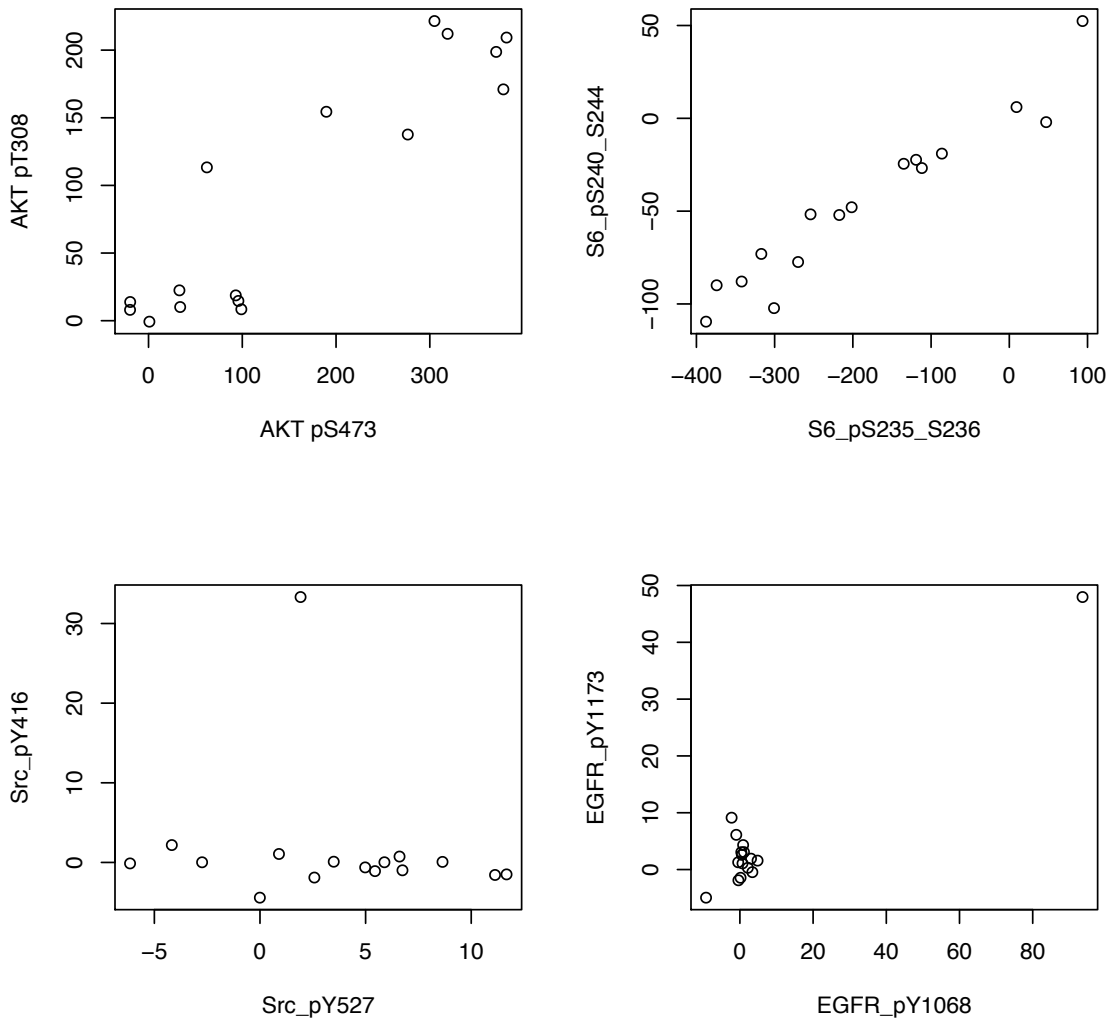


Figure 26. Multicollinearity analysis for BT549 cells. Note the possible differential usage of EGFR pY1173 versus pY1068.

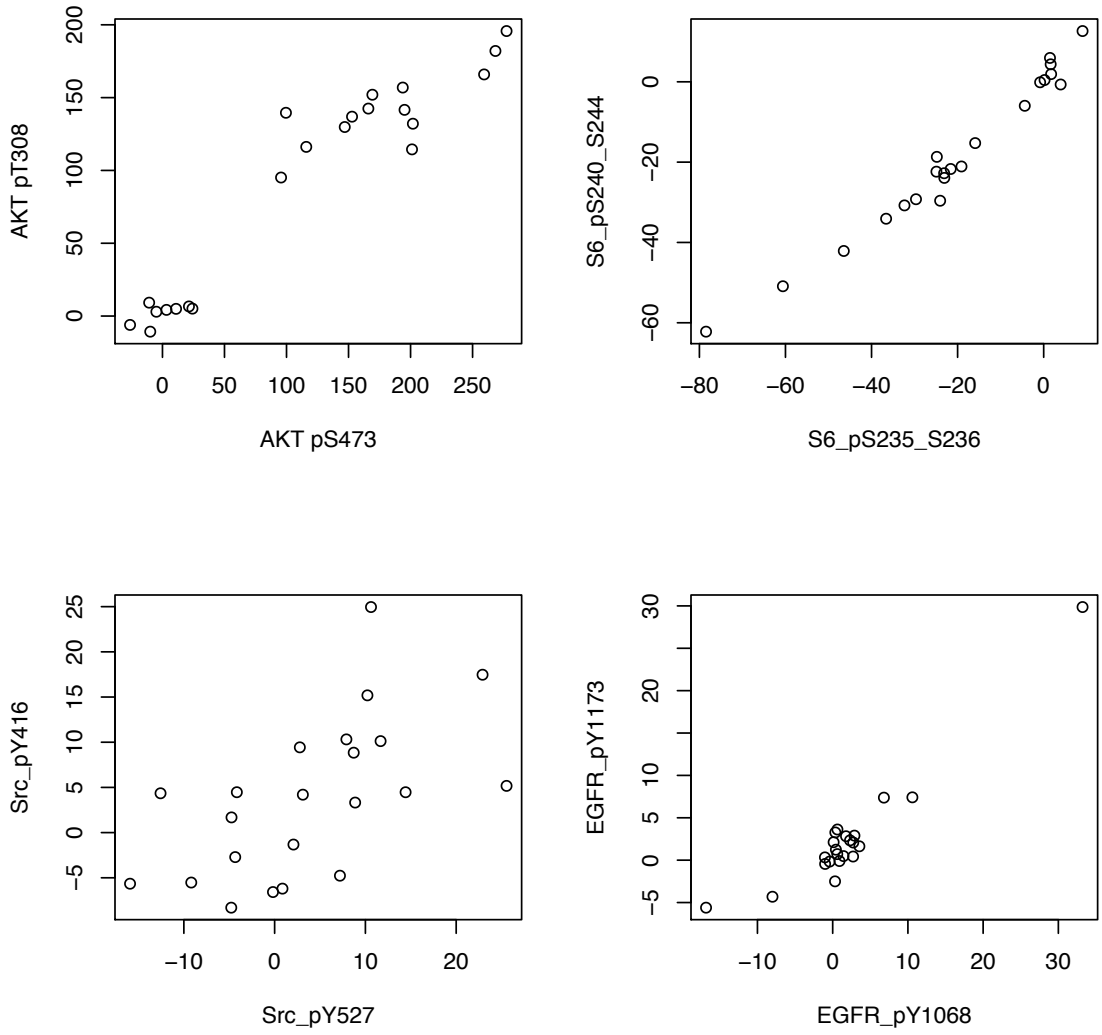


Figure 27. Multicollinearity analysis for BT20 cells. Note the high correlation between the two S6 phosphosites.

Visualizing the network models. Figure 28 shows an example of the PLS-PM model derived for the UACC812 early condition. Negative, or inhibitory relations are highlighted in red, while positive relations are highlighted in blue. Of interest are the negative relationship between EGFR and MEK (MAP2K1), and ERK (MAPK1). Some proteins, while highly expressed (PEA15), have very little influence on downstream neighbors (MAPK1).

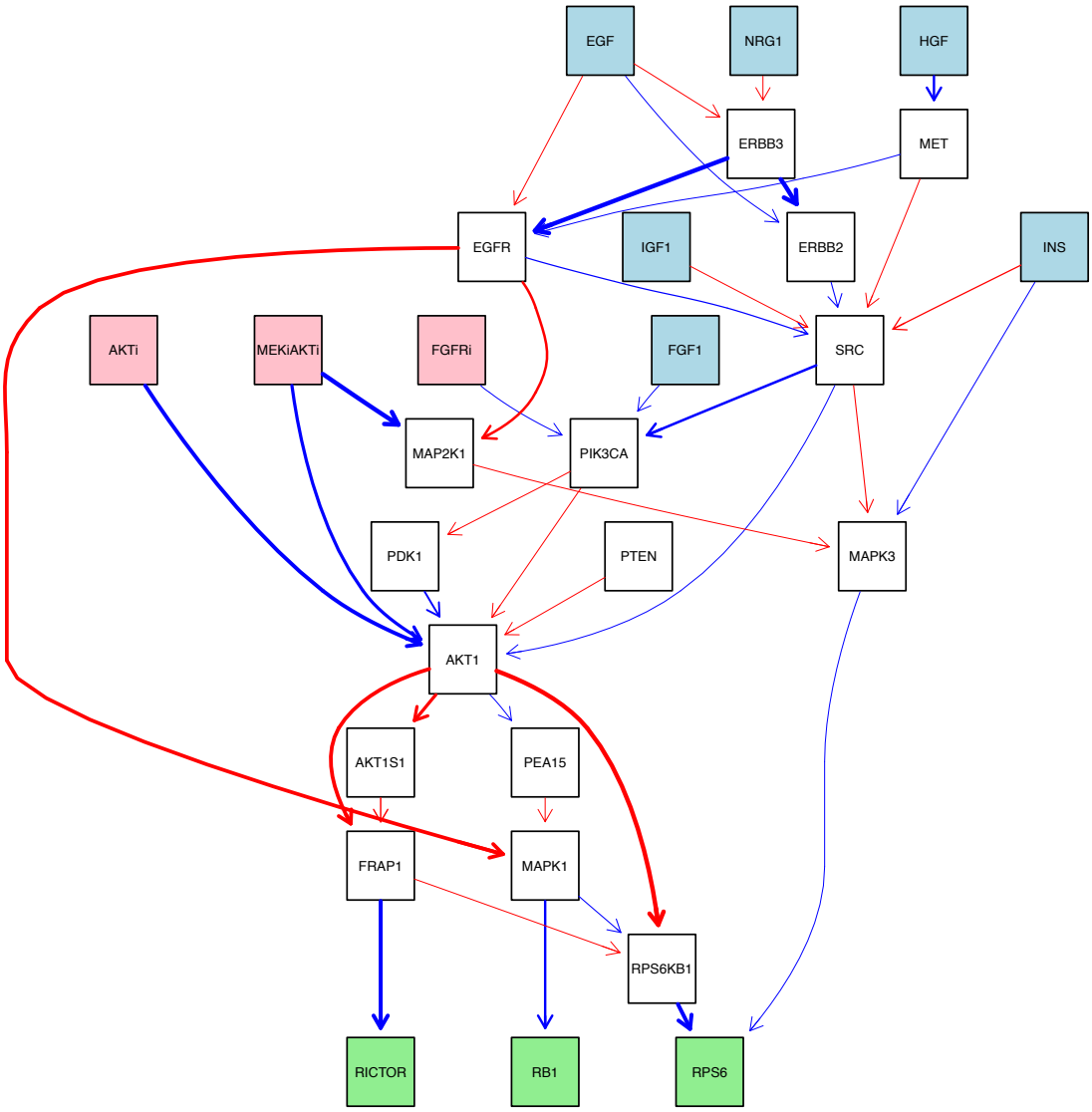


Figure 28. Calculated PLS-PM coefficients for UACC812 early condition. Blue boxes indicate exongenous inputs, Pink boxes are inhibitors in training set, and Green boxes are outputs of system. Thickness of arrows indicate strength of contribution in predicting the child node, with blue lines as positive relations and red lines as negative relations.

R-squared values show where networks characterize the data best. We will first examine the upstream protein receptors (Table 6) and their R-squared values, which characterize the percentage of variability we capture in the model. For the UACC812 cell line, which is HER2 positive, we account for 80.3% of the variability for EGFR, and 75.3% variability

for HER2 (ERBB2) for the early network. This is to be expected, as UACC812 is HER2 positive and expected to be addicted to HER2 and EGFR signaling. For the mid-stream signaling proteins AKT1 and MEK (MAPK1 and MAPK3), we capture a high amount of variability across multiple cell lines, especially for the UACC812 late and BT589 conditions for AKT1.

For the downstream proteins (Table 7), we capture a large amount of variability in the S6, S6 Kinase and Rictor proteins for the cell lines that contain these proteins across all cell lines.

Table 6. Well characterized upstream nodes (green) in PLS-PM networks. Values are R-Squared values.

GeneID	BT20early	BT20late	BT549early	BT549late	MCF7early	MCF7late	UACC812ear	UACC812late
EGFR(1956)	0.0352	0.0961	0.3532	0.0349	0.0154	0.1498	0.8039	0.5427
ERBB2(2064)	0.0522	0.0345	0.1752	0.088	0.1161	0.0278	0.7534	0.6363
MAPK1(5594)	0.8526	0.7627	0.2705	0.2858	0.7128	0.4338	0.5373	0.6266
AKT1(207)	0.6811	0.7033	0.7526	0.5975	0.4276	0.4398	0.6343	0.6029

Table 7. Well characterized downstream nodes (green) in PLS-PM networks. Values are R-Squared values.

GeneID	BT20early	BT20late	BT549early	BT549late	MCF7early	MCF7late	UACC812ear	UACC812late
RICTOR(253260)	0.5944	0.7699	0.8241	0.5445	NA	NA	0.609	0.8355
RPS6(6194)	0.7877	0.7889	0.8565	0.664	0.326	0.5489	0.655	0.5055
RPS6KB1(6198)	0.8671	0.8285	0.7497	0.4687	0.2012	0.1721	0.7273	0.8564

One strength of the PLS-PM method is that it also highlights nodes for which we have incomplete knowledge, or whose variability we do not adequately characterize using our network. These nodes include MAP2K1, PDK1, PIK3CA, PTEN, MAPK1, MAPK3, AKT1S1, PEA15, MTOR, PEA15, and RB1.

Ranking of predictions based on DREAM8 challenge. We submitted our sets of predictions to the DREAM8 challenge to see how our results would rank among submitted approaches. Briefly, 162 cell line/phosphoprotein pairs were used to score the results across the four cell lines. Two scores were generated for the set of predictions: Root Mean Squared Error (RMSE), which is an absolute measure of how close the prediction is to the actual data,

and mean z-score, which is an assessment of average rank across the cell-line protein pairs. In summary, the actual submission ranked 8th overall in terms of mean RMSE, and 9th overall in terms of mean z-score among the 15 entries (Table 8 – green indicates our submissions). However, a submission that underestimated the AUC values by two degrees of magnitude ranked higher at 5th place, and whose submission was dominated by the PBS traces. This suggests that the model actually performs worse than the initial guess of the traces. Unfortunately, with only one round of evaluation and limited feedback metrics from that round, a formal sensitivity analysis could not be conducted for the model, which would highlight exactly where the model was failing.

Table 8. Comparison of results with DREAM8 challenge final scoring. Results are sorted by mean Root Mean Squared Error.

Entry	meanRMSE	meanZscore
GuanLab	0.4528	-3.5864
Jcheng	0.4756	-2.9375
ALAK	0.4852	-3.6461
StochasticChaos	0.4853	-3.5982
TeamPineUnderAUC	0.4865	-2.9129
DynamoBios	0.5032	-1.9478
Opiyo	0.5071	-3.6103
sakev	0.5126	-3.0667
TeamPine	0.5294	-2.2756
Hatric	0.5718	-1.6207
StuartLab	0.6727	1.6516
CGR	0.8201	-1.1935
HD_Systems	1.353	2.20783
Information	1.951	47.3282
Frey	4.8708	144.3765
SBIT	8.0787	185.6077

2.6 Discussion

2.6.1 The importance of characterizing cross-phenotype response

RPPA characterizes the cross-phenotype response Lapatinib response in high sensitivity cells. The results of subaims 1-1 and 1-2 characterize the cross-phenotype response in cellular systems that results from applying Lapatinib to these cells. We will discuss each of the systems in turn, and what proteins in each system were affected.

The first cell system affected is the mitogenic signaling system. Akt phosphorylation (at both the pS473 and pT308 sites) is affected, as are the SHC bridging adaptor protein. These proteins are all downregulated in response to Lapatinib. One consequence of suppressing these upstream proteins is to downregulate mTOR, S6 Kinase, and S6, effectively reducing the level of protein translation in the system. However, it should be noted that the mTOR antibody did not pass the quality filter, and should be interpreted with caution. In some of the Lapatinib cell lines (AU565, BT474, UACC812), we see complete downregulation of S6 even after 72 hours, but in other cell lines (SKBR3 and MDAM175), we see recovery to DMSO control levels, suggesting that there may be additional regulation involved.

The other downstream consequence of Lapatinib is to downregulate three transcription factors: STAT5, STAT6, and YB-1. STAT5 and STAT6 are participants in the Jak-Stat signaling pathway, but are also affected by mitogenic signaling through Akt/MAPK. Overexpression of STAT5 is implicated in Breast Cancer.[86]

Overexpression in YB-1 is implicated in gastric cancer, but it is also implicated in Lapatinib sensitivity.[87] Crucially, YB-1 is also involved in transcription of the HER2 receptor, suggesting that it may be a key long-term feedback loop to understanding the resistance to Lapatinib. Transcription Factor binding site analysis indicates that a key binding motif of YB-1 is

found in the promoter region of HER2, but not in EGFR or HER3, which is additionally supported by knocking down YB-1, which affects HER2 levels, but not EGFR or HER3.[87] Transcriptional silencing of YB-1 actually increases Lapatinib resistance in gastric cell lines. It should be noted that additional transcription factors (such as Jun, Fos or Elk) for which we do not have validated antibodies could be affected by Lapatinib. However, we are unable to see these effects without validated antibodies.

The second cell system that is affected by Lapatinib is the apoptotic system, through Bim. Bim is a pro-apopotic protein that is repressed by the MAPK/AKT mitogenic pathways, which suggests that Lapatinib's effect of downregulating the AKT pathway results in the upregulation of Bim's expression.[88,89] Apoptosis is one possible outcome. Normally, Bim integrates the response of the cell from signals such as anoikis (loss of attachment to surrounding cells), UV damage, Ca2+ Flux, or Cytokine deprivation. Bim is also highly correlated with Lapatinib GI50, suggesting that it may be a key output in drug sensitivity.

The third cell system that is affected by Lapatinib is the Cell Cycle system. A key Cyclin that is upregulated in cancer, Cyclin B1 is downregulated when Lapatinib is applied.[90] It should be noted that the Cyclin B1 antibody did not pass the quality filter and thus should be interpreted with caution. Additionally, Rb phosphorylation is also downregulated, which as we have seen is associated with the G1 -> S transition, suggesting that one effect of Lapatinib is to halt cells at this transition.[9] As we have seen Rb1 phosphorylation is also highly correlated with Lapatinib GI50, providing further evidence that this may be a key output in Lapatinib sensitivity.

Finally, Rad51 is a DNA damage response protein, and is downregulated, suggesting that one possible side effect of Lapatinib is reducing the DNA damage response, which may potentially be a deleterious effect.

We do not know the transcriptional effects of Lapatinib. One drawback to focusing on the protein response is that we do not know the long-term transcriptional effects of Lapatinib. One study, that of O'Neill et all did study the transcriptional response to Lapatinib over time.[62] They identified a panel of 5 genes, RB1CC1, FOXO3A, NR3C1, Cyclin D1 (CCND1), and ERBB3 (HER3) that were differentially expressed. However, cell line growth conditions may differ between this study and O'Neill's, so caution should be used in interpreting these results in light of ours.

One caveat with these results is the choice of cell lines may be driving the correlations. For example, the cell lines in the Lapatinib dataset consist of ten HER2+ cell lines and five non-HER2 cell lines. The GI50s are clustered into two these distinct groups, which may be driving the correlations. Additionally, treating the highly sensitive cell lines as if they were replicates may obscure the individual responses. Actual technical or biological replicates may clarify the phosphoprotein response.

Another caveat is that the ANOVA results highlight similarities between the high-sensitivity cell lines, but not individual differences. These individual differences may include potential long-term responses that are compensatory in nature, and thus may affect drug sensitivity for that particular cell line.

2.6.2 Visualization of RPPA data on Pathways

Although projecting the RPPA data onto known pathways and connections was a useful visualization tool, it was also helpful to point out that data for many of the cell lines do not necessarily fit the canonical pathways in KEGG. For example, the cascade of MTORC2 -> S6K -> S6 may not be accurate in UACC812, as phosphorylated mTOR shows a late response while S6K and S6 show a more immediate response. However, other long-term regulatory loops involving

proteins such as IRS1 and the PI3K response were highlighted by this approach for cell lines such as UACC812 and could potentially be interesting candidates for further study.

Additionally, visualizing the RPPA data and the mutation data onto protein-protein interaction networks (Figure 21) raised the question that we pose in Aim 2, of whether mutations that surround a node can act like "surrogate" mutations.

2.6.3 The Utility of AUC as a Summary Measure

We have shown the utility of AUC with two different applications (Drug sensitivity and predicting the protein response to stimuli). In the first application, we integrated over the entire time interval in the data and showed that these AUCs are correlated with GI50s. This analysis highlights outputs in the RPPA data that are correlated with drug sensitivity, suggesting a possible new interpretation of GI50 in HER2 positive cells.

In the second application, we used AUCs as measures of pathway activity in a locally linear model and showed that they reasonably capture variability in upstream nodes in the mitogenic signaling network such as EGFR, HER2, AKT and MAPK, as well as capturing high variability in downstream nodes such as Rictor, S6 Kinase, and S6. Our inability to capture variability in the middle nodes is in line with many other modeling approaches, as the signaling pathways are not well characterized.[91,92] [REFS]

2.6.4 Differential Usage of Phosphosites Across Cell Lines

One observation that came out of the Aim 1-4 modeling is that different cell lines utilize phosphosites differentially. Within the BT549 and MCF7 cell line, the Src pY527 phosphosite was used preferentially to the pY416 site. pY416 is required for activation of Src Kinase activity, whereas pY527 phosphorylates the tail portion of the protein, essential for binding to Src's own

SH2 domain, keeping the protein in an inactive conformation. The preference for pY527 suggests that the Src protein is kept inactive under conditions of stimulation/inhibition.[93]

Binary, all or nothing, responses were also noted for some phosphoproteins. Most notably, a binary response was noted within the BT20 cell line for both the AKT pT308 and AKT pS473 phosphosites. Within UACC812, Src pY416 appears to have a binary response, whereas the pY527 site is continuous.

Any future modeling effort of these cell lines will need to account for differential phosphosite usage, and account for the binary behavior in the phosphosite response as well. Thus model interactions of RTK signaling should include this level of interactions. However, we are aware of no knowledge bases that capture interactions on the level of phosphosites.

2.6.5 Predicting the phosphoprotein response using PLS-PM modeling

While we note that PLS-PM is a powerful approach to predicting protein response data, there are many improvements that can be made to the models generated. It is also important to note that there are many caveats to using PLS-PM modeling to predict the phosphoprotein trajectories. The first is error propagation. Because we calculate the nodes in topological order, upstream predictions greatly affect those predictions downstream. Thus, our efforts to improve our models should focus on improving our characterization of these upstream nodes.

Another improvement that can be made is to assess the impact and improve how inhibition is done in the model, most notably the multiplier used in inhibition. One method to improve this parameter would be specifically quantifying the effects of inhibitors using alternate datasets.

A number of improvements can be made to the networks themselves. First of all, non-phosphoprotein data can be included in the network. This may be especially useful in modeling the response of the upstream nodes and increasing the variability captured for those upstream nodes.

Tenenhaus note that the estimates for the Latent Variables can be used in different types of regression as well. Quadratic regression or logistic regression may improve our estimates of model relations, especially in the case of threshold or binary seeming AUCs.[76]

Once the networks have been improved, the networks themselves can be used as the basis for ODE models. Each PLS-PM model highlights nodes for which we characterize the data best, and models using only those nodes that have the highest R-squared values could be a potential starting point for an ODE model.

2.6.6 PLS-PM Results on DREAM8 Challenge

While placing 8th place overall in terms of RMSE, our model currently does worse than the initial basis for our guess, the Stimuli/DMSO traces for each Stimuli condition (5th place). Ideally, a sensitivity analysis of parameters would have been made to assess why the model was performing badly, but could not be conducted due to limited evaluation (1 round) and metrics available from that round.

Given the extremely time compressed nature of the DREAM8 submission process, our entry was a late submission and did not benefit from the longer, iterative improvement of the other models submitted. Only two submissions could be made in the timeframe allotted to us. Unfortunately this means that a formal sensitivity analysis of model parameters, such as the inhibition mechanism could not be conducted. The difficulty of assessing and improving a

model given just two DREAM8 metrics, the Root Mean Square Error (RMSE) and Z-score is difficult.

We suggest that for further modeling challenges, better metrics be made available to contestants in order to improve their models. Indicators of local error (RMSE for a given phosphoprotein for a given drug and stimulus) would be ideal, and a more continuous evaluation process be made. If errors on that level could not be made, at least providing errors on the Cell Line level would have been extremely helpful in localizing errors. Unfortunately, certain portions of the evaluation framework were not automated, which required manual intervention on the evaluation side.

2.6.7 Considerations for future analysis of RPPA data

In Table 9, we summarize limitations of the RPPA platform and experiments and their consequences for downstream computational analysis. Most importantly, for our analysis, *the dynamics of the antibodies are affected by the control treatment*, DMSO, which required the DMSO effects be subtracted from the actual treatment in order to understand the true dynamics of applying the treatment. Thus, all predictions must be made on a relative scale in this case.

In addition, *all concentrations measured with RPPA as currently used in our data are relative and concentrations across antibodies cannot be compared*. This impacts models, especially models that require an initial steady-state value such as ODE models. This means that additional information about protein abundances are needed in the model. The absolute values can be derived from RPPA through the use of standard samples with known protein concentrations, but this is not currently done.

Another issue is the *lack of replicates in the time series data*. This impacts the analysis in two ways. The first way is the need for robust summarization methods, as it is difficult to assess whether a spike in a timepoint is spurious or not. In order to deal with this issue, we use summarization methods like area under the curve (AUC), which is more robust to spikes in the data. The second way the lack of replicates impact the analysis is that power must be generated by using cell lines as replicates in the statistical analysis.

Next, the list of validated antibodies is *currently limited to around 300 antibodies* that are derived largely from signaling pathways and cellular systems that are known to be altered in cancer. While this list is quite large, it is far from comprehensive. Our view of the signaling process is biased towards these antibodies and thus generation of new hypotheses is limited to this list, unlike the unbiased approach of gene expression microarrays.

Finally, as was seen with the UACC812 data for the DREAM8 challenge, if a cell line is split between two different arrays, batch effects can result. Such batch effects require normalization.

One final suggestion is that the raw RPPA data, including the positive control surface and the standards should be given to users as well as the preprocessed data. While the current analysis pipeline minimizes many sources of variation, understanding the role of the positive control surface in reducing spatial variation on antibodies themselves is important from the user perspective.

Table 9. Limitations of the RPPA data and consequences for analysis.

Data Type	Limitations of Data Platform	Consequence for Analysis
RPPA	Concentrations are relative; values cannot be compared across antibodies	Steady State values for abundances are needed; ODE models can't be initialized without this information
	Dynamics must be gauged with a reference	Model must incorporate relative effects; use of AUC to compare dynamics
	Splitting samples between arrays can cause batch effects	Normalization is needed to mitigate batch effects
	Lack of Replicate Timecourses	Treat cell lines as replicates using GI50 as a guide for grouping
	List of Antibodies is Limited	Understanding of effects on signaling is limited to these antibodies

2.7 Conclusion

In this chapter, we have highlighted four approaches for characterizing system cross-phenotype response in RPPA data: ANOVA analysis across cell lines, AUC analysis across cell lines, Visualization of time series in directed and undirected networks, and finally utilizing AUCs to build predictive models of dynamic data. Each of these approaches highlights aspects of the data that are useful in understanding drug sensitivity.

The ANOVA analysis highlights the extent of cross-phenotype response in Lapatinib sensitive cell lines, showing that Mitogenic signaling proteins (and transcription factors), Cell-Cycle proteins, Apoptotic Proteins, and double stranded DNA damage response proteins are affected by Lapatinib. Specifically, the feedback loop of YB-1 (a transcription factor) and its transcription regulatory target (HER2) is highlighted as a potential source of Lapatinib resistance.

Our AUC analysis adds to the characterization of cross-phenotype response in that it directly posits a linear relation between several proteins and drug sensitivity.

Visualization of time-series data on pathways and networks has potential for highlighting the individual cell line response to Lapatinib. In particular, the IRS1 response is seen as potentially important in regulating the long-term PI3K response in UACC812 cell lines.

Finally, we highlight a novel technique for predicting cross-phenotype response in timecourse data, Partial Least Squares Path Modeling. The strength of this technique is that it utilizes multiple linear regression in a network framework to estimate linear relationships between variables. However, the AUC predictions from our model were actually worse than a submission that was largely the stimuli/DMSO traces. In order to understand why, a formal sensitivity analysis with using good local error metrics needs to be conducted in order to understand the specific failings of the model.

Chapter 3: Aim 2 – Assessing the Impact of Genetic Mutations in Cell Lines using Protein/Protein interaction networks

3.1 Research Question and Aim

Do groups of mutations collaborate? Can we detect these groups using network analysis? Are these groupings associated with drug sensitivity?

These questions inform my aim, which is:

Aim 2: Identify potential oncogenic collaborations that can inform the application of targeted drugs.

3.2 Introduction

Tumorigenesis is a multi-step process that commonly occurs over decades. Given a long enough lifespan, each of us will eventually acquire cancer as it is a result of multi-systemic breakdown and genomic alterations that are acquired to heredity, exposure to environmental factors, as well as the buildup of internal genomic alterations due to mistakes in our DNA repair machinery.[9] As we have seen in Chapter 1, tumor cells must acquire many hallmarks and capabilities in order to proliferate and ultimately metastasize to other portions of the body.[3,4] The acquisition of these capabilities happens through a decades long process of selection through multiple clonal expansions.

One method of acquiring such capabilities (other than acquiring direct mutations) is oncogenic collaboration. Oncogenic collaboration is defined as a synergistic interaction between two or more mutations that transforms cells to tumorigenic.[9] These mutations by themselves may be insufficient to transform the cells; it is only in combination that they are transformative.[94]

In this aim, we propose another type of oncogenic collaboration that goes beyond the two gene collaborative model. Based on observations of mutation distribution that came from visualizing mutations on a protein-protein network, we ask the question whether the node-centric distribution of these mutations has biological significance in drug sensitivity. Such a model potentially integrates many of the genes that we consider as passenger mutations and provides a new perspective on tumorigenesis.

3.2 Background

In sub-aim 1-3, we observed through the results of visualizing RPPA data with mutations an interesting phenomenon: that certain nodes had a high density of mutations surrounding

them (Figure 21). Oftentimes, the node itself would not be mutated, which led us to question whether the surrounding mutations had some sort of role in regulating that protein.

In this background section, we will first discuss the detection of Oncogenes and Tumor Suppressor Genes (3.2.1), as well as their further classification and annotation (3.2.2). We then discuss the current landscape of Breast Cancer Mutations as found by the recent Cancer Genome Atlas (TCGA) paper on Breast cancer (3.2.3). We then discuss previous work using networks and mutations in order to provide motivation for our proposal for a new model of oncogene collaboration, which we call surrogate mutations (3.2.4).

3.2.1 How Do We Detect Oncogenes and Tumor Suppressors?

The question is then how oncogenes and tumor suppressor genes can be detected using current technology. In this section, we will discuss current approaches for detecting oncogenes and tumor suppressor genes. Oncogenes are an easier target to find, requiring integration of copy number data and next generation sequencing (NGS) data. Detecting tumor suppressor genes is a more difficult problem, requiring the integration of not just the Next Generation Sequencing Data, but also methylation data, as well as special algorithms to detect LOH. We do not discuss detection of TSGs in this dissertation due to this complexity.

Detection of oncogenes using Next Generation Sequencing. Next generation sequencing methods result in millions of short fragmentary sequences (called reads) that must then be assembled by aligning these reads onto a scaffold sequence[95] By comparing the sequence with either the scaffold sequence or a set of normal samples, the likelihood of a mutation occurring at a location can be called.

It is beyond the scope of this dissertation to talk about the total impact of sequencing methods on the downstream results. We will discuss two experimental parameters that we

deem most important. Next generation sequencing methods have two parameters that impact downstream analysis: *coverage* and *depth*.[96] Briefly, coverage can be defined by whether we have fragments that cover a desired interval of the genome, and depth can be defined as how many fragments (and thus evidence) we have covering that desired part of the genome.

Sequence coverage can be partially controlled by selecting portions of the genome to sequence, either by molecular cloning, or microarray based methods, such as Exon Capture. However, there are also biases to sequence coverage that need to be corrected for that are inherent to NGS methods. Notably, there is a bias to sequence coverage associated in GC (the percentage of Guanines and Uracils in the sequence) content.[97] This bias affects the quantitation, and thus must be corrected for.

Sequence depth can be controlled by increasing the number of reads applied to the genome in question. Sequence depth greatly impacts our confidence in a somatic mutation. Depending on the expected frequency of the genomic feature, more sequence depth may be required to be confident in a feature. For example, since SNPs are defined as having a high-expected frequency in a population, the depth of coverage required is lower than a somatic mutation, which by definition is a rare event in the population.

The choice of reference affects what alterations are called as mutations. Most often, tumor samples are compared to matched normal samples in the same tissue in order to assess possible functional mutations.[9,98] One caveat to this approach is that we have seen that tumorigenesis is a long, multi-step approach, meaning that the 'normal' samples may have acquired functional mutations. Thus, all mutations that contribute to the cancer may not be found with this approach.

With cell lines, no such matched samples exist, thus the reference sequence may come from a sequencing project such as the Human Genome Project. This choice of reference may mean that many more genomic differences will be found, and thus the number of assessed mutations and SNPs may be much larger than using a matched sample.[9]

The raw short reads are then aligned to a reference genomic sequence using any number of fast alignment programs, which utilize algorithms such as the Burrows-Wheeler transform to align as many short reads as possible to the reference genome.[99] The output of this step is a genome sequence where each location in the genome is quantified by the number of supporting reads (read depth), as well as an overall score for the quality at that location. For example, in the case of a heterozygous mutation, a particular location may contain equal numbers of reads for the wild-type and mutation. This output is then fed to a mutation-calling algorithm.

Based on the proportion of reads supporting the reference sequence and the proportion of reads supporting the mutated sequence, and the expected frequency in the population, a position in the sequence can be called as homozygous reference (and thus not a mutation), heterozygous mutant, or homozygous mutant.[99] This calling is often done through the use of Bayesian methods, which can incorporate a prior (expected frequency in the population) and sequence read information into a probability of each type of mutation.[99,100]

If the sequence is then in a protein coding region, the mutation is then translated to a protein sequence. The mutation may then be called as synonymous (does not affect protein coding sequence) or non-synonymous (affects protein coding sequence). Other mutations may be found in promoter regions, which can theoretically affect the binding of transcription factors.

Many biases exist in NGS methods. The sequencing method itself has a known bias towards GC-rich content in the genome, which needs to be corrected for.[97,101] The base calling algorithm can also be biased due to improper selection of parameters, and so the effect of parameters on downstream results should be assessed. Batch effects have also been observed when comparing across sequencing batches, and these effects also need to be corrected for.[101]

Copy number detection of amplified oncogenes. Copy number amplifications and deletions can be theoretically determined at the genomic level by counting the number of copies that exist in the tumor genome using NGS. However, the majority of methods utilize array-based platforms such as genomic microarrays to detect copy number variations (CNVs). There are two different methods in use: Comparative Genomic Hybridization (CGH) and single sample genomic hybridization.[102] The main difference between the two methods is what is hybridized to the array. In CGH, a reference sample is hybridized along with the sample of interest. Thus the ratio of hybridization between the two samples is generated.[102] In single sample, the sample of interest is hybridized to the array, which must be compared to the hybridization of reference samples run on separate arrays.[102] We will focus on copy number analyses using single sample arrays.

The first experimental parameter that affects results is the *degree of coverage of the genome* offered by the DNA array. Coverage refers to the genomic markers used in the design of the arrays. Coverage affects the granularity (that is, localization) of the copy number call. Early expression arrays for copy number detection, ArrayCGH, limited the resolution of copy number alterations to the megabase range, due to the limited number of genomic markers used in their design.[102] However, with the design of arrays with better genomic coverage have been made. Two examples of higher coverage arrays are the tiling arrays and SNP (single nucleotide

polymorphism) arrays. Tiling arrays attempt to cover as much of a genome as possible; one consequence of this is that annotation of results to the genome can be difficult and biased towards the reference genome used in their construction. SNP arrays attempt to cover well-characterized SNPs that are observed in a large percentage of the population. However, SNP arrays have also been used for copy number prediction by using the SNPs as genomic markers. SNP arrays are obviously biased towards regions that are dense in SNPs.

However one complication to this analysis is in order to call whether a genomic region is amplified or deleted, its expression must be compared to a reference. In tumors, the reference is straightforward: much like somatic mutation calls, the reference is a matched normal sample. In cell lines, however, no such matched normal samples exist, and a composite reference derived from multiple normal samples must be used. One source of normal samples is the HapMap project, which was an early project to uncover common haplotypes (small genomic regions that are varied together) across human populations.[103,104] Obviously, the selection of these normal samples affects the amplification/deletion calls. In order to make an absolute copy number call for the sample, the ploidy (number of copies) of the reference must be known.

Another aspect that affects results is the algorithm used in making the copy number calls. We will discuss a common algorithm used for copy number calling, Circular Binary Segmentation (CBS).[105] Essentially, CBS looks for change-points across a series of copy number calls, separating the chromosome into regions of constant copy number ratios. These change-points may represent transitions from regions that are 0 (no difference from reference) to negative values (copy number deletions) or positive values (copy number amplifications) or any other combination of these three states (Figure 29). The Binary Segmentation algorithm finds change-points through an iterative algorithm. The first pass is over the entire chromosome, and

asks at each marker whether the distributions of copy number ratios before and after the marker are statistically different. If this is the case, then the marker is considered a change-point. The algorithm is then run over the two segments before and after the change-point, and then over the new segments until no new change-points are found. The Circular Binary Segmentation algorithm is similar, except it treats the chromosome as circular by fusing the start and endpoints of the chromosome and searches for two change-points at once. The output of CBS is a set of genomic regions (or segments) with their Copy Number Ratio, or ratio of number of copies in the sample to the number of copies in the reference sample. More often, the log of the Copy Number ratio is used to aid in interpretation, such that positive number signify amplifications and negative numbers signify deletions. CBS is weak when the range of expression values between markers is small, and it can generate too many segments if the data is noisy.[105]

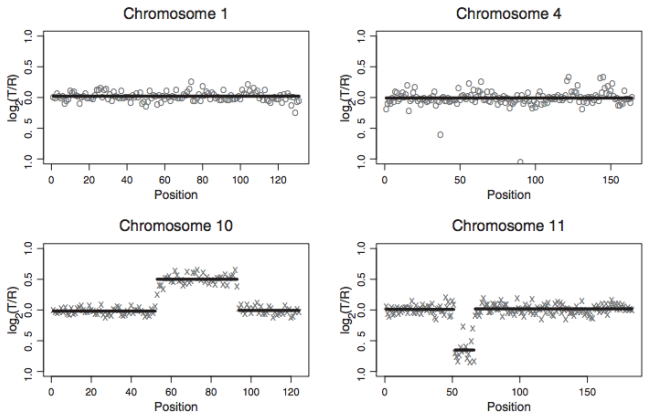


Figure 29. Example of CBS region calls. Reproduced from Ohlsen.[105]

These copy number regions can then be compared across samples to find regions that statistically vary across many samples. The reasoning behind such a method is to identify commonly amplified and deleted regions in a cancer type or subtype. One algorithm to do this comparison is GISTIC (Genomic Identification of Significant Targets in Cancer).[106] Given a set of

per-sample copy number regions derived from an algorithm such as CBS, GISTIC asks the question of whether a genomic region shows a wider than expected amount of CNV. This call is made by comparing the amount of variation in a region to its variation in a permuted set of regions, where labels have been shuffled. By defining a significance threshold over the distribution of the permuted set, significance can be decided. The output of GISTIC is then a series of genomic regions that show higher than expected CNV.

Once these alterations are detected using the above algorithms, we can then use various annotation and visualization methods to make sense of them. One such method is Gene Set Enrichment Analysis, in which we treat the mutations as a bag of genes and ask whether certain annotations in this bag of genes, such as pathways or gene functions are statistically overrepresented.

The approach we take is to analyze the distribution of mutations on a protein-protein interaction network and ask whether this distribution is statistically significant.

3.2.2 Assessing the Functional Impact of Mutations

Because thousands of mutations may be called by the sequencing step, it is important to filter out probable passenger mutations from driver mutations.[107] This can be done by various sequence analysis methods.[108] Most analysis methods will first classify sequence mutations as being exonic (within protein-coding sequence) or intronic (outside of protein coding sequence, in genomic regulatory regions, splicing regions, or other regions). As a first pass, most analysis methods will only focus on the exonic mutations.

Exonic mutations are further classified into synonymous (resulting in the same coded protein when translated to protein sequence) or non-synonymous (resulting in a different translated protein sequence) mutations. For example, synonymous mutations in the proteins

may be screened out as they do not have an impact on protein structure. It should be noted that while it may not affect the protein level, synonymous substitutions may affect the rate of transcription, suggesting a possible effect on the transcriptional level.[109] Additionally, the type of non-synonymous substitution may be assessed, such as a hydrophobic amino acid being substituted for a hydrophilic amino acid. This substitution may be considered to be more impactful than a substitution of one hydrophilic amino acid for another.

The vast majority of methods for analyzing functional impact of mutations in cancer tend to utilize machine learning in one of two contexts: the first is an conservation/evolutionary context using comparative methods, and the second context is frequency analysis of mutations across cancer samples.[110] This second context can be done with various focuses: 1) a Whole Genome approach, 2) A Pathway/Gene Set focus, 3) a Gene-centric approach, or 4) a Network-based approach. We will discuss the first three focuses here; the network-based approach is discussed in section 3.2.4.

Both the evolutionary and frequentist approaches attempt to assess the impact of a mutation by comparing its frequency of mutation to an observed background model of mutations. The rationale behind this comparison is that Driver mutations confer an evolutionary advantage onto tumor cells and thus are acquired at a faster rate than non-deleterious, silent mutations.

In the evolutionary/conservation context, deleterious somatic mutations are identified by comparing gene sequences across organisms and noting changes in highly conserved positions. The most well known of these approaches is SIFT (Sorts Intolerant From Tolerant substitutions).[111,112] By comparing across a number of reference genomes across species, how deleterious a mutation can be estimated. If a protein residue is highly conserved across

genomes across multiple species, a mutation at that site is considered highly deleterious. However, if a protein residue is not highly conserved, then that residue is considered as tolerant of mutations. Other approaches include PolyPhen, which assesses the impact of an amino acid substitution on the predicted protein structure.[113]

In contrast, frequency based approaches compare mutations observed across multiple cancer samples and compare their frequency to a null model of mutations. There are a number of pipelines for identifying driver mutations. One of the most well known is MutSigCV, which uses synonymous (or silent) mutations as an estimate for the background mutation rate. Because mutations are rare, MutSigCV also incorporates slient mutations from genes that are nearby in the genome into the estimation of the background mutation rate.[114]

The impact of these filtering steps is to drastically reduce the number of mutations and copy number alterations studied. Choices made during these steps may reduce false positives at the expense of false negatives. Thus, these choices impact us in our study, which is largely exploratory.

3.2.3 The Current Mutational Landscape of Breast Cancer

We will now look at current studies of the genomic variation in breast cancer. The vast majority of known and confirmed mutations in breast cancer are in four proteins: p53, which is a transcriptional regulator, PI3K, PTEN, and GATA3.[98] Both p53 and GATA3 are transcription factors, whereas PI3K and PTEN are kinases that are involved in phosphorylating and dephosphorylating PIP2/PIP3, a molecule involved in signaling.

The Cancer Genome Atlas consortium looked for driver mutations in breast cancer using two approaches: 1) a frequentist approach that integrated multiple data types (using their MutSigCV algorithm) and 2) a network approach that looked for mutually exclusive mutations

(using a network-based approach called MEMo).[98] Using MutSigCV, the highest frequency mutated genes across all subtypes of breast cancer (across 501 samples) were: PIK3CA (36%), TP53 (37%), GATA3 (11%), and MAP3K1 (8%). Within the HER2 enriched subtype (57 samples), the most frequently mutated genes was slightly different: TP53 (72%), PIK3CA (39%), and MLL3 (7%). This study is one of the largest studies to screen for breast cancer mutations, with 825 patient samples.

In Table 10, we can see the incidence of mutations and copy number alterations in the HER2 subtype for two different studies: the TCGA breast cancer study, with an N of 57, and a database of somatic mutations called COSMIC (Catalogue of Somatic Mutations), a sample size of 299 samples.[115] Within the HER2 subtype, several patterns emerge. The PI3K/AKT signaling pathway, which is connected upstream to HER2 is mutated in several genes: PIK3CA, PI3KR1, both components of the PI3K protein, PTEN, AKT1, and INPP4B. Additionally, MAPK signaling, which is also connected upstream to HER2 signaling is also affected. Several downstream signaling events are also targeted, including Cell Cycle Regulation and various transcription factors.

These frequently mutated genes offers part of the picture of how cancer evades various regulatory and signaling mechanisms. But as we will see, a large portion of the mutations for an individual or cell line are unique to the cell line and do not fall within these pathways. The question then, is if these unique mutations confer any other hallmarks to the cancer cell.

Table 10. List of frequently mutated genes in the HER2 subtype. Columns 2 and 3 are derived from the TCGA breast cancer study, and column 4 from COSMIC (catalogue of somatic mutations in cancer).

Gene	HER2 copy incidence (n=57)	HER2 mutation incidence (n=57)	HER2 mutation COSMIC (n=299)	Function/Pathway
PIK3CA		39%	20%	PI3K/AKT signaling
TP53		72%	31%	Transcriptional Regulation
MAP3K1		4%		MAPK signaling
MAP2K4		2%		MAPK signaling
GATA3		2%		Transcriptional Regulation
MLL3		7%		DNA binding/leukemia assoc
CDH1		5%	2%	Cellular Binding/Structure
PTEN		2%		PI3K/AKT signaling
PIK3R1		4%		PI3K/AKT signaling
AKT1		2%		PI3K/AKT signaling
RUNX1		4%		Transcriptional Regulation
CBFB		2%		Transcriptional Regulation
TBX3		0%		Transcriptional Regulation
NCOR1		0%		Transcriptional Repressor
CTCF		2%		Transcriptional Repressor
FOXA1		2%		Transcriptional Activator
SF3B1		4%		Splicing Factor
CDKN1B		2%		Cell Cycle
RB1		0%		Cell Cycle
AFF2		5%		Implicated in Breast Cancer
NF1		0%		MAPK signaling
PTPN22		5%		Immune Response
PTPRD		4%		Signaling
ERBB2	71%		5%	HER2 signaling
MDM2	30%			Negative regulator of p53
CCND1	38%			Cell Cycle
CDK4	24%			Cell Cycle
INPP4B	30%			PI3K/AKT signaling
NOTCH1			6%	Notch signaling
AKAP9			5%	Binds to Protein Kinase A
NUP214			3%	Nuclear Pore Complex

3.2.3 Previous Work: Network-Based Approaches to Finding Oncogenic Collaborations

A number of network-based approaches of assessing the functional impact of mutations have been used on cancer genomic data. These network-based approaches essentially search for oncogenic collaboration by highlighting important interactions within the network of interest. Gulati et al summarize a number of these network-based approaches.[116] Most approaches utilize a protein/protein interaction network such as the Human Protein Reference Database (HPRD), or STRING, although there are a few that are utilize known transcriptional networks. By annotating mutations on these networks, they hope to ascribe certain properties to these mutations, such as connectivity, or path distance to important signaling proteins. We summarize three network-based approaches: MEMo, HotNet, and DriverNet.

MEMo (Mutual Exclusivity Modules in Cancer) is an approach that integrates network information into a search for mutually exclusive mutations across a cancer population.[37] It has been applied to glioblastoma, and also breast cancer mutations observed in TCGA.[98] Essentially, it limits the search for mutually exclusive mutations to those genes that only have direct protein-protein interactions. Limiting the search using the PPI network is useful because exhaustively searching all combinations of genes is computationally prohibitive. Within the breast cancer results, a number of pathways with significant mutually exclusive mutations across methylation, copy number variation, gene expression, and exome muations were observed. These pathways included PI3K/Akt, RB1 (involved in cell cycle), and TP53 (cell survival). Additionally, MEMo has been used to analyse Breast Cancer patients within The Cancer Genome Atlas consortium.[98]

Vandin et al describe a network-based algorithm called HotNets for detecting significant mutations in cancer patient populations.[118] They tested this algorithm using glioblastoma (GBM) data derived from TCGA and lung adenocarcinoma data as well. Using protein-protein interaction networks, they derived a 'neighborhood of influence' (essentially a subnetwork influenced by a mutation) for a mutation based on a diffusion-based process across the network. Essentially, a protein was considered to be influenced by a mutation if there was a direct path (or flow) through the network between the two nodes. Secondly, a threshold for significance of the subnetworks was defined using false discovery rate (FDR), by both permuting node labels and randomly distributing mutations in the subnetworks. Vandin et al found that in the GBM dataset significant subnetworks that recapitulated the Ras/PI3K pathway, a pathway previously found significantly mutated in cancer.

DriverNet is another network-based approach that attempts to relate mutations to expression using a network derived from known pathways.[119] Briefly, a influence network of genes is derived from known pathways and mutations are imposed onto nodes in this network. Next, expression data is overlaid onto the network and a bipartite graph is then constructed with the two types of nodes being expression and gene nodes. Those genes in bipartite graph that are connected to the most expression events are classified as drivers. This approach has the advantage of leveraging network structure, but again, drivers in the network are identified by frequency.

We suggest a simpler approach than MEMo, DriverNet or the HotNet approach. We examine a subset of genes involved in signaling and examine the possible role of neighboring mutations in their regulation as possible oncogenic collborators. We term the nodes in the subset that have a higher than expected number of mutated neighbors as *surrogate mutations*.

Because these signaling proteins have high connectivity, we define a statistical background model for deciding whether or not the number of neighbor mutations a signaling protein has is greater than expected. We suggest that surrogate mutations are a potential new model of oncogenic collaboration and they provide a potential role for many unique mutations that have been previously classified as passenger mutations. Additionally, we show that surrogate mutations are predictive of drug sensitivity and thus may be useful in the selection of appropriate targeted therapies.

3.3 Methods

3.3.1 Datasets Used

Three related datasets were used for this study. These datasets will be referred to as the mutation dataset, the GI50 values, and the expression dataset. The GI50 data was described in Heiser, et al.[21] This dataset consists of preprocessed data for the DREAM7 Drug Sensitivity Challenge. All GI50 values were measured by fitting three sets of dilution data to a dosage response curve.[15]

The mutation, copy number, and expression data were described in detail at the DREAM7 challenge website.[120] The mutation dataset was derived from Exon Capture sequencing using the Agilent SureSelect system with additional preprocessing steps. All sequences were aligned to the hg19 reference genome and filtered for quality. Because there was no matched normal sample for the tumor cell lines, mutations were called using a different procedure. Allele counts that went through the mutation calling pipeline required a high base quality (>= 10), a high neighborhood base quality (>= 10), and high mapping quality (>=20) for associated reads. The likelihoods of all possible genotypes at a site were calculated, and use as the input for a Bayesian model that incorporated the prior probability for the reference call, and

incorporating the heterozygous rate of the human genome. All heterozygous or homozygous mutants alleles were then filtered by the following metrics: genotype quality (>= 100), total depth (>= 8), and mutant allele strand bias (p-value < 0.005). Additionally, all mutations were filtered by whether the SNP occurred in DbSNP, a database of SNP variants.

The expression dataset is derived from Affymetrix Human 1.0 Exon Arrays. Gene level summaries were calculated using quantile normalization using the aroma.affymetrix package.

The Copy Number dataset consisted of Affymetrix SNP 6.0 chips which were normalizing using the same aroma.affymetrix package as the gene expression data. Copy number regions were defined by the use of a circular binary segmentation algorithm using the Bioconductor DNAcopy package. A reference sample of 20 normal tissue samples was used to call the copy number regions. The CNtools package was used to map Copy Number changes to the Gene Level. An absolute threshold for Amplifications and Deletions was derived by examining expression across all cell lines (Figure 27).

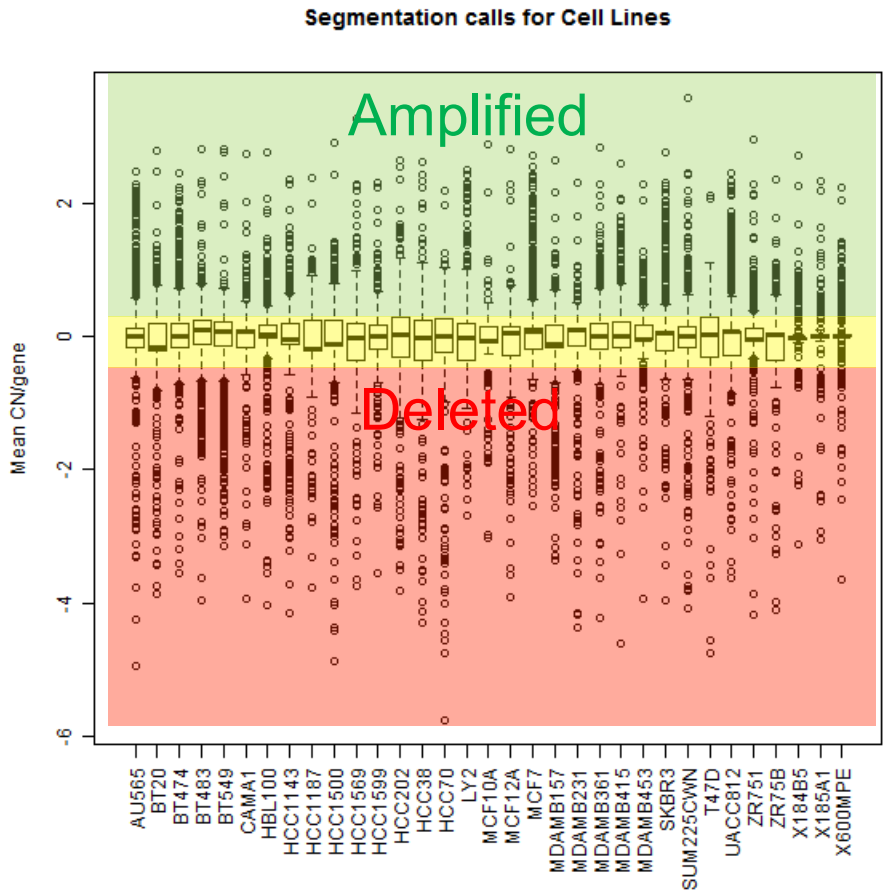


Figure 30. Calling amplified and deleted segments for the Copy Number Dataset.

3.3.2 Methods Aim 2-1: Annotating Mutations and Copy Number Alterations

The mutations (which were mapped to dbSNP), were also run through SIFT (as discussed in the Background chapter) in order to call them as deleterious or not.

Additionally, mutations were mapped to PFAM domains in order to assess their impact. PFAM domains are conserved segments of protein sequence that have known functions, such as enzymatic function or sequence recognition.[121] We have already encounted the SH2 domain, which is used by RTKs to specifically bind a phosphosite given a three amino-acid downstream sequence. Many other domains exist, such as SH3 domains. PFAM domains are summarized as Hidden Markov Models, or sequence-dependent probabilistic models based on all known sequences of domains.

Note that we do not use either of these annotations as filtering criteria for the data. They are used to annotate mutations in order to provide more information about these mutations to the user.

3.3.3 Methods Aim 2-1: Calculation of Surrogate Node Scores

The BioNet and igraph Bioconductor packages were used in the calculation of surrogate mutation scores.[122,123] Determination of mutated neighbors for a surrogate node was derived from a protein-protein interaction graph from the Human Protein Reference Database (HPRD).[44,73,74]

First, all genes in the dataset were mapped to the PPI graph using Entrez Gene IDs. This set of genes consists of the surrogate query set (surrogate nodes of interest) and all genes that were mutated in each cell line (cell line mutations).

Second, an adjacency matrix was derived from the protein/protein interaction graph. This matrix is a square matrix whose row and column labels consist of every node in the network. An entry of 1 for [gene1,gene2] in the adjacency matrix indicates there is an interaction between gene1 and gene2. Note that the protein/protein interactions in the HPRD data set are not directional. Thus the adjacency matrix is symmetrical about the diagonal. A given column or row for gene i in the adjacency matrix indicates all of the interactions that exist between surrogate gene i and all the other nodes in the network. Intersecting all genes with a 1 in this column with the list of cell line mutations gives both the number of mutations and the identity of all mutations that interact with the surrogate node of interest.

Significance of a surrogate mutation was determined through the use of permutations (Figure 31). The background model used was that of a randomly mutated network. The null hypothesis chosen was that the number of mutations for a surrogate node is not greater than

the number of mutations for that surrogate node in a random network. The alternative hypothesis is that the number of mutations for a surrogate node is greater than this background model. Specifically, blank PPI networks were randomly mutated with equal probablility and with replacement to produce a single permuted network. The number of mutations distributed to a background network was identical to the number of mutations observed within the cell line of interest. This random mutation process was done for 10000 permutations, generating 10000 background networks. For a surrogate node of interest, a distribution of number of neighboring mutations can be built for the set of background mutations and compared to the actual observed number. The p-value is then calculated by the fraction of background networks with the actual observed number of mutations or higher.

We explored the impact of two multiple comparison adjustment methods on our p-value distributions: the Benjamini-Hochberg method of adjustment, also known as q-value[71], and the Benjamini-Yekutieli (BH) method of adjustment, which adjusts for multiple comparisons when the data is nested.[124] For both methods, we adjusted the p-values in the TCGA-plus set within a cell line. However, we summarize the results across all cell lines. Our initial set before adjustment consisted of 741 significant surrogate mutations of the 7920 total mutations across all cell lines. After q-value adjustment, 53 of these mutations were still considered significant. After BY adjustment, only 4 mutations were considered significant. However, in light of the multiple comparison problem, neither approach is completely satisfactory. Each surrogate mutation is dependent on a varying number of other surrogate mutations due to the structure of the PPI network. Thus, correcting for this type of nested comparison is a difficult statistical problem and outside the scope of this dissertation.

3.3.4 Methods Aim 2-1: Initial Gene Set

The initial set of genes used in the query was the set of genes that were mapped to the RPPA antibodies. This set was initially chosen in order to integrate with the RPPA measurements described in Aim 1, and that they were chosen as a set of signaling proteins of interest.

However, this set was arbitrarily chosen and we desired a set that had a stronger basis. To this end, we used a set of TCGA genes as the start of a basis set.

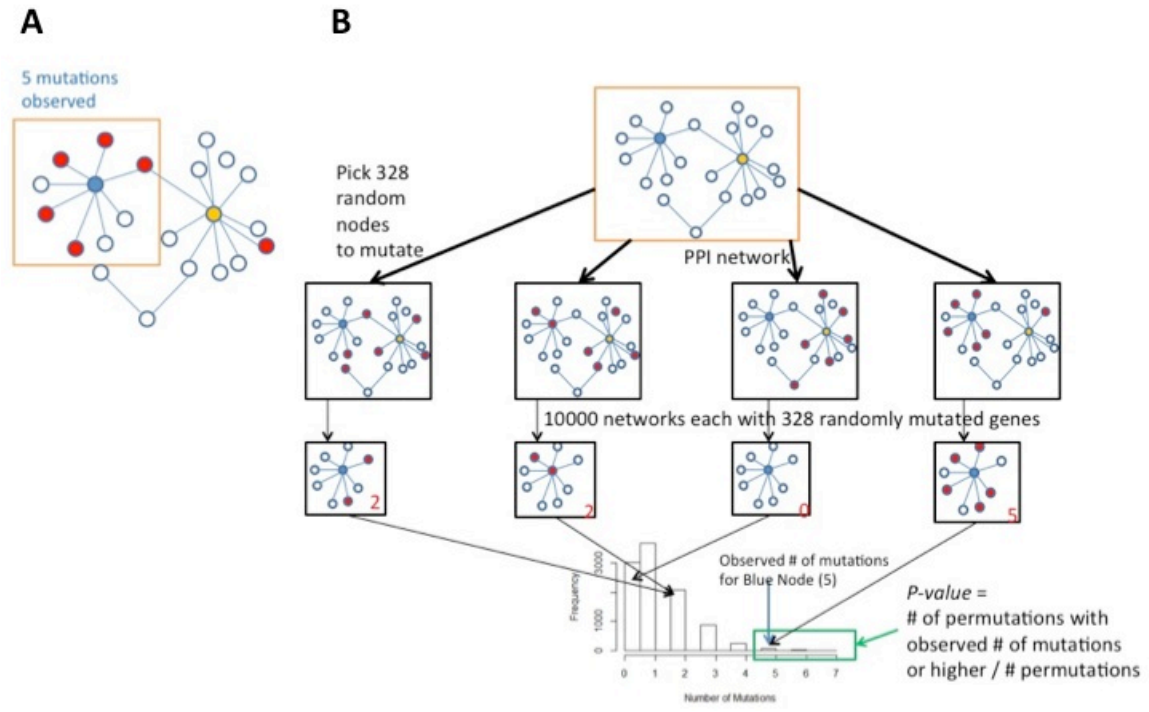


Figure 31. Methodology of Aim 2. In order to determine significance of the 2nd Order blue node (A), a permutation analysis is run in B). The blank PPI network is randomly mutated with the same number of mutations as the observed network of figure 31A.

3.3.5 Methods Aim 2-1: Determination of TCGAplus Query Set

Because the selection of the RPPA gene set was arbitrary, we decided on using the mutations in the TCGA Breast Cancer study as a basis set. Two lists of genes were used as the initial seeds for the query set: 1) genes highly mutated in all samples, and 2) genes highly

mutated within a specific subtype. These sets were derived from a frequency analysis of mutations for The Cancer Genome Atlas for Breast Cancer. [98] These 59 genes were mapped to the PPI network, forming a seed network (blue nodes in figure 32). From this seed network, a set of immediate neighbors were tabulated and ranked by the number of connections to the seed network. A threshold was drawn (greater than 1 connection in this case) and those neighboring nodes with greater than the threshold (red nodes) were included into the query set, resulting in a list of 180 genes total.

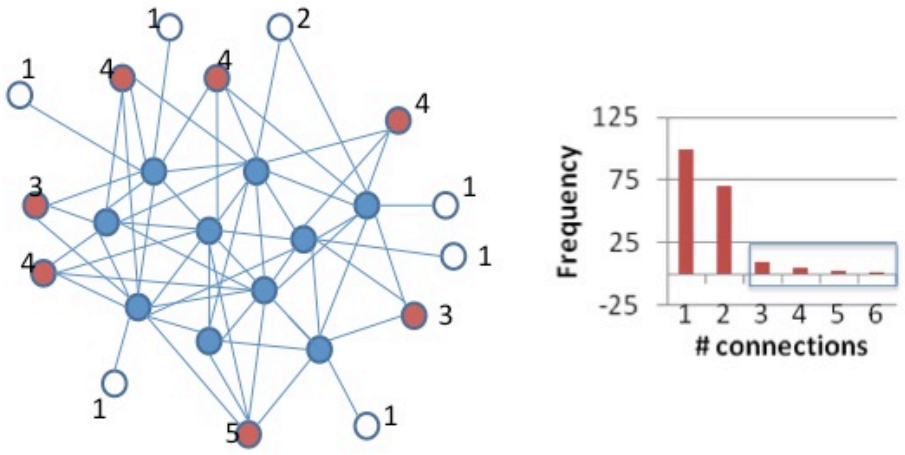


Figure 32. TCGAplus extension. Blue nodes are TCGA nodes, while white and red nodes are 1st order neighbors. The distribution of 1st order neighbor connectivities to TCGA nodes are calculated and a threshold of connectivity is called (> 2 in this toy example case).

3.3.6 Methods Aim 2-1: Surrogate Expression Scores

Additionally, the acquisition of mutations surrounding a surrogate node may not be of functional significance if the surrogate node and its neighbors are not expressed. In determining whether drug sensitivity is associated with surrogate features, we tried to incorporate an expression measure for the surrogate node.

Yang, et al describe a robust method for summarizing genes at the pathway level using rank for microarray expression data.[125] They report the average rank (AR) of the members of a

pathway as summarization method, normalized by the total number of genes on the array. AR has several desirable properties. First of all, it is a non-parametric metric of expression, which makes it a robust way to compare expression across multiple conditions. Additionally, the AR value ranges from 0 (downregulated) to 1 (upregulated).

Two modifications of AR were used to summarize the contributions of neighbors to the surrogate node: Weighted Average Rank (WAR) and Weighted Average Rank across Cells (WARC).

$$WAR_{self} = w_{self} * RANK_{self} + \sum_{i}^{n} \frac{w_{neighbors}}{n} * RANK_i$$

Briefly, for a node of interest, the contributions of itself and its neighbors to the surrogate expression value are determined by two sets of weights: w_{self}, which is the weighting of the node of interest, and $w_{neighbors}$, which is the total weight assigned to all the neighbors, spread equally among the neighbors. For WAR, the ranks are derived across all of the genes in a cell line; for WARC, the ranks are derived from the ranking of a gene's expression across all of the cell lines.

Both WARC and WAR were used as expression modifiers to modify surrogate scores by multiplying WARC or WAR times the binary surrogate feature (0 if not significant, 1 if significant). These modified surrogate scores were used as inputs to the Random Forest classifer used for validation, described below.

3.3.7 Methods Aim 2-2: Validation of Surrogate Mutations Using Random Forest Classifiers

Surrogate scores are only useful if they can help in clarifying the roles of mutations in drug sensitivity. In order to validate their usefulness, we used them as input features to a Random Forest (RF) classifier. Briefly, Random Forest classifiers are a highly robust approach to classification that are used in multiple bioinformatics applications.[126] They can be considered an ensemble machine learning method, in which a classification is decided by the votes of multiple machine learners.[127] The multiple machine learners in this case are decision trees. Each decision tree is grown from a separate bootstrap sample of the data, roughly divided into a training set (2/3 of the data) and a test set (1/3 or the data). Because each classifier sees a slightly different portion of the data because of the bootstrap approach, Random Forest classifiers are shown to avoid overfitting and tend to be more generalizable than other machine learning algorithms such as Neural Networks.[128]

One advantage of the RF approach is that a similar measure to cross-validation error is automatically calculated. This error is known as the "Out Of Bag" (OOB) error and represents the incremental discrimination error as a tree is added to the forest, up to the total number of trees in the forest.[128] OOB error has been shown to be a stable estimate of the cross-validation error, and thus we use it as an estimate for the generalizability of our random forests.

Cell lines were sorted by GI50 and divided into high and low sensitivity lines through the use of equiprobable binning (Figure 33). This method was chosen because the Random Forest algorithm is biased towards the largest group in unbalanced groupings and we wished to minimize this bias. There are some drugs for which the GI50 was not equally distributed, such as Methotrexate (Figure 34).

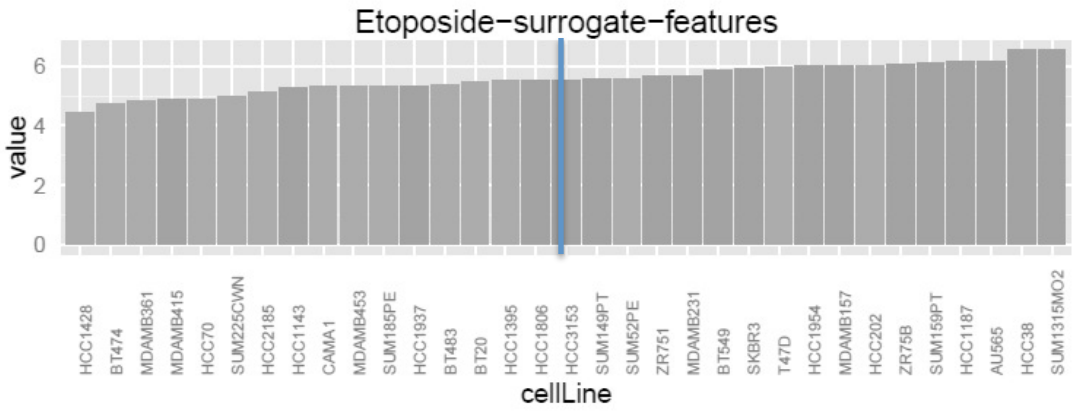


Figure 33. Binning of GI50 values in Etoposide. Location of boundary is indicated by the blue line.

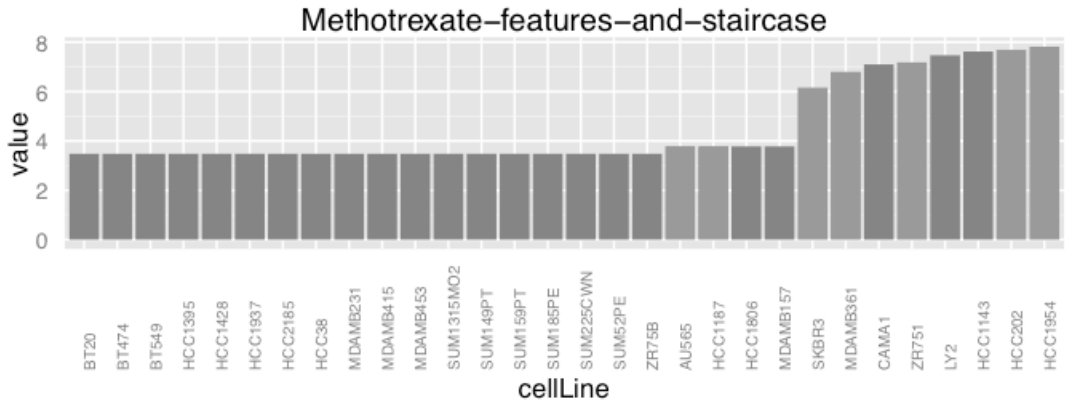


Figure 34. Methotrexate, a drug with uneven GI50 bins.

Table 11 shows the features that were used as inputs to Random Forests and their data types. We compared the performance of five sets of features: 1) Mutations called in genes within the TCGA plus set, 2) Surrogate mutations observed in the TCGA plus set, 3) Surrogate Features multiplied by the WARC expression measure, and two combinations of Surrogate and Mutation features. 4) Separate combination consisted of a concatenation of the two sets of features, while 5) combinedFlat: consisted of an "OR" union (1 if there was either a Surrogate or Mutated Feature". Of the 180 genes in the TCGA plus set, only those genes that had 2 surrogate mutations or more were used, narrowing the set to 70 genes. Only 106 genes had observed mutations in the TCGA data across all 38 cell lines.

As a comparison, three sets of features derived from the 50 PAM50 genes were used (Table 12): 1) mutations observed in the PAM50 genes, 2) the actual expression values for the PAM50 genes, and 3) a concatenation of 1) and 2).

Additionally, Breiman has noted that the overall OOB error for a Random Forest can be improved if the features are filtered by a variable importance measure such as Mean Gini Importance. Mean Gini importance is an impurity measure that measures the probability that a random feature would be incorrectly labeled if this feature was randomly labeled according to the distribution of labels in the dataset. Using this suggestion, we initially ran Random Forests with all features, and then filtered the features by the top 20% Mean Gini Importance before rerunning the Random Forest algorithm on these filtered features.

Table 11. Input features for Surrogate Random Forests.

Feature Name	Description	Total Features	Feature Type
Mutated Features	Observed Mutation in TCGA plus set or Not	106	Binary (0,1)
Surrogate Features	Observed Surrogate Mutation, Yes/NO	70	Binary (0,1)
Surrogate x Expression	Observed Surrogate Mutation x Expression	70	Continuous, -1 to 1
Separate Combination	Concatenated Surrogate Mutated Features	176	Binary (0,1)
Flat Combination	Observed or Surrogate Mutation, Yes/No	106	Binary (0,1)

Table 12. Input features for PAM50 classifiers

Feature Name	Description	Total Features	Feature Type
PAM50 Mutation	Observed Mutation in PAM50 or not	50	Binary (0,1)
PAM50 Expression	Expression value in PAM50 gene	50	Continuous
PAM50 Mut + Exp	Concatenated Mutation + Expression data	100	Binary, Continuous

3.4 Results

3.4.1 Results for Aim 2-1

In this section, we will discuss the statistical findings for surrogate features across the multiple cell lines.

Anatomy of a Surrogate Mutation. In Figure 35A an example of a Surrogate Mutation, SHC1 can be seen for the cell line SKBR3. The Blue Nodes are mutations, with mutations that were called by SIFT as deleterious in darker blue. Copy number amplifications are Pink, while Copy Number Deletions are green. Below the subnetwork (Figure 35B) is the background distribution calculated by permutation analysis. The actual number of surrounding mutations, 17, can be compared with the background distribution. Additionally, if the mutation was mapped to a PFAM domain or clan, we have included that annotation in the diagram.

RPPA set results. Visualizing significant mutations (p-value less than 0.05) for the 104 RPPA gene set across all 38 cell lines shows that surrogate mutations occur across multiple cell lines (Figure 36). In the matrix representation, a filled box indicates that there was a significant surrogate mutation observed in that gene for that cell line. Some genes were more frequent than others; for example BRCA1, ERBB2, and the estrogen receptor gene ESR1 occurred in over 21 of the cell lines.

TCGA plus results. Figure 37 shows all significant mutations for the 180 gene TCGA plus set. Only those genes that had 2 or more surrogate mutations are shown. 39 of the genes in the RPPA set also occurred in the TCGA plus set.

High Frequency Candidates. A number of genes in the TCGA plus set had a high frequency among the 49 cell lines, most notably BRCA1, implicated as a frequently mutated gene in breast cancer patients, and ESR1, an estrogen receptor gene implicated in cancer (Table 13). UBB is a protein involved in protein degradation, SMARCB1 is involved in relieving repressive chromatin structures, PHB mutations have been associated with breast cancer and is thought to be associated in transcriptional activity.

Table 13. High Frequency (> 10) Surrogate Mutations

Gene	Frequency
BRCA1(672)	23
ESR1(2099)	19
SMARCB1(6598)	19
UBB(7314)	18
PHB(5245)	17
PML(5371)	17
PRKAR2A(5576)	17
E4F1(1877)	15
MAPK9(5601)	15
PIK3R3(8503)	14
GSK3B(2932)	13
MNDA(4332)	13
PTK2(5747)	13
ABL1(25)	12
MAPK8IP3(23162)	12
CDK6(1021)	11
PPARBP(5469)	11
STK11(6794)	11

Surrogate Mutations in Drug Targets are not associated with Drug Sensitivity. The TCGA list of surrogate mutations was cross-referenced with the list of drug targets to produce a list of surrogate mutations that were also drug targets. These surrogate mutations were mapped to 23 drugs. Of the 23 drugs, very few showed a high association with drug sensitivity (Figure 38).

Expression does not improve the association. Since surrogate mutations are dependent on expression, it was hoped that integrating an expression measure for the subnetwork would weight surrogate nodes that were upregulated or downregulated, making the now weighted nodes correlate with drug activity. So WARC was used as an expression multiplier on the above surrogate mutation/drug targets at multiple weightings. However, for

the 11 drugs that had expression measures and surrogate drug targets, no increase in association was noted (Figure 39).

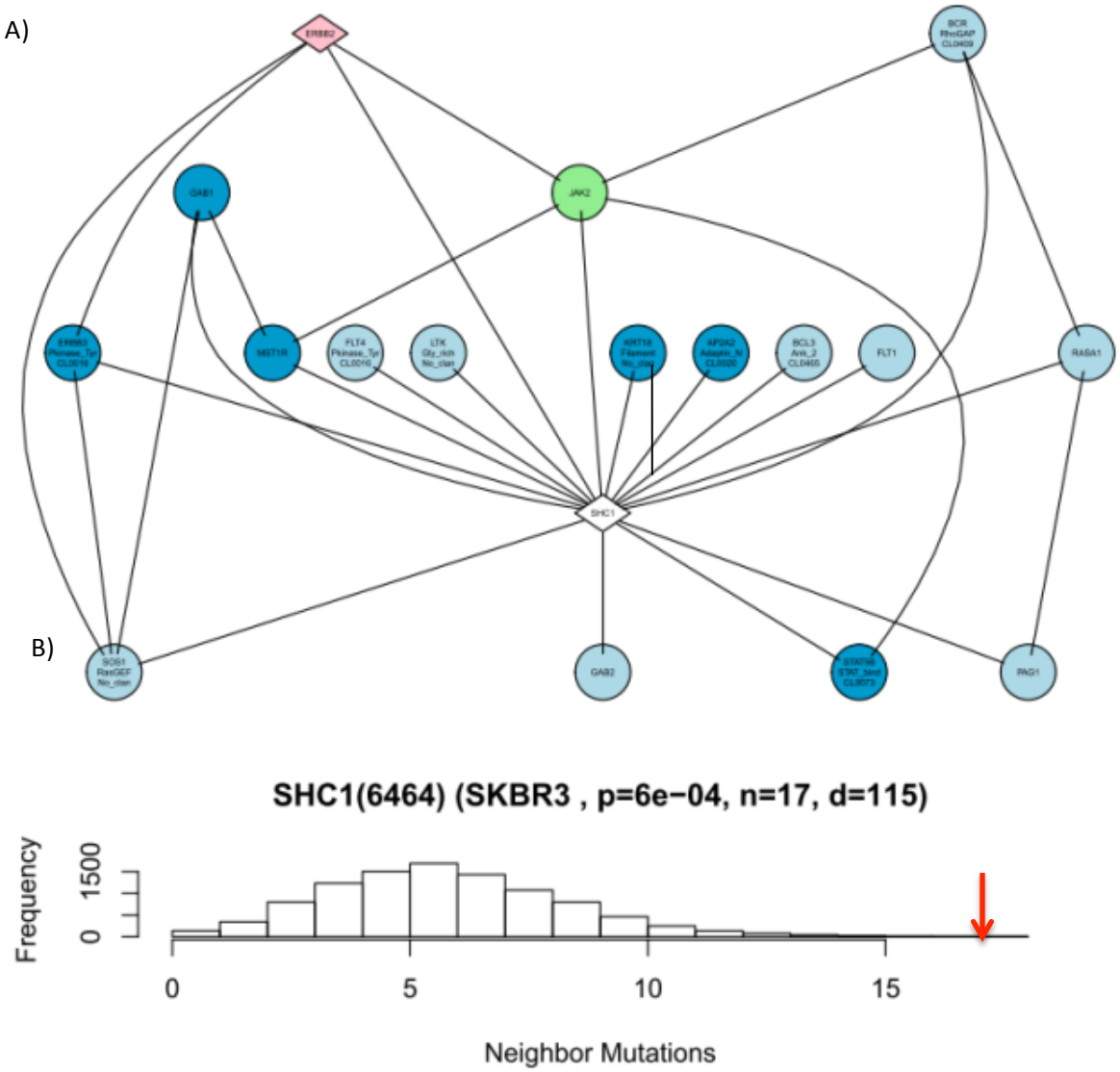


Figure 35. A) Current evidence for 2nd order nodes. A) Example of a 2nd order node, with mutations in blue (deleterious mutations called by SIFT are dark blue), pink is focal copy number amplification, green is deletion. n refers to the number of neighboring mutations, d is the total number of neighboring interactions and p is the p-value. B) shows the null distribution for that node calculated by permutation analysis. This node is highly significant for the background model, with an n of 17 (red arrow).

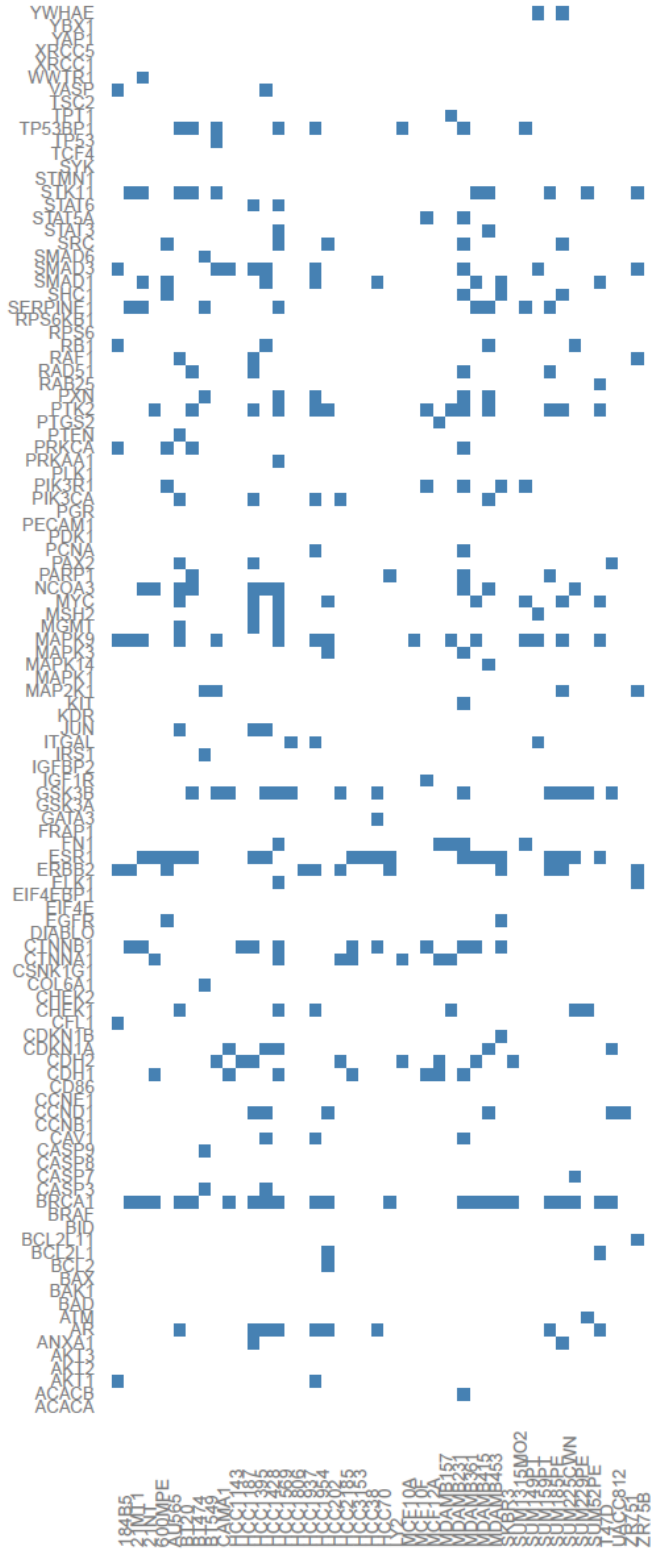


Figure 36. Distribution of significant surrogate mutations across the 49 cell lines for RPPA set. Blue boxes indicate a significant surrogate mutation (pvalue < 0.05), white indicates insignificant.

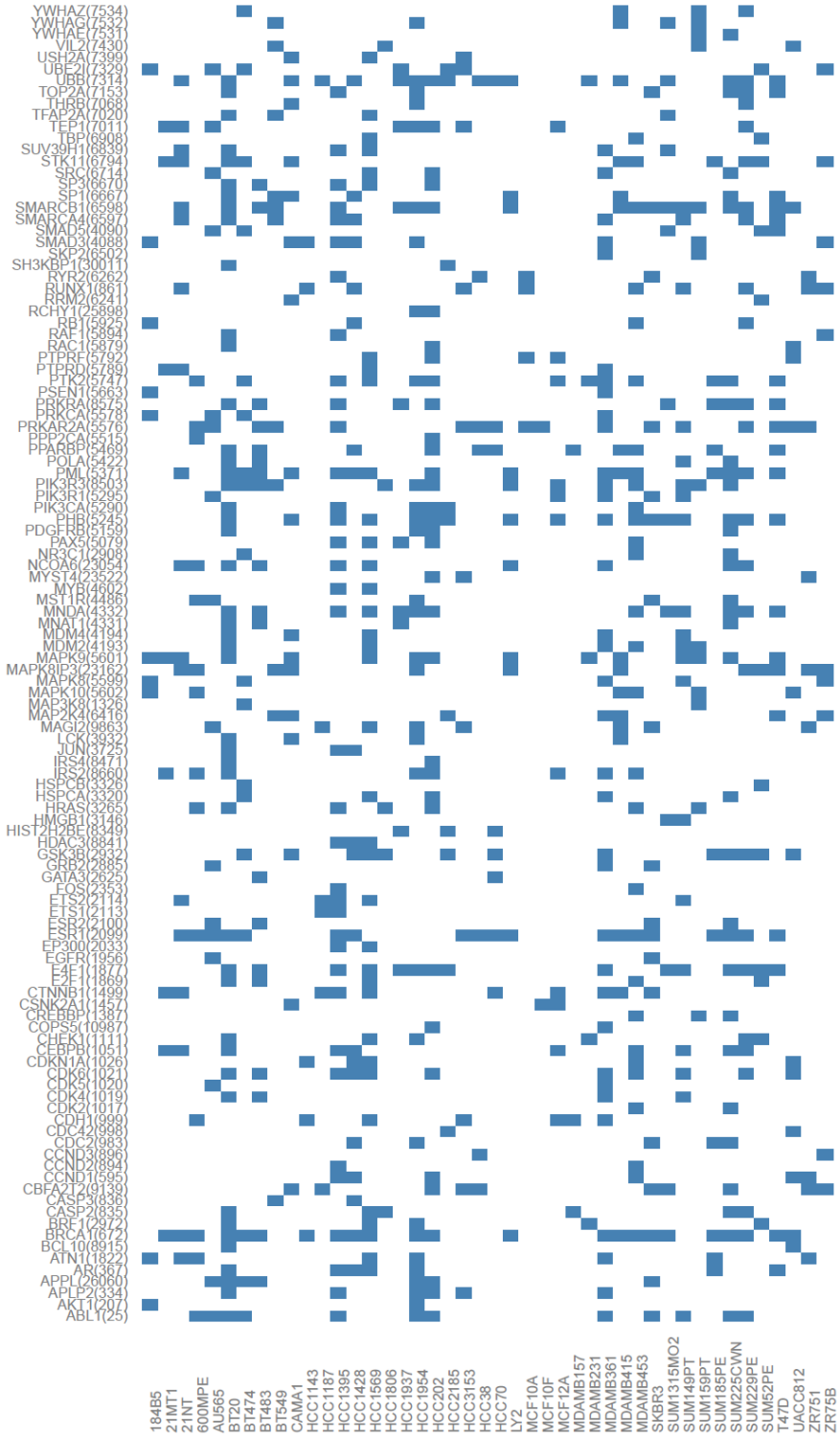


Figure 37. Surrogate mutation results for TCGA Plus Gene Set

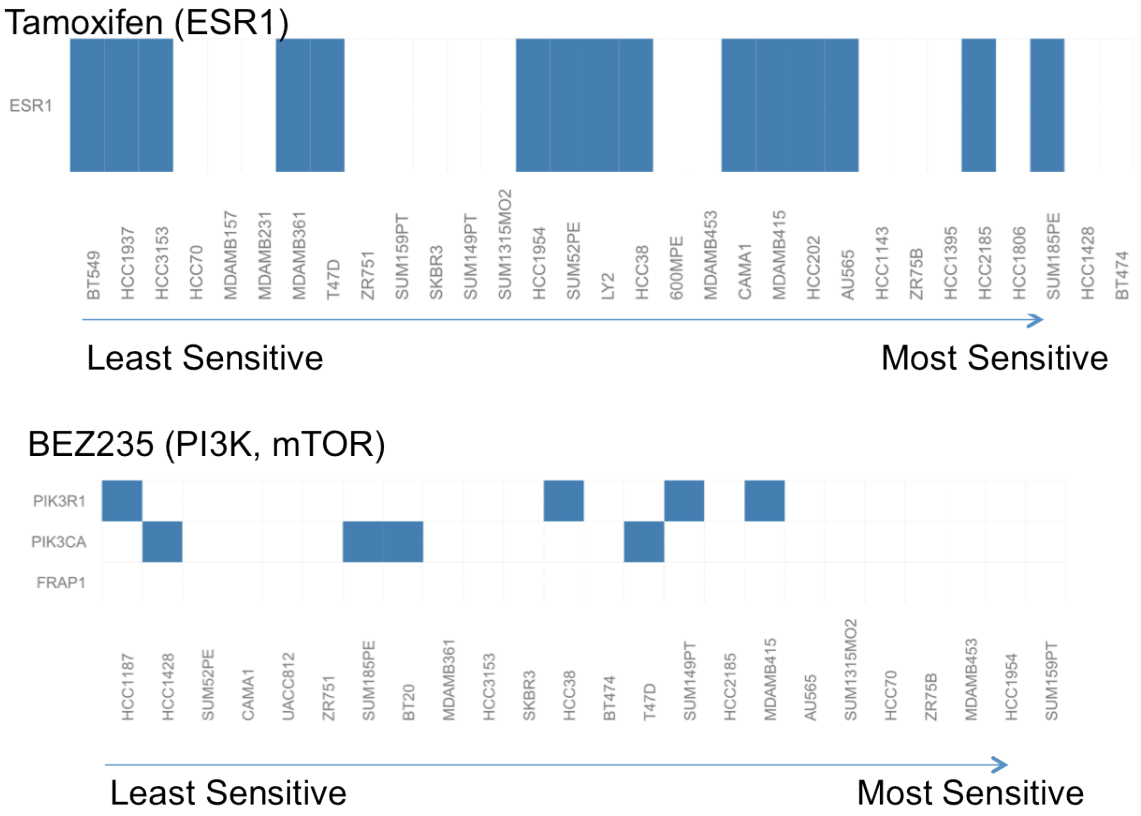


Figure 38. Surrogate mutations in drug target are not indicative of sensitivity. A box indicates a surrogate mutation was observed in one of the gene targets. Cell Lines are sorted by least sensitive to most sensitive.

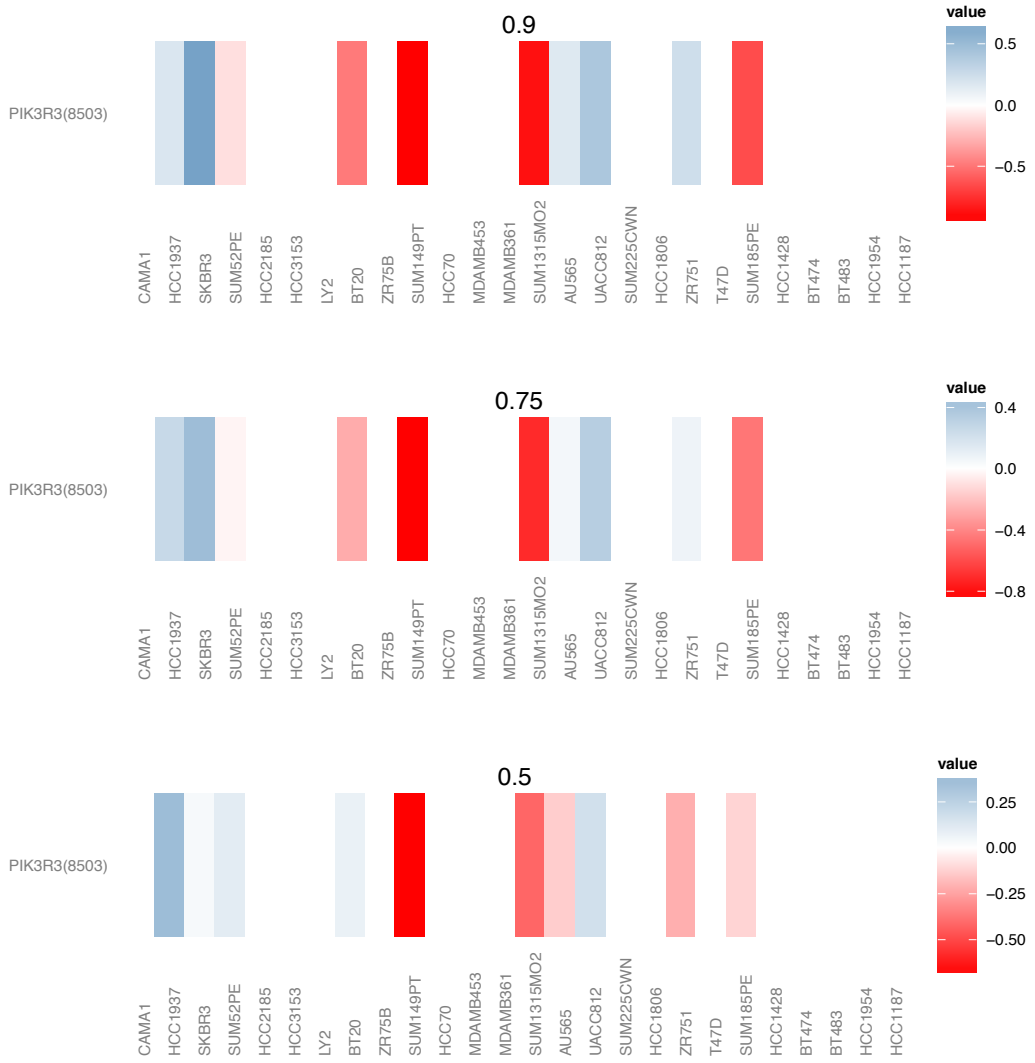


Figure 39. Integrating expression does not improve the picture. The drug AS-252424 (target PI3K) is shown, with WARC expression multipliers for each cell line (blue is Upregulated, and red is downregulated).

3.4.2 Results Aim 2-2. Validation of Surrogate Mutation scores

Filtering features by importance improves performance of Random Forests. Figure 40A shows the overall performance for each random forest learner over all 74 drugs for the surrogate mutation set. The mean error of this first run is 48.8%. After filtering the features by importance (Figure 40B), the mean error for the 74 random forests improves considerably, to

32.1%, indicating that filtering by importance is a necessary step. Thus, all results reported incorporate this step.

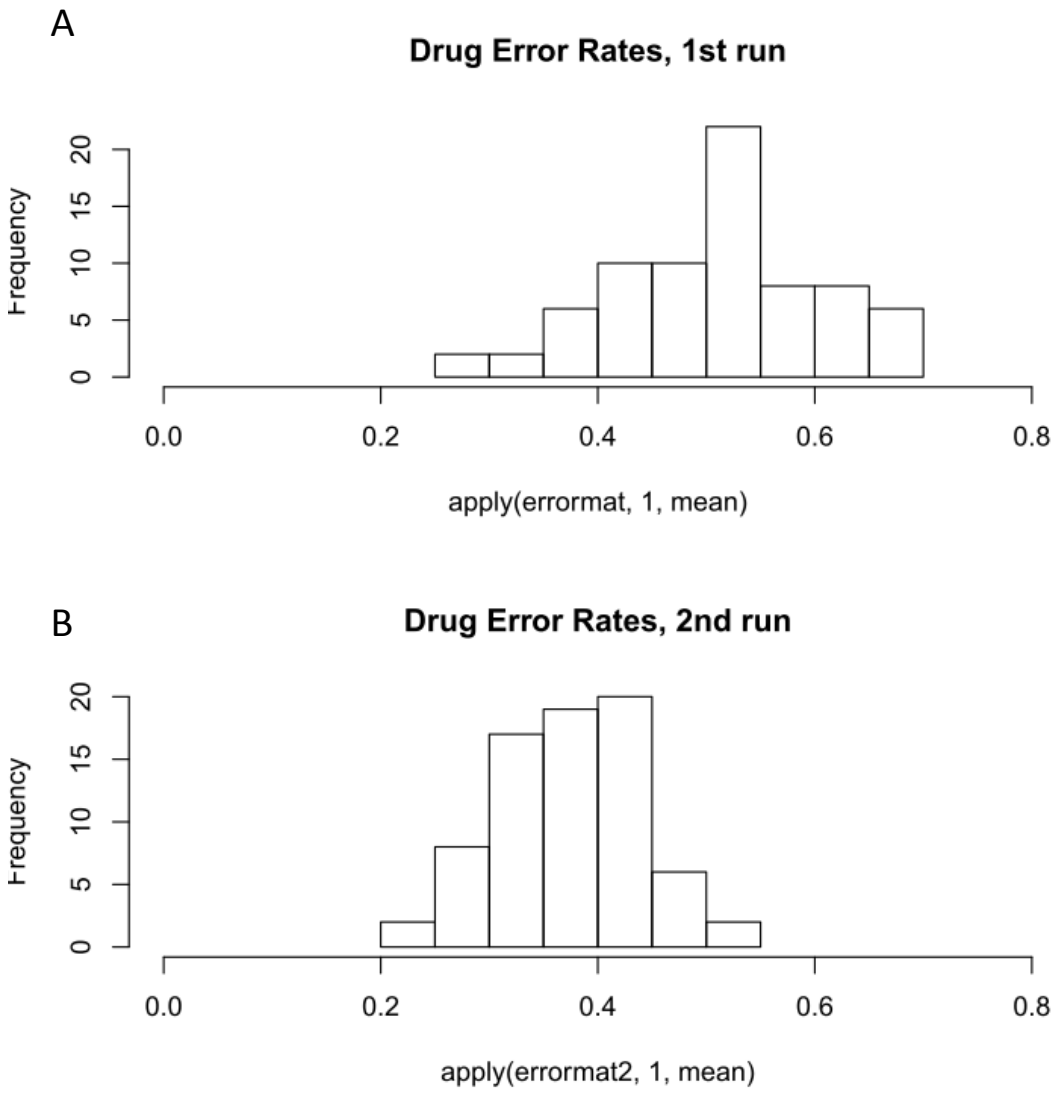


Figure 40. Filtering features by Mean Gini Importance improves classification rates. A) All 70 surrogate features. B) Filtering by importance.

Surrogate features classify better than mutated features. Figure 41 shows the comparison of the Random Forests derived from the mutated versus surrogate features. Each histogram shows the error rates of each Random Forest derived for the 74 drugs used. The mutated features classify with a mean error rate of 43.51%, whereas the surrogate features classify with a mean error rate of 32.11%, a notable improvement. Additionally, for the

combinedFlat features, the error rate improves to 30.9%, a slight improvement. We chose the combinedFlat features as the best set of features for further comparisons.

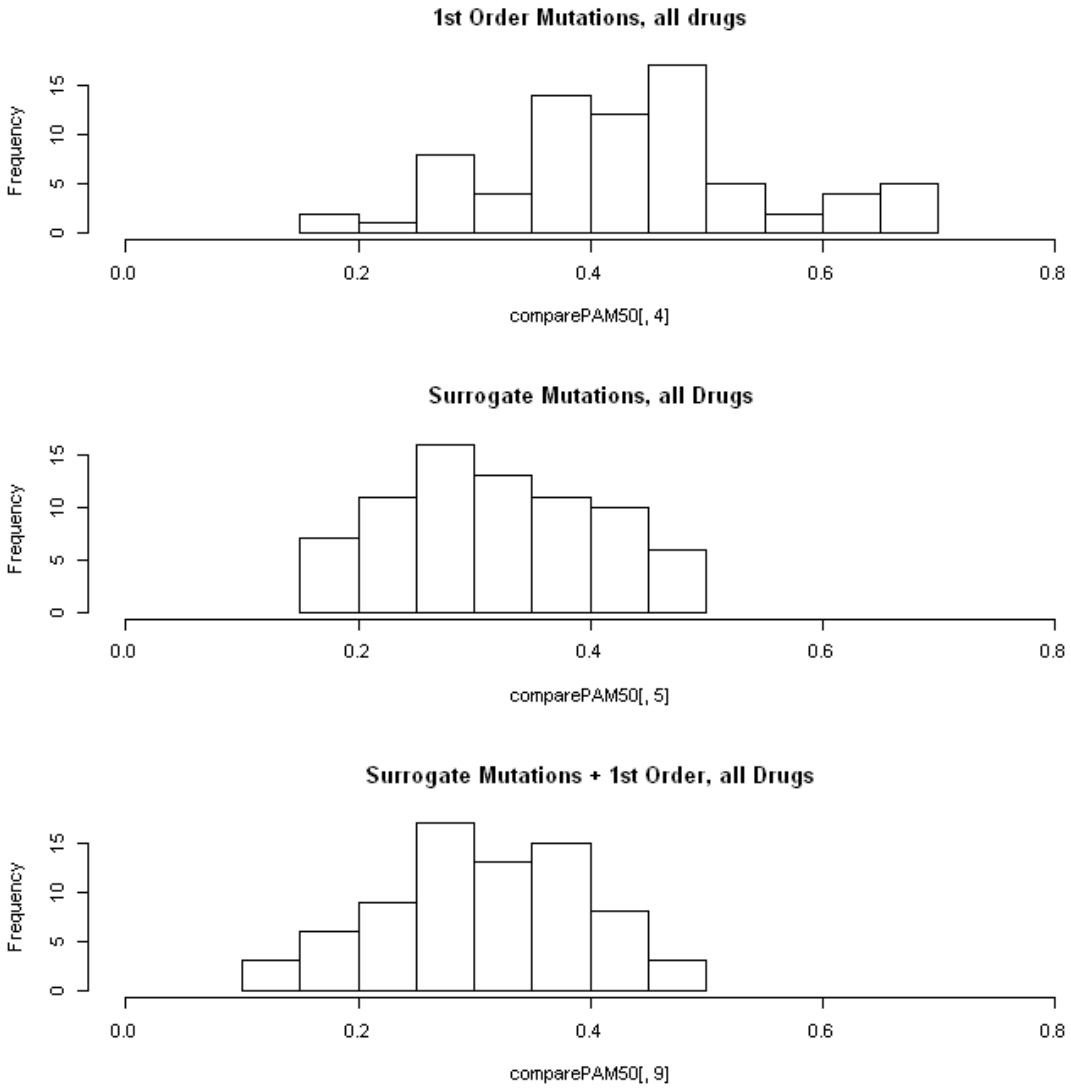


Figure 41. Comparison of Mutated versus Surrogate Features.

PAM50 versus Surrogate Mutation Results. In terms of mean OOB error, both the PAM50 expression forests and the combinedFlat features were roughly equal (29.1% error rate for PAM50, 30.9% error rate for combinedFlat). We compared those drugs for which the combinedFlat features and the PAM50 features were clearly outclassifying the other (within at

least a margin of error of two misclassifications). Table 14 shows all drugs for which the combinedFlat Forests outclassifed the PAM50 forests. One trend within the combinedFlat predictions is that the combinedFlat features clearly outclassifies for the platinum-based drugs (Oxaliplatin, Carboplatin and Cisplatin).

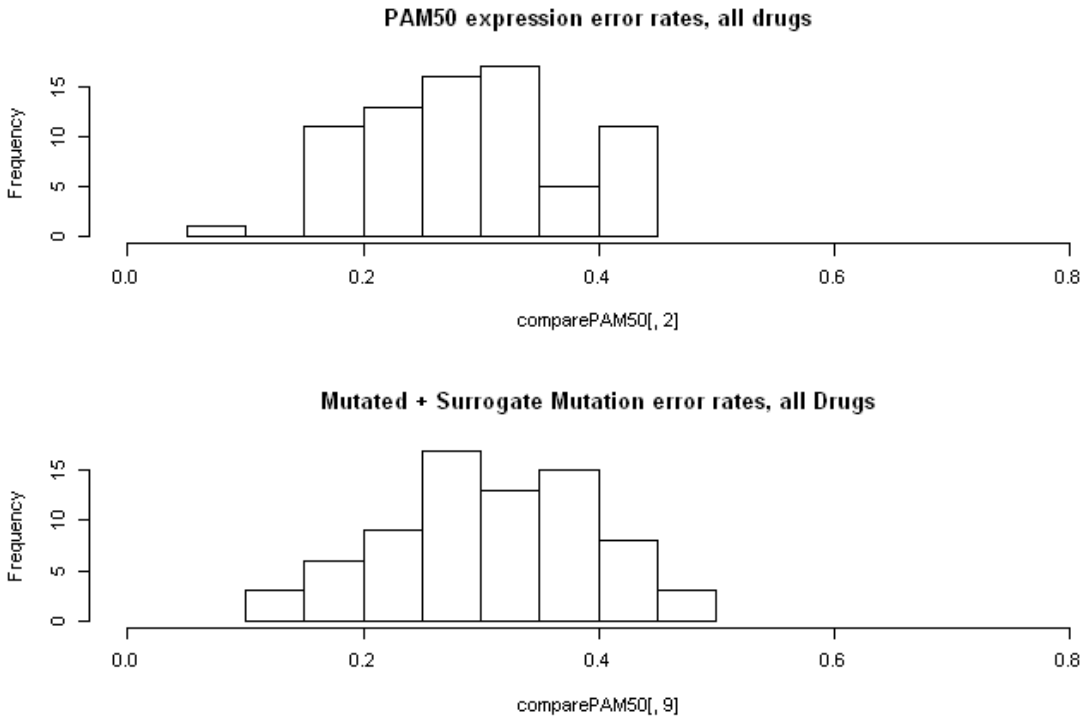


Figure 42. PAM50 error rates versus Surrogate Mutation error rates.

Table 14. Compounds for which surrogate mutations score better than PAM50 expression, by at least 2 votes.

COMPOUNDS	GeneTarget	PAM50exp	combinedFlat	Call
Sunitinib Malate	VEGFR	0.412	0.353	SURROGATE
AG1024	IGF1R	0.303	0.242	SURROGATE
TCS2312 dihydrochloride	CHK1	0.419	0.355	SURROGATE
AG1478	EGFR	0.167	0.100	SURROGATE
Methotrexate	DHFR	0.233	0.167	SURROGATE
NSC663284	CDC25S	0.333	0.267	SURROGATE
Topotecan	TOP2A	0.448	0.379	SURROGATE
GSK1487371	PIK3CG	0.321	0.250	SURROGATE
Lestaurtinib(CEP-701)	TRKA, FLT3	0.321	0.250	SURROGATE
ICRF-193	PLK1	0.407	0.333	SURROGATE
MLN4924	NAE	0.407	0.333	SURROGATE
ZM.447439	AURKA, AURKB, AURKC	0.423	0.346	SURROGATE
Cisplatin	DNA cross-linker	0.333	0.242	SURROGATE
Oxaliplatin	DNA cross-linker	0.273	0.182	SURROGATE
17-AAG	HSP90	0.200	0.100	SURROGATE
Bortezomib	NFKB1, NFKB2	0.286	0.171	SURROGATE
Trichostatin A	HDAC	0.371	0.257	SURROGATE
Ispinesib	KSP	0.412	0.265	SURROGATE
Carboplatin	DNA cross-linker	0.364	0.212	SURROGATE
AS-252424	PIK3R3	0.423	0.192	SURROGATE

Combining PAM50 scores with Surrogate scores. Because the distributions of the PAM50 expression forests and the combinedFlat forests look so dissimilar, we attempted to combine information from both of these forests to result in an improved learner. One method to combine two sets of random Forests is to take a weighted average of the votes from each random forest (Figure 43). We tried six sets of weights for the combinedFlat and PAM50expression random forests (Figure 44). Note that the error rates we report here are slightly different than the previous section, as they are direct error rates, rather than the OOB error. Overall, we note no improvement in the mean error of the combinedFlat random forests, as the lowest mean (30.1%) for all weight combinations was for the (1,0), or 100% combinedFlat, 0% PAM50 expression combination.

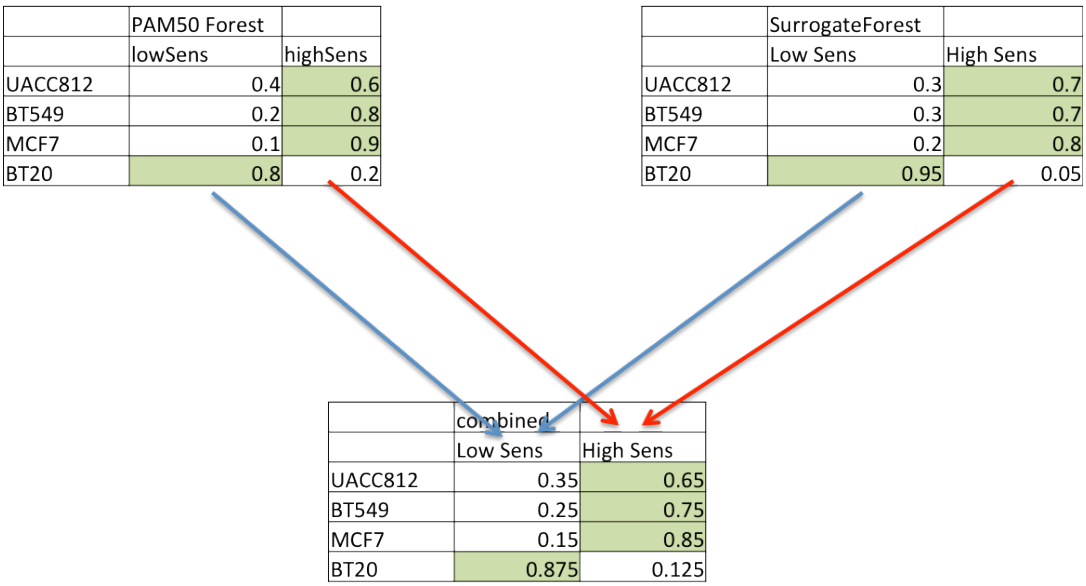


Figure 43. Combining PAM50 and Surrogate Forests via weighted average of votes.

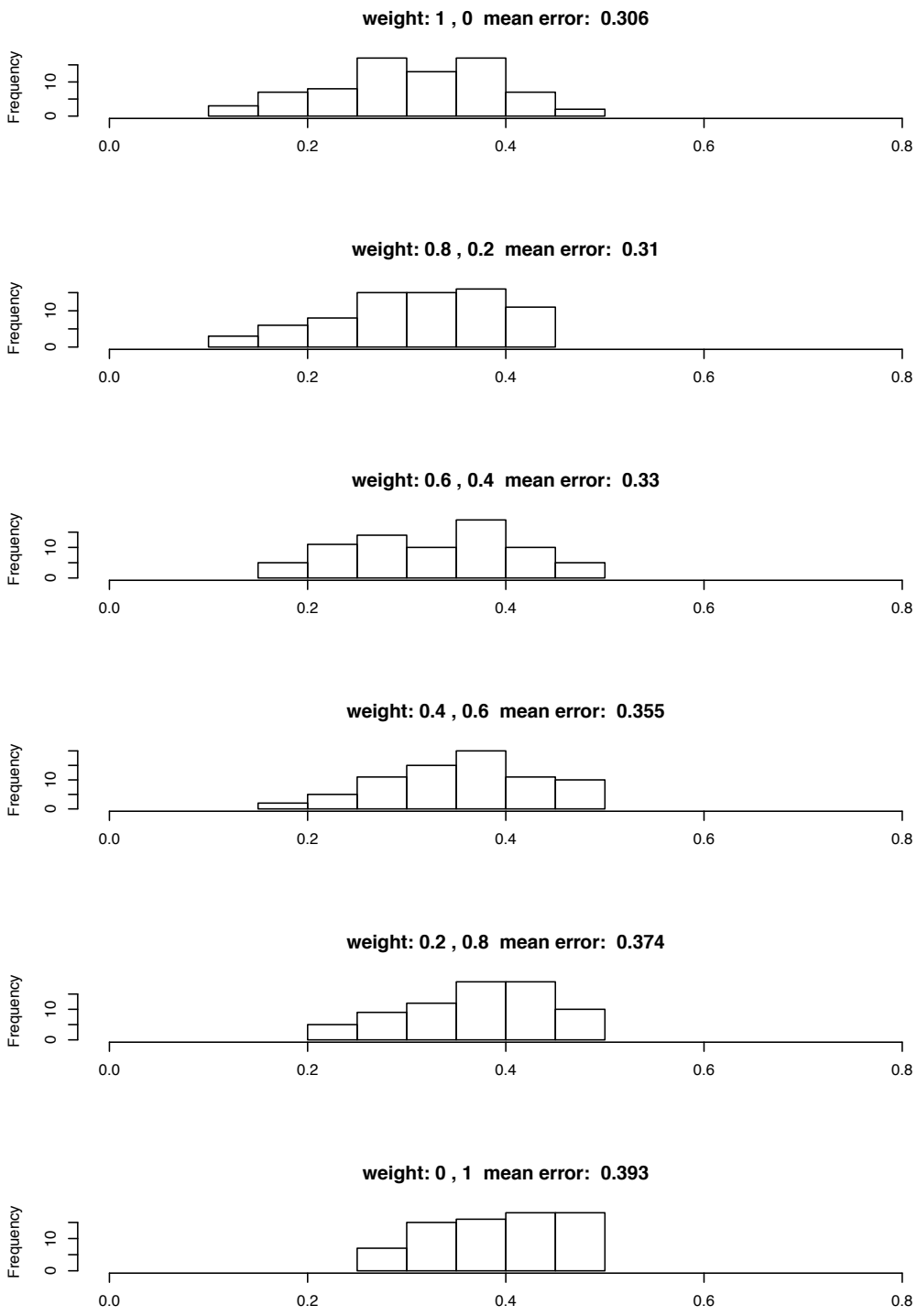


Figure 44. Results for Combined Votes with combinedFlat surrogate features and PAM50 expression. The top graph represents 100% combined Flat, 0% PAM50, the bottom graph represents 0% combined Flat, 100% PAM50.

Comparison of Surrogate Mutations with Subtype Specificity. One potential caveat to the use of Surrogate Mutation Scores is the potential of overlapping with Subtype Specificity. That is, the underlying biology of the significant Surrogate score subnetworks might be subtype specific.

One way to answer this caveat is to compare the clusterings generated by surrogate mutations to the PAM50 subtypes (Luminal A, Luminal B, Basal, HER2E). PAM50 calls for the cell lines were obtained from Neve, et al. Average-link Hierarchical Clustering was used on the TCGA plus surrogate mutation matrix (Figure 45) for each drug sorted by Mean Gini Importance. For each drug, the dendrograms produced by the clustering algorithm were cut to produce four clusters in order to directly compare with the four PAM50 subtypes.

As a measure of agreement, Cohen's kappa was used to compare the PAM50 clusters and the Surrogate Mutation clusters. Essentially, kappa summarizes the proportion of pairs of genes that co-occur across clusters to those that do not with a correction for chance using the hypergeometric distribution. Thus, two clusterings that have a high kappa are considered to be in close agreement to each other. A general rule of thumb of Kappa given by Altman is that 0.7 and above is considered good agreement, 0.7-0.4 is adequate agreement and 0.4 and below is considered poor agreement.[129]

Since the majority of the clusterings had poor agreement with the subtype groupings (kappa <= 0.35), we examined the OOB error rates of the Random Forests with the lowest kappa, less than 0.1. These 31 learners constituted 40.1% of the Random Forests (Table 15). Of these 31 Random Forests, 8 surrogate forests classified (by at least 2 more correct) better than the PAM50 expression forests, 13 PAM50 forests classified better than the surrogate forests, and 16 were roughly equal (within 1 or 2 correct votes on either side).

Of the 8 surrogate forests that classified the best, three are platinum-based drugs (Carboplatin, Cisplatin, and Oxaliplatin), which induce DNA cross-linking, leading to an apoptotic response in the cells. These drugs do not have a specific target, which may suggest why the network approach works better in these cases. ICRF-193 is a topoisomerase inhibitor. Trichostatin-A inhibits histone deactylase (HDAC). AS-252424 inhibits PI3K.

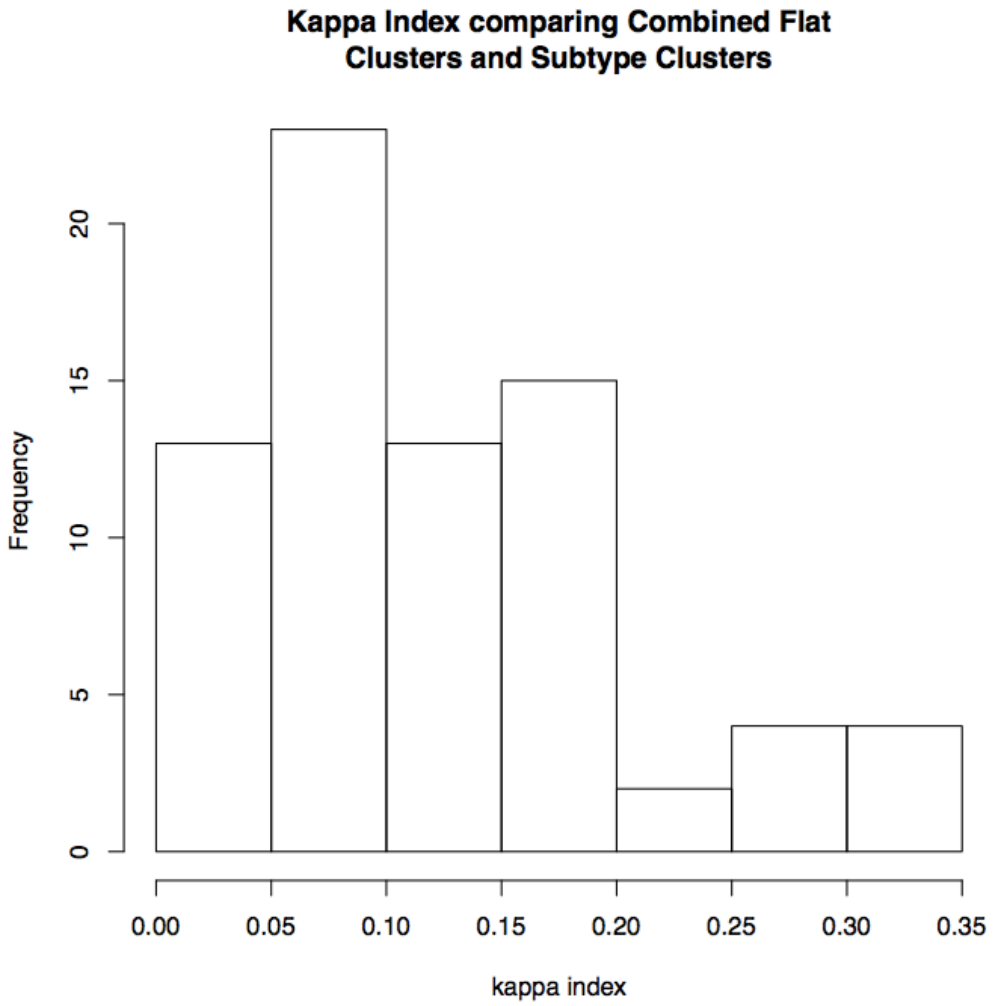


Figure 45. Distribution of Kappa values between combined flat clusters and PAM50 subtypes for all 74 drugs.

Table 15. Low overlap candidates with classification error.

Drug	Target	combinedFlat	pam50exp	Better RF
GSK1838705	IGF1R	0.457	0.229	PAM50
GSK1070916	Aurora kinase	0.400	0.200	PAM50
Glycyl.H.1152	ROCK	0.444	0.259	PAM50
XRP44X	Ras-Net (Elk-3)	0.407	0.222	PAM50
GSK2126458	PI3K	0.355	0.194	PAM50
GSK2126458	PI3K	0.355	0.194	PAM50
X5.FdUR	DNA	0.323	0.194	PAM50
VX.680	Aurora kinase	0.370	0.259	PAM50
LBH589	HDAC, pan inibitor	0.321	0.214	PAM50
Lapatinib	ERBB2	0.133	0.067	PAM50
Ixabepilone	Microtubule	0.353	0.294	EQUAL
AZD6244	MEK	0.261	0.217	EQUAL
Ibandronate sodium salt	farnesyl diphosphate synthase	0.276	0.241	EQUAL
PD173074	FGFR3	0.276	0.241	EQUAL
Purvalanol.A	CDK1	0.333	0.300	EQUAL
Vorinostat	Histone deacetylase	0.257	0.229	EQUAL
Etoposide	Topoisomerase II	0.242	0.242	EQUAL
TCS.JNK.5a	JNK	0.304	0.304	EQUAL
NU6102	CDK1/CCNB	0.290	0.323	EQUAL
Tamoxifen	ESR1	0.290	0.323	EQUAL
GSK461364	PLK1, Topoisomerase II	0.233	0.267	EQUAL
GSK923295	CENP-E	0.200	0.233	EQUAL
Temsirolimus	mTOR	0.300	0.333	EQUAL
Bosutinib	Src	0.296	0.333	EQUAL
TCS.2312.dihydrochloride	CHK1	0.355	0.419	SURROGATE
ICRF.193	PLK1, Topoisomerase II	0.333	0.407	SURROGATE
Cisplatin	DNA cross-linker	0.242	0.333	SURROGATE
Oxaliplatin	DNA cross-linker	0.182	0.273	SURROGATE
Trichostatin.A	Histone deacetylase	0.257	0.371	SURROGATE
Carboplatin	DNA cross-linker	0.212	0.364	SURROGATE
AS.252424	PI3K gamma	0.192	0.423	SURROGATE

3.5 Discussion

3.5.1 A New Model of Oncogenic Collaboration

In this aim, we have proposed a new model of Oncogenic Collaboration based on observations on how mutations are distributed in a protein-protein interaction network. It aims to go beyond simple two gene models of oncogenic collaboration, such as the "Ras-like" and "Myc-like" collaboration to a node-centric collaboration model that can incorporate the effect of unique patient mutations.

We proposed a simple network based background distribution of mutations in order to decide which surrogate mutations are potentially significant, and showed that these surrogate mutations are frequent across our panel of Breast Cancer Cell lines. We have also showed that these surrogate mutations are useful in predicting drug sensitivity of cell lines, with an average error rate comparable to the current standard of predicting drug sensitivity, that of PAM50 subtypes using Random Forest discriminators.

This model has potential implications for mechanisms of tumorigenesis. As we have seen in section 1.7, tumorigenesis is a multi-step process, in that tumor cells must acquire multiple oncogenic mutations or lose Tumor Suppressor function in order to proliferate and metastasize. Our model suggests another route to loss of function, in that mutations surrounding an oncogene or tumor suppressor gene could affect that oncogene or TSG. The acquisition of such neighboring mutations could be a subtler strategy of the tumor cell to deregulate these oncogenes or TSGs. In fact, Berger et al suggest that the current "two-hit" model of Tumor Suppressor Genes could be replaced with a model of continuous haploinsufficiency, citing such subtle changes as we have mentioned.[130]

Not adjusting for multiple comparisons has consequences. By not adjusting for multiple comparisons we must accept a much higher false positive rate for our surrogate mutations. This is not ideal, given that surrogate mutation subnetworks are highly overlapping. One possibility is to a method to adjust for nested comparisons, such as that of Benjamini-Yekulteili.[124] However, as we have noted, BY adjustment is not completely appropriate, as each surrogate mutation is dependent on other surrogate mutations to a varying extent. Finding an appropriate method for correcting for multiple comparisons is beyond the scope of this dissertation.

Are Surrogate Features Useful in Predicting Drug Sensitivity? We have shown that the Surrogate Mutation information classifies cell lines into high and low sensitivity classes at least as well as PAM50 expression information. The surrogate scores have an added benefit in that they are based on binary features – mutation and copy number – rather than the continuous features of PAM50 expression.

Surrogate Mutations capture different information than PAM50 molecular subtypes. We have also shown that even when the groupings between the two sets of features – Surrogate and PAM50 are very different, that Surrogate Features classify equally or better than PAM50 67% of the time.

Improvements to the Surrogate Computation. A number of improvements to the surrogate features can be made. Incorporating epigenetic information such as methylation of genes can be incorporated into a formal framework for identifying the Tumor Suppressor genes. However, one hurdle to integrating gene methylation is the lack of a good reference for the cell lines. The use of HapMap samples for methylation calling could be potentially used, with all of the drawbacks of using such a reference with the copy number samples.

Calculating Surrogate Mutations for the TCGA data. Another direction is to take the TCGA breast Cancer Mutations and Copy Number data and run it through the surrogate analysis workflow. Running this analysis would provide further evidence that Surrogate Mutations exist in patients and are not simply an artefact of cell lines.

3.5.2 Limitations of the Data has consequences for analysis

Because our downstream integrative analysis is dependent on upstream choices in analysis of each of the data types, we must understand the consequences of each of the choices made in each of these data types. In Table 16, we outline these limitations and consequences.

Copy number data limitations and consequences. For the copy number data, we are essentially limited to large genomic regions and those that are statistically varying across a high percentage of samples due to the choice of Circular Binary Segmentation and MutSigCV. Thus, our analysis does not cover smaller copy number amplifications that may arise due to alternative mechanisms such as segmental duplications. Thus we may overestimate the number of false negatives by using these large genomic regions.

Additionally, the reference chosen in the copy number analysis is derived from the average copy number profile of 20 HapMap samples. These samples may be biased in terms of population, which may limit our generalization to other populations.

By using an absolute call across cell lines to call a region, we are potentially increasing Sensitivity at the expense of specificity. However, examination of the normalized data showed similar copy number ratio distributions across the cell lines, so such an absolute call seems to be justified.

Mutation Data limitations and consequences. The first large consequence of the mutation data is that the NGS was limited to exonic regions by using an Exon Capture method. Thus we are limited to mutations within exonic regions. This choice may increase sensitivity (reducing false positives) at the expense of sensitivity (reducing false negatives). Other studies that have focused on whole genome sequencing have shown that mutations associated with disease also occur in promoter regions[131,132] and intronic regions.[133] Thus, by limiting the mutation calling to exons, we are potentially leaving out other mutations that could participate in this oncogenic collaboration.

The consequences of this limitation on the analysis are that our number of mutations are potentially limited, and by limiting the mutations, the networks may be limited in depth. Thus, we again lose potential surrogate mutation candidates by focusing on exonic mutations.

Drug Sensitivity measurement and consequences for analysis. We have seen that the dosage response curve varies in different measured sensitivity parameters across drug class.[15] Effectively, this variation has several consequences for the analysis. For a given drug class, the spread of GI50 can be effectively much smaller, and thus the classification problem is made more difficult. Repeating the Random Forest analysis with other measured parameters (such as HS, IC50, and EC50) could be potentially illuminating.

Table 16. Limitations of the use of Copy Number, Mutation, and Drug Sensitivity Data in these analyses.

Data Type	Limitations of Data Platform	Consequence for Analysis
Copy Number	Analysis focuses on large genomic regions (CBS)	Loss of oncogenes/TSGs with smaller indels
	Analysis only focuses on varying regions	Loss of oncogenes/TSGs with smaller indels
	Reference is 20 HapMap samples	Reference may already have amplified/deleted regions; loss of specificity
Mutation Analysis	Mutation is called using human reference genome	Loss of specificity; may have a larger number of false positives
	Mutations are Exon Focused	Loss of mutations in regulatory regions; loss of mutations due to gene fusions
	Base calling is done using a Bayesian algorithm that incorporates expected frequency	Potential loss of rare variants
	Coverage may not be deep enough to capture rare variants	Loss of rare variants
Drug Sensitivity (GI50)	Measure incorporates growth relative to untreated sample	Growth and sensitivity of population to growth inhibition are conflated
	Dosage response curves vary over drug class	GI50 may not be the most discriminating measure of sensitivity in drug class

3.6 Conclusion

In this chapter, we have shown a method for integrating mutations using networks, that of surrogate mutations. We suggest that these surrogate mutations may be another form of oncogenic collaboration. Surrogate mutations integrate mutations that were possibly considered previously as unique passenger mutations.

While surrogate mutations are not directly predictive of targeted drug sensitivity, we have shown them to be able to predict drug sensitivity when used as features in a Random Forest classifier, with a mean OOB error rate of 30.1% over 74 drugs, compared to the OOB rate for the PAM50 classifier of 29.1%. Over all 74 drugs, the surrogate forests classify better or equal to the PAM50 forests 66% of the time. We have additionally shown that the intrinsic

molecular subtypes determined by PAM50 are different from the groupings produced by the surrogate mutations, with all clusterings from the surrogate mutations having a kappa of 0.4 or less, indicative of poor agreement between the two sets of groupings.

Chapter 4: Discussion and Conclusions

In the previous two chapters, we have highlighted approaches for characterizing cross-phenotype response and proposed a new type of oncogenic collaboration. In terms of characterizing cross-phenotype response, we have highlighted a statistical approach (ANOVA) using high sensitive cell lines as replicates, and the use of AUC as an effective method for identifying outputs of signaling that are correlated with drug sensitivity. In terms of oncogenic collaboration, we have shown that surrogate mutations are at least as predictive of drug sensitivity as the current best set of features, PAM50 expression.

The remainder of this discussion is organized as follows. We will first highlight some key points about how the analysis of the data types affects our analysis. Then for surrogate mutations we will discuss challenges to the use of surrogate mutations can as markers of drug sensitivity in precision medicine. Finally, for the RPPA data, we will discuss methodologies for incorporating mutations in modeling, which could potentially provide mechanistic enlightenment of oncogenic/TSG mutations and their effect on drug cross-phenotype response.

4.1 Limitations of Each Data Type Has Consequences For The Analysis

Integration of multiple data types for prediction and discrimination will always be dependent on the upstream analysis choices for each individual data type. Given these difficulties, we have presented a first-pass attempt at integrating these data types. We have discussed many of these limitations in sections 2.6 and 3.5. We will summarize some of the key limitations and the subsequent consequences on our integrative analysis.

The drug sensitivity data is limited in that it utilizes GI50 as a metric. This is problematic for two reasons: 1) as we have seen, because GI50 uses relative growth as its y-axis, it conflates growth inhibition with expected growth under no inhibition and 2) the dosage response curves upon which GI50 are based vary in parameters due to drug class. The downstream effect of this on our analysis are that GI50 effects across cell lines for a drug class may be very narrow in range, making discrimination using our surrogate features more difficult. Thus, a potential future direction is to repeat this analysis using other drug sensitivity metrics derived from the curve such as HS and EC50.

For the RPPA data, we have focused on a relative effect approach, subtracting out the DMSO traces in our analysis. This was necessitated by the fact that DMSO by itself induces a response in the cell lines, and that any drug treatment that uses it as a delivery mechanism will conflate its effects with the actual drug response.

Additionally, due to the lack of replication, a robust summarization of the dynamics was needed, especially due to potentially spurious spikes in the data. For this summarization, we have used AUC in our analysis. AUC is reasonably robust to single spikes in the data and is a useful measure of finding outputs in the RPPA data that correlate with drug sensitivity.

One of the requirements made by the committee for this dissertation was to undertake the DREAM8 challenge. The choice of AUC as a robust metric necessitated the use of PLS-PM as a modeling strategy. The search for an appropriate modeling strategy was difficult. PLS-PM is not meant for prediction of the node trajectories as defined by the DREAM8 challenge but we have outlined a method in order to utilize it for prediction. By its nature, however, error propagation for later nodes can be high, which can affect downstream predictions.

One final improvement to the analysis was mentioned in Section 3.5.1, that of incorporating methylation data into the analysis. However, as we have seen, incorporation of methylation data means that we must integrate it into a formal Tumor Suppressor framework, integrating not only mutation calls, but also Loss of Heterozygosity calls with the Methylation Data. Such a framework is a future direction in itself.

4.2 Surrogate Mutations and Personalized Medicine

Challenges to the use of Surrogate Mutations in Precision medicine. The first challenge to the use of surrogate mutations is to confirm that they exist in patient data. Thus, repeating the analysis for patient samples such as TCGA is an important direction to pursue. Using patient samples increases mutation and copy number sensitivity (through the use of matched references), though at the expense of specificity. Additionally, more samples are available, which may increase our confidence that Surrogate Mutations are potentially present in the patient population.

Second, as they currently exist, surrogate mutations are a construct, not a truly validated genomic feature. Additional studies validating promising surrogate mutations are needed in the cell lines. In such studies, we need to quantify the number of unique patient-specific mutations in order to truly answer the question of whether surrogate mutations are integrators of these unique mutations.

Finally, as we have seen, tumor cell populations are continuously evolving under selection pressure to survive, and thus we must take heterogeneity of the tumor population into question. An unanswered question is whether heterogenous tumor cells actually support each other, and thus inter-clonal subpopulation networks may be of interest.

4.3 The Future: Incorporating the effects of mutations into ODE Models

In this section we will discuss a potential future application of the RPPA data: building deterministic models of cellular signaling and cross-phenotype response such as those based on Ordinary Differential Equations (ODEs). We outline one approach of incorporating driver mutations, and highlight a mechanistic approach to understanding feedback mechanisms incorporating mutation data, that of rate ratios.

A number of issues exist with building ODE models themselves. A major issue with ODE models is the identifiability of a model with respect to its Reference Behavior Pattern. That is, how unique is the parameter set with respect to producing a particular fit to the data? Gutenkunst et al, showed that over at least 15 biological models, a wide variety of parameters can fit the output.[134] They called such systems "sloppy" with respect to the parameter space.

One approach to handling both model sloppiness and observed heterogeneity is the use of ensembles of models. Rather than trust the output of a single model, we can spawn many models with observed variations in both protein levels (corresponding to observed heterogeneity) and rate constants (corresponding to parameter variation), and observe how such changes in both protein level and reaction rates affect the model output. The LHS approach can help us understand the global response of a model, and avoid the trap of local minima with gradient-based approaches. Instead of reporting a single trace, we report the observed range of outputs of the model and note levels and rate parameters that cause the largest changes in output.

Marino et al outlined a general method for global sensitivity analysis of parameters in a model using a sampling scheme called Latin Hypercube Sensitivity (LHS) analysis.[135] LHS analysis marries Monte Carlo methodology with a focused sampling scheme over the parameter space.

Each parameter is divided into a set of intervals. From each interval, a parameter is randomly sampled (Figure 46).

LHS provides us with a framework for testing the effects of mutations on the model. We will now discuss two studies that utilized ODE models for testing the effects of mutations on signaling models.

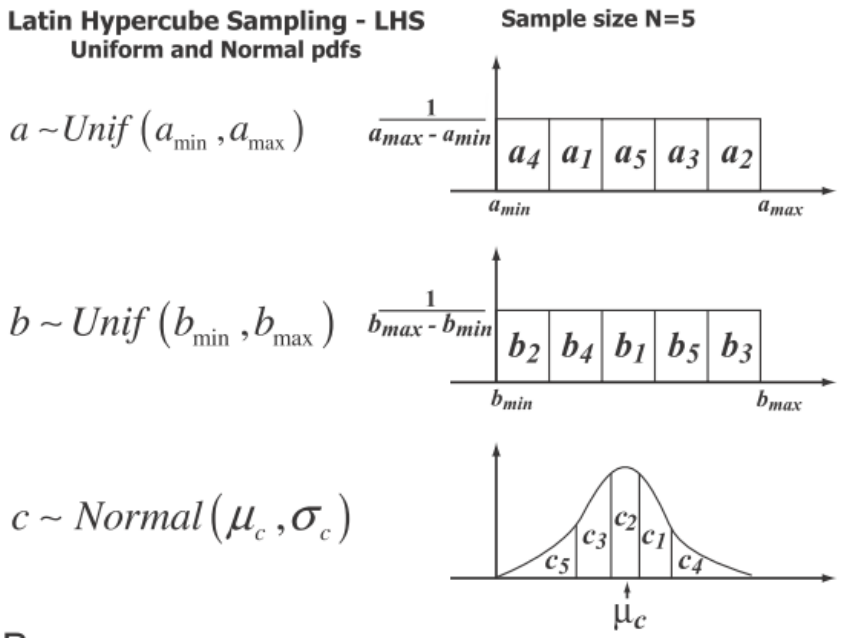


Figure 46. The Latin Hypercube Sampling framework for global parameter sensitivity analysis. Reproduced from Marino.[135]

Cheng, et al modeled and quantified the systemic impacts of 40 missense mutations observed in neuro-cardial-facial-cutaneous syndrome using a simplified ODE model in the MAPK pathway using a two step approach[136]. In the first step, each mutation in each protein was mapped onto a structural model of the protein and its effect on the free energy of the protein was assessed using molecular dynamic methods. Secondly, rate constants associated with the protein were perturbed in a simplified ODE model of the MAPK pathway and the effect on model output was assessed. The model output of interest was three different qualities of

phosphorylated ERK expression: amplitude, duration, and peak time differences. They were able to show that K-Ras mutations showed a very different impact in terms of modeling output compared to other mutations seen in the pathway.

Another future direction is the use of models to derive balance metrics that summarize the balance of activity in compensatory feedback loops. Harrison, et al. have probed the PI3K/PTEN feedback loop in detail, using a ordinary differential equation (ODE) model to relate missense mutations to sensitivity to the drug Pertuzumab.[137] Pertuzumab's target, the HER2 receptor, is upstream of the PI3K/PTEN feedback loop. Harrison's group summarized the enzymatic activity through this feedback loop with a single metric they termed Γ.(figure 47). Importantly, Γ, normalized to wild-type activity of these enzymes, integrates both the effects of mutation and of quantity. Sensitivity to Pertuzumab is related to gamma; when gamma is less than 1, such as when mutations increase the PI3K activity or reduce PTEN activity, the system is resistant to Pertuzumab, requiring much higher doses to inhibit AKT activity. Conversely, if gamma is greater than 1, the system is much more sensitive to Pertuzumab. Thus, gamma is a feedback-oriented balance metric that is indicative to drug sensitivity to Pertuzumab.

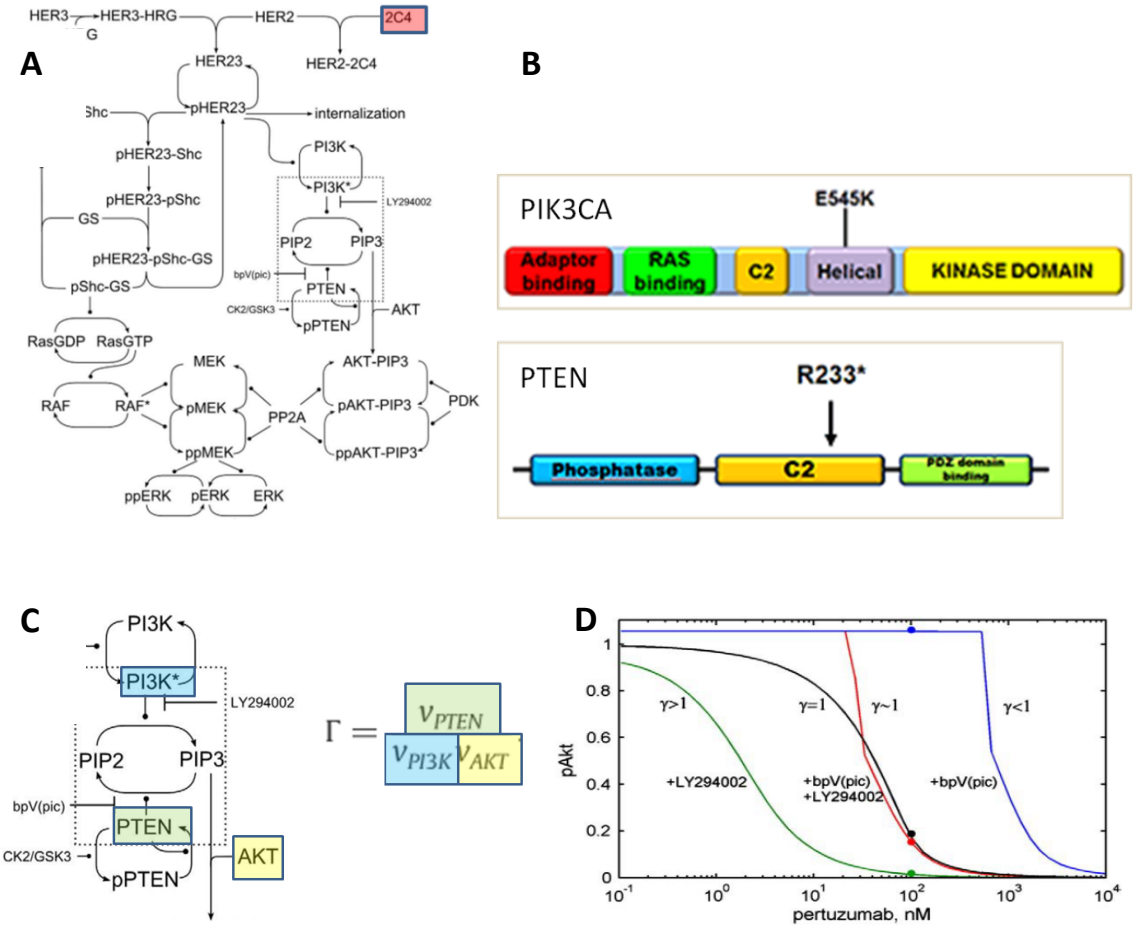


Figure 47. A balance of PTEN, PI3K and AKT activities determines pertuzumab sensitivity. A. Schematic of PI3K/MAPK pathway. Note that only two of the HER receptors, HER2 and HER3 are represented. Pertuzumab (abbreviated as 2C4) is highlighted in red, and affects the HER2 receptor. B. Two mutually exclusive mutations in PIK3CA, a component of the PI3K protein, and PTEN, and their position within the protein domains. C. A balance of PTEN, PI3K, and AKT activities known as gamma determines pertuzumab sensitivity. D. Effect of gamma on drug sensitivity. When gamma is greater than 1 (blue curve), in the case of both PTEN and PIK3CA mutations, the drug response curve is shifted to the right, meaning the signaling system is much less sensitive to Pertuzumab. Figure is adapted from Harrison[137] and mycancergenome.org.

As further proof of the effectiveness of gamma as a predictor of Pertuzumab sensitivity, Harrison measured total protein levels of AKT, PTEN, and PI3K as a proxy for the activity levels of each for 12 ovarian cancer cell lines. The protein level of PI3K was multiplied 2-fold if the most common mutations in PIK3CA (such as the E545K mutation) were present. These modified protein levels were then combined as an experimental proxy for gamma (γ_{exp}) and plotted

against the amount of growth inhibition observed when Pertuzumab was applied to the cells at a fixed concentration. As seen in the plot below, γ_{exp} and % growth inhibition, are highly correlated, indicating that γ_{exp} is a useful measure to predict drug sensitivity across the ovarian cancer cell populations.

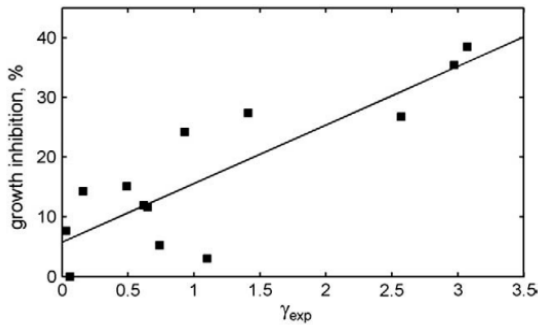


Figure 48. Correlation between growth inhibition and γ_{exp} for 12 ovarian cancer lines. Figure is reproduced from Harrison.[137]

From this example, it is clear that the balance of feedback loops is a driver of drug sensitivity. The question becomes how to 1) identify which of the many feedback loops are important to drug sensitivity, and 2) understand how genetic mutations modify these feedback loops and affect drug sensitivity.

One potential hope for the future is that patient sensitive balance metrics can be derived from these models. Such an "instrument panel" of balance metrics, derived from a combination of genomic features and ODE models may illuminate future doctors as to appropriate treatments at a patient's point in disease progression and genetic background.

4.4 Conclusions

In this dissertation, we have proposed methods for tackling two problems in drug sensitivity: 1) characterizing cross-phenotype response of targeted drugs and 2) suggested a new model for oncogenic collaboration in tumor cells.

In terms of characterizing cross-phenotype response, we have highlighted a statistical approach (ANOVA) using high sensitive cell lines as replicates, and the use of AUC as an effective method for identifying outputs of signaling that are correlated with drug sensitivity. AUC is a useful metric in characterizing cross-phenotype response in the RPPA data, especially due to its robustness in light of the lack of replication in the data.

In terms of oncogenic collaboration, we have shown that surrogate mutations are at least as predictive of drug sensitivity as the current best set of features, PAM50 expression. Surrogate mutations have the potential in integrating the larger number of unique patient specific mutations that have been previously classified as passenger mutations. We suggest that surrogate mutations may be indicative of subtler methods of regulating oncogenes.

We have accomplished these by tuning our integrative methods to the best of our ability and in light of the current limitations of the data platforms and the upstream analysis choices. Our methods are meant as a first-pass approach to the data. It is our hope that future researchers will build on our work, leading to a greater understanding of cross-phenotype response and oncogene collaboration in tumor cells, thus increasing our ability to decide on the appropriate, precise treatment of cancer.

Appendix – Additional Supplemental Figures and Tables

Session Information for Subaim 1-1 to 1-3

The following session information summarizes all packages with version numbers used in subaims 1-1 to 1-3:

```
> sessionInfo()
R version 2.15.2 (2012-10-26)

Platform: x86_64-w64-mingw32/x64 (64-bit)
locale:

[1] LC_COLLATE=English_United States.1252
[2] LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.1252

attached base packages:

[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:

[1] Bolstad_0.2-25

loaded via a namespace (and not attached):

 [1] BiocGenerics_0.4.0 colorspace_1.2-2   dichromat_2.0-0    digest_0.6.3
 [5] ggplot2_0.9.3.1    graph_1.36.2       grid_2.15.2        gtable_0.1.2
 [9] KEGGgraph_1.14.0   labeling_0.2       MASS_7.3-23        munsell_0.4
 [13] plyr_1.8           proto_0.3-10       RColorBrewer_1.0-5 reshape2_1.2.2
 [17] scales_0.2.3       stats4_2.15.2      stringr_0.6.2      tools_2.15.2
 [21] XML_3.98-1.1
```

Session Information for Subaim 1-4

The following session information summarizes the versions and all packages used in subaim 1-4:

```
> sessionInfo()
R version 2.15.3 (2013-03-01)
Platform: x86_64-apple-darwin9.8.0/x86_64 (64-bit)

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] grid      stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
```

```
 [1] plspm_0.3.7        diagram_1.6        shape_1.4.0        amap_0.8-7
BioNet_1.16.0
 [6] RBGL_1.34.0        Biobase_2.18.0     BiocGenerics_0.4.0 igraph_0.6.5-1
Rgraphviz_2.2.1
[11] graph_1.36.2

loaded via a namespace (and not attached):
 [1] AnnotationDbi_1.20.7 colorspace_1.2-1     DBI_0.2-5            dichromat_2.0-0
digest_0.6.3
 [6] ggplot2_0.9.3.1      gtable_0.1.2         igraph0_0.5.6-2      IRanges_1.16.6
labeling_0.1
[11] MASS_7.3-23          munsell_0.4          parallel_2.15.3      plyr_1.8
proto_0.3-10
[16] RColorBrewer_1.0-5   reshape2_1.2.2       RSQLite_0.11.2       scales_0.2.3
stats4_2.15.3
[21] stringr_0.6.2        tools_2.15.3
```

Session Information for Aim 2

```
> sessionInfo()
R version 2.15.2 (2012-10-26)

Platform: x86_64-w64-mingw32/x64 (64-bit)

locale:

[1] LC_COLLATE=English_United States.1252

[2] LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.1252

attached base packages:

[1] grid      stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:

[1] Rgraphviz_2.2.1    BioNet_1.16.0      RBGL_1.34.0        graph_1.36.2
[5] Biobase_2.18.0     BiocGenerics_0.4.0 foreach_1.4.1

loaded via a namespace (and not attached):

 [1] AnnotationDbi_1.20.7 codetools_0.2-8      colorspace_1.2-2
 [4] DBI_0.2-7            dichromat_2.0-0      digest_0.6.3
 [7] ggplot2_0.9.3.1      gtable_0.1.2         igraph0_0.5.7
[10] IRanges_1.16.6       iterators_1.0.6      labeling_0.2
[13] MASS_7.3-23          munsell_0.4          parallel_2.15.2
[16] plyr_1.8             proto_0.3-10         RColorBrewer_1.0-5
[19] reshape2_1.2.2       RSQLite_0.11.4       scales_0.2.3
[22] stats4_2.15.2        stringr_0.6.2        tools_2.15.2
```

Table 17. R-Square values for endogenous variables in all cell line networks. NA means that that node was not included in that cell line model. Green cells are nodes for which at least 50% variability in the training data was accounted for.

GeneID	BT20early	BT20late	BT549early	BT549late	MCF7early	MCF7late	UACC812ear	UACC812late
ACACA(31)	0.1399	0.5744	NA	NA	NA	NA	NA	NA
AKT1(207)	0.6811	0.7033	0.7526	0.5975	0.4276	0.4398	0.6343	0.6029
AKT1S1(84335)	0.4033	0.2762	0.194	0.0973	0.002	0.023	0.4013	0.3491
BAD(572)	0.2988	0.2129	0.0089	0.0467	0.0157	0.077	NA	NA
EGFR(1956)	0.0352	0.0961	0.3532	0.0349	0.0154	0.1498	0.8039	0.5427
ERBB2(2064)	0.0522	0.0345	0.1752	0.088	0.1161	0.0278	0.7534	0.6363
FRAP1(2475)	0.3965	0.0991	0.5158	0.2933	0.2297	0.1468	0.2488	0.2816
MAP2K1(5604)	0.5005	0.0971	NA	NA	0.2162	0.1683	0.1407	0.3361
MAPK1(5594)	0.8526	0.7627	0.2705	0.2858	0.7128	0.4338	0.5373	0.6266
MAPK3(5595)	0.6135	0.4453	0.0919	0.0862	0.2223	0.0527	0.0718	0.4523
MET(4233)	0.1066	0.0905	0.1483	0.1149	0.056	0.0967	0.222	0
PDK1(5163)	0.0489	0.1722	NA	NA	NA	NA	0.029	2.00E-04
PEA15(8682)	0.0162	0.0061	0.0688	0.7165	NA	NA	3.00E-04	0.0366
PIK3CA(5290)	0.4065	0.5204	0.2119	0.6647	0.6073	0.579	0.2413	0.2156
PRKAA1(5562)	0.0036	0.0858	NA	NA	0.1694	0.2167	NA	NA
PRKAA2(5563)	0.0036	0.0858	NA	NA	0.1694	0.2167	NA	NA
RB1(5925)	0.0187	0.0017	NA	NA	0.6	0.3529	0.2218	0.1136
RICTOR(253260)	0.5944	0.7699	0.8241	0.5445	NA	NA	0.609	0.8355
RPS6(6194)	0.7877	0.7889	0.8565	0.664	0.326	0.5489	0.655	0.5055
RPS6KA1(6195)	0.8501	0.8963	NA	NA	0.3363	0.1197	NA	NA
RPS6KB1(6198)	0.8671	0.8285	0.7497	0.4687	0.2012	0.1721	0.7273	0.8564
SRC(6714)	0.4011	0.2592	0.596	0.0897	0.3103	0.6678	0.2207	0.396
ERBB3(2065)	NA	NA	NA	NA	NA	NA	0.016	0.0053

Table 18. All compounds and their classification error rates by PAM50 and surrogate mutations.

COMPOUNDS	GeneTarget	PAM50exp	combinedFlat	Call
Fascaplysin	CDK1	0.235294118	0.5	PAM50
GSK1838705	IGF1R	0.228571429	0.457142857	PAM50
GSK1070916	AURKA, AURKB, AURKC	0.2	0.4	PAM50
Sigma.AKT1.2.inhibitor	AKT1, AKT2, AKT3	0.166666667	0.366666667	PAM50
Glycyl H1152	ROCK	0.259259259	0.444444444	PAM50
XRP44X	ELK3	0.222222222	0.407407407	PAM50
BIBW2992	EGFR, ERBB2	0.173913043	0.347826087	PAM50
Nutlin.3a	MDM2	0.2	0.371428571	PAM50
GSK2126458	PIK3R1, PIK3CA	0.193548387	0.35483871	PAM50
Epirubicin	TOP2A	0.3	0.433333333	PAM50
5-FdUR	DNA	0.193548387	0.322580645	PAM50
Vinorelbine	TUBB	0.303030303	0.424242424	PAM50
PD.98059	MAP2K1	0.307692308	0.423076923	PAM50
VX-680	AURKA, AURKB, AURKC	0.259259259	0.37037037	PAM50
LBH589	HDAC	0.214285714	0.321428571	PAM50

Rapamycin	MTOR	0.25	0.357142857	PAM50
5-FU	TYMS	0.178571429	0.285714286	PAM50
TGX.221	PIK3CA	0.166666667	0.266666667	PAM50
GSK2119563	PIK3CA	0.322580645	0.419354839	PAM50
Paclitaxel	BCL2, TUBB1	0.310344828	0.379310345	PAM50
Lapatinib	EGFR, ERBB2	0.066666667	0.133333333	PAM50
Geldanamycin	HSP90	0.258064516	0.322580645	PAM50
SB-3CT	MMP2, MMP9	0.387096774	0.451612903	PAM50
GSK1120212	MAP2K1	0.28125	0.34375	PAM50
Ixabepilone	TUBB3	0.294117647	0.352941176	PAM50
AZD6244	MAP2K1	0.217391304	0.260869565	PAM50
L.779450	RAF1, ARAF	0.296296296	0.333333333	EQUAL
Doxorubicin(FD)	TOP2A	0.344827586	0.379310345	EQUAL
Ibandronate sodium salt	FDPS	0.24137931	0.275862069	EQUAL
PD173074	FGFR3	0.24137931	0.275862069	EQUAL
Purvalanol A	CDK1	0.3	0.333333333	EQUAL
Triciribine	AKT1, AKT2, AKT3	0.176470588	0.205882353	EQUAL
SAHA (Vorinostat)	HDAC	0.228571429	0.257142857	EQUAL
BEZ235	FRAP1, MTOR, PIK3R1, PIK3CA	0.44	0.44	EQUAL
CPT-11(FD)	VirDNA-topo-I_N	0.413793103	0.413793103	EQUAL
Docetaxel	BCL2, TUBB1	0.310344828	0.310344828	EQUAL
Etoposide	TOP2A	0.242424242	0.242424242	EQUAL
Gemcitabine	TYMS, RRM1, CMPK	0.366666667	0.366666667	EQUAL
Pemetrexed	TYMS, GART, ATIC, DHFR	0.296296296	0.296296296	EQUAL
Sorafenib	VEGFR	0.333333333	0.333333333	EQUAL
TCS JNK 5a	JNK	0.304347826	0.304347826	EQUAL
TPCA-1	IKK2	0.285714286	0.285714286	EQUAL
CGC-11144	DNA	0.382352941	0.352941176	EQUAL
Erlotinib	EGFR	0.294117647	0.264705882	EQUAL
CGC-11047	DNA	0.424242424	0.393939394	EQUAL
Iressa	EGFR	0.21875	0.1875	EQUAL
NU6102	CCNB, CDK1	0.322580645	0.290322581	EQUAL
Oxamflatin	HDAC	0.290322581	0.258064516	EQUAL
Tamoxifen	ESR1	0.322580645	0.290322581	EQUAL
GSK1059615	PIK3R1, PIK3CA	0.266666667	0.233333333	EQUAL
GSK461364	PLK1	0.266666667	0.233333333	EQUAL
GSK923295	CENPE	0.233333333	0.2	EQUAL
Temsirolimus	FRAP1	0.333333333	0.3	EQUAL

SKI-606(Bosutinib)	SRC	0.333333333	0.296296296	EQUAL
Sunitinib Malate	VEGFR	0.411764706	0.352941176	SURROGATE
AG1024	IGF1R	0.303030303	0.242424242	SURROGATE
TCS2312 dihydrochloride	CHK1	0.419354839	0.35483871	SURROGATE
AG1478	EGFR	0.166666667	0.1	SURROGATE
Methotrexate	DHFR	0.233333333	0.166666667	SURROGATE
NSC663284	CDC25S	0.333333333	0.266666667	SURROGATE
Topotecan	TOP2A	0.448275862	0.379310345	SURROGATE
GSK1487371	PIK3CG	0.321428571	0.25	SURROGATE
Lestaurtinib(CEP-701)	TRKA, FLT3	0.321428571	0.25	SURROGATE
ICRF-193	PLK1	0.407407407	0.333333333	SURROGATE
MLN4924	NAE	0.407407407	0.333333333	SURROGATE
ZM.447439	AURKA, AURKB, AURKC	0.423076923	0.346153846	SURROGATE
Cisplatin	DNA cross-linker	0.333333333	0.242424242	SURROGATE
Oxaliplatin	DNA cross-linker	0.272727273	0.181818182	SURROGATE
17-AAG	HSP90	0.2	0.1	SURROGATE
Bortezomib	NFKB1, NFKB2	0.285714286	0.171428571	SURROGATE
Trichostatin A	HDAC	0.371428571	0.257142857	SURROGATE
Ispinesib	KSP	0.411764706	0.264705882	SURROGATE
Carboplatin	DNA cross-linker	0.363636364	0.212121212	SURROGATE
AS-252424	PIK3R3	0.423076923	0.192307692	SURROGATE

Bibiliography

1. Ross, J. S. *et al.* The HER-2 Receptor and Breast Cancer: Ten Years of Targeted Anti–HER-2 Therapy and Personalized Medicine. *The Oncologist* **14,** 320 –368 (2009).
2. Slamon, D. *et al.* Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* **235,** 177 –182 (1987).
3. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100,** 57–70 (2000).
4. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144,** 646–674 (2011).
5. Avraham, R. & Yarden, Y. Feedback regulation of EGFR signalling: decision making by early and delayed loops. *Nat. Rev. Mol. Cell Biol.* **12,** 104–117 (2011).
6. Vogel, C. L. *et al.* Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. *J. Clin. Oncol.* **20,** 719–726 (2002).
7. Weinstein, I. B. & Joe, A. K. Mechanisms of disease: Oncogene addiction--a rationale for molecular targeting in cancer therapy. *Nat Clin Pract Oncol* **3,** 448–457 (2006).
8. Holbrook, J. D. *et al.* Deep sequencing of gastric carcinoma reveals somatic mutations relevant to personalized medicine. *Journal of Translational Medicine* **9,** 119 (2011).
9. Weinberg, R. A. *Biology of cancer.* (Garland Science, 2013).
10. Wolff, A. C. *et al.* American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. *J. Clin. Oncol.* **25,** 118–145 (2007).
11. Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27,** 1160–1167 (2009).

12. Nielsen, T. O. *et al.* A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor positive breast cancer. *Clin Cancer Res* **16,** 5222–5232 (2010).

13. Kuo, W.-L. *et al.* A systems analysis of the chemosensitivity of breast cancer cells to the polyamine analogue PG-11047. *BMC Med* **7,** 77 (2009).

14. Monks, A. *et al.* Feasibility of a high-flux anticancer drug screen using a diverse panel of cultured human tumor cell lines. *J. Natl. Cancer Inst.* **83,** 757–766 (1991).

15. Fallahi-Sichani, M., Honarnejad, S., Heiser, L. M., Gray, J. W. & Sorger, P. K. Metrics other than potency reveal systematic variation in responses to cancer drugs. *Nat Chem Biol* **9,** 708–714 (2013).

16. Special Concetration Parameters GI50, TGI and LC50. at <http://dtp.nci.nih.gov/docs/compare/compare_methodology.html#specon>

17. Neve, R. M. *et al.* A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* **10,** 515–527 (2006).

18. Korkola, J. & Gray, J. W. Breast cancer genomes--form and function. *Curr. Opin. Genet. Dev.* **20,** 4–14 (2010).

19. Burdall, S. E., Hanby, A. M., Lansdown, M. R. & Speirs, V. Breast cancer cell lines: friend or foe? *Breast Cancer Research* **5,** 89 (2003).

20. Gazdar, A. F., Gao, B. & Minna, J. D. Lung cancer cell lines: Useless artifacts or invaluable tools for medical science? *Lung Cancer* **68,** 309–318 (2010).

21. Heiser, L. M. *et al.* Subtype and pathway specific responses to anticancer compounds in breast cancer. *PNAS* (2011). doi:10.1073/pnas.1018854108

22. Hao, Y. *et al.* Gain of Interaction with IRS1 by p110α-Helical Domain Mutants Is Crucial for Their Oncogenic Functions. *Cancer Cell* **23,** 583–593 (2013).

23. Benedict, W. F. *et al.* Patient with 13 chromosome deletion: evidence that the retinoblastoma gene is a recessive cancer gene. *Science* **219,** 973–975 (1983).

24. Dryja, T. P. *et al.* Homozygosity of chromosome 13 in retinoblastoma. *N. Engl. J. Med.* **310,** 550–553 (1984).

25. Boland, C. R. & Ricciardiello, L. How many mutations does it take to make a tumor? *Proc Natl Acad Sci U S A* **96,** 14675–14677 (1999).

26. Pawson, T. Specificity in signal transduction: from phosphotyrosine-SH2 domain interactions to complex cellular systems. *Cell* **116,** 191–203 (2004).

27. Liu, B. A., Engelmann, B. W. & Nash, P. D. The language of SH2 domain interactions defines phosphotyrosine-mediated signal transduction. *FEBS Lett.* **586,** 2597–2605 (2012).

28. Amin, D. N. *et al.* Resiliency and Vulnerability in the HER2-HER3 Tumorigenic Driver. *Sci Transl Med* **2,** 16ra7–16ra7 (2010).

29. Campbell, M. R., Amin, D. & Moasser, M. M. HER3 comes of age: new insights into its functions and role in signaling, tumor biology, and cancer therapy. *Clin. Cancer Res.* **16,** 1373–1383 (2010).

30. Stemke-Hale, K. *et al.* An integrative genomic and proteomic analysis of PIK3CA, PTEN, and AKT mutations in breast cancer. *Cancer Res.* **68,** 6084–6091 (2008).

31. Dalle Pezze, P. *et al.* A dynamic network model of mTOR signaling reveals TSC-independent mTORC2 regulation. *Sci Signal* **5,** ra25 (2012).

32. Vinayak, S. & Carlson, R. W. mTOR inhibitors in the treatment of breast cancer. *Oncology (Williston Park, N.Y.)* **27,** 38–44, 46, 48 passim (2013).

33. Ewen, M. E. & Lamb, J. The activities of cyclin D1 that drive tumorigenesis. *Trends Mol Med* **10,** 158–162 (2004).

34. Yamaguchi, H., Chang, S.-S., Hsu, J. L. & Hung, M.-C. Signaling cross-talk in the resistance to HER family receptor targeted therapy. *Oncogene* (2013). doi:10.1038/onc.2013.74

35. Sterman, J. *Business Dynamics*. (McGraw-Hill, Inc., 2000).

36. Burnette, W. N. 'Western blotting': electrophoretic transfer of proteins from sodium dodecyl sulfate--polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radioiodinated protein A. *Anal. Biochem.* **112,** 195–203 (1981).

37. Heritage Provider Network. HPN-DREAM breast cancer network inference challenge. at <https://www.synapse.org/#!Wiki:syn1720047/ENTITY/56061>

38. Branzei, D. & Foiani, M. Regulation of DNA repair throughout the cell cycle. *Nat. Rev. Mol. Cell Biol.* **9,** 297–308 (2008).

39. Nehrt, N., Peterson, T., Park, D. & Kann, M. Domain landscapes of somatic mutations in cancer. *BMC Genomics* **13,** S9 (2012).

40. Gatza, M. L. *et al.* A pathway-based classification of human breast cancer. *PNAS* **107,** 6994–6999 (2010).

41. Heiser, L. M. *et al.* Integrated analysis of breast cancer cell lines reveals unique signaling pathways. *Genome Biol* **10,** R31 (2009).

42. Khatri, P., Sirota, M. & Butte, A. J. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS Comput Biol* **8,** e1002375 (2012).

43. Vogelstein, B. & Kinzler, K. W. Cancer genes and the pathways they control. *Nat. Med.* **10,** 789–799 (2004).

44. Peri, S. *et al.* Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* **13,** 2363–2371 (2003).

45. Franceschini, A. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **41,** D808–815 (2013).

46. Tibes, R. *et al.* Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. *Mol Cancer Ther* **5,** 2512–2521 (2006).

47. Yalow, R. S. & Berson, S. A. IMMUNOASSAY OF ENDOGENOUS PLASMA INSULIN IN MAN. *J Clin Invest* **39,** 1157–1175 (1960).

48. Cheng, K. W., Lu, Y. & Mills, G. B. Assay of Rab25 function in ovarian and breast cancers. *Meth. Enzymol.* **403,** 202–215 (2005).

49. Akbani, Rehan. Conversation about RPPA preprocessing. (2014).

50. Lu, Y. Conversation about RPPA Process. (2014).

51. Neeley, E. Shannon. Models for the Preprocessing of Reverse Phase Protein Arrays. (2009). at <http://scholarship.rice.edu/bitstream/handle/1911/61900/3362375.PDF?sequence=1>

52. Hennessy, B. T. *et al.* A Technical Assessment of the Utility of Reverse Phase Protein Arrays for the Study of the Functional Proteome in Non-microdissected Human Breast Cancers. *Clin Proteomics* **6,** 129–151 (2010).

53. Mills, G. B. Conversation about RPPA variability. (2014).

54. Array Pro Analyzer Overview. at <http://www.mediacy.com/index.aspx?page=ArrayProOverview>

55. Ju, Z. Conversation about RPPA Quality Metrics. (2014).

56. Hu, J. *et al.* Non-parametric quantification of protein lysate arrays. *Bioinformatics* **23,** 1986–1994 (2007).

57. Neeley, E. S., Kornblau, S. M., Coombes, K. R. & Baggerly, K. A. Variable slope normalization of reverse phase protein arrays. *Bioinformatics* **25,** 1384–1389 (2009).

58. Neeley, E. S., Baggerly, K. A. & Kornblau, S. M. Surface Adjustment of Reverse Phase Protein Arrays using Positive Control Spots. *Cancer Inform* **11,** 77–86 (2012).

59. Antibody Validation. at <http://www.mdanderson.org/education-and-research/resources-for-professionals/scientific-resources/core-facilities-and-services/functional-proteomics-rppa-core/antibody-validation/index.html>

60. Troncale, S. *et al.* NormaCurve: a SuperCurve-based method that simultaneously quantifies and normalizes reverse phase protein array data. *PLoS ONE* **7,** e38686 (2012).

61. Untch, M. & Luck, H.-J. Lapatinib - Member of a New Generation of ErbB-Targeting Drugs. *Breast Care (Basel)* **5,** 8–12 (2010).

62. O'Neill, F. *et al.* Gene expression changes as markers of early lapatinib response in a panel of breast cancer cell lines. *Mol. Cancer* **11,** 41 (2012).

63. Hegde, P. S. *et al.* Delineation of molecular mechanisms of sensitivity to lapatinib in breast cancer cell lines using global gene expression profiles. *Mol Cancer Ther* **6,** 1629–1640 (2007).

64. Imami, K. *et al.* Temporal profiling of lapatinib-suppressed phosphorylation signals in EGFR/HER2 pathways. *Mol Cell Proteomics* mcp.M112.019919 (2012). doi:10.1074/mcp.M112.019919

65. Kim, H.-P. *et al.* Lapatinib, a Dual EGFR and HER2 Tyrosine Kinase Inhibitor, Downregulates Thymidylate Synthase by Inhibiting the Nuclear Translocation of EGFR and HER2. *PLoS ONE* **4,** e5933 (2009).

66. Gilmer, T. M. Lapatinib: Functional Genomics Study Leads to Insights into Mechanism of Action. *Mol Cancer Ther* **10,** 2025–2025 (2011).

67. Hamilton, J. D. *Time series analysis*. (Princeton University press, 1994).

68. Camillo, B. D., Toffolo, G., Nair, S. K., Greenlund, L. J. & Cobelli, C. Significance analysis of microarray transcript levels in time series experiments. *BMC Bioinformatics* **8,** S10 (2007).

69. Bar-Joseph, Z., Gitter, A. & Simon, I. Studying and modelling dynamic biological processes using time-series gene expression data. *Nat Rev Genet* **13,** 552–564 (2012).

70. Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3,** Article3 (2004).

71. Dabney, A., Storey, J. D. & Barnes, G. *qvalue: Q-value estimation for false discovery rate control*.

72. Curran, J. *Bolstad: Bolstad functions.* at <http://cran.r-project.org/package=Bolstad>

73. Mishra, G. R. *et al.* Human protein reference database--2006 update. *Nucleic Acids Res.* **34,** D411–414 (2006).

74. Keshava Prasad, T. S. *et al.* Human Protein Reference Database--2009 update. *Nucleic Acids Res.* **37,** D767–772 (2009).

75. Kanehisa, M. The KEGG database. *Novartis Found. Symp.* **247,** 91–101; discussion 101–103, 119–128, 244–252 (2002).

76. Tenenhaus, M., Vinzi, V. E., Chatelin, Y.-M. & Lauro, C. PLS path modeling. *Computational Statistics & Data Analysis* **48,** 159–205 (2005).

77. *Handbook of Partial Least Squares*. at <http://www.springer.com/statistics/computational+statistics/book/978-3-540-32825-4>

78. Zhang, X. *et al.* A PLSPM-Based Test Statistic for Detecting Gene-Gene Co-Association in Genome-Wide Association Study with Case-Control Design. *PLoS One* **8,** (2013).

79. Xue, F. *et al.* A latent variable partial least squares path modeling approach to regional association and polygenic effect with applications to a human obesity study. *PLoS ONE* **7,** e31927 (2012).

80. Li, F. *et al.* A powerful latent variable method for detecting and characterizing gene-based gene-gene interaction on multiple quantitative traits. *BMC Genet.* **14,** 89 (2013).

81. Janes, K. A. *et al.* A Systems Model of Signaling Identifies a Molecular Basis Set for Cytokine-Induced Apoptosis. *Science* **310,** 1646–1653 (2005).

82. Sanchez, G. *PLS Path Modeling with R*. at <http://www.gastonsanchez.com/PLS_Path_Modeling_with_R.pdf>

83. Withanage, C., Ton Hien Duc, T., Choi, H.-J. & Park, T. Dynamic Partial Least Square Path Modeling for the Front-end Product Design and Development. *J. Mech. Des.* **134,** 100907–100907 (2012).

84. Sanchez, G. *plspm: Tools for Partial Least Squares Path Modeling.* (2013). at <http://CRAN.R-project.org/package=plspm>

85. Cormen, T. H., Stein, C., Rivest, R. L. & Leiserson, C. E. *Introduction to Algorithms*. (McGraw-Hill Higher Education, 2001).

86. Barash, I. Stat5 in breast cancer: potential oncogenic activity coincides with positive prognosis for the disease. *Carcinogenesis* **33,** 2320–2325 (2012).

87. Shibata, T. *et al.* Y-box Binding Protein-1 Contributes to Both HER2/ErbB2 Expression and Lapatinib Sensitivity in Human Gastric Cancer Cells. *Mol Cancer Ther* **12,** 737–746 (2013).

88. Faber, A. C., Ebi, H., Costa, C. & Engelman, J. A. Apoptosis in targeted therapy responses: the role of BIM. *Adv. Pharmacol.* **65,** 519–542 (2012).

89. Gillings, A. S., Balmanno, K., Wiggins, C. M., Johnson, M. & Cook, S. J. Apoptosis and autophagy: BIM as a mediator of tumour cell death in response to oncogene-targeted therapeutics. *FEBS J.* **276,** 6050–6062 (2009).

90. Suzuki, T. *et al.* Nuclear cyclin B1 in human breast carcinoma as a potent prognostic factor. *Cancer Science* **98,** 644–651 (2007).

91. Heinrich, R., Neel, B. G. & Rapoport, T. A. Mathematical models of protein kinase signal transduction. *Mol. Cell* **9,** 957–970 (2002).

92. Ventura, A. C., Jackson, T. L. & Merajver, S. D. On the Role of Cell Signaling Models in. *Cancer Res* **69,** 400–402 (2009).

93. Cary, L. A., Klinghoffer, R. A., Sachsenmaier, C. & Cooper, J. A. Src Catalytic but Not Scaffolding Function Is Needed for Integrin-Regulated Tyrosine Phosphorylation, Cell Migration, and Cell Spreading. *Mol Cell Biol* **22,** 2427–2440 (2002).

94. Chen, Z. *et al.* Crucial role of p53-dependent cellular senescence in suppression of Pten-deficient tumorigenesis. *Nature* **436,** 725–730 (2005).

95. Metzker, M. L. Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11,** 31–46 (2010).

96. Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics* **11,** 685–696 (2010).

97. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* **40,** e72 (2012).

98. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).

99. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18,** 1851–1858 (2008).

100. Larson, D. E. *et al.* SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28,** 311–317 (2012).

101. Taub, M. A., Corrada Bravo, H. & Irizarry, R. A. Overcoming bias and systematic errors in next generation sequencing data. *Genome Med* **2,** 87 (2010).

102. Curtis, C. *et al.* The pitfalls of platform comparison: DNA copy number array technologies assessed. *BMC Genomics* **10,** 588 (2009).

103. Consortium, T. I. H. 3. Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (2010).

104. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).

105. Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5,** 557–572 (2004).

106. Beroukhim, R. *et al.* Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *PNAS* **104,** 20007–20012 (2007).

107. Carter, H. *et al.* Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res.* **69,** 6660–6667 (2009).

108. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucl. Acids Res.* (2011). doi:10.1093/nar/gkr407

109. Agashe, D., Martinez-Gomez, N. C., Drummond, D. A. & Marx, C. J. Good codons, bad transcript: large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. *Mol. Biol. Evol.* **30,** 549–560 (2013).

110. Lee, W., Yue, P. & Zhang, Z. Analytical methods for inferring functional effects of single base pair substitutions in human cancers. *Hum Genet* **126,** 481–498 (2009).

111. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4,** 1073–1081 (2009).

112. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31,** 3812–3814 (2003).

113. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7,** 248–249 (2010).

114. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).

115. Forbes, S. A. *et al.* The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet* **Chapter 10,** Unit 10.11 (2008).

116. Gulati, S., Cheng, T. M. K. & Bates, P. A. Cancer networks and beyond: Interpreting mutations using the human interactome and protein structure. *Semin. Cancer Biol.* (2013). doi:10.1016/j.semcancer.2013.05.002

117. Ciriello, G., Cerami, E. G., Sander, C. & Schultz, N. Mutual Exclusivity Analysis Identifies Oncogenic Network Modules. *Genome Res.* (2011). doi:10.1101/gr.125567.111

118. Vandin, F., Clay, P., Upfal, E. & Raphael, B. J. Discovery of mutated subnetworks associated with clinical data in cancer. *Pac Symp Biocomput* 55–66 (2012).

119. Bashashati, A. *et al.* DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biology* **13,** R124 (2012).

120. *Sage Bionetworks DREAM Breast Cancer Prognosis Challenge*. at <http://www.the-dream-project.org/challenges/sage-bionetworks-dream-breast-cancer-prognosis-challenge>

121. Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Research* **40,** D290–D301 (2011).

122. Beisser, D., Klau, G. W., Dandekar, T., Mueller, T. & Dittrich, M. BioNet: an R-Package for the Functional Analysis of Biological Networks. *Bioinformatics* btq089 (2010). doi:10.1093/bioinformatics/btq089

123. Csardi, G. & Nepusz, T. The igraph Software Package for Complex Network Research. *InterJournal* **Complex Systems,** 1695 (2006).

124. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29,** 1165–1188 (2001).

125. Yang, H., Cheng, C. & Zhang, W. Average Rank-Based Score to Measure Deregulation of Molecular Pathway Gene Sets. *PLoS ONE* **6,** e27579 (2011).

126. Riddick, G. *et al.* Predicting in vitro drug sensitivity using Random Forests. *Bioinformatics* **27,** 220–224 (2011).

127. *The Elements of Statistical Learning - Data Mining, Inference, and Prediction, Second Edition*. at <http://www.springer.com/statistics/statistical+theory+and+methods/book/978-0-387-84857-0>

128. Breiman, L. Random Forests. *Machine Learning* **45,** 5–32 (2001).

129. Ashby, D. Practical statistics for medical research. Douglas G. Altman, Chapman and Hall, London, 1991. No. of pages: 611. Price: £32.00. *Statistics in Medicine* **10,** 1635–1636 (1991).

130. Berger, A. H., Knudson, A. G. & Pandolfi, P. P. A continuum model for tumour suppression. *Nature* **476,** 163–169 (2011).

131. Horn, S. *et al.* TERT Promoter Mutations in Familial and Sporadic Melanoma. *Science* **339,** 959–961 (2013).

132. Huang, F. W. *et al.* Highly Recurrent TERT Promoter Mutations in Human Melanoma. *Science* **339,** 957–959 (2013).

133. Anczukow, O. *et al.* BRCA2 Deep Intronic Mutation Causing Activation of a Cryptic Exon: Opening Towards a New Preventive Therapeutic Strategy. *Clin Cancer Res* clincanres.1100.2012 (2012). doi:10.1158/1078-0432.CCR-12-1100

134. Gutenkunst, R. N. *et al.* Universally Sloppy Parameter Sensitivities in Systems Biology Models. *PLoS Comput Biol* **3,** e189 (2007).

135. Marino, S., Hogue, I. B., Ray, C. J. & Kirschner, D. E. A Methodology For Performing Global Uncertainty And Sensitivity Analysis In Systems Biology. *J Theor Biol* **254,** 178–196 (2008).

136. Cheng, T. M. K. *et al.* A Structural Systems Biology Approach for Quantifying the Systemic Consequences of Missense Mutations in Proteins. *PLoS Computational Biology* **8,** e1002738 (2012).

137. Goltsov, A. *et al.* Compensatory effects in the PI3K/PTEN/AKT signaling network following receptor tyrosine kinase inhibition. *Cellular Signalling* **23,** 407–416 (2011).