COMPUTATIONAL BIOLOGY APPROACHES IN DRUG REPURPOSING AND GENE ESSENTIALITY SCREENING

Santosh Philips

Submitted to the faculty of the University Graduate School

in partial fulfillment of the requirements

for the degree

Doctor of Philosophy

in the School of Informatics and Computing,

Indiana University

August 2016

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Lang Li, PhD, Chair

Todd C. Skaar, PhD

Doctoral Committee

Yunlong Liu, PhD

June 20, 2016

Sarath C. Janga, PhD

Xiaowen Liu, PhD

DEDICATION

To my parents for their prayers, blessings and in making me the person I am, to my wife and son for being a part of this journey and adding cheer during the challenging times and to my sister. Thank you for your unconditional love, encouragement and support, without which this would not have been possible.

ACKNOWLEDGEMENT

My journey through this PhD has been often exciting, challenging but always fruitful. There was never a point at which I did not learn something new; it could be from how not to do something or from how to improve on something. A very productive journey. That said, it was not just me but the many I met along the way who gave a word of encouragement or stood by me as I moved through this journey that helped make this a reality; I thank you all.

I would like to extend my heartfelt thank you to my mentor Dr. Lang Li, who has been a tremendous role model and a great support. Your advice and guidance has been priceless and I consider it an honor to have worked under your supervision.

I would like to thank Dr. Todd Skaar for his invaluable advice, guidance and for stimulating and feeding my interest in pharmacogenomics research.

I would like to thank Drs. Yunlong Liu, Sarath Janga and Xiaowen Liu for their time and valuable comments.

I would like to thank Heng-Yi for his help with the computational problems I encountered.

I would like to thank Elizabeth Bunge for her exceptional help with the administrative part of my PhD.

Santosh Philips

COMPUTATIONAL BIOLOGY APPROACHES IN DRUG REPURPOSING AND GENE ESSENTIALITY SCREENING

The rapid innovations in biotechnology have led to an exponential growth of data and electronically accessible scientific literature. In this enormous scientific data, knowledge can be exploited, and novel discoveries can be made. In my dissertation, I have focused on the novel molecular mechanism and therapeutic discoveries from big data for complex diseases. It is very evident today that complex diseases have many factors including genetics and environmental effects. The discovery of these factors is challenging and critical in personalized medicine. The increasing cost and time to develop new drugs poses a new challenge in effectively treating complex diseases. In this dissertation, we want to demonstrate that the use of existing data and literature as a potential resource for discovering novel therapies and in repositioning existing drugs. The key to identifying novel knowledge is in integrating information from decades of research across the different scientific disciplines to uncover interactions that are not explicitly stated. This puts critical information at the fingertips of researchers and clinicians who can take advantage of this newly acquired knowledge to make informed decisions.

This dissertation utilizes computational biology methods to identify and integrate existing scientific data and literature resources in the discovery of novel molecular targets and drugs that can be repurposed. In chapters 1 of my dissertation, I extensively sifted through scientific literature and identified a novel interaction between Vitamin A and

CYP19A1 that could lead to a potential increase in the production of estrogens. Further in chapter 2 by exploring a microarray dataset from an estradiol gene sensitivity study I was able to identify a potential novel anti-estrogenic indication for the commonly used urinary analgesic, phenazopyridine. Both discoveries were experimentally validated in the laboratory. In chapter 3 of my dissertation, through the use of a manually curated corpus and machine learning algorithms, I identified and extracted genes that are essential for cell survival. These results brighten the reality that novel knowledge with potential clinical applications can be discovered from existing data and literature by integrating information across various scientific disciplines.

Lang Li, PhD, Chair

TABLE OF CONTENTS

List of Tables xi

List of Figures xiii

List of Abbreviations xv

Chapter 1: A translational bioinformatic approach in identifying and validating an interaction between Vitamin A and CYP19A1 1

1.1 Introduction 1

1.2 Materials and methods 2

1.2.1 CYPs and their super-regulators 2

1.2.2 Identification of compounds that influence CYP regulators 3

1.2.3 Influence of retinoic acid on DAX1/CYP19A1 gene expression 4

1.3 Results 5

1.3.1 CYPs and their regulatory network 5

1.3.2 Influence of ATRA on CYP19A1 (Aromatase) and DAX1 genes in the various cell lines 7

1.4 Discussion 10

Chapter 2: Discovery of the antiestrogenic properties of phenazopyridine 13

2.1 Introduction 13

2.1.1 Drug repurposing: new uses for old drugs. 13

2.1.2 Gene Expression Profiling 15

2.1.3 Estrogens and antiestrogens 17

2.2 Methods 18

2.2.1 Identification of drugs having a similar gene expression profile as that of estradiol 18
2.2.2 Effect of phenazopyridine on MCF-7 cell proliferation 19
2.2.3 Effect of estradiol on phenazopyridine induced inhibition of cell proliferation 20
2.2.4 Influence of phenazopyridine and estradiol on progesterone receptor expression 21
2.2.5 Effect of phenazopyridine on tumor growth in athymic nude mice 22
2.3 Results 23
2.3.1 Drugs with similar gene expression signature as estradiol 23
2.3.2 Phenazopyridine inhibits the proliferation of MCF-7 cells 24
2.3.3 Effect of estradiol on phenazopyridine induced cell proliferation 25
2.3.4 Phenazopyridine effect on PGR expression 26
2.3.5 Phenazopyridine effect of tumor growth in vivo 27
2.4 Discussion 28
Chapter 3: Genes essential for cell survival 32
3.1 Introduction 32
3.1.1 Knowledge in literature 32
3.1.2 Text mining for knowledge discovery 34
3.1.3 Model evaluation 39
3.1.4 RNA interference 42
3.2 Methods 44
3.2.1 Abstract selection and corpus construction 44

3.2.2 Training and testing datasets 45
3.2.3 Selection of algorithm 47
3.2.3 Training and testing the model 52
3.2.4 Generation of the screening dataset 52
3.2.5 Extraction of RNAi relevant abstracts 53
3.2.6 Creation of dictionary for entity recognition 53
3.2.7 Entity tagging and cell-gene information extraction 53
3.2.8 Validation of the essential genes 54
3.3 Results 54
3.3.1 Identification of siRNA relevant abstracts and corpus creation 54
3.3.2 Evaluation of the classifiers 55
3.3.3 Evaluating the performance of the top 3 models 56
3.3.4 Genes essential for cancer cell survival 59
3.3.5 Validation of Genes predicted to be essential 72
3.4 Discussion 72
References 78
Curriculum Vitae

LIST OF TABLES

Table 1: F-statistic along with the p-value for the effect of ATRA on the expression of CYP19A1 (aromatase) and DAX1 across the each cell lines 8

Table 2 : Examples of drugs that have been successfully repurposed 15

Table 3: Top 20 up and 20 down regulated estradiol sensitive genes 19

Table 4: The various combinations of treatments that were used to study the effect of estradiol and phenazopyridine on PGR expression 22

Table 5: Drugs with similar (+ mean) or opposite (- mean) gene expression signature to that of estradiol 24

Table 6: Confusion matrix for binary text classification 40

Table 7: Composition of the training and testing sets used to test the various weka classifiers 46

Table 8: The % accuracy of classification after evaluating each classifier on a given dataset using 10 fold stratified cross validation 56

Table 9: The % accurately classified by the top three models after training and testing.. 57

Table 10: Classifier errors for the classifier's tested on dataset 5 58

Table 11: Performance metrics across the various classifiers tested on dataset 5 for abstracts classified as RNAi 58

Table 12: The number of abstracts that were processed per year and the number of abstracts that were identified as relevant to RNA interference studies 61

Table 13: The number of times a given gene and cell line were studied together 62

Table 14: Frequency of the number of genes being studied in a given cell line 63

Table 15: Frequency of the number of cell lines used to study a given gene 63

Table 16: The genes amongst the top 20 that are known to be cancer genes and their roles in the various processes required for cellular function 65

Table 17: Genes targeted for treating various cancers along with the respective drugs used 66

Table 18: Genes targeted for treating various cancers along with the respective drugs used 67

Table 19: The table shows the number of genes among the top 20 genes that were studied in a given cancer type 69

LIST OF FIGURES

Figure 1: Compound - CYP Regulator Network 6

Figure 2: Compound - CYP Regulator - CYP Network 7

Figure 3: Relative fold change of aromatase gene in response to various concentrations of ATRA 9

Figure 4: Relative fold change of DAX1 gene in response to various concentrations of ATRA 9

Figure 5: A comparison of traditional de novo drug discovery and development versus drug repositioning 14

Figure 6: Success rates from first-in-man to registration..... 14

Figure 7: Concept behind the Connectivity Map 16

Figure 8: Plate layout for the treatment of MCF-7 cells with phenazopyridine. 20

Figure 9: Phenazopyridine inhibits the proliferation of MCF-7 cells 25

Figure 10: Estradiol does not inhibit the antiproliferative activity of phenazopyridine ... 26

Figure 11: Phenazopyridine partially inhibits the estradiol induced PGR expression..... 27

Figure 12: Effect of Phenazopyridine on the implanted MCF-7 tumor growth in athymic mice 28

Figure 13: Growth of MEDLINE abstracts between 2001 and 2014 33

Figure 14: The continuum of understanding..... 33

Figure 15: The basic steps involved in text mining 36

Figure 16: The WEKA interface and available tools 38

Figure 17: The WEKA explorer 39

Figure 18: Mechanism of RNA interference 43

Figure 19: Structure of the WEKA ARFF file 47

Figure 20: The filtered classifier option allowing for the simultaneous selection of a classifier and filter 50

Figure 21: The stringToWord filter and the various options available to process the text 51

Figure 22: AUC Receiver Operator Characteristics for the SMO-5 model 59

Figure 23: Top 10 most studied cell lines 64

Figure 24 : The top 20 genes predicted to be essential for cell survival 64

Figure 25: Genes that are co-expressed and interact with the top 20 genes that were identified in different cancer types 70

Figure 26: The gene network of the top hit gene AKT1 from our study and the other genes that are co-expressed or interact with it 71

Figure 27: p53 response to cellular stress can lead to cell survival or cell death 74

LIST OF ABBREVIATIONS

CYP	Cytochrome P450
ATRA	All Trans Retinoic Acid
ANOVA	Analysis of Variance
CMap	Connectivity Map
FDA	Food and Drug Administration
AI	Aromatase Inhibitor
SERM	Selective Estrogen Receptor Modulators
ERD	Estrogen Receptor Downregulators
FSH	Follicle Stimulating Hormone
DMEM	Dulbecco's Modified Eagle Medium
FBS	Fetal Bovine Serum
IR	Information Retrieval
IE	Information Extraction
NER	Named Entity Recognition
WEKA	Waikato Environment for Knowledge Analysis
ARFF	Attribute Relation File Format
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
TPR	True Positive Rate
FPR	False Positive Rate

RNAi	Ribonucleic Acid Interference
siRNA	Small Interfering Ribonucleic Acid
shRNA	Short Hairpin Ribonucleic Acid
GO	Gene Ontology
NCBI	National Center for Biotechnology Information
PMID	PubMed Identifier
XML	Extensible Markup Language
NA	Not Applicable
ASCII	American Standard Code for Information Interchange
RNS	Random Negative Set
SMO	Sequential Minimal Optimization
SVM	Support Vector Machine
IDF	Inverse Document Frequency
TF	Term Frequency
HGNC	HUGO Gene Nomenclature Committee
DPSC	Donnelly - Princess Margaret Screening Centre
ROC	Receiver Operator Characteristic

Chapter 1: A translational bioinformatic approach in identifying and validating an interaction between Vitamin A and CYP19A1

1.1 Introduction:

The Cytochrome P450 (CYP) system consists of 57 enzymes, which are further classified into 18 families and 43 subfamilies based on sequence similarity[1]. They play a crucial role in the metabolism of various chemicals both endogenous and exogenous[2]. The members of the first three CYP families 1-3 are mainly involved in the metabolism of exogenous compounds such as medications, whereas the members of the other families are involved largely in the metabolism of endogenous compounds such as cholesterol, bile acids, steroid hormones and fatty acids. A given CYP enzyme can metabolize multiple substrates and a given substrate can be metabolized by multiple CYPs. Mutations or the absence of genes encoding the CYP enzymes can not only result in altered drug response but can also make an individual more susceptible to human disease such as glaucoma[3-7] and elevated cholesterol[8].Even a single mutation has the potential to alter the structure of these enzymes, resulting in altered activity or substrate specificity[9]. Furthermore, the co-administration of multiple drugs can influence the enzymes involved in their metabolism either through induction or inhibition [10]. Age and sex as well can influence CYP activity, studies have shown that CYP3A4 activity is higher in adults compared to fetus [11] and that women metabolize CYP3A4 faster than men [12, 13]. A wealth of information on the CYP variants is available at the Human Cytochrome P450 Allele Nomenclature website [14]. Despite the presence of this large amount of information it is still challenging to optimize therapy to meet an individual's needs, especially with the increasing usage of supplements and herbal medications.

One of the challenges of personalized medicine is to identify or fine tune drug combinations without drastically affecting the metabolic pathway of either. There are numerous studies that have shown that the activity of these enzymes are influenced by various upstream regulatory mechanisms [13, 15-24], which in turn can potentially influence drug response. Despite the use of various patient characteristics there still exist a substantial amount of variations in drug response and mainly due to the nature and combinations of sources of variation. One such factor is the increasing use of dietary supplements that are not always taken into account while drugs are prescribed, which can potentially alter Cytochrome P450 activity [25, 26]. The mining of previously published literature across various disciplines has been very useful and effective in identifying potential drug interaction and rofecoxib is an excellent example, where the drugs toxic effect was present in literature before the drug was recalled [27]. Thus translational bioinformatics methods summarizing the literature data have proven to be an effective way of uncovering interactions that could be beneficial or harmful. In our study we were able to identify a correlation between retinoic acid and aromatase gene expression through bioinformatics analyses of existing databases. We were able to functionally validate this bioinformatic prediction in three different cell lines using physiological concentrations of retinoic acid. Our studies show that retinoic acid substantially alters the expression of the aromatase gene.

1.2 Materials and methods:

1.2.1 CYPs and their super-regulators

For this study we choose seven CYP subfamilies (CYP1A, CYP2A, CYP2B, CYP2C, CYP2D, CYP2E, and CYP3A) that are mainly responsible for metabolizing

more than 90% of drugs as well as CYP19A1 that is largely involved in the biosynthesis of estrogens. In order to identify compounds that indirectly affected the CYP activity/expression we used previously published endogenous CYP regulators[13, 15-24] from the literature and those that had a significant influence over the expression as well as activity of these CYP enzymes from the human liver bank gene expression quantitative loci data set[28]. These endogenous CYP regulators that had a direct effect on CYP enzymes were used as seed to identify compounds (Super-Regulators) that in turn influenced their regulation.

1.2.2 Identification of compounds that influence CYP regulators.

The endogenous CYP regulators were uploaded into Metacore (Thomson Reuters, NY, USA), a web based computational tool backed by text mining capabilities to build a highly interconnected network of CYP regulators and compounds that influenced their expression. Each node in the network represented a CYP regulator or a compound and the edges represented the interaction between the two denoting either an activation or inhibition. Not all CYP regulators were associated with upstream compounds. The CYP regulators that were not associated with any compounds and compounds that had fewer than 3 edges were eliminated from further analysis. The CYP regulators from the above list (compounds (edges =>3) – CYP regulator) were used to build a second network along with the 10 CYP enzymes, namely CYP1A2, CYP2A6, CYP2B6, CYP2C19, CYP2C8, CYP2C9, CYP2D6, CYP2E1,CYP3A4, AND CYP19A1 to further confirm the interactions between them. The results from the above two networks namely the [Compounds (edges =>3) - CYP regulator] and [CYP regulator- CYP enzyme] was merged to form the final network using Cytoscape[29] to represent the overall interaction

between the compounds, CYP regulators and CYP enzymes. Using this network the path from a given compound to its terminal leaf, which was either a CYP regulator or CYP enzyme was traced, thus predicting the interaction between the compound and CYPs.

1.2.3 Influence of retinoic acid on DAX1/CYP19A1 gene expression:

1.2.3.1 Cell culture and treatment:

Three cell lines namely, JEG3 (Placental Cancer), HeLa (Cervical Cancer), and LNCAP (Prostate Cancer) were chosen to study the expression of the CYP19A1 (Aromatase) and DAX1 genes. The cells were plated in six T25 flasks at a density of 0.25 X 10^6 cells/ flask and grown in DMEM with 10% FBS. After 24 hours the media was removed and the cells were washed 3 times with DMEM containing 10% charcoal stripped FBS and cells were then allowed to grow in the new media. The media was replaced with fresh media every 24 hours for two more days. After 72 hours of initial media change the cells in each of the six flasks were treated with either vehicle (0.01% Ethanol) or All Trans Retinoic Acid (ATRA) (Sigma-Aldrich, USA) at 0.1nM. 1nM, 10nM, 100nM and 1000nM respectively.

1.2.3.2 RNA extraction, cDNA synthesis and Gene Expression:

After 24 hours of treatment the cells were harvested and RNA extracted using miRNeasy Kit (Qiagen Inc., USA) according to the manufactures protocol. The RNA was then quantified using Quant-IT Kit (Life Technologies, USA) on the Qubit Fluorometer (Life Technologies, USA) according to the manufactures protocol. The cDNA was synthesized using the QuantiTect Reverse Transcription Kit (Qiagen Inc. USA) according to the manufactures protocol from 1ug of RNA. The gene expression for CYP19A1 and DAX1 was measured with the respective Taqman Gene Expression Assays (Life

Technologies, USA) on the iCycler instrument (Bio-Rad Inc., USA) in accordance with the manufactures protocol.

1.3 Results:

1.3.1 CYPs and their regulatory network

The initial network between the compounds and CYP regulators consisted of 868 edges between the compounds and CYP regulators, with 15 large clusters (Figure 1). The top 15 clusters were around the following CYP regulators namely, ESR1, PXR, PPARalpha, LXRalpha, GCRalpha, LXRbeta, AHR, PPARgamma, PPARbeta, VDR, FXR, RXRbeta, LHR, CAR and TRbeta. The number of compounds that formed these cluster ranged from 187 to 7, thus representing the extent to which a single CYP regulator can be influenced by multiple compounds. This network was further reduced by eliminating compounds that had less than 3 edges with other CYP regulators. The final regulatory network mainly consisted of the 134 edges between 42 nodes which included 9 CYPs, 16 CYP regulators and 17 compounds (Figure 2). All of the drug metabolizing CYP enzymes including the CYP19A1 had either a compound or CYP regulator upstream. Only CYP2E1 did not have any compound or CYP regulators associated with it. The resultant network clearly indicated the potential influence of retinoic acid (Vitamin A) on the expression of CYP19A1 (Aromatase) through DAX1.

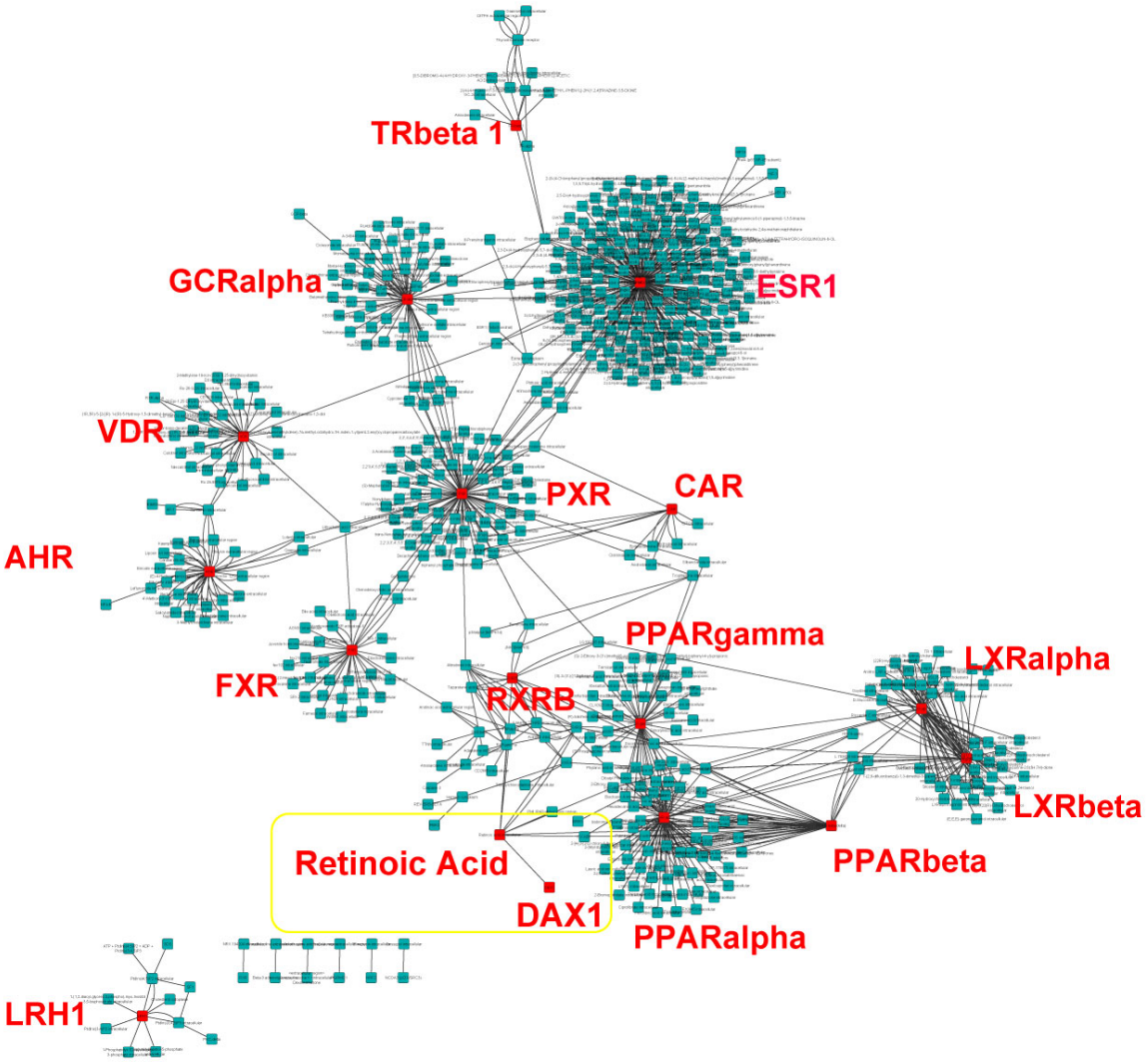

Figure 1: Compound - CYP Regulator Network. The overall network consisting of 868 interaction between the various compounds and CYP regulators that they affect. The density of the cluster is proportional to the number of compounds that influence its activity, and the nodes between clusters represents the compounds that influence more than one CYP regulator. The top 15 clusters were formed around ESR1, PXR, PPARalpha, LXRalpha, GCRalpha, LXRbeta, AHR, PPARgamma, PPARbeta, VDR, FXR, RXRbeta, LHR, CAR and TRbeta (the respective nodes are highlighted in red and interaction between retinoic acid and DAX1 is highlighted by the yellow rectangle).

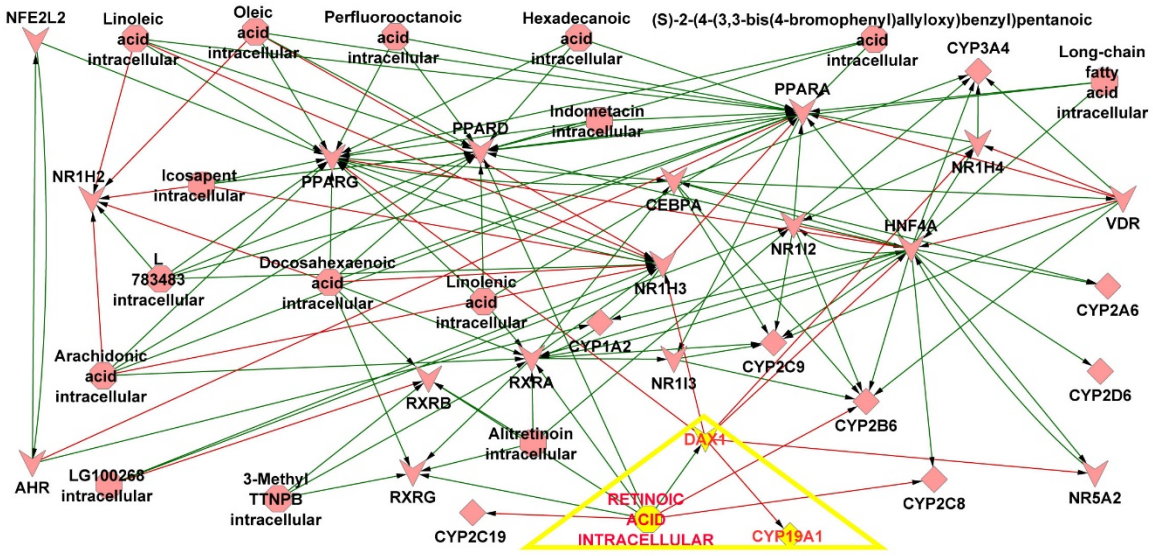

Figure 2: Compound - CYP Regulator - CYP Network. The above network shows the interaction between the Compounds, CYP regulators and CYPs. Green arrows represent up regulation and red arrows represent down regulation of the respective gene. The interaction between retinoic acid, DAX1 and CYP19A1 is highlighted by the yellow triangle.

1.3.2 Influence of ATRA on CYP19A1 (Aromatase) and DAX1 genes in the various cell lines:

The above hypothesis that retinoic acid alters the expression of aromatase gene was experimentally verified using three different cells lines, namely JEG3, HeLa and LNCaP. Each cell line was treated with various concentrations of ATRA ranging from 0.1nM to 1uM, which included the physiological concentration at which retinoic acid is found in humans. The physiological concentration of retinoic acid in human plasma is around 4.9 ng/ml and all-trans retinoic acid (ATRA) accounts for ~75% of the total[30].

After a 24 hour treatment period the cells were harvested, RNA extracted and the expression was measured for CYP19A1 and DAX1 using the respective Taqman gene expression assays. The expression of aromatase gene increased proportionally with increasing concentrations of ATRA and tapered off at 10nM ATRA (Figure 3). DAX1 expression was observed only in the HeLa cell line showing a decrease in activity with increasing concentration of ATRA (Figure 4). The above experiments were performed in triplicates on different days for each cell line.

Cell Line	Gene	F(5/12)	p-value
HeLa	CYP19A1	9.775	0.0007
	DAX1	8.1982	0.0014
JEG3	CYP19A1	2.8328	0.0647
LNCaP	CYP19A1	1.4003	0.292

Table 1: F-statistic along with the p-value for the effect of ATRA on the expression of CYP19A1 (aromatase) and DAX1 across the each cell lines

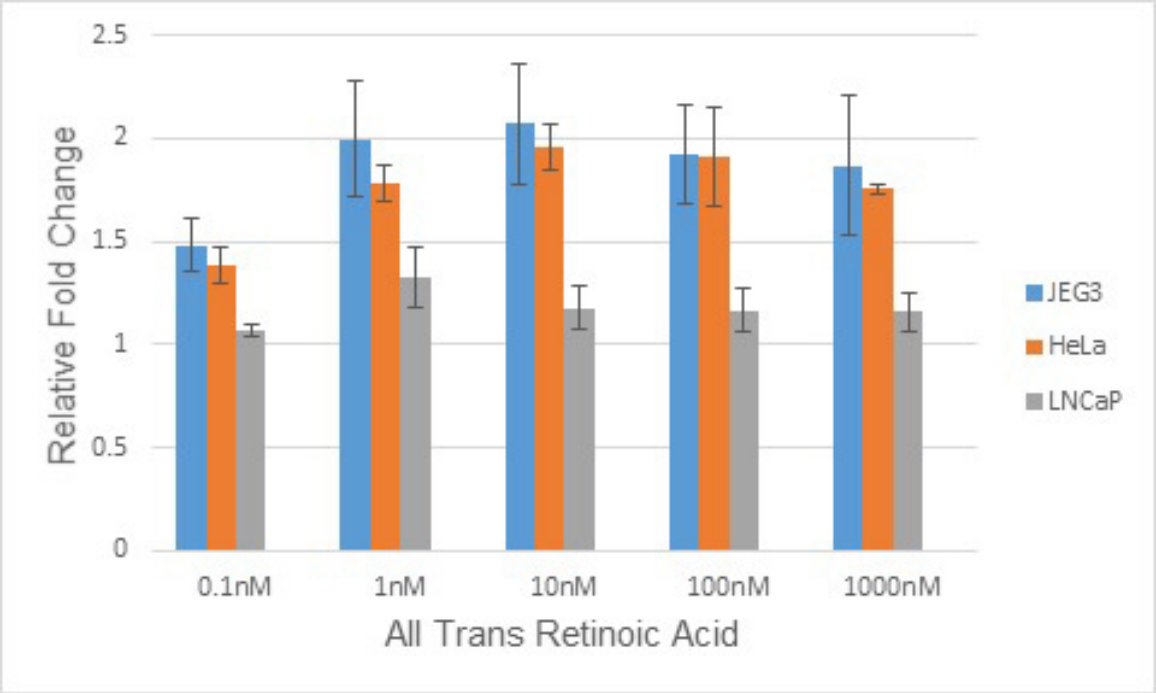

Figure 3: Relative fold change of aromatase gene in response to various concentrations of ATRA.

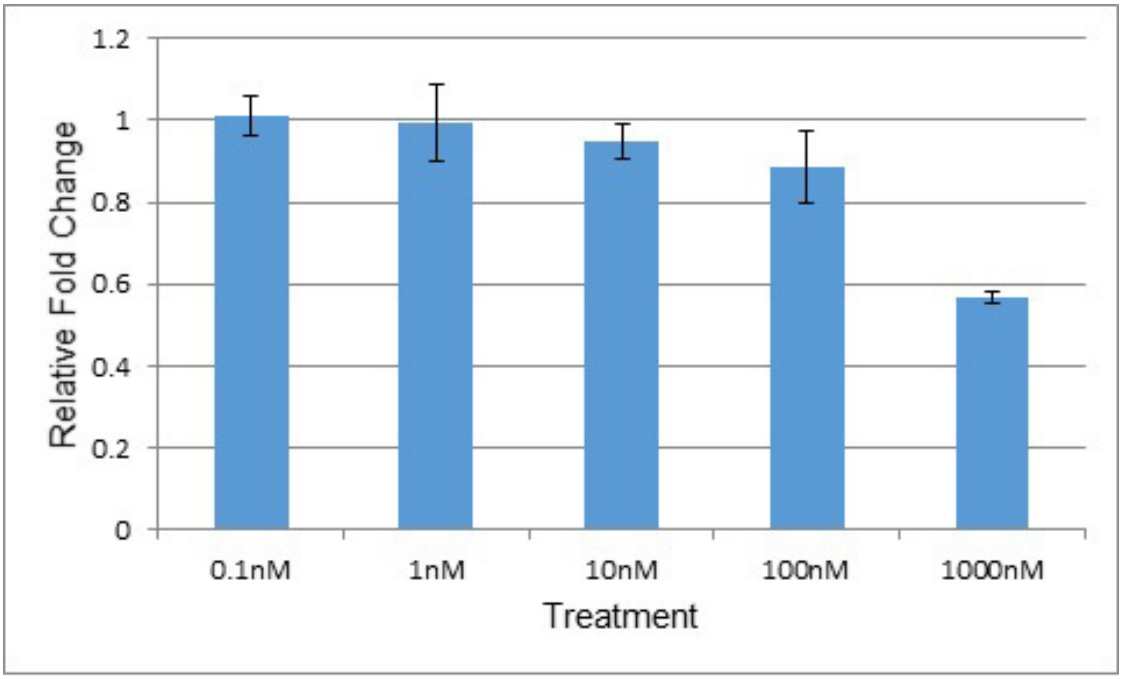

Figure 4: Relative fold change of DAX1 gene in response to various concentrations of ATRA.

A one-way between treatments ANOVA was conducted to compare the effect of retinoic acid on the expression of CYP19A1 and DAX1 in the three different cell lines. There was a significant effect of retinoic acid on the expression of CYP19A1 and DAX1 at the $p<0.05$ level in the HeLa cell line (Table 1). Further, post hoc comparison using the Tukey test showed that the fold change for treatments (0.1nM, 1nM, 10nM and 100nM) were significantly different from treatment at 1000nM. Thus indicating that the expression of aromatase gene proportionally increased with an increasing concentration of retinoic acid reaching 100nM, which included the physiological concentration at which retinoic acid is present in the human body.

1.4 Discussion:

In the current study we were able to identify 868 interactions between various chemical compounds and cytochrome P450 regulators. We choose to follow the interaction of retinoic acid on aromatase enzyme because of its possible significant application towards personalized medicine in endocrine therapy. The cell lines chosen for this study are known to express CYP19A1 and DAX1 (HeLa). In the cell experiments, we found that retinoic acid up-regulates the aromatase enzyme. Retinoic acid a metabolite of Vitamin A is very commonly found in various foods and dietary supplements. Aromatase is a key enzyme involved in the biosynthesis of estrogens [31, 32], which can catalyze the progression of estrogen-dependent breast cancers. The levels of aromatase activity and mRNA expression are higher in the breast cancer tissue than in normal tissue [33-35]. In addition to the ovarian supply of estrogens, aromatase enzyme is also involved in the local production of estrogens through the conversion of circulating adrenal androgens [36], thus having an immense potential to fuel estrogen receptor

positive breast cancer. DAX 1(dosage-sensitive sex reversal adrenal hypoplasia congenital critical region on the X-chromosome gene 1) is an orphan member of the nuclear receptor family [37, 38], and functions as global anti-steroid factor and represses the expression of many enzymes involved in the steroidogenic pathway, including aromatase [39, 40]. The expression of DAX 1 has been reported in breast cancers [41, 42], although it's exact mechanism is not fully understood. Aromatase inhibitors were developed and widely utilized to treat endocrine tumors[43], especially breast cancer with estrogen receptor positive patients. Therefore, the up regulation of aromatase would in turn result in higher levels of estrogens, and could possibly stimulate the endocrine tumor growth. Most importantly, the usage of Vitamin A could reduce the effectiveness of aromatase inhibitor treatment for cancer. Given the fact that Vitamin A is so commonly found in various food/supplement source, the chance for its potential influence is higher, thus we chose to validate its influence on aromatase expression.

Our initial bioinformatics finding using the MetaCore database revealed that the presence of retinoic acid caused an up regulation of DAX1 and which in turn caused the down regulation of aromatase. In our functional study, only one cell line, namely the HeLa cells showed expression for DAX1. HeLa cells treated with retinoic acid showed a down regulation of DAX1 and an up regulation of aromatase expression. Though not in the exact same direction as the bioinformatics prediction, the overall experimental outcome in HeLa cells does bring out the possibility that when present, DAX1 expression is inversely related to CYP19A1 expression in the presence of retinoic acid. The other two cell lines did not show any expression for DAX1 but did show that retinoic acid up regulated CYP19A1 expression. DAX1 is an orphan member of the nuclear receptor

superfamily of transcription factors, whose disruption has been linked with increase expression of aromatase enzyme [39, 44]. Even though the cell lines chosen for this study are from extra gonadal sites it clearly shows that retinoic acid has a significant influence on aromatase expression in the presence or absence of DAX1. This finding is of importance as the use of aromatase inhibitors in treating breast cancer is widespread [45, 46]. Though aromatase inhibitors are a gold standard in treating ER-positive breast cancer, resistance to this therapy still requires the use of other modes of suppression of intra-tumoral estrogen production[36], thus calling for further investigation into the underlying cause of resistance.

In this part of the dissertation, we have shown that the use of curated literature data is valuable in discovering novel drug enzyme interactions, and potential clinically significant drug interactions. Our primary contribution is the established feasibility of this translational bioinformatics approach in detecting novel drug interaction signals.

Chapter 2: Discovery of the antiestrogenic properties of phenazopyridine

2.1 Introduction:

2.1.1 Drug repurposing: new uses for old drugs.

Despite the increased resources and research put in by pharmaceutical companies, there is a huge concern about the number of new chemical entities that finally make it to the market as approved drugs [47-50]. This has led to an increasing interest in repurposing or repositioning existing drugs. Drug repurposing is the process of finding new indications for existing drugs and comes with the advantage of having established clinical and pharmacokinetic data associated with the old drug. In addition, it has reduced costs and a shorter time to market of 3 -12 years than the traditional de novo drug discovery and development process that can take 10 -17 years and incur huge costs (Figure 5) [51]. The average success rate across different therapeutic areas for a new chemical entity to emerge as an approved drug is approximately 11% and varies within each therapeutic area as shown in Figure 6. Approximately 90% of blockbuster drugs have additional indications were they can be potentially used for an indication other than the one for which they were originally approved[52]. There are many examples of drugs that are currently being used for an indication other than the one they were initially tested for, which include the classical sildenafil (Viagra) a PDE5 inhibitor which was initially tested for treating angina but repurposed to treat male erectile dysfunction[53] and thalidomide a TNF α inhibitor which was initially used as a sedative and in treating nausea and insomnia, today is used mainly in the treatment of multiple myeloma and complications associated with leprosy[54]. Table 2 shows some of the drugs that have been effectively repurposed along with their targets and new indications[55].

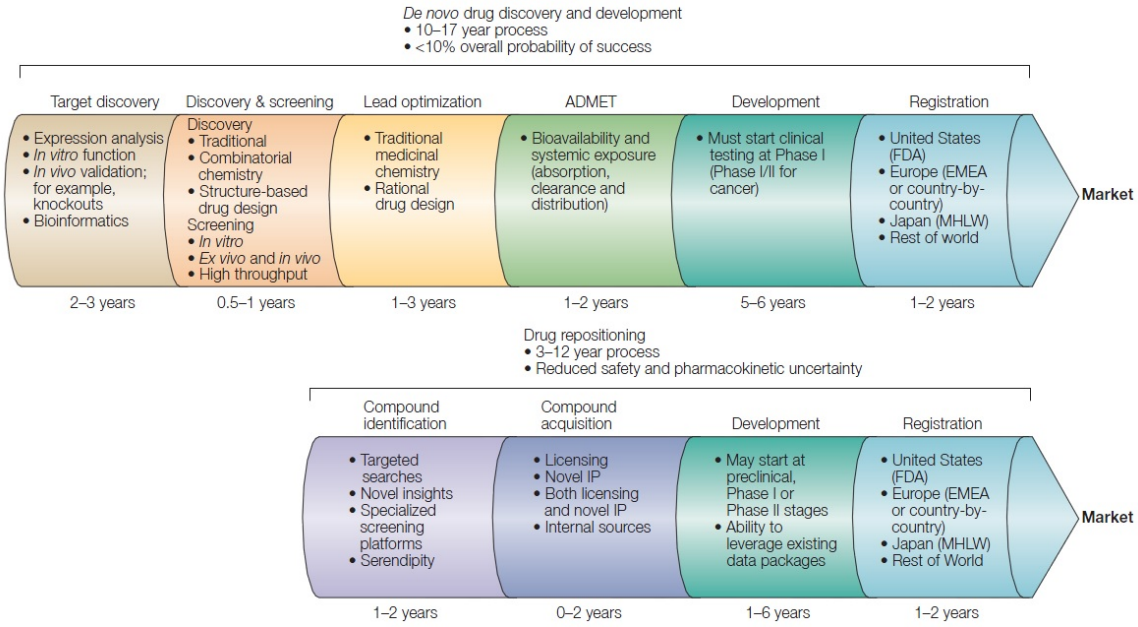

Figure 5: A comparison of traditional de novo drug discovery and development versus drug repositioning.

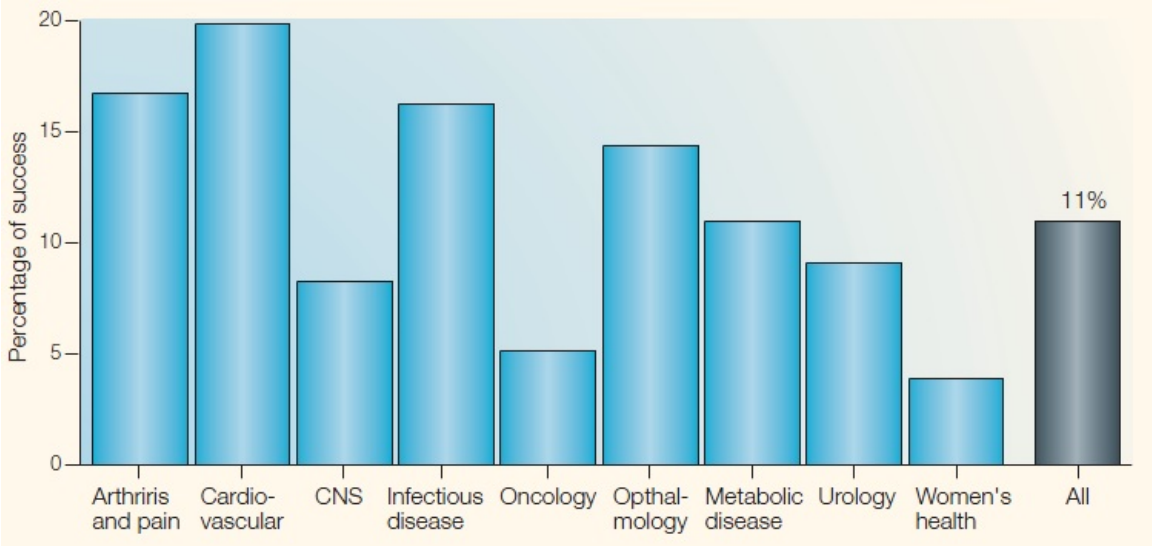

Figure 6: Success rates from first-in-man to registration. The overall success rate is 11%.

Drug name	Original target	Original indication	New target	New indication
Successful repositionings from approved drugs				
Duloxetine	Serotonin and norepinephrine reuptake	Depression	Serotonin and norepinephrine reuptake	Stress urinary incontinence, fibromyalgia, chronic musculoskeletal pain
Everolimus	mTOR	Immunosuppressant	Unchanged	Pancreatic neuroendocrine tumors
Imatinib	BCR-ABL	CML	KIT, PDGFRA	GIST
Minoxidil	Unknown	Hypertension	Unknown	Hair loss
Nelfinavir	HIV-1 protease	AIDS	Inhibits AKT pathway	In clinical trials for multiple cancers
Sildenafil	PDE5	Angina	Unchanged	Erectile dysfunction, pulmonary arterial hypertension
Sunitinib	Multiple kinases	GIST, renal cell carcinoma	Unchanged	Pancreatic neuroendocrine tumors
Trastuzumab	HER2	HER2-positive breast cancer	Unchanged	HER2-positive metastatic gastric cancer
Successful repositionings from investigational drugs				
Crizotinib	MET kinase	Clinical trials for anaplastic large-cell lymphoma	*EML4-ALK* oncogene	NSCLC
Thalidomide	Unknown	Morning sickness (withdrawn)	Inhibits tumor necrosis factor α production	Leprosy
Thalidomide	Unknown	Morning sickness (withdrawn)	Inhibits angiogenesis	Multiple myeloma
Zidovudine	Reverse transcriptase	Failed clinical trials for cancer	Reverse transcriptase	AIDS

Table 2 : Examples of drugs that have been successfully repurposed.

2.1.2 Gene Expression Profiling:

The completion of the Human Genome Project and the advances in information and sequencing technology has led to the generation of immense amount of genetic data and related knowledge on how genes are involved in the various physiological processes as well as their influence on disease pathogenesis and drug response. This massive amounts of data can be exploited to identify pathways that are shared by diseases and influenced by drug action. Thus an understanding of the extent to which a gene is expressed and the conditions under which they are expressed can be used to understand the biological roles of the proteins and enzymes they encode as well as the manner in which they interact to maintain or disrupt homeostasis. Gene expression profiling has

been used for decades as a way to identify genes that are involved in various biological activities, susceptibility to disease, in identifying novel targets, in predicting toxicity of novel compounds and in patient drug response. A novel approach to exploit the gene expression data was developed at the broad institute, namely the connectivity map [56, 57]. The connectivity map is based on the assumption that a given drug can be potentially used to treat a disease, if the disease gene expression signature is negatively correlated with the drugs gene expression profile. This database has the gene expression profiles from cultured human cells treated with varying concentrations of 1309 FDA approved bioactive small molecules. The main goal was to describe all biological states – physiological, disease or induced with a chemical, thus providing a tool for researchers studying a drug candidate, gene or disease condition to compare their signature with the ones present in the database to discover unexpected novel connections as shown in Figure 7.

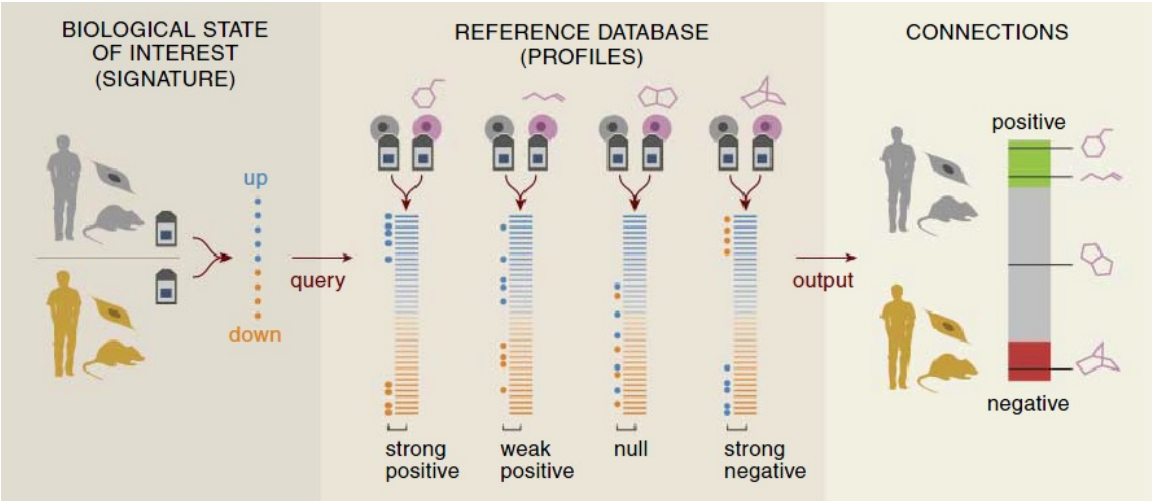

Figure 7 : Concept behind the Connectivity Map.

2.1.3 Estrogens and antiestrogens:

Estrogens are predominantly present in females and are crucial for the sexual and reproductive development and are mainly produced by the ovaries and to some extent in the fat cells and adrenal glands. They are also present in males but at lower levels than in females and are produced by the adrenal glands and testes. There are three hormones which are representative of estrogens, namely estrone, estradiol and estriol. Despite their predominant role in reproduction they also have many beneficial roles such as in neuroprotection [58-60], cardiovascular health [61-63], and bone health [64-67]. Apart from their many roles and benefits they bring to the human body, estrogens are also known to fuel the growth of breast, ovarian and uterine cancers [68, 69]. For decades antiestrogens have been used to either block the action of estrogens or lower their levels in the body through the use of aromatase inhibitors (AI), selective estrogen receptor modulators (SERM) or estrogen receptor downregulators (ERD). AI's such as anastrazole, exemestane and letrozole inhibit the production of estrogens by blocking the aromatase enzyme which is involved in the conversion of androgens to estrogens. SERMs such as tamoxifen, toremifene, droloxifene, raloxifene and arzoxifene bind to estrogen receptors thereby preventing the action of estrogens. ERDs such as fulvestrant work by downregulating estrogen receptors and promote their degradation thus preventing estrogens from stimulating cell growth.

Estrogens are also used in combined oral contraceptives as an effective means to prevent pregnancy. Combined oral contraceptives contain estrogens and progestin. Estrogens play a crucial role in oral contraceptives as it has been shown that estrogen containing contraceptives are more effective than progestin only pill. Estrogens decrease

the secretion of FSH which inhibits follicular development and helps prevent ovulation and fertility.

2.2 Methods:

2.2.1 Identification of drugs having a similar gene expression profile as that of estradiol:

A total of top 40 Estradiol regulated genes, 20 up regulated and 20 down regulated (Table 3) was obtained from a previous research work that studied the effects of endoxifen, 4-hydroxy-tamoxifen and estradiol on the global gene expression patterns in MCF7 cells using the Affymetrix U133A GeneChip Array [70]. The 40 estradiol sensitive genes selected for this study were mapped to their respective probe ID's by querying the Affymetrix probe database through the NetAffy application. The final probe set representing the 20 up and 20 down regulated genes was queried against the CMap database that contains the drug-exposure gene expression data for 1309 compounds measured on cultured human cancer cell lines, to identify compounds that shared a similar or opposite gene expression profile to that of estradiol.

Upregulated	Downregulated
AREG	BAGE
CA2	BCAS1
CA8	BLNK
CAP2	C18orf1
CDC20	C2orf54
CXCL12	CALCOCO1
EGR3	CALML5
FHL1	CSGALNACT1
GREB1	DAB2
MDC1	DDIT4
MYBL1	FBN2
NPY1R	GABARAPL3
NR5A2	GABBR2
PDZK1	HOP
RAI14	IGFBP3
RPLP2	PEX14
SERPINA3	PLSCR4
SGK3	PNRC1
SNRPA1	PSCA
SOX3	SAMD4A

Table 3: Top 20 up and 20 down regulated estradiol sensitive genes.

2.2.2 Effect of phenazopyridine on MCF-7 cell proliferation:

MCF-7 cells were grown in DMEM medium supplemented with 10% fetal bovine serum in a T-75 flask maintained at 37ºC in 5% CO_2 incubator. Once the cell culture reached around 80% confluency, the media was removed, cells trypsinized and re-suspended in DMEM medium. The cell count was measured using the Beckman Z Coulter Counter. A solution containing 10,000 cells/ml was prepared. The cells were plated in 96-well plates at a seeding density of 1000 cells/well and allowed to grow for 24

hours. Phenazopyridine was prepared at the following concentrations, 100uM, 31.6uM, 10uM, 3.2uM, 1uM, 0.32uM, 0.1uM and 0.032uM. Each treatment was performed in 6 wells including the vehicle (0.1% ethanol) which had 12 wells (Figure 8). The cells were treated for a total of 7 days with the media being replaced initial after 72hours and then every 48 hours for the remainder of days. On the 7th day the media was removed and the cells stained with a solution containing 0.5% crystal violet and 25% methanol for 10 minutes. Following which the wells were rinsed with water 7 times and then 100ul of citrate buffer was added to each well. The absorbance was measured at 570nm using the Synergy 2 plate reader.

Plate	1	2	3	4	5	6	7	8	9	10	11	12
A												
B		100uM	31.6uM	10uM	3.2uM	0uM	0uM	1uM	0.32uM	0.1uM	0.032uM	
C		100uM	31.6uM	10uM	3.2uM	0uM	0uM	1uM	0.32uM	0.1uM	0.032uM	
D		100uM	31.6uM	10uM	3.2uM	0uM	0uM	1uM	0.32uM	0.1uM	0.032uM	
E		100uM	31.6uM	10uM	3.2uM	0uM	0uM	1uM	0.32uM	0.1uM	0.032uM	
F		100uM	31.6uM	10uM	3.2uM	0uM	0uM	1uM	0.32uM	0.1uM	0.032uM	
G		100uM	31.6uM	10uM	3.2uM	0uM	0uM	1uM	0.32uM	0.1uM	0.032uM	
H												

Figure 8: Plate layout for the treatment of MCF-7 cells with phenazopyridine.

2.2.3 Effect of estradiol on phenazopyridine induced inhibition of cell proliferation:

This experiment was carried out to see if estradiol had any effect on the phenazopyridine induced inhibition of MCF-7 cell proliferation. The experimental procedure was followed as described above, except that the media was replaced after the initial 24 hours with DMEM supplemented with 10% charcoal stripped calf serum and included the treatments with estradiol prepared at two concentrations, 1nM and 10nM. The MCF-7 cells were plated in two 96 well plates at a seeding density of 1000cells/well.

Each plate received the drug treatment for phenazopyridine as shown in Figure 8 along with either 1nM or 10nM of estradiol.

2.2.4 Influence of phenazopyridine and estradiol on progesterone receptor expression:

MCF-7 cells were plated in four T-25 flask at a seeding density of 0.25 X 10^6 cells in DMEM supplemented with 10% fetal bovine serum (FBS). After 24 hours the cells were rinsed 3 times with DMEM containing 10% charcoal stripped calf serum and the media was changed to DMEM supplemented with 10% charcoal stripped calf serum. The cells were allowed to grow in the new media for an additional two days with media being replaced every 24 hours. After 72 hours of initial media change the cells were treated with different combinations of the drugs as shown in Table 4 and allowed to grow for an additional 24 hours. The cells were harvested and RNA extracted using the miRNeasy (Qiagen) and quantified using the Quant-iT RNA kit (Life technologies) on the Qubit Fluorometer (Life Technologies, USA) according to the manufactures protocol. The cDNA was synthesized using QuantiTect Reverse Transcription Kit (Qiagen Inc. USA) and the expression of PGR and GAPDH was measured on the iCycler (BioRad) using the respective Taqman Gene Expression assays (Life Technologies).

Flask T-25	Estradiol (10nM)	Phenazopyridine (10nM)
1	X	X
2	✓	X
3	✓	✓
4	X	✓

Table 4: The various combinations of treatments that were used to study the effect of the estradiol and phenazopyridine on PGR expression.

2.2.5 Effect of phenazopyridine on tumor growth in athymic nude mice:

A total of 50 athymic nude mice were used for the experiment. The animals were caged and experiments performed at the Indiana University in vivo therapeutics core laboratory following the approval and guidelines of the University. The MCF-7 cells were grown, harvested and stored as pellets containing 3 X 10^6 cells/pellet. To initiate the tumors, the pellet were implanted into the mammary fat pad of the athymic nude mice and allowed to grow for 5 weeks. In addition, to promote the growth of the estrogen receptor positive MCF-7 tumor, the mice were also implanted with estradiol pellets (Innovative Research of American, Florida). The tumor volume was measured on an average every 5 days by measuring the two longest perpendicular diameters of the tumor. Following the 5 weeks after MCF-7 cell and estradiol pellet implantation the mice that survived were divided into 4 groups, each consisting of 10 mice. The animals were separated into the individual groups maintaining the same average body weight across the groups. Each group received one of the following treatments: vehicle, 0.2mg/kg/day, 2 mg/kg/day or 20mg/kg/day for a period of two weeks. The tumor volumes were measured

on an average every 4 days. At the end of the treatment the mice were sacrificed and the tumors were harvested. One half of the tumors was stored in RNA later and the other half was fixed in formalin and embedded in paraffin blocks.

2.3 Results:

2.3.1 Drugs with similar gene expression signature as estradiol:

The gene expression profiles for drugs that are similar to that of estradiol was retrieved and Table 5 shows the top 9 compounds. The drugs that have a positive mean value are those that have a similar gene profile to that of estradiol and those that have a negative mean value have an opposite effect. All the results had a significant p-value associated with them. From the table it can be seen that the top four drugs namely, genistein, fulvestrant, estradiol and LY-294002 are drugs that have documented evidence of their estrogenic and antiestrogenic activities. They thus served as internal controls in validating the authenticity of this approach. It has been previously shown that genistein has similar action to that of estradiol in enhancing proliferation of MCF-7 breast cancer cells[71]. The antiestrogenic activity of fulvestrant has been well demonstrated with its effective use in treating breast cancer [72-81]. The eighth drug namely phenazopyridine with a mean of -0.53 is predicted to have a strong antiestrogenic activity with a significant p-value of 0.0019 (Table 5).

rank	cmap name	mean	p-value
1	Genistein	0.43	0
2	Fulvestrant	-0.42	0
3	Estradiol	0.34	0
4	LY-294002	-0.29	0
5	Tanespimycin	-0.32	0
6	Trichostatin A	-0.26	0
7	Tretinoin	-0.31	0.0004
8	Phenazopyridine	-0.53	0.0019
9	Sirolimus	-0.25	0.0022

Table 5: Drugs with similar (+ mean) or opposite (- mean) gene expression signature to that of estradiol.

2.3.2 Phenazopyridine inhibits the proliferation of MCF-7 cells:

In order to validate the antiestrogenic activity of phenazopyridine, we analyzed the effect of the drug on the estrogen receptor positive cell line, MCF-7. These cells are estrogen dependent for their growth and thus an antiestrogen would be expected to interfere with the growth of these cells. The cells were grown in 96 well plates in FBS for 7 days and treated with different concentrations of phenazopyridine following which they were stained and absorbance measured. The results showed a decline in number of viable cells which indicated that phenazopyridine inhibited MCF-7 cell proliferation in a dose dependent manner (Figure 9). Thus providing evidence for a possible antiestrogenic action exerted by phenazopyridine on the estrogen dependent MCF-7 cells.

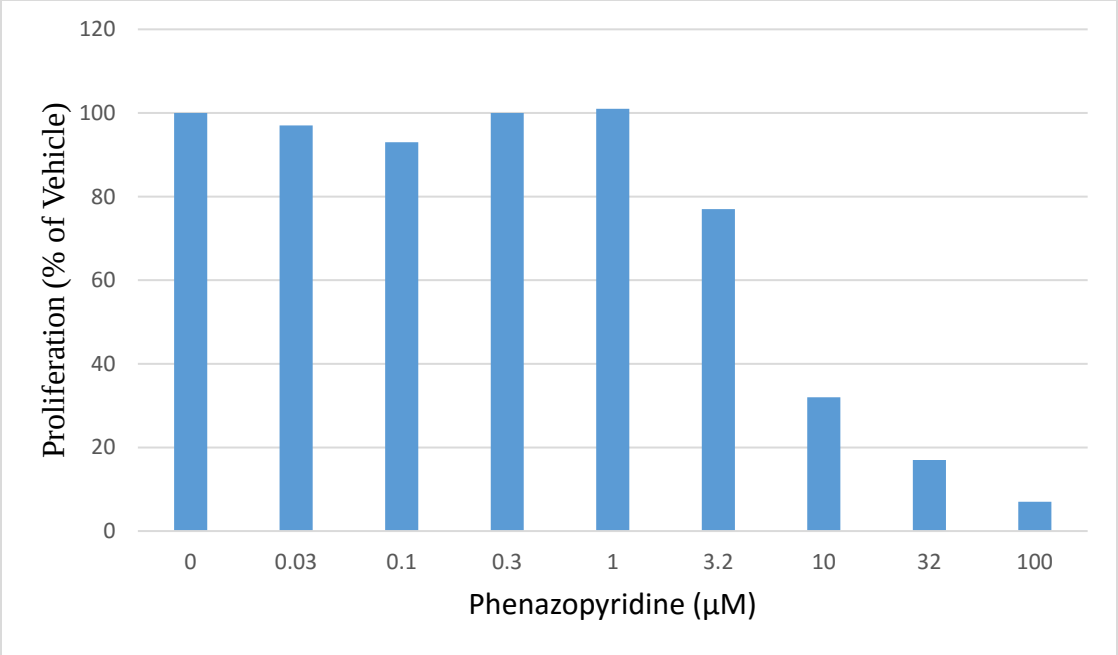

Figure 9: Phenazopyridine inhibits the proliferation of MCF-7 cells.

2.3.3 Effect of estradiol on phenazopyridine induced cell proliferation:

It is well known that estradiol fuels the growth and proliferation of MCF-7 cells [82-87]. The goal here was to find out if estradiol had an effect on phenazopyridine's inhibitory action on the proliferation of MCF-7 cells. As shown in the Figure 10 estradiol at 1nM and 10nM did not interfere with the inhibitory action exerted by phenazopyridine.

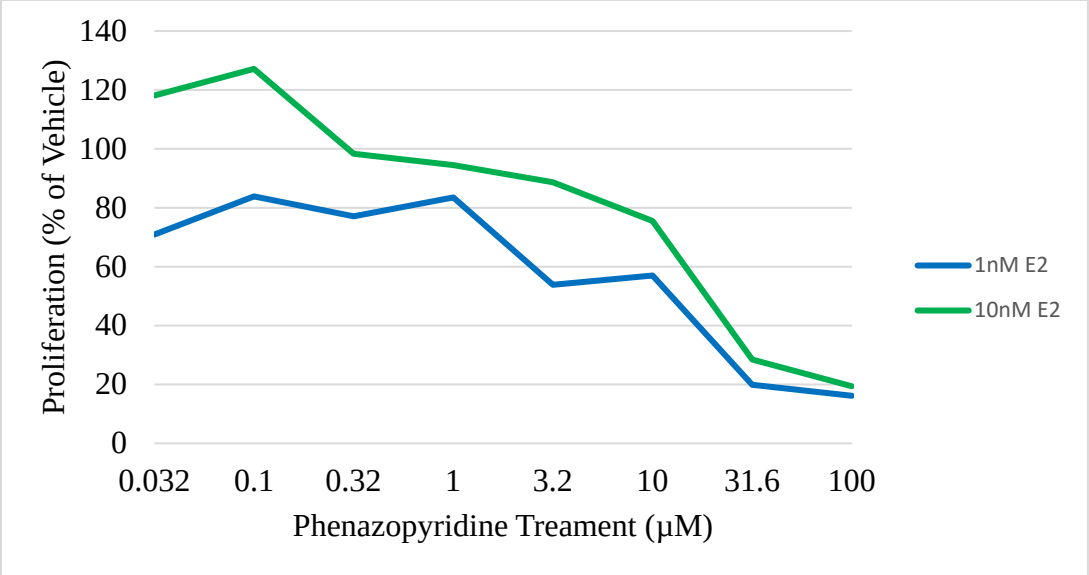

Figure 10: Estradiol does not inhibit the antiproliferative activity of phenazopyridine.

2.3.4 Phenazopyridine effect on PGR expression:

It has been previously shown that there exists a positive correlation between estrogen and progesterone receptor (PGR) expression [88, 89]. This fact was exploited in our study to see the effect of phenazopyridine on the estradiol sensitive gene. Our results showed that phenazopyridine was able to decrease the estradiol induced PGR expression by approximately 15%, which shows that it has some influence on the estradiol induced PGR expression (Figure 11).

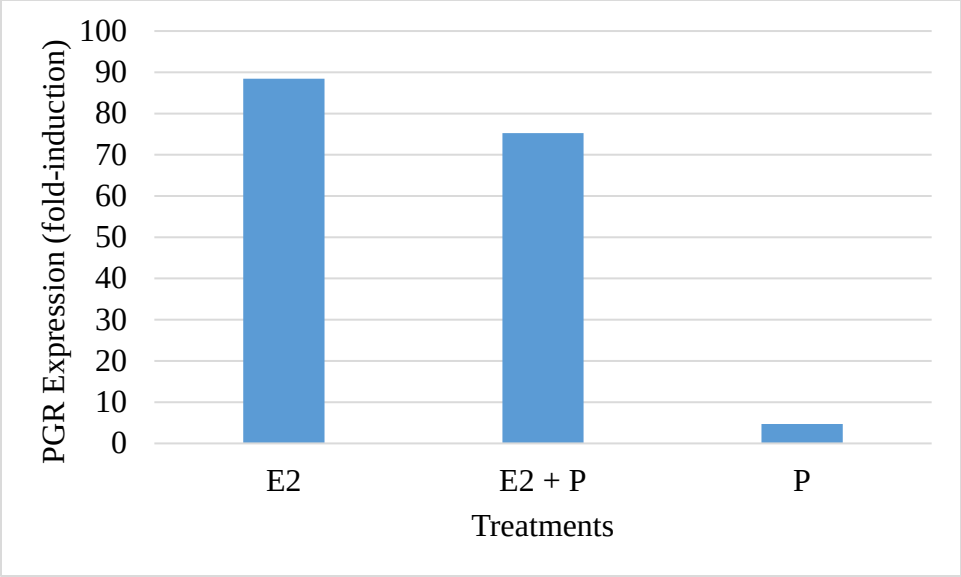

Figure 11: Phenazopyridine partially inhibits the estradiol induced PGR expression.

2.3.5 Phenazopyridine effect of tumor growth in vivo:

To determine if phenazopyridine had similar effects in vivo, nude mice were injected with MCF-7 cells along with estradiol pellets. The mice were then treated with three different concentrations of phenazopyridine and the tumor volumes were measured. The vehicle and low doses (0.2 & 2mg/kg/day) had no effect on the tumor growth. The highest concentration (20mg/kg/day) had an initial effect in inhibiting tumor growth, but lost its effect after a period of 7 days suggesting that the tumor may have readjusted in some manner to overcome the initial inhibition from phenazopyridine (Figure 12).

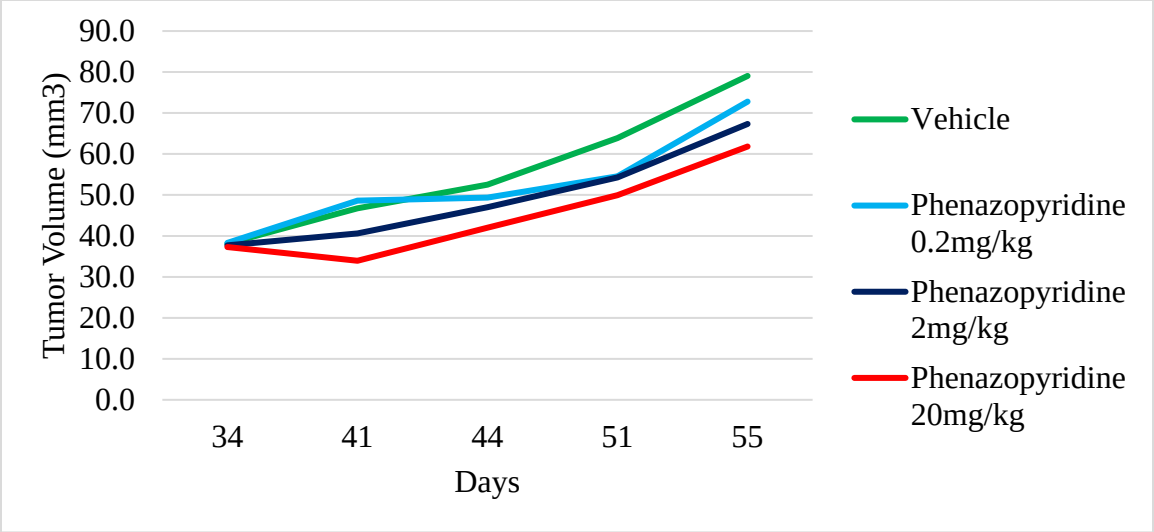

Figure 12: Effect of Phenazopyridine on the implanted MCF-7 tumor growth in athymic mice.

2.4 Discussion:

In this part of the dissertation we have shown the effectiveness of using gene expression profiles in identifying new indications for existing drugs. Drug repurposing is one of the most efficient ways to identify new indications for existing drugs as well as for those that have failed clinical trials. Since most drugs hit multiple targets other than the target they are intended for [90], it is very likely that a given drugs side effect may hold the key to treatment of complex and rare disease. One of the most outstanding case of drug repurposing is that of thalidomide. It was originally developed to treat morning sickness during pregnancy, but had to be withdrawn due to its tragically dangerous side effects. A drug that was so dangerously toxic was several years later accidentally discovered to be beneficial in treating complications of leprosy and later in treating multiple myeloma. This brings out the fact that even the worst of side effects can find beneficial use in another disease state. The advances in technology has made it possible

to swift through large data sets, may it be from expression or literature to identify interconnected pathways that can be exploited to drive novel therapeutic interventions.

In this study we were able to discover the antiestrogenic potential existent in a urinary analgesic drug. Phenazopyridine is commonly prescribed to relieve urinary tract symptoms such as pain, burning, irritation and discomfort as well as urgent and frequent urination caused by urinary tract infections, surgery, injury or examination procedures[91]. Through in vitro proliferation studies on estrogen dependent MCF-7 cells we were able to confirm a potential antiestrogenic indication for this urinary analgesic. Further, we tested the effect of phenazopyridine in athymic nude mice that were injected with MCF7 cells and observed a reduction of the tumor volume during the initial phase of treatment at the highest dose.

The fact that estradiol did not interfere with phenazopyridine's antiproliferative effect could suggest that phenazopyridine may not be exerting its action at the receptor level. It is well know that there exists a positive correlation between estradiol and progesterone receptor expression. A study has previously shown that the estrogen induced transcription of the progesterone receptor gene does not equal estrogen receptor occupancy[92]. From our results, phenazopyridine does show a partial inhibition of estrogen induced progesterone receptor expression. This and the above fact suggests that phenazopyridine could be hitting targets downstream of the estrogen receptor in mediating its antiproliferative action.

The in vivo study showed an initial decrease in tumor volume at the highest dose of phenazopyridine (20mg/kg/day) that lasted for only 7 days, following which the tumor continued to growth. This observation does not come as a surprise, as most cancers show

resistance to treatments. The resistance could be either primary where it exists prior to treatment or acquired where it develops following treatment [93, 94]. Compared to the in vitro environment, the in vivo environment comes with many levels of complexity and there are many more factors that come into play and additional resources available for the drug resistance variant cell type to utilize and over grow rapidly thereby acquiring resistance. This could be a possible explanation to resistance observed to phenazopyridine after initial treatment at 20mg/kg/day. In addition, the presence of the estradiol pellets may have played a role in further fueling this drug resistance acquired tumor growth. This resistance to drug observed here, confirms the difficulty that arise in treating cancers or the resistance seen to treatment as the tumors self-adjust to the tumor environment and are able to increase their immunity towards treatment.

The fact that phenazopyridine showed a positive antiestrogenic action in vitro but was only partially seen in vivo brings out the complexity of the in vivo system compared to the basic unit a cell that was used in the in vitro studies. It clearly shows the necessity in identifying the network of molecular events that play a role in tumor progression in vivo when there are more resources and biological components that can influence and help the tumor cells adapt and evade drug action.

Apart from its possible role in inhibiting cell proliferation, the predicted antiestrogenic activity may also interfere with combination oral contraceptives that are composed of estrogen and progestin. It can be argued that an antiestrogen could interfere with the estrogen levels and thus affect the effectiveness of the oral contraceptive pill. Not taking a pill on the prescribe schedule can lead to the ineffectiveness of the therapy, similarly an interference in estrogen levels through an antiestrogen, in this case

phenazopyridine which is primarily prescribed as a urinary analgesic could potentially reduce the effectiveness of oral contraceptives. Since our study is the first to identify a potential antiestrogenic action for phenazopyridine this drug interaction would never have been considered before. Retrospective studies will need to be performed to measure the frequency at which the two drugs have been prescribed together and if any unexpected outcomes were reported. This could be used to point out the possibility of phenazopyridine's antiestrogenic property that could interfere with the oral contraceptives mechanism of action.

These results prove the immense potential that informatics can play in discovering novel interaction and in identifying new indications to failed and existing drugs, thereby opening venues for novel therapeutic interventions. In this case a urinary analgesic that holds the potential to be used in treating cancer, if not by itself possibly in combination with other drugs, and its possible interaction with oral contraceptives. We can conclude that from the informatics and preliminary experimental analysis, phenazopyridine may be a potential candidate for development as a novel therapeutic agent with antiestrogenic properties. The precise underlying mechanism of it antiestrogenic action and its effect being overcome in the in vivo environment of athymic mice calls for further experimental validation by domain experts to elucidate the pathways that are activated there by blocking its anti-estrogenic activity in vivo.

Till date there has been no reports of phenazopyridine being used for an indication other than as a urinary analgesic. Our discovery of the antiestrogenic action of phenazopyridine is novel and has the potential of being used as an antiestrogen and possibly in treatment of cancers that are estrogen dependent.

Chapter 3: Genes essential for cell survival

3.1 Introduction:

3.1.1 Knowledge in literature:

There is no lack for data or scientific literature as they continue to grow at an exceedingly exponential rate, yet there is this unquenchable thirst for knowledge. Figure 13 shows the growth in MEDLINE abstracts between 2001 and 2014. The knowledge that can lead to new discoveries, aid in making clinical decisions and designing efficient therapeutic strategies are hidden within this huge mass of data and literature. It has been shown decades earlier that the medical literature holds hidden knowledge that can be exploited in treating complex diseases [95-100]. In spite of the availability of this huge amounts of literature two thirds of the questions that clinicians raise about patient care in their practices remain unanswered[101]. These question most often could be classified into a small set of generic questions[102] but require a diverse set of answers based on the clinicians specialty. With the advances in technology and the completion of the human genome we have data, but the challenge lies in how to identify the crucial knowledge that can lead to a better understanding of the disease pathology and equip the clinician to make informed decisions as to the best course of therapeutic action. In addition the various factors that can influence or contribute to disease susceptibility or progression poses a challenge to scientist in finding a preventative or therapeutic solution for these diseases [103-105]. The challenges in finding a cure are proportionally increasing with complexity presented by the disease. The question most commonly asked when dealing with huge amounts of data is, how the low value data can be transformed to high value knowledge that can be applied to treating complex diseases more effectively.

As shown in the Figure 14, data can be transformed into wisdom through the various stages employing various processes along the way[106].

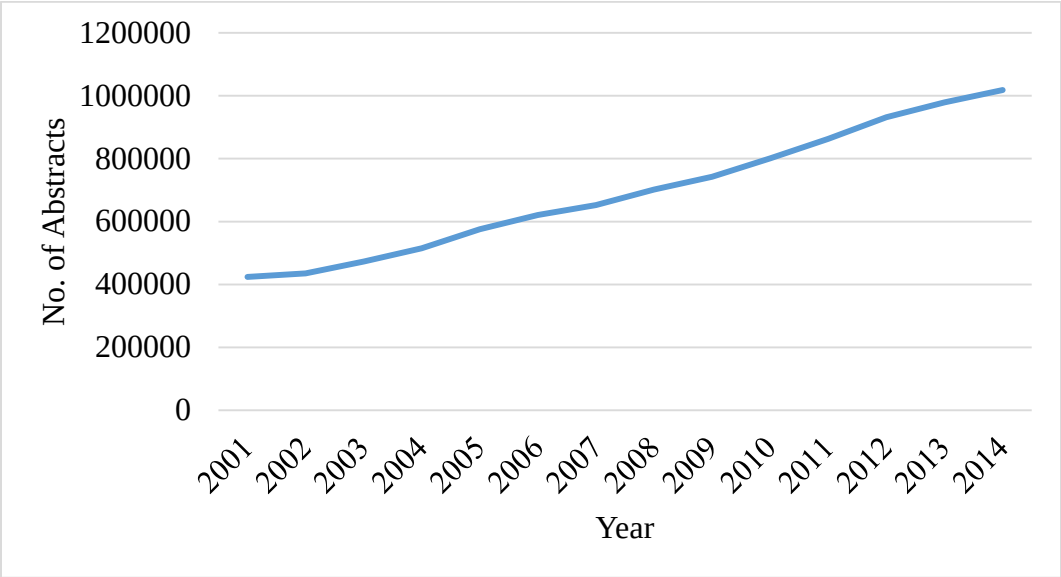

Figure 13: Growth of MEDLINE abstracts between 2001 and 2014.

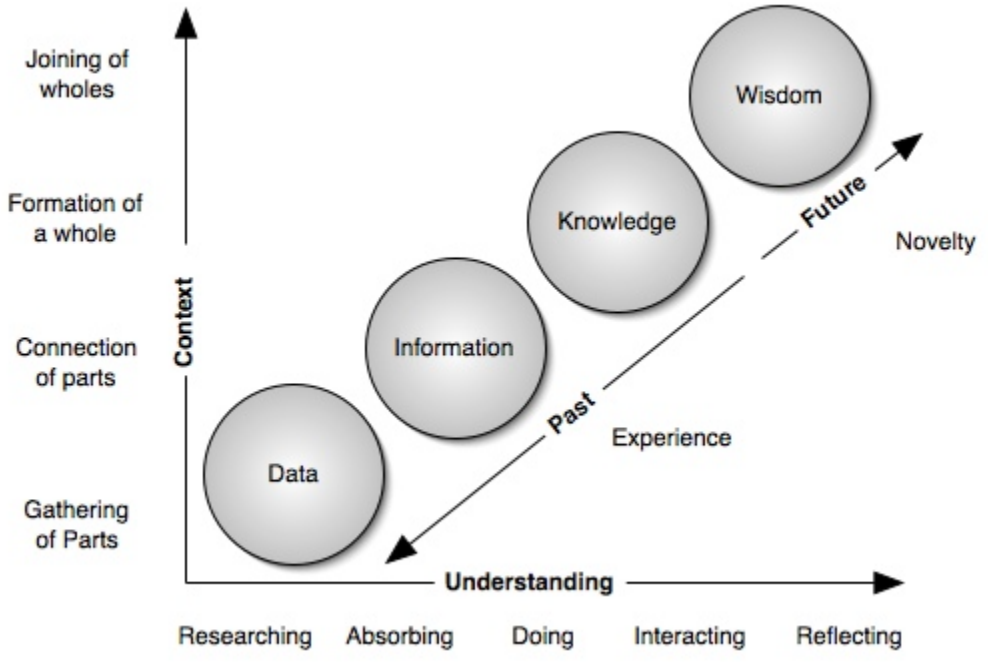

Figure 14: The continuum of understanding

The heterogeneous nature of the scientific literature across multiple disciplines is something that can be exploited to identify crucial knowledge that underlies the essence of survival. The free availability of this unstructured text makes it the biggest and most widely used for the identification of new knowledge. It would be highly impossible for a human to devour this huge amount of literature to identify the dots that connect various components within a pathway that can be targeted to effectively treat a disease, especially when the information is present in non-interacting articles. Manual curation is a possibility with the advantage of being accurate, but comes at a high cost of time, labor and finding expertise in multiple disciplines. The use of computers and more specifically machine learning algorithms that can be trained to identify relevant literature and then extract the relationships between entities of interest to produce clinically applicable knowledge is gaining popularity in the race to find cures. The later though highly scalable with the ever increasing growth of literature is error prone due to the complexity of natural languages used. The ultimate goal of information access is to help the user or practitioner in finding relevant documents that satisfy their information needs so they can gain wisdom and apply it to their practice. The challenge still remains, how can we apply the continuum of understanding[106] in finding wisdom from the huge amounts data.

3.1.2 Text mining for knowledge discovery:

Text mining can be defined as the process of using computers to mine through large amounts of unstructured text to discover and extract knowledge. At the core of text mining is the ability to identify scientific information across diverse research fields and connect these pieces of information to generate hypotheses and discover new previously unknown connections that can be useful in developing therapeutic diagnostic tests or lead

to knowledge of prevention or treatment of diseases. There is no lack for data, but connecting the information across diverse disciplines is challenging [107-109].

As shown in Figure 15, text mining basically consists of the following steps, namely, (i) information retrieval (IR), (ii) information extraction (IE), (iii) knowledge discovery and (iv) hypothesis generation [110]. Information retrieval (IR) is the very first step in the text mining process and involves the retrieval of texts that are relevant to the user's topic of interest. This step is initiated through the use of topic specific keywords to query the bibliographic databases such as MEDLINE. PubMed is the most commonly used IR system by majority of researchers, which houses more than 25 million references to scientific literature. Information extraction involves locating textual mentions of entities of interest and their relations with the aim of representing them in a structured format for easy of analysis. This step is backed by named entity recognition (NER), which identifies the relevant entities such as drug names, gene names, disease names or those that the user is interested from a predefined set of categories. NER is a very critical step in the information extraction task and falls into three main categories, namely: (i) dictionary-based[111, 112], (ii) rule-based[113], and machine learning based[114]. Knowledge discovery involves the creation of new knowledge by connecting the information extracted from the unstructured text in the preceding steps. Hypothesis generation involves identifying chance connections that hold the potential to be a solution to a problem. The hypothesis thus generated can be further validated through conventional experimentation. Text mining has many applications from business to medical research and has been applied to various cancer domains such as breast[115-118], lung[119-121], pancreas[122], prostate[118, 119, 123-125], and ovary[126].

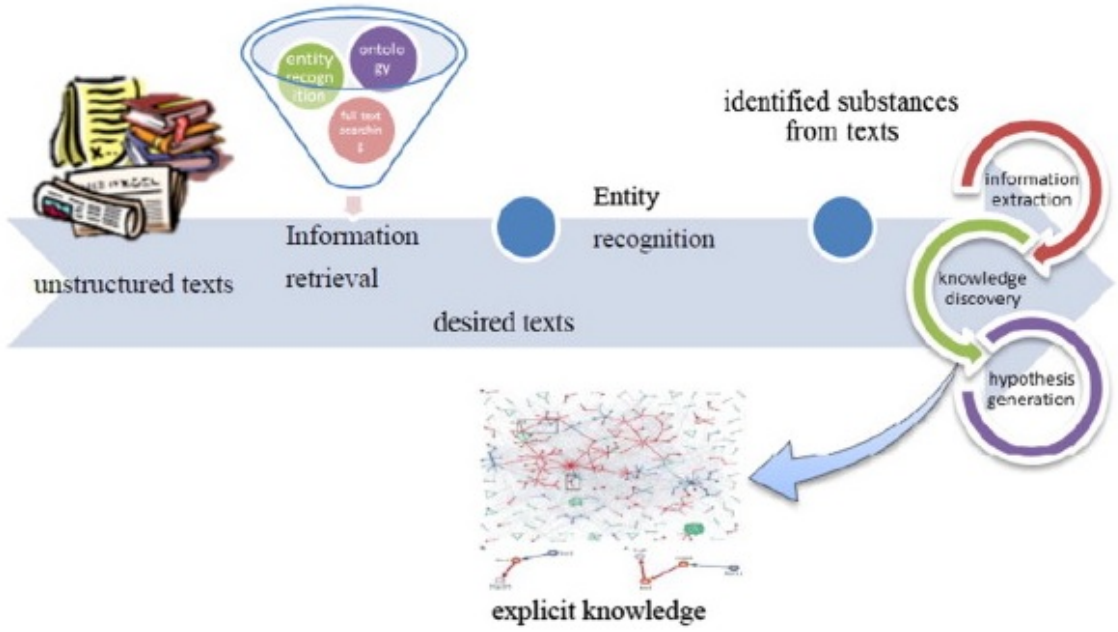

Figure 15: The basic steps involved in text mining.

Machine learning as the name states is the process where computer programs are designed to learn unique features about a dataset so they can make predictions about a previously unseen dataset. These methods are trained to distinguish between the relevant and irrelevant text documents based on the word content of these documents. The most common application is the development of a classifier that can distinguish texts into two or more classes based on the attributes measured in each distinct text class.

WEKA (Waikato Environment for Knowledge Analysis) workbench is a collection of state of the art machine learning algorithms and data preprocessing tools developed at the University of Waikato[127]. It includes methods for main stream data mining problems such as regression, classification, clustering, association rule mining

and attribute selection. Apart from the various algorithms it also provides methods for evaluation such as cross validation and standard numeric performance measures such as accuracy and root mean squared error. The availability of multiple algorithms and filters provides maximum flexibility to test on various datasets. It accepts input in its native file format called attribute relation file format (arff). It is implemented in java and runs on almost any platform. As shown in Figure 16 weka has multiple interfaces to it. The Explorer is the main and most commonly used interface in weka. It consists of six panels as shown in Figure 17. The preprocess panel is used for loading the dataset and preprocessing the data using one of weka's inbuilt filters. If the data involves a classification or regression problem this can be handled in the classifier panel, which provides the various classification algorithms. The major algorithms for classification and regression in weka are, support vector machines, decision trees, rule sets, bayesian classifiers, logistic and linear regression, multi-layer perceptron and nearest-neighbor. It also has meta-learners such as bagging, boosting, stacking. The third and fourth panel contain algorithms for clustering such as k-means and those to generate association to identify relationships between groups of attributes in the data. The fifth panel offers methods that can be used to identify attributes that are predictive of other attributes in the data. The sixth panel provides visualization tools for plotting the attributes representative of the data.

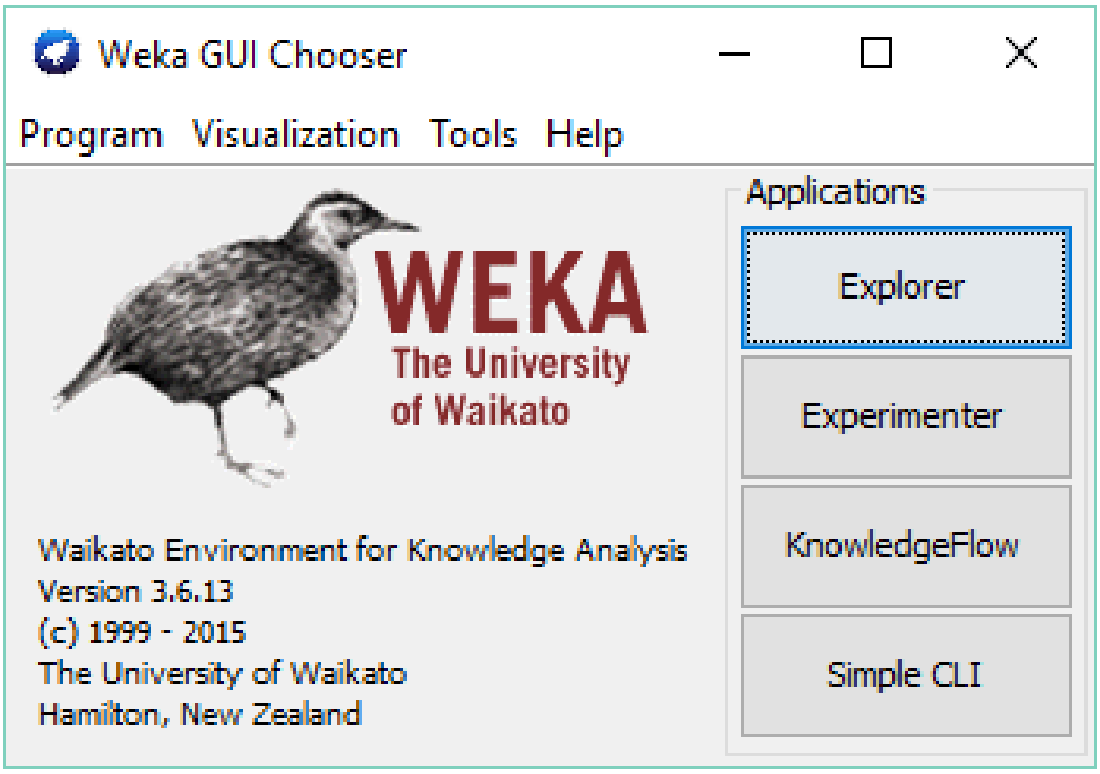

Figure 16: The WEKA interface and available tools.

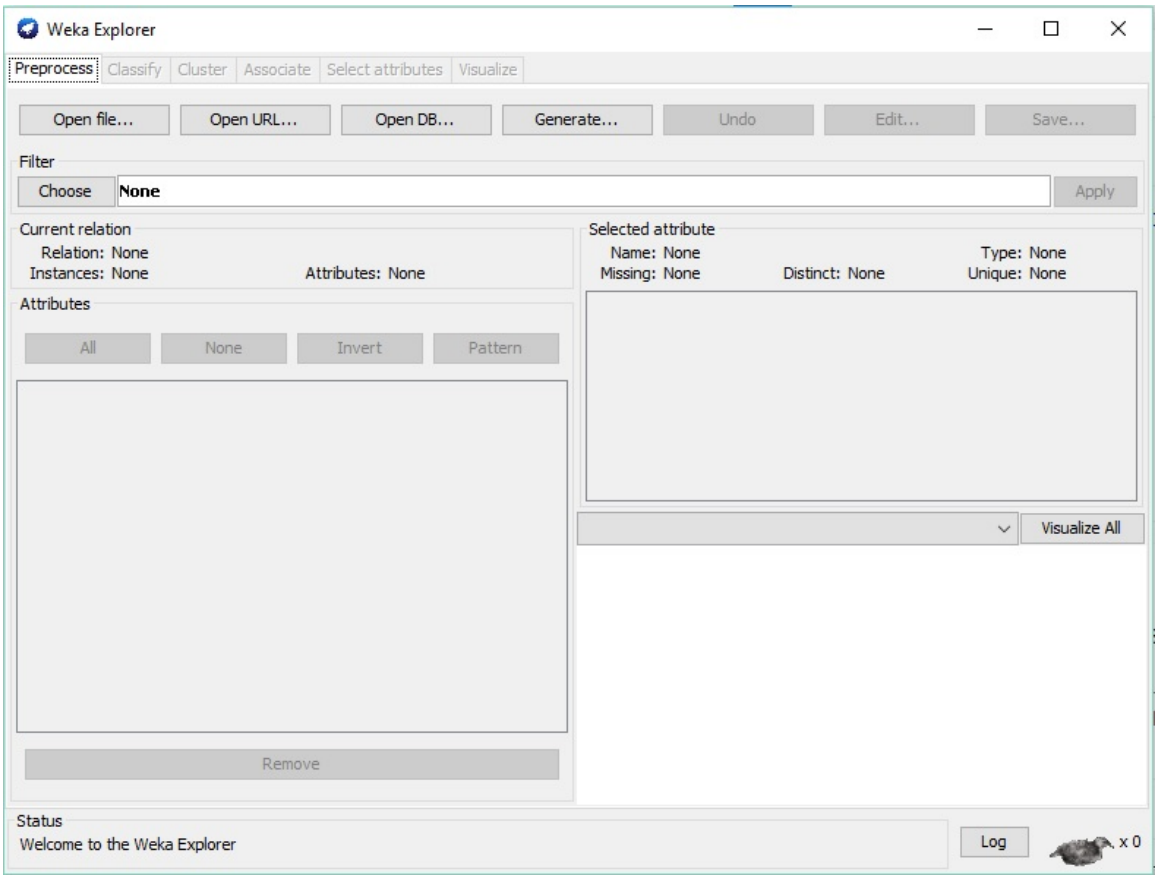

Figure 17: The WEKA explorer.

3.1.3 Model evaluation:

Evaluating a model is an essential step as it provides information on the performance as well as how well a model may perform on future unseen datasets. Cross validation is a very good evaluation method when the available data is limited. Especially when the data generated has to be manually selected by an expert. Before running the cross validation the number of folds or how the data needs to be partitioned is decided. In a 10 fold cross validation, the data is partitioned into 10 equal parts and then 9 parts are used for training and one part is used for testing. This process is repeated 10 times, so each part has been used exactly once for testing. If the representative classes are not

equally distributed, this could lead to overrepresentation of a class in the test set because there was not enough representation of the class in the training set. Use of stratified cross validation can to an extent safeguard against such a situation. In the case of stratified cross validation the data is randomly divided into equal parts with the class representation maintained in close approximation to the entire dataset. The error estimate is calculated for each fold and finally the average of the individual errors is reported.

There are many performance measures that can be used to evaluate a classifier, some of which are as follows. Given a classifier that classifies text documents into two classes namely class A (positive) and B (negative). The correctness of classifier can be evaluated by calculating the number of instances that were correctly classified into class A (true positive), the number of instances that were correctly classified as not belonging to class A but belonging to class B (true negative), the number of instances that were wrongly classified as belonging to class A (false positive) and the number of instances that were wrongly classified as belonging to class B (false negative). These four values make up the confusion matrix as shown in Table 6.

		Predicted Value	
		Positive	**Negative**
Actual Value	**Positive**	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Table 6: Confusion matrix for binary text classification.

From the confusion matrix the following measures can be calculated:

Accuracy: This provides the overall effectiveness of the classifier and is calculated as follows (TP + TN) / (TP + FP + FN + TN).

Precision: This provides the proportion of documents classified as positive that are indeed truly positive and is calculated as follows, (TP) / (TP + FP).

Recall (sensitivity): This provides the proportion of documents from the overall dataset that were classified as positive and is calculated as follows, (TP) / (TP + FN).

F-score: This provides the accuracy of the binary classification and is calculated from the precision and recall as follows 2 * ((precision * recall) / (precision + recall)).

Specificity: This provides the effectiveness of the classifier in identifying the negative instances and is calculated as follows (TN) / (FP + TN).

The performance of a classifier can also be obtained from its error rate. This error rate needs to be estimated from a new previously unseen data, namely the test dataset. In this way the performance of the model is evaluated on the data that had no part in its training. The test data set should be representative of the classification problem at hand, basically it should be similar to the training dataset. To truly estimate the performance of the classifier it is very important to make sure that there is no overlap of data between the training and test datasets. Some of the commonly used methods are kappa statistic, mean absolute error, root mean squared error, relative absolute error and root relative squared error.

3.1.4 RNA interference:

RNA interference is a very powerful biological process that involves the silencing of gene expression in eukaryotic cells [128-133]. It is indeed a natural host defense mechanism by which exogenous genes, such as viruses are degraded [134-136]. Figure 18, shows the mechanism through which gene silencing is achieved by RNA interference[137]. With the emergence of the RNA interference technology, scientist have been able to study the consequences of depleting the expression of specific genes that code for pathological proteins and are able to observe the resultant cellular phenotypes, which can provide insights into the significance of the gene. Diseases that are associated or driven by genes, such as cancer, autoimmune disease and viral disease can take advantage of RNA interference to generate a new class of therapeutics. Synthetic RNAi can be developed to trigger the RNA interference machinery to produce the desired silencing of genes [138-140]. The power of this process can be harnessed to identify and validated drug targets and also in the development of targeted gene specific medicine.

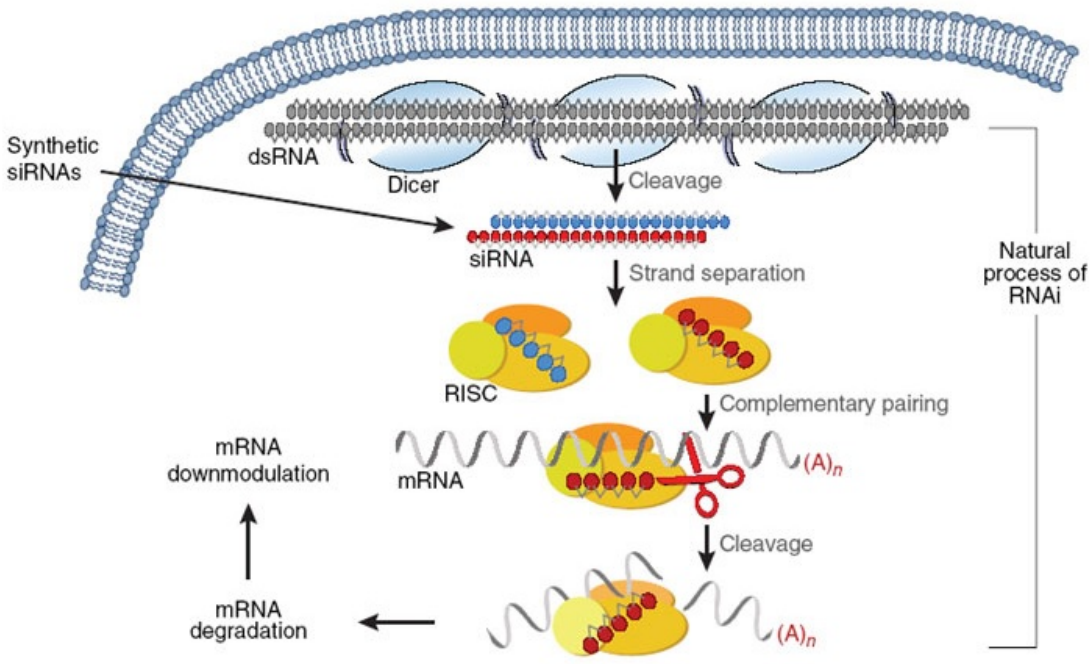

Figure 18: Mechanism of RNA interference.

One of the benefits of RNA interference technology, is that it provides information about the function of genes within an organism and helps us in identifying essential genes. Essential genes are those that are very important towards the survival of a cell or organism[141]. Identification of the minimum essential genes required for a cell to survive and being able to generate distinct sets that can represent normal versus cancer cell survival will not only enhance our understanding on what causes a normal cell to progress into a cancerous cell but will also provide the precise location of the gene that is the driving force of uncontrolled cell proliferation. This crucial knowledge can guide in the development of targeted cancer treatments. For example, it is very evident today, that breast cancer is no longer a single disease but heterogeneous in nature requiring different prognosis and treatments [142-145]. Since tumors are highly heterogeneous in nature,

there may be more than one gene that needs to be targeted within the heterogeneous population of cells, which makes the treatment of cancer so complex. By identifying these essential genes, one can use them as building blocks to capture the heterogeneity of the tumor environment and improve the clinical decision making in treating them more effectively and with precision.

3.2 Methods:

3.2.1 Abstract selection and corpus construction:

The Medline database was queried for abstracts that studied the effects of siRNA or drugs on cell lines using the following boolean query structure [(siRNA or shRNA or drug) AND (cell line name)] across 6 different cell lines, namely MCF7, MCF10A, SKBR3, HS578T, BT20, and MDAMB231 The results of the query were downloaded as a list of PMIDs (PubMed Identifier). The articles for the list of PMIDs were download in the XML format from NCBI and processed as follows. The XML for each of the retrieved article was parsed to extract the PMID, article title and article abstract. These files formed the initial unfiltered set of abstracts and were converted to a pdf format to aid in the manual process of scanning them to select the most relevant abstracts to construct the text corpus. In addition these abstracts were further divided among four other individuals consisting of a high school student and three master's level students for manual scanning and classification. The abstracts were read and then grouped under four categories as follows:

i. RNAi: These abstracts had siRNA/shRNA being studied, along with the cell line used and the resultant cell phenotype.

ii. Drug: These abstracts had a drug being studied, along with the cell line used and the resultant cell phenotype.

iii. Drug-Drug: These abstracts had a drug interaction being studied, along with the cell line used and the resultant cell phenotype.

iv. NA (Not Applicable): If the abstract did not fall into any of the above categories it was labelled as NA.

For an abstract to be placed in any of the categories (i) – (iii) they needed to have all three components, namely siRNA or drug and cell line and resultant cell phenotype. If one of these components were not clearly stated or was missing, the abstract was placed in the NA category. Close to 2000 abstracts were manually screened using the above criteria.

3.2.2 Training and testing datasets:

The list of abstracts as PMIDs for each category was grouped together and converted into individual xml files respectively. These xml files were then processed to generate the individual text files represented by the PMIDs within each group. The text files for the abstracts classified as RNAi were used as the positive set and the remaining were used as the negative set. The training and testing datasets consisted of various combinations as shown in the Table 7. The text files representing the training and testing datasets were converted into the WEKA native file format, namely ARFF (attribute relation file format) using the java TextDirectoryLoader class. The ARFF file is an ASCII text file that describes a list of instances that share a set of attributes. The header of the file contains the name of the relation, the list of attributes along with their types, and the class information followed by the data. The classes were labelled as RNAi for the

abstracts from the positive set and Non_RNAi for the abstracts from the negative set. The structure of the .arff file is as shown in Figure 19. The final training set consisted of 120 RNAi abstracts in the RNAi class and a total of 1700 abstracts from drug, drug-drug, NA and RNS in the Non_RNAi class. The testing set consisted of 101 RNAi abstracts in the RNAi class and a total of 1700 abstracts from drug, drug-drug, NA and RNS in the Non_RNAi class.

Set	Training		Testing		Data
	Positive	Negative	Positive	Negative	
1	100	300	100	300	r,d,dd,g
2	100	100	100	100	r,d,dd,na
3	100	300	100	300	r,d,dd,na
4	100	400	100	400	r,d,dd,na,g
5	120	1700	101	1700	r,d,dd,na,rns

Table 7: Composition of the training and testing sets used to test the various weka classifiers. [r: RNAi abstracts, d: drug only abstract, dd: drug interaction abstracts, na: not applicable, rns: random negative set]

```
@relation c__weka_set_5_train

@attribute text string
@attribute filename string
@attribute @@class@@ {Non_RNAi,RNAi}

@data
'Title = Growth inhibition by dehydrothyrsiferol - a non-Pgp
modulator, derived from a marine red alga - in human breast
cancer cell lines.\n \nAbstract = The novel marine terpenoid
dehydrothyrsiferol (DHT) has been isolated from a Canarian red
alga Laurencia viridis sp.\nnov (Ceramiales, Rhodomelaceae) (1).
\nIts cytotoxicity against three human breast cancer cell lines,
namely T47D, ZR-75-1, and Hs578T was examined and compared with
the chemotherapeutic compound doxorubicin and the mitosis-
inhibitor colchicine.\nPrimary breast carcinomas exhibit MDR1
gene expression (3).\nAs the investigated mammary cell lines did
not exhibit rhodamine 123 efflux we proved in a P-glycoprotein
(Pgp) overexpressing human epidermoid cancer cell line that the
marine metabolite does not modulate Pgp mediated drug transport.
\nTherefore, it could be used in Pgp expressing cancer cells
without interference.\n','Non_RNAi\\10087323.txt',Non_RNAi
'Title = Abnormal levels and minimal activity of the dsRNA-
activated protein kinase, PKR, in breast carcinoma cells.\n
\nAbstract = The interferon induced, dsRNA-activated, protein
kinase, PKR, is a key regulator of translational initiation,
playing an important role in the regulation of cell
proliferation, apoptosis and transformation.\nPKR levels
correlate inversely with proliferative activity in several human
tumor systems.\nThis inverse relationship breaks down in human
invasive ductal breast carcinomas which exhibit high levels of
```

Figure 19: Structure of the WEKA ARFF file.

3.2.3 Selection of algorithm:

The training and testing set consists of the abstracts as shown in Table 7. Evaluation is key to identifying the best classifier that can perform the given task with the highest accuracy. With the limited amount of data for training and testing, the 10 fold stratified cross validation was chosen as the most appropriate method for evaluating the various classifiers. The dataset was evaluated using the following 7 classifiers, namely, ZeroR, NaiveBayes, K-nearest neighbor, J48, Random Forest, Support Vector Machine

and OneR. These are some of the most commonly used algorithms for text classification, except for ZeroR which was used here to get a baseline. The filtered classifier belonging to the WEKA meta classifier was used, since it has the advantage of simultaneous selection of a classifier and filter to evaluate the model. The various classifiers mentioned above were tested along with the string to word vector filter, as shown in Figure 20. The string to word filter converts string attributes into a set of attributes that represent the word occurrences from the text contained within the strings. The set of attributes is determined from the training data set. The string to word filter provides various options as shown in Figure 21 and is explained here. The IDFtransform or inverse document frequency transform calculates the word frequency of the term within the document set. The TFtransform or term frequency transform calculates the frequency of a given term with a document. The combination of IDF and TF reflects the importance of a word to a document within the corpus. The attributeIndices specifies the range of attributes to be acted upon. The attributeNamePrefix provides the option to add a prefix to the attribute names created. The doNotOperateOnPerClassBasis if set to true, results in the maximum number of words and the minimum term frequency not being enforced on a per class basis but would be based on the documents present in all the classes. The invertSelection sets the attribute selection, if false then only the selected attributes within the range will be processed, if true only non-selected attributes will be processed. The lowerCaseTokens allows for the conversion of all words to lower case before being added to the dictionary. The minTermFrequency allows for selecting the minimum term frequency and operates on a per class basis. The normalizeDocLength, sets whether the word frequencies for an instance should be normalized or not. The outputWordCounts if

set to true gives the exact count of the words rather than just their presence or absence. The periodicPruning specifies the rate at which to periodically prune the dictionary. The stemmer allows for the selection of a stemming algorithm to stem the words. The available stemmers are IteratedLovinsStemmer, LovinsStemmer or SnowBallStemmer. The stopWords option allows for the use of a stop word dictionary to eliminate stop words. The tokenizer option allows for the selection of the tokenizing algorithm to be used. The available options are, alphabeticTokenizer, nGramTokenizer and wordTokenizer. The useStopList provide the option to use an additional list of words to eliminate. If used, the word in the stop list are eliminated. The wordsToKeep option specifies the number of word to keep per class. The 10 fold stratified cross validation option was selected and the data from the training set (Table 7) was evaluated to identify the best classifier.

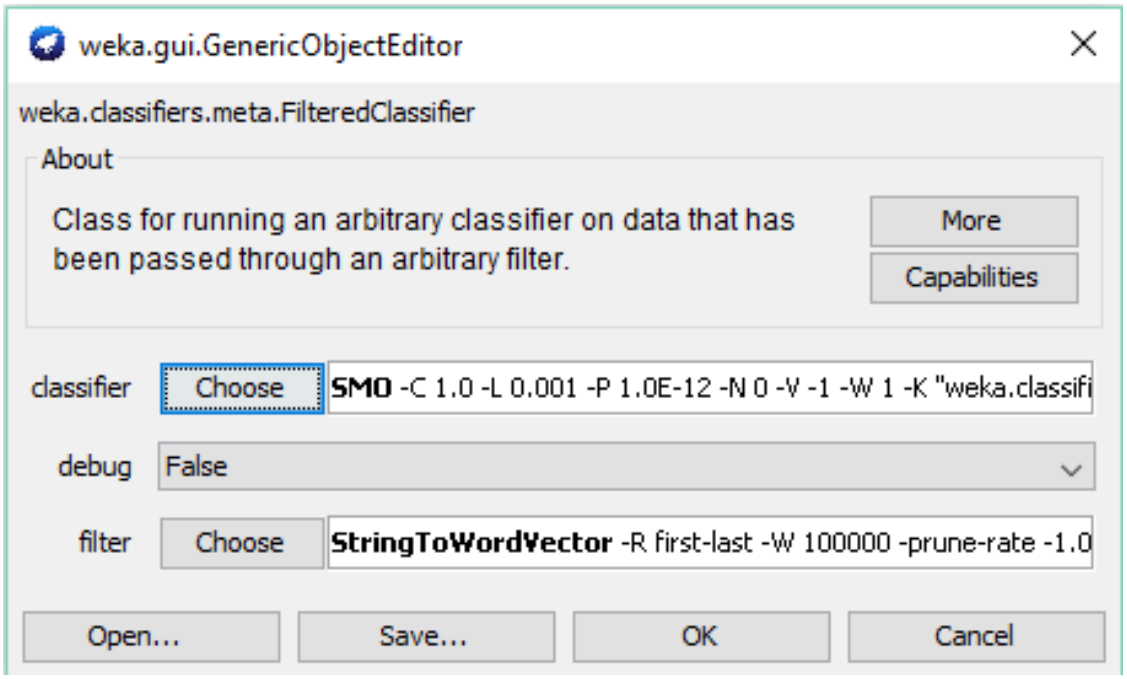

Figure 20: The filtered classifier option allowing for the simultaneous selection of a classifier and filter.

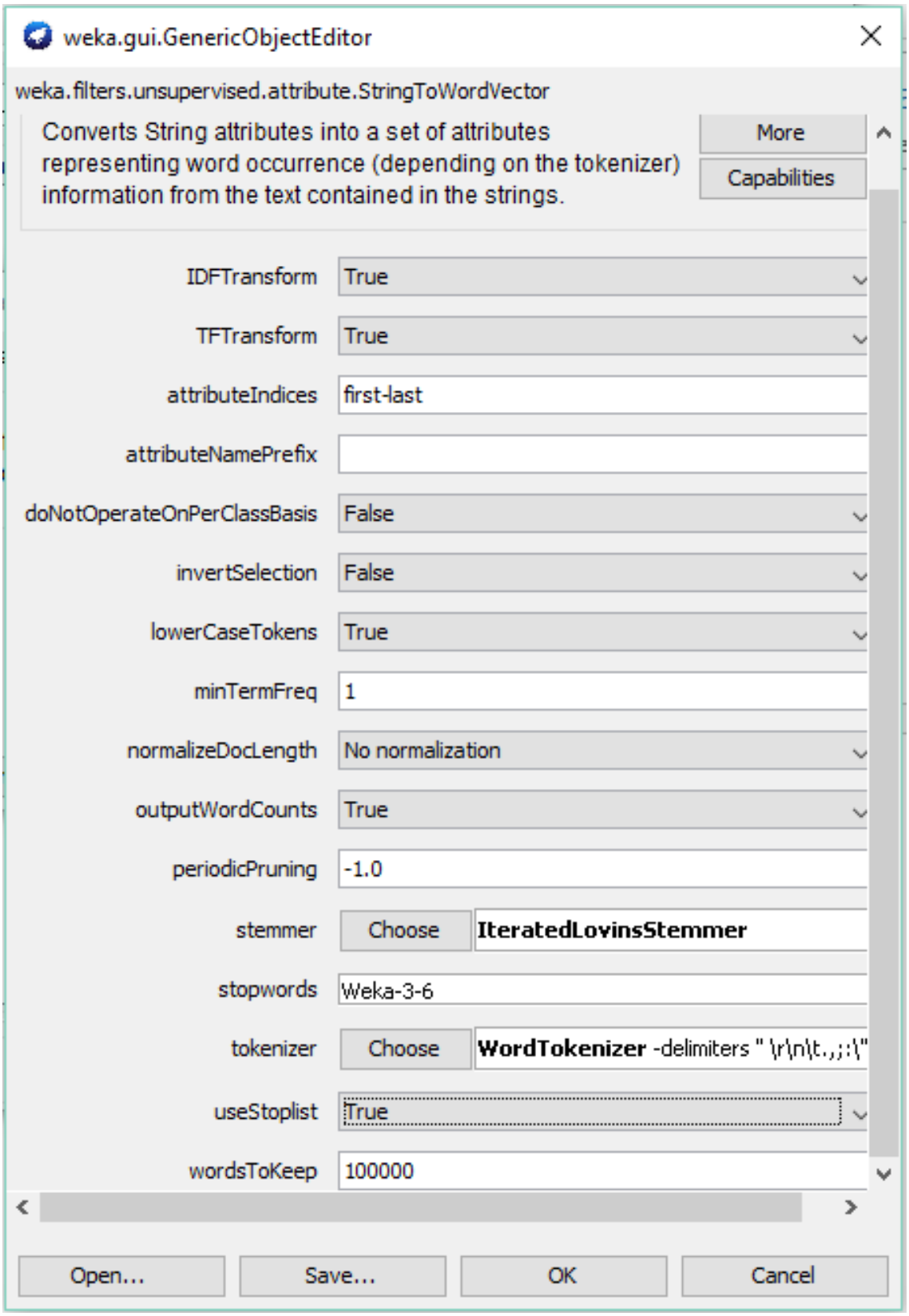

Figure 21: The stringToWord filter and the various options available to process the text.

3.2.3 Training and testing the model:

Based on the classification accuracy of the above 5 models, the top three were selected for training and testing These models were trained and then tested on the dataset shown in Table 7. The highest performing model namely SMO trained on Set 4 (SMO-4) was chosen as the model to be used on the unknown dataset. The model was further improved by adding a randomly generated set, to improve the classification of abstracts. A random number generating script was used to randomly select 10,000 numbers between 10000000 and 25000000. The numbers thus obtained were used as PMIDs to download the respective abstracts. These abstracts were processed and converted to the attribute relationship file format. The 10,000 abstracts were tested using the SMO-4 model. The abstracts that were classified as RNAi by SMO-4 were eliminated. The remaining abstracts formed the random negative dataset. This step ensures that the random negative set is free of positive RNAi instances. The randomly generated dataset was included in the dataset 5. The dataset shown in Table 7 was used to evaluate a new model using the filtered classifier (SMO/StringToWordVector) and named as SMO-5. The performance of SMO seemed to be better and consistent and was chosen as the model of choice for further analysis.

3.2.4 Generation of the screening dataset:

The abstracts for the years 1975 – 2015 was downloaded from the MEDLINE database. The abstracts were downloaded and converted to individual text files retaining just the PMID, title and abstract information using the procedure as mentioned earlier. The text files were grouped by year and then converted to the attribute relationship file format using the WEKA TextDirectoryLoader class. The individual .arff weka input files

were updated to reflect the classes that were used to generate the classification model (SMO-5), namely RNAi and Non_RNAi.

3.2.5 Extraction of RNAi relevant abstracts:

The weka arff files containing the abstracts for each year from 2001 to 2015 was classified using the SMO-5 classification model on the Bigred2, a Cray XE6/XK7 supercomputer with a hybrid architecture comprising of 1020 computing nodes. A total of 10.5 million abstracts were processed to be classified as RNAi or Non_RNAi. The resultant file containing the PMID's along with the classification as RNAi or Non_RNAi was further processed to extract the PMIDs of abstracts classified as RNAi. The abstracts for these PMID's were retrieved and converted to XML format retaining the PMID, article title and abstract using the methods previously described.

3.2.6 Creation of dictionary for entity recognition:

A perl module was created to house the dictionaries for gene names and cell line names. The list of gene names along with their aliases was downloaded from HGNC (HUGO Gene Nomenclature Committee) [146] and the list of cell lines names along with their aliases was downloaded from cellosaurus [147]. These list were further processed to form the final dictionary with cell line names and gene names normalized to their official names/symbols. These dictionaries are very comprehensive with the Gene dictionary containing 161863 entries and the cell line dictionary containing 73370 entries.

3.2.7 Entity tagging and cell-gene information extraction:

The abstracts that were classified as RNAi were further processed and the gene and cell line mentions were tagged with the normalized name of the cell line or gene name using the dictionary that was created as mentioned above. Once tagged the abstracts

were further processed to extract the cell line name and gene names. These were stored in a table format to preserve the genes studied in a given cell line within a given abstract.

3.2.8 Validation of the essential genes:

The extracted genes were ranked in descending order of number of studies associated. The genes that were studied on an average of 100 or more times were extracted and the cell lines in which these genes were studied on average of 20 or more times were extracted as well. In addition the top 20 most studied genes, the median 20 genes and the bottom 20 genes were extracted. The correctness of the extracted cell gene associations was verified by selecting the relevant PMIDs and manually scanning for the presence of the cell and gene information that was extracted. The top genes predicted to be essential for cell survival was queried against the network of cancer genes[148] to identify their relevance to cancer and were also queried against the Therapeutics Target Database [149] to identify if they were drug targets. The genes were also queried against the DPSC database [150] at a threshold p-value of < 0.05 to check for them being reported as essential genes.

Finally a network of the essential genes was built to check for interaction within themselves and with other genes using the GeneMANIA software[151].

3.3 Results:

3.3.1 Identification of siRNA relevant abstracts and corpus creation:

From the approximately 2000 abstracts that were manually screened 221 belonged to the RNAi class and 1644 belonged to the Non_RNAi class, which included abstracts from drug, drug-drug or the not applicable class as described in the methods section. The

average inter classification agreement among individuals who manually read the abstracts was 0.75.

Since these abstracts were initially downloaded based on the specific cell lines prior to the manual scan, there were duplicate abstracts among the cell lines, which were eliminated from the dataset. The above mentioned datasets formed the text corpus to be used for RNAi text classification. This dataset was further divided into training and testing data for evaluating and training the models for RNAi text classification.

3.3.2 Evaluation of the classifiers:

Evaluation is key to identifying the best model for a given task. In order to get an estimate of the generalization error each of the classifiers chosen was evaluated using the 10 fold stratified cross validation. In this way the performance of the model can be evaluated and decision made based on the classification accuracy of a given classifier for a given dataset. The classifiers were evaluated and the results as percent correctly classified are as shown in Table 8. The zeroR classifier is used here to determine the baseline performance and as a benchmark for the other classification methods used. The zeroR classifier is the simplest classification method and does not have any predictability power. It simply builds a frequency table of the given data and selects the most frequent value as its prediction. It can be noted from the Table 8 for zeroR that the percentage accurately predicted is the same as the percentage of the class that is most abundantly present. This is used as a bench mark and any value that falls below the accuracy of a zeroR classification would not be considered a good classifier, it would only be as good as a random guess.

From Table 8, it can be observed that the composition and balance between the positive and negative set does affect the accuracy results of some of the classifiers. Overall the J48, NaiveBayes and SMO seemed to be consistent across the various datasets and more immune to the varying changes between the dataset size and composition.

Classifiers	Set 1	Set 2	Set 3	Set 4	Set 5
ZeroR	75	50	75	80	93.41
NaiveBayes	93	89	93.25	92.4	95
KNN	77	74	81	83.2	94.23
J48	95	95	94.5	96.6	98.46
RandomForest	91	95	84.75	82.8	93.41
SMO	94.25	94.5	94.5	96	98.35
OneR	88.75	78	88.75	91	96.09

Table 8: The % accuracy of classification after evaluating each classifier on a given dataset using 10 fold stratified cross validation.

3.3.3 Evaluating the performance of the top 3 models:

The top 3 classifier models with the highest accuracy of prediction for a given dataset was chosen for further analysis to determine the final model to be selected for RNAi text classification. Each of the top 3 performing models evaluated on a given dataset was further trained on the respective datasets that were used for their evaluation in the 10 fold stratified cross validation, following which they were tested on a previously

unseen dataset, namely the test dataset. The performance results from training and testing are as shown in Table 9.

Set 1	Train	Test	Set 2	Train	Test	Set 3	Train	Test
J48	99.5	94.5	**J48**	99	93	**J48**	99.5	94.5
SMO	100	96.25	**RandomForest**	100	97.5	**SMO**	100	94.5
NaiveBayes	98	86.5	**SMO**	100	93	**NaiveBayes**	98.5	84.25
Set 4	**Train**	**Test**	**Set 5**	**Train**	**Test**			
J48	99	92.4	**J48**	99.5	99.2			
SMO	100	93	**SMO**	100	98.5			
NaiveBayes	94.2	89	**oneR**	96.6	97.1			

Table 9: The % accurately classified by the top three models after training and testing. These models were previously evaluated using the 10 fold cross validation.

In addition to the performance measures such as percent correctly classified, precision and recall, using the error rate is a good way of measuring the classifiers performance. Based on the classifiers prediction of whether an abstract belongs to the RNAi or Non_RNAi class, the proportion of error made over the whole dataset can be calculated thus giving the overall performance of a classifier. It can been seen from Table 10 that J48 and SMO have the best performance according to the five error metrics. They have the lowest values for the mean absolute error, root mean squared error, relative absolute error and root relative squared error and the highest value for the kappa statistic making them the models of choice.

It can be noted that J48 and SMO performed the best. Since SMO was consistently better across the various datasets and SVM being a preferred, faster performing and reliable classifier for text classification, it was chosen for further analysis. The various performance metrics for abstracts classified as RNAi are shown in detail for

the classifiers tested on dataset 5 in Table 11 and the classifier errors are shown in Table 10. The J48 and SMO models performed the best with the SMO model being faster in time taken to build the model. In addition the ROC curve (Figure 22) for the SMO-5 proves its efficiency as a very good classification model.

Classifier Error	ZeroR	NaiveBayes	KNN	J48	RandomForest	SMO	OneR
Kappa statistic	0.00	0.66	0.23	0.87	0.00	0.86	0.57
Mean absolute error	0.12	0.05	0.0 : 6	0.02	0.09	0.02	0.04
Root mean squared error	0.25	0.22	0.24	0.12	0.20	0.13	0.20
Relative absolute error	100%	40.55%	47.10%	15.63%	73.11%	13.33%	31.55%
Root relative squared error	100%	90.13%	96.73%	48.74%	79.35%	51.73%	79.59%

Table 10: Classifier errors for the classifier's tested on dataset 5.

Classifiers	Time (sec)	TPR	FPR	Precision	Recall	F-Measure	ROC Area
ZeroR	2.45	0.00	0.00	0.00	0.00	0.00	0.50
NaiveBayes	28.82	0.83	0.04	0.59	0.83	0.69	0.95
KNN	3.22	0.14	0.00	0.90	0.14	0.25	0.57
J48	116.14	0.83	0.00	0.93	0.83	0.88	0.92
RandomForest	70.66	0.00	0.00	0.00	0.00	0.00	0.99
SMO	6.46	0.82	0.01	0.93	0.82	0.87	0.91
OneR	12.94	0.42	0.00	0.98	0.42	0.59	0.71

Table 11: Performance metrics across the various classifiers tested on dataset 5 for abstracts classified as RNAi. [Time in seconds to build the model, True Positive Rate (TPR), False Positive Rate (FPR), Receiver Operator Characteristic (ROC)].

Figure 22: AUC Receiver Operator Characteristics for the SMO-5 model.

3.3.4 Genes essential for cancer cell survival:

A total of 10.5 million abstracts from the years 2001 to 2015 were tested using the SMO_5 model which resulted in 32164 abstracts being classified as RNAi (Table 12). These abstracts spanned over 1467 cancer cell lines and 4373 genes. There was a total of 25891 cell gene associations identified (Table 13), out of which 97% of the associations between a cell line and a gene occurred 5 or less times. Only 2 gene-cell line pairs were studied more than 90 times. Among the 1467 cell lines 88% of them had at least 1 or up to 25 genes studied in a given cell line (Table 14). Among the 4373 genes 96% of them were studied in at least 1 or up to 25 different cell lines (Table 15).

The top 10 cell lines extracted namely, MCF7, MDA-MB-231, HELA, A549, HEPG2, HCT116, LNCAP, HEK293, SGC7901 and SW480 (Figure 23) had on an average 300 or more associated gene studies and represented Breast, Lung, Colon,

Gastric, Liver, Cervical, Prostate and Kidney cancers, which are some of the most common cancers that affect men and women. On analyzing the cell lines and genes extracted from these abstracts, the top 20 genes, namely AKT1, TP53, CDH1, CCND1, VEGFA, BCL2, EGF, CDKN1A, EPHB2, BIRC5, MYC, EGFR, SNAI1, VIM, BAX, IFI27, AHSA1, SRC, JUN and STAT3 had on an average 100 studies or more associated across different cell lines as shown in Figure 24. Among the top 20 genes, 9 of them are known cancer genes that have a role in cellular function as shown in Table 16 [148]. These functions are defined in the biological process branch of the Gene Ontology (GO) levels 5 and 6. Out of the top 20 genes queried against the DPSC database, 15 of the genes were found to be essential among the four cancer types, namely breast, colon, ovarian and pancreas. In addition 11 out of the 20 genes have active drugs that are being studied in clinical trials or being researched as a potential therapeutic target, some of which have been approved. (Table 17, Table 18) [149].

Year	Medline	RNAi
2001	424042	101
2002	435427	180
2003	472745	425
2004	514910	745
2005	575403	1101
2006	620688	1503
2007	652232	1724
2008	701623	1996
2009	742510	2308
2010	801061	2707
2011	862838	3070
2012	931619	3923
2013	978796	4048
2014	1018012	4498
2015	796876	3835
Total	10528782	32164

Table 12: The number of abstracts that were processed per year and the number of abstracts that were identified as relevant to RNA interference studies.

No. of Cell Gene Associations	Frequency
0	0
5	25198
10	461
15	99
20	52
25	25
30	15
35	8
40	4
45	10
50	1
55	1
60	6
65	5
70	1
75	2
80	0
85	1
90	2
95	0
100	0

Table 13 : The number of times a given gene and cell line were studied together.

Genes / Cell Line	Frequency
0	0
25	1291
50	73
100	54
200	30
300	10
400	3
500	1
600	2
700	0
800	1
900	1
1000	0
1100	1

Table 14: Frequency of the number of genes being studied in a given cell line.

Cell Lines/ Gene	Frequency
0	0
25	4209
50	96
100	46
150	10
200	5
250	3
300	1

Table 15: Frequency of the number of cell lines used to study a given gene.

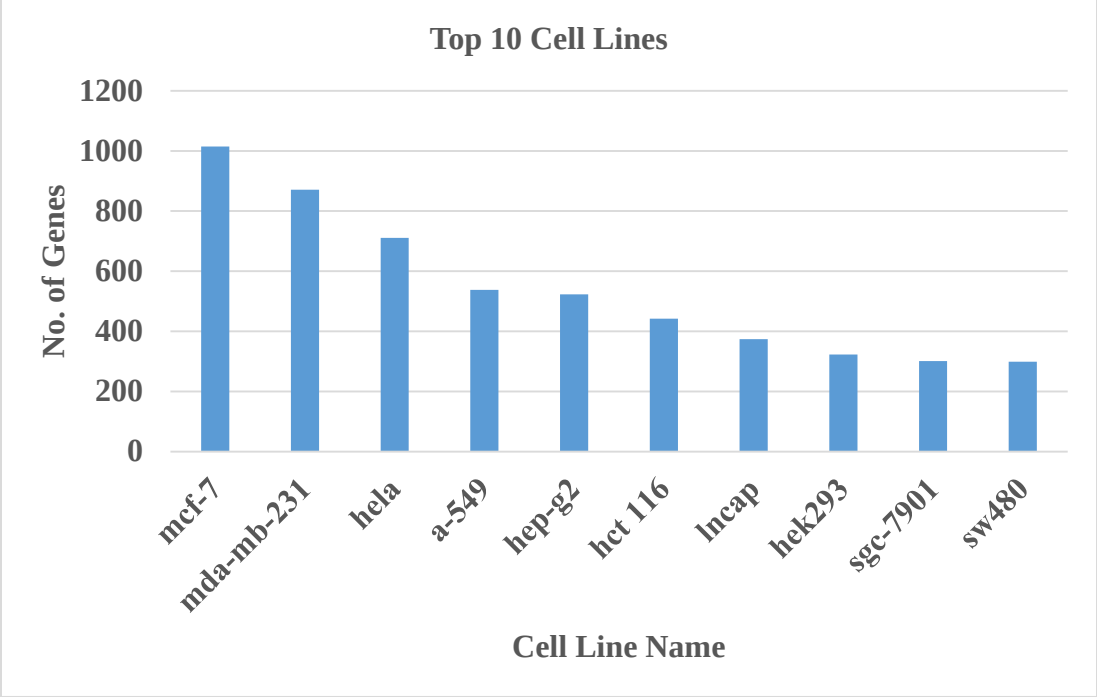

Figure 23: Top 10 most studied cell lines

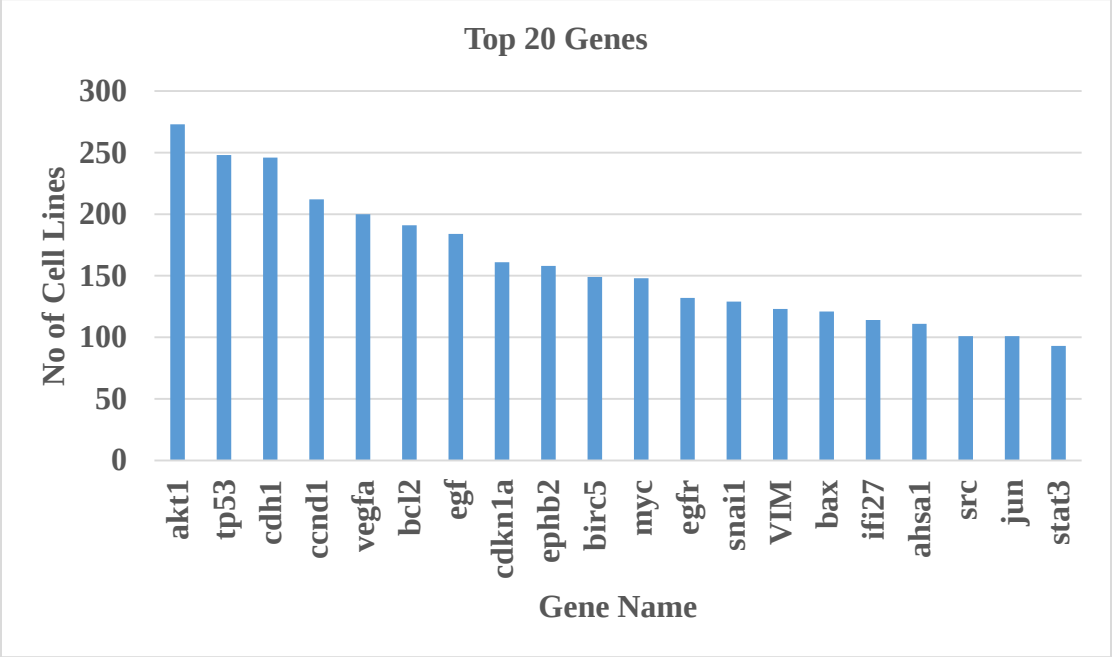

Figure 24: The top 20 genes predicted to be essential for cell survival.

Functional Class	AKT1	TP53	CDH1	CCND1	BCL2	CDKN1A	MYC	EGFR	JUN
Cell cycle	X	X		X	X	X	X	X	
Cell motility and interactions			X					X	
Cell response to stimuli	X	X		X	X	X			
Cellular metabolism	X	X		X		X	X	X	X
Cellular processes	X	X	X	X	X			X	X
Development	X	X		X	X	X		X	X
DNA/RNA metabolism and transcription		X					X		X
Immune system response	X				X	X			
Multicellular activities	X				X				
Regulation of intracellular processes and metabolism	X	X	X	X	X	X	X	X	X
Regulation of transcription	X	X					X		X
Signal transduction	X	X		X	X	X		X	

Table 16: The genes amongst the top 20 that are known to be cancer genes and their roles in the various processes required for cellular function. The presence of the X mark indicates that they are involved in that particular functional process of the cell.

Gene	Drug	Cancer Type	Status	Gene	Drug	Cancer Type	Status
AKT1	Enzastaurin	Non-Hodgkin's lymphoma	Phase 3	AKT1	TCN-P	Acute myeloid leukemia	Phase 1
	Rigosertib	Solid tumours	Phase 3		XL418	Solid tumours	Phase 1
	AZD5363	Cancer	Phase 2	SRC	Dasatinib	Chronic myelogenous leukemia	Approved
	CI-1033	Lymphoma	Phase 2		Herbimycin A	Cancer	Approved
	Enzastaurin	Glioblastoma multiforme	Phase 2		AZD0530	Osteosarcoma; Hematological malignancies; Solid tumours	Phase 2
	GSK2110183	Leukemia	Phase 2		Dasatinib	Solid tumours; Multiple myeloma	Phase 2
	GSK2141795	Colorectal cancer	Phase 2		Bosutinib	Advanced breast cancer	Phase 1/2
	GSK2141795	Lymphoma	Phase 2		KX-02	Brain cancer	Phase 1
	MK-2206	Colon cancer; Rectal cancer	Phase 2		KX01	Acute myeloid leukemia	Phase 1
	RG7440	Solid tumours	Phase 2		KX2-361	Glioblastoma	Phase 1
	RX-0201	Solid tumours	Phase 2		KX2-391	Prostate cancer	Phase 2
	Trametinib + 2141795	Cancer	Phase 2	BCL2	ABT-263	Advanced SCLC; Relapsed or refractory CLL; Lymphoid malignancies	Phase 2
	Triciribine prodrug	Cancer	Phase 1/2		PNT-2258	Cancer	Phase 2
	VQD-002	Haematological malignancies; Leukemia; Non-small-cell lung cancer	Phase 1/2		Beclanorsen	Leukaemia	Phase 1/2
	AR-12	Lymphoma	Phase 1		Obatoclax	Solid tumours	Phase 2
	ARQ 092	Inoperable/unresectable late-stage solid tumors	Phase 1	VEGFA	Ziv-aflibercept	Metastatic colorectal cancer	Approved
	Afuresertib	Multiple myeloma	Phase 1		RG7221	Colorectal cancer	Phase 2
	BAY1125976	Cancer	Phase 1		CVX-241	Cancer	Phase 1
	Enzastaurin	Brain cancer; Central nervous system tumors	Phase 1	MYC	AVI4126	Cancer	Phase 2
	GDC-0068	Solid tumours	Phase 1		DCR-MYC	Hepatocellular carcinoma	Phase 1/2
	GSK690693	Hematologic malignancies	Phase 1		Resten-NG	Cancer	Phase 2
	LY2780301	Cancer	Phase 1	TP53	Cenersen	Acute myeloid leukemia	Phase 2
	LYS-6KAKT1	Cancer	Phase 1	CDH1	Research Target	Gastric Cancer	Research
	MSC2363318A	Solid tumours	Phase 1	BAX	Research Target	Prostate cancer	Research
	Perifosine	Solid tumours	Phase 1	EPHB2	Research Target	Colorectal cancer	Research
	SR13668	Cancer	Phase 1	BIRC5	Research Target	Cancer	Research

Table 17: Genes targeted for treating various cancers along with the respective drugs used.

Gene	Drug	Cancer Type	Status	Drug	Cancer Type	Status
EGFR	BIBW 2992	Non-small-cell lung cancer	Approved	CetuGEX	Cancer	Phase 2
	CP-358774	Non-small cell lung cancer	Approved	Erlotinib	Colon cancer; Glioma; Breast cancer; Head and neck cancer	Phase 2
	Cetuximab	Colorectal cancer	Approved	Gefitinib	Urethral cancer; Bladder cancer; Ovarian cancer; Prostate cancer; Breast cancer	Phase 2
	Erlotinib	Non-small-cell lung cancer	Approved	HER1-VSSP vaccine	Cancer	Phase 2
	IMC-C225	Colorectal cancer	Approved	HM-78136B	Solid tumours	Phase 2
	Lapatinib	Breast cancer	Approved	MEHD-7945A	Solid tumours	Phase 2
	Panitumumab	Colorectal cancer	Approved	MM-121	Breast cancer	Phase 2
	Tyverb/Tykerb	Refractory breast cancer	Approved	Matuzumab	Gastric cancer	Phase 2
	Vandetanib	Solid tumours	Approved	Panitumumab	Locally advanced head and neck cancer	Phase 2
	Icotinib hydrochloride	Non-small-cell lung cancer	Registered	Pazopanib + Tyverb/Tykerb	Inflammatory breast cancer	Phase 2
	AZD9291	Melanoma	Phase 3	Pelitinib	Lymphoma	Phase 2
	CO-1686	Non-small cell lung cancer	Phase 3	SYM-004	Head and neck cancer	Phase 2
	DE-766	Gastric cancer	Phase 3	Sym004	Colorectal cancer	Phase 2
	DE-766	Non-small cell lung cancer	Phase 3	Sym004	Metastatic colorectal cancer	Phase 2
	Dacomitinib	Non-small-cell lung cancer	Phase 3	TT-100	Non-small-cell lung cancer	Phase 2
	Erlotinib	Pancreatic cancer	Phase 3	Tyverb/Tykerb	Head and neck squamous cell cancer	Phase 2
	Gefitinib	Head and neck cancer	Phase 3	VATALANIB	Solid tumours	Phase 2
	HKI-272	Breast cancer	Phase 3	EGF816	Non-small-cell lung cancer	Phase 1/2
	NERATINIB	Lymphoma	Phase 3	EMD 55900	Gliomas	Phase 1/2
	Necitumumab	Colorectal cancer	Phase 3	SN-32793	Non-small-cell lung cancer	Phase 1/2
	Necitumumab	Non-small-cell lung cancer	Phase 3	Varlitinib	Breast cancer	Phase 1/2
	Necitumumab	Squamous non-small cell lung cancer	Phase 3	ABT-806	Cancer	Phase 1
	PF-299804	Glioblastoma	Phase 3	AMG 595	Glioblastoma	Phase 1
	Rindopepimut	Glioblastoma	Phase 3	AST-1306	Cancer	Phase 1
	Tykerb	Refractory metastatic breast cancer; Renal cell cancer	Phase 3	CUDC-101	Solid tumours	Phase 1
	Tyverb/Tykerb	Gastric cancer	Phase 3	Cipatinib	Cancer	Phase 1
	Zalutumumab	Head and neck cancer	Phase 3	HER-2/HER-1 vaccine	Solid tumours	Phase 1
	Indium-111	Solid tumours	Phase 2/3	Anti-HER3/EGFR DAF	Metastatic epithelial tumors	Phase 1
	ABT-414	Glioblastoma	Phase 2	BIBX-1382	Chronic lymphocytic leukaemia	Phase 1
	AP26113	Cancer	Phase 2	IMGN289	Solid tumours	Phase 1
	ASP8273	Non-small-cell lung cancer	Phase 2	JNJ-26483327	Cancer	Phase 1
	BMS-599626	Solid tumours	Phase 2	MM-151	Solid tumours	Phase 1
	Bevacizumab + Erlotinib	Non-small-cell lung cancer	Phase 2	MR1-1	Brain cancer	Phase 1
	CI-1033	Lymphoma	Phase 2	S-222611	Malignant tumor	Phase 1b

Table 18: Genes targeted for treating various cancers along with the respective drugs used.

The top 20 genes studied on an average 20 or more times in a given cell line was extracted and the cell lines were associated to their respective cancer types. The Table 19 shows the number of genes among the top 20 genes associated with a given cancer type. All of the top 20 genes were studied in breast cancer, indicating the complexity of this disease and the network of genes that may play a role in the progression of this cancer. As shown in Figure 22, generated using GeneMANIA [151] these 20 genes are co-expressed and interact with other genes that are required for cell cycle progression and metabolic activities of the cell. Figure 26 shows the top hit gene namely AKT1 and the other genes that are co-expressed or have an interaction with it.

Cancer Type	Genes Studied
Breast	20
Lung	5
Colon	3
Liver	3
Cervix	2
Gastric	1
Ovary	1
Pancreas	1
Prostate	1

Table 19: The table shows the number of genes among the top 20 genes that were studied in a given cancer type. All 20 of them were studied in breast cancer.

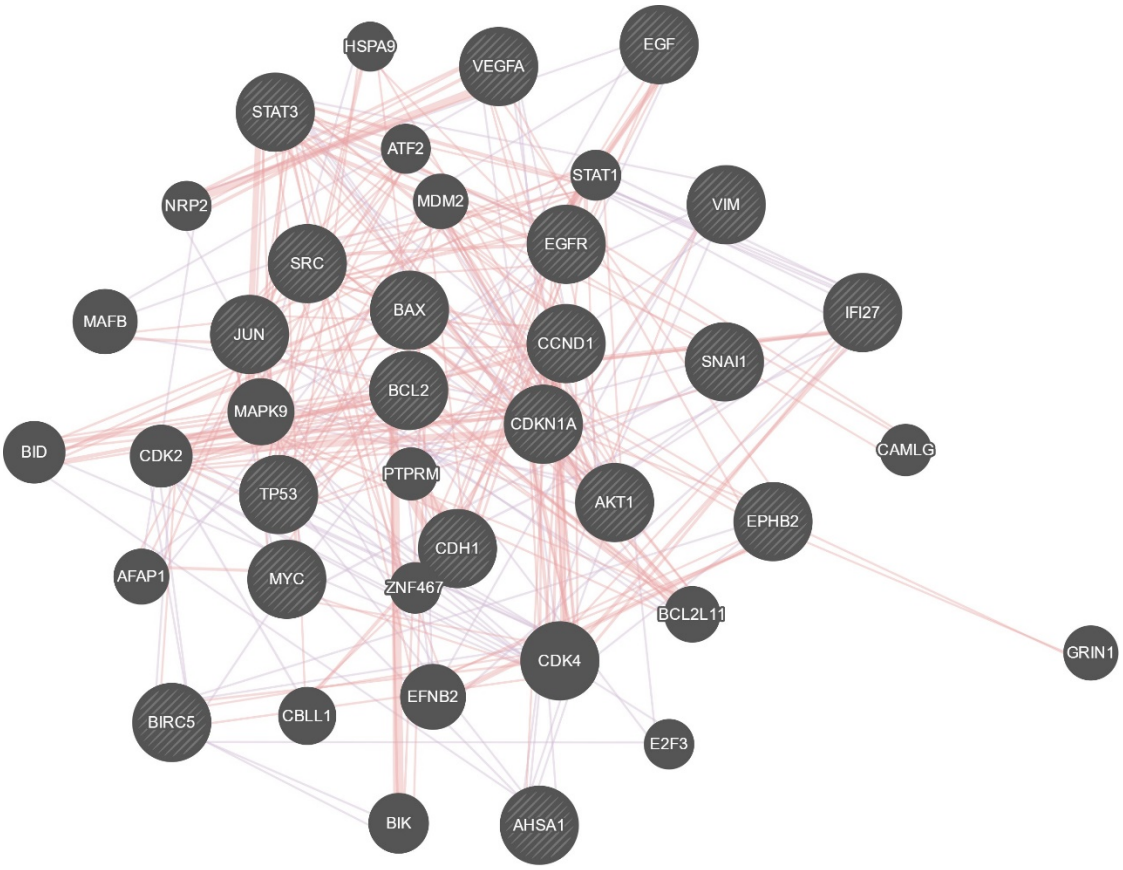

Figure 25: Genes that are co-expressed and interact with the top 20 genes that were identified in different cancer types. The blue lines indicate genes that are co-expression and the pink lines indicate an interaction between the genes.

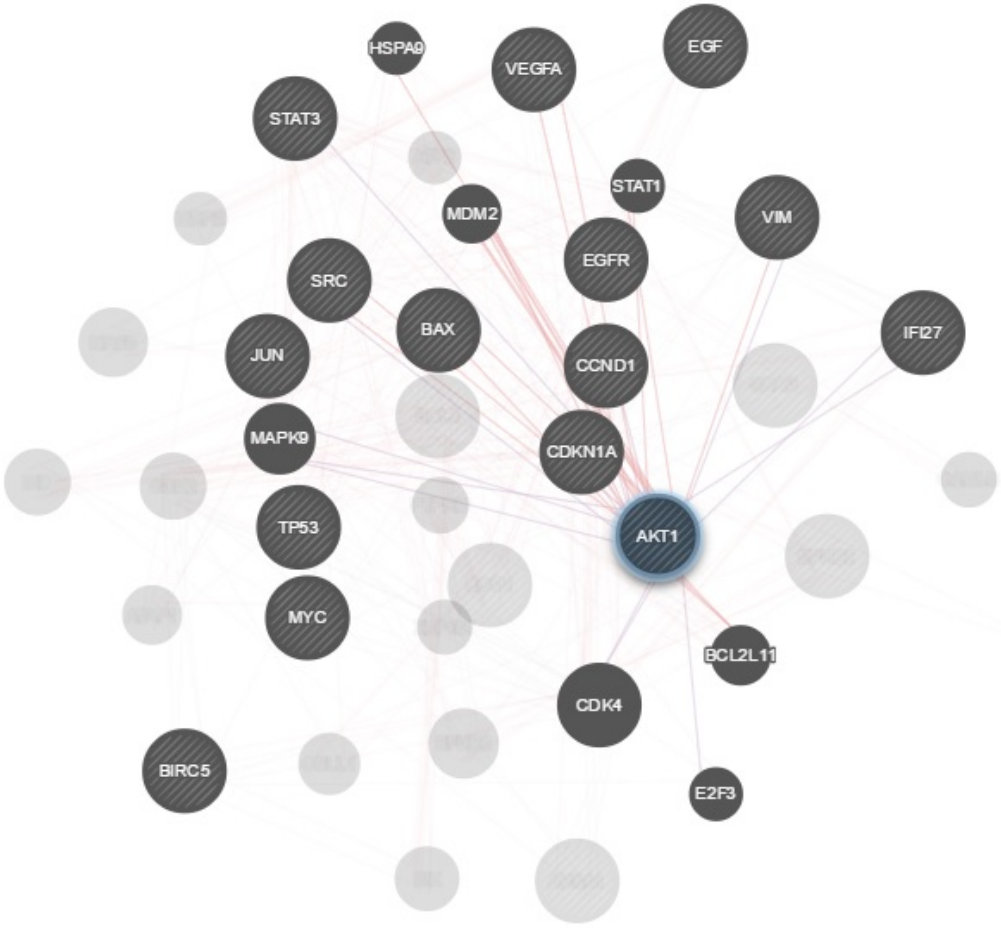

Figure 26: The gene network of the top hit gene AKT1 from our study and the other genes that are co-expressed or interact with it. The blue lines indicate genes that are co-expression and the pink lines indicate an interaction between the genes.

3.3.5 Validation of Genes predicted to be essential:

The top 20 genes, the median 20 genes and bottom 20 genes were extracted and were manually verified from the respective abstracts for their essentiality in cell survival. The top 20 genes were all found to be essential towards cell survival. Among the median 20 there were around four that were false positives and among the bottom 20 there were two that were false positives and four that were genes found to be essential in a non-human species.

3.4 Discussion:

In multicellular organisms, cell death is a critical process by which the damaged cells or those that pose a threat to the organism are destroyed through a tightly regulated process of cell destruction [152-154]. This process is very essential for the overall health and survival of the organism as it gets rid of the cells that may interfere with its normal function [155]. It is clear that a crucial balance between cell proliferation and cell death should be maintained and tipping to one side could lead to a diseased state. Cancer, the uncontrolled proliferation of cells is one of the most complex and challenging disease to treat as it involves many underlying molecular mechanisms and moreover these mechanisms are shared alike by cancerous as well as normal cells. This sharing makes it difficult to therapeutically target cancerous cells without damaging the normal cells. Most of the chemotherapeutic agents available today are relatively nonspecific and cause considerable damage to the surrounding normal cells, leading to severe adverse events. Thus identifying those molecular mechanisms that are essential only to the survival of cancerous cells but not normal cells holds the key to effective cancer treatments. In addition the heterogeneity of cancer calls for a systematic identification of genes that are

essential for the growth of these diverse set of cells and the resultant cancer phenotype which can aid in the identification of potential drug targets.

Our top hit, AKT is a major signaling hub for various downstream substrates and is known to be critical for cell growth and survival [156-158]. It is involved in the progression of many human cancers [159-161]. There are various therapeutic interventions that are currently being targeted towards the inhibition of AKT [162-164]. Perifosine, MK-2206, RX-0201, PBI-05204 and GSK2141795 are some of the potential AKT inhibitors being investigated in several cancers[164].The role of AKT in promoting cell proliferation and survival in hormone responsive MCF-7 breast cancer cells has been previously studied[83]. The investigational drug, MK-2206 has been found to be effective in treating breast cancer[165]. It has been shown that increased levels of AKT in certain cell lines is associated with acquired resistance to antiestrogenic therapy and an inhibition of AKT led to a pronounced growth inhibition of the cell lines[166]. With a wide array of involvement in cell survival and cancer progression, AKT is a potential drug target in cancer therapy, yet finding an optimal way to inhibit AKT has been elusive. Identifying the genes that are essential for cell survival and those that drive tumor resistance are critical pieces of information for developing targeted therapies to prevent the progression of cancer.

p53 has been widely studied and is best known for its tumor suppressing ability through the initiation of apoptosis. The p53 gene once hailed as a potential therapeutic target to halt cancer is met with complexity as many of its functions remain unclear. As shown in Figure 27, its ability to regulate the same cellular processes both positively and negatively makes it hard to predict the outcomes of its activation[167].

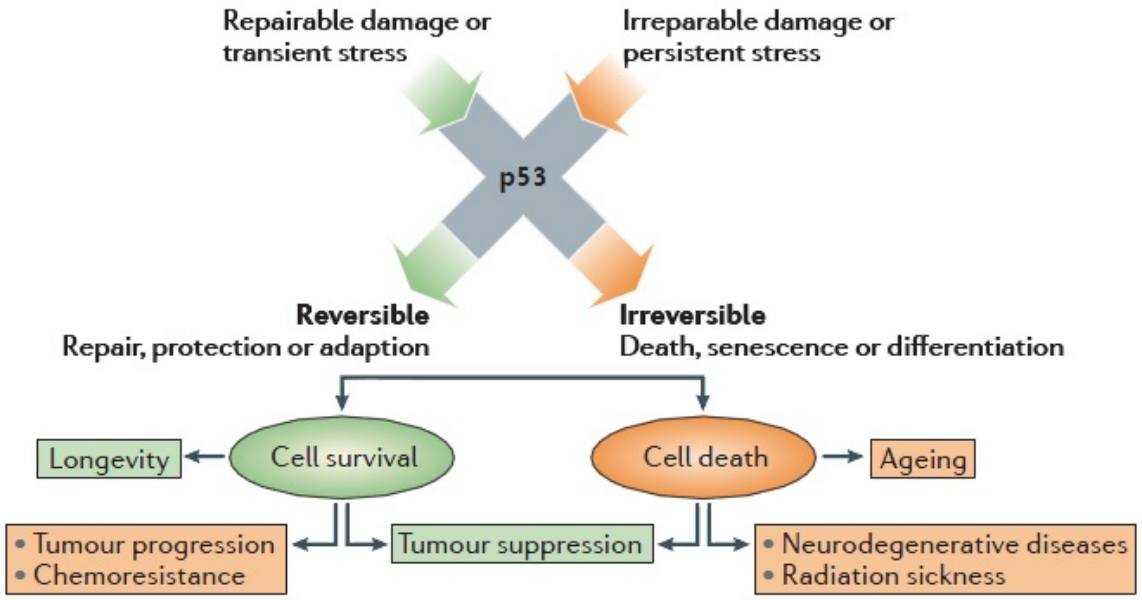

Figure 27: p53 response to cellular stress can lead to cell survival or cell death

Moreover, the median 20 and bottom 20 genes, though not frequently studied may hold the answers to treating cancers that respond poorly to therapy. For example the NFAT gene from our bottom 20 gene list has been found to be involved in many solid tumors and malignancies[168-170]. This and many other genes extracted during this process can be exploited for their role in cancer.

Most of the top essential genes identified and extracted through the large scale scanning of PubMed abstracts are involved in the survival pathways and in various malignancies – AKT1[162-164, 166, 171-175], TP53[167, 176, 177], CDH1[178, 179], CCND1[180], VEGFA[181, 182], BCL2[183, 184], ITK[185], CDKN1A[186], EPHB2[187, 188], BIRC5[189], MYC[190], EGFR[191, 192], VIM[193, 194], BAX[195], AHSA1[196], and SRC[197]. Figure 25 and Figure 26, show the number of genes that are either co-expressed or interact with the essential genes identified in this

study. This suggests that the growth and survival of cancer cells is sustained by a network of genes that come into harmony to fuel the cancer progression. This clearly brings out the importance in not only targeting essential genes, but also those that may be closely involved but not very evident as to their role in fueling cancer. This calls for an extensive mining of data and literature in search of genes that are less known but critical in cellular processes, as these could play a crucial role in the progression of complex disease just as rare SNPs do. The co expression of a gene may not mean that it is or has an influence on the essential gene identified here. But it could mean that in the absence of the targeted essential gene, the co-expressed gene could possible play a role in promoting cell survival, a fact that cannot be ruled out. The complexity of effectively treating cancers unfolds as the network of genes linked to essential genes grow. Identifying the potential interaction that exists between these genes and their individual roles in cell survival or the extent of their influence within a pathway can shed light into developing targeted therapies that destroy cancerous cells but leave the normal cells intact.

This approach in scanning millions of abstracts to identify top genes that are essential for survival is a feat that is not possible by an individual researcher or a group, just because of the sheer volume of literature that needs to be processed and the connections between entities to be made. Using machine learning algorithms, has not only helped narrow down the search and provided information about essential genes in different cancer types but also provided the building blocks to generate a network of interconnected genes and processes, which can be used to generate hypothesis that can be experimentally validated to improve our understanding of what triggers and maintains the growth of cancerous cells. This comprehensive list of genes that are predicted to be

essential in various cancer types can be used as an informational tool by researchers who wish to identify more genes that may be crucial to answer the questions they may have in treating a specific type of cancer. Moreover when the top essential genes do not provide all the answers that a research is seeking, they can expand their targeted gene list by utilize this resource to look up the less frequently studied genes which might prove to be more critical just as rare variants are in finding answers to treating complex diseases. Since genes that are essential are typically involved in biological processes that are critical to a cell, the identification of essential genes in other species through this process can be used as a method of identifying novel targets that would have otherwise gone unnoticed.

There are a few limitations to the method used here. Even though majority of the genes found to be essential are identified and associated with their respective cancer cell lines, there have been instances where a gene or gene alias was the same as that of a commonly used word in English and got tagged incorrectly leading to a false positive. Another limitation of this process is that it cannot identify instances where a gene was specifically found to be not essential for a given cell line.

It is very evident thus far that the efficacy of a therapeutic intervention is multifactorial in nature and in many cases the source of therapeutic disruption could be from an unsuspected source as shown in this dissertation. As in the case of vitamin A which is abundantly present in a range of daily consumed foods to multivitamin pills, its presence can lead to the upregulation of aromatase enzyme, which could possibly lead to an increase in estrogen production there by interfering with the efficacy of aromatase inhibitors. Further the identification of an antiestrogenic property for the urinary

analgesic phenazopyridine previously unknown could be exploited to alter estrogen action and can be further explored for any potential interactions with oral contraceptives that could influence their efficacy. Identifying genes that are essential for a cells survival can be utilized to target them in precisely shutting of resources that help cancer cells to survive. More importantly, the mining of scientific literature helps connect information across non interacting scientific articles as well as uncovering hidden knowledge leading to hypothesis with the potential for clinical applications in treating complex diseases.

References

1. Ingelman-Sundberg M: **The human genome project and novel aspects of cytochrome P450 research**. *Toxicology and applied pharmacology* 2005, **207**(2 Suppl):52-56.
2. Nelson DR, Koymans L, Kamataki T, Stegeman JJ, Feyereisen R, Waxman DJ, Waterman MR, Gotoh O, Coon MJ, Estabrook RW *et al*: **P450 superfamily: update on new sequences, gene mapping, accession numbers and nomenclature**. *Pharmacogenetics* 1996, **6**(1):1-42.
3. Bejjani BA, Lewis RA, Tomey KF, Anderson KL, Dueker DK, Jabak M, Astle WF, Otterud B, Leppert M, Lupski JR: **Mutations in CYP1B1, the gene for cytochrome P4501B1, are the predominant cause of primary congenital glaucoma in Saudi Arabia**. *American journal of human genetics* 1998, **62**(2):325-333.
4. Chavarria-Soley G, Sticht H, Aklillu E, Ingelman-Sundberg M, Pasutto F, Reis A, Rautenstrauss B: **Mutations in CYP1B1 cause primary congenital glaucoma by reduction of either activity or abundance of the enzyme**. *Human mutation* 2008, **29**(9):1147-1153.
5. Choudhary D, Jansson I, Schenkman JB: **CYP1B1, a developmental gene with a potential role in glaucoma therapy**. *Xenobiotica; the fate of foreign compounds in biological systems* 2009, **39**(8):606-615.
6. Stoilov I, Akarsu AN, Sarfarazi M: **Identification of three different truncating mutations in cytochrome P4501B1 (CYP1B1) as the principal cause of primary congenital glaucoma (Buphthalmos) in families linked to the GLC3A locus on chromosome 2p21**. *Human molecular genetics* 1997, **6**(4):641-647.
7. Vasiliou V, Gonzalez FJ: **Role of CYP1B1 in glaucoma**. *Annual review of pharmacology and toxicology* 2008, **48**:333-358.
8. Pullinger CR, Eng C, Salen G, Shefer S, Batta AK, Erickson SK, Verhagen A, Rivera CR, Mulvihill SJ, Malloy MJ *et al*: **Human cholesterol 7alpha-hydroxylase (CYP7A1) deficiency has a hypercholesterolemic phenotype**. *The Journal of clinical investigation* 2002, **110**(1):109-117.
9. Kobayashi K, Takahashi O, Hiratsuka M, Yamaotsu N, Hirono S, Watanabe Y, Oda A: **Evaluation of influence of single nucleotide polymorphisms in cytochrome P450 2B6 on substrate recognition using computational docking and molecular dynamics simulation**. *PloS one* 2014, **9**(5):e96789.
10. **Drug Interactions** [http://medicine.iupui.edu/clinpharm/ddis/]
11. Ince I, Knibbe CA, Danhof M, de Wildt SN: **Developmental changes in the expression and function of cytochrome P450 3A isoforms: evidence from in vitro and in vivo investigations**. *Clinical pharmacokinetics* 2013, **52**(5):333-345.
12. Soldin OP, Chung SH, Mattison DR: **Sex differences in drug disposition**. *Journal of biomedicine & biotechnology* 2011, **2011**:187103.
13. Zanger UM, Schwab M: **Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation**. *Pharmacology & therapeutics* 2013, **138**(1):103-141.

14. **The Human Cytochrome P450 (CYP) Allele Nomenclature Database** [http://www.cypalleles.ki.se/]
15. Cai Y, Konishi T, Han G, Campwala KH, French SW, Wan YJ: **The role of hepatocyte RXR alpha in xenobiotic-sensing nuclear receptor-mediated pathways**. *European journal of pharmaceutical sciences : official journal of the European Federation for Pharmaceutical Sciences* 2002, **15**(1):89-96.
16. Gerbal-Chaloin S, Daujat M, Pascussi JM, Pichard-Garcia L, Vilarem MJ, Maurel P: **Transcriptional regulation of CYP2C9 gene. Role of glucocorticoid receptor and constitutive androstane receptor**. *The Journal of biological chemistry* 2002, **277**(1):209-217.
17. Honkakoski P, Negishi M: **Regulation of cytochrome P450 (CYP) genes by nuclear receptors**. *The Biochemical journal* 2000, **347**(Pt 2):321-337.
18. Jover R, Moya M, Gomez-Lechon MJ: **Transcriptional regulation of cytochrome p450 genes by the nuclear receptor hepatocyte nuclear factor 4-alpha**. *Current drug metabolism* 2009, **10**(5):508-519.
19. Li T, Chen W, Chiang JY: **PXR induces CYP27A1 and regulates cholesterol metabolism in the intestine**. *Journal of lipid research* 2007, **48**(2):373-384.
20. Lim YP, Huang JD: **Interplay of pregnane X receptor with other nuclear receptors on gene regulation**. *Drug metabolism and pharmacokinetics* 2008, **23**(1):14-21.
21. Nebert DW, Dalton TP, Okey AB, Gonzalez FJ: **Role of aryl hydrocarbon receptor-mediated induction of the CYP1 enzymes in environmental toxicity and cancer**. *The Journal of biological chemistry* 2004, **279**(23):23847-23850.
22. Plant N: **The human cytochrome P450 sub-family: transcriptional regulation, inter-individual variation and interaction networks**. *Biochimica et biophysica acta* 2007, **1770**(3):478-488.
23. Wang D, Jiang Z, Shen Z, Wang H, Wang B, Shou W, Zheng H, Chu X, Shi J, Huang W: **Functional evaluation of genetic and environmental regulators of p450 mRNA levels**. *PloS one* 2011, **6**(10):e24900.
24. Waxman DJ: **P450 gene induction by structurally diverse xenochemicals: central role of nuclear receptors CAR, PXR, and PPAR**. *Archives of biochemistry and biophysics* 1999, **369**(1):11-23.
25. Bailey RL, Gahche JJ, Lentino CV, Dwyer JT, Engel JS, Thomas PR, Betz JM, Sempos CT, Picciano MF: **Dietary supplement use in the United States, 2003-2006**. *The Journal of nutrition* 2011, **141**(2):261-266.
26. Nekvindova J, Anzenbacher P: **Interactions of food and dietary supplements with drug metabolising cytochrome P450 enzymes**. *Ceska a Slovenska farmacie : casopis Ceske farmaceuticke spolecnosti a Slovenske farmaceuticke spolecnosti* 2007, **56**(4):165-173.
27. Shetty KD, Dalal SR: **Using information mining of the medical literature to improve drug safety**. *Journal of the American Medical Informatics Association : JAMIA* 2011, **18**(5):668-674.
28. Yang X, Zhang B, Molony C, Chudin E, Hao K, Zhu J, Gaedigk A, Suver C, Zhong H, Leeder JS *et al*: **Systematic genetic and genomic analysis of cytochrome P450 enzyme activities in human liver**. *Genome research* 2010, **20**(8):1020-1036.

29. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome research* 2003, **13**(11):2498-2504.
30. Napoli JL: **Quantification of physiological levels of retinoic acid**. *Methods in enzymology* 1986, **123**:112-124.
31. Fishman J, Goto J: **Mechanism of estrogen biosynthesis. Participation of multiple enzyme sites in placental aromatase hydroxylations**. *The Journal of biological chemistry* 1981, **256**(9):4466-4471.
32. Thompson EA, Jr., Siiteri PK: **The involvement of human placental microsomal cytochrome P-450 in aromatization**. *The Journal of biological chemistry* 1974, **249**(17):5373-5378.
33. Miller WR: **Aromatase activity in breast tissue**. *The Journal of steroid biochemistry and molecular biology* 1991, **39**(5B):783-790.
34. Miller WR, Anderson TJ, Jack WJ: **Relationship between tumour aromatase activity, tumour characteristics and response to therapy**. *The Journal of steroid biochemistry and molecular biology* 1990, **37**(6):1055-1059.
35. Utsumi T, Harada N, Maruta M, Takagi Y: **Presence of alternatively spliced transcripts of aromatase gene in human breast cancer**. *The Journal of clinical endocrinology and metabolism* 1996, **81**(6):2344-2349.
36. Sasano H, Miki Y, Nagasaki S, Suzuki T: **In situ estrogen production and its regulation in human breast carcinoma: from endocrinology to intracrinology**. *Pathology international* 2009, **59**(11):777-789.
37. Bardoni B, Zanaria E, Guioli S, Floridia G, Worley KC, Tonini G, Ferrante E, Chiumello G, McCabe ER, Fraccaro M *et al*: **A dosage sensitive locus at chromosome Xp21 is involved in male to female sex reversal**. *Nature genetics* 1994, **7**(4):497-501.
38. Zanaria E, Muscatelli F, Bardoni B, Strom TM, Guioli S, Guo W, Lalli E, Moser C, Walker AP, McCabe ER *et al*: **An unusual member of the nuclear hormone receptor superfamily responsible for X-linked adrenal hypoplasia congenita**. *Nature* 1994, **372**(6507):635-641.
39. Wang ZJ, Jeffs B, Ito M, Achermann JC, Yu RN, Hales DB, Jameson JL: **Aromatase (Cyp19) expression is up-regulated by targeted disruption of Dax1**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**(14):7988-7993.
40. Zazopoulos E, Lalli E, Stocco DM, Sassone-Corsi P: **DNA binding and transcriptional repression by DAX-1 blocks steroidogenesis**. *Nature* 1997, **390**(6657):311-315.
41. Chae BJ, Lee A, Bae JS, Song BJ, Jung SS: **Expression of nuclear receptor DAX-1 and androgen receptor in human breast cancer**. *Journal of surgical oncology* 2011, **103**(8):768-772.
42. Conde I, Alfaro JM, Fraile B, Ruiz A, Paniagua R, Arenas MI: **DAX-1 expression in human breast cancer: comparison with estrogen receptors ER-alpha, ER-beta and androgen receptor status**. *Breast cancer research : BCR* 2004, **6**(3):R140-148.

43. Chen GG, Zeng Q, Tse GM: **Estrogen and its receptors in cancer**. *Medicinal research reviews* 2008, **28**(6):954-974.
44. Lanzino M, Maris P, Sirianni R, Barone I, Casaburi I, Chimento A, Giordano C, Morelli C, Sisci D, Rizza P *et al*: **DAX-1, as an androgen-target gene, inhibits aromatase expression: a novel mechanism blocking estrogen-dependent breast cancer cell proliferation**. *Cell death & disease* 2013, **4**:e724.
45. Miller WR: **Aromatase inhibitors and breast cancer**. *Minerva endocrinologica* 2006, **31**(1):27-46.
46. Puhalla S, Jankowitz RC, Davidson NE: **Adjuvant endocrine therapy for breast cancer: don't ditch the switch!** *Journal of the National Cancer Institute* 2011, **103**(17):1280-1282.
47. Munos B: **Lessons from 60 years of pharmaceutical innovation**. *Nat Rev Drug Discov* 2009, **8**(12):959-968.
48. Lawerence S: **Drug output slows in 2006**. *Nat Biotechnol* 2007, **25**:1073.
49. Kola I, Landis J: **Can the pharmaceutical industry reduce attrition rates?** *Nat Rev Drug Discov* 2004, **3**(8):711-715.
50. Moreno L, Pearson AD: **How can attrition rates be reduced in cancer drug discovery?** *Expert Opin Drug Discov* 2013, **8**(4):363-368.
51. Ashburn TT, Thor KB: **Drug repositioning: identifying and developing new uses for existing drugs**. *Nat Rev Drug Discov* 2004, **3**(8):673-683.
52. Gelijns AC, Rosenberg N, Moskowitz AJ: **Capturing the unexpected benefits of medical research**. *N Engl J Med* 1998, **339**(10):693-698.
53. Ghofrani HA, Osterloh IH, Grimminger F: **Sildenafil: from angina to erectile dysfunction to pulmonary hypertension and beyond**. *Nat Rev Drug Discov* 2006, **5**(8):689-702.
54. Ladizinski B, Shannon EJ, Sanchez MR, Levis WR: **Thalidomide and analogues: potential for immunomodulation of inflammatory and neoplastic dermatologic disorders**. *J Drugs Dermatol* 2010, **9**(7):814-826.
55. Li YY, Jones SJ: **Drug repositioning for personalized medicine**. *Genome Med* 2012, **4**(3):27.
56. Lamb J: **The Connectivity Map: a new tool for biomedical research**. *Nat Rev Cancer* 2007, **7**(1):54-60.
57. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN *et al*: **The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease**. *Science* 2006, **313**(5795):1929-1935.
58. Brann DW, Dhandapani K, Wakade C, Mahesh VB, Khan MM: **Neurotrophic and neuroprotective actions of estrogen: basic mechanisms and clinical implications**. *Steroids* 2007, **72**(5):381-405.
59. Garcia-Segura LM, Azcoitia I, DonCarlos LL: **Neuroprotection by estradiol**. *Prog Neurobiol* 2001, **63**(1):29-60.
60. Green PS, Simpkins JW: **Neuroprotective effects of estrogens: potential mechanisms of action**. *Int J Dev Neurosci* 2000, **18**(4-5):347-358.
61. Knopp RH, Paramsothy P, Retzlaff BM, Fish B, Walden C, Dowdy A, Tsunehara C, Aikawa K, Cheung MC: **Sex differences in lipoprotein metabolism and**

dietary response: basis in hormonal differences and implications for cardiovascular disease. *Curr Cardiol Rep* 2006, **8**(6):452-459.
62. Baker L, Meldrum KK, Wang M, Sankula R, Vanam R, Raiesdana A, Tsai B, Hile K, Brown JW, Meldrum DR: **The role of estrogen in cardiovascular disease**. *J Surg Res* 2003, **115**(2):325-344.
63. Guetta V, Cannon RO, 3rd: **Cardiovascular effects of estrogen and lipid-lowering therapies in postmenopausal women**. *Circulation* 1996, **93**(10):1928-1937.
64. Lanfranco F, Zirilli L, Baldi M, Pignatti E, Corneli G, Ghigo E, Aimaretti G, Carani C, Rochira V: **A novel mutation in the human aromatase gene: insights on the relationship among serum estradiol, longitudinal growth and bone mineral density in an adult man under estrogen replacement treatment**. *Bone* 2008, **43**(3):628-635.
65. Imai Y, Youn MY, Kondoh S, Nakamura T, Kouzmenko A, Matsumoto T, Takada I, Takaoka K, Kato S: **Estrogens maintain bone mass by regulating expression of genes controlling function and life span in mature osteoclasts**. *Ann N Y Acad Sci* 2009, **1173 Suppl 1**:E31-39.
66. Riggs BL: **The mechanisms of estrogen regulation of bone resorption**. *The Journal of clinical investigation* 2000, **106**(10):1203-1204.
67. Vaananen HK, Harkonen PL: **Estrogen and bone metabolism**. *Maturitas* 1996, **23 Suppl**:S65-69.
68. Ito K, Utsunomiya H, Niikura H, Yaegashi N, Sasano H: **Inhibition of estrogen actions in human gynecological malignancies: new aspects of endocrine therapy for endometrial cancer and ovarian cancer**. *Molecular and cellular endocrinology* 2011, **340**(2):161-167.
69. Tyson JJ, Baumann WT, Chen C, Verdugo A, Tavassoly I, Wang Y, Weiner LM, Clarke R: **Dynamic modelling of oestrogen signalling and cell fate in breast cancer cells**. *Nat Rev Cancer* 2011, **11**(7):523-532.
70. Lim YC, Li L, Desta Z, Zhao Q, Rae JM, Flockhart DA, Skaar TC: **Endoxifen, a secondary metabolite of tamoxifen, and 4-OH-tamoxifen induce similar changes in global gene expression patterns in MCF-7 breast cancer cells**. *J Pharmacol Exp Ther* 2006, **318**(2):503-512.
71. Hsieh CY, Santell RC, Haslam SZ, Helferich WG: **Estrogenic effects of genistein on the growth of estrogen receptor-positive human breast cancer (MCF-7) cells in vitro and in vivo**. *Cancer Res* 1998, **58**(17):3833-3838.
72. Al-Mubarak M, Sacher AG, Ocana A, Vera-Badillo F, Seruga B, Amir E: **Fulvestrant for advanced breast cancer: a meta-analysis**. *Cancer Treat Rev* 2013, **39**(7):753-758.
73. Chia S, Gradishar W: **Fulvestrant: expanding the endocrine treatment options for patients with hormone receptor-positive advanced breast cancer**. *Breast* 2008, **17 Suppl 3**:S16-21.
74. Ciruelos E, Pascual T, Arroyo Vozmediano ML, Blanco M, Manso L, Parrilla L, Munoz C, Vega E, Calderon MJ, Sancho B *et al*: **The therapeutic role of fulvestrant in the management of patients with hormone receptor-positive breast cancer**. *Breast* 2014, **23**(3):201-208.

75. Dodwell D, Pippen J: **Time to response: comparison of fulvestrant and oral endocrine agents**. *Clin Breast Cancer* 2006, **7**(3):244-247.
76. Dodwell D, Vergote I: **A comparison of fulvestrant and the third-generation aromatase inhibitors in the second-line treatment of postmenopausal women with advanced breast cancer**. *Cancer Treat Rev* 2005, **31**(4):274-282.
77. Hancock CM: **Fulvestrant antiestrogen for treatment of breast cancer**. *Clin J Oncol Nurs* 2003, **7**(2):201-202.
78. Jelovac D, Macedo L, Goloubeva OG, Handratta V, Brodie AM: **Additive antitumor effect of aromatase inhibitor letrozole and antiestrogen fulvestrant in a postmenopausal breast cancer model**. *Cancer Res* 2005, **65**(12):5439-5444.
79. Journe F, Body JJ, Leclercq G, Laurent G: **Hormone therapy for breast cancer, with an emphasis on the pure antiestrogen fulvestrant: mode of action, antitumor efficacy and effects on bone health**. *Expert Opin Drug Saf* 2008, **7**(3):241-258.
80. Robertson JF, Come SE, Jones SE, Beex L, Kaufmann M, Makris A, Nortier JW, Possinger K, Rutqvist LE: **Endocrine treatment options for advanced breast cancer--the role of fulvestrant**. *European journal of cancer* 2005, **41**(3):346-356.
81. Valachis A, Mauri D, Polyzos NP, Mavroudis D, Georgoulias V, Casazza G: **Fulvestrant in the treatment of advanced breast cancer: a systematic review and meta-analysis of randomized controlled trials**. *Crit Rev Oncol Hematol* 2010, **73**(3):220-227.
82. Falany JL, Macrina N, Falany CN: **Regulation of MCF-7 breast cancer cell growth by beta-estradiol sulfation**. *Breast Cancer Res Treat* 2002, **74**(2):167-176.
83. Ahmad S, Singh N, Glazer RI: **Role of AKT1 in 17beta-estradiol- and insulin-like growth factor I (IGF-I)-dependent proliferation and prevention of apoptosis in MCF-7 breast carcinoma cells**. *Biochem Pharmacol* 1999, **58**(3):425-430.
84. Castoria G, Barone MV, Di Domenico M, Bilancio A, Ametrano D, Migliaccio A, Auricchio F: **Non-transcriptional action of oestradiol and progestin triggers DNA synthesis**. *EMBO J* 1999, **18**(9):2500-2510.
85. Dupont J, Le Roith D: **Insulin-like growth factor 1 and oestradiol promote cell proliferation of MCF-7 breast cancer cells: new insights into their synergistic effects**. *Mol Pathol* 2001, **54**(3):149-154.
86. Keshamouni VG, Mattingly RR, Reddy KB: **Mechanism of 17-beta-estradiol-induced Erk1/2 activation in breast cancer cells. A role for HER2 AND PKC-delta**. *The Journal of biological chemistry* 2002, **277**(25):22558-22565.
87. Lai A, Sarcevic B, Prall OW, Sutherland RL: **Insulin/insulin-like growth factor-I and estrogen cooperate to stimulate cyclin E-Cdk2 activation and cell Cycle progression in MCF-7 breast cancer cells through differential regulation of cyclin E and p21(WAF1/Cip1)**. *The Journal of biological chemistry* 2001, **276**(28):25823-25833.
88. Horwitz KB, Koseki Y, McGuire WL: **Estrogen control of progesterone receptor in human breast cancer: role of estradiol and antiestrogen**. *Endocrinology* 1978, **103**(5):1742-1751.

89. Shyamala G, Schneider W, Guiot MC: **Estrogen dependent regulation of estrogen receptor gene expression in normal mammary gland and its relationship to estrogenic sensitivity**. *Receptor* 1992, **2**(2):121-128.
90. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, Jensen NH, Kuijer MB, Matos RC, Tran TB *et al*: **Predicting new molecular targets for known drugs**. *Nature* 2009, **462**(7270):175-181.
91. **Phenazopyridine** [https://www.nlm.nih.gov/medlineplus/druginfo/meds/a682231.html]
92. Lee YJ, Gorski J: **Estrogen-induced transcription of the progesterone receptor gene does not parallel estrogen receptor occupancy**. *Proceedings of the National Academy of Sciences of the United States of America* 1996, **93**(26):15180-15184.
93. Lippert TH, Ruoff HJ, Volm M: **Current status of methods to assess cancer drug resistance**. *Int J Med Sci* 2011, **8**(3):245-253.
94. Meads MB, Gatenby RA, Dalton WS: **Environment-mediated drug resistance: a major contributor to minimal residual disease**. *Nat Rev Cancer* 2009, **9**(9):665-674.
95. Swanson DR: **Fish oil, Raynaud's syndrome, and undiscovered public knowledge**. *Perspect Biol Med* 1986, **30**(1):7-18.
96. Swanson DR: **Two medical literatures that are logically but not bibliographically connected**. *J Am Soc Inf Sci* 1987, **38**(4):228-233.
97. Swanson DR: **A second example of mutually isolated medical literatures related by implicit, unnoticed connections**. *J Am Soc Inf Sci* 1989, **40**(6):432-435.
98. Swanson DR: **Online search for logically-related noninteractive medical literatures: a systematic trial-and-error strategy**. *J Am Soc Inf Sci* 1989, **40**(5):356-358.
99. Swanson DR: **Medical literature as a potential source of new knowledge**. *Bull Med Libr Assoc* 1990, **78**(1):29-37.
100. Swanson DR: **Literature-based Resurrection of Neglected Medical Discoveries**. *J Biomed Discov Collab* 2011, **6**:34-47.
101. Del Fiol G, Workman TE, Gorman PN: **Clinical questions raised by clinicians at the point of care: a systematic review**. *JAMA Intern Med* 2014, **174**(5):710-718.
102. Ely JW, Osheroff JA, Gorman PN, Ebell MH, Chambliss ML, Pifer EA, Stavri PZ: **A taxonomy of generic clinical questions: classification study**. *BMJ* 2000, **321**(7258):429-432.
103. Hunter DJ: **Gene-environment interactions in human diseases**. *Nat Rev Genet* 2005, **6**(4):287-298.
104. Naylor S, Chen JY: **Unraveling human complexity and disease with systems biology and personalized medicine**. *Per Med* 2010, **7**(3):275-289.
105. Schwartz DA: **The importance of gene-environment interactions and exposure assessment in understanding human diseases**. *J Expo Sci Environ Epidemiol* 2006, **16**(6):474-476.
106. **Understanding and Performance** [http://nwlink.com/~donclark/performance/understanding.html]

107. Antezana E, Kuiper M, Mironov V: **Biological knowledge management: the emerging role of the Semantic Web technologies**. *Brief Bioinform* 2009, **10**(4):392-407.
108. Butte AJ: **Translational bioinformatics: coming of age**. *Journal of the American Medical Informatics Association : JAMIA* 2008, **15**(6):709-714.
109. Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H, Doherty D, Forsberg K, Gao Y, Kashyap V *et al*: **Advancing translational research with the Semantic Web**. *BMC Bioinformatics* 2007, **8 Suppl 3**:S2.
110. Zhu F, Patumcharoenpol P, Zhang C, Yang Y, Chan J, Meechai A, Vongsangnak W, Shen B: **Biomedical text mining and its applications in cancer research**. *J Biomed Inform* 2013, **46**(2):200-211.
111. Yang Z, Lin H, Li Y: **Exploiting the performance of dictionary-based bio-entity name recognition in biomedical literature**. *Comput Biol Chem* 2008, **32**(4):287-291.
112. Hettne KM, Stierum RH, Schuemie MJ, Hendriksen PJ, Schijvenaars BJ, Mulligen EM, Kleinjans J, Kors JA: **A dictionary to identify small molecules and drugs in free text**. *Bioinformatics* 2009, **25**(22):2983-2991.
113. Narayanaswamy M, Ravikumar KE, Vijay-Shanker K: **A biological named entity recognizer**. *Pac Symp Biocomput* 2003:427-438.
114. Jin Y, McDonald RT, Lerman K, Mandel MA, Carroll S, Liberman MY, Pereira FC, Winters RS, White PS: **Automated recognition of malignancy mentions in biomedical literature**. *BMC Bioinformatics* 2006, **7**:492.
115. Buckley JM, Coopey SB, Sharko J, Polubriaginof F, Drohan B, Belli AK, Kim EM, Garber JE, Smith BL, Gadd MA *et al*: **The feasibility of using natural language processing to extract clinical information from breast pathology reports**. *J Pathol Inform* 2012, **3**:23.
116. Burnside ES, Davis J, Chhatwal J, Alagoz O, Lindstrom MJ, Geller BM, Littenberg B, Shaffer KA, Kahn CE, Jr., Page CD: **Probabilistic computer model developed from clinical data in national mammography database format to classify mammographic findings**. *Radiology* 2009, **251**(3):663-672.
117. Krallinger M, Leitner F, Valencia A: **Analysis of biological processes and diseases using text mining approaches**. *Methods Mol Biol* 2010, **593**:341-382.
118. Strauss JA, Chao CR, Kwan ML, Ahmed SA, Schottinger JE, Quinn VP: **Identifying primary and recurrent cancers using a SAS-based natural language processing algorithm**. *Journal of the American Medical Informatics Association : JAMIA* 2013, **20**(2):349-355.
119. D'Avolio LW, Nguyen TM, Farwell WR, Chen Y, Fitzmeyer F, Harris OM, Fiore LD: **Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (ARC)**. *Journal of the American Medical Informatics Association : JAMIA* 2010, **17**(4):375-382.
120. Hripcsak G, Austin JH, Alderson PO, Friedman C: **Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports**. *Radiology* 2002, **224**(1):157-163.
121. Nguyen AN, Lawley MJ, Hansen DP, Bowman RV, Clarke BE, Duhig EE, Colquist S: **Symbolic rule-based classification of lung cancer stages from free-**

text pathology reports. *Journal of the American Medical Informatics Association : JAMIA* 2010, **17**(4):440-445.
122. Zhao D, Weng C: **Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction**. *J Biomed Inform* 2011, **44**(5):859-868.
123. Chun HW, Tsuruoka Y, Kim JD, Shiba R, Nagata N, Hishiki T, Tsujii J: **Automatic recognition of topic-classified relations between prostate cancer and genes using MEDLINE abstracts**. *BMC Bioinformatics* 2006, **7 Suppl 3**:S4.
124. Deng X, Geng H, Bastola DR, Ali HH: **Link test--A statistical method for finding prostate cancer biomarkers**. *Comput Biol Chem* 2006, **30**(6):425-433.
125. Heintzelman NH, Taylor RJ, Simonsen L, Lustig R, Anderko D, Haythornthwaite JA, Childs LC, Bova GS: **Longitudinal analysis of pain in patients with metastatic prostate cancer using natural language processing of medical record text**. *Journal of the American Medical Informatics Association : JAMIA* 2013, **20**(5):898-905.
126. Tate AR, Martin AG, Ali A, Cassell JA: **Using free text information to explore how and when GPs code a diagnosis of ovarian cancer: an observational study using primary care records of patients with ovarian cancer**. *BMJ Open* 2011, **1**(1):e000025.
127. Frank E, Hall M, Trigg L, Holmes G, Witten IH: **Data mining in bioinformatics using Weka**. *Bioinformatics* 2004, **20**(15):2479-2481.
128. Fire A: **RNA-triggered gene silencing**. *Trends Genet* 1999, **15**(9):358-363.
129. Hammond SM, Caudy AA, Hannon GJ: **Post-transcriptional gene silencing by double-stranded RNA**. *Nat Rev Genet* 2001, **2**(2):110-119.
130. Manoharan M: **RNA interference and chemically modified small interfering RNAs**. *Curr Opin Chem Biol* 2004, **8**(6):570-579.
131. Sharp PA: **RNA interference--2001**. *Genes Dev* 2001, **15**(5):485-490.
132. Tuschl T: **RNA interference and small interfering RNAs**. *Chembiochem* 2001, **2**(4):239-245.
133. Almeida R, Allshire RC: **RNA silencing and genome regulation**. *Trends Cell Biol* 2005, **15**(5):251-258.
134. Ding SW, Voinnet O: **Antiviral immunity directed by small RNAs**. *Cell* 2007, **130**(3):413-426.
135. Li H, Li WX, Ding SW: **Induction and suppression of RNA silencing by an animal virus**. *Science* 2002, **296**(5571):1319-1321.
136. Obbard DJ, Gordon KH, Buck AH, Jiggins FM: **The evolution of RNAi as a defence against viruses and transposable elements**. *Philos Trans R Soc Lond B Biol Sci* 2009, **364**(1513):99-115.
137. Bumcrot D, Manoharan M, Koteliansky V, Sah DW: **RNAi therapeutics: a potential new class of pharmaceutical drugs**. *Nat Chem Biol* 2006, **2**(12):711-719.
138. Caplen NJ, Parrish S, Imani F, Fire A, Morgan RA: **Specific inhibition of gene expression by small double-stranded RNAs in invertebrate and vertebrate systems**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**(17):9742-9747.

139. Elbashir SM, Harborth J, Lendeckel W, Yalcin A, Weber K, Tuschl T: **Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells**. *Nature* 2001, **411**(6836):494-498.
140. Lewis DL, Hagstrom JE, Loomis AG, Wolff JA, Herweijer H: **Efficient delivery of siRNA for inhibition of gene expression in postnatal mice**. *Nature genetics* 2002, **32**(1):107-108.
141. Juhas M, Eberl L, Glass JI: **Essence of life: essential genes of minimal genomes**. *Trends Cell Biol* 2011, **21**(10):562-568.
142. Cancer Genome Atlas N: **Comprehensive molecular portraits of human breast tumours**. *Nature* 2012, **490**(7418):61-70.
143. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z *et al*: **Supervised risk predictor of breast cancer based on intrinsic subtypes**. *J Clin Oncol* 2009, **27**(8):1160-1167.
144. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA *et al*: **Molecular portraits of human breast tumours**. *Nature* 2000, **406**(6797):747-752.
145. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS *et al*: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**(19):10869-10874.
146. **HUGO Gene Nomenclature Committee** [http://www.genenames.org/]
147. **The Cellosaurus: a cell line knowledge resource** [http://web.expasy.org/cellosaurus/]
148. **Network of Cancer Genes** [http://ncg.kcl.ac.uk/]
149. **Therapeutic Target Database** [http://bidd.nus.edu.sg/group/cjttd/]
150. **Donnelly - Princess Margaret Screening Centre** [http://dpsc.ccbr.utoronto.ca/cancer/]
151. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT *et al*: **The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function**. *Nucleic acids research* 2010, **38**(Web Server issue):W214-220.
152. Duprez L, Wirawan E, Vanden Berghe T, Vandenabeele P: **Major cell death pathways at a glance**. *Microbes Infect* 2009, **11**(13):1050-1062.
153. Fulda S, Gorman AM, Hori O, Samali A: **Cellular stress responses: cell survival and cell death**. *Int J Cell Biol* 2010, **2010**:214074.
154. Hotchkiss RS, Strasser A, McDunn JE, Swanson PE: **Cell death**. *N Engl J Med* 2009, **361**(16):1570-1583.
155. Vicencio JM, Galluzzi L, Tajeddine N, Ortiz C, Criollo A, Tasdemir E, Morselli E, Ben Younes A, Maiuri MC, Lavandero S *et al*: **Senescence, apoptosis or autophagy? When a damaged cell must decide its path--a mini-review**. *Gerontology* 2008, **54**(2):92-99.
156. Datta SR, Brunet A, Greenberg ME: **Cellular survival: a play in three Akts**. *Genes Dev* 1999, **13**(22):2905-2927.
157. Song G, Ouyang G, Bao S: **The activation of Akt/PKB signaling pathway and cell survival**. *J Cell Mol Med* 2005, **9**(1):59-71.

158. Manning BD, Cantley LC: **AKT/PKB signaling: navigating downstream**. *Cell* 2007, **129**(7):1261-1274.
159. Fresno Vara JA, Casado E, de Castro J, Cejas P, Belda-Iniesta C, Gonzalez-Baron M: **PI3K/Akt signalling pathway and cancer**. *Cancer Treat Rev* 2004, **30**(2):193-204.
160. Vivanco I, Sawyers CL: **The phosphatidylinositol 3-Kinase AKT pathway in human cancer**. *Nat Rev Cancer* 2002, **2**(7):489-501.
161. Altomare DA, Testa JR: **Perturbations of the AKT signaling pathway in human cancer**. *Oncogene* 2005, **24**(50):7455-7464.
162. Alexander W: **Inhibiting the akt pathway in cancer treatment: three leading candidates**. *P T* 2011, **36**(4):225-227.
163. LoPiccolo J, Blumenthal GM, Bernstein WB, Dennis PA: **Targeting the PI3K/Akt/mTOR pathway: effective combinations and clinical considerations**. *Drug Resist Updat* 2008, **11**(1-2):32-50.
164. Pal SK, Reckamp K, Yu H, Figlin RA: **Akt inhibitors in clinical development for the treatment of cancer**. *Expert Opin Investig Drugs* 2010, **19**(11):1355-1366.
165. Ma CX, Sanchez C, Gao F, Crowder R, Naughton M, Pluard T, Creekmore A, Guo Z, Hoog J, Lockhart AC *et al*: **A Phase I Study of the AKT Inhibitor MK-2206 in Combination with Hormonal Therapy in Postmenopausal Women with Estrogen Receptor-Positive Metastatic Breast Cancer**. *Clin Cancer Res* 2016, **22**(11):2650-2658.
166. Frogne T, Jepsen JS, Larsen SS, Fog CK, Brockdorff BL, Lykkesfeldt AE: **Antiestrogen-resistant human breast cancer cells require activated protein kinase B/Akt for growth**. *Endocr Relat Cancer* 2005, **12**(3):599-614.
167. Kruiswijk F, Labuschagne CF, Vousden KH: **p53 in survival, death and metabolic health: a lifeguard with a licence to kill**. *Nat Rev Mol Cell Biol* 2015, **16**(7):393-405.
168. Mancini M, Toker A: **NFAT proteins: emerging roles in cancer progression**. *Nat Rev Cancer* 2009, **9**(11):810-820.
169. Muller MR, Rao A: **NFAT, immunity and cancer: a transcription factor comes of age**. *Nat Rev Immunol* 2010, **10**(9):645-656.
170. Pan MG, Xiong Y, Chen F: **NFAT gene family in inflammation and cancer**. *Curr Mol Med* 2013, **13**(4):543-554.
171. Chen L, Kang QH, Chen Y, Zhang YH, Li Q, Xie SQ, Wang CJ: **Distinct roles of Akt1 in regulating proliferation, migration and invasion in HepG2 and HCT 116 cells**. *Oncol Rep* 2014, **31**(2):737-744.
172. Irie HY, Pearline RV, Grueneberg D, Hsia M, Ravichandran P, Kothari N, Natesan S, Brugge JS: **Distinct roles of Akt1 and Akt2 in regulating cell migration and epithelial-mesenchymal transition**. *J Cell Biol* 2005, **171**(6):1023-1034.
173. Ju X, Katiyar S, Wang C, Liu M, Jiao X, Li S, Zhou J, Turner J, Lisanti MP, Russell RG *et al*: **Akt1 governs breast cancer progression in vivo**. *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**(18):7438-7443.

174. Roy HK, Olusola BF, Clemens DL, Karolski WJ, Ratashak A, Lynch HT, Smyrk TC: **AKT proto-oncogene overexpression is an early event during sporadic colon carcinogenesis**. *Carcinogenesis* 2002, **23**(1):201-205.
175. Testa JR, Tsichlis PN: **AKT signaling in normal and malignant cells**. *Oncogene* 2005, **24**(50):7391-7393.
176. Lukin DJ, Carvajal LA, Liu WJ, Resnick-Silverman L, Manfredi JJ: **p53 Promotes cell survival due to the reversibility of its cell-cycle checkpoints**. *Mol Cancer Res* 2015, **13**(1):16-28.
177. Singh B, Reddy PG, Goberdhan A, Walsh C, Dao S, Ngai I, Chou TC, P OC, Levine AJ, Rao PH *et al*: **p53 regulates cell survival by inhibiting PIK3CA in squamous cell carcinomas**. *Genes Dev* 2002, **16**(8):984-993.
178. Graziano F, Humar B, Guilford P: **The role of the E-cadherin gene (CDH1) in diffuse gastric cancer susceptibility: from the laboratory to clinical practice**. *Ann Oncol* 2003, **14**(12):1705-1713.
179. Pecina-Slaus N: **Tumor suppressor gene E-cadherin and its role in normal and malignant cells**. *Cancer Cell Int* 2003, **3**(1):17.
180. Fasanaro P, Magenta A, Zaccagnini G, Cicchillitti L, Fucile S, Eusebi F, Biglioli P, Capogrossi MC, Martelli F: **Cyclin D1 degradation enhances endothelial cell survival upon oxidative stress**. *FASEB J* 2006, **20**(8):1242-1244.
181. Byrne AM, Bouchier-Hayes DJ, Harmey JH: **Angiogenic and cell survival functions of vascular endothelial growth factor (VEGF)**. *J Cell Mol Med* 2005, **9**(4):777-794.
182. Carmeliet P: **VEGF as a key mediator of angiogenesis in cancer**. *Oncology* 2005, **69 Suppl 3**:4-10.
183. Adams JM, Cory S: **The Bcl-2 protein family: arbiters of cell survival**. *Science* 1998, **281**(5381):1322-1326.
184. Cory S, Huang DC, Adams JM: **The Bcl-2 family: roles in cell survival and oncogenesis**. *Oncogene* 2003, **22**(53):8590-8607.
185. Sagiv-Barfi I, Kohrt HE, Czerwinski DK, Ng PP, Chang BY, Levy R: **Therapeutic antitumor immunity by checkpoint blockade is enhanced by ibrutinib, an inhibitor of both BTK and ITK**. *Proceedings of the National Academy of Sciences of the United States of America* 2015, **112**(9):E966-972.
186. Price JG, Idoyaga J, Salmon H, Hogstad B, Bigarella CL, Ghaffari S, Leboeuf M, Merad M: **CDKN1A regulates Langerhans cell survival and promotes Treg cell generation upon exposure to ionizing irradiation**. *Nat Immunol* 2015, **16**(10):1060-1068.
187. Gao Q, Liu W, Cai J, Li M, Gao Y, Lin W, Li Z: **EphB2 promotes cervical cancer progression by inducing epithelial-mesenchymal transition**. *Hum Pathol* 2014, **45**(2):372-381.
188. Jubb AM, Zhong F, Bheddah S, Grabsch HI, Frantz GD, Mueller W, Kavi V, Quirke P, Polakis P, Koeppen H: **EphB2 is a prognostic factor in colorectal cancer**. *Clin Cancer Res* 2005, **11**(14):5181-5187.
189. Lamers F, Schild L, Koster J, Versteeg R, Caron HN, Molenaar JJ: **Targeted BIRC5 silencing using YM155 causes cell death in neuroblastoma cells with low ABCB1 expression**. *European journal of cancer* 2012, **48**(5):763-771.

190. Conacci-Sorrell M, Ngouenet C, Anderson S, Brabletz T, Eisenman RN: **Stress-induced cleavage of Myc promotes cancer cell survival**. *Genes Dev* 2014, **28**(7):689-707.
191. Ha SY, Choi SJ, Cho JH, Choi HJ, Lee J, Jung K, Irwin D, Liu X, Lira ME, Mao M *et al*: **Lung cancer in never-smoker Asian females is driven by oncogenic mutations, most often involving EGFR**. *Oncotarget* 2015, **6**(7):5465-5474.
192. Normanno N, De Luca A, Bianco C, Strizzi L, Mancino M, Maiello MR, Carotenuto A, De Feo G, Caponigro F, Salomon DS: **Epidermal growth factor receptor (EGFR) signaling in cancer**. *Gene* 2006, **366**(1):2-16.
193. Costa VL, Henrique R, Danielsen SA, Duarte-Pereira S, Eknaes M, Skotheim RI, Rodrigues A, Magalhaes JS, Oliveira J, Lothe RA *et al*: **Three epigenetic biomarkers, GDF15, TMEFF2, and VIM, accurately predict bladder cancer from DNA-based analyses of urine samples**. *Clin Cancer Res* 2010, **16**(23):5842-5851.
194. Shirahata A, Hibi K: **Serum vimentin methylation as a potential marker for colorectal cancer**. *Anticancer Res* 2014, **34**(8):4121-4125.
195. Ouyang H, Furukawa T, Abe T, Kato Y, Horii A: **The BAX gene, the promoter of apoptosis, is mutated in genetically unstable cancers of the colorectum, stomach, and endometrium**. *Clin Cancer Res* 1998, **4**(4):1071-1074.
196. Shao J, Wang L, Zhong C, Qi R, Li Y: **AHSA1 regulates proliferation, apoptosis, migration, and invasion of osteosarcoma**. *Biomed Pharmacother* 2016, **77**:45-51.
197. Sen B, Johnson FM: **Regulation of SRC family kinases in human cancers**. *J Signal Transduct* 2011, **2011**:865819.

CURRICULUM VITAE

SANTOSH PHILIPS

EDUCATION

PhD Bioinformatics — Aug 2016
Indiana University, Indiana, USA

MS Bioinformatics — Dec 2004
Indiana University, Indiana, USA

BS Pharmacy — Nov 1999
Bangalore University, Karnataka, India

WORK EXPERIENCE

Bioinformatics Scientist — Jul 16 - present
Hematology/Oncology, Indiana University School of Medicine

Doctoral Research Student — Aug 13 – Jun 16
Center for Computational Biology and Bioinformatics**.** Indiana University

Research Analyst — Feb 06 – Dec 13
Clinical Pharmacology, Indiana University School of Medicine

Research Technician — Oct 04 – Jan 06
Clinical Pharmacology, Indiana University School of Medicine

Programmer — Feb 02 – Dec 02
Electrical and Computer Engineering, IUPUI

PEER REVIEWED PUBLICATIONS

1. Saab R, Zouk AN, Mastouri R, Skaar TC, **Philips S**, Kreutz RP. AMPD1 polymorphism and response to regadenoson. Pharmacogenomics. 2015 Nov; 16(16):1807-15

2. Oesterreich, Steffi, Henry, N Lynn, Kidwell, Kelley, Von Poznak, Catherine H, Skaar, Todd C, Dantzer, Jessica, Li, Lang, Thomas Hangartner, M Peacock, Nguyen, Anne T, Rae, James M, Desta, Zeruesenay, **Philips, Santosh**, Carpenter, Janet S, Storniolo, Anna M, Stearns, Vered, Hayes, Daniel F, Flockhart, David A. Associations between genetic variants and the effect of Letrozole and Exemestane on bone mass and bone turnover. Breast Cancer Res Treat.2015 Nov;154(2):263-73

3. Santa-Maria CA, Blackford AL, Nguyen AT, Skaar TC, **Philips S**, Oesterreich S, Rae JM, Desta Z, Robarge JD, Henry NL, Storniolo AM, Hayes DF, Blumenthal RS, Ouyang P, Post WS, Flockhart DA, Stearns V. Association of Variants in Candidate Genes with Lipid Profiles in Women with Early Breast Cancer on Adjuvant Aromatase Inhibitor Therapy. Clin Cancer Res. 2015 Oct; 22(6):1395-402

4. Kimberly Burgess, **Santosh Philips**, Eric Benson, Zeruesenay Desta, Andrea Gaedigk, Matthew Segar, Yunlong Liu, Todd C. Skaar. Age-related changes in microRNA expression and pharmacogenes in human liver. Clinical Pharmacology and Therapeutics. 2015 Aug; 98(2): 205-15

5. **Santosh Philips**, Jing Zhou, Zhigao Li, Todd C. Skaar, Lang Li. A translational bioinformatic approach in identifying and validating an interaction between Vitamin A and CYP19A1. BMC Genomics. 2015 Jun 16(Suppl 7):S17

6. N Lynn Henry, Heang-Ping Chan, Chirayu P Goswami, Lang Li, Todd C Skaar, James M Rae, Zereunesay Desta, Nagi Khouri, Renee Pinsky, Steffi Oesterreich, Chuan Zhou, Lubomir Hadjiiski, **Santosh Philips**, Jason Robarge, Anne T Nguyen, Anna M Storniolo, David A Flockhart, Daniel F Hayes, Mark A Helvie, Vered Stearns. Aromatase Inhibitor-Induced Modulation of Breast Density: Clinical and Genetic Effects. British Journal of Cancer. 2013 Oct;109(9):2331-9

7. Anuradha Ramamoorthy, Yunlong Liu, **Santosh Philips**, Zeruesenay Desta, Hai Lin, Chirayu Goswami, Andrea Gaedigk, Lang Li, David A. Flockhart, and Todd C. Skaar. Regulation of miRNA expression by rifampin in human hepatocytes. Drug Metabolism and Disposition. 2013 Oct;41(10):1763-8

8. N. Lynn Henry, Todd C. Skaar, Jessica Dantzer, Lang Li, Anne T. Nguyen, James M. Rae, Zeruesenay Desta, Steffi Oesterreich, **Santosh Philips**, Janet S. Carpenter, Anna M. Storniolo, Vered Stearns, Daniel F. Hayes, David A. Flockhart. Genetic Associations With Toxicity-related Discontinuation of Aromatase Inhibitor Therapy for Breast Cancer. Breast Cancer Research and Treatment. 2013 Apr;138(3):807-16

9. L. Weng, D. Ziliak, H. K. Im, E. R. Gamazon, **S. Philips**, A. T. Nguyen, Z. Desta, T. C. Skaar, the Consortium on Breast Cancer Pharmacogenomics (COBRA), D. A. Flockhart, and R. S. Huang. Genome-wide discovery of genetic variants affecting tamoxifen sensitivity and their clinical and functional validation. Annals of Oncology. 2013 Jul;24(7):1867-73

10. Wu HY, Karnik S, Subhadarshini A, Wang Z, **Philips S**, Han X, Chiang C, Liu L, Boustani M, Rocha LM, Quinney SK, Flockhart D, Li L. An integrated pharmacokinetics ontology and corpus for text mining. BMC Bioinformatics. 2013 Feb 1;14:35

11. Haas DM, Dantzer J, Lehmann AS, **Philips** S, Skaar TC, McCormick CL, Hebbring SJ, Jung J, Li L. The impact of glucocorticoid polymorphisms on markers of neonatal respiratory disease after antenatal betamethasone administration. Am J Obstet Gynecol. 2013 Mar;208(3):215 e1-6

12. Haas DM, Lehmann AS, Skaar T, **Philips S**, McCormick CL, Beagle K, Hebbring SJ, Dantzer J, Li L, Jung J. The impact of drug metabolizing enzyme polymorphisms on outcomes after antenatal corticosteroid use. Am J Obstet Gynecol. 2012 May; 206(5):447.e17-24

13. **Philips S**, Richter A, Oesterreich S, Rae JM, Flockhart DA, Perumal NB, Skaar TC. Functional characterization of a genetic polymorphism in the promoter of the ESR2 gene. Horm Cancer. 2012 Apr;3(1-2):37-43

14. Hartmaier RJ, Richter AS, Gillihan RM, Sallit JZ, McGuire SE, Wang J, Lee AV, Osborne CK, O'Malley BW, Brown PH, Xu J, Skaar TC, **Philips S**, Rae JM, Azzouz F, Li L, Hayden J, Henry NL, Nguyen AT, Stearns V, Hayes DF, Flockhart DA, Oesterreich S. A SNP in steroid receptor coactivator-1 disrupts a GSK3β phosphorylation site and is associated with altered tamoxifen response in bone. Mol Endocrinol. 2012 Feb; 26(2):220-7

15. **Philips S**, Rae JM, Oesterreich S, Hayes DF, Stearns V, Henry NL, Storniolo AM, Flockhart DA, Skaar TC. Whole genome amplification of DNA for genotyping pharmacogenetics candidate genes. Front Pharmacol. 2012; 3:54

16. Yoon HH, Catalano P, Gibson MK, Skaar TC, **Philips S**, Montgomery EA, Hafez MJ, Powell M, Liu G, Forastiere AA, Benson AB, Kleinberg LR, Murphy KM. Genetic variation in radiation and platinum pathways predicts severe acute radiation toxicity in patients with esophageal adenocarcinoma treated with cisplatin-based preoperative radiochemotherapy: results from the Eastern Cooperative Oncology Group. Cancer Chemother Pharmacol. 2011 Oct; 68(4):863-70

17. Yoon HH, Catalano PJ, Murphy KM, Skaar TC, **Philips S**, Powell M, Montgomery EA, Hafez MJ, Offer SM, Liu G, Meltzer SJ, Wu X, Forastiere AA, Benson AB, Kleinberg LR, Gibson MK. Genetic variation in DNA- repair pathways and response to radiochemotherapy in esophageal adenocarcinoma: a retrospective cohort study of the Eastern Cooperative Oncology Group. BMC Cancer. 2011 May 17; 11:176

18. Haas DM, Daum M, Skaar T, **Philips S**, Miracle D, Renbarger JL. Human breast milk as a source of DNA for amplification. J Clin Pharmacol. 2011 Apr; 51(4):616-9

19. Hayes DF, Skaar TC, Rae JM, Henry NL, Nguyen AT, Stearns V, Li L, **Philips S**, Desta Z, Flockhart DA; Consortium on Breast Cancer Pharmacogenomics (COBRA). Estrogen receptor genotypes, menopausal status, and the effects of tamoxifen on lipid levels: revised and updated results. Clin Pharmacol Ther. 2010 Nov; 88(5):626-9

20. Borges S, Desta Z, Jin Y, Faouzi A, Robarge JD, **Philips S**, Nguyen A, Stearns V, Hayes D, Rae JM, Skaar TC, Flockhart DA, Li L. Composite Functional Genetic and Comedication CYP2D6 Activity Score in Predicting Tamoxifen Drug Exposure Among Breast Cancer Patients. J Clin Pharmacol. 2010 Apr; 50(4):450-8

21. NL Henry, A Nguyen, F Azzouz, J Robarge, L Li, **S Philips**, TC Skaar, AM Storniolo, DA Flockhart, DF Hayes, and V Stearns, for the Consortium on Breast Cancer Pharmacogenomics investigators. Lack of association of estrogen receptor polymorphisms and change in bone mineral density with tamoxifen therapy. British Journal of Cancer. 2010 Jan 19; 102(2):294-300

22. Lynn Henry N, Rae JM, Li L, Azzouz F, Skaar TC, Desta Z, Sikora MJ, **Philips S**, Nguyen AT, Storniolo AM, Hayes DF, Flockhart DA, Stearns V; Consortium on Breast Cancer Pharmacogenomics Investigators. Association between CYP2D6 genotype and tamoxifen-induced hot flashes in a prospective cohort. Breast Cancer Res Treat. 2009 Oct; 117(3):571-5

23. Carpenter JS, Yu M, Wu J, Von Ah D, Milata J, Otte JL, Johns S, Schneider B, Storniolo AM, Salomon R, Desta Z, Cao D, Jin Y, **Philips S**, Skaar TC. Evaluating the role of serotonin in hot flashes after breast cancer using acute tryptophan depletion. Menopause. 2009 Jul-Aug; 16(4):644-52

24. Million Arefayene, **Santosh Philips,** Donghua Cao, Sudharani Mamidipalli, Zeruesenay Desta, David A. Flockhart, David S. Wilkesc, and Todd C.Skaar. Identification of genetic variants in the human indoleamine 2,3-dioxygenase (IDO1) gene, which have altered enzyme activity. Pharmacogenetics and genomics 2009 June; 19(6): 464-76

25. Jin Y, Hayes DF, Li L, Robarge JD, Skaar TC, **Philips S**, Nguyen A, Schott A, Hayden J, Lemler S, Storniolo AM, Flockhart DA, Stearns V. Estrogen receptor genotypes influence hot flash prevalence and composite score before and after tamoxifen therapy.J Clin Oncol. 2008 Dec 20; 26(36):5849-54

26. Bharucha AE, Skaar T, Andrews CN, Camilleri M, **Philips S**, Seide B, Burton D, Baxter K, Zinsmeister AR. Relationship of cytochrome P450 pharmacogenetics to the effects of yohimbine on gastrointestinal transit and catecholamines in healthy subjects.Neurogastroenterol Motil. 2008 Aug; 20(8):891-9

27. Jin Y, Wang YH, Miao J, Li L, Kovacs RJ, Marunde R, Hamman MA, **Philips S**, Hilligoss J, Hall SD. Cytochrome P450 3A5 genotype is associated with verapamil response in healthy subjects. Clin Pharmacol Ther. 2007 Nov; 82(5):579-8