BE 568
Final Project

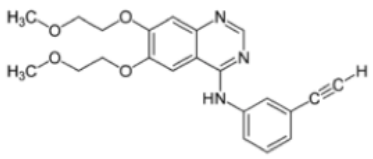
Group Members: Elias Exarchos, Maria Barrios, Suma Cheekati, and Devanshi Patel

Contributions: Contributions: all four members of the team met up to work through the computational process for one drug together. Maria and Elias (M&E) took care of the computational portion of this project. All the WEKA results, their explanations, and their significance were analyzed by M&E. The prognostic gene list was also developed by M&E; after literature analysis, M&E manually added some genes to the list. Suma and Devanshi then performed GSEA and DAVID analysis on the three drugs, researching the enriched pathways and genes in readings and literature and writing about their significance and relevance to cancer. All team members helped to edit and read over the paper completely before submission, and actively participated in the conclusion of the project.

I. Introduction

Cancer is a disease that is currently causing a significant number of human deaths around the world. Discovering effective treatments is a challenge due to the nature of the disease. Cancer cells are constantly diving and rapidly spreading, causing tumors and metastases. The main problem researchers encounter is being able to successfully engineer tissue specific drugs that actively target diseased tissue, while attempting to leave normal tissue intact (differential targeting). Some of the main methods of differential targeting include surgical resection of primary tumor (the cancer tumor is excised along with some neighboring normal tissue), chemotherapy and radiotherapy, where DNA-damaging agents are used to exploit the much greater proliferative rates of cancer cells. Cancer cells are more susceptible to these damaging agents since they go through more cycles of cell division in comparison to normal cells. But in recent years, targeted chemotherapy methods have been gaining more popularity with the success of drugs such as Novartis and Genentech. Targeted cancer therapies are drugs or other substances that help stop the spread of cancer by interfering with specific proteins that are involved in cell signaling pathways. These pathways control the basic functions of cells such as cell division and cell death. For this reason, targeted cancer therapies can stop cancer progression and help induce apoptosis. Traditional chemotherapy destroys cells that are spreading and dividing rapidly while targeted therapies are more specific, selective, and target exact intracellular pathways, saving more healthy cells. Since targeted cancer therapies affect specific pathways, they only have an effect on the cancer cells that are involved with those particular pathways. By investigating and predicting specific pathways and genes that are affected by three drugs, Erlotinib, Sorafenib, and Topotecan, we will be able to deduce if these targeted compounds are effective for treating cancer. Before continuing with the analysis, it is important to get a brief background on the compounds in question:

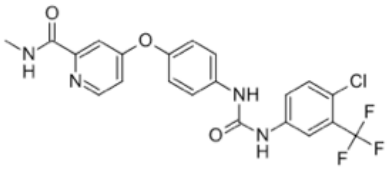
Erlotinib (Trade name Tarceva)



This drug is used to treat non-small cell lung cancer (NSCLC), pancreatic cancer and several other types of cancer. It treats non-small cell lung cancer that has spread to nearby tissues or to other parts of the body in patients who have already been treated with at least one other chemotherapy medication with no effect. It has been shown to be effective in lung cancer patients with or without EGFR mutations, but appears to be more effective in patients with EGFR mutations. It is used in combination with the medication gemcitabine [Gemzar] to treat pancreatic cancer that has spread.

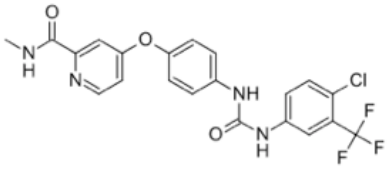
Erlotinib is a reversible tyrosine kinase inhibitor, which acts on the epidermal growth factor receptor (EGFR), which is highly expressed and occasionally mutated in various types of cancer. It binds in a reversible fashion to the adenosine triphosphate (ATP) binding site of the receptor. By inhibiting the ATP, formation of phosphotyrosine residues in EGFR is not possible and the signal cascades are not initiated. However, patients typically develop resistance within 8–12 months from starting treatment. Over 50% of resistance is due to a "gatekeeper mutation", in which the ATP binding pocket of the EGFR kinase domain substitutes a small polar threonine residue with a large nonpolar methionine residue (T790M), preventing Erlotinib from binding through steric hindrance. Due to the unwanted developments of resistance in patients treated with Erlotinib, a more promising approach to treating these patients would be a "drug cocktail" (combination therapy) to surpass resistance.
Erlotinib is marketed in the United States by Genentech, OSI Pharmaceuticals, and by Roche in the rest of the world. The drug is a tablet taken by mouth usually once a day. The most common side effects include a rash on the head and neck (which may in fact indicate clinical benefit), diarrhea, loss of appetite, and fatigue.

Sorafenib (Trade name Nexavar)



Sorafenib is used to treat advanced renal cell carcinoma (RCC, which begins in the kidneys and is also known as primary kidney cancer), unresectable hepatocellular carcinoma (HCC, an advanced primary liver cancer that cannot be treated with surgery), and radioactive iodine resistant advanced thyroid carcinoma. Sorafenib prolongs cancer-progression-free survival in patients with advanced RCC in whom previous therapy has failed, and shows a 44% improvement in median overall survival in patients with liver cancer. In thyroid cancer, there was significant improvement in progression-free survival but not in overall survival, with frequent side effects.
Sorafenib is a kinase inhibitor of several tyrosine protein kinases (VEGFR and PDGFR) and Raf kinases (more so of C-Raf than B-Raf), and also of some intracellular serine/threonine kinases (such as C-Raf, wild-type B-Raf, and mutant B-Raf). It works by blocking the actions of abnormal proteins that signal cancer cell multiplication. It induces autophagy, which may suppress tumor growth, but may also cause drug resistance.
Sorafenib is co-developed and co-marketed by Bayer and Onyx Pharmaceuticals. It comes in 200mg tablets, with a recommended starting dose of 2 tablets per day. The most common side effects are diarrhea, fatigue, rashes and itching, hand-foot syndrome (redness, pain, swelling, or blisters on hands or feet), and change in thyroid hormone levels.

Topotecan (Trade name Hycamtin)



Topotecan is a chemotherapeutic agent that has been approved to treat ovarian cancer that has spread after other treatments have failed, cervical cancer that cannot be treated with surgery or radiation therapy in combination with cisplatin, and small cell lung cancer (SCLC). It is in experimental use for neuroblastomas, brainstem gliomas, and Ewing's sarcoma.
Topotecan is a topoisomerase inhibitor that is a water-soluble derivative of camptothecin and is used in the form of hydrochloride for treatment. Camptothecin is a natural product extracted from the bark of the tree, Camptotheca acuminata. The drug works by blocking the action of an enzyme in cells called topoisomerase-I, which keeps DNA in the proper shape when cells are dividing. Blocking this enzyme leads to breaks in the DNA, which leads to cell death. Because

cancer cells divide more rapidly than normal cells, they are more likely than normal cells to be affected by Topotecan.
GlaxoSmithKline released Topotecan as the first topoisomerase-I inhibitor for oral use, though it can also be given by an infusion into a vein (IV) over 30 minutes. The typical schedule is once a day for 3 to 5 days, which is usually repeated every 3 weeks, but the dose and schedule depends on many factors, including body size, blood counts, and the type of cancer. The most common side effects are myelosuppression, diarrhea, low blood count, and susceptibility to infection.

For this project, different tools were used to perform analysis and obtain results. These tools include WEKA (refer to ***Discussion*** section), GenePattern, GSEA, and DAVID.

Gene Pattern:
Gene Pattern is a platform consisting of various tools to perform genomic analysis. One of these tools includes Differential Expression Analysis (also known as marker selection) which finds genes that are differentially expressed between distinct phenotypes. Gene Pattern uses either the signal-to noise ratio or t-test statistic to determine differential expression and ranks the genes on the value based on the same statistic (usually signal-to-noise). Within differential analysis, we used Comparative Maker Selection, which ranks the genes based on the value of the statistic being used to assess differential expression and uses permutation testing to compute the significance (p-value) of the rank assigned to each gene.

Gene Set Enrichment Analysis (GSEA):
GSEA is a computational method that takes in sets of genes and determines whether they are statistically significant as well as the differences between the two phenotypes. Three key elements of the GSEA method are: calculation of an enrichment score (ES), estimation of significance level of ES, and adjustment for multiple hypothesis testing. The MSigDB (Molecular Signatures Database) is used with the GSEA computational software since it contains a wide collection of annotated gene sets.

In the Molecular Signatures Database, *Compute Overlaps* is a tool that evaluates the overlap and statistical significance between the gene set the user provides and the other genes sets in MSigDB. Examining overlaps can specify various commonalities between the gene sets such as processes, pathways, and biological functions. The description of the overlapping gene sets in the GSEA figures included, include a link to the gene set page, number of genes in the gene set, a description of the gene set, as well as the number of genes in the overlap between this gene set and our inputted gene set. Furthermore, the p-value is from hypergeometric distribution (k-1, K, N, n), which applies to sample that cannot be replaced since they are from populations whose elements are dichotomous variables. In this case, the k is the number of genes in the intersection of the query set with a set from MSigDB, K is the total numbers of genes known, N is the total number of all known human gene symbols, and n is the number of genes in the query set. The FDR q-value is the false discovery rate analog of hypergeometric p-value after correction for multiple hypothesis testing. The column of color bars range from light green to black in shading where the lighter color indicates more significant q-values (<0.05) and black indicates insignificant q-values ($>= 0.05$). Finally, an overlap matrix is also shown where the genes in the overlapping gene sets are specified. The rows correspond to the list of genes in the inputted gene set while the columns correspond to the overlapping genes.

DAVID:
Generally speaking, DAVID is a free public resource that allows rapid functional annotation of lists of genes. DAVID is designed around the "DAVID Gene Concept" which is a graph theory evidence- based method that essentially agglomerates species- specific gene/ protein identifiers

from a variety of public genomic resources such as NCBI, PIR, and Uniprot/SwissProt. In our project, we used the *Functional Annotation Tool* provided in DAVID, which basically provides typical batch annotation and gene –GO term (gene ontology) enrichment analysis to highlight the most relevant GO terms associated with the provided gene list. The last version we used in the project kept the same enrichment analytic algorithm but with extended annotation content coverage, increasing to over 40 annotation categories. These annotation categories included GO terms, protein- protein interactions, protein functional domains, bio-pathways, literature, and so on. In the Annotation Summary results page, the pathways annotation categories contains percentage (number of involved genes divided by the total number of genes) and genes from the inputted list involved in that pathway. The gene number is also portrayed graphically using a blue bar (refer to ***Supplement 7***). The chart report is an annotation-term focused view which provides lists, annotation terms, and their associated genes under study. The Fisher Exact statistics is calculated in order to avoid over counting duplicated genes. It is calculated based on the corresponding DAVID gene IDs. The chart report also consists of the same percentage and gene involved information as well as enrichment terms and related terms associated with the inputted gene list (refer to ***Supplement 8***).

II. Methods

Note: this methods section only outlines what was done for one drug, Erlonitib. However, the same exact steps were followed for Sorafenib and Topotecan.

Dividing the Cell Lines and Assigning Classifiers

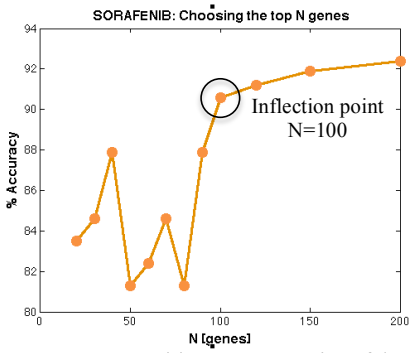
The lowest 25% of the negative scores were chosen as the *responsive* group and the upper 75% of the negative scores, and the positive scores, as the *non-responsive* group. The cell lines were divided this way because negative scores signify percent of inhibition (to a degree of responsiveness), whereas positive scores indicate percentage of cell line growth (to a degree of non-responsiveness). Therefore the responsive group was made up from the most negative scores (these cell lines are most responsive to the drug). Using MATLAB, the responsive cell lines were classified as 1, and the non-responsive cell lines were classified as 0. The CLS file was edited with these changes.

Identifying Differentially Expressed Genes

In order to identify differentially expressed genes in responsive and non-responsive cell lines, GenePattern was used. The CLS file created above and the GCT file that was provided were uploaded into the *Comparative Marker Selection* section of Genepattern. The output file that GenePattern provided was edited in order list the genes from highest absolute score to lowest absolute score –the highest absolute scores represents the most differentially expressed genes.

Choosing N and Pruning the Data

In order to pick top N genes for later use in classification, different N's were tried (N=20,30 …100). A .cvs file with the GCT data for N genes was made, and this file was uploaded into WEKA. Using Logistic Regression as our primary classifier (with 5 cross validations) on the file with N genes provided us with a certain accuracy Rate. This rate was recorded for each corresponding different value of N. These values were plotted, and the inflection point was chosen (refer to figure below).



For Erlonitib N was chosen to be 50 (Sorafenib N=100, Topotecan N=140). The Logistic regression model with 50 genes was inspected, and the top 16 coefficients were chosen (15 for Sorafenib, and 14 for Topotecan). We did not want to pick too many genes (in order to truly pick the most differently expressed genes), but we also did not want to pick too few (in order to find a more significant enrichment analysis). The final file contained the 16 (or 15 or 14) genes. These were uploaded into WEKA. The accuracy of several machine learning algorithms were compared using 5 and 10 cross-validation. From WEKA, true positive, true negative, false positive, false negative, responsive area under ROC (responsive AUC), and non-responsive area under ROC (non-responsive AUC) were obtained and recorded (refer to ***Tables 3,6,9***)

GNEA and DAVID

In order to understand over-represented biological processes and pathways as well as which specific genes are involved in those particular pathways, we used the GSEA and DAVID tools. Taking the GenePattern output mentioned earlier, we ordered the list of genes using the q-value variable (ordering them from smallest to largest). Initially, the entire odf (GenePattern output) file was inserted into GSEA to be analyzed but because of the overwhelming number of genes, this led to more background noise which yielded irrelevant pathways. Therefore, we selected the top 300 genes and inserted them into the GSEA investigate gene sets tool. For drugs Erlonitib and Sorafenib, 300 genes seemed to be the ideal number that results in the most significant over-represented pathways. For the drug Topotecan however, we increased the input gene number from 300 to 600 genes in order to get the most significant results. For all of three drugs, we computed overlap between the top genes we inputted against the C2:CP:KEGG gene sets. Furthermore, we used DAVID to further support the GSEA results in which particular pathways are the most over- represented. We inputted the same genes for all three drugs that we used with GSEA into DAVID's functional Annotation tool. In the resulting Annotation Summary, we selected the KEGG gene sets to get the specific overlapped pathways.

III. Erlotinib Results

Table 1: Chosen Prognostic Genes

ZNF529	INHBE	DIRAS1	C14orf43	EGFR
DMTF1	NKX2-5	KIAA1522	KIAA1586	
PCMT1	C14orf37	LSR	PALM	
FBXO10	NTNG2	LAMA5	PTPN3	

Table 1 *–The 17 prognostic genes that were used for Erlotinib*

Maria Barrios 5/8/14 10:40 PM
Deleted: *This table lists*

Table 2: Algorithm Accuracy for Erlotinib

		Logistic Regression	Naïve Bayes	Neural Network
% Accuracy	**5 Cross Validation**	80.2	87.9	90.1
	10 Cross Validation	84.6	89.0	91.2

Table 2 *–Accuracy of the algorithms that were used. Logistic regression was used a baseline. Neural Networks with 10 cross validation proved to be the best classifier on the data set that corresponds to the genes in* ***Table 1****.*

Maria Barrios 5/8/14 10:40 PM
Deleted: *This table documents the*

Table 3: True Positive, False Positive, True Negative, True Positive Rates

		TP (Sensitivity)	TN (Specififcity)	FP	FN	Responsive AUC	Non-Responsive AUC
Logistic Regression	**5 Cross Validation**	0.783	0.809	0.217	0.191	0.872	0.834
	10 Cross Validation	0.739	0.882	0.261	0.118	0.912	0.873
Naïve Bayes	**5 Cross Validation**	0.913	0.868	0.087	0.132	0.939	0.941
	10 Cross Validation	0.913	0.882	0.087	0.118	0.944	0.944
Neural Networks	**5 Cross Validation**	0.739	0.956	0.261	0.044	0.952	0.952
	10 Cross Validation	0.783	0.956	0.217	0.044	0.945	0.945

Table 3 *–True positive (TP), false positive (FP), true negative (TN), and true positive (TP) rates for each algorithm performed on the Erlotinib data represented in* ***Table 1****. The sensitivity and specificity of the predictions are given by TP and FP, respectively. The Area under the responsive and non-responsive ROC is also documented –this is known as the area under the curve (AUC)*

Maria Barrios 5/8/14 10:41 PM
Deleted: *This table documents*

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value	FDR q-value
KEGG_PATHWAYS_IN_CANCER [328]	Pathways in cancer	9		$9.54\ e^{-5}$	$6.53\ e^{-3}$
KEGG_CELL_ADHESION_MOLECULES_CAMS [134]	Cell adhesion molecules (CAMs)	6		$1.05\ e^{-4}$	$6.53\ e^{-3}$
KEGG_ECM_RECEPTOR_INTERACTION [84]	ECM-receptor interaction	5		$1.05\ e^{-4}$	$6.53\ e^{-3}$
KEGG_GLYCOSPHINGOLIPID_BIOSYNTHESIS_LA_LACTO_AND_NEOLACTO_SERIES [26]	Glycosphingolipid biosynthesis - lacto and neolacto series	3		$3.9\ e^{-4}$	$1.82\ e^{-2}$
KEGG_RIG_I_LIKE_RECEPTOR_SIGNALING_PAT PATHWAY [71]	RIG-I-like receptor signaling pathway	4		$6.53\ e^{-4}$	$2.42\ e^{-2}$
KEGG_ADHERENS_JUNCTION [75]	Adherens junction	4		$8.02\ e^{-4}$	$2.42\ e^{-2}$
KEGG_FOCAL_ADHESION [201]	Focal adhesion	6		$9.12\ e^{-4}$	$2.42\ e^{-2}$
KEGG_SNARE_INTERACTIONS_IN_VESICULAR_TR_TRANSPORT [38]	SNARE interactions in vesicular transport	3		$1.21\ e^{-3}$	$2.8\ e^{-2}$
KEGG_APOPTOSIS [88]	Apoptosis	4		$1.46\ e^{-3}$	$3\ e^{-2}$
KEGG_BLADDER_CANCER [42]	Bladder cancer	3		$1.62\ e^{-3}$	$3\ e^{-2}$

Figure 1 *–GSEA results for Erlotinib. These are the enriched gene sets, with their corresponding statistics.*

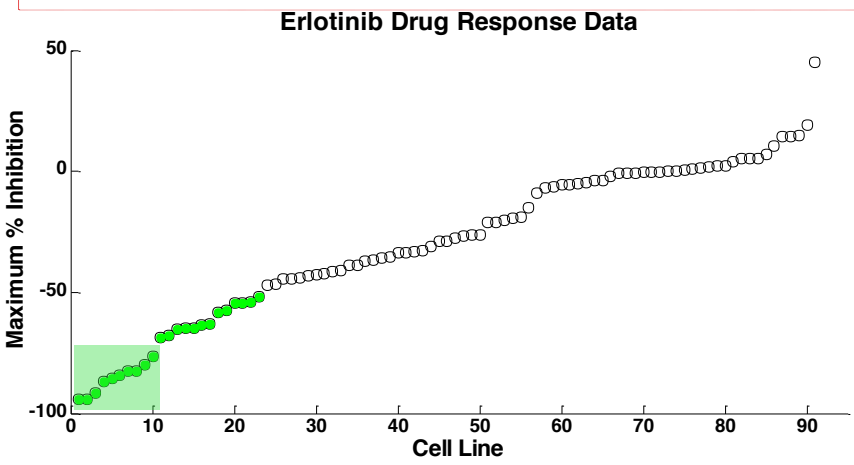
Maria Barrios 5/8/14 10:41 PM
Deleted: *this Figure illustrates the*

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
☐	KEGG_PATHWAY	ECM-receptor interaction	RT	6	0.2		7.6E-3	5.1E-1
☐	KEGG_PATHWAY	Cell adhesion molecules (CAMs)	RT	6	0.2		4.4E-2	8.8E-1
☐	KEGG_PATHWAY	Glycosphingolipid biosynthesis	RT	3	0.1		5.2E-2	8.1E-1
☐	KEGG_PATHWAY	RIG-I-like receptor signaling pathway	RT	4	0.1		8.7E-2	8.8E-1

Figure 2 *–Results from DAVID for Erlotinib. These are the enriched gene sets, with their corresponding statistics. Note that that the ECM-receptor Interaction pathway, the Cell Adhesion Molecules pathway, and the Glycosphingolipid Biosynthesis pathway are found in DAVID* ***and*** *GSEA results.*

Maria Barrios 5/8/14 10:41 PM
Deleted: *this Figure illustrates the*

Maria Barrios 5/8/14 10:33 PM
Comment [1]: The figure below was added



Drug Response E: *Drug response curve for the 91 cell lines treated with Erlotinib. The green markers indicate the 23 cell lines (25%) that were classified as responsive and the light green box highlights the cell lines that should have been chosen as responsive taking into account the observed inflection point, which ultimately should have yielded better classifier models.*

IV. Sorafenib Results

Table 4: Chosen Prognostic Genes

AVPR1A	KCNIP1	SHH	OR7C1
ANP32D	CCDC33	ADGB	C3orf30
RPS16	NA	EBF2	TRH
RACGAP1P	CCT8	ADAMTS8	VEGFA

Table 4 *–16 prognostic genes that were used for Sorafenib. The NA gene refers to a probe AFFX-r2-Bs-dap-3_at. This probe did not have a gene match up.*

Maria Barrios 5/8/14 10:41 PM
Deleted: *This table lists the*

Table 5: Algorithm Accuracy for Sorafenib

		Logistic Regression	Naïve Bayes	Neural Network
% Accuracy	**5 Cross Validation**	89	93.4	96.7
	10 Cross Validation	90.1	95.6	97.8

Table 5 *–Accuracy of the algorithms that were used. Logistic regression was used a baseline. Neural Networks with 10 cross validations proved to be the best classifier on the data set that corresponds to the genes in* ***Table 4****.*

Maria Barrios 5/8/14 10:41 PM
Deleted: *This table documents the*

Table 6: True Positive, False Positive, True Negative, True Positive Rates

		TP (Sensitivity)	TN (Specififcity)	FP	FN	Responsive AUC	Non-Responsive AUC
Logistic Regression	**5 Cross Validation**	0.826	0.912	0.174	0.088	0.974	0.647
	10 Cross Validation	0.783	0.941	0.217	0.059	0.969	0.523
Naïve Bayes	**5 Cross Validation**	0.739	1	0.261	0	0.783	0.928
	10 Cross Validation	0.826	1	0.174	0	0.870	0.985
Neural Networks	**5 Cross Validation**	0.870	1	0.130	0	0.948	0.786
	10 Cross Validation	0.913	1	0.087	0	0.965	0.740

Table 6 *–True positive (TP), false positive (FP), true negative (TN), and true positive (TP) rates for each algorithm performed on the Sorafenib data represented in* ***Table 4****. The sensitivity and specificity of the predictions are given by TP and FP, respectively. The Area under the responsive and non-responsive ROC is also documented –this is known as the area under the curve (AUC)*

Maria Barrios 5/8/14 10:42 PM
Deleted: *This table documents*

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value	FDR q-value
KEGG_LONG_TERM_DEPRESSION [70]	Long-term depression	5		$3.63\ e^{-5}$	$6.52\ e^{-3}$
KEGG_TIGHT_JUNCTION [134]	Tight junction	6		$8.39\ e^{-5}$	$6.52\ e^{-3}$
KEGG_FOCAL_ADHESION [201]	Focal adhesion	7		$1.05\ e^{-4}$	$6.52\ e^{-3}$
KEGG_REGULATION_OF_ACTIN_CYTOSKELETON [216]	Regulation of actin cytoskeleton	7		$1.64\ e^{-4}$	$7.63\ e^{-3}$
KEGG_PATHWAYS_IN_CANCER [328]	Pathways in cancer	8		$3.9\ e^{-4}$	$1.32\ e^{-2}$
KEGG_LEUKOCYTE_TRANSENDOTHELIAL_MIGRATRATION [118]	Leukocyte transendothelial migration	5		$4.27\ e^{-4}$	$1.32\ e^{-2}$
KEGG_NEUROTROPHIN_SIGNALING_PATHWAY [126]	Neurotrophin signaling pathway	5		$5.76\ e^{-4}$	$1.37\ e^{-2}$
KEGG_MAPK_SIGNALING_PATHWAY [267]	MAPK signaling pathway	7		$5.88\ e^{-4}$	$1.37\ e^{-2}$
KEGG_FC_EPSILON_RI_SIGNALING_PATHWAY [79]	Fc epsilon RI signaling pathway	4		$8.4\ e^{-4}$	$1.74\ e^{-2}$
KEGG_PROSTATE_CANCER [89]	Prostate cancer	4		$1.31\ e^{-3}$	$2.31\ e^{-2}$

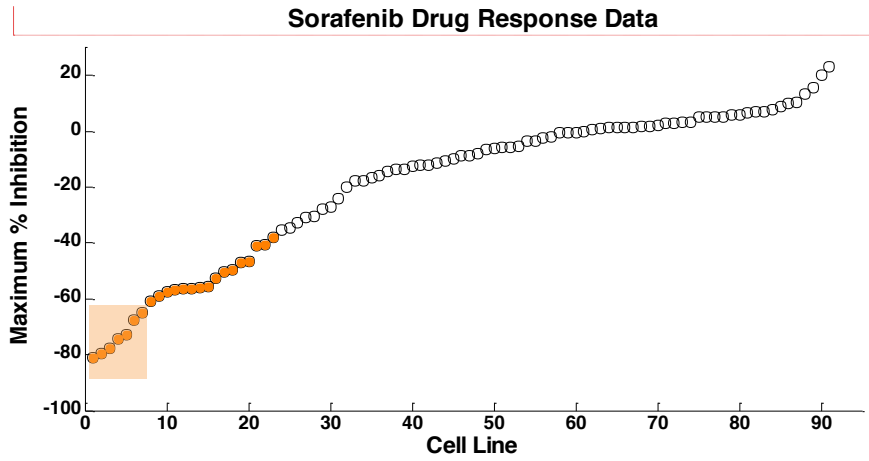
Figure 3 *–GSEA results for Sorafenib. These are the enriched gene sets, with their corresponding statistics.*

Maria Barrios 5/8/14 10:42 PM
Deleted: *this Figure illustrates the*

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	KEGG_PATHWAY	Long-term depression	RT	5	0.2	2.3E-2	9.2E-1	
	KEGG_PATHWAY	Tight junction	RT	6	0.2	5.9E-2	9.7E-1	
	KEGG_PATHWAY	Focal adhesion	RT	7	0.3	9.5E-2	9.8E-1	

Figure 4 *–Results from DAVID for Sorafenib. These are the enriched gene sets, with their corresponding statistics. Note that that the Long Term Depression Pathway, Tight Junction Pathway, and Focal Adhesion Pathway are found in DAVID* ***and*** *GSEA results.*

Maria Barrios 5/8/14 10:35 PM
Deleted: *this Figure illustrates the*



Maria Barrios 5/8/14 10:35 PM
Comment [2]: The figure below was added

Drug Response S: *Drug response curve for the 91 cell lines treated with Sorafenib. The orange markers indicate the 23 cell lines (25%) that were classified as responsive and the light orange box highlights the cell lines that should have been chosen as responsive taking into account the observed inflection point, which ultimately should have yielded better classifier models.*

V. Topotecan Results

Table 7: Chosen Prognostic Genes

ATP5S	NDUFB6	ANXA7	CAPN11
SDC3	ERGIC1	NA	NA
ZNF770	MINK1	TBC1D10B	TOP1
KIAA0494	SMG6	DIO1	FAN1

Table 7 *–16 prognostic genes that were used for Topotecan. The NA genes refers to a probes 28660_at and 284080_at, respectively. These probes did not have a gene match up.*

Maria Barrios 5/8/14 10:42 PM
Deleted: *This table lists the*

Table 8: Algorithm Accuracy for Topotecan

		Logistic Regression	Naïve Bayes	Neural Network
% Accuracy	**5 Cross Validation**	75.8	93.4	85.7
	10 Cross Validation	76.9	92.3	89.0

Table 8 *–Accuracy of the algorithms that were used. Logistic regression was used a baseline. Naïve Bayes with 5 cross validations proved to be the best classifier on the data set that corresponds to the genes in* ***Table*** *7.*

Maria Barrios 5/8/14 10:43 PM
Deleted: *This table documents the*

Table 9: True Positive, False Positive, True Negative, True Positive Rates

		TP (Sensitivity)	TN (Specififcity)	FP	FN	Responsive AUC	Non-Responsive AUC
Logistic Regression	**5 Cross Validation**	0.652	0.912	0.348	0.088	0.836	0.839
	10 Cross Validation	0.652	0.809	0.348	0.191	0.862	0.855
Naïve Bayes	**5 Cross Validation**	0.870	0.956	0.130	0.044	0.970	0.970
	10 Cross Validation	0.826	0.956	0.174	0.044	0.967	0.967
Neural Networks	**5 Cross Validation**	0.565	0.956	0.435	0.044	0.939	0.939
	10 Cross Validation	0.696	0.956	0.304	0.044	0.946	0.946

Table 9 *–True positive (TP), false positive (FP), true negative (TN), and true positive (TP) rates for each algorithm performed on the Topotecan data represented in* ***Table*** *7. The sensitivity and specificity of the predictions are given by TP and FP, respectively. The Area under the responsive and non-responsive ROC is also documented –this is known as the area under the curve (AUC)*

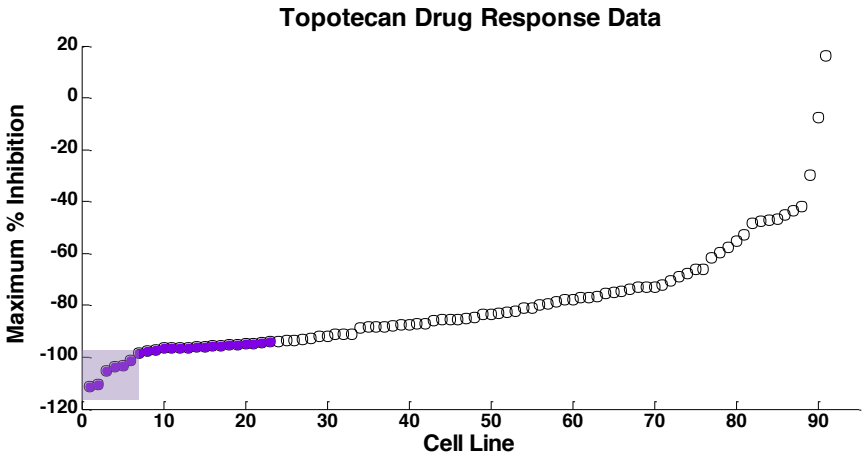
Maria Barrios 5/8/14 10:43 PM
Deleted: *This table documents*

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value	FDR q-value
KEGG_FOCAL_ADHESION [201]	Focal adhesion	11		$1.1\ e^{-5}$	$1.96\ e^{-3}$
KEGG_REGULATION_OF_ACTIN_CYTOSKELETON [216]	Regulation of actin cytoskeleton	11		$2.15\ e^{-5}$	$1.96\ e^{-3}$
KEGG_MAPK_SIGNALING_PATHWAY [267]	MAPK signaling pathway	12		$3.16\ e^{-5}$	$1.96\ e^{-3}$
KEGG_CHEMOKINE_SIGNALING_PATHWAY [190]	Chemokine signaling pathway	9		$2.06\ e^{-4}$	$9.57\ e^{-3}$
KEGG_CELL_ADHESION_MOLECULES_CAMS [134]	Cell adhesion molecules (CAMs)	7		$5.8\ e^{-4}$	$1.81\ e^{-2}$
KEGG_FC_GAMMA_R_MEDIATED_PHAGOCYTOSIS [97]	Fc gamma R-mediated phagocytosis	6		$5.89\ e^{-4}$	$1.81\ e^{-2}$
KEGG_SNARE_INTERACTIONS_IN_VESICULAR_TR_TRANSPORT [38]	SNARE interactions in vesicular transport	4		$6.81\ e^{-4}$	$1.81\ e^{-2}$
KEGG_VASOPRESSIN_REGULATED_WATER_REABSABSORPTION [44]	Vasopressin-regulated water reabsorption	4		$1.19\ e^{-3}$	$2.77\ e^{-2}$
KEGG_AXON_GUIDANCE [129]	Axon guidance	6		$2.56\ e^{-3}$	$4.6\ e^{-2}$
KEGG_GAP_JUNCTION [90]	Gap junction	5		$2.7\ e^{-3}$	$4.6\ e^{-2}$

Figure 5 *–GSEA results for Topotecan. These are the enriched gene sets, with their corresponding statistics.*

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
☐	KEGG_PATHWAY	Regulation of actin cytoskeleton	RT		12	0.2	3.4E-2	9.9E-1
☐	KEGG_PATHWAY	Focal adhesion	RT		11	0.2	4.9E-2	9.6E-1
☐	KEGG_PATHWAY	SNARE interactions in vesicular transport	RT		4	0.1	8.6E-2	9.8E-1

Figure 6 *–Results from DAVID for Topotecan. These are the enriched gene sets, with their corresponding statistics. Note that that the Regulation of Actin Cytoskeleton Pathway, Focal Adhesion Pathway, and the SNARE Interactions in Vesicular Transport Pathway are found in DAVID* ***and*** *GSEA results.*



Drug Response T: *Drug response curve for the 91 cell lines treated with Topotecan. The purple markers indicate the 23 cell lines (25%) that were classified as responsive and the light purple box highlights the cell lines that should have been chosen as responsive taking into account the observed inflection point, which ultimately should have yielded better classifier models.*

Maria Barrios 5/8/14 10:40 PM
Comment [3]: Figure added below

Maria Barrios 5/8/14 10:43 PM
Deleted: *this Figure illustrates the*

Maria Barrios 5/8/14 10:43 PM
Deleted: *this Figure illustrates the*

VI. Discussion

WEKA and Prognostic Genes

This project was tailored towards building a prognostic model that identifies a small number of prognostic genes. With the expression levels of these genes in specific cell lines, one should (theoretically) be able to classify a cell line as responsive or non-responsive to the drug in question. Once N was chosen (refer to ***Choosing N and Pruning the Data*** in the ***Methods*** Section), WEKA was used to find the top prognostic genes. WEKA is a machine-learning software that contains algorithms for data analysis and predictive modeling. All of the machine learning tools in WEKA (i.e. Naïve Bayes, Logistic Regression, Neural Networks…etc.) are statistically based algorithms. Because of this, every time WEKA runs a machine-learning algorithm, it reports the accuracy statistics of the analysis.

The **Logistic Regression** model is a type of probabilistic statistical classification model that is used to predict the outcome of a categorical dependent variable (i.e., a class label) based on one or more predictor variables (i.e., features). In our case the class label was either Responsive or Nonresponsive and the features were the expression levels of variable probes. The probabilities describing the possible outcomes of a single trial are modeled, as a function of the predictor variables, using a logistic function. The **Naïve Bayes** classifier model is a simple probabilistic classifier based on applying Bayes' theorem with strong independence assumptions. It assumes that the value of a particular feature, or predictor variable, is completely conditionally independent to the presence or absence of any other feature, given the class label. Naïve Bayes is a learning algorithm with greater bias, but lower variance, than Logistic Regression. The **Neural Networks** model maps sets of input data onto a set of appropriate outputs. It utilizes a supervised learning technique called back-propagation for training the network and can distinguish data that are not linearly separable. The Neural Networks model (specifically the multilayer perceptron) consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Note that it contains an input layer, an output layer, and some number of hidden layers.

Maria Barrios 5/8/14 10:21 PM
Deleted: unrelated

In order to make the model more accurate, the top N genes were further "trimmed" –all with the intent of increasing prediction accuracies (using logistic regression as the primary classifier for this part). All of these top prognostic genes were "hand-picked". What does this mean? The logistic regression model (on the top N genes) assigned a beta coefficient (positive or negative) to each gene. Because both large positive and negative coefficients are important, the absolute values of the coefficients were analyzed. An arbitrary cut-off was determined –all of the genes with beta coefficients below this cut-off were discarded, and those above the cut-off were considered to be part of the prognostic gene pool. This cut-off was tailored to increase the prediction accuracies from the logistic regression model for the top N genes to the logistic regression model to the "hand-picked" prognostic genes. For Erlotinib, the prediction accuracy increased from 73.6%(N=50) to 84.6%. (N=17). For Sorafenib, the prediction accuracy increased from 87.9% (N=100) to 90.1% (N=16). Lastly, for Topotecan, the prediction accuracy increased from 81.3% (N=140) to 82.4% (N=16). The fact that the prediction accuracies increased shows that the "hand-picking" was done thoroughly. Note that, upon further research, it was discovered that some specific genes were targeted and/or affected by the drugs. If these genes were not included in the final top prognostic genes, they were manually added to the list since they were known to be important ahead of time (and were wrongfully discarded by the arbitrary choosing of the beta cut-off). The genes that were manually added were EGFR for Erlotinib, VEGFA for Sorafenib, and TOP1 and FAN1 for Topotecan.

When identifying the best classifier, all of the accuracies increased even more! For Erlonitib and Sorafenib, the best classifier was Neural Networks at a 10 cross-validation, with accuracies of 91.2% and 97.8%, respectively. The 10-fold cross-validation partitioned the data into 90% training and 10% testing. For Topotecan, the best classifier was Naïve Bayes at a 5 cross-validation with an accuracy of 93.4%. The 5-fold cross-validation partitioned the data into 80% training and 20% testing.

From a venture capitalists point of view, it is important that our classifier models, using the top prognostic genes that we chose, perform better than classifiers using the target genes alone. In response to this, we can say with confidence that our prognostic gene list performs better than the target gene alone. For Erlotinib, our model produced an accuracy of 91.2%, which surpassed the accuracy given by the target gene alone. When using EGFR (Erlotinib's target), an accuracy of 76.9% was yielded using both logistic regression and Neural Networks model at both 5 and 10 fold-cross-validations. For Sorafenib, our model produced an accuracy of 97.8%, which one again, surpassed the accuracy given by one of the target genes alone. When using VEGFA (one of Sorafenib's main targets), an accuracy of 74.7% was yielded using both logistic regression and Neural Networks model at both 5 and 10 fold-cross-validations. Lastly, for Topotecan, our model produced an accuracy of 93.4%, which surpassed the accuracy given by the target gene alone. When using TOP1 (Topotecan's target), an accuracy of 74.7% was yielded using both logistic regression and Neural Networks model at both 5 and 10 fold-cross-validations. To sum up these results, our prognostic gene sets for all three drugs performed better than target gene classifier models at predicting the responsiveness of the given cell lines.

For this project, the sensitivity and specificity (reported in WEKA) of a prediction measured the proportion of actual responsive and non-responsive cell lines that were correctly identified as such, respectively. Because specificity and sensitivity cannot simultaneously be improved, the prognostic model could only be tailored to maximize sensitivity or specificity. Therefore, once the prediction accuracies were maximized, the sensitivities and specificities of each model were analyzed. Based on ***Table 3***, ***Table 6***, and ***Table 9***, it can be seen that the true negative rate (specificity) is always larger and much closer to 1 than the true positive rate (sensitivity). With these results, it can be concluded that the prognostic models for each of the three drugs maximized specificity.

Based on Prognostic Genes: Gene Discussion for each Drug
Erlotinib: EGFR and LAMA5
EGFR, or epidermal growth factor receptor, is a cell surface protein that binds to epidermal growth factor, which leads to receptor dimerization and tyrosine autophosphorylation and cell proliferation. Mutations with this gene are common with various types of cancer (lung cancer). In cancers such as lung cancer and pancreatic cancer, higher than normal numbers of these receptors are observed on their cell surfaces, which makes sense since higher numbers of EGFR results in higher levels of cancer cell proliferation. The drug, Erlotinib, targets these tyrosine kinase proteins (EGFR) by inhibiting their signals that cause cancer cells to multiply. This drug functions by stopping or slowing down the spread of cancer cells, blocking the receptor so that they cannot signal the cancer cells to divide and grow. GenePattern yielded a score (for Erlotinib) of -3.2308 for the EGFR gene. This negative score tells us that EGFR expression levels are up-regulated in the Responsive cell lines. This up-regulation of EGFR in Responsive cell lines was expected since Erlotinib is administered in hope of inhibiting the up-regulation of EGFR.

Maria Barrios 5/8/14 10:22 PM
Deleted: GenePattern yielded a score (for Erlotinib) of -3.2308 for the EGFR gene which is most certainly what is expected. The down-regulation of EGFR in cancer patients treated with Erlotinib illustrates that Erlotinib is effective in stopping, or at least slowing down, the unwanted growth of cancer cells.

LAMA5, or laminin-alpha 5, is a gene that codes the LAMA5 protein, which is (as mentioned above) implicated in a wide variety of biological processes including cell adhesion, differentiation, migration, signaling, neurite outgrowth and metastasis. The laminins, a family of heterotrimeric extracellular glycoproteins, affect tissue development and integrity in such diverse organs such as the lung, kidney, pancreas, and skin. Note that it is thought that laminins mediate the attachment, migration, and organization of cells into tissues during embryonic development by interacting with other extracellular matrix components. Now, the LAMA5 protein is a major component of basement membranes, which have various functions, one of which is acting as a protective barrier against foreign objects or malignant cells (tumor/cancer cells) between the surface of organs and the internal tissues. Also, the basement membrane located on the interior of a blood vessel is involved in the process of angiogenesis. Vascular basement membrane components have been found to be involved in the regulation of tumor angiogenesis. One would expect cancer patients to have abnormal and unwanted basement membrane compositions, which would lead to the permission of unwanted cancer cells to travel into deep tissues and organs, as well as allow unwanted tumor angiogenesis to occur. GenePattern yielded a score (for Erlotinib) of -4.3531 for the LAMA5 gene – this indicates that LAMA5 is up-regulated in Responsive cell lines. With these results we can gather that before treatment with Erlotinib, the cell lines that were later classified as responsive, had a normal basement membrane composition (which is strange since membrane composition is altered in cancer). Erlotinib does not target LAMA5 (despite the significance that LAMA5 was given through WEKA and GSEA analysis) therefore we cannot make any conclusions on the effect Erlotinib has on LAMA5. However, due to LAMA5's important function in metastasis and angiogenesis, any down-regulation of this gene should be carefully monitored to prevent (or stop) metastasis.

Sorafenib: VEGFA and RACGAP1

Upon treatment with Sorafenib, genes (functionally) involved in angiogenesis, apoptosis, transcription regulation, signal transduction, protein biosynthesis and modification are predominantly up-regulated, while genes involved in cell cycle control, DNA replication, recombination, and repair, cell adhesion, metabolism, and transport are mainly down-regulated upon treatment (Cervello). Based on this knowledge, the prognostic genes VEGFA and RACGAP1 will be analyzed, and their role within the Sorafenib domain will be explored.

In the literature it has been found that Sorafenib "potently inhibits the proangiogenic vascular endothelial factor receptor VEGFR2" (Wilhelm). VEGFR2 is the main vascular endothelial growth factor (VEGF) receptor. The binding of VEGFA to its surface activates VEGFR2 –this binding mediates agiogenesis –the growth of new blood vessels from pre-existing vessels (Wilhelm). VEGFA is a protein in humans that is encoded by the VEGFA gene –VEGFA was one of our prognostic genes for Sorafenib (refer to ***Table 4***). This gene had a positive score (1.30) in the GenePattern output file –this suggests that VEGFA is up-regulated in Non-Responsive cell lines before treatment with Sorafenib. These results are not what we expected. Because Sorafenib inhibits the activity of VEGFR2 (and therefore inhibits the activity of VEGFA), we expected VEGFA to be up-regulated in Responsive cell lines, so that treatment with Sorafenib could inhibit that up-regulation, in hopes of stopping angiogenesis. When dividing the cell lines, the lowest 25% of the cell lines were chosen as the Responsive group. However, as can be clearly seen in ***Drug Response S***, there is a much clearer inflection point that should have been taken into account when choosing the responsive cell lines. Therefore, by choosing the lowest 25% as responsive, we included some cell lines that were most likely non-responsive –these cell lines influenced the statistics given by GenePattern, and could explain why the score we obtained did not match our expectations.

Maria Barrios 5/8/14 10:22 PM

Deleted: which indicates that the basement membranes in the tissues of study are compromised and that the drug, Erlotinib, is not effective in promoting healthy basement membrane function. As discussed, the composition of basement membranes is of great importance in lieu of different types of cancer, and their components (such as LAMA5) are attractive candidate targets for potential cancer therapies. Although Erlotinib is effective in slowing down the unwanted growth of cancer cells (by targeting the EGFRs), it can be deduced that other types of treatment that promote healthy structure of basement membranes must be combined with Erlotinib to improve the effects of Erlotinib on cancer patients.

Maria Barrios 5/8/14 10:23 PM

Deleted: This gene had a high positive score in the GenePattern output file –this suggests that VEGFA is up-regulated in Responsive patients upon treatment with Sorafenib. These results match the expected results that were found in the *Molecular Mechanisms of Sorafenib Action in Liver Cancer* research paper: "upon treatment with Sorafenib, genes functionally involved in angiogenesis . . . are predominantly up-regulated" (Cervello). That Sorafenib inhibits the activity of VEGFR2 makes sense, since we want to stop angiogenesis in cancer. However, the up-regulation of VEGFA upon treatment with Sorafenib causes some confusion, since the up-regulation of VEGFA promotes angiogenesis. For this very reason, we suggest a drug cocktail in the scenario, in order to combat these unwanted effects. Despite this, it makes sense that the drug targets genes involved with angiogenesis (such as VEGFA) –if angiogenesis is threatened, metastasis can be stopped or at least delayed.

RACGAP1 encodes Rac GTPase-activating protein 1, which belongs to the GTPase-activating protein (GAP) family (Rac). This protein plays a regulatory role in controlling cell growth and differentiation of hematopoietic cells. The up-regulation of RACGAP1 promotes cell growth, therefore one would expect the up-regulation of RACGAP1 in cancer patients. This gene had a negative score (-0.072) in GenePattern –this suggests that RACGAP1 was up-regulated in Responsive cell lines. This up-regulation in responsive cell lines was expected since Sorafenib is expected to inhibit RACGAP1, in order to stop the proliferation of cancer cells: According to the *Molecular Mechanisms of Sorafenib Action in Liver Cancer* research paper, "RACGAP1 was specifically down-regulated upon treatment [with Sorafenib]". As seen, the literature supports with our results.

Topotecan: TOP1 and FAN1

As explained before, Topotecan works by blocking the action of Topoisomerase-I. Topoisomerase-I is an enzyme that ensures proper shape of DNA when cells are dividing. The TOP1 gene encodes topoisomerase-I. This gene was manually added to the Topotecan prognostic genes due the important role it has within the domains of Topotecan. TOP1 was given a negative score (-2.22) by GenePattern –this suggests that TOP1 was up-regulated in the Responsive cell lines. This up-regulation of TOP1 in Responsive cell lines was expected since Topotecan is administered in hope of inhibiting the expression of TOP1. The down-regulation of TOP1 leads to breaks and imperfections in the DNA, which leads to cell death. Therefore, treating cancer with Topotecan targets the DNA of replicating cells, killing any cells whose DNA has failed to be regulated by Topoisomerase-I. This form of action, however, can also kill normal cells, since Topoisomerase-I regulates the shape of DNA in all replicating cells. Therefore, Topotecan can also kill normal cells. However, because cancer cells divide more rapidly than normal cells, they are more likely than normal cells to be affected by Topotecan.

FAN1, or FANCD2/FANCI-associated nuclease, is a protein-coding gene that codes for the FAN1 protein which is required for the maintenance of chromosomal stability. This gene plays a role in DNA repair of interstrand cross-links (ICL) by being recruited to sites of DNA damage by monoubiquinated FANCD2. Specifically, it is involved in the repair of ICL-induced DNA breaks by being required for efficient homologous recombination. Depletion (or down-regulation) of FAN1 causes DNA damage sensitivity and genome instability. GenePattern yielded a score of -3.73 –this indicates that FAN1 was up-regulated in Responsive Cell-lines. The up-regulation of FAN1 in Responsive cell lines makes sense, since cancer cells want to replicate without any problems that would attract apoptotic signals (such as DNA damage). Now, with Topotecan, TOP1 is blocked, therefore any DNA abnormality produced by the up-regulation of FAN1, would no longer go unnoticed, and the body would produce the necessary apoptotic signals.

Maria Barrios 5/8/14 10:23 PM

Deleted: According to the *Molecular Mechanisms of Sorafenib Action in Liver Cancer* research paper, "RACGAP1 was specifically down-regulated". This gene had a negative score in the GenePattern output file, which suggests that the gene is down-regulated in Responsive cancer patients upon treatment with Sorafenib –a fact that agrees with the literature. RACGAP1 encodes Rac GTPase-activating protein 1, which belongs to the GTPase-activating protein (GAP) family (Rac). This protein plays a regulatory role in controlling cell growth and differentiation of hematopoietic cells. That RACGAP1 is down-regulated upon treatment with Sorafenib allows us to deduce that RACGAP1 promotes cell growth –for this reason, the inhibition of RACGAP1 helps to stop the proliferation of cancer cells by inhibiting cell growth.

Maria Barrios 5/8/14 10:23 PM

Deleted: TOP1 is down-regulated in Responsive patients upon treatment with Topotecan (hence its negative score on GenePattern).

Maria Barrios 5/8/14 10:24 PM

Deleted: Now, depletion of FAN1 causes DNA damage sensitivity and genome instability, which is the case for the patient(s) treated with Topotecan. GenePattern yielded a score of -3.7263 for FAN1. This down-regulation of FAN1 indicates that Topotecan is effectively forcing the cancer cells to die due to DNA damage, etc.

GNEA and DAVID Discussion

Erlotinib

For the first drug, Erlotininb, the pathway and gene analysis produced very interesting results. In GSEA, the most significant enriched pathway was ***Pathways in Cancer***. Since the drug treats NSCLC, pancreatic cancer, and many other types of cancer, this was very fitting. The genes in this family are involved in the cancer biology and are seen mostly in the following gene families: oncogenes, transcription factors, cytokines, growth factors, translocated cancer genes, and protein kinases. There were 9 genes in this overlap with the lowest p-value of 9.54E-5. The next most significant pathway was ***Cell adhesion molecules (CAMs)***, with 6 genes and a p-value of 1.05E-4. This pathway was also seen in DAVID, though it had a less significant p-value (4.4E-2). Cell adhesion molecules are glycoproteins on the cell surface that play a role in processes such as hemostasis, the immune response, inflammation, embryogenesis, and development of neuronal

tissue. Most relevantly, cancer metastasis tumors use cell adhesion to produce new tumors in the body, which spread throughout the circulatory system. During metastasis, cell-adhesion and cell-migration dysfunction allows cells to migrate, and focal adhesion, where cells form integrin mediated attachment sites, allows cells to pull forward. Interestingly, **Focal adhesion** is another gene set that was enriched in our analysis containing 6 genes, but had a lower significance. Aside from their relationship with migrating cells, focal adhesions generally control cell behavior with the regulatory effects of extracellular matrix (ECM) adhesion. Again, this relates to another pathway seen in both GSEA and DAVID, ***ECM-receptor interaction***. The ECM is the extracellular part of the cell composed of macromolecules that control cell activities such as adhesion, migration, differentiation, proliferation, and apoptosis, as well as provide structural support. In cancer, there is increased synthesis of ECM components or release of ECM cleavage products which contain many growth factors such as FGF or VEGF which spreads cancer growth. The ECM creates a niche for tumor formation and its components help spread the cancer (Zent, 2010). This pathway has 6 genes, with the second lowest p-value in GSEA, and the lowest p-value in DAVID, despite a slightly higher p-value of 7.6E-3. ***Glycosphingolipid biosynthesis*** is another pathway shared between GSEA and DAVID with 3 genes in overlap. Glycosphingolipids (GSLs) are types of glycolipids that have the amino alcohol sphingosine, and its biosynthesis, degradation, and intracellular transport are highly regulated. They can be highly expressed in tumors, causing an antibody response in which they act as antigens. GSLs can be adhesion molecules in tumor cell metastasis, and can also control signal transduction in tumor growth and movement, making it a target for cancer therapy (Hakomori, 1997). The last pathway seen in both analysis tools was ***RIG-I-like Receptor Signaling Pathway*** with 4 genes, and the 5th most significant pathway in GSEA and the least most significant pathway in DAVID. RIG-I-like receptors act as a sensor for viruses and regulate immune responses. Studies have shown evidence that RLR activation can cause rapid apoptosis in many cancer cell. For example, in dsRNA-transfected breast cancer cells, RLR activation initiated 49 extrinsic and intrinsic apoptotic signaling pathways (Yang, 2013). GSEA and DAVID showed similar overlapping and significant pathways. GSEA is a more sensitive tool than DAVID, which may explain why there are more significant pathways and why most pathways have a more significant p-value.

Erlotinib is an EGFR inhibitor drug and accordingly, **EGFR** was one of the most significant enriched genes. It is seen in four of the pathways: Pathways in Cancer, Adherens Junction, Focal Adhesion, and Bladder Cancer. EGFR is a cell-surface receptor in the epidermal growth factor receptor and known oncogene. Mutations cause EGFR up-regulation in many cancers. The mutations produce constant activation, which in turn produces uncontrolled cell division. Erlotinib directly targets the EGFR which prevents activation of the signaling pathways and improves response rates in selected NSCLC patients. The most common mutations associated with sensitivity to EGFR TKIs are exon 19 deletions and the L858R point mutation –these have response rates of >70% in patients treated with Erlotinib (Lung Cancer Mutation Panel). EGFR was found to act as a strong prognostic indicator in head and neck, ovarian, cervical, bladder and oesophageal cancers (Nicholson, 2001). **CDH1** or cadherin 1, type 1/E-cadherin (epithelial) was seen as the most significant gene and was present in the following pathways: Pathways in Cancer, Adherens Junction, CAMs, and Bladder Cancer. It is a member of the cadherin superfamily and is a calcium-dependent cell-cell adhesion glycoprotein. Mutations are correlated with gastric, breast, colorectal, thyroid, and ovarian cancers. Loss of function or expression of CDH1 is seen to contribute to cancer progression by increasing proliferation, invasion, and metastasis. A study has shown that methylation of E-cadherin promoter is associated with risk of lung cancer (Zeng, 2013). **LAMA5** or laminin alpha 5 mediate the attachment, migration, and organization of cells into tissues during embryonic development by interacting with other extracellular matrix components. It is seen in three pathways: Pathways in Cancer, ECM-Receptor Interaction, and Focal Adhesion. A study demonstrated frequent epigenetic inactivation of LAMA5-encoding

genes in lung cancers (Rani, 2013). Other prominent genes this pathway include **ITGA2B**, **CASP8**, **ERBB2**, and more.

Sorafenib

One significant pathway which overlapped with Sorafenib is ***Focal Adhesion*** pathway. The pathway consists of total 201 genes with 7 genes that overlapped with the inputted gene set resulting in a p-value of $1.05e^{-4}$. The pathway k/K color bar has a shade of very light green portraying a significant q- value of $6.52e^{-3}$ which is less than 0.05. The analysis from DAVID consisted of focal adhesion pathway as well, which further proves the importance of this pathway in correlation with Sorafenib. The results in DAVID were similar (portray a 7 gene overlap) but they provide a p-value of $9.5e^{-2}$. In general, Focal Adhesions are groups of actin filaments that are attached to the transmembrane receptors of the integrin family. They are located at the contact points of the cell- extracellular matrix. These cell matrix adhesions are involved in various biological processes such as cell proliferation, cell motility, cell differentiation, cell survival and gene expression regulation. A key mediator of integrin signaling is **FAK** (focal adhesion kinase) which is a cytoplasmic tyrosine kinase. In the article about signal transduction by focal adhesion kinase in cancer, it was described how FAK plays a huge role in tumor progression and metastasis. The FAK- integrin signaling has been shown to promote tumorigenesis by activating signaling pathways through phosphorylation and protein- protein interactions (Zhao et. al, 2009). Other research also indicated that FAK is a relevant target of the drug Sorafenib since it's knockdown slightly prevents the inhibitory efforts of the drug on cell migration and actin polymerization. Sorafenib dephosphorylates FAK which overall plays a big part tumor progression due to its regulation of cancer cells and their microenvironments (Xargay et. al, 2013).

Another significant pathway is the ***MAPK signaling pathway*** which consisted of 267 numbers of genes, of which 7 overlapped with Sorafenib. The pathway did not result in the DAVID tool analysis but in GSEA, it proved to be significant since it has a 5.88 e^{-4} p-value and k/K color bar has a shade of very light green portraying a significant q- value of $1.37e^{-2}$.. MAPK (mitogen – activated protein kinase) pathways are high conserved kinases modules that are involved in various cellular functions such as cell proliferation, differentiation, migration and apoptosis. They connect the extracellular signals to the machinery that control these specific cellular functions. The MAPK pathways are composed of a three- tier kinase module where the **MAPKKK** (mitogen- activated protein kinase kinase kinase) activates **MAPKK** (mitogen-activated protein kinase kinase) which in turn activates the **MAPK** (Dhillon et. al, 2007). To date, Mammals have been characterized to express at least four distinctly regulated groups of MAPKS which includes the following: extracellular signal – related kinases (ERK) -1/2, Jun amino- terminal kinases (JNK1/2/3), p38 proteins (p38alpha/beta/gamma/delta) and ERK5, which are activated by specific MAPKKs. But the complexity very high and is increasing because each MAPKK can be activated by more than one MAPKKK (Dhillon et. al, 2007). Since the pathway helps regulate numerous cellular functions associated with cell division, alterations in the pathway have been associated with various forms of cancer as of today. The drug Sorafenib has be proven to inhibit the MAPK pathway, specifically in malignant peripheral nerve sheath cells (**MPNST**). Since the drug inhibits growth as well as the MAPK pathway in patients with MPNST, research states that this drug will prove to be an effective therapy (Ambrosini et. al, 2008). Some findings also suggest that the drug will be a good tool to help manage Barrett's associated dysplasia and adenocarcinoma. They stated that Sorafenib essentially nullifies the MAPK activation resulting in a significant cell growth inhibition in the Barrett's esophageal adenocarcinoma cell line (Keswani et. al, 2008).

Furthermore, two specific genes from the Sorafenib gene set that were present in the most pathways are **EGF** and **MAPK10**. EGF and MAPK10 both appear in about 5 different pathways, each of which include the Focal Adhesion Pathway and MAPK pathway discussed earlier. The

EGF receptor, once activated, results in autophosphorylation of key tyrosine residues which allows proteins to bind through their domains and activate the downstream signaling cascades. These signaling cascades include the **RAS**/extracellular signal regulated kinase (**ERK**) pathway, the phosphatidylinositol 3 – kinase (**PI3**) pathway and the Janus kinase/ Signal transducer and activator of transcription (**JAK/STAT**) pathway. All three pathways act together and in a coordinated way to help promote cell survival, hence, EGF is very essential (Hooper, 2014). The MAPK10 stands for mitogen activated protein kinase10 which is a member of the MAPK family. This protein is a neuronal – specific form of c-Jun N- terminal kinases (JNKs). Upon phosphorylation, the MAPK10 regulate roles in the signaling pathways during neuronal apoptosis (**NCBI** gene). It overall responds to activation of environmental stress and pro- inflammatory cytokinesis by phosphorylating a number of transcriptions factors and is important because it is required for stress – induced neuronal apoptosis.

Topotecan

For Topotecan, the most significant pathway in GSEA was ***Focal adhesion*** with 11 genes and a p-value of 1.1E-5. It was the second most significant pathway in DAVID with a higher p-value, as we've seen with most pathways in DAVID. This pathway was described earlier when it was seen in the drug Erlotinib. Another pathway for this drug that was also seen in Erlotinib was ***Cell adhesion molecules***, with 7 genes. Adhesion molecules seem to play a larger role in this drug than in Erlotinib, with more enriched genes in the two adhesion pathways. The second most significant pathway in GSEA was ***Regulation of Actin Cytoskeleton*** with 11 genes and a p-value of 2.15E-5. It was the most significant pathway in DAVID with a count of actually 12 genes. The actin cytoskeleton is a structural support skeleton in the cell's cytoplasm. Localized polymerization of actin filaments is the force behind cancer cell migration and it has been seen that molecules that link migratory signals to the actin cytoskeleton are up-regulated in metastatic cancer cells. Key regulatory proteins of the actin cytoskeleton such as WASP family proteins, Arp2/3 complex, LIM-kinase, cofilin, and cortactin could be involved (Yamaguchi, 2007). The ***MAPK Signaling Pathway*** is the next most significant pathway in GSEA with 12 genes and a p-value of 3.16E-5. The mitogen-activated protein kinase (MAPK) cascade is a chain of proteins in the cell that communicates a signal from a receptor on the surface of the cell to the DNA in the nucleus of the cell. It is involved in many functions including cell proliferation, differentiation and migration. The proteins communicate by adding phosphate groups to a neighboring protein, which acts as an "on" or "off" switch, and if there are mutations in a proteins, it gets fixed in an "on" or "off" position, which is necessary for the development of many cancers. The ***Chemokine Signaling Pathway*** was the 4th most significant pathway in GSEA with 9 genes. Chemokines are small cytokines or signaling proteins that induce chemotaxis and play a role in regulating immune cell recruitment. De-regulated expression and activity of chemokine signaling pathways have been demonstrated in cancer progression. These molecules also seem to regulate angiogenesis and epithelial cell growth and survival, making them important for regulating the tumor microenvironment. ***Snare Interactions in Vesicular Transport***, though a less significant pathway with 4 genes, was seen in both GSEA and DAVID. SNARE proteins are small and abundant, and mediate vesicle fusion.

RAC2 or ras-related C3 botulinum toxin substrate 2 is the most significant gene seen in this drug. It is seen in almost all of the enriched pathways. It is a small signaling G protein, and is a member of the Rac subgroup of the Rho family of GTPases. It regulates events such as cell growth, cytoskeletal reorganization, and the activation of protein kinases, and function as binary switches in the regulation of various cellular activities. It has been implicated as contributing to a variety of cancers. Activating mutations of RAC GTPases are found in low frequency in many cancers. Rac2 also regulates the actin cytoskeleton during breast cancer metastasis (Li, 2013). Another

significant gene is **VAV3** or vav3 guanine nucleotide exchange factor (GEF) which is a GEF for Rho family GTPases and it activates pathways leading to actin cytoskeletal rearrangements and transcriptional alterations. It is seen in the following four pathways: Focal Adhesion, Regulation of Actin Cytoskeleton, Chemokine Signaling Pathway and Gamma R Mediated Phagocytes. Vav3-mediated signaling pathway has been demonstrated to be a possible target for prostate cancer metastasis and as a useful marker for predicting the outcomes of patients with gastric cancer (Lin, 2012). A group of significant genes are the mitogen-activated protein kinase genes: **MAPK10**, **MAP3K3**, and **MAP4K1** and are all associated with the MAPK pathway which is implicated in cancer. MAPK10, also known as JNK3, is the most significant of these and can be involved in proliferation, differentiation, transcription regulation and development. It has been identified as a novel epigenetic marker for kidney cancer (Yoo, 2011).

V. Conclusion

It was found that our algorithm(s) in finding top prognostic genes (using WEKA) lacked the capability to find the main gene(s) targeted by each drug, but it was capable of identifying the most differentially expressed, cancer-related genes. This limitation resulted in the need for the "hand-picking" of certain prognostic genes as discussed earlier, after researching the main targets of each drug. This hand-picking of prognostic genes ultimately benefitted our classifying algorithms by yielding higher accuracies, as expected. For Erlotinib, based on the analysis of GSEA, DAVID, and our research, EGFR was found to be unanimously important for the action of Erlotinib in treating cancer, as it is an EGFR inhibitor drug. Note that GenePattern was especially useful for confirming the inhibitory effects on EGFR by Erlotinib, as well as confirming the effects of the other two drugs on genes such as TOP1 and VEGFA. For Sorafenib and Topotecan, the GSEA DAVID results illustrated that Focal Adhesion pathways played a significant role, as well as the MAPK pathway; both of which have been proven to play a key role in the progression of cancer as shown in the literature. Note that Topotecan did not yield as significant results as the other two drugs in the gene set enrichment analyses. The enrichment results were poorer using the same N number of significant genes as the other two drugs for Topotecan. This leads us to believe that the effects of treatment with Topotecan are not as strong compared to those of Erlotinib and Sorafenib. From what we have seen in our project, our analysis techniques are effective in confirming the desired effects of various cancer therapies (Erlotinib, Sorafenib, and Topotecan) as well as calling for other various therapies to be used in conjunction with these three drugs to maximize the effectiveness of targeted therapy.

Furthermore, as noted in the ***Methods*** section, given the drug response data for each cell line, the cell lines were divided into the two groups, responsive and nonresponsive, in such a way that the lowest 25% of the negative scores were classified as responsive. To further enhance the validity of our results, more careful and thorough methods for creating the two groups could be tried (i.e. unsupervised clustering, etc.). Refer to Drug Response E, S, and T. This would allow us to implement the same set of steps used in this project on different divisions of drug response scores in order to find the ideal division of cell lines that would ultimately lead to the highest attainable accuracies in our classifier tests.

VII. Supplements

Suplement 1: Confusion Matrices

	5 Cross Validation			10 Cross Validation		
Logistic Regression		**NR**	**R**		**NR**	**R**
	NR	55	13	**NR**	60	8
	R	5	18	**R**	6	17
Naïve Bayes		**NR**	**R**		**NR**	**R**
	NR	59	9	**NR**	60	8
	R	2	21	**R**	2	21
Neural Networks		**NR**	**R**		**NR**	**R**
	NR	65	3	**NR**	65	3
	R	6	17	**R**	5	18

Supplement 1 *–This supplements documents the confusion matrices of all the algorithms run on Erlonitib data. It is from these values that the TP, TN, FP, FN rate were obtained from.*

Suplement 2: Confusion Matrices

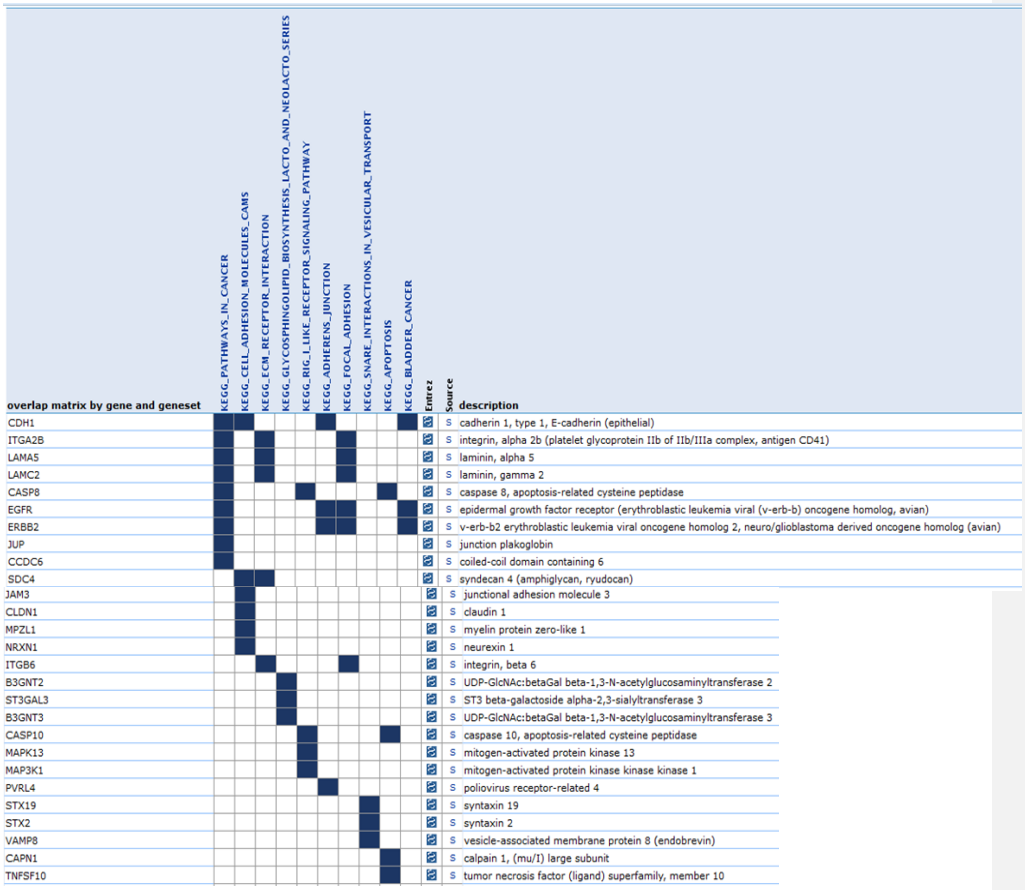
	5 Cross			10 Cross		
Logistic Regression		**R**	**NR**		**R**	**NR**
	R	19	4	**R**	18	5
	NR	6	62	**NR**	4	64
Naïve Bayes		**R**	**NR**		**R**	**NR**
	R	17	6	**R**	19	4
	NR	0	68	**NR**	0	68
Neural Networks		**R**	**NR**		**R**	**NR**
	R	20	3	**R**	21	2
	NR	0	68	**NR**	0	68

Supplement 2 *–This supplements documents the confusion matrices of all the algorithms run on Sorafenib data. It is from these values that the TP, TN, FP, FN rate were obtained from.*

Suplement 3: Confusion Matrices

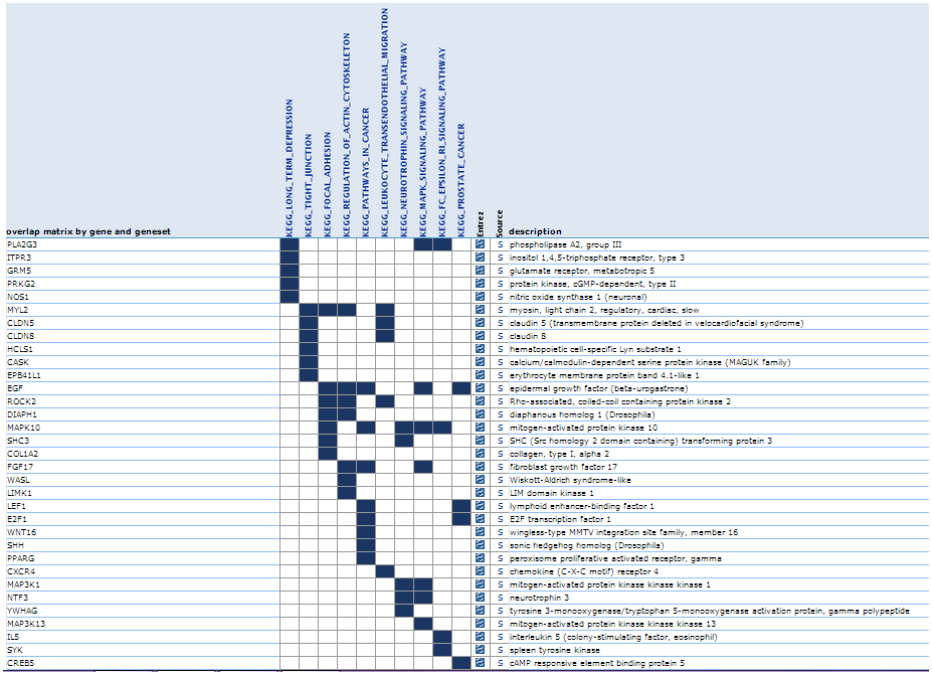
	5 Cross			10 Cross		
Logistic Regression		**R**	**NR**		**R**	**NR**
	R	15	8	**R**	15	8
	NR	14	54	**NR**	13	55
Naïve Bayes		**R**	**NR**		**R**	**NR**
	R	20	3	**R**	19	4
	NR	3	65	**NR**	3	65
Neural Networks		**R**	**NR**		**R**	**NR**
	R	13	10	**R**	16	7
	NR	3	65	**NR**	3	65

Supplement 3 *–This supplements documents the confusion matrices of all the algorithms run on Topotecan data. It is from these values that the TP, TN, FP, FN rate were obtained from.*

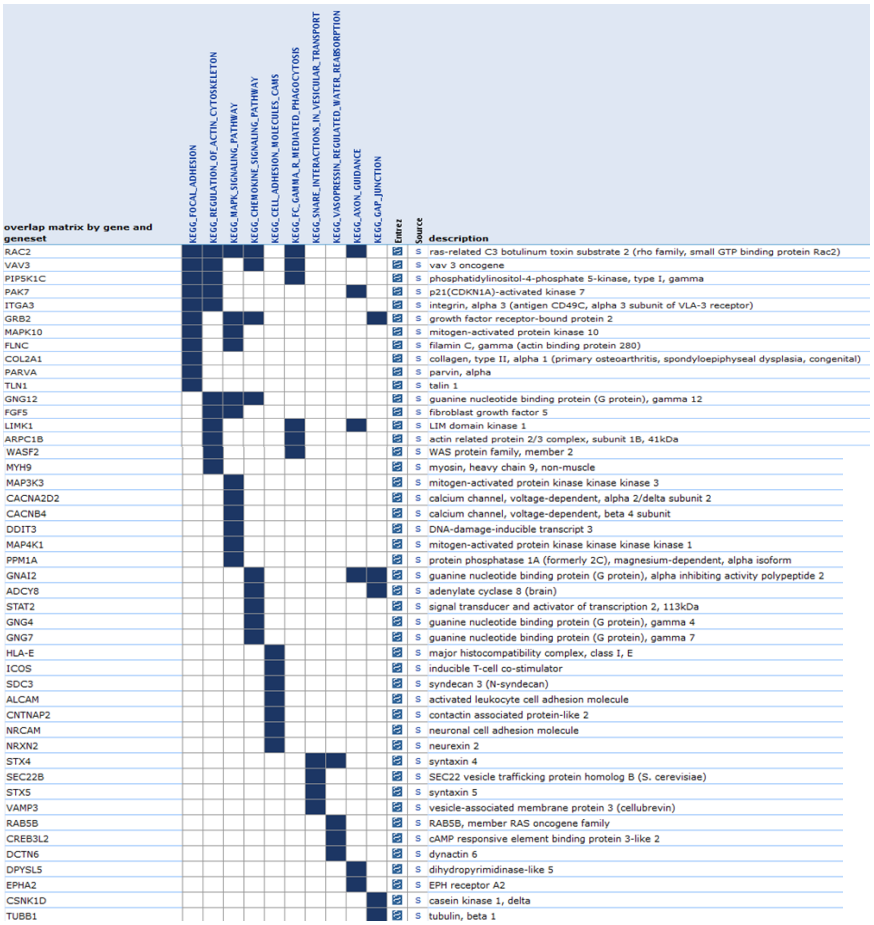


overlap matrix by gene and geneset	KEGG_PATHWAYS_IN_CANCER	KEGG_CELL_ADHESION_MOLECULES_CAMS	KEGG_ECM_RECEPTOR_INTERACTION	KEGG_GLYCOSPHINGOLIPID_BIOSYNTHESIS_LACTO_AND_NEOLACTO_SERIES	KEGG_RIG_I_LIKE_RECEPTOR_SIGNALING_PATHWAY	KEGG_ADHERENS_JUNCTION	KEGG_FOCAL_ADHESION	KEGG_SNARE_INTERACTIONS_IN_VESICULAR_TRANSPORT	KEGG_APOPTOSIS	KEGG_BLADDER_CANCER	Entrez	Source	description
CDH1	■	■				■				■		S	cadherin 1, type 1, E-cadherin (epithelial)
ITGA2B	■		■				■					S	integrin, alpha 2b (platelet glycoprotein IIb of IIb/IIIa complex, antigen CD41)
LAMA5	■		■				■					S	laminin, alpha 5
LAMC2	■		■				■					S	laminin, gamma 2
CASP8	■				■				■			S	caspase 8, apoptosis-related cysteine peptidase
EGFR	■					■	■			■		S	epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian)
ERBB2	■					■	■			■		S	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)
JUP	■											S	junction plakoglobin
CCDC6	■											S	coiled-coil domain containing 6
SDC4		■	■									S	syndecan 4 (amphiglycan, ryudocan)
JAM3		■										S	junctional adhesion molecule 3
CLDN1		■										S	claudin 1
MPZL1		■										S	myelin protein zero-like 1
NRXN1		■										S	neurexin 1
ITGB6			■				■					S	integrin, beta 6
B3GNT2				■								S	UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 2
ST3GAL3				■								S	ST3 beta-galactoside alpha-2,3-sialyltransferase 3
B3GNT3				■								S	UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 3
CASP10					■				■			S	caspase 10, apoptosis-related cysteine peptidase
MAPK13					■							S	mitogen-activated protein kinase 13
MAP3K1					■							S	mitogen-activated protein kinase kinase kinase 1
PVRL4						■						S	poliovirus receptor-related 4
STX19								■				S	syntaxin 19
STX2								■				S	syntaxin 2
VAMP8								■				S	vesicle-associated membrane protein 8 (endobrevin)
CAPN1									■			S	calpain 1, (mu/I) large subunit
TNFSF10									■			S	tumor necrosis factor (ligand) superfamily, member 10

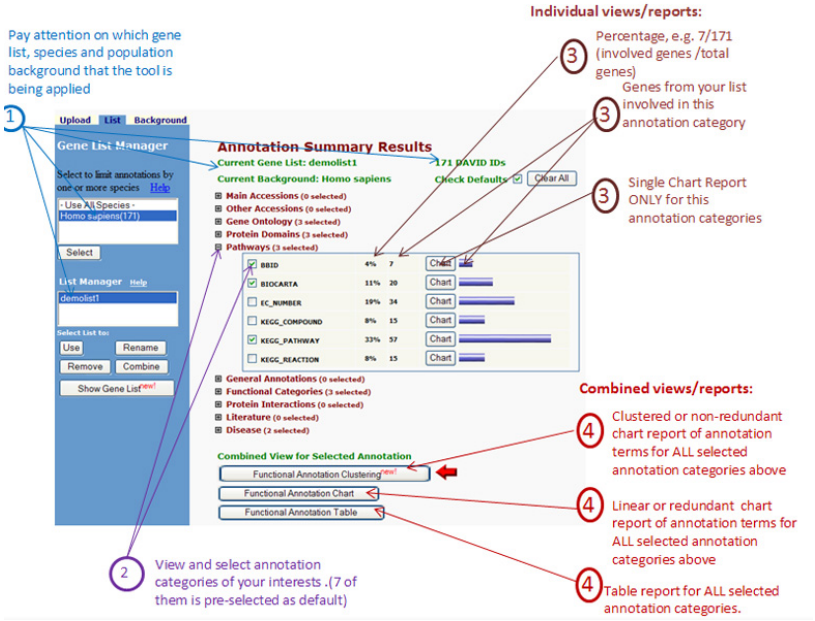
Supplement 4 *–This supplements shows the Eroltinib Enriched Gene List from GSEA*



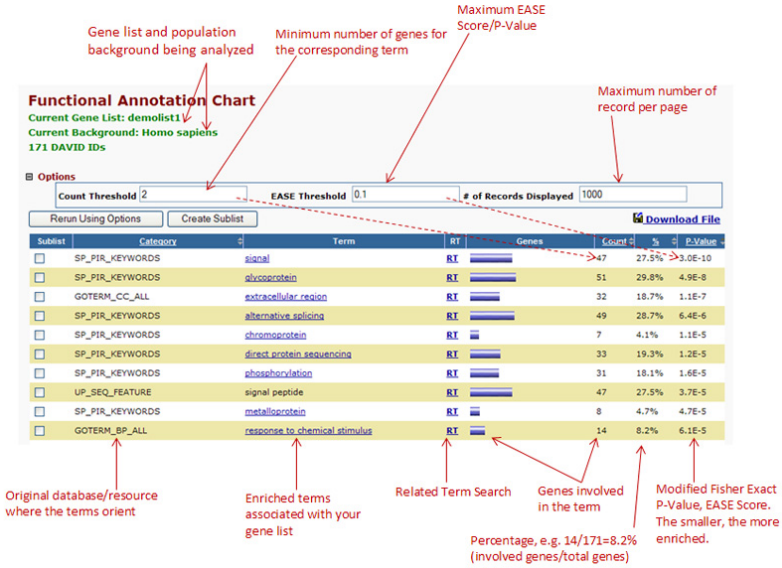
Supplement 5 *–This supplements shows the Sorafenib Enriched Gene List from GSEA*



Supplement 6 *–This supplements shows the Topotecan Enriched Gene List from GSEA*



Supplement 7 *– Functional Annotation Summary Results In DAVID*



Supplement 8 *– Functional Annotation Chart in DAVID*

MATLAB Code

```
% PROGRAM 1
% Load the drug response data
% Note all cell lines are the same for
the three drug tests
load('erlonitibCellLines.mat')
erlonitibData = [-43.08368 -82.32113
5.12151 -68.47551 -51.7345 -26.41511 -
62.86543 0.1386896 1.516382 -42.60942 -
5.432577 -44.40665 -3.865206 -79.92746 -
94.3258 -19.15372 -1.00116 -85.47997
1.746782 10.36488 -0.8061247 -26.11557 -
0.3308865 -41.24062 -65.39386 -63.64347 -
4.80092 -2.104016 14.14728 -35.96056
44.99139 -40.98888 4.081686 -37.31577 -
53.94451 -5.396407 0.01895391 -54.20909 -
38.7101 14.6441 -30.90703 -42.25717 -
35.42674 -76.51304 -47.11651 14.34459 -
21.01797 -20.4345 -36.80983 -57.39882 -
33.73491 -46.71859 -44.45197 -33.84303 -
21.09889 -54.4102 -6.610373 -32.86922
2.232733 -64.62232 -82.49171 -19.36131 -
44.19146 -0.3340859 -58.4826 5.463198 -
33.21539 -84.20258 -94.33957 -67.59582 -
28.6947 0.2885649 0.7389826 -27.6804 -
0.8786044 -14.86487 -6.69327 -3.718819
5.218784 -86.95715 -26.91757 6.808955 -
0.4910699 2.088601 -38.7407 -91.42735 -
64.72824 -4.977474 -28.80327 -8.991793
19.12427];
sorafenibData = [-81.19079 -56.24479
3.101964 -17.68352 -6.615608 -64.97485
20.18308 -74.41914 -50.36484 8.653348 -
5.671657 -13.76845 -0.6293455 23.1311 -
32.82162 -10.75931 6.502373 10.03283 -
11.43284 -55.64251 7.145658 -5.474973 -
34.5413 -8.652964 -77.71759 -5.610258 -
17.78753 -55.818 5.99298 1.471072
7.116985 -6.248536 -67.52592 1.488508 -
15.85543 1.741804 10.16371 -52.4842 -
16.66059 5.161666 -2.425484 -35.4911 -
27.16766 -27.96682 -1.886012 -0.1828269 -
56.34416 -14.27553 -7.855617 5.19112 -
8.653709 -13.51376 -12.49035 4.9415
13.22719 -59.09224 1.621373 -57.53521 -
46.73141 -56.85644 -30.66334 -72.72377 -
40.5203 -41.05743 -79.5486 2.888413
7.877636 5.910589 -0.6731039 -47.15472 -
24.27414 -12.1934 -30.92502 -19.86929 -
9.770709 0.5375397 1.187648 15.43999 -
49.56928 -0.5788 2.098569 -12.18904 -
60.79331 1.463271 -3.646381 0.9068974 -
3.362509 2.989052 5.219263 -38.01049
3.224429];
topotecanData = [-97.29373 -93.20826 -
75.59962 -80.88937 -94.43704 -73.7702 -
77.75194 -96.28535 -85.42012 -87.8224 -
87.29766 -87.60235 -77.94633 -72.09923 -
73.12481 -88.17322 -88.44751 -65.96295 -
41.92655 -72.97452 -83.39963 -52.90018 -
87.23193 -88.59472 -92.09533 -103.2793 -
88.48399 -95.76366 -103.4716 -94.60352 -
94.09624 -79.78586 -93.42007 -96.32133 -
47.36174 -59.73095 -7.641711 -95.31171 -
29.63369 -87.39281 -96.20318 -85.49323 -
74.44997 -92.06783 -85.43598 -101.232 -
80.98301 -69.14299 -111.2544 -73.18568 -
92.67866 -110.3661 -83.23412 -74.92393 -
93.64024 -95.95497 -46.92022 -93.85157 -
95.17377 -105.0908 -94.93001 16.234 -
96.58159 -66.09457 -95.48691 -78.68719 -
82.77236 -97.71704 -77.11539 -95.8827 -
84.75957 -91.12349 -82.35303 -61.7276 -
57.72153 -79.45873 -48.32278 -45.00325 -
43.58299 -91.1289 -85.76198 -67.64258 -
98.38835 -55.16548 -76.72397 -85.14671 -
91.0312 -77.04993 -83.34422 -70.46362 -
47.62931];

% Sort the drug response data in
ascending order and get the indices of
% these places
[sortedEData ePlaces] =
sort((erlonitibData));
[sortedSData sPlaces] =
sort((sorafenibData));`
[sortedTData tPlaces] =
sort((topotecanData));

% Assign 1's (Responsive) to the lowest
25% of the data (most negative
% %inhibitions) and 0's (Nonresponsive to
the rest)
erlonitibClass = zeros(1,91);
for i = 1:91
if i <=23
    erlonitibClass(ePlaces(i))=1;
    topotecanClass(tPlaces(i))=1;
    sorafenibClass(sPlaces(i))=1;
else
    erlonitibClass(ePlaces(i))=0;
    topotecanClass(tPlaces(i))=0;
    sorafenibClass(sPlaces(i))=0;
end
end
% Now change the cls file and upload to
GenePattern
% Then we use the gene pattern output
file to give us [drug
name]GenesRanked.mat

% PROGRAM 2
% Note that this section of the code was
changed every time it was run in
% order to find the inflection points
(changing N) on WEKA

% Choose number of significantly genes
expressed for WEKA
N = 140;

% Finding the top N genes in the GCT file
for i = 1:N

findGCT(i)=find(strcmp(GCTText(:,1),erlon
itibGenesRanked((i+1),2)));
end
% The erlonitibGenesRanked vcariable was
changed per trial for each drug

% Create matrices to compile new GCT file
for WEKA
topNGCTData=zeros(N,91);
topNGCTText=cell(N,2);
for i = 1:N
topNGCTData(i,:) = GCTData((findGCT(i)-
1),3:93);
topNGCTText{i,1} = GCTText{findGCT(i),1};
topNGCTText{i,2} = GCTText{findGCT(i),2};
```

```
end


% PROGRAM 3
% Create final (top significant genes)
CLS files to input into WEKA to test the
accuracies of
% various classification algorithms
(Naive Bays, Neural Networks, and
% Losistic Regression)

% 16 CHOSEN ERLONITIB
for i = 1:11

finalErlonitib(i)=find(strcmp(topNGCTText
(:,1),erlonitibTop50LogisticProbes(i,1)))
;
end
for i = 1:5

finalErlonitib(11+i)=find(strcmp(topNGCTT
ext(:,1),erlonitibTop50LogisticProbes(45+
i,1)));
end

for i = 1:16

erlonitibFinalData(i,:)=topNGCTData(final
Erlonitib(i),:);

erlonitibFinalText{i,1}=topNGCTText{final
Erlonitib(i),1};

erlonitibFinalGeneSymbols{i,1}=topNGCTTex
t{finalErlonitib(i),2};
end

% 15 CHOSEN SORAFENIB
for i = 1:7

finalSorafenib(i)=find(strcmp(topNGCTText
(:,1),sorafenibTop100LogisticProbes(i,1))
);
end

for i = 1:8

finalSorafenib(7+i)=find(strcmp(topNGCTTe
xt(:,1),sorafenibTop100LogisticProbes(92+
i,1)));
end

for i = 1:15

sorafenibFinalData(i,:)=topNGCTData(final
Sorafenib(i),:);

sorafenibFinalText{i,1}=topNGCTText{final
Sorafenib(i),1};

sorafenibFinalGeneSymbols{i,1}=topNGCTTex
t{finalSorafenib(i),2};
end

% 21 CHOSEN TOPOTECAN
for i = 1:8

finalTopotecan(i)=find(strcmp(topNGCTText
(:,1),topotecanTop140LogisticProbes(i,1))
);
end
for i = 1:13

finalTopotecan(8+i)=find(strcmp(topNGCTTe
xt(:,1),topotecanTop140LogisticProbes(127
+i,1)));
end

for i = 1:21

topotecanFinalData(i,:)=topNGCTData(final
Topotecan(i),:);

topotecanFinalText{i,1}=topNGCTText{final
Topotecan(i),1};

topotecanFinalGeneSymbols{i,1}=topNGCTTex
t{finalTopotecan(i),2};
end
```

Works Cited

Amrosini, G., Cheema, H.S., Seelman, S., Teed, A., Sambol, E.B., Singer, S., & Schwartz, G.K. (2008). Sorafenib inhibits growth and mitogen-activated protein kinase signaling in malignant peripheral nerve sheath cells. Mol Cancer Ther, 7(4):890-6. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18413802

Cervello, Melchiorre, Dimcho Bachvarov, Nadia Lampiasi, Antonella Cusimano, Antonia Azzolina, James A. McCubrey, and Giuseppe Montalto. "Molecular Mechanisms of Sorafenib Action in Liver Cancer Cells." N.p., 1 Aug. 2012. Web. 27 Apr. 2014.

Dijk, M. V., Goransson, S. A., & Stromblad, S. (2013). Cell to extracellular matrix interactions and their reciprocal nature in cancer. *Experimental Cell Research*, *319*(11), 1663-1670.

Dhillon, A.S, Hagan, S., Rath, O., Kolch, W., (2007).MAP kinase signalling pathways in cancer. Oncogene, 26, 3279-3290. Retrieved from http://www.nature.com/onc/journal/v26/n22/full/1210421a.html

Erlotinib: MedlinePlus Drug Information. (n.d.). *U.S National Library of Medicine*. Retrieved April 16, 2014, from http://www.nlm.nih.gov/medlineplus/druginfo/meds/a605008.html

Hakomori, S., & Zhang, Y. (1997). Glycosphingolipid antigens and cancer therapy. *Chemistry & Biology*, *4*(2), 97-104.

Hooper, C.(2014). Epidermal growth factors and cancer. Abcam. Retrieved from http://www.abcam.com/index.html?pageconfig=resource&rid=10723

Li, H., Jin, T., Xu, X., Hao, X., Guo, H., Wang, Y., et al. (2013). Association between GÎ±i2 and ELMO1/Dock180 connects chemokine signalling with Rac activation and metastasis. *Nature Communications* , *4*, 1706.

Lin, K., Wang, L., Hseu, Y., Fang, C., Yang, H., Kumar, K. J., et al. (2012). Clinical Significance of Increased Guanine Nucleotide Exchange Factor Vav3 Expression in Human Gastric Cancer. *Molecular Cancer Research*, *10*(6), 750-759.

Lung Cancer Mutation Panel (EGFR, KRAS, ALK). (n.d.). *Lung Cancer Mutation Panel (EGFR, KRAS, ALK)*. Retrieved April 26, 2014, from https://www.questdiagnostics.com/testcenter/testguide.action?dc=TS_LungCancerMutation_Panel

Keswani, R.N., Chumsangsri, A., Mustafi, R., Delgado, J., Cohen, E.E, & Bissonnette, M. (2008). Sorafenib inhibits MAPK- mediated proliferation in a Barrett's esophageal adenocarcinoma cell line. Dis Esophagus, 21(6):514-21. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18840136

MAPK10 mitogen-activated protein kinase 10[Homo sapiens (human)]. NCBI Gene. (2014). Retrieved from http://www.ncbi.nlm.nih.gov/gene/5602

Nicholson, R., Gee, J., & Harper, M. (2001). EGFR and cancer prognosis. *European Journal of Cancer*, *37*, 9-15.

"Rac GTPase Activating Protein 1." *GeneCards*. Weizmann Institute of Science, n.d. Web. 27 Apr. 2014.

Rani, V., Mccullough, M., & Chandu, A. (2013). Assessment of Laminin-5 in Oral Dysplasia and Squamous Cell Carcinoma. *Journal of Oral and Maxillofacial Surgery*, *71*(11), 1873-1879.

Sorafenib: MedlinePlus Drug Information. (n.d.). *U.S National Library of Medicine*. Retrieved April 16, 2014, from http://www.nlm.nih.gov/medlineplus/druginfo/meds/a607051.html

Tarceva is a once-a-day pill. (n.d.). *TarcevaÂ® (erlotinib) Tablets for Advanced-Stage NSCLC & Advanced-Stage Pancreatic Cancer*. Retrieved April 16, 2014, from http://www.tarceva.com/patient/

Topotecan. (n.d.). *Topotecan*. Retrieved April 16, 2014, from http://www.cancer.org/treatment/treatmentsandsideeffects/guidetocancerdrugs/topotecan

Wilhelm, Scott M., Lila Adnane, Philippa Newell, Augusto Villanueva, Joseph M. Llovet, and Mark Lynch. "Preclinical Overview of Sorafenib, a Multikinase Inhibitor That Targets Both Raf and VEGF and PDGF Receptor Tyrosine Kinase Signaling." N.p., 7 Oct. 2008. Web. 27 Apr. 2014.

Yamaguchi, H., & Condeelis, J. (2007). Regulation of the actin cytoskeleton in cancer cell migration and invasion. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, *1773*(5), 642-652.

Yang, K., Wang, J., Xiang, A. P., Zhan, X., Wang, Y., Wu, M., et al. (2013). Functional RIG-I-like receptors control the survival of mesenchymal stem cells. *Cell Death and Disease*, *4*(12), e967.

Yoo, K. H., Park, Y., Kim, H., Jung, W., & Chang, S. (2011). Identification of MAPK10 as a novel epigenetic marker for chromophobe kidney cancer. *Pathology International*, *61*(1), 52-54.

Xargay- Torrent, S., Lopez-Guerra, M., Montraveta, A., Saborit-Villarroya, I., Rosich, L., Navarro, A.,Perez-Galan, P., Roue, G., Campo, E., & Colomer, D. (2013). Sorafenib inhibits cell migration and stroma-mediated bortezomib resistance by interfering B- cell receptor signaling and protein translation in mantel cell

lymphoma. Clin Cancer Res, 19(3):586-97. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23231952

Zeng, H., Zheng, R., Zhang, S., He, J., & Chen, W. (2013). Lung cancer incidence and mortality in China, 2008. *Thoracic Cancer*, *4*(1), 53-58.

Zhao, J., Guan, J.L. (2009). Signal transduction by focal adhesion kinases in cancer. Cancer Metastasis, 28(1-2):35-49. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19169797