

UC Santa Cruz

UC Santa Cruz Electronic Theses and Dissertations

Title

Tools for Extracting Actionable Medical Knowledge from Genomic Big Data

Permalink

<https://escholarship.org/uc/item/53k1x0f5>

Author

Goldstein, Theodore Charles

Publication Date

2013

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

**TOOLS FOR EXTRACTING ACTIONABLE MEDICAL
KNOWLEDGE FROM GENOMIC BIG DATA**

A dissertation submitted in partial satisfaction
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING

by

Theodore C. Goldstein

June 2013

The Dissertation of Theodore C.
Goldstein is approved:

Professor David Haussler

Professor Josh Stuart

Professor Christopher Benz

Professor Eric Collisson

Tyrus Miller, Ph.D.
Vice Provost and Dean of Graduate Studies

Copyright © by

Theodore C. Goldstein

2013

Table of Contents

List of Figures	v
List of Tables.....	viii
Abstract.....	ix
1 Introduction and Overview.....	1
1.1 Cancer Biomarker-Guided Therapy	4
1.2 How Can the Activitome Improve Cancer Treatment?	7
1.2.1 Improving Surgical Outcomes.....	8
1.2.2 Guiding Therapy Choice	9
1.2.3 Guide Cytotoxic Drug Dosage.....	10
1.2.4 Hormone Molecular-Targeted Drug Therapies.....	15
1.2.5 Kinase Molecular-Targeted Drug Therapies	16
1.2.6 Drug Resistance and Targeted Therapies.....	18
1.2.7 Immune System and Antibody-based Targeted Therapies.....	20
1.3 How Does Cancer Circumvent Current Therapies?.....	22
1.4 How Can Knowledge of Pathways Provide a Foundation for Discovery?	23
1.5 A Pathway-based Method for Inferring Molecular Activity.....	26
1.6 PARADIGM-Derived Features Predict Patient Outcomes More Reliably than Gene Expression.....	30
1.7 Conclusion	31
2 Methods	33
2.1 Background.....	33
2.1.1 Definitions	34
2.1.2 Dichotomous Event	35
2.1.3 Significance Analysis of Microarrays	36
2.1.4 Pearson's R Correlation	38
2.2 OCCAM Method	40
2.2.1 The Risk of Using OCCAM Naively.....	41
2.3 DIPSC Method	43
2.3.1 DIPSC Formally.....	45
2.3.2 DIPSC Self-Correlation and Normalization	46
2.4 DIPSC vs. OCCAM Validation Using Simulated Data.....	46
2.4.1 Other Quality Metrics	49
2.5 Summary	50
3 Clinical-Translational Results.....	52
3.1 SU2C Breast Cancer Drug Sensitivity Activitome.....	52
3.1.1 Claudin-low Activitome Signature vs. Basal Activitome Signature	55
3.1.2 Drug Biomarkers	59
3.2 Activitome Signatures Provide Clues about Cellular Targets	60

3.3	Towards Patient-specific Therapeutic Models.....	60
3.4	Comparing Activitome Signatures Reveals Novel Connections Between Mutations and Drug Response in Luminal Breast Cancers.....	63
3.4.1	Summary of my contributions	65
3.5	TCGA Breast Cancer (BRCA).....	67
3.6	TCGA Colorectal Cancer.....	71
3.7	TCGA Kidney and Renal Cancers	75
3.8	TCGA Lung Adenocarcinoma	77
3.9	TCGA Uterine and Endometrial Cancer	82
3.10	The TP53 PANCAN Activitome Correlation with Drug Resistance.....	84
3.11	Discussion.....	87
4	Future: MedBook Knowledge & Decision Support System.....	89
4.1	The Computational Opportunity	93
4.2	The MedBook Vision.....	94
4.2.1	The Need for Data Integration	95
4.2.2	Biomedical Evidence Graph (BMEG)	97
4.2.3	Molecular Disease Models (MDM) Database	99
4.2.4	Therapeutic plans and Outcomes Database.....	100
4.2.5	Activitome Signatures Database	101
4.2.6	Individual Profiles.....	101
4.2.7	There's an App for That	102
4.2.8	Federated System	106
4.2.9	Validation of New Methods.....	107
4.3	Patient, Health Care Provider And Researcher Use Scenarios	107
4.3.1	Patient and Health Care Provider Use Scenario	108
4.3.2	Clinical Researcher Use Scenario	110
4.3.3	Researcher Discovery Use Case	111
4.3.4	Virtual Tumor Board.....	114
4.4	Current MedBook Status.....	115
4.4.1	Integrating High-throughput Genomic Sequencing into a Medical Context.....	116
4.4.2	Data Analysis Workflows.....	119
4.4.3	Visualization of Data by Dynamic Heatmaps.....	120
4.4.4	OCCAM and DIPSC Tools	122
4.4.5	DipStick Tool	123
4.4.6	UCSC Activitome Signature Database	124
4.5	Business Model	125
4.5.1	Rules of Engagement.....	126
4.5.2	System Implementation Considerations	127
4.6	Conclusion.....	128
	Glossary.....	130
	References	134

List of Figures

Figure 1 Molecular structure of 5-FU compared to the pyrimidines.	12
Figure 2 PARADIGM model for integrative data analysis. A. Factor graph model oriented around a single gene. Hidden states in a tumor sample (purple ellipses) are denoted: genomic copies (G), epigenetic promoter state (E), mRNA transcripts (T), peptide (P), and active protein (A). Regulation gene expression (green ellipses) includes transcriptional (RT), translational (RP), and post-translational (RA) control. Sample data (green circles, constrain gene states through factors (boxes). B. Toy example of a MYC/MAX-associated pathway. Two transcription factors (MYC and MAX) form a complex (MYC/MAX) that is inhibited by PAK2, a protein kinase. MYC/MAX activates two target genes (CCNB1, ENO1) and inactivates a third (WNT5A). C. Single patient data converted to inferred activities for toy simplified pathway. Measurements and inferred levels are higher (red), lower (blue), or comparable (purple) to levels in matched normal. Belief propagation infers the kinase is inactive based on inferred higher activity of MYC/MAX. Figure copyright AACR 2013, used with permission.	28
Figure 3 Central dogma of biology extended with regulators and assays.	30
Figure 4 Activitome Signatures are an n-dimensional vector (only three dimensions shown). Each vector is a point in an n dimensional space where dimension is the difference in activity between two groups. For example, the Aromatase Inhibitor signature compares genes of two groups: the group of samples derived from patients that responded positively to aromatase inhibitors compared to the group of samples that are derived from patients who did not respond to aromatase inhibitors.	34
Figure 5 Pictorial illustration of Correlation. X and Y are a vector of sample data. Each sample is a vector of PARADIGM IPLs.	38
Figure 6 OCCAM Pipeline model. OCCAM finds & ranks differential pathway activity. GI50 depicted lower left (or some other measurement) dichotomizes the cell lines into + and – sensitive/resistant classes. Cell line expression, and copy number are processed through PARADIGM to obtain a score for each pathway concept. Then using SAM as a classifier, OCCAM processes the pathway events detected and the dichotomization to create a differential activity score, depicted as a heatmap lower right.	41
Figure 7 Depiction where samples S3 and S4 are on both sides of the correlation, which confounds the result.	42
Figure 8 Depiction of Bootstrap Sampling, where the process of resampling called a dip is repeated D times, where typically D=1000. Samples in two different groups in a dip do not overlap.	44
Figure 9 OCCAM and DIPSC chart of the evaluation of systematic GNW gene knockout simulation. The chart shows the p-value significance score. The X-axis is simulated genes, sorted by DIPSC p-value. Only genes below the green line have significance of < 0.05. OCCAM (blue line) incorrectly shows 80 genes are significant, whereas DIPSC correctly identifies the nine objectively true gene knockouts whose signatures are significant (those within the solid green box).	49
Figure 10 DIPSC Self-correlations using 100 significant mutated genes in the TCGA dataset. The x-axis is the sample count; the y-axis is the R value. Strength of self-correlation	

improves with sample count and strength of correlation. Strongest signal is luminal A at $R=0.7$, p-value is near zero. Correlations below 0.1 are insignificant (p-value > 0.5).	50
Figure 11 Drug Resistance and Sensitivity Activitome Signature Heatmap. Red indicates sensitive and Blue indicates resistant signatures. Inset boxes list signatures for each of the basal, luminal and claudin low and Her2 breast cancer subtypes.	53
Figure 12 Cell line subtypes have unique network features. In the above figure created by my colleague Sam Ng, each node in the graph represents a different pathway “concept” corresponding to either a protein (circles), a multimeric complex (hexagons), or a cellular process (squares). The size of each node was drawn in proportion to the differential activity score such that larger nodes correspond to pathway concepts with activities more correlated with subtype (basal, luminal, etc) versus non-subtype cell lines. Color indicates whether the concept is positively correlated (red) or negatively correlated (blue) with the basal subtype. Links represent different interactions, including protein-protein level interactions (dashed lines) and transcriptional (solid lines). See text for discussion.....	54
Figure 13 Differential Activities Across Drugs. Identifying pathways that contribute to drug resistance is a critical part of this research. The research found that the ERBB2AMP cell lines were resistant to the aurora kinase inhibitor VX-680 (panel C, upper), and further that sensitivity to this compound was not associated with amplification at 20q13 (AURKA). This raises the possibility that this resistance may be mediated through CCNB1, which is co-regulated with AURKB by FOXM1. Of the four subtypes, ERBB2AMP is the only one that shows substantial downregulation of CCNB1 (panel C). This proposed mechanism is supported by the observation that in primary tumors, CCNB1 gene expression is significantly correlated with AURKB gene expression.	58
Figure 14 DipStick activitome signatures correlation to luminal B vs. luminal A dichotomization activitome signature.....	65
Figure 15 Concordance between TCGA patient and Gray Cell line. Heatmap of non-redundant PARADIGM activities for both cell line and TCGA samples. Cluster dendrogram represents Euclidian distance between samples and was created using Eisen Cluster and drawn using Java Treeview. Each row represents a network feature, each column represents a sample (tumor or cell line). Colored bars below dendrogram indicate sample subtype (upper) and sample cohort (lower). Overall, cell lines and tumors are intermixed, and subtypes tend to cluster together, indicating that the tumors and cell lines share many of the same network features. From PNAS Figure S4 [Heiser 2011], permission granted to authors via http://www.pnas.org/site/aboutpnas/rightperm.xhtml	68
Figure 16 DipStick correlation of TCGA breast cancer mutation, histology, and subtype events.	69
Figure 17 Volcano plot of TCGA BRCA patient subtypes activitome signature (major heading) correlated to Gray cell line drug sensitivity signatures (inset labels). Red indicates the correlation that patients with this cancer subtype will respond to this drug. Purple is anti-correlated, blue dots indicates unlabeled drugs whose p-value is not significant at the p-value < 0.05 level.	70
Figure 18 TCGA Colorectal cancer correlation of correlations of activitome signatures. Gene names indicate mutation activitome signatures, all others are clinical assay activitome signatures or location activitome signatures. Signatures show distinct upper vs. lower colon division. Most mutations correlate with a specific region.	72

Figure 19 TCGA Kidney and Renal Cancer associates by grade. Horizontal access labels are the same as vertical labels. Mutations are annotated _MUT and deletions are annotated _DEL.	75
Figure 20 Lung adenocarcinoma (LUAD) follows three basic subtypes: squamoid, magnoid and bronchoid.	78
Figure 21 PathMark diagrams on TCGA Lung identify differential pathways between subtypes [Ng 2012].....	80
Figure 22 Dynamic Heatmap of LUAD clusters with phenotypes annotated with significant pathways.....	81
Figure 23 DIPSC heatmap of uterine and endometrial cancer reveal clustering by mutation to subtype and grade.....	83
Figure 24 Gray cell line drug sensitivity correlated against TP53 activitome: colorectal cancer (CRC), bladder cancer (BLCA), kidney renal cell cancer (KIRC), breast cancer (BRCA), glioblastoma brain cancer (GBM), lung adenocarcinoma cancer (LUAD), uterine cancer/endometrial cancer (UCEC) and “pan” cancer (PAN). Red indicates correlated, blue indicates anti-correlated.....	86
Figure 25 Prostate Cancer West Coast Dream Team Rapid Learning Community.....	91
Figure 26 Physician Oncologist view of MedBook. The focus is on the patient care with scientific data views available, but not foremost in view.	102
Figure 27 Patient Centered MedBook. Focus is on the patient care and managing the complex events and personal decisions using the best medically recognized data.	103
Figure 28 Clinical Researcher view of MedBook. Focus balanced between patient care and research questions, scientific data views are in context of patient information presentations.....	111
Figure 29 Academic Researcher view of MedBook. Patient names are anonymized and the tools favor research questions. Scientific data views mapped to analysis are foremost.	113
Figure 30 Virtual Tumor Board View of MedBook, which extends the conversation to include drug genomic profiling in the conversation.....	114
Figure 31 User community and current state of MedBook completion.....	115
Figure 32 Clinical Genomic Pipeline segregates mechanical raw data analysis associated with genomic data from medical knowledge based analysis.	117
Figure 33 MedBook architecture consists of three zones: Red zone (HIPPA), Blue zone (genomic data), Green zone (anonymized data) and access to exascale computers.	118
Figure 34 DIPSC tool in the workflow in MedBook	120
Figure 35 Dynamic Heatmap of patients by gene expression with annotations.....	121

Figure 36 MedBook DIPSC tool requires setting the phenotype and clinical feature file and produces correlations, p-values and variance data sets.....	122
Figure 37 DipStick tool (left) and DipStick figure (right).	123
Figure 38 UCSC Activitome Signature Database Tool in MedBook.....	124
Figure 39 UCSC Activitome Signature Database Differential App allows the user to run various methods comparing signatures to samples.	125

List of Tables

Table 1 Gray cell line activitome signature subtype compared to colorectal location and histological type activitome signature.	73
Table 2 Gray cell line drug activitome signature compared to TCGA colorectal cancer mutation and clinical phenotype activitome signature. All entries are significant.	74
Table 3 Top TCGA KIRC mutational signatures correlated to Gray lab breast cancer drug sensitivity signatures.	76
Table 4 Top correlations between Gray cell drug and subtype signatures and TCGA Uterine cancer and endometrial cancer signatures.	84

Abstract

TOOLS FOR EXTRACTING ACTIONABLE MEDICAL KNOWLEDGE FROM GENOMIC BIG DATA

By

Theodore C. Goldstein

Cancer is an ideal target for personal genomics-based medicine that uses high-throughput genome assays such as DNA sequencing, RNA sequencing, and expression analysis (collectively called *omics*); however, researchers and physicians are overwhelmed by the quantities of big data from these assays and cannot interpret this information accurately without specialized tools. To address this problem, I have created software methods and tools called *OCCAM* (OmiC data Cancer Analytic Model) and *DIPSC* (Differential Pathway Signature Correlation) for automatically extracting knowledge from this data and turning it into an actionable knowledge base called the *activitome*. An *activitome* signature measures a mutation's effect on the cellular molecular pathway. As well, *activitome* signatures can also be computed for clinical phenotypes. By comparing the vectors of *activitome*

signatures of different mutations and clinical outcomes, intrinsic relationships between these events may be uncovered. OCCAM identifies activitome signatures that can be used to guide the development and application of therapies. DIPSC overcomes the confounding problem of correlating multiple activitome signatures from the same set of samples. In addition, to support the collection of this big data, I have developed *MedBook*, a federated distributed social network designed for a medical research and decision support system. OCCAM and DIPSC are two of the many apps that will operate inside of MedBook. MedBook extends the Galaxy system with a signature database, an end-user oriented application platform, a rich data medical knowledge-publishing model, and the Biomedical Evidence Graph (BMEG). The goal of MedBook is to improve the outcomes by learning from every patient.

Acknowledgements

I want to thank my committee, my many excellent teachers, my colleagues on the UCSC Cancer Team, my collaborators at: Five-3 Genomics, Oregon Health Sciences University, Washington University, UCSF, Stand Up 2 Cancer, Prostate Cancer Foundation, American Association of Cancer Research, National Cancer Institute The Cancer Genome Atlas project, my dear friends Robert Baertsch, Peter Hornick, Uri Lopatin, Jena Pincott, Jon Ziegler, Robert Zimmerman, my family, and especially my wife Jessica Bernhardt, and my mother Beatrice Goldstein without whom none of this would have been possible.

1 Introduction and Overview

The world of cancer diagnosis and treatment is undergoing a paradigm shift. Historically, cancer diagnosis and treatment decisions were based mainly on observations through a microscope of a dense mass of misshaped cells. This approach is increasingly being augmented by (and will eventually be replaced by) molecular diagnostic assays. These new high-throughput assays measure millions of data points representing cancer cell DNA mutations and copy number, RNA expression, and protein expression. These data provide a more direct window on cancer's molecular cause. The new diagnostic paradigm however needs a comprehensive knowledge base that connects cancer's molecular disruptions and clinical factors to therapeutic biomarkers that will guide treatment and improve outcomes [Haussler 2012]. This thesis will show how tools I developed called *OCCAM* and *DIPSC*, in a framework called *MedBook*, systematically create a knowledge base of cancer's activity signatures called the *activitome* that can guide treatment and assist in the development of new therapies.

Just as *genome* is a portmanteau of gene and ome, *activitome* is the portmanteau of *activity* and *ome*. *Activitome signatures* are a form of *biomarker*, indicators of underlying biological state. An *activitome* signature consists of a vector of probabilities of the molecular differential diagnostic effect of a molecular event or clinical phenomenon. Since they are measurable mathematical objects, *activitome* signatures also can be correlated and compared each other to better understand the underlying causal phenomenon. Most biomarkers cannot be compared directly. The

phenomena may be clinical phenotypes, inherited genotypes, somatic mutations, or any other dichotomized characteristic. The most important activitome signatures will be derived directly from patient outcome data where drug effectiveness is clinically known.

There has been recent criticism of biological science for the proliferation of *omes*: invented biological terminology for trivial phenomena that cannot be measured. Simultaneously, there is recognition that there are also undiscovered significant omes that, if they are useful and can be quantified, will enhance biology and medical care [Baker 2013].

In the first section, I hope to convince you that the activitome is both worthy of standing and necessary to make progress in the understanding and the treatment of cancer. Cancer's cell mutations are concrete physical events whose manifestations usually matter. The mutations' effects can now be measured as molecular activities. Since the activitome is a concrete mathematical measure, it *is* quantifiable, and it enables the application of statistical and computational methods that are extremely helpful in diagnosing and treating cancer. Therefore, the activitome deserves to be called an *ome*. The activitome is a knowledge base of mutations and clinical phenotypes that have a mathematical signature.

In section 2, I present the OCCAM and DIPSC methods that I developed to create activitome signatures. Section 2 is also intended to be a methods paper that will be submitted to a journal such as *Bioinformatics*.

In Section 3, I describe my results, which created the first version of the

activitome. More than just a mathematical construct, activitome signatures will guide therapy selection. As part of the Stand Up 2 Cancer Dream team, I used my tools to analyze drug sensitivity and resistance from breast cancer cell line data provided by the LBNL/OHSU Gray Laboratory. The results were published in *PNAS* as *Subtype and Pathway Specific Responses to Anti-Cancer Compounds in Breast Cancer* [Heiser 2011].

I then analyzed patient data from Washington University and co-authored a *Nature* paper with Matt Ellis, Elaine Mardis, Li Ding, and others on *Whole-Genome Analysis Informs Breast Cancer Response to Aromatase Inhibition* [Ellis 2012].

Section 3 also contains my first-author *American Association of Cancer Research Journal of Clinical Oncology* paper entitled: *Extracting Medical Knowledge from High-Throughput Genomic Data* [Goldstein 2013] that summarizes the contribution of activitome signatures to the *Nature* paper. This section also discusses results from the National Cancer Institute's *The Cancer Genome Atlas* project *Nature* marker papers in which I am also listed as a co-author. The list below describes various cancers (the abbreviation), and the marker papers written about each one:

- Colorectal Cancer (CRC) *Comprehensive Molecular Characterization of Human Colon and Rectal Cancer*, *Nature* July 2012.
- Breast Cancer, (BRCA or TCGABRCA), *Comprehensive molecular portraits of human breast tumours*, *Nature* October 2012.
- Uterine and Endometrial Cancer (UCEC), *Integrated Genomic Characterization of Endometrial Carcinoma*, *Nature*, May 2013.
- Lung Adenocarcinoma Cancer (LUAD), submitted to *Nature*.
- Clear Cell Kidney Renal Cancer (KIRC), submitted to *Nature* (now in press).

In Section 4, I offer my vision for how the activitome will translate into actionable medical treatment on a knowledge web platform I call *MedBook*.

The remainder of this overview introduces the foundation for this thesis by answering these questions:

- Why does cancer therapy need to be guided by biomarkers?
- How can activitome signatures improve cancer treatment?
- How does cancer avoid being treated by current therapies?
- How can the knowledge of pathways provide a foundation for discovery?
- How does a pathway-based method infer molecular activity?

1.1 Cancer Biomarker-Guided Therapy

Improving patient outcomes and the difficulties of the pharmaceutical industry in finding new drugs to treat cancer are driving the paradigm shift to biomarker-guided therapy. The pharmaceutical industry model of the previous fifty years was to invent drugs that could be safely administered to all patients with an illness. President Nixon promised “magic bullets” that would eradicate cancer the way that antibiotics cured pathogenic bacterial infections. Instead, cancer has proven to be a more complicated disease process and the pharmaceutical industry has failed to find the magic bullets to cure cancer. Making the situation worse, pharmaceutical companies have also experienced an exponential loss in productivity of new drug development over the last fifty years [Scannell 2012].

There may never be a magic bullet for cancer. If new therapies are only appearing slowly, oncologists have to make better use of current therapies. They will need to work harder to find the right combination and delivery sequence of therapies to effectively treat most cancers. Biomarkers can guide oncologists to administer the

best combination of therapies to an individual. Biomarkers will also enable better application of both existing drugs and new drugs by targeting a drug directly to the appropriate patient population [Dalton 2005]. The FDA already lists 37 gene-based biomarkers approved for drug labels [FDA 2013]. Clinical biomarkers such as Genomic Health's Oncotype DX provide information on cancer subtype and some prognostic information, but do not provide therapeutic guidance beyond a single therapy. These biomarkers are often inadequate for finding therapies for the most critically ill patients [Kennecke 2012]. A good example of a successful biomarker is the test for the HER2+ (ERBB2) breast cancer subtype, called HercepTest. This test determines whether Genentech's trastuzumab (Herceptin) may work for a patient [Piccard-Gebhard 2005]. The HercepTest biomarker provides a positive predictive value of just 57% to select the 15-25% of breast cancer patients who will *likely* benefit and the 75-85% who will *likely not* benefit from trastuzumab. Despite these low numbers, HercepTest is a frequently cited model for diagnostic driven treatment.

In contrast to trastuzumab's success, there have been many notable failures of drug selection when there have not been biomarkers to guide therapy. Drugs such as Genentech's bevacizumab (Avastin) had their approval removed by the FDA for breast cancer. Patients who took the drug did not live any longer than those who did not take the drug. [Allison 2008]. Yet, in other cancers, such as colorectal cancer, bevacizumab has been shown to be effective [Saltz 2008].

A colorectal study found that the median duration of survival was 20.3 months in the treatment group given irinotecan, bolus fluorouracil, and leucovorin (IFL) and

bevacizumab, as compared with 15.6 months in the group given IFL plus placebo. The median duration of progression-free survival was 10.6 months in the group given IFL plus bevacizumab, as compared with 6.2 months in the group given IFL plus placebo [Hurvitz 2004].

It has been suggested that there may be a sub-population of breast cancer patients for whom bevacizumab would also provide a similar benefit, however a biomarker is needed to identify these patients. One such biomarker may be an activitome signature that is created for colorectal cancer and other cancers where bevacizumab successfully treats patients.

The greatest immediate benefit of using the activitome may be to assist oncologists with determining the right combination of existing therapies to cope with the constantly shifting ground of cancer biology. The necessary drug combination shifts over time because cancers are constantly mutating by evolutionary selection of minor pre-existent cancer clones bearing mutations that allow the cells to survive therapies. These mutations may result in new pathway connections or altered proteins that adversely affect therapeutic success. This means that the oncologist should consider the tumor to be an evolving collection of diseases, and not just one disease to be treated in isolation.

Mutations are often considered either to be *driver* mutations, which actually cause tumor growth, or *passenger* mutations that are merely along for the ride and do not facilitate cancer progression. Recently however, it has been observed that the increased burden of many passenger mutations can also alter cancer's progression

[McFarland 2013]. If medicine is to gain a full understanding of tumorigenesis and progression, then *ideally* all mutations should be considered when selecting the right therapy. The problem rapidly becomes very complicated. The oncologist will need to treat multiple diverging clonal subpopulations of tumors, each with different disjoint mutations competing for resources. This heterogeneity is caused at least in part by the stochastic nature of genome damage that occurs during passage through telomere crisis [Collisson 2012]. These *mutation inventions* select for mutants that avoid the patient's internal immune system as well as those that avoid external therapies and allow proliferation to other tissues. Because cancer continually evolves within the patient, multiple combination therapies may be necessary to prevent a cancer from becoming multi-drug resistant [NatBio 2000]. By charting the trajectory of changing activitome signatures over time, it may be possible to anticipate how the cancer is likely to evolve next and thus enable a careful and curative selection of combination therapies in advance of wide-scale proliferation of multiple clonal malignant subpopulations of cancer cells.

1.2 How Can the Activitome Improve Cancer Treatment?

Activitome signatures simplify pathway alteration complexity into a simple vector that can be easily mathematically compared. High activitome signature correlation suggests that mutations in different genes cause the same downstream biochemical pathway disruptions. This implies that different causal mutations may be therapeutically treated the same. To understand this idea in detail, I review cancer treatment in the context of potential activitome signatures to guide treatment.

Cancer is currently treated with a combination of surgery, radiation, cytotoxic chemotherapy, and molecular pathway targeted drug therapy. Soon the roster may include gene therapy and viral therapy. The patient's endogenous immune system's role in working with exogenous treatments is not well understood. This subsection reviews cancer treatment, the limited use of existing biomarker technologies and proposes possible roles that activitome signatures may play in the future in each of these treatments and how to mobilize the endogenous immune system for best outcome.

1.2.1 Improving Surgical Outcomes

Surgical resection of a cancer is the simplest and often the most direct form of treatment. Surgery is used both for diagnosis and for excision of tumor masses. Surgery may be supplemented with other adjuvant treatments, such as radiation, chemotherapy, and molecular-targeted drug therapy. When the tumor is compact, and safe margins to surrounding tissue can be achieved, surgery is the preferred therapy.

Most decisions regarding surgery are made based on tumor location as determined by imaging studies, except when inherited cancer predisposition indicates a high risk for developing cancer in other remote locations as well. In this regard, Myriad's *BRCAnalysis* test, revealing an inherited lifetime risk of developing breast, ovarian and other cancers, is the first and only biomarker test currently used to determine the extent of surgical excision as recommended by the *National Comprehensive Cancer Network* (NCCN), an alliance of the world's leading cancer centers [NCCN 2013]. Breast Conserving Surgery (BCS) is the standard treatment for

women with early-stage non-BRCA mutation-bearing breast cancer, as recommended by current guidelines. However, BCS may put certain patients (e.g. BRCA mutation carriers) at risk of recurrence or new cancer development. Activitome signatures that could stratify breast cancer patients before surgery into high, moderate, and low risk for local recurrence or metastases would therefore have important clinical implications.

1.2.2 Guiding Therapy Choice

Radiation and chemotherapies target rapidly dividing proliferating cells. However, cell proliferation is a poor therapeutic target because healthy dividing cells in children (all tissues) and adults (many tissues) will sustain collateral damage. Activitome signatures may be able to identify the pathways that are amplified by specific lesions and be used to select therapies that suppress just those pathways to minimize collateral damage to cells.

Not all patients benefit from radiation therapy and there are no identified biomarkers that predict which patients will benefit from radiation therapy. Elevated expression of Ape1/ref-1 proteins and damage to P53 pathways may confer resistance to radiation therapy [Robertson 2001]. As well, patients with a high potential of distant metastases should receive chemotherapy in addition to radiation therapy [Langlands 2013] because chemotherapy has broad effect and radiation only kills localized tumors within the radiation field. Radiation therapy also stimulates inflammation that, for some patients, is known to accelerate cancer and cause cardiovascular disease so it is best to minimize radiation exposure when possible.

It may be possible to create activitome signatures that model when radiation therapy will lead to further complications such as consequential secondary cancer tumors or distant metastases [Saito 1997]. Activitome signatures could help select the right combination of surgery, radiation therapy, cytotoxic drug therapy, and molecular-targeted drug therapy. For example, the PI3K/Akt pathway indirectly signals apoptosis when activated by radiation [Toulany 2007]. The PI3K/Akt pathway is also the target of the drug rapamycin (mTOR) [Franke 2003]. Activitome signatures of this pathway may be able to determine whether this pathway is intact and whether it can be sensitized to radiation by administering an adjuvant [Albert 2006].

1.2.3 Guide Cytotoxic Drug Dosage

Activitome signatures will guide chemotherapy decisions by characterizing the toxicity of the drug as well as the sensitivity of tissues to the toxicity. Chemotherapy must spare healthy cells and preferentially kill cancer cells. Chemotherapy is preferred when a cancer has spread throughout an organ beyond the possibility of its complete surgical excision, when the cancer has metastasized to distal regions, or when the cancer occurs in an organ that is inaccessible to either radiation or surgery [Smith 2010].

One of the most effective chemotherapies is *methotrexate*, a cytotoxic drug that kills dividing cells. This includes both the cancer and the necessary normal replacement cells in the patient's body, such as those in the digestive tract, red blood cells, and the immune system lymphocytes. Methotrexate works by preferentially

binding and immobilizing the dihydrofolate reductase (DHFR) enzyme, which synthesizes purines for nucleic acid bases and many amino acids. Cancer cells that have multiple copies of the DHFR gene are the most drug resistant [Stark 1984] possibly because the methotrexate dosage needed to treat all sites would be lethal to the patient.

Amplification of genes (increased copies) may be a result of both germ line and cancer-caused recombination or transposon-specific enzymes. Increased copy number of DHFR is detectable by using gene sequencing or gene labeling techniques as a biomarker to guide methotrexate administration. However, upregulation of expression DHFR may also occur because of random regulatory epigenomic changes triggered by cancer that gene copy number measurements cannot detect. Activitome modeling will be able to identify patients that have mutations, copy number, or epigenomic-originated upregulated DHFR; which may then favor or disfavor treatment with methotrexate.

Another analogous chemotherapy is the anti-metabolite *5-fluorouracil (5-FU)* used to treat many cancers ranging from easily treated basal cell carcinoma to usually fatal pancreatic cancer. Gene amplification among different human populations also plays a role in 5-FU resistance. Choosing the best mix of 5-FU and methotrexate has been the subject of many oncologists work without biomarker guidance.

I will go into depth on the 5-FU pathways because it illustrates the complexity that simple gene sequencing cannot model but activitome signatures can. As well, there has been success in using other expensive assays to monitor patients using rare

high-performance liquid chromatographic (HPLC) assays [Bocci 2006], which measure the ratio of uracil/dihydrouracil in the blood, which can also predict drug uptake and patient response.

Few pathways and the drugs that target them are as well understood as 5-FU and TYMS (thymidylate synthase, the enzymatic target of 5-FU). The 5-FU molecule is a pyrimidine and similar to cytosine and uracil (the C and U in RNA) and cytosine and thymine (the C and T in DNA). The similarity is obvious in Figure 1. The drug 5-FU targets the TYMS gene. TYMS generates thymidine monophosphate, which in turn makes the thymidine triphosphate used in both DNA synthesis and repair.

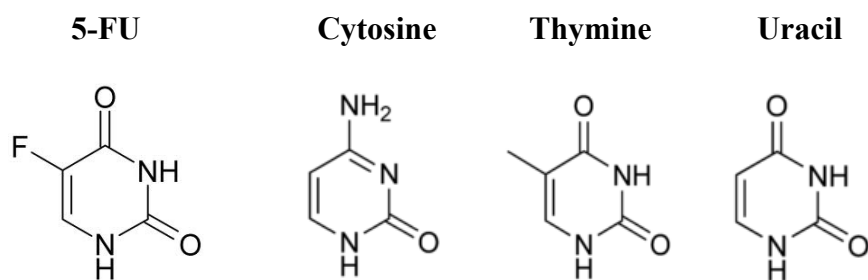


Figure 1 Molecular structure of 5-FU compared to the pyrimidines.

The 5-FU molecule interferes with DNA synthesis, which is required for cancer proliferation. The biochemical reactions in the pathway do not detect the fluorine before 5-FU is transformed into cytotoxic metabolites that are precursors for DNA and RNA synthesis. The DNA and RNA molecules that incorporate 5-FU components are no longer functional, ultimately leading to cell death in cells with functional P53 pathways. Even if the P53 pathway is not functional, 5-FU and

methotrexate synergistically inhibits DHFR regeneration, which brings purine synthesis to a halt [Benz 1982]. Without purine synthesis the cancer cannot proliferate.

The human body has excellent detoxification systems and not surprisingly treats 5-FU as a poison to be eliminated or neutralized. Unfortunately, approximately 85% of an intravenous dose of 5-FU is inactivated in the liver by dihydropyrimidine dehydrogenase (DPD), an enzyme that exhibits up to 20-fold variation in activity among individuals. Therefore, in some patients, almost no 5-FU is getting to the cancer cells! When 5-FU is inactivated by DPD, it forms the inactive metabolite dihydro-5-FU (FUH₂). This then provides a physical block to the remaining 15% of the biologically available 5-FU [Walters 2003] [Goetz 2004].

Given what is known so far, one would think that patients with low levels of DPD would be fortunate and show enhanced benefit from 5-FU therapy. Unfortunately, patients with low drug metabolism have increased toxic side effects. For example, many side effects of 5-FU occur because the DPD is now inactivated. Patients with low DPD activity cannot effectively inactivate 5-FU, which leads to excess 5-FdUMP in healthy tissue, causing gastrointestinal, hematopoietic, and neurological toxicities that may be fatal.

The mystery of this 5-FU host toxicity returns to the TYMS gene. Genomic studies found a polymorphism within the 5'-promoter enhancer region of TYMS that consists of tandem repeats of 28 base pairs consisting of the sequence CCGCGCCACTTGGCCTGCCTCCGTCCCG. The vast majority of Caucasian

subjects carry either a double (TSER*2) or a triple (TSER*3) repeat for this polymorphism. Although four, five, and nine copies of the repeated sequence have been also described. Allele frequencies in Southwest Asian subjects are similar to those found in Europeans.

However, the homozygous triple repeat is almost twice as common in Chinese subjects (67%) as in Caucasian subjects (38%). Patients homozygous for the TSER*3 genotype have elevated intra-tumor TYMS messenger RNA levels, elevated TYMS protein levels, and poorer response rates after 5-FU chemotherapy compared with TSER*2 homozygotes. In a study of 65 patients with locally advanced rectal cancer treated with 5-FU, of patients with the TSER*3/*3 homozygous genotype only 22% improved. However 60% of patients with either the TSER*2/*2 (homozygous) or TSER*2/*3 (heterozygous) genotypes improved. Similar reports of patients with the various TSER alleles have been observed in other cancers, including gastric cancer [Pullarkat 2001]. Thus, the greater the number of repeats, the worse the patient outcome is. It is likely this is more than just a gene dosage effect and remains an active area of study.

Oncologists are in a difficult position. 5-FU is often a second line therapy, used only when other therapies have failed or in cases where there are no other options such as for pancreatic cancer [Azmy 2013]. 5-FU does work well as a topical cream for basal cell carcinoma since the liver has no opportunity to detoxify it. It is understood now that the more copies of the TSER gene, the greater the product of TSER in the body and the lower the likelihood of positive effect from 5-FU.

Myriad Genomics has made a rapid targeted sequencing test called *TheraGuide 5-FU* to detect copies of the TSER gene [Winder 2010]. TheraGuide only sequences two genes: DPD and TYMS. However, TheraGuide has no visibility into any of the many other resistance mechanisms nor does it assay the rest of the TYMS pathway. An increased number of copies of the DNA promoter region is just one way to increase the activity of TSER. Cancer may also increase the activity by increasing the expression of promoters or decrease the activity of inhibitors. Activitome signatures can recognize the entire activity of the pathway.

Activitome signatures could monitor the chemotherapy progress, alerting when drug resistance mechanisms emerge so that counter-measures to drug resistance may be taken at the proper time. All the chemotherapy drugs have terrible side effects because they are broad-spectrum cytotoxics. Determining how to best minimize the dosage is another application for activitome management of the tumor.

1.2.4 Hormone Molecular-Targeted Drug Therapies

It is in molecular-targeted therapies where activitome signatures have their first utility as described in [Goldstein 2013]. Targeted therapies offer significant hope to reverse cancer without the greater side effects of cytotoxic chemotherapy. The hormone chemical system is a particularly rich area for molecular-targeted drug targets. As well, the hormone system is also a rich visible system for activitome signatures.

Hormones are a chemical global message system that coordinates growth, digestion, sexual reproduction, and other long-term growth-related functions

including menstruation across many cells in the body. Hormones also stimulate the growth of many cancers. Hormone therapy may attempt to slow cancer by either upregulating or blocking hormones directly or interfering with the attachment of hormones to receptors [Hanahan 2000]. With breast cancer, the female hormones estrogen and progesterone may promote the growth of breast cancer cells. Hormone therapy blocks the body's naturally occurring estrogen and fights the cancer's growth [Ellis 2012].

1.2.5 Kinase Molecular-Targeted Drug Therapies

Closely linked to hormone pathways, *kinases* are another exciting target for targeted therapies. Kinases are dynamic protein enzymes that store chemical state by transferring a phosphate group from ATP to a protein in a process called *phosphorylation*. Kinases act as switches with an “on” state and an “off” state. Kinases are important drug targets because they are components in cancer and other disease pathways including growth pathways, apoptosis pathways, and inflammation pathways. One important family of activitome signatures will likely specifically target EGFR. The Human EGFR family consists of four structurally analogous receptors: HER1 (EGFR), HER2, HER3, and HER4. These receptors regulate cell growth for wound healing, child development and growth, and adult secondary sex characteristics.

The pathway cascade begins when a growth factor binds to the receptor. This may trigger activation MAPK, AKT1, and PI3K kinases that may commence the cancer’s cell division and proliferation. The human *kinome* (knowledge base of

kinases) consists of at least 518 different kinases [Manning 2002]. When one of the kinases in the pathway is mutated, it can be fixed in the “on” or “off” position, which may lead to cancer. If there were a complete set of drugs to inhibit many (if not all) of the 518 kinases, there would need to be a specific set of activitome signatures to guide the application of the drugs. Drugs may treat aberrantly phosphorylated kinases. New techniques allow the phosphorylation state to be observed. Activitome signatures may fully explain the downstream effect of phosphorylated kinase on the cell. Many kinases are involved in inter-cell communication, and therefore are accessible at the cell surface.

A very important subset of the kinases is the *tyrosine kinases* that attach phosphate groups to tyrosine amino acid residues (other kinases attach phosphate groups to the amino acids serine and threonine). Mutations in a kinase gene may cause the kinase to lock in the *on* position. This may lead to unregulated growth of the cell. Therefore, kinase inhibitors can be effective cancer treatments. Activitome signatures will be made of the proteins that surround the tyrosine kinase receptor, EGFR.

The first of the new generation of kinase targeted drugs is Gleevec (*imatinib*), which targets the blood cancer Chronic Myeloid Leukemia (CML), and in particular, targets a lesion called the *Philadelphia chromosome* (named in 1960 for the city in which it was discovered) [Talpaz 2002]. The Philadelphia chromosome is a translocation between chromosome 9 and 22 in the middle of two important genes, the BCR gene and the ABL gene.

The BCR gene's purpose is unknown. The ABL gene encodes a kinase involved in cellular growth. The fusion of BCR-ABL, however, is known to cause uncontrolled cancer. When the BCR-ABL fusion lesion occurs, leucocyte cells proliferate without responding to cytokines (their normal regulatory mechanism). The drug *imatinib* blocks the ABL tyrosine kinase domain. Imatinib also inhibits tyrosine kinase activity of the platelet-derived growth factor (PDGF) receptor and c-Kit protein [Pierotti 2011], but it does not affect other members of the type III receptor kinase family, such as Flt-3 and Fms. Therefore, the side effects are few [Sierra 2010].

Imatinib has utility beyond CML [Schoenewolf 2011]. A large study found imatinib effective for some patients who had mucosal melanoma (a radically different indication than CML) which is frequently caused by mutations to c-Kit [Kim 2008]. It would be useful to have a comprehensive database of the activitome signatures of patients for whom imatinib was effective so that new patients may be treated. Activitome signatures could help link pathways to specific gene family members that will be key in future drug development.

1.2.6 Drug Resistance and Targeted Therapies

Cancers can develop drug resistance to targeted therapies. Early stage cancers may still have well-formed stable genomes. Later stage cancers have hypermutated unstable genomes. The drug resistance emerges partly because the binding sites of the drug are altered by mutations [Yap 2010][Sierra 2010]. The general solution inspired by poly-chemotherapy HIV treatments to drug resistance may work in cancer. When

the mix of HIV virus subtypes change, the clinicians change the cocktail of drugs [Daley 2003]. However, the HIV virus genome is small and a given patient will have only a changing proportion of a limited number of alternative virus subtypes.

Cancer's genome is the human genome (therefore very large). For example, imatinib drug resistance typically forms because lesions in imatinib's binding site prevent imatinib from binding but still allow the EGFR to bind. Imatinib resistance mechanisms include [Hochhaus 2002]:

- Alterations affect Gleevec binding sites to BCR-ABL and C-Kit
- Primitive CML cells demonstrate enhanced BCR-ABL expression
- Upregulation of Src-related LYN kinase, SHH, Wnt/ β -catenin, anti-apoptotic factor Bcl-2, MDR1/ABCB1 and ABCG2/BCRP
- Upregulation of multi-drug efflux transporters and/or cellular organic cation transporter Oct-1 contribute to the CML recurrence
- Upregulation of the drug efflux pumps breast cancer resistance protein (ABCG2) and P-glycoprotein
- Gleevec uptake by leukemic stem cells may require expression of human organic cation transporter (hOCT1)
- In vitro culture of primitive CML cells in the presence of Gleevec leads to accumulation of quiescent cells
- The bone marrow microenvironment may protect leukemic stem cells from the effects of Gleevec [De Giorgi 2011]

Each of these resistance mechanisms will require the alteration of the cocktail of drugs administered with Gleevec. Unlike for HIV, however, sequencing assays do not tell the complete story. Patient melanoma cells may not show any c-Kit mutations, nor respond to Gleevec in vitro. The only way to detect resistance is to profile the activating tumor growth-driving pathways [Schoenewolf 2011].

Unfortunately, the stem cell that generates the CML cells is not itself vulnerable to Gleevec [Jamieson 2004]. However, when a treatment is found that

targets the CML progenitor, activitome signatures may be useful in determining the optimal cocktail. An interesting special case may occur when the leukemic cells are hiding in protected bone compartments; it may be necessary to administer bone marrow mobilization cytokines to make the cells accessible to the treatment. It is thought that the bone marrow stem cells provide a microenvironment that allows HIV, and tuberculosis, and possibly cancer cells to evade immune system surveillance, possibly by having an enriched environment of hematopoietic progenitor cells [De Giorgi 2011]. The activitome database will amass the catalog of treatments, outcomes, and special case therapies.

1.2.7 Immune System and Antibody-based Targeted Therapies

Antibody-based treatments enlist the patient's immune system to target the cancer. Not all patients benefit, however, because cancer disables normal immune system response. Activitome signatures may be able to detect whether the immune system is functional or whether antibody targeting may be contraindicated if the cancer has suppressed the immune system.

Monoclonal antibody therapy avoids the difficulty of creating small molecular drugs and focuses exclusively on cell surface and secreted molecules. They are manufactured antibodies produced by the immune cells of a target animal or by using other molecular engineering techniques to create the antibody protein. The B cells are extracted and grown in culture to produce therapeutic quantities of the antibody protein that is then purified and administered to a patient intravenously. An example antibody therapy is Herceptin (trastuzumab). Herceptin binds to the extracellular

segment of the HER2 receptor and prevents dimerization with HER3 preventing EGFR pathway activation.

Numerous tyrosine kinase inhibitors however target the EGFR pathway. Two other well-known drugs used to treat non-small cell lung cancer are gefitinib (Iressa) and erlotinib (Tarceva)[Kobayashi 2005] [Godin-Heymann 2008]. Patients can become resistant to these drugs because of lesions in the EGFR target. The most common lesion is a single amino acid change from the EGFR T790M lesion which prevents binding of both gefitinib and erlotinib [Kobayashi 2005]. The bulkier methionine residue at position 790 sterically hinders the interaction with an inhibitor, effectively preventing binding to the EGFR kinase domain while preserving the oncogenic catalytic activity. Therefore, EGFR is a strong prognostic indicator for outcomes, and upregulated pathways indicate poor prognosis [Nicholson 2001].

Targeted therapies also fail because of pathway *crosstalk*. Crosstalk occurs when two powerful signaling pathways interact, leading to complex information processing downstream of the original receptors that may circumvent receptor-targeted cancer therapies. Crosstalk contributes to drug resistance by increasing the robustness of even damaged pathways. Receptor crosstalk occurs between MET, IGF-1R and EGFR and enhanced IGF-1R activates the PI3K/AKT pathway, which can be tumorigenic. Cancer cells that mutate their EGFR receptor survive and are thus become EGFR target drug resistant. There is reason to believe that many endocrine receptors have considerable cross reactivity. IGF-1 (the ligand of IGF-1R), IGFBP-3, IGFBP-1, insulin, leptin, and adiponectin have been associated with breast cancer

incidence and thus may be other examples of pathway crosstalk that activitome signatures may be able to detect.

1.3 How Does Cancer Circumvent Current Therapies?

Cancer cells evolve and compete until a dominant cell population emerges from the process [Greaves 2012]. Then the dominant cancer cell will clonally expand to become the bulk of the tumor mass. Mutations provide cancer with survival mechanisms against the endogenous cell tumor suppression systems, host immune systems and against external drug therapies. A summary of the tactics that cancer cells evolve to resist drugs are:

- Mutations that appear in the proteins which are the target of the drug prevent drug binding
- Pathways route around the target of the drug due to pathway crosstalk
- New pathway mechanisms upregulate (increase) and the drug then does not reach the target either because the delivery pathway is blocked, or because it is effluxed (eliminated) from the cell too quickly to be effective

The activitome will catalog these and other resistance mechanisms that emerge because of lesions in the DNA replication and checking pathways. Combination therapies will need to include countermeasures that meet and anticipate emerging mutations.

Constantly emerging mutations is one of the profound reasons why cancer is so robust and is able to produce drug resistant cells. Normal cells have spontaneous random mutations at a rate of less than 10^{-8} per base pair (per cell division), whereas tumors from the same individual can have a mutation rate of 10^{-6} per base pair [True 2006]. This figure was closely validated by another team which determined that the

average frequency of random mutations in tumor samples was 2.2×10^{-6} per base pair [Bielas 2006]. They measured that a cancer cell contains more than a thousand random mutations. A typical radiologically detectable tumor contains over 10^9 cells, while the entire tumor may harbor as many as 10^{12} DNA base pair alterations (such as substitutions, deletions, and small insertions). This gives every cancer a unique opportunity to invent novel pathways and pathway connections that never have been and never will be studied by science.

Thus, activitome signatures provide a higher-level abstraction that allows individual mutations to be clustered together into families of similar activity.

Section 3 on Clinical-Translational Results will show how many mutations have similar activitome signatures. Of course, many mutations may be irrelevant to cancer. Most mutations are passenger mutations, and only some of these will have any consequences. Unless it is known in the database, however, the mutation that is going have morbid consequences is indistinguishable from benign mutations until the damage is done.

1.4 How Can Knowledge of Pathways Provide a Foundation for Discovery?

The knowledge of pathways is a fundamental study at UCSC and elsewhere, and is another important pillar supporting this thesis. Pathways are a series of chemical reactions occurring within a cell that may process materials, information, or both. Activitome signatures summarize a cancer's pathway alteration. More than just chemical reactions, pathways are robust adaptable network control systems that have evolved to allow cells and organisms to cope with dynamic stress, infection, wounds,

and starvation. Most pathways process information redundantly and tolerate most changes caused by gene mutations. This plasticity is useful for species evolution, where changes may confer fitness advantage. The same plasticity, however, makes cancer difficult to cure.

Theodore Dobzhansky said, “*Nothing in biology makes sense except in the light of evolution*” [Dobzhansky 1973]. This includes cancer. Cancer recapitulates evolution where dysfunctional pathways (whether signaling or metabolic) confer an advantage to a cancer cell that improve the fitness of the cancer cell’s lineage to flourish at the expense of the healthy cells and the organism as a whole [Salk 2010] [Greaves 2012]. Pathway models are useful to researchers for reasoning about cancer because many mutations across multiple genes in a given pathway will produce the same improvement to fitness. For example, mutations can disable many genes in a tumor suppressor pathway. The resulting tumor suppressor pathway no longer prevents dysfunctional growth pathways.

Over 325 pathway databases [PathGuide 2011] delineate biochemical pathways in organisms ranging from bacteria to humans. The knowledge encoded in pathway databases contains descriptions of functioning pathways and many dysfunctional pathways. The goal in cancer systems biology is to infer the cause and treatment of cancer’s molecular processes from the structure of functional and dysfunctional pathways. UCSC’s cancer pathway projects draw largely from the following pathway databases:

- BioCarta
- Reactome

- HumanCyc
- Kyoto Encyclopedia of Genes and Genomes
- MetaCore
- National Cancer Institute Pathway Interaction Database
- PathwayCommons
- WikiPathways

Individual published pathways were integrated into a single superimposed pathway referred to as the *SuperPathway*. Redundancies were removed so that highly upregulated *subpathways (novel subsets of the SuperPathway)* could be readily identified. The UCSC team was motivated to create the SuperPathway because curated pathways only partially overlap. An example is EGFR, PI3 kinase, and MEK are curated as separate pathways. In the cell, they are basic components of a larger system that becomes dysfunctional. Recognition of this fact led to the concept of a SuperPathway that unites separate pathways and eliminates redundancies. It is only by computing all pathways together that it is possible to gain a complete understanding of a cancer's dysfunction. The role of bioinformatics in biology is to allow us to see the entire forest and not just the individual trees.

While, these databases are still largely incomplete, they do contain information about the complex biochemical interactions of the regulatory and metabolic state of a cell. It is not possible to measure many key elements of these pathways because they are transitory or inaccessible to current biochemical assay. Some other elements of these networks are deduced from biochemical principles. Despite the incomplete knowledge of the state of the network, UCSC and others are able to make useful inferences concerning individual elements that are derived from

the state of the network as a whole [Bell 2012][Varadan 2012].

The final critical insight of the UCSC approach to pathway cancer biology is that every node in the pathway graph is informed by Francis Crick’s *Central Dogma of Biology* [Crick 1958], that describes the flow of information from DNA to RNA and RNA to protein. Each node in the pathway may consist of multiple “dogma” nodes that represent DNA copy number, RNA transcription, and protein expression. With this background, it is now possible to put all of these elements together using the *Pathway Recognition Algorithm using Data Integration on Genomic Models* method (PARADIGM) [Vaske 2010].

1.5 A Pathway-based Method for Inferring Molecular Activity

The input to OCCAM and DIPSC consists of integrated pathway activity levels (IPL) that come from the PARADIGM method. PARADIGM, developed at UCSC, makes inferences from large-scale graphs of pathway networks. PARADIGM is implemented on a *factor graph* graphical model that computes the marginal distribution using a deterministic message-passing algorithm called Bayesian Belief Propagation (BBP) [Pearl 1982]. BBP is a computationally intensive method that uses the underlying factor graph to set the internal probabilities of the graphical model (indicated as squares in the diagrams below) to a configuration that has a high likelihood of occurring according to the observed data [Kschischang 2001].

Message passing can be computationally unbounded. Thus far, the PARADIGM SuperPathway has successfully scaled to over 10,000 pathway *concepts* (genes, complexes, abstract diagnostic elements). Since it is a graph and not a tree,

PARADIGM uses the *loopy* form of the BBP sum-product algorithm. Convergence is determined by iteration until change falls below a small threshold. The graph models three states for each node in the graph: upregulated, neutral, or downregulated.

Microarrays, genome sequencing, and any other physical assays are likely to yield exceptional outlier data that may skew results. PARADIGM helps to overcome this problem because the process of computing the *maximum likelihood* of all nodes in the graph probabilistically evaluates and smoothes outlier measurements.

Figure 2 illustrates how gene activities can be inferred for a pared-down model, simpler than reality. The PARADIGM graphical model is centered on a particular gene as shown in detail in Figure 2A. Gene expression, copy number and other measurements of a tumor sample are connected into a graphical model as observed variables (purple ellipses). Unobserved states of gene expression and activity are connected into the graph as hidden variables (green ellipses).

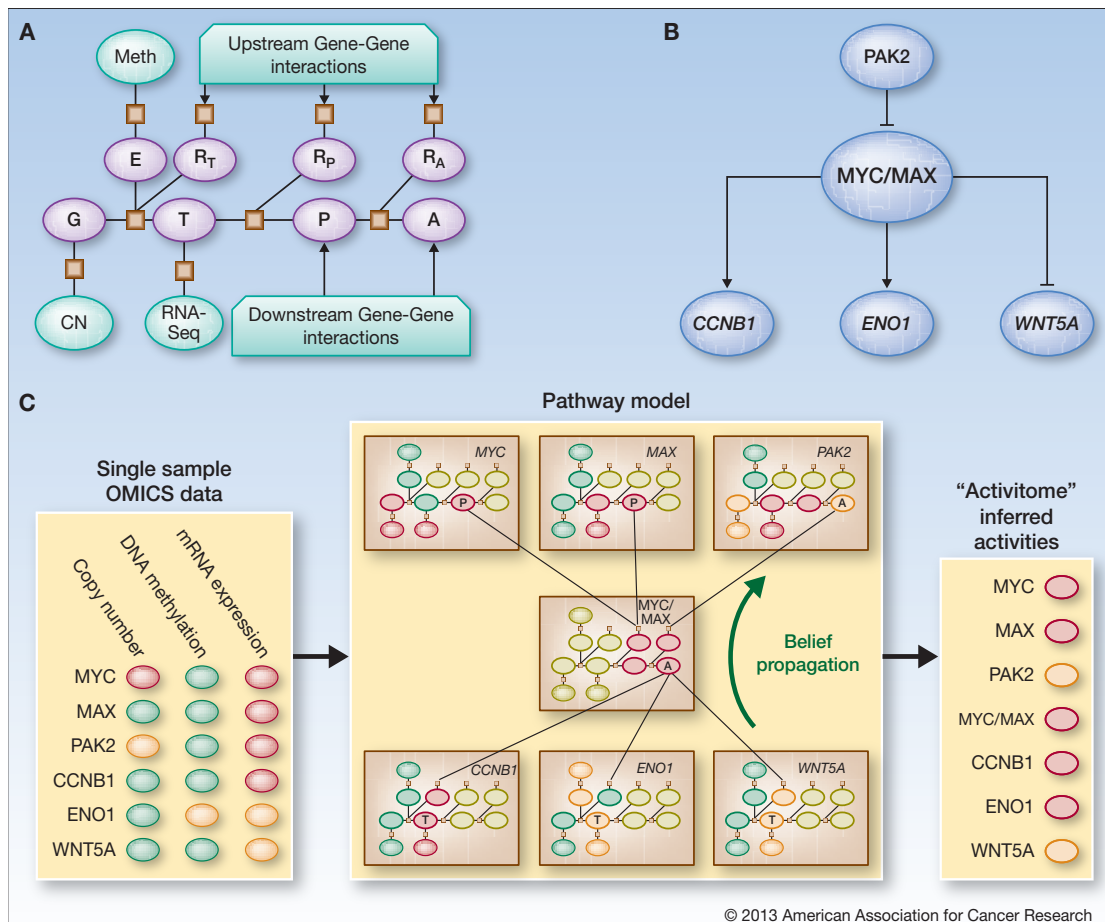


Figure 2 PARADIGM model for integrative data analysis. **A.** Factor graph model oriented around a single gene. Hidden states in a tumor sample (purple ellipses) are denoted: genomic copies (G), epigenetic promoter state (E), mRNA transcripts (T), peptide (P), and active protein (A). Regulation gene expression (green ellipses) includes transcriptional (RT), translational (RP), and post-translational (RA) control. Sample data (green circles, constrain gene states through factors (boxes)). **B.** Toy example of a MYC/MAX-associated pathway. Two transcription factors (MYC and MAX) form a complex (MYC/MAX) that is inhibited by PAK2, a protein kinase. MYC/MAX activates two target genes (CCNB1, ENO1) and inactivates a third (WNT5A). **C.** Single patient data converted to inferred activities for toy simplified pathway. Measurements and inferred levels are higher (red), lower (blue), or comparable (purple) to levels in matched normal. Belief propagation infers the kinase is inactive based on inferred higher activity of MYC/MAX. Figure copyright AACR 2013, used with permission.

The simplified pathway in Figure 2B shows a small pathway involving a single kinase, PAK2, that post-translationally inhibits the MYC/MAX complex. The transcription factor complex MYC/MAX in turn activates CCNB1 and ENO1, and *represses* WNT5A. Figure 2C illustrates how belief propagation could set the internal *active-state* of PAK2 based on the downstream evidence. In this example, it finds that

the PAK2 kinase is inactivated using the evidence downstream that suggests that MYC/MAX, a complex that is inhibited by PAK2, is active. This is because one of its known activated targets (CCNB1) is highly expressed while one of its repressed targets (WNT5A) has lower expression. Note that even though ENO1 is an activated target of the complex, it is not highly expressed. The model can explain away this apparent discrepancy using the information that ENO1's promoter seems to be epigenetically silenced and therefore the lower expression of ENO1 does not imply a lower activity of the MYC/MAX transcription factor complex.

The set of all inferred quantities of gene-encoded features appears as a nested wiring diagram graph (right hand side of Figure 2C), which integrated across patient samples are a quantitative state description of a tumor. This not only represents expression states of genes, but also multimeric protein complexes, gene family roles, and higher-level cellular processes that encapsulate the metabolic function and the information transmission aspect of genes and proteins.

PARADIGM is extensible to include new proteomic assays such as reverse phase protein array (RPPA) (shown in Figure 3), which will provide even more reliable insight into regulatory pathways, because the state of the measured proteins can be directly observed. RPPA will provide direct observation of a protein state, which means more evidence for the inference engine.

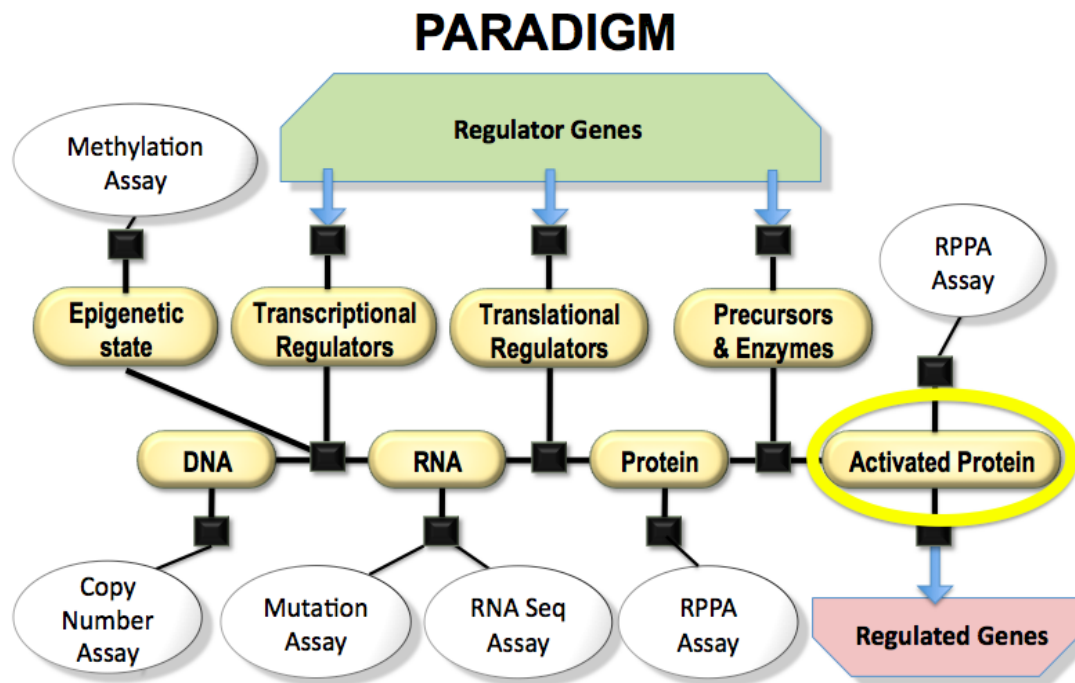


Figure 3 Central dogma of biology extended with regulators and assays.

PARADIGM infers activity on a per sample basis. To make an aggregate activitome signature from individual samples, a data aggregation integration method such as my OCCAM method is necessary.

1.6 PARADIGM-Derived Features Predict Patient Outcomes More Reliably than Gene Expression

Specific pathway components in an activitome signature can reveal important aspects of tumor biology. For example, PARADIGM uncovered previously unknown crosstalk between proliferation pathways and DNA damage repair pathways regulated by distinct isoforms of FOXM1 in the FOXM1 transcription factor pathway networks in ovarian serous cancers, published in *Nature* with The Cancer Genome Atlas (TCGA) marker paper [TCGA Ovarian 2011]. This crosstalk may explain in

part why these tumor cells proliferate in response to DNA repair signals.

When PARADIGM was applied to the TCGA glioblastoma multiforme (GBM) dataset, higher accuracy predictors of overall survival in the patients could be obtained compared to those only using gene expression signatures [Vaske 2010]. This strongly suggests that the pathway-level information provides biologically relevant clues about the intrinsic state of tumor cells. Thus, using pathway-inferred levels to build activitome signatures is preferred over gene expression for predicting biomedical outcomes.

1.7 Conclusion

In this overview, I have motivated a new comprehensive knowledge domain called the *activitome*, which consists of *activitome signatures*. These signatures consist of vectors of probabilities that characterize the effect of a mutation or a clinical phenotype in cancer. I have surveyed cancer's biomarkers, treatments, and especially mutations. Mutations are the proximal cause of cancer. I have also reviewed the PARADIGM method, which is the bioinformatic foundation of my research.

Section 2 describes two closely related bioinformatic methods called OCCAM and DIPSC that I developed to create activitome signatures using PARADIGM integrated pathway level vectors. Section 3 presents the results that include my most significant first author paper, which first introduced the activitome in print. I worked with Dr. Matt Ellis at Washington University to find activitome signatures related to *aromatase* (estrogen-related) *inhibitor* drug resistance in luminal

breast cancer. Section 3 also summarizes the steps in building the activitome from SU2C and TCGA results.

Bioinformatic tools, by themselves, do not save lives. Section 4 describes how these tools fit into my next work, *MedBook*, a medical research, and decision support environment I designed. MedBook extends the Galaxy system [Goecks 2010] with the activitome signature database, a biomedical evidence graph and other features described in section 4. The goal of MedBook is to improve the outcomes by learning from every patient.

2 Methods

Activitome signatures measure the downstream effect that may be causative or at least correlative to a mutation or a clinical phenomenon. I have developed two related methods embodied in software tools, called *OCCAM* (OmiC data Cancer Analytic Model) and *DIPSC* (Differential Pathway Signature Correlation), which create and compare activitome signatures. Researchers can use these tools to understand the relative correlation of the effect of clinical phenotype observations and tumorigenic events. This allows us to detect situations where seemingly different cancer events alter molecular pathways in ways that result in the same phenotype. In my approach, I reduce complex phenotypes to an essential label such as *P53 mutation* or *5-FU resistant*. Patient samples are in any of the two possible (*dichotomous*) event states: the positive event state or the negative event state (if the event state is unknown, the sample is excluded from analysis).

2.1 Background

The key insight is that the uniform treatment of and the ability to compare gene expression or IPL gene signatures as defined by differential of two groups of samples distinguished by dichotomous state variables provides the ability to profoundly connect phenotypes to underlying molecular events. This is graphically depicted in Figure 4, which shows a number of signatures derived from gene mutation events, drug sensitivity events, and clinical features.

Activitome Signatures that are “close together” in vector space are similar

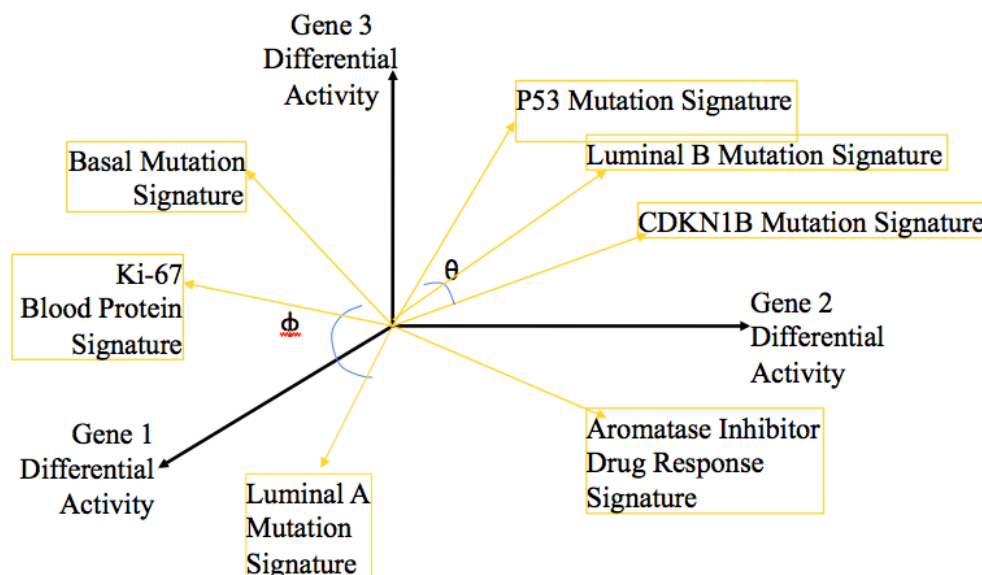


Figure 4 Activitome Signatures are an n-dimensional vector (only three dimensions shown). Each vector is a point in an n dimensional space where dimension is the difference in activity between two groups. For example, the Aromatase Inhibitor signature compares genes of two groups: the group of samples derived from patients that responded positively to aromatase inhibitors compared to the group of samples that are derived from patients who did not respond to aromatase inhibitors.

OCCAM and DIPSC combine methods of dichotomization, *Significance*

Analysis of Microarrays (SAM) [Tusher 2001] and Pearson’s R correlation. These are reviewed in the following sections.

2.1.1 Definitions

It would be useful to know if two events such as mutations or other clinical variables denoted by labels, *A* and *B*, are similar by knowing the correlation of their downstream effect on genomic pathways. If samples labeled *A* and *B* do not overlap, this is simple:

Let X be a matrix of integrated pathway levels, represented as an n by m matrix of PARADIGM IPLs. IPLs can be treated similarly to gene expression values (IPLs are preferred over gene expression values, see section 1.5). $X(i)$ is the row of IPLs for gene i . Here there are m genes in the rows and n patient samples in the columns.

X_{A+} , X_{A-} are random variables defined by the sample columns labeled $A+$ and $A-$. Unlabeled samples are excluded. Missing values in rows are inferred as the mean if below a small threshold (other options are possible including *K-nearest neighbor*, which provides similar results in this method).

$X_{A+}(i)$ is the values of gene $X(i)$ labeled A . $X_{A-}(i)$ is the values of gene $X(i)$ not labeled A .

X_B similarly except label B .

2.1.2 Dichotomous Event

Phenotypes that are represented by categorical, continuous, or discrete variables are dichotomized into an event state. The event may derive from a high-throughput genomic sequencing where the nucleotides of the P53 gene are known to have a genomic mutation or a clinical observation such as a pathologist evaluation of Ki-67 state. A laboratory cell line will have annotations for subtype such as luminal breast cancer and growth inhibition measurements in response to a drug such as GI50 (the concentration of drug necessary to inhibit growth to the 50% level, which is no growth) and a collection of mutations and other lesions. A patient sample may be annotated with clinical information, such as subtype, patient age, response to

treatment and other continuous and discrete variables. It is useful to have a clinician's assistance to help with determining which clinical information is likely to be useful, for example, that there is a clinical threshold for Ki-67 blood marker at 14%. When this type of information is not available, OCCAM automatically selects the mean of the sample range.

Formally:

Subtype	$X_{subtype}^+ =$ samples which are clinically diagnosed with the subtype
	$X_{subtype}^- =$ all other samples
Mutation	$X_{gene}^+ =$ samples which contain a significantly mutated gene
	$X_{mutation}^- =$ all samples not in $X_{mutation}^+$
Clinical	$X_{continuous\ clinical\ variable}^+ =$ samples $\geq$ critical clinical value
	$X_{continuous\ clinical\ variable}^- =$ samples $<$ critical clinical value

The critical clinical value for clinical variables for survival time, age, tumor size and others are simply the median or the mean value of the values of the cohort. In other cases, such as the Ki-67 blood protein marker, the critical value is derived from clinical practice.

2.1.3 Significance Analysis of Microarrays

The classic *Student's t test* is the usual method for comparing the difference in variation between samples. A problem occurs for genes with low expression levels. When their variance is very small, their z-score will be very large which will allow genes with low expression to aberrantly bias the results. Instead of the *t test*, OCCAM uses the Significance Analysis of Microarrays (SAM) method [Tusher 2001][Efron 2007] to derive activitome signatures. SAM generates a vector of fold differences in

genes in reactivity between the dichotomous classes similar to the z-scores in a statistical *t test*. Rather than using genes with low expression to calculate the signature, SAM uses a small strictly positive variance mitigating factor denoted s_0 , added to the denominator of the usual t statistic.

SAM output thus allows a user to find the most over-differentiated and under-differentiated genes. SAM and other gene comparison technologies suffer from the problem of making multiple comparisons. This can be addressed by an appropriate statistical adjustment called the *Bonferroni correction*. Unfortunately, the Bonferroni correction is overly conservative [Bender 1999][Tusher 2001] and would obviate most findings. Instead SAM provides a *false discovery rate* (FDR), a ratio of how many genes may have been discovered because of random chance compared to the number which are likely valid. An OCCAM activitome signature is a vector of SAM scores, therefore multiple hypothesis correction does not need to be applied to each individual score.

Formally:

$\mathbf{S}_A = \mathbf{SAM}(\mathbf{X}_{A+}, \mathbf{X}_{A-})$, is a random variable vector of *differential* scores (similar to *t*-test z-score) from the Significance Analysis of Microarray [Tusher 2001].

$$S_A(i) = \frac{\bar{X}_{A+}(i) - \bar{X}_{A-}(i)}{s(i) + s_0}$$

$$s(i) = \sqrt{\left(\frac{1}{n_{A+}} + \frac{1}{n_{A-}}\right) \left\{ \sum_{j \in A+} [X_j(i) - \bar{X}_A(i)]^2 + \sum_{j \in A-} [X_j(i) - \bar{X}_{A-}(i)]^2 \right\}}$$

S_0 = variance mitigating factor is chosen by an empirical bayes method to minimize the coefficient of variation of $\mathbf{S}$, see [Schwender 2003].

2.1.4 Pearson's R Correlation

A variety of statistics can be used to compare a vector of features. The easiest to interpret is Pearson's R correlation [Pearson 1900], which provides a metric of the similarity between two variables, X and Y . OCCAM applies the Pearson's R correlation to compare vectors to identify the correlation between events. The resulting correlation provides a measure of how associated or related two medical events are, such as whether a mutation is associated with drug resistance. An R-value of one indicates that X and Y are maximally linearly correlated. Obtaining unbiased correlations may sometimes be more complex than it first appears. An unbiased correlation may be created because the data that is being correlated contains the same sample on both sides of the correlation. This is illustrated in the following discussion that refers to using a metaphorical scale of correlation, which is introduced here but used later to develop the DIPSC method.

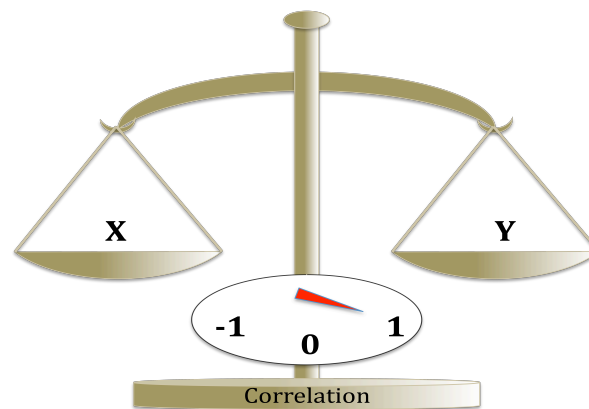


Figure 5 Pictorial illustration of Correlation. X and Y are a vector of sample data. Each sample is a vector of PARADIGM IPLs.

An R-value of one (1) indicates perfect correlation. An R-value of negative one (-1) indicates that X and Y are *anti-correlated*. Whereas values close to zero (0) indicate that X and Y are uncorrelated. Since the number of data points in the vector being correlated is typically large (greater than 200), correlations as low as 0.1 may be significant at the p-value < 0.05 level. The formula to calculate Pearson's R is a geometric summarization of the difference in vectors.

$$R = \frac{\sum_{i=1}^n [(X_i - \bar{X})^2 (Y_i - \bar{Y})^2]}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}, \text{ where } \mathbf{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad \bar{X} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{Y} = \frac{1}{n} \sum_{i=1}^n y_i$$

This is a well-studied formula and many subtle enhancements are possible. For example, it may be useful in the future to weigh the individual genes by SAM's false discovery rate.

$R_{A,B}$ is a random variable whose value is between -1 and 1 representing the similarity of samples [Pearson 1900]. The background distribution follows the correlation distribution [Fisher 1915], which is closely related to the normal distribution. Pearson's correlation formula with signature variables is:

$$R = \text{Corr}(A, B) = \frac{\sum_{i=1}^n [(S_A(i) - \bar{S}_A)^2 (S_B(i) - \bar{S}_B)^2]}{\sqrt{\sum_{i=1}^n (S_A(i) - \bar{S}_A)^2} \sqrt{\sum_{i=1}^n (S_B(i) - \bar{S}_B)^2}}$$

The normalization transformation of the vectors to the unit sphere (z-scores) such that the mean is zero and the standard deviation for each vector is one, allows the more efficient dot product to be used to calculate the correlation.

2.2 OCCAM Method

A continuous clinical phenotype, such as the concentration of which growth is inhibited by 50% (which is called GI50) of cell lines from the OHSU Gray lab is dichotomized into two groups by the mean. The resulting activitome signature may be used as a biomarker for indication of sensitivity to this drug. In this example, I use negative log values of the GI50, so that greater value means the cells are more sensitive to the drug. Lower values are an indication that the cell is more resistant. The second step is to compare and correlate this signature to other signatures. Beyond the scope of this thesis, there is work to do for others to use these signatures on individual patient data to predict treatment benefit.

The entire data flow schema of OCCAM is shown in Figure 6. Expression data such as that would be obtained from exon microarrays, exome sequencing, copy number assays, and protein assays are passed to PARADIGM method (described earlier) to produce the IPL matrix of X . GI50, Subtype and other clinical variables are passed to the dichotomizer to define the $A+$ and $A-$ class labels. The resulting IPL values are then fed to **SAM** to produce individual difference scores, which in aggregate are the S_A signature vector. For all vectors, A and B correlations values are calculated:

$$R_{A,B} = Corr(S_A, S_B)$$

Heatmaps and other visualizations such as these seen in later chapters are produced from these correlations.

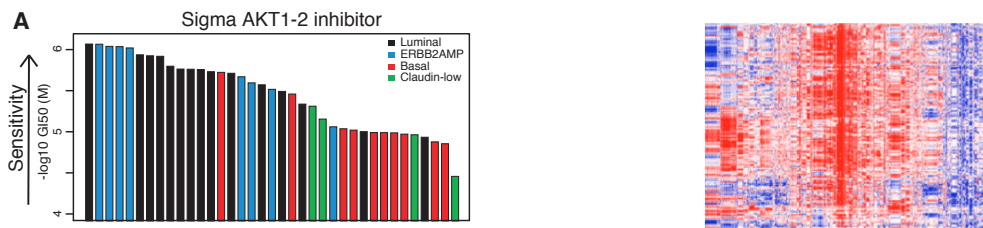
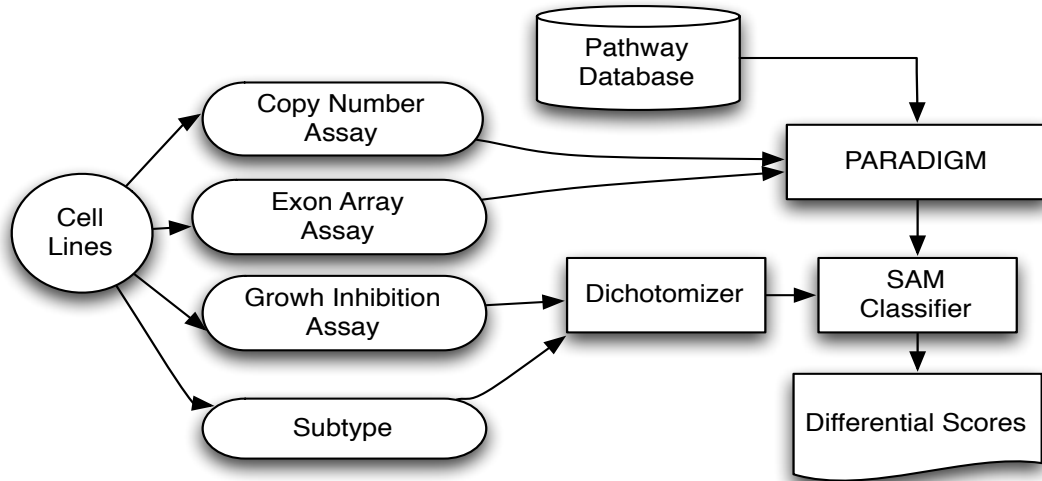


Figure 6 OCCAM Pipeline model. OCCAM finds & ranks differential pathway activity. GI50 depicted lower left (or some other measurement) dichotomizes the cell lines into + and – sensitive/resistant classes. Cell line expression, and copy number are processed through PARADIGM to obtain a score for each pathway concept. Then using SAM as a classifier, OCCAM processes the pathway events detected and the dichotomization to create a differential activity score, depicted as a heatmap lower right.

2.2.1 The Risk of Using OCCAM Naively

During a discussion of OCCAM results on TCGA breast cancer data, Charles Perou, a UNC researcher, made the important observation that the OCCAM analysis may be biased and confounded by the same samples being on both sides of the correlation. To deconfound the effect of many events on the same sample, I

developed an enhancement to OCCAM called *Differential Pathway Signature Correlation* (DIPSC) which solves this problem. The DIPSC method deconfounds sample co-occurrence by using a novel sub-sampling bootstrap-based approach. The bootstrap technique also provides the confidence interval and the variability of the correlation statistic that is integral to quality analysis.

Correlation is a powerful tool, but it must be used correctly. If a sample has multiple labels (as multiple mutations), **A** and **B**, then the sample will be on both sides of the correlation. This will confound the signal and bias the correlation measure higher.

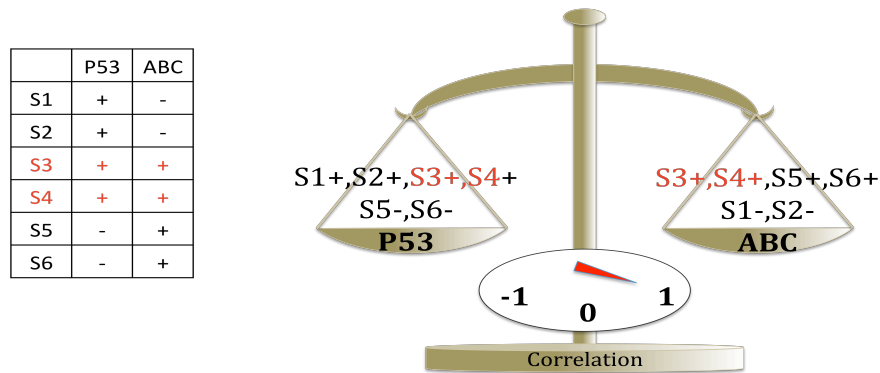


Figure 7 Depiction where samples S3 and S4 are on both sides of the correlation, which confounds the result.

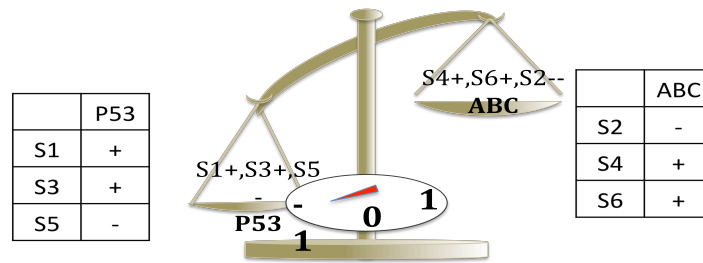
In most scientific fields outside of oncology and genetics, the solution to this problem is usually to obtain a larger cohort of samples. However, mutation driven diseases such as cancer require that every sample of every mutation be leveraged in the cohort. In cancer research, every patient may have a unique cancer because of his or her unique combination of mutations. Therefore, the need to be able to accurately analyze small sample sizes is more important in cancer biology than

other fields. Therefore many samples will have multiple mutations and thus are labeled both A and B , e.g. $(X_A \cap X_B) \neq \emptyset$. It is still useful to use the Pearson correlation statistic $R_{A,B}$, as long as the confounding caused by having the same sample on both sides of the correlation is removed. This problem is solved using the DIPSC method.

2.3 DIPSC Method

DIPSC overcomes this problem by computing signatures using a bootstrap [Efron 1990] formalism that always keeps the cohorts segregated. The bootstrap is a family of techniques based on resampling from an empirical distribution to obtain an approximation of a statistic. The bootstrap can be used to estimate the mean, the variance, and the confidence intervals of the correlations. I use the word *dip* to mean the more complex term that the literature calls a *sub-sampling replica*. The idea is to split X into a series of two virtual cohorts called α and β , such that every dip has the property $(X^\alpha \cap X^\beta) = \emptyset$. For the moment, the discussion will imagine that label A (both A+ and A- samples) will be measured in cohort α and label B will be measured in cohort β .

Solution, Resample 1000 Using Bootstrap



No Sample is on both sides of the scale

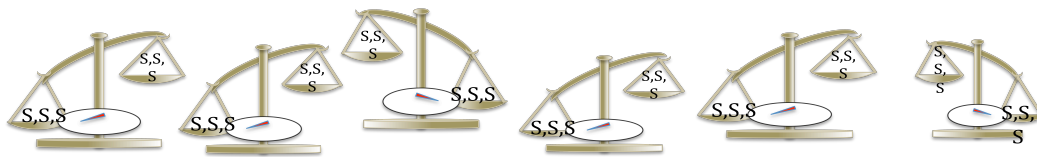


Figure 8 Depiction of Bootstrap Sampling, where the process of resampling called a dip is repeated D times, where typically $D=1000$. Samples in two different groups in a dip do not overlap.

DIPSC randomly splits the patient cohort in half. In each half, two different activitome signatures are calculated from two distinct contrasts (see Figure 8). A contrast corresponds to the dichotomy defined by the presence versus the absence of a particular label such as a mutation or a clinical outcome. The activitome signatures computed from each disjoint half are compared to one another. This guarantees that the comparison of the signatures is not confounded by any overlapping set of samples. The procedure of randomly splitting the cohort, re-deriving the activitome signatures with SAM, and comparing the signatures is repeated 1000 times, each repeat referred to as a *Dip* in the bootstrap process, to ensure statistical significance [Efron 1990]. The final correlation is then computed as a mean and standard deviation across the samplings.

2.3.1 DIPSC Formally

Here is a formal definition of the DIPSC method:

Let $\mathbf{X}^{boot} = \{ (\mathbf{X}^{\alpha_1}, \mathbf{X}^{\beta_1}), (\mathbf{X}^{\alpha_2}, \mathbf{X}^{\beta_2}), \dots, (\mathbf{X}^{\alpha_D}, \mathbf{X}^{\beta_D}) \}$ be D independent dips

drawn with probability 0.5 from $\mathbf{X}$, where $n/2$ samples are in $\mathbf{X}^{\alpha^*}$ and the remainder

are in $\mathbf{X}^{\beta^*}$. Various statistics such as the mean, variance, confidence intervals, and

correlation can be calculated because the empirical distribution function (edf) places mass $1/D$ at each observation [Efron 1997].

I define $\mathbf{R}^*_{A,B}$, the random variable estimate of the correlation of $\mathbf{A}$ and $\mathbf{B}$ as follows:

$\mathbf{X}^{\alpha_i}_{A+}$ = are the $\mathbf{A}+$ samples in the i -th α dip. Similarly for $\mathbf{B}$.

$\mathbf{X}^{\alpha_i}_{A-}$ = are the samples known not to be $\mathbf{A}$ in the i -th α dip. Similarly for $\mathbf{B}$.

$$\mathbf{S}^{\alpha_i}_A = \text{SAM}(\mathbf{X}^{\alpha_i}_{A+}, \mathbf{X}^{\alpha_i}_{A-}) \quad \mathbf{S}^{\beta_i}_B = \text{SAM}(\mathbf{X}^{\beta_i}_{B+}, \mathbf{X}^{\beta_i}_{B-})$$

$$\mathbf{R}^*_{A,B} = \frac{1}{2D} \left[\sum_{i=1}^D \text{Corr}(\mathbf{S}^{\alpha_i}_A, \mathbf{S}^{\beta_i}_B) + \sum_{i=1}^D \text{Corr}(\mathbf{S}^{\alpha_i}_B, \mathbf{S}^{\beta_i}_A) \right]$$

where Corr is the Pearson Correlation function defined above.

Since the samplings are symmetric, without loss of generality, it is possible to use the dip in both directions to double the number of samplings. Put another way, the $\mathbf{B}$ label in the α cohort can be correlated against the $\mathbf{A}$ label in the β cohort. This doubles the statistical power for the same number of samplings.

The 95% confidence intervals can be obtained directly from the bootstrap distribution [Efron 1990]. As in other bootstrap estimates, $\mathbf{R}^*_{A,B}$ is biased by the

noise. My simulations show that $R^*_{A,B} \rightarrow R_{A,B}$ as $D \rightarrow \infty$. In practice $D \rightarrow M$ converges when $M = 1000$. In very rare cases, $M=10,000$ may be necessary. This gives rise to the following useful *self-correlation* metric.

2.3.2 DIPSC Self-Correlation and Normalization

Since there are X_{A+} and X_{A-} samples in both the α and β dips, it is possible to perform a *self-correlation* within the bootstrap itself.

$$R^{auto*}_{A,A} = \sum_{i=1}^D \text{Corr}(S_A^{\alpha_i}, S_A^{\beta_i})$$

If there is no signal in the samples, R^{auto*} has a coefficient near zero and is therefore uncorrelated noise. The *self-correlation* metric is the maximum correlation possible and thus can be used as the normalization constant to scale correlations, which is shown in the various DipStick figures (such as Figure 14).

2.4 DIPSC vs. OCCAM Validation Using Simulated Data

How much better does DIPSC actually model the underlying correlation than OCCAM? The task is to demonstrate that OCCAM sometimes incorrectly identifies activitome signatures as similar that are not objectively similar because of sample co-occurrence. The simulation below shows that DIPSC deconfounds this problem and only identifies the objectively true similar signature.

I developed a simulation to test this hypothesis using the Gene Net Weaver toolset (GNW) [Schaffter 2011]. GNW is a highly regarded toolset for simulating

hierarchical scale-free network cell pathway models for many purposes including the DREAM4 In-Silico Network reverse engineering challenge [Greenfield 2010]. This tool allows me to generate gene expression values from a cellular network. The GNW tool purpose is to simulate an arbitrary set of knockouts to a cellular biochemical pathway network. An in-silico knockout is equivalent to a loss of function mutation in cancer. The GNW tool provides the Dream4 gene network, a eukaryotic cell pathway similar to one used by PARADIGM. GNW uses ordinary differential equations and various random noise parameters that are similar to the randomness found in many assays of gene networks.

The comparison process I developed analyzes a *simulated clinical study* where I use OCCAM and DIPSC to deduce which mutations are similar in a set of virtual patients each with a random set of mutations, by computing an activitome signature for each mutation and then correlating the signatures. Unlike a real clinical study, I can also obtain an *objective truth* because my simulated clinical study can be run on virtual patients in which I can limit the in-silico pathway alteration to a single knock out.

The objective truth is obtained by creating 101 reference activitome signatures in 101 virtual patients (one knockout gene per patient) independent of OCCAM and DIPSC in isolation. The activitome signature for Gene #0 knockout is designated the reference signature. I then correlate each of the other 100 signatures to Gene #0 and compute the p-value. The resulting objective truth in this run of the simulation is that nine (9) of the 100 activitome signatures have a p-value < 0.05 to Gene #0 (the

p-value of 0.05 is a good choice because of the noise parameters chosen by the Dream4 challenge). The objective truth for this specific network is that nine genes have similar knockout (mutation) activity to the reference gene. Why nine? This number represents the redundancy of Gene #0 in this simulated network.

Then, to compare OCCAM and DIPSC to each other and the objective truth, I performed random dual knockouts of 100 genes on the same virtual patients and created 3000 in-silico patient expression vectors using GNW by running the simulation 30 times. I used these 3000 GNW expression vectors to compute the 100 knockout activitome gene signatures using both OCCAM and DIPSC.

The similarity results depicted in the figure below show the striking contrast between the OCCAM and DIPSC methods. The figure shows the p-value for each of the gene signature plotted along the horizontal axis sorted by DIPSC's correlation p-value. The nine significantly correlated genes are indicated below in the green box. OCCAM incorrectly identified an additional 71 signatures as significantly correlated, whose p-value < 0.05 is shown as being below the dashed green line, but not in the green box. Whereas, DIPSC only identified the nine known activitome gene mutation signatures, the objective truth. The reason OCCAM identified an additional 71 knockout signatures as significant is because of sample co-occurrence as explained earlier.

OCCAM is still an excellent tool for accurately comparing non-overlapping sample signatures. DIPSC is computationally more complicated and is only necessary when signatures from overlapping samples need to be analyzed.

OCCAM vs. DIPSC Signature Similarity

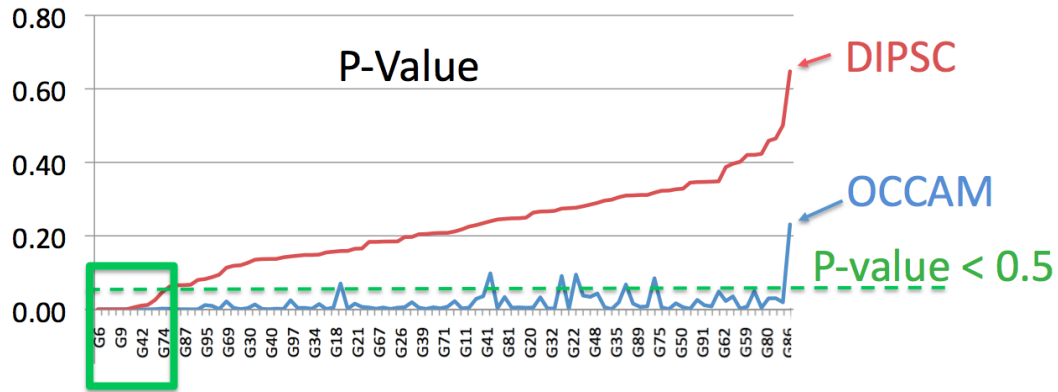


Figure 9 OCCAM and DIPSC chart of the evaluation of systematic GNW gene knockout simulation. The chart shows the p-value significance score. The X-axis is simulated genes, sorted by DIPSC p-value. Only genes below the green line have significance of < 0.05 . OCCAM (blue line) incorrectly shows 80 genes are significant, whereas DIPSC correctly identifies the nine objectively true gene knockouts whose signatures are significant (those within the solid green box).

2.4.1 Other Quality Metrics

Not all events generate valid activitome signatures. A valid activitome signature for a given event derived from one set of samples should correlate (with $p\text{-value} < 0.05$) with the activitome signature derived from another similar set of samples. This *self-correlation* provides a baseline indication of whether there is signal or whether the noise dominates the signal. Like any other noise measure, the self-correlation is influenced by the number of samples with a given phenotype. Below in Figure 10 is the self-correlation of each of the top 100 mutations and whether the patient was luminal A or B in the TCGA luminal breast cancer 347 patient cohort (185 luminal A, 162 luminal B). The self-correlation also provides the normalization constant for display in the DipStick figure shown in Figure 14.

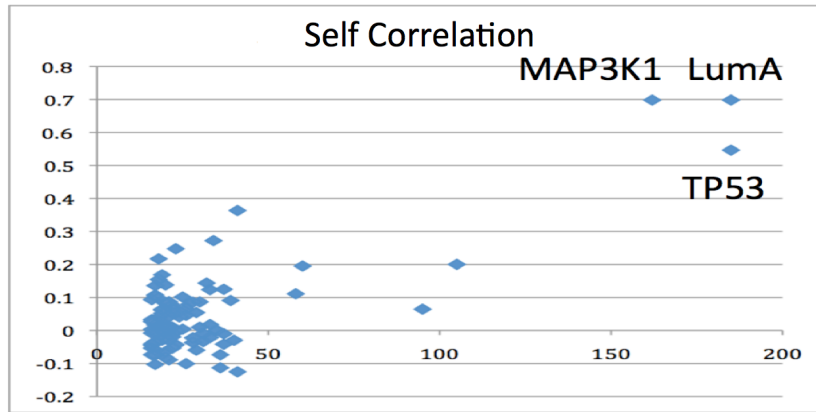


Figure 10 DIPSC Self-correlations using 100 significant mutated genes in the TCGA dataset. The x-axis is the sample count; the y-axis is the R value. Strength of self-correlation improves with sample count and strength of correlation. Strongest signal is luminal A at $R=0.7$, p-value is near zero. Correlations below 0.1 are insignificant (p-value > 0.5).

The strongest signal is the luminal A versus luminal A (as well as luminal B vs. luminal B) signal at $R=0.7$ and MAP3K1. Interestingly, TP53 has a lower self-correlation signal at $R=0.55$, even though there are more TP53 mutated samples than MAP3K1 samples, which may indicate that TP53 mutations have multiple modalities, for example both gain of function mutations and loss of function mutations.

2.5 Summary

In this section, I have presented the methods and the underlying mathematics for OCCAM and DIPSC. These methods create, compare, and correlate activitome signatures. OCCAM uses Efron and Tibshirani's SAM method to create these signatures. The DIPSC method compares signatures while accounting for the confounding that stems from sample overlap. Mutations in different genes may cause disruptions in the same pathway, which may lead to similar disruptions in the activitome signature. By comparing the vectors of activitome

signatures of different mutations and clinical outcomes, intrinsic connections between these events may be uncovered. These are useful for guiding therapy, applying treatments for one cancer to another cancer, and developing new therapies.

3 Clinical-Translational Results

Without translational applicability, inferring gene activities with pathway knowledge would be no more than an academic exercise. In this section, I describe how the knowledge encoded in the activitome can be clinically translated to improve patient outcomes.

3.1 SU2C Breast Cancer Drug Sensitivity Activitome

One of the most important efforts of the Stand Up 2 Cancer Breast Cancer Dream Team project was to better understand drug sensitivity and resistance. Utilizing cell line data provided by the Gray lab, I applied OCCAM and DIPSC to identify the associations between molecular subtypes, pathways, and drug response.

This spectacular data was derived from wet-lab experiments of growth inhibition (GI50) assays of 77 compounds, testing across 54 cell lines. The Gray lab provided GI50, exon expression, and copy number data to UCSC. I used PARADIGM to integrate genome copy number and transcription data with approximately 1400 curated signal transduction, transcriptional and metabolic pathways to identify differences in pathway activity among the subtypes and drugs in the cell line panel. The pathways most associated with disease subtype and drug sensitivity and resistance were used to identify new drug targets and biomarkers that could guide clinical practice [Heiser 2011].

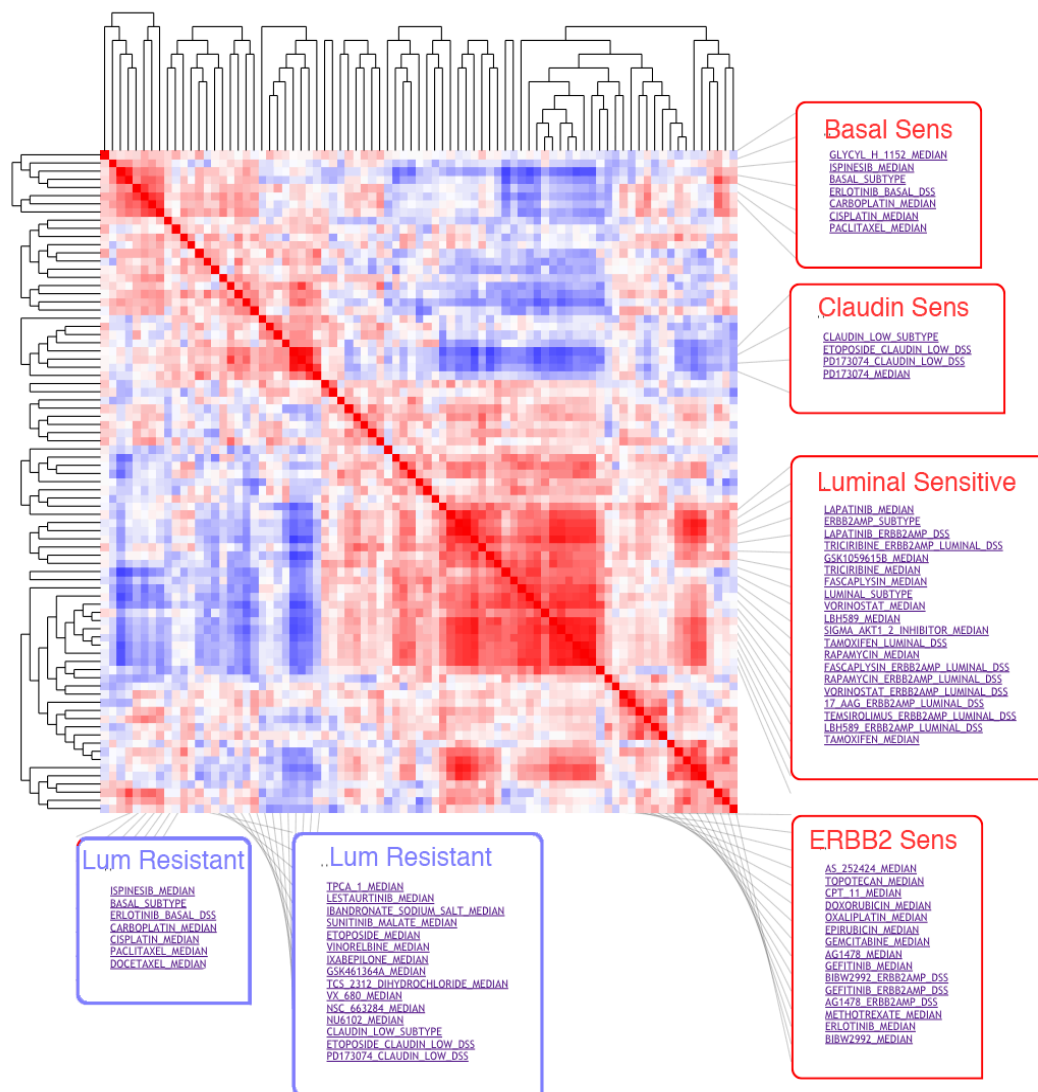


Figure 11 Drug Resistance and Sensitivity Activitome Signature Heatmap. Red indicates sensitive and Blue indicates resistant signatures. Inset boxes list signatures for each of the basal, luminal and claudin low and Her2 breast cancer subtypes.

Using heatmaps hierarchically clustered by Cluster 3.0 [Eisen 1998] [de Hoon 2004] in MedBook's dynamic heatmap tool transformed the DIPSC data into the visualization in Figure 11, summarizing drug resistance and sensitivity among the cell lines. The largest correlation group is the luminal sensitive drug group containing familiar luminal breast cancer treatments including lapatinib and tamoxifen, as well

as drugs that are not known to be used for breast cancer such as rapamycin. The figure also shows that drugs such as vinorelbine and etoposide are not good choices for luminal breast cancer. For basal subtype breast cancer, DNA damaging agents such as cisplatin and carboplatin; or microtubule inhibitors such as paclitaxel could be good choices. Erlotinib, which targets EGFR, may be a good choice for some basal breast cancers. Using the SuperPathway reveals pathway activities that provide information about mechanisms of response. A key subpathway from each of these subtypes is depicted in Figure 12. These pathways are visual components of the activitome signature that describes the cancer subtype.

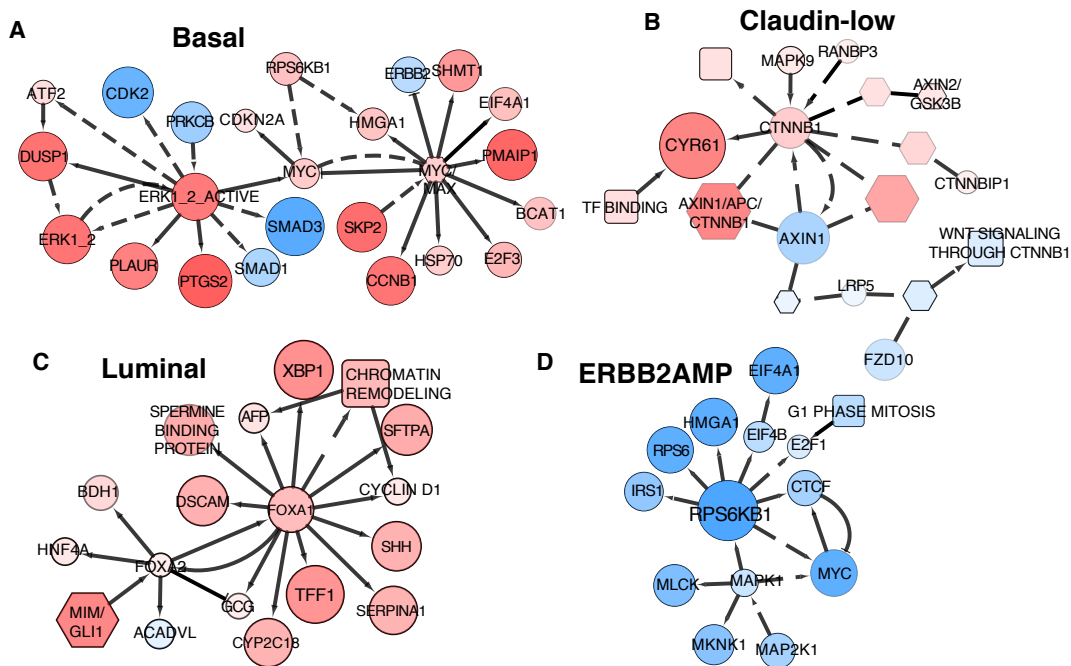


Figure 12 Cell line subtypes have unique network features. In the above figure created by my colleague Sam Ng, each node in the graph represents a different pathway “concept” corresponding to either a protein (circles), a multimeric complex (hexagons), or a cellular process (squares). The size of each node was drawn in proportion to the differential activity score such that larger nodes correspond to pathway concepts with activities more correlated with subtype (basal, luminal, etc) versus non-subtype cell lines. Color indicates whether the concept is positively correlated (red) or negatively correlated (blue) with the basal subtype. Links represent different interactions, including protein-protein level interactions (dashed lines) and transcriptional (solid lines). See text for discussion.

Interactions were included in the map only if they interconnect pathway concepts whose absolute level of differential activity are both higher than the mean absolute level. Figure 12A shows a basal-associated PathMark [Ng 2012] activitome subnet in which is visible the upregulation of the MYC/MAX subpathway associated with proliferation, angiogenesis, and oncogenesis; and upregulation of the ERK1/2 subpathway controlling cell cycle, adhesion, invasion, and macrophage activation.

The FOXM1 and DNA damage subpathways also were upregulated in the basal cell lines. FOXM1 normally controls cell cycle progression during S and G2/M phases through its interaction with Plk1, cyclin B2, Nek2, and CENPF, and plays an important role in the maintenance of chromosomal segregation and genomic stability.

3.1.1 Claudin-low Activitome Signature vs. Basal Activitome Signature

A pathway comparison of the claudin-low activitome signature to the basal activitome signature shows many of the same pathways upregulated and downregulated. One exception is the beta-catenin (CTNNB1) pathway, which is upregulated in claudin-low as compared to basal (Figure 12B). This suggest that MYC activity may be further differentially activated in claudin-low subtype samples as beta-catenin is known to activate MYC, leading to de-differentiating programs that make the cell behave more like a stem-cell. Comparison of the luminal cell lines with all others showed downregulation of an ATF2 network, which inhibits tumorigenicity in melanoma, and upregulation of FOXA1/FOXA2 networks, which control transcription of ER-regulated genes, and are implicated in good prognosis for luminal breast cancers (Figure 12C). Comparison of ERBB2AMP cell lines with all others

showed many network features common to luminal cells—not surprising because most ERBB2AMP cells are of the luminal subtype.

A validation of my method was to rediscover the mechanisms of action for known drugs. ERBB2AMP cell lines are sensitive to geldanamycin, a highly toxic inhibitor of HSP90, and show upregulation in the ERBB2-HSP90 subpathway. This observation is consistent with the mechanism of action for geldanamycin: it binds to and inhibits HSP90, leading to proteasomal degradation of ERBB2.

Significant downregulation of pathways are also known to be implicated in drug resistance. This is depicted in Figure 12D, which shows downregulation centered on RPS6KBP1 in ERBB2AMP cell lines. RPS6KBP1, a known drug target, is the ribosomal protein S6 kinase (70 kDa) beta/polypeptide 1 [Mills 2001]. RPS6KBP1 is part of the S6K map kinase-signaling cascade that is involved in both proliferation and translation of proteins. S6K is also part of the tumor suppressor pathway and interacts with PTEN phosphatase. Loss of the PTEN pathway may contribute to tumorigenicity. My results depicted in Figure 12C show that RPS6KBP1 is significantly downregulated in ERBB2AMP cell lines and that this may contribute to both tumorigenicity and to drug resistance.

As I described earlier, another result revealed by the activitome signature is that the basal cell lines are preferentially sensitive to cisplatin, a DNA damaging agent. To understand why this may be, I observe that upregulation of the DNA-damage response subpathway includes ATM, CHEK1 and BRCA1, key players associated with response to cisplatin (Figure 13A). The effects of ATM and CHEK1

are known in other cancers but have not been previously shown to be involved in breast cancer. A key prediction from our data is that PTGS2 (prostaglandin-endoperoxide synthase) also known as COX2 may be a useful biomarker to guide NSAID COX2 targeted therapy for basal cell breast cancer.

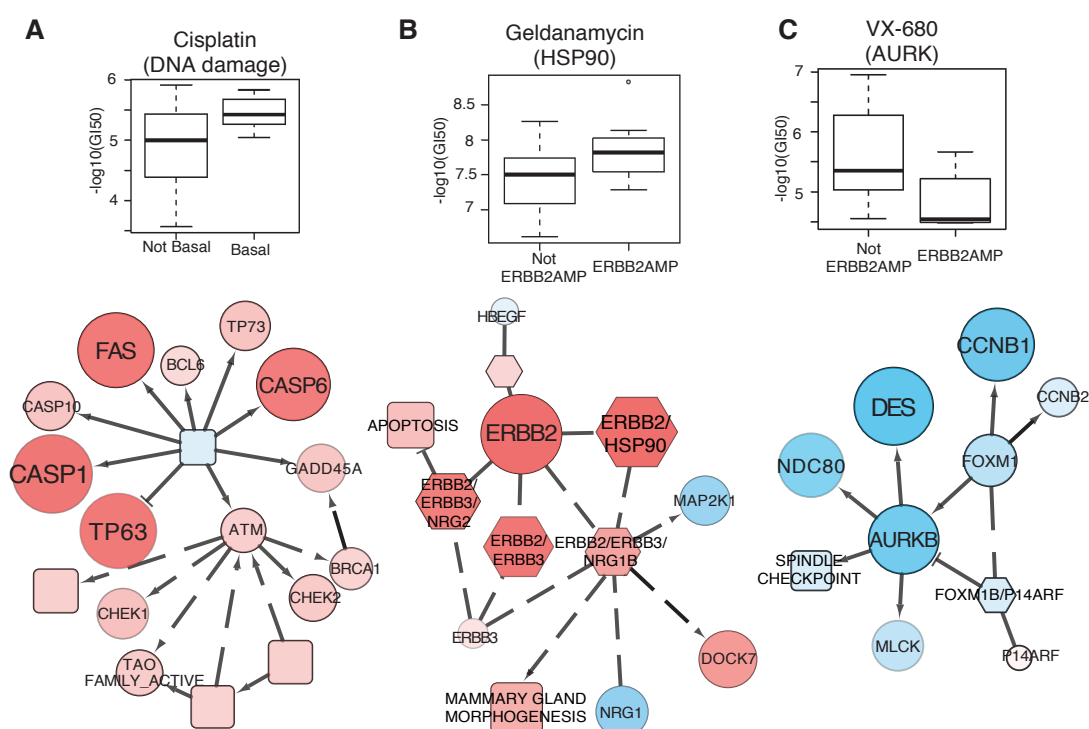


Figure 13 Differential Activities Across Drugs. Identifying pathways that contribute to drug resistance is a critical part of this research. The research found that the ERBB2AMP cell lines were resistant to the aurora kinase inhibitor VX-680 (panel C, upper), and further that sensitivity to this compound was not associated with amplification at 20q13 (AURKA). This raises the possibility that this resistance may be mediated through CCNB1, which is co-regulated with AURKB by FOXM1. Of the four subtypes, ERBB2AMP is the only one that shows substantial downregulation of CCNB1 (panel C). This proposed mechanism is supported by the observation that in primary tumors, CCNB1 gene expression is significantly correlated with AURKB gene expression.

Pathway activities predict responses to therapies and correlate with subtype.

In Figure 13A, created by my collaborator Laura Heiser, basal breast cancer cell lines preferentially respond to the DNA damaging agent cisplatin. The lower panel shows basal cell lines having enhanced activity in pathways associated with the DNA damage response, providing a possible mechanism by which cisplatin acts on these cell lines. In Figure 13B upper panel, ERBB2AMP cell lines are sensitive to the HSP90 inhibitor geldanamycin. In the lower panel, the ERBB2-HSP90 network is upregulated in ERBB2AMP cell lines. In Figure 13C upper panel, ERBB2AMP cell lines are resistant to the aurora kinase inhibitor VX-680. In the lower panel, resistance

may be mediated through co-regulation of AURKB and CCNB1, which could be targeted by small molecule adjuvant drugs. Combining drugs to neutralize drug resistance will increase the benefits of existing therapies. Similar approaches have worked successfully for decades in anti-infectives such as *augmentin* [Rollinson 1982]. As well, combination drug therapy may allow drug administration with lower toxicity and therefore fewer side effects to the patient.

3.1.2 Drug Biomarkers

The basal-luminal contrasting activitome signature and SuperPathway together were used to identify significantly large sub-networks that connect high-scoring pathway components to drug biomarkers. Sub-networks were created by retaining any interaction that connected two features, where both of which, had absolute association scores higher than the average absolute association score. Among the largest of the hubs in the resulting network was a central DNA damage hub with the second highest connectivity (55 regulatory interactions; 1% of the network) and TP53 with the 14th highest connectivity (26 connections; 0.5% of the network). The sub-network identified several pathways of interest including the FOXM1-related network. Several genes upstream of FOXM1 are known targets of available drugs, including PLK3, suggesting polo-kinase inhibitors may disrupt basal tumors. Indeed, basal breast cancer cell lines treated with polo-kinase inhibitors were found to sensitize basal cells to a higher degree compared to luminal cells, consistent with the prediction encoded in the network.

The cell line activitome signatures may contain cell line artifacts that would not exist were the same signatures derived from patients. Equally challenging is when I apply the breast cancer cell line activitome signatures to other body organs because the tissue specific background expression may confound the signal.

Nevertheless, these preliminary results are interesting and provide insights that will be more fully appreciated once there is a set of tissue-specific expression deconvolutions. Fortunately, these are being developed as part of the TCGA PANCAN study [Stuart 2013].

3.2 Activitome Signatures Provide Clues about Cellular Targets

Gene expression signatures have been used to identify putative drug targets. Some examples include the Connectivity-Map project pursued by the Golub lab at the Broad Institute, the Ailun project by the Butte lab at Stanford [Chen 2007], and the Disease Diagnostic Gene Expression database by the Zhou lab at USC [Huang 2010]. In the same way, activitome-based signatures can be used to build predictive signatures. Activitome signatures provide potentially much richer information because the pathway interactions can reveal cryptic signals such as active transcription factors and signaling molecules that could go unnoticed by looking at gene expression alone.

3.3 Towards Patient-specific Therapeutic Models

Clearly, activitome development for clinical care is in an early stage of development. The data-driven discovery approach requires pooling data from many patients and data sources to build functional inferences. The challenge ahead of us is

to develop single sample predictors that can guide therapeutic decisions in individual cases with the aid of the activitome. The knowledge base I envision is a constantly refreshed collection of signatures to represent all known cancer subtypes and treatment responses. To identify a patient-specific therapy then would require two conceptual steps: 1) From the database of signatures make an identification of the most representative signatures for a particular patient sample; and 2) Refine the model to best fit the particular set of genomic, epigenomic, and proteomic changes observed in the patient's data.

This approach leverages the statistical power of multiple samples to define the starting subtype models but also encompasses the flexibility to adapt to a particular form of the disease. Just as in gene expression-based models, activitome-based models will require the careful acquisition of samples from well-conducted clinical trials that are sufficiently powered for the full suite of genomic and proteomic analysis methods executed at the CLIA clinical laboratory grade testing level. While clinical utility is still several years away, the activitome approach described here provides the basis for achieving this goal. It underscores the value of the deeply collaborative environment provided by our rapidly growing bioinformatics and computational biology discipline, and many teams of clinicians, and genome and proteome centers that provide us with data to analyze. To accomplish this goal requires assembling the activitome, the first steps of which are described in the next section.

The activitome is only useful if its collection of signatures can match cancer's genomic alterations to effective treatments and outcomes. Section 3.4 will begin explaining the approach in the narrow confines of one carefully managed study that took place at one medical center. Activitome signatures will be most useful, however, if there is an effective interpretable comprehensive knowledge base that connects the signatures to the clinical knowledge and mutation phenotypes derived from many studies across many medical centers.

Fortunately, several large initiatives are under way to create large-scale catalogs of cancer information. Unfortunately, the outcome data provided by these surveys are limited. I have begun to populate the activitome signature database using data from The Cancer Genome Atlas projects and Stand Up 2 Cancer Breast Cancer Dream Team. However, the ultimate solution for providing the outcome data will be realized by the MedBook vision. In the mean time, I have been very fortunate to have cell line data that provided a useful proxy for many patients and was an excellent tool to help develop OCCAM and DIPSC. Other interesting cell lines which may be incorporated in the future include the classic NCI-60 multi-cancer survey and the comprehensive 1000+ Cancer Cell Line Encyclopedia (CCLE).

However, created cell lines are not perfect. The problem with cell line data is that it provides a biased proxy for patients with tumors that were amenable to making cell lines. These patients usually did poorly which is why they were referred to academic care centers that had the facilities to make cell lines. The true activitome should come from direct point of care for patients at both major

care centers and community care centers, and should include both data from successfully and unsuccessfully treated patients as well as from healthy non-patients.

3.4 Comparing Activitome Signatures Reveals Novel Connections Between Mutations and Drug Response in Luminal Breast Cancers

Activitome signatures can also be used to connect mutations, clinical outcomes, and other “events” present in tumor samples. It is often of interest to know whether a particular mutation is associated with elevated risk or the possibility of developing resistance to a particular treatment option. Molecular signatures derived from such events can be used as proxies to predict such tendencies. As an example, pathway-based activitome signatures were successfully used to analyze a set of patient tumors of the luminal breast cancer subtype (both luminal A and luminal B subtypes) [Ellis 2012]. In this study, clinical and genomic data from samples were assessed from a neoadjuvant aromatase inhibitor (AI) clinical trial designed to assess the responsiveness of patients to these estrogen-lowering agents [Ellis 2011]. This study at Washington University used a rigorous clinical model, which had separate discovery and validation patient cohorts [Ellis 2012][Goldstein 2013].

OCCAM was used to build a predictive model for AI therapy and to develop links between gene mutations and clinical outcomes. These outcomes were chosen from an unbiased group of luminal breast cancer patients. PARADIGM analysis revealed that multiple pathways are affected by a phalanx of mutations including caspase/apoptosis, ErbB signaling, Akt/PI3K/mTOR signaling, TP53/RB signaling

and MAPK/JNK pathways. Several “hubs” such as ESR1 and FOXA1 were activated cohort-wide while other hubs exhibited high but differential changes in aromatase-inhibitor-resistant tumors including MYC, FOXM1 and MYB.

DIPSC was applied to the luminal samples using the PARADIGM-derived activitome signatures to uncover phenomena that underlie the resistance of some cancers to aromatase inhibitors. All pairs of associations were scored across both discovery and validation cohorts. Any association that was significant in the discovery cohort was only reported if it was also significant in the validation cohort. An example of the association of mutations to subtype is illustrated in the DIPSC dipstick plot shown in Figure 14, which plots the correlation of all activitome signatures against the luminal B vs. luminal A dichotomy activitome signature. From this visualization, one can immediately see what patient groups lead to common signatures. The analysis revealed, for example, that mutated MALAT1 (a small non-coding RNA) had activitome signatures similar to TP53 mutations and are associated with both high Ki-67 and high preoperative endocrine prognostic index (PEPI) scores that are indicators of resistance to drug treatment. Ki-67 is a prognostic indicator of proliferation in breast cancers. Because MALAT1 is mutated in only a handful of samples, several analyses used to detect such relationships were not performed because of the low samples size, whereas DIPSC was able to leverage the robustness of the PARADIGM IPLs to find a significantly indicative pattern. On the other hand, PIK3CA, MLL3, and CDH1 do not correlate for either luminal subtype. ATR, and MAP2K4 are correlated for luminal A and MAP3K1 mutations are strongly

correlated with luminal A signatures. Thus, correlating cancer outcome signatures to mutation signatures revealed novel connections not available to standard approaches. This provides a facile way to identify genomic events that distinguish even closely related subtypes of a disease (such as Luminal A versus Luminal B).

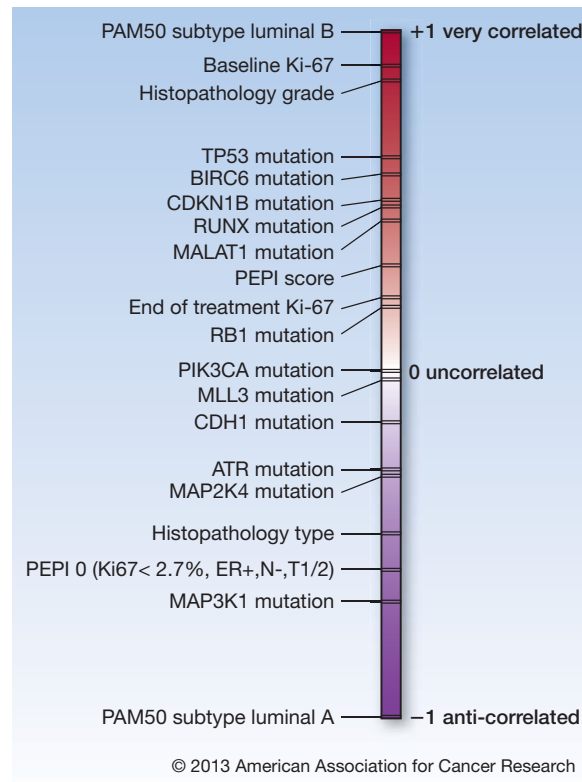


Figure 14 DipStick activitome signatures correlation to luminal B vs. luminal A dichotomization activitome signature.

3.4.1 Summary of my contributions

In this study, I found that activitome mutation signatures from different mutations accurately correlate with patient responses to aromatase inhibitors. In particular, Luminal B signature correlated with TP53, RB1, RUNX1 and MALAT1 Signatures (P-Value < -4.4×10^{-6}). MALAT1, a long noncoding RNA, is a surprise and has not been previously implicated to have a role in breast cancer though it has a

role in other cancers. I observe that MALAT1 mutations may cause similar loss of cell growth regulation to TP53 mutation, RB mutation, and RUNX1 mutation.

MALAT1 is a known RNA regulatory gene with a role in lung cancer [Gutschner 2012]. Its name derives from metastasis associated lung adenocarcinoma transcript 1. It also may have a role in normal cell motility regulation, which would also explain its role in metastatic cancer. PathMark analysis identified a number of functional hubs as consistent aspects of aromatase inhibitor resistant tumors. In particular, novel observations of ESR1 and FOXA1 were found to be activated cohort-wide while other hubs exhibited high but differential changes in aromatase-inhibitor-resistant tumors including MYC, FOXM1 and MYB.

It should be possible to test the role of MALAT1 using RNA inhibitor techniques to knock the MALAT1 gene down in existing cell lines to see whether it creates an activitome signature that more closely correlates with the MALAT1 signature.

Both the high level activitome signatures and the hub observations used clinical trial style investigations that first find hubs in the discovery cohort and then assure the discoveries in the validation cohorts. This discipline builds trust in bioinformatic techniques and in particular increased the confidence of my collaborators in the OCCAM method. In general, I observe that genomic studies require patient consent, high quality samples, and detailed outcomes from clinical trials. The way forward is to develop a series of prognostic activitome signatures.

These signatures need to be independently validated in additional clinical trial cohorts.

3.5 TCGA Breast Cancer (BRCA)

Reflecting on both the cell line project and the clinical trial project provokes a number of questions including: how well do the cell lines match the tumors in people and can the cell line drug activitome signatures be used to predict drug sensitivity in actual breast cancer patients? The first question was answered by comparing the cell line data to TCGA breast cancer data.

There is a general concordance of subtype between the SU2C BRCA drug samples and the TCGA Breast Cancer BRCA samples visible in Figure 15. The TCGA BRCA marker paper was a significant undertaking and accomplishment by many groups. Overall, 510 tumors from 507 patients were subjected to whole-exome sequencing, identifying 30,626 somatic mutations comprised of 28,319 point mutations. I analyzed the 100 most significantly mutated genes and their connection to a handful of phenotypes. Drug information and patient outcome data was not available, however subtype, mutation, copy number and gene expression information were used in this analysis. TCGA team members used the MuSiG package [Dees 2012] to determine the significance of the observed mutation rate of each gene based on the background mutation rate, and identified 35 significantly mutated genes with false discovery rate (FDR) <5%. My analysis compared this list as well as pathological type and breast cancer subtype to identify associations between mutated

genes and the subtype. Luminal-like cancers are distinguished from basal-like cancers by the presence of a hyperactivated FOXA1–ER complex (see Figure 15).

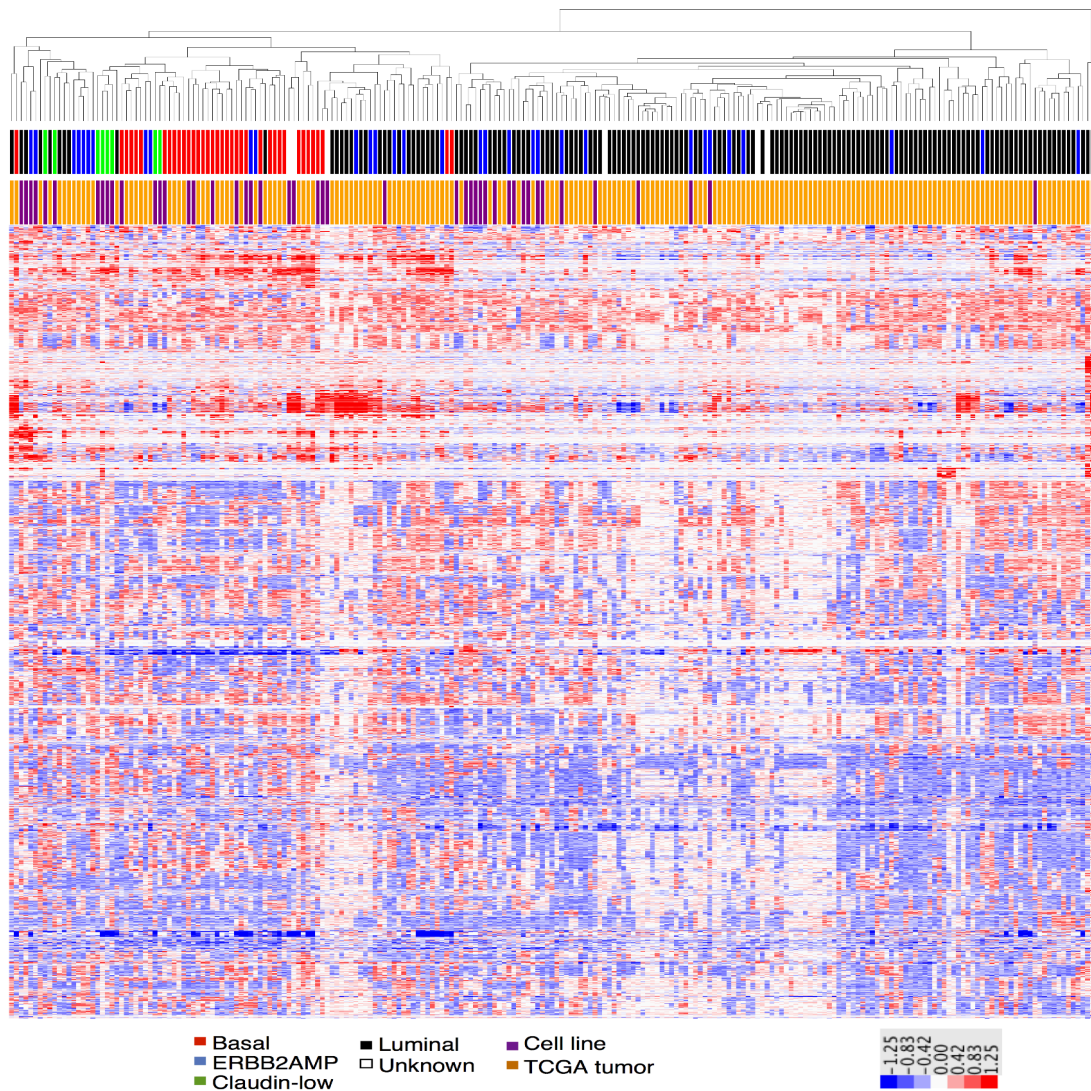


Figure 15 Concordance between TCGA patient and Gray Cell line. Heatmap of non-redundant PARADIGM activities for both cell line and TCGA samples. Cluster dendrogram represents Euclidian distance between samples and was created using Eisen Cluster and drawn using Java Treeview. Each row represents a network feature, each column represents a sample (tumor or cell line). Colored bars below dendrogram indicate sample subtype (upper) and sample cohort (lower). Overall, cell lines and tumors are intermixed, and subtypes tend to cluster together, indicating that the tumors and cell lines share many of the same network features. From PNAS Figure S4 [Heiser 2011], permission granted to authors via <http://www.pnas.org/site/aboutpnas/rightperm.xhtml>

In the TCGA data set, expression features of basal-like tumors include a characteristic signature containing keratins and high expression of genes associated

with cell proliferation. The analysis of basal versus luminal tumors emphasized the importance of hyperactivated FOXM1 as a transcriptional driver of this enhanced proliferation signature. PARADIGM also identified hyperactivated MYC and HIF1- α /ARNT network hubs as key regulatory features of basal cancers. Even though chromosome 8q24 is amplified across all subtypes, high MYC activation seems to be a basal-like characteristic. The DipStick in Figure 16 shows the DIPSC correlation of mutations to TCGA BRCA Luminal A vs. all other subtype dichotomy.

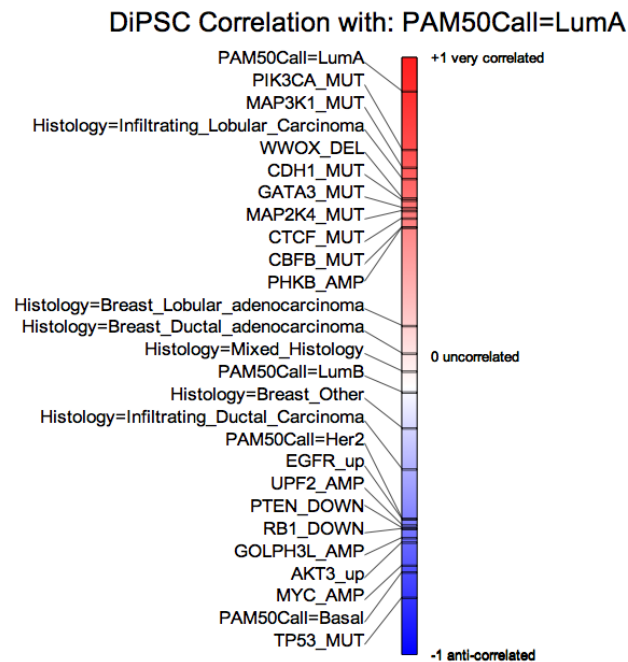


Figure 16 DipStick correlation of TCGA breast cancer mutation, histology, and subtype events.

The DipStick figure, while attractive, does not show the significance. So, I recently re-analyzed Gray cell line activitome drug signatures against the TCGA BRCA data set and created a *Volcano Plot* (see Figure 17) which places correlation along the X-axis and p-value significance along the Y-axis.

The red labels and points indicate significant positive correlation. The blue points (no labels) indicate no significant correlation. The purple labels and points indicate significant anti-correlation.

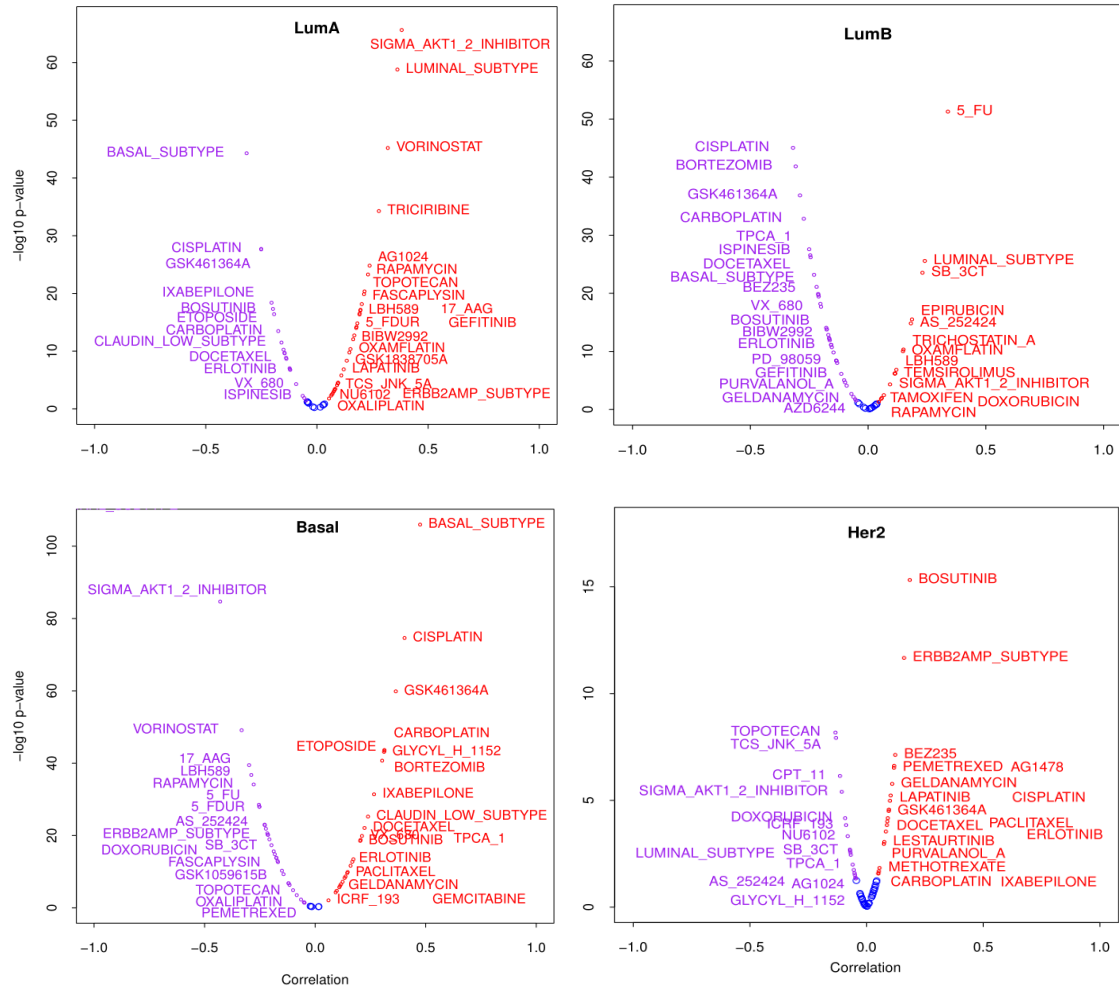


Figure 17 Volcano plot of TCGA BRCA patient subtypes activitome signature (major heading) correlated to Gray cell line drug sensitivity signatures (inset labels). Red indicates the correlation that patients with this cancer subtype will respond to this drug. Purple is anti-correlated, blue dots indicates unlabeled drugs whose p-value is not significant at the p-value < 0.05 level.

Interestingly, Sigma's AKT 1-2 inhibitor, which is indicated for luminal A breast cancer, correlates more strongly with the TCGA luminal samples and the

luminal cell line LUMINAL_SUBTYPE signature. This is because cell line data combines luminal A and B samples.

3.6 TCGA Colorectal Cancer

Colorectal cancer, like breast cancer, is another complex cancer where it would be useful to have a well-understood set of subtypes to drive treatment. Despite the diversity in anatomical origin or mutation levels, nearly 100% of these tumors have changes in MYC transcriptional targets, both those promoted by and those inhibited by MYC. These findings are consistent with patterns deduced from genetic alterations and suggest an important role for MYC in colorectal cancer.

The DIPSC analysis of the TCGA colorectal data set shown in Figure 18 correlates activitome signatures into six major areas distinguished by the anatomical region of the mutation and basic colorectal subtype micro-satellite instable (MSI) phenotype vs. micro-satellite stable (MSS) phenotype, all of which correlate better than $R=0.2$ and with p-value near zero. However, one exception is the Beta Catenin CTTNB1 mutation activitome signature that has a distinct signature, and is unlike any of the other tissue locations, possibly because it occurs in all colorectal cancer tissues.

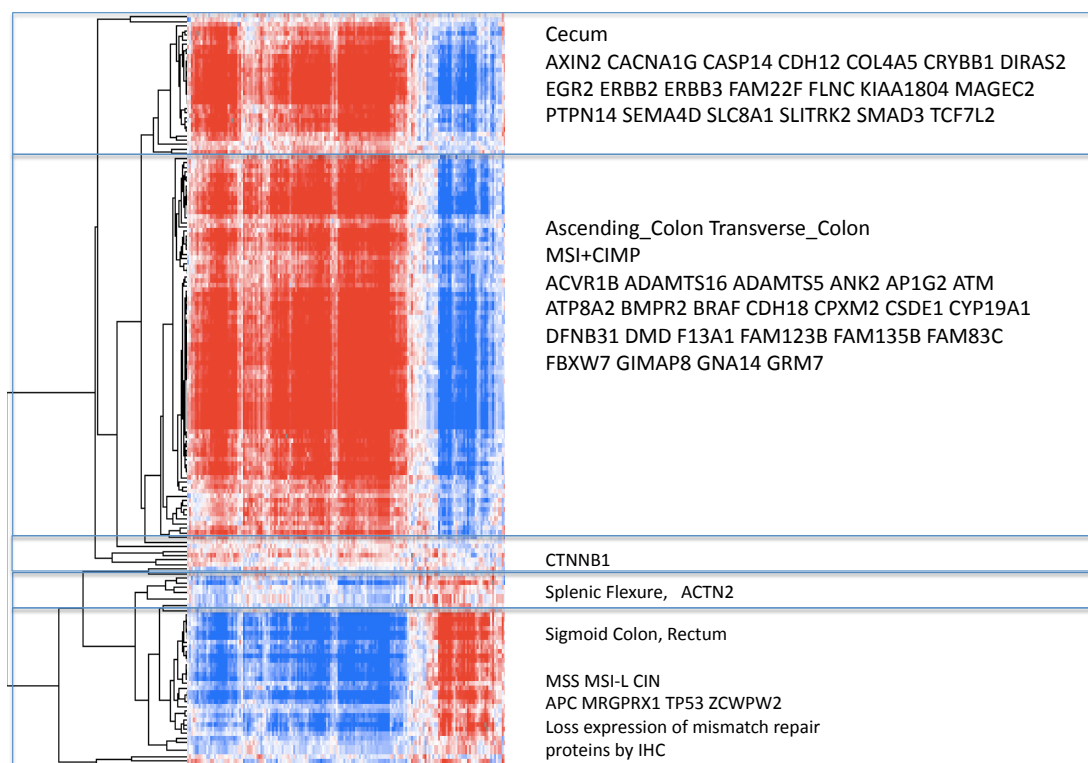


Figure 18 TCGA Colorectal cancer correlation of correlations of activitome signatures. Gene names indicate mutation activitome signatures, all others are clinical assay activitome signatures or location activitome signatures. Signatures show distinct upper vs. lower colon division. Most mutations correlate with a specific region.

New insights come from correlating mutational activitome signatures with drug activitome signatures. I was interested in whether there may be a correspondence between cancer subtypes in different organs. In Table 1, I show the most significant activitome correlations between breast cancer cell line subtypes and the colorectal cancer location and tissue. The correlation for most subtypes is marginal. The only p-value that survives Bonferroni correction is luminal to rectal adenocarcinoma. Larger cohorts than TCGA may reveal stronger correlations.

Gray BRCA Cell line Subtype	TCGA COLOREAD Patient Location and Histology	Pearson's R	Uncorrected p-value	Bonferroni corrected p-value
BASAL	Colon Adenocarcinoma	0.08	5.0E-04	2.9E-13
	Colon Mucinous Adenocarcinoma	0.02	0.40	
CLAUDIN LOW	Colon Mucinous Adenocarcinoma	0.07	3.3E-03	
	Rectal Adenocarcinoma	0.01	0.61	
ERBB2AMP	Rectal Mucinous Adenocarcinoma	0.08	5.5E-04	
	Rectal Adenocarcinoma	0.03	0.18	
LUMINAL	Rectal Adenocarcinoma	0.20	2.7E-17	
	Colon Adenocarcinoma	0.01	0.73	

Table 1 Gray cell line activitome signature subtype compared to colorectal location and histological type activitome signature.

However, the same correlation analysis of the drug activitome signatures from the Gray cell line against the colorectal activitome signatures tells a different story. It reveals many significant correlations that do survive Bonferroni correction. In Table 2, I show the correlations between activitome drug signatures and other activitome signatures derived from colorectal location and mutational event that have very significant Bonferroni corrected p-value. This implies that drug activitome signatures may be viable across different cancers. For example, my predictions suggest that vorinostat, an HDAC inhibitor, may be useful for treating rectal adenocarcinoma. This is consistent with my finding that vorinostat is implicated from analysis of the luminal breast cancer and my observation that rectal cancer has significant similarity to luminal breast cancer relative to basal breast cancer. A useful next step would be to validate the drug activitome signatures in other organs either with historical data or clinical trials. Fisher's correlation distribution background model implies that a correlation of greater than 0.01 are significant at the p-value < 0.05 level.

Gray BRCA Drug Signature	Clinical Phenotype or mutation signature	R	Uncorrected p-value	Bonferroni p-value
PD 98059	Number of first degree relatives with cancer diagnosis (implies genetic disposition)	0.34	5.1E-50	5.6E-46
AG1024	Rectum	0.33	1.6E-45	1.7E-41
AG1024	Histological type=Rectal Adenocarcinoma	0.31	7.3E-41	7.9E-37
VORINOSTAT	Rectum	0.29	1.2E-34	1.3E-30
SIGMA AKT1 2 INHIBITOR	Splenic Flexure	0.29	1.2E-34	1.3E-30
TCS 2312				
DIHYDROCHLORIDE	Rectosigmoid Junction	0.29	1.2E-34	1.3E-30
VINORELBINE	Rectosigmoid Junction	0.28	3.5E-33	3.8E-29
VORINOSTAT	Histological type=Rectal Adenocarcinoma	0.27	1.5E-31	1.6E-27
AG1024	Loss expression of mismatch repair proteins by ihc=YES	0.27	5.7E-30	6.2E-26
IXABEPILONE	Manacles=MSI CIMP	0.27	5.8E-30	6.3E-26
PURVALANOL A	Number of first degree relatives with cancer diagnosis	0.26	1.7E-29	1.9E-25
VORINOSTAT	Loss expression of mismatch repair proteins by ihc=YES	0.26	2.9E-29	3.2E-25
AG1024	fga 1	0.26	5.4E-29	5.9E-25
FASCAPLYSIN	Splenic Flexure	0.25	2.4E-26	2.6E-22
AG1024	sidedness=left	0.25	3.3E-26	3.6E-22
IXABEPILONE	Rectosigmoid Junction	0.24	1.5E-25	1.6E-21
VORINOSTAT	mrnaclust=Invasive	0.24	1.9E-25	2.1E-21
TCS 2312	histological type=Rectal Mucinous			
DIHYDROCHLORIDE	Adenocarcinoma	0.24	1.1E-24	1.2E-20
FASCAPLYSIN	number of first degree relatives with cancer diagnosis	0.24	2.7E-24	2.9E-20
L 779450	number of first degree relatives with cancer diagnosis	0.24	4.7E-24	5.1E-20
IXABEPILONE	FAM83C	0.24	4.8E-24	5.2E-20
GSK461364A	DMD	0.24	5.0E-24	5.4E-20
NU6102	KIAA1804	0.23	3.2E-22	3.5E-18
ICRF 193	FAM47C	0.23	3.4E-22	3.7E-18
GSK461364A	CPXM2	0.23	4.3E-22	4.7E-18
VORINOSTAT	fga 1	0.22	9.7E-22	1.1E-17
IXABEPILONE	Transverse Colon	0.22	1.1E-21	1.2E-17
VX 680	Hepatic Flexure	0.22	2.2E-21	2.4E-17
IXABEPILONE	msi nch=MSI-H	0.22	2.2E-21	2.4E-17
VORINOSTAT	number of first degree relatives with cancer diagnosis	0.22	2.6E-21	2.8E-17
SIGMA AKT1 2 INHIBITOR	number of first degree relatives with cancer diagnosis	0.22	2.9E-21	3.2E-17
IXABEPILONE	msi mda=MSI-H	0.22	4.1E-21	4.5E-17
GSK1838705A	Splenic Flexure	0.22	4.2E-21	4.6E-17
RAPAMYCIN	mrnaclust=Invasive	0.22	1.2E-20	1.3E-16

Table 2 Gray cell line drug activitome signature compared to TCGA colorectal cancer mutation and clinical phenotype activitome signature. All entries are significant.

3.7 TCGA Kidney and Renal Cancers

Kidney and renal cancers (KIRC) signatures have unusually good separation by tumor grade (see Figure 19). The associations to tumor grade suggest the possibility that mutations in KIRC follow an accumulative progressive evolution, but not necessarily one implied by the figure. Note that lymph node involvement event signatures seem to have a distinct pattern apart from grade.

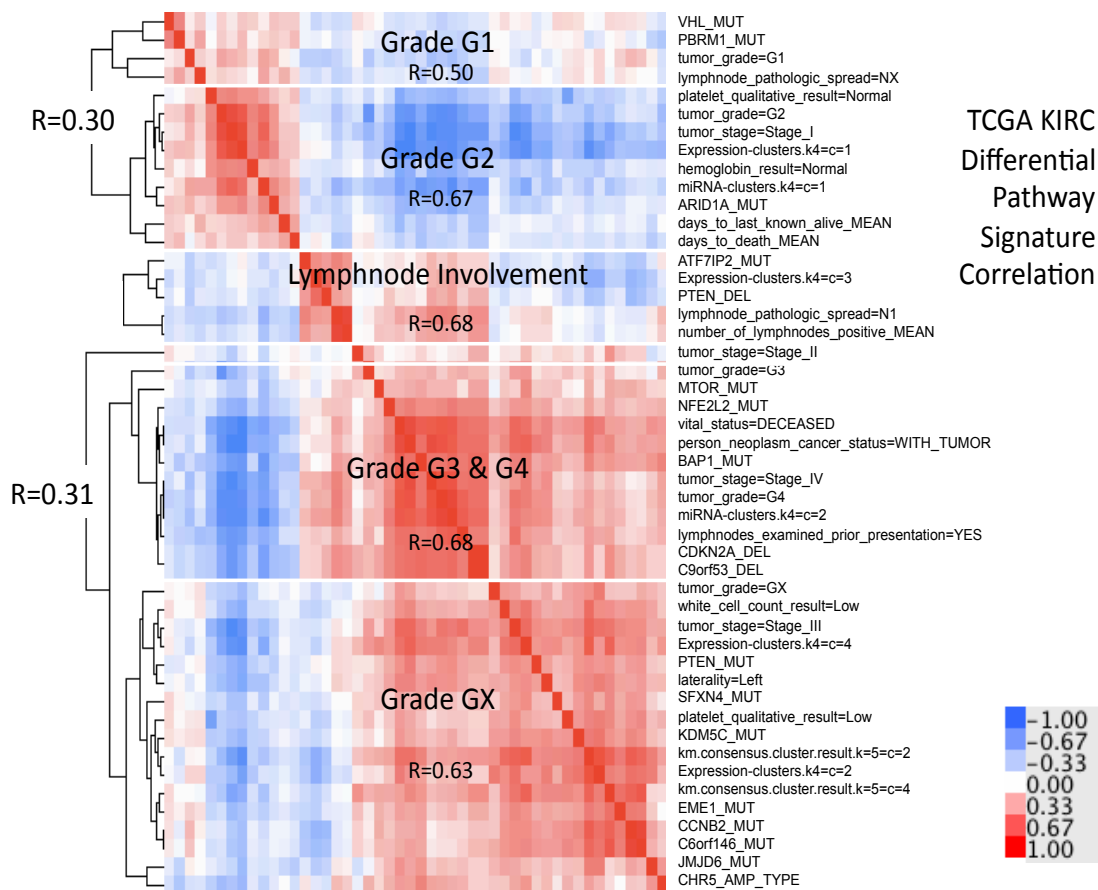


Figure 19 TCGA Kidney and Renal Cancer associates by grade. Horizontal access labels are the same as vertical labels. Mutations are annotated _MUT and deletions are annotated _DEL.

I also observe that CDKN2A mutation and the C9orf53 mutation (which is the promoter region of CDKN2A) are not in grade G1, but they are significantly correlated in grade G2, G3, G4, and GX. Stage 1 kidney cancer associates with just two of the mutations, VHL and PBRM1, both well known in the COSMIC database. The most significant activitome signatures that correlate with kidney and renal cancer mutation correlations are shown in Table 3.

Cell Line Drug Signature	TCGA Mutation	Pearson's Correlation	p-value	Bonferroni Corrected p-value
ICRF_193	RUNX1	0.22	9.72E-23	6.80E-19
GSK461364A	DNAH17	0.18	9.92E-16	6.94E-12
L_779450	ENSG00000245549	0.13	9.88E-09	6.91E-05
5_FU	MLL3	0.12	9.79E-08	6.85E-04
AG1478	LOC442421	0.10	9.70E-06	6.78E-02
5_FDUR	AHNAK2	0.10	9.92E-06	6.94E-02

Table 3 Top TCGA KIRC mutational signatures correlated to Gray lab breast cancer drug sensitivity signatures.

Cancer of the kidney and renal system are particularly morbid and patients have few options. Unlike colorectal or breast cancer, the activitome drug signatures compiled reveals only six correlations that are significant. Most of these are experimental, including Sigma's apoptosis inducer ICRF-193 drug, GSK's PLK inhibitor GSK461364A drug, Tocris Bioscience RAF inhibitor L779450, and the EGFR targeting AG1478. The only approved drugs that are significant in this survey are the classic 5-FU and 5-FDUR (the 5-FU deoxyuridine variant) drugs. The Sigma and the GSK drug are currently undergoing pre-clinical tests.

3.8 TCGA Lung Adenocarcinoma

Lung adenocarcinoma (LUAD) is often diagnosed late when it is an unresectable disease that cannot be cured with current treatment. It is largely driven by driver oncogenes that are the targets for the latest generation of therapeutic compounds. The TCGA data for LUAD was derived from RNA SEQ of a cohort of 230 patients in three previously well-known subtypes: squamoid, magnoid and bronchoid. The DIPSC analysis associates the BRAF and ERBB4 with the Squamoid subtype. Simple clustering associates magnoid subtype with mutations in KRAS and STK11 and PIK3CA lesions. However, DIPSC analysis makes a distinction that Non-smoker LUAD is similar to bronchoid subtype and is associated with mutations in MET and EGFR. The DIPSC analysis of the TCGA Lung Adenocarcinoma dataset is summarized in Figure 20 that illustrates the associations between mutations and the various subtypes.

I observe from the DIPSC analysis that squamoid associates with BRAF mutations in current smokers. Bronchoid associates with both non-smokers and long time smokers, and especially with the ERBB2 gene mutations. ERBB2 mutations occur with low frequency in cancer. This suggests that long time smokers might have survived due to some protective effect that eventually fails and allows the cells that harbor the ERBB2 mutation to proliferate.

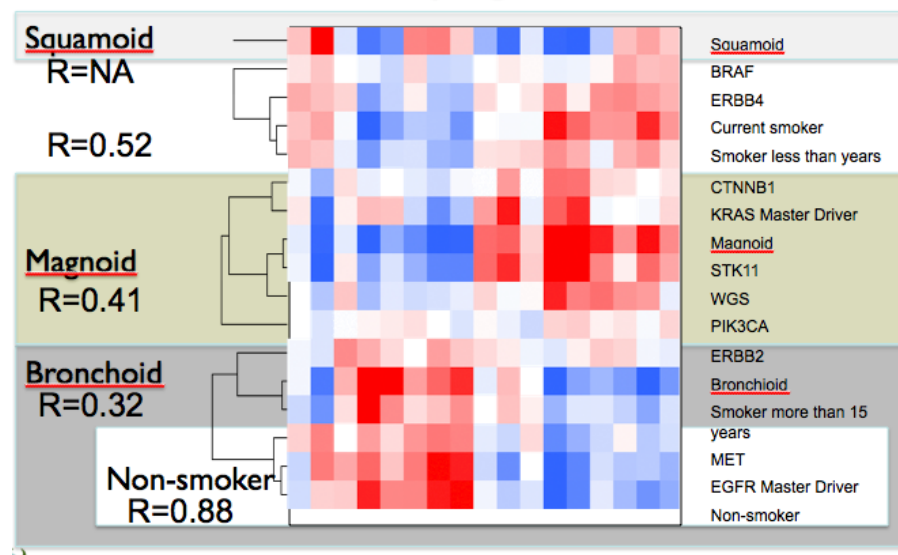


Figure 20 Lung adenocarcinoma (LUAD) follows three basic subtypes: squamoid, magnoid and bronchoid.

Once alerted to these distinctions, it is useful to use differential gene analysis from OCCAM's SAM output to focus on the specific genes that contribute to the signal. These differential values can be projected and visualized directly onto a graphical network using the Cytoscape tool as was done by my colleague Sam Ng using PathMark analysis [Ng 2012]. Figure 21 shows P53 upregulated in bronchioid but downregulated in magnoid and squamoid subtypes. Whereas FoxM1 is the

reverse, FoxM1 is downregulated in bronchioid but upregulated in magnoid and squamoid subtypes.

Static heatmaps convey only part of the story. Distinctions of the subtypes such as an association of magnoid and immunosuppression can be observed on the dynamic heatmap captured in Figure 22, which is available at the following URL:

<http://soe.ucsc.edu/~ted/TCGA/LUAD/>

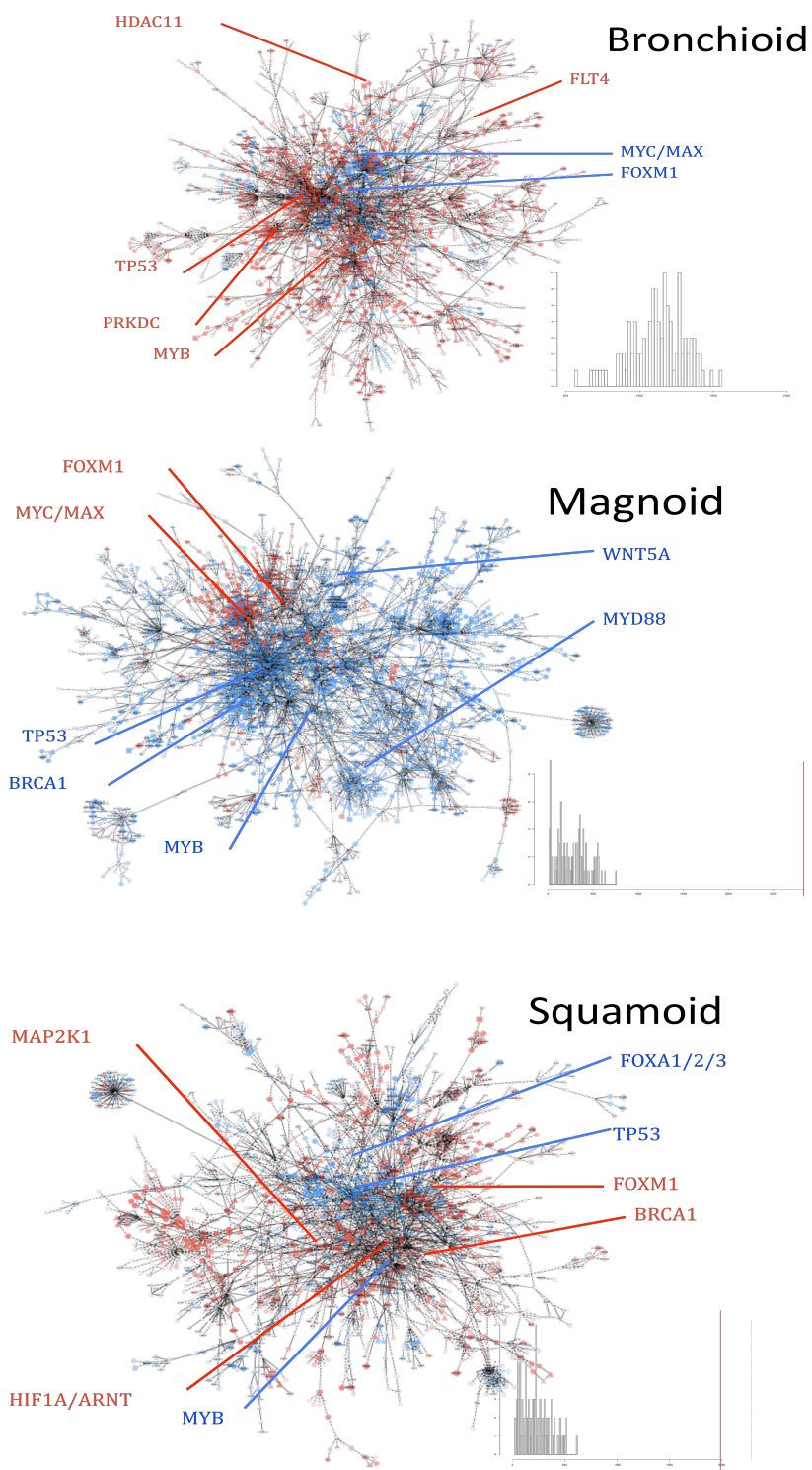


Figure 21 PathMark diagrams on TCGA Lung identify differential pathways between subtypes [Ng 2012].

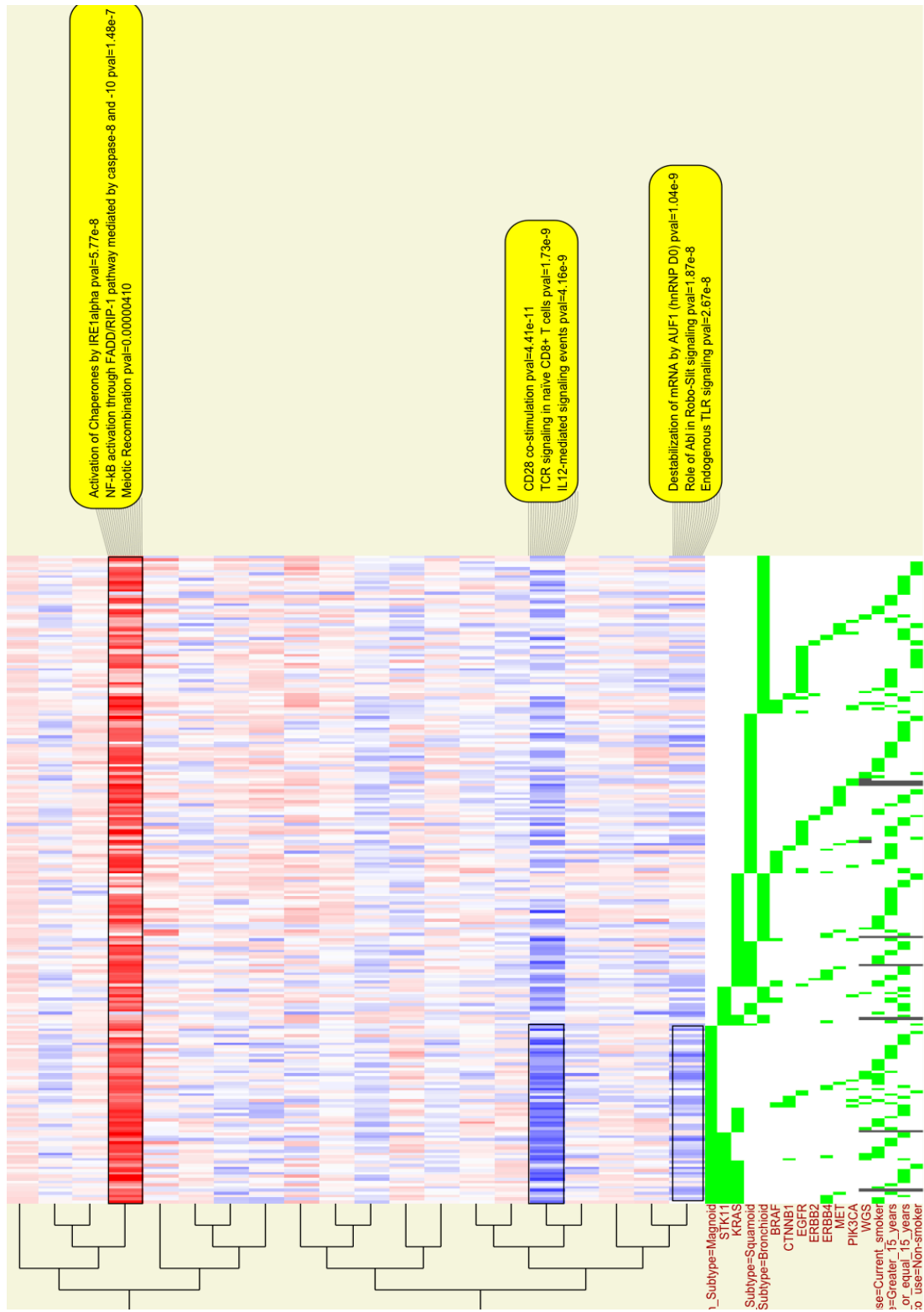


Figure 22 Dynamic Heatmap of LUAD clusters with phenotypes annotated with significant pathways.

3.9 TCGA Uterine and Endometrial Cancer

Similar to colorectal cancer, uterine and endometrial cancer (UCEC) has subtypes defined by hypermutation and micro-satellite instable (MSI) and microsatellite stable (MSS) phenotypes. As well, UCEC has pathologist determined subtypes called *endometrial subtype* and *serous subtype*. High microsatellite instable (MSI-H) phenotype is associated with the hyper-mutator phenotype. Low microsatellite instability is associated with mixed histology and microsatellite stable phenotype is correlated with serous-like correlations. Low mutation and serous are radically different in their pattern and MSS may in fact be a different disease entirely given the opposite polarity of the correlation visible in Figure 23 of serous and mixed histology vs. all of the other groups.

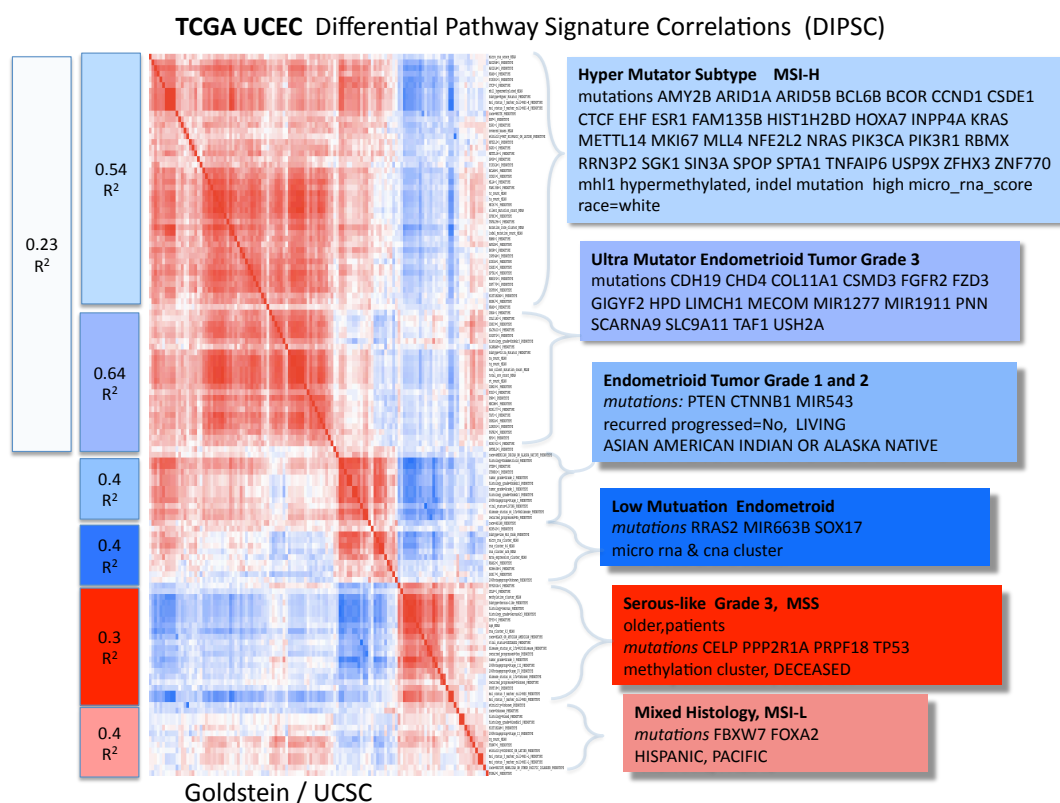


Figure 23 DIPSC heatmap of uterine and endometrial cancer reveal clustering by mutation to subtype and grade.

There are many significant correlations between UCEC and Gray cell line drug activitome signatures implying that there is a strong relationship between breast cancer and uterine cancer as shown in Table 4.

Cell Line Drug Signature	TCGA UCEC Event	Pearson's Correlation	p-value	Bonferroni Corrected p-value
TRICIRIBINE	FOXA2	0.57	7.8E-30	6.4E-26
TRICIRIBINE	BCL6B	0.57	1.6E-29	1.3E-25
FASCAPLYSIN	BCL6B	0.56	2.8E-29	2.3E-25
FASCAPLYSIN	SMTNL2	0.56	3.4E-29	2.8E-25
VORINOSTAT	SMTNL2	0.56	6.5E-29	5.4E-25
AG1024	SMTNL2	0.55	4.3E-27	3.5E-23
TRICIRIBINE	SMTNL2	0.52	3.4E-24	2.8E-20
FASCAPLYSIN	FOXA2	0.51	7.0E-24	5.7E-20
GEFITINIB	BCL6B	0.51	1.0E-23	8.2E-20
SIGMA AKT1 2 INHIBITOR	SMTNL2	0.50	6.3E-22	5.2E-18

GEFITINIB	FOXA2	0.48	8.8E-21	7.2E-17
SIGMA AKT1 2				
INHIBITOR	cna cluster k4	0.47	4.9E-20	4.0E-16
SORAFENIB	MLL4	0.47	8.2E-20	6.8E-16
SIGMA AKT1 2				
INHIBITOR	MLL4	0.47	2.9E-19	2.4E-15
VORINOSTAT	CTCF	0.46	4.4E-19	3.7E-15
RAPAMYCIN	MLL4	0.46	1.7E-18	1.4E-14
VORINOSTAT	BCL6B	0.46	1.9E-18	1.5E-14
SIGMA AKT1 2				
INHIBITOR	BCL6B	0.46	1.9E-18	1.6E-14
GEFITINIB	MLL4	0.45	2.6E-18	2.2E-14
5 FDUR	FZD3	0.45	4.0E-18	3.3E-14

Table 4 Top correlations between Gray cell drug and subtype signatures and TCGA Uterine cancer and endometrial cancer signatures.

3.10 The TP53 PANCAN Activitome Correlation with Drug Resistance

It is clear that many similar phenotypes are caused by numerous mutations occur across cancers. The TCGA pancancer (PANCAN) project has recently begun to reanalyze the data from multiple other cancer groups together. As of this writing, the *pan-cancer* (the commonalities of cancer considered across all tissues of origin, abbreviated PANCAN) mutations are still in flux, however, I decided to do a quick analysis that compared just the TP53 mutation across all of the cancers individually, and in PANCAN collectively.

The following Figure 24 is a correlation of the Gray cell line drug and subtype activitome signatures correlated against the TP53 activitome signatures derived from the TCGA data sets from colorectal cancer, bladder cancer, kidney / renal cell cancer, breast cancer, glioblastoma brain cancer, lung adenocarcinoma cancer, uterine endometrial cancer and pan-cancer. There is tremendous correlation among the different cancers.

As can be observed in the top panel of the heatmap below, many DNA damaging agents are correlated, which is a positive control given that replication is

the key hallmark of cancer, as well as TP53's known role in regulating the DNA damage pathway. However, colorectal is a sharp outlier that anticorrelates with the other cancers. The TP53 vs. drug response pattern is completely different for colorectal cancer. This is possibly due to coincident APC mutations which alter TP53. It is too early in this investigation to tell what this means, but it raises some interesting questions for follow up as the TCGA pancancer data matures.

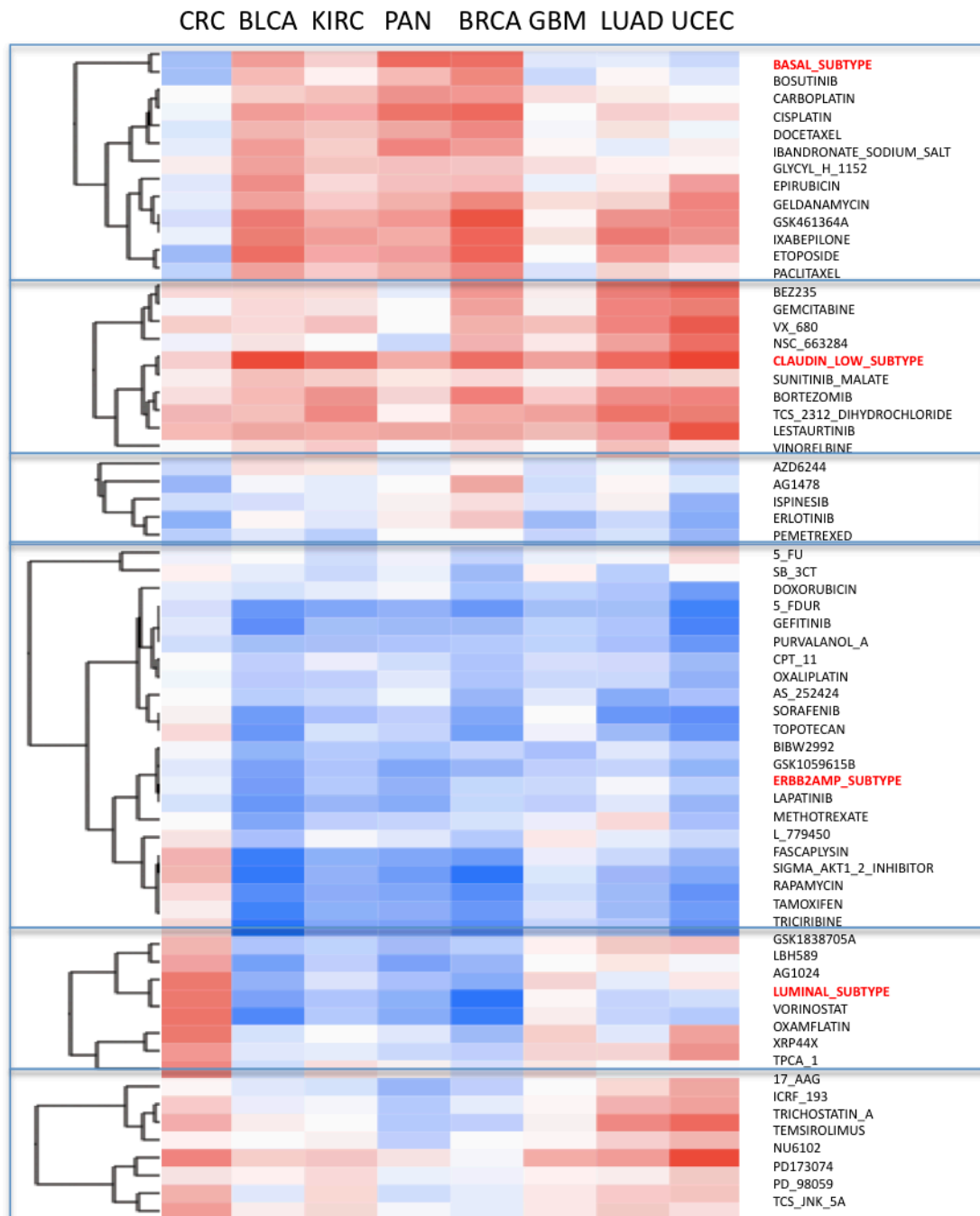


Figure 24 Gray cell line drug sensitivity correlated against TP53 activitome: colorectal cancer (CRC), bladder cancer (BLCA), kidney renal cell cancer (KIRC), breast cancer (BRCA), glioblastoma brain cancer (GBM), lung adenocarcinoma cancer (LUAD), uterine cancer/endometrial cancer (UCEC) and “pan” cancer (PAN). Red indicates correlated, blue indicates anti-correlated.

3.11 Discussion

It is still early in The Cancer Genome Atlas (TCGA) data gathering process. The overall pattern of the activitome is beginning to emerge. The cancer data currently shows that some event signatures such as TP53 and FOXM1 will be shared among many, but not all cancers. I tried to see if there exist correlations between uterine and colorectal cancers using the MSI signature and hypermutation signature. Thus far, the signatures show no correlation. Further data may be necessary. However I suspect that the signal is dominated by tissue of origin. Drug activitome signatures seem to have very good cross tissue of origin correlation. But it is still early to say that there are luminal-like cancers besides breast cancer. I have shown that OCCAM and DIPSC are methods that provide a way to survey the broad activitome signature landscape. As new tissue deconfounding principles are developed, OCCAM and DIPSC will incorporate these methods, which will assist in boosting the activitome signal from cancer's noise. To gain the confidence of the medical community, OCCAM and DIPSC signatures must be validated using classical medical processes as was done in the Washington University study, which had separate discovery and validation cohorts.

Thus far, I have only used very well known, well-proven, linear models of correlation to compare signatures. Many more sophisticated (though less proven) machine-learning algorithms exist such as *support vector machines* method and *random forest* method that may be able to tease greater signal out of the same data. The DIPSC sub-sampling technique would be useful as well to deconfound the

sample overlap problem. These techniques may require novel mathematical methods and further analysis to combine results. For my purposes, because correlation is a linear metric, taking the mean of all values successfully combines subsample correlations.

In this section, I have shown the beginnings of the creation of the activitome. Activitome signatures have been useful to identify drug targets and to collect similar mutations into clusters that provide insight into subtype, tumor stage and grade. These insights are the useful knowledge encoded in the activitome that can be clinically used to improve patient outcomes. Most significantly, drug activitome signatures derived from cell line experiments show how drugs in one cancer may be repurposed for treating another cancer. This may result in better survival for critically ill patients. Of course, since cell lines and animal models have artifacts distinct from patients, superior data and knowledge would be better derived from outcome information from actual patients, perhaps by using MedBook—the subject of the next section.

4 Future: MedBook Knowledge & Decision Support System

High-throughput genomic sequencing costs are rapidly decreasing and producing volumes of big data. As Eric Green, director of the National Human Genome Research Institute has said: *The major bottleneck in genome sequencing is no longer data generation—the computational challenges around data analysis, display and integration are now rate limiting* [Green 2011]. Medical bioinformatics must turn this genomic big data into knowledge and rapidly integrate the clinically relevant knowledge into clinical practice so that patient outcomes are improved. The challenge then is how to use big data and its required analysis tools both for long-term research and immediate patient treatment.

To meet this challenge I have developed **MedBook**, an open-source federated knowledge web system designed to extract actionable knowledge from genomic data. The tools described in my thesis have been built to run in MedBook, a project that is a component of The Stand Up 2 Cancer *Prostate Cancer West Coast Dream Team* (WCDT), a team of prostate cancer oncologists and researchers from six universities along the west coast of North America. This section presents both the MedBook vision and the current state of the work.

MedBook’s goal is to connect the cancer research world to new knowledge learned from treating every patient, and to feed that knowledge back into the cancer treatment *knowledge base* in a timely manner. Without MedBook, or a system like it, cancer care will remain balkanized with knowledge isolated in individual cancer care centers only very slowly disseminating to the wider world of medicine.

MedBook is modeled on principles of social networking and built on top of the Galaxy workflow system [Goecks 2010] that already supports hundreds of genomic tools. Galaxy already provides web-based access to data analysis tools for biologists and bioinformaticians. MedBook adds several important new ideas:

- Biomedical Evidence Graph (BMEG)
- An end user apps platform
- A rich data publishing model that combines discussion with auditable re-computable datasets
- Social networking and collaboration

While there are numerous individual online communities for patients, doctors and researchers, none bring all these constituencies together in one place. Patients may fail to enroll in randomized Phase III clinical trials because they and their doctors do not know such a trial exists. As well, in the face of no information, doctors sometimes randomly select a therapy; at a recent think tank of the leading prostate cancer oncologists, doctors described using a coin flip when trying to decide which drug to use. The default is to choose a drug based on minimizing side effects rather than on how well the agent fits the tumor's vulnerabilities. **This shows that there is an urgent need to apply machine-learning techniques to selecting therapies for patient care.**

The advantage of an intentional online community such as MedBook is that the patient is uniting with doctors and researchers in common cause to create a complete bench-to-bedside integration of research with therapy. The SU2C West Coast Cancer Dream Team is an intentional rapid learning cancer community (see Figure 25) that will allow scientists, clinicians, and eventually patients to collaborate

through social media to deliver individualized treatments and improve outcomes. In the future, MedBook apps will allow doctors and patients to:

- Select optimal therapy based on genomic and electronic health record data
- Match patients to clinical trials
- Capture and transform the data into new knowledge
- Share information between clinic and research

MedBook's goal is to treat each patient in accord with the best available knowledge and to continually update that knowledge to benefit future patients, in other words, *computer-assisted evidence-based medicine*. Achieving this goal requires collecting and analyzing the molecular and clinical data from every cancer patient and patient outcomes, and bringing the therapeutic team's collective knowledge to bear on each treatment decision.

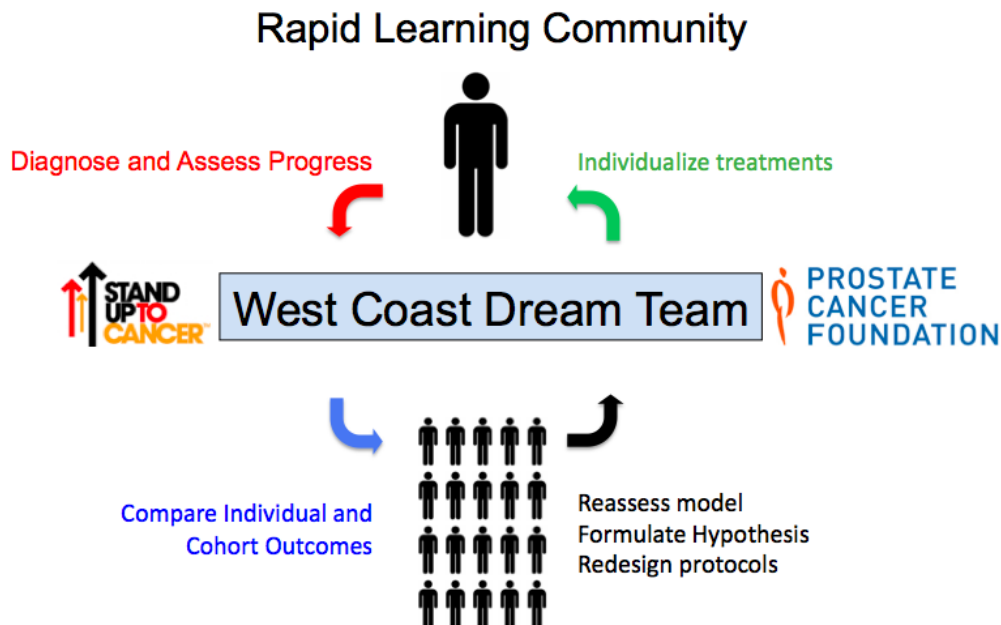


Figure 25 Prostate Cancer West Coast Dream Team Rapid Learning Community

Modern molecular biology supports the hypothesis that cancer is actually hundreds or thousands of rare diseases, and every patient's tumor is, to some extent, unique. In some idealistic vision, every individual mutation would be identified, catalogued, and a risk assessment performed singly and in combination with all other mutations. This is computationally unbounded and therefore unrealistic. Therapies must be developed and verified for some group of patients with like mutations. Instead of considering individual point mutations or gene mutations, MedBook will assist by aggregating similar dysfunctional signatures to increase statistical power. Assuming that the data is normally distributed, to achieve a p-value < 0.05 requires a cohort of $n=30$ patients with a similar activitome signature. A cohort this size will be necessary to have the statistical power necessary for verification. A much larger n is necessary to develop a new therapy. It costs over a billion dollars to develop a new drug. Therefore, even with the exorbitant prices of drugs at \$100,000 for a course of treatment, a cohort of 10,000 patients is required to justify developing a new drug. Very few cancers have this many patients. This is another reason why repurposing existing drugs into new cocktails is attractive.

New technologies in DNA sequencing now make it possible to decode every sample's genome, which may provide invaluable clues about cancer onset, progression, and treatment. Subsequent to the sequencing of the first human genome, the Cancer Genome Hub (CGHub) at UCSC became the nation's archive for all TCGA genome sequence information. UCSC leads the way in cataloging and

securely storing all personal genomic data produced by other large-scale public cancer projects.

Despite significant scientific progress, the clinical translation of scientific knowledge is still frustratingly slow. To accelerate the translation process from knowledge gained to clinical practice, the cancer research and care communities need a sample and patient-centric strategy that focuses research directly on helping individuals obtain better outcomes. The process requires collaborative tools that facilitate data sharing and data interpretation into knowledge within and across the scientific teams, as well as with clinicians and patients. The ultimate objective of a rapid learning community is to share knowledge and best practices so what is learned from one patient is disseminated in time to help the next.

4.1 The Computational Opportunity

High-throughput biology produces high volume, so called *big data*, which requires high-throughput computing for both research and clinical purposes. Cancer medicine will use *exascale* (10^{18} floating point operations) computing for shotgun assembly, pathway simulation (such as PARADIGM), and data mining techniques such as OCCAM, PATHMARK, DIPSC, HOTNET, and other programs (see Figure 33). All of these techniques require the full computation be done on the true data, as well random data is generated on null models to determine p-values and confidence intervals associated with every measure.

Exascale computation may use 1000 to 100,000 or more 3ghz CPUs to process data for hours evaluating cohorts or selecting the right drug for a patient.

Until recently, this amount of computing was only available for atomic bomb simulation or code breaking. Fortunately, the commercial purposes of indexing the world-wide-web or performing data mining for retail shopping has made exascale computing commercially available. Exascale computing is (or will soon be) available from cloud vendors including Amazon, Google, Microsoft, and others, who have come along to provide this level of computing as a service. Semiconductor and computer manufactures including Intel, Nvidia, AMD, Dell, and Oracle/Sun are preparing to sell these systems to medical centers, which they believe will want to operate exascale-computing infrastructures.

4.2 The MedBook Vision

The combination of the demand for better medical treatments and the opportunity provided by high-throughput genomic assays and exascale computing technologies support creating MedBook, a platform for creating and using the cancer knowledge base. The problem begins with the fact that most knowledge of cancer (as is most scientific knowledge) is represented as PDF documents. The corpus of documents and data is vast and needs organization. Even worse, the connections between documents and data are sparse and often lost. The production process from raw data to scientific findings may be described in publications, but it is usually insufficient to actually re-perform the analysis. At best, publications only allow for a modest amount of findings and data to be published.

The world-wide-web begins to help in solving this problem, since it is now possible to index and access vast quantities of data. However, even most of TCGA's

cancer data on the web is not directly processable by further bioinformatic analysis. What is needed is a mechanically replayable automated workflow that transforms the raw data from observations and machine assays into interpretable figures; while maintaining patient privacy and auditability. Systems such as Sage Bionetwork's Synapse [Friend 2013] and Penn State University's Galaxy system [Goecks 2010] have been developed to address this part of the problem.

What is still needed are links between the machine playable description of the process, along with the textual interpretation that unites raw (and transformed) data, the data analysis process, and the science into one unified *Biomedical Evidence Graph* (BMEG). This proposed world-wide-web-based graph will unite the HTML pages, URLs, relational and post-relational databases, application programming interfaces (APIs), cross language portable (JSON) objects, search engines, and the Synapse and Galaxy workflow systems into a set of data integrated standards.

4.2.1 The Need for Data Integration

An example of the need for the Biomedical Evidence Graph involves imatinib (Gleevec), originally developed in 1998 by Novartis for Chronic Myelogenous Leukemia (CML). Imatinib was among the first of a new generation of targeted therapies developed to block a specific enzyme—in this case, BCR-ABL—in a cancer-signaling pathway. Shortly after its development, it was experimentally observed that Gleevec also blocked a structurally similar enzyme called c-Kit, which was a primary cause of gastro-intestinal stromal tumors, a rare fatal form of sarcoma.

Imatinib was subsequently tested on many forms of cancer, including a 2003 trial for melanoma that appeared to have failed [Kim 2008].

Unfortunately, five years passed before the world learned from this trial that careful analyses showed it wasn't a failure but it only worked for a subgroup of patients who had tumors on interior mucosal surfaces. Their tumors all had a mutated c-Kit gene. Immediately upon publication, in 2008, results of a definitive trial testing Imatinib on melanoma patients whose tumors had been pre-tested for c-Kit mutations demonstrated a major response. Normally, a randomized trial would have needed perhaps a million patients to demonstrate statistical significance for a drug that affected at most 3 percent of melanoma patients. However, because of the knowledge of the c-Kit gene drug sensitivity from the previous trial, doctors were able to immediately apply these findings to real patients. Finding effective therapies like this would be significantly accelerated if there were a web-based cancer knowledge cloud based on the BMEG that was able to display this type of knowledge as soon as it was entered.

Web-based social media networks, such as Facebook.com, Google+, and cloud-computing platforms such as Salesforce.com, facilitate rapid collaboration across geographically separated people and information. However, the existing commercial web platforms such as Salesforce.com and Google.com do not solve the unique problems that arise in translational medical research, which include the integration of clinical and genomic data, and the need for medical information privacy. Moreover, software applications used in cancer research, such as the Cancer

Genome Browser, do not address the needs of clinicians and patients, who need intuitive tools that can help them make scientifically, informed treatment decisions. MedBook will use social media to link the data and activity of community oncologists, research oncologists, patients, and cancer researchers in a user friendly, secure, scientifically rigorous framework.

4.2.2 Biomedical Evidence Graph (BMEG)

The massive scale of FaceBook’s social graph is inspiring. It represents the connections between hundreds of millions of people and things and events that they care about, yet is accessible to most people using a simple web-based interface. FaceBook users and applications seamlessly navigate this graph to discover common threads of interest and new relationships between people. Fundamental phenomena such as the “six degrees of separation” can be directly seen on a global scale in the FaceBook social graph [Ugander 2011].

MedBook will create a similar structure to the social graph that knits together the informal anecdotal and formal scientific data called the Biomedical Evidence Graph (BMEG). The BMEG is a large persistent representation of the relationships between the scientific evidence from tools such as hierarchical clustering, biclustering, OCCAM, DIPSC, and machine learning classifiers.

Nodes in the BMEG will represent the original scientific assays and inferred findings. MedBook will initially draw data from the The Cancer Genome Atlas, Stand up 2 Cancer and other collaborations. One distinction from the FaceBook graph is that MedBook may use a Bayesian framework, such as *HyperPrior* [Tian 2009], that maintains a prior evidence probabilistic score on arcs in the graph. The BMEG is a hypergraph (graph of graphs) whose constituent graphs represent:

- Patients and their treatments
- Samples and their high-throughput genome sequencing and expression assays
- Inter and Intracellular Biology Pathways
- Experimental methods and parameters
- Clinical pathological state
- Human and machine generated hypotheses
- Logical arguments, corrections and other annotations
- Molecular Disease Models (MDMs) that define disease subtype ontologies
- Classifiers of patient sub-type based on best of breed machine learning
- Meta signatures that recognize common patterns across disease types

The benefit of the BMEG is that it will enable researchers to query across studies and conduct meta-analyses across cancer subtype and cohorts at any time and on the fly. Queries such as, “what mutations drive drug resistance” or “does surgery in combination with chemotherapy have a better long-term survival than either alone,” will be able to be asked and answered.

Just as FaceBook consists of many different extensible feeds and document types, the BMEG is composed of a variety of databases and documents linked and tagged together:

- Molecular Disease Models database
- Therapeutic plans and Outcomes database
- Activitome Signatures database
- Individual and Group Patient Profile database

The following sections elaborate each component.

4.2.3 Molecular Disease Models (MDM) Database

The Molecular Disease Models (MDM) database represents the current state of knowledge about cancer and its subtypes, their associated molecular diagnostics, targeted drug therapies and drug resistance. The MDM aggregates the knowledge found in databases such as NCBI, CGHub, and COSMIC into a browsable interface with a queryable programming interface. The MDM connects information about pathways, targets, tests, drugs, and trials previously scattered across hundreds of publications and databases, as well as the expertise of experienced oncologists. The MDM is an experimental work in progress, but a solid foundation for rapid learning. As oncologists use the MDM to treat patients, the knowledge about subtypes and their associated therapies will be continually refined based on how individual patients respond. Subtypes will be split as appropriate to correspond to treatment responders and nonresponders, and new subtypes added to accommodate previously unseen tumor types.

4.2.4 Therapeutic plans and Outcomes Database

The most important use of MedBook will be to answer the question: “Which treatment is best for a given patient?” This is a critical question and no existing framework currently supports the aggregation of the volumes of data to create a traceable and logically coherent answer to this question. The library of therapeutic plans will represent both the on-label and the off-label plans. The Outcomes Database, accessed through the Outcomes App, will track the actual treatments given and the resulting patient outcomes. The insurance companies, government payers, and pharmaceutical companies will especially like to have access to this data. The Outcomes Database will be created by data mining electronic health record systems, and with the Outcomes Reporting App. This is a key component for improving quality and future patient outcomes. In the SU2C Prostate Cancer project, we are integrating with UCSF’s OnCore patient clinical trial system to harvest data and track the intermediate outcomes of biomarkers such as Gleason scores used in prostate cancer.

One use of MedBook is to assist with clinical trials and especially adaptive clinical trial knowledge-management. Rolling adaptive clinical trials informed by high-throughput sequencing will allow drugs to “graduate” from the trial into the clinic and new agents to enter the trial without the time consuming process of ending one trial and starting the next. The I-SPY trial series [Barker 2009] is an example of this new kind of trial that is not yet currently funded for sequencing. Although this trial could be run without sequencing, the amount of valuable knowledge about

changes in molecular signatures triggered by the therapy that could be learned with sequencing in this setting is irreplaceable and UCSC is working to raise funds for sequencing and analysis.

4.2.5 Activitome Signatures Database

As described in the previous sections, the activitome signature database collects signatures, which are mathematical descriptions of the differential diagnostic effects of a mutation, and other molecular event or clinical phenomenon. This database is a collection of biomarkers that connect the patient and group profile databases to identify molecular targets and prognostic factors used to predict appropriate therapies. In the SU2C Prostate cancer project, I am working on TMPRSS2-ERG gene fusion and Gleason score signatures. These are being evolved from data from previous studies and recent morbidly ill patients enrolled in the SU2C Prostate Cancer project.

4.2.6 Individual Profiles

Every patient, even anonymized patients, will have a patient profile that aggregates their information. There will be multiple specialized views of a patient. The treating physician team will see a view of the profile designed to guide the optimal treatment for a patient. The patient will have their own specialized view that presents the data in a way appropriate for them. Health care provider and research institutions can adapt MedBook for local standards of care.

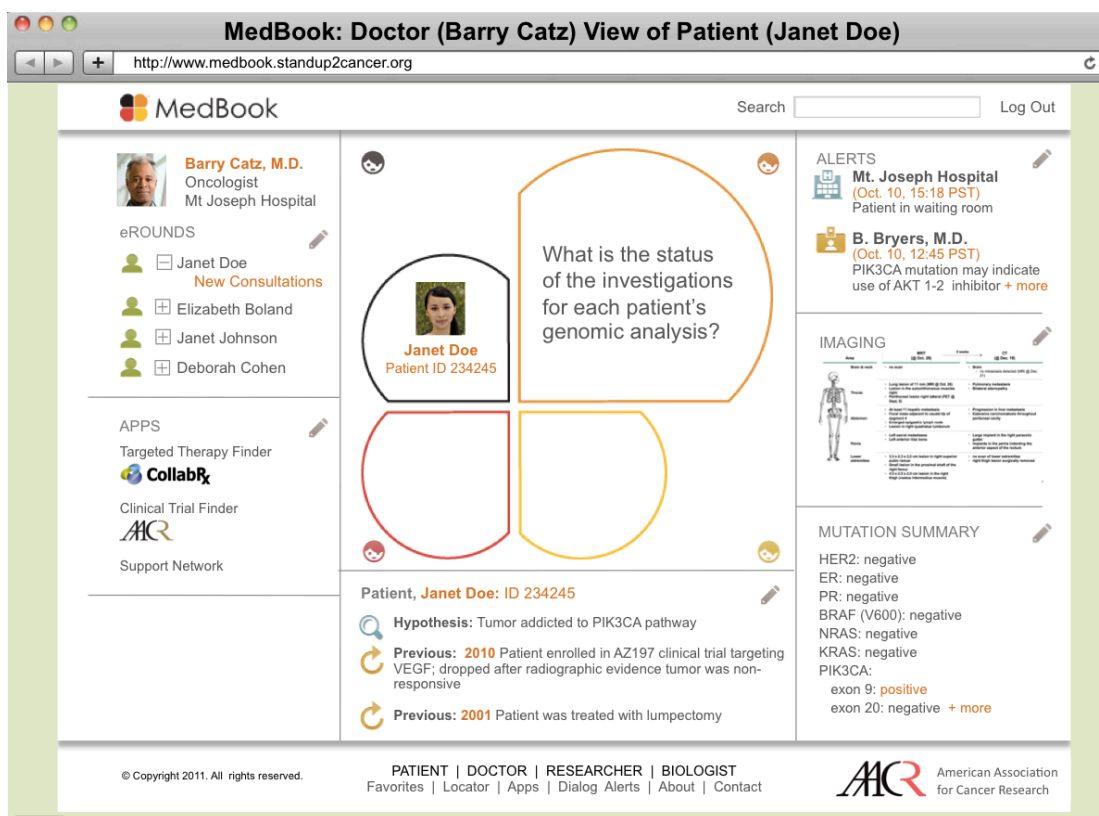


Figure 26 Physician Oncologist view of MedBook. The focus is on the patient care with scientific data views available, but not foremost in view.

The individual patient profile is also anonymized and exportable to other clinicians and researchers for consultation purposes and research benefit (See Figure 28) for data mining, using statistics and machine learning of aggregate data. The per-patient page is already used in the SU2C Prostate Cancer project, where we have captured images of patient samples. These images are useful for discussion between researchers to identify the ideal tissue to be assayed.

4.2.7 There's an App for That

There is great need for many systems to mine and use the data. MedBook is an open extensible social medium and a computational web platform for rapid learning

that will have numerous *apps*. Apps, short for applications, are a useful unit of user experience for both smartphones and web applications. As well, there is a popular notion that apps are purchased, easily discovered, and widely available. Ultimately, MedBook will be operated as part of a clinical care environment, where diagnostic genomic applications will receive insurance and other payer reimbursement. Currently MedBook apps are user interface front ends to canned workflows.

Apps will run on behalf of users to create profiles for biological samples such as biopsies, blood specimens or cell cultures, as well as to exchange messages, upload medical information and run applications that analyze data collected from these samples. Some apps will extend the base platform and provide services to other apps.

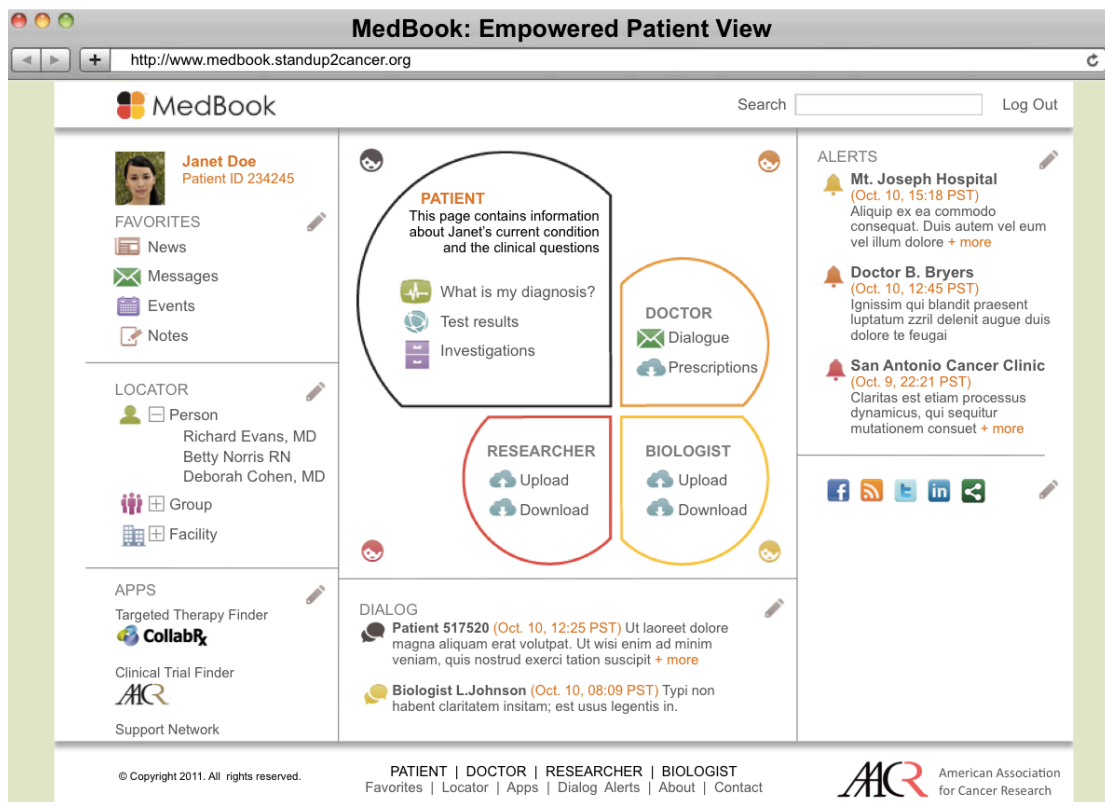


Figure 27 Patient Centered MedBook. Focus is on the patient care and managing the complex events and personal decisions using the best medically recognized data.

All data is anonymized and retains institutional review board controls. Physicians and researchers will share sample data, biopsy information, patient staging plans, high-throughput genomic data, and their personal experiences in a safe structured environment covering research, disease course, and treatment. This aspect of the social media model allows both the capture of unstructured information as well as collaborative discussions, annotation, and editing of structured information.

The MedBook platform applies the simple, but powerful social media concept of *tags* to create groups. MedBook tags are hierarchical ontological keywords that create membership in groups. MedBook will use tags on groups of patient samples with persistent pattern-mining algorithms to find common attributes, such as recurrent mutations or activity signatures of genetic pathways. Algorithms will be run to create predicted groupings of patient samples. For example, cancer subtypes based on gene expression profiles could create possibly new ways of grouping patients. Set-theoretic mathematical operations allow straightforward but sophisticated queries about disease samples. For example, a researcher could pose a question to determine the overall proportion of breast cancer ER+ patients with PIK3CA mutations and see how many of those patients progressed on tamoxifen treatment using set intersections.

Today, it is extremely difficult to pull all the data from various and numerous geographically dispersed studies together into a common format to determine the answer to this question. The idea of grouping patient samples and making these group memberships available as both vetted and predicted attributes for samples allows the results of one analysis application to build off of another. This will occur across all

cancer patient samples wherever they are, regardless of how the data were collected. When all data is in the same format, all researchers, clinicians, and patients can use and benefit from it.

Structured data enables computational analysis. Clinical researchers can upload the data from patient cohorts. Individual patients can share their experiences with other patients and clinical researchers. Patients and their physicians will also have a platform that explains the meaning of their molecular tests. Alternative hypotheses can be presented, discussed and experiments proposed that will support or refute these hypotheses. Clinical researchers will run bioinformatics analyses that summarize the data without revealing individual patient identity or personally identifying DNA mutation information and return their findings to the patient's physician. Physicians and clinical researchers will finally have an electronic forum to discuss the impact of this information on patient health. Because samples are understood more fully by comparison to a statistically significant cohort of phenotypically alike samples, the larger the MedBook database is, the more valuable every item in it becomes.

Although learning from structured data is the goal, learning from unstructured gene expression without knowledge of the important variables is also important. By aggregating data from many sources and applying advanced clustering and other unbiased methods, medicine can fill in missing links in our understanding of the knowledge graph.

4.2.8 Federated System

MedBook is built on a loosely coupled federated database system to operate across multiple data centers with synchronization, sharing, and distribution of applications. For example, there could be MedBook instances at medbook.ucsc.edu (physically located at UC San Diego's super computer center), medbook.ucla.edu, medbook.ucdavis.edu, medbook.ucsf.edu, medbook.ohsu.edu, and medbook.med.ubc.ca. When the data is small (less than terabytes), the data and its provenance can be distributed from one instance to another. When the data is large, or laws such as HIPAA require it, MedBook will send the workflows to the data.

Updates and sharing of data will proceed according to cryptographically controlled rules and will be managed by the Gateway Security Model that I developed, used in both Apple's iPhone and Google's Android phone operating systems [Goldstein 1996].

The *MedBook* platform will be branded and promoted as a public service of many medical centers. It will be run in a certified secure data center but will be accessible on the web using electronic health record security standards. Especially sensitive data will require multi-factor industrial authentication before it can be viewed. As it gains acceptance, MedBook could be networked with medical centers and their electronic health records and provide a platform for patient-centric multi-center collaboration.

Like FaceBook and other social network platforms, the user experience starts with looking at and viewing profiles of anonymized patients and their samples. The

experience then moves on to groups of demographically similar samples, as well as to cohorts of commonly treated patients. Researchers and bioinformaticians will collaborate to create new apps that build on these data and the data of other apps to provide interpretation and actionable information. Many of these apps are currently under development by affiliated healthcare organizations, but lack a sufficiently powerful platform that connects patients, their data, and physicians.

4.2.9 Validation of New Methods

Before research oncologists (not to mention community oncologists) will start using MedBook in the clinic, they will need to see new methods recapitulate known signatures and diagnostic methods. MedBook provides a platform to allow clinicians to compare new methods with existing methods. Only after demonstrating increased power to explain existing datasets with new methods, will the research clinicians redirect resources to validate hypotheses generated by the bioinformatics community. This demonstration will have to consist of improved patient outcomes such as delayed recurrence and improved survival curves.

4.3 Patient, Health Care Provider And Researcher Use Scenarios

To make the description of the system more concrete, here are several example use cases developed for the OHSU Knight Cancer Center. These are drawn from a very large number of possible scenarios. I am designing the system such that the primary mechanism of interaction between users and MedBook is through apps. The example use cases give illustrations of just a handful of

potential apps, which represent only a small portion of the entire spectrum of what the apps should be capable of accomplishing.

4.3.1 Patient and Health Care Provider Use Scenario

J.D., at age 41, detected a lump in her right breast and was diagnosed with node positive, triple negative, invasive ductal carcinoma. She was treated with lumpectomy, radiation, adriamycin, cyclophosphamide, and taxol. A local advocate contacts J.D. about the OHSU Knight Cancer Institute breast cancer initiative and she enrolls, consenting to genomic testing and submitting initial health records.

At age 45, J.D. was diagnosed with a metastatic recurrence to lung and was treated with capecitabine, but her disease progressed. She and her physician turn to explore options beyond the standard of care using the MedBook system.

J.D.'s use case 1: Patient-initiated trial matching query. J.D. wants to find the best clinical trial for her. This decision involves many different factors. Once her data is uploaded, all the information necessary to make this decision is represented in MedBook. MedBook makes the data available to apps, which help her and her physician search for trials for which she is eligible and that are matched to her clinical features, cancer subtype, and specific cancer genome mutations. She indicates her preference also for trial locations that are near her home or family. She discusses the trade-offs with her physician. She also uses MedBook to connect with patients in her similar situation.

J.D.'s tumor showed mutations in *TP53* and *BRCAl* that met inclusion conditions for a particular clinical trial, which fortunately was open at a center near her home.

She signed consent for participation in this clinical trial and began trial therapy. After 8 weeks, there was no evidence of response as shown by radiographic assay; instead scans showed progressive disease. She stopped study interventions so that she could pursue another therapy. A new needle biopsy was obtained from J.D. to track mutational progression.

J.D.'s use case 2: Physician-initiated therapy query. J.D.'s physician uses the completed genomic analysis of J.D.'s tumor already in MedBook to identify other potential interventions as it becomes clear that J.D. was not doing well on the trial.

The Activitome Differential pathway signature app analyzes the genomic and expression data to identify molecular signatures of signaling pathway activity that are found in J.D.'s tumor. Activitome Differential identifies that the Notch pathway is dysregulated. A full analysis of upregulated and downregulated pathways is posted on MedBook, annotating J.D.'s case. In addition, the profile is now included in the cohort of tumors from which Activitome Differential draws its analysis.

A case report app examines the clinical records of patients who also have dysregulation of the Notch pathway. One case is found where the patient responded well to an off-label use of drug B, already approved for lung cancer.

J.D.'s physician receives notifications of this opportunity for therapy beyond the standard of care. J.D.'s physician can now review all of the relevant data with

genomic tumor profiles similar to J.D.'s. The physician wants to consult with peers, so he posts a case report on J.D. asking for assistance in making the decision to try drug B off label.

4.3.2 Clinical Researcher Use Scenario

BCNM Inc. has developed a small molecule inhibitor of *FGFR2*. BCNM wishes to perform a clinical trial of the molecule in female patients with amplification of *FGFR2* who have metastatic breast cancer. BCNM applies to identify patients with amplification of *FGFR2* in the MedBook system; 251 patients are identified of which 81 are currently seeking to enroll in trials specifically targeting their disease. The 81 eligible patients are geographically distributed over a substantial area, and only 62 are within a two-hour drive of the 14 medical centers contracted to enroll for the trial. BCNM uses the MedBook system to contact these women and create a virtual cohort. Of these patients, 56 agree to the terms of BCNM's trial enrollment criteria, which include a biopsy of at least one metastatic site to confirm *FGFR2* amplification in the metastasis. Results from the biopsies are resubmitted to MedBook per the use agreement that BCNM signed. BCNM sets an alert in the MedBook system to identify additional patients who meet their enrollment criteria, as a total of 75 patients are needed to meet the requirements of the trial design. 41 women successfully enroll in the blinded trial (above the 30 needed for BCNM interim analysis requirements). Overall, the patients who received the *FGFR2* inhibitor showed a

mean time to progression 58% longer than the control, a marginally significant effect, which encourages BCNM to continue their trial.

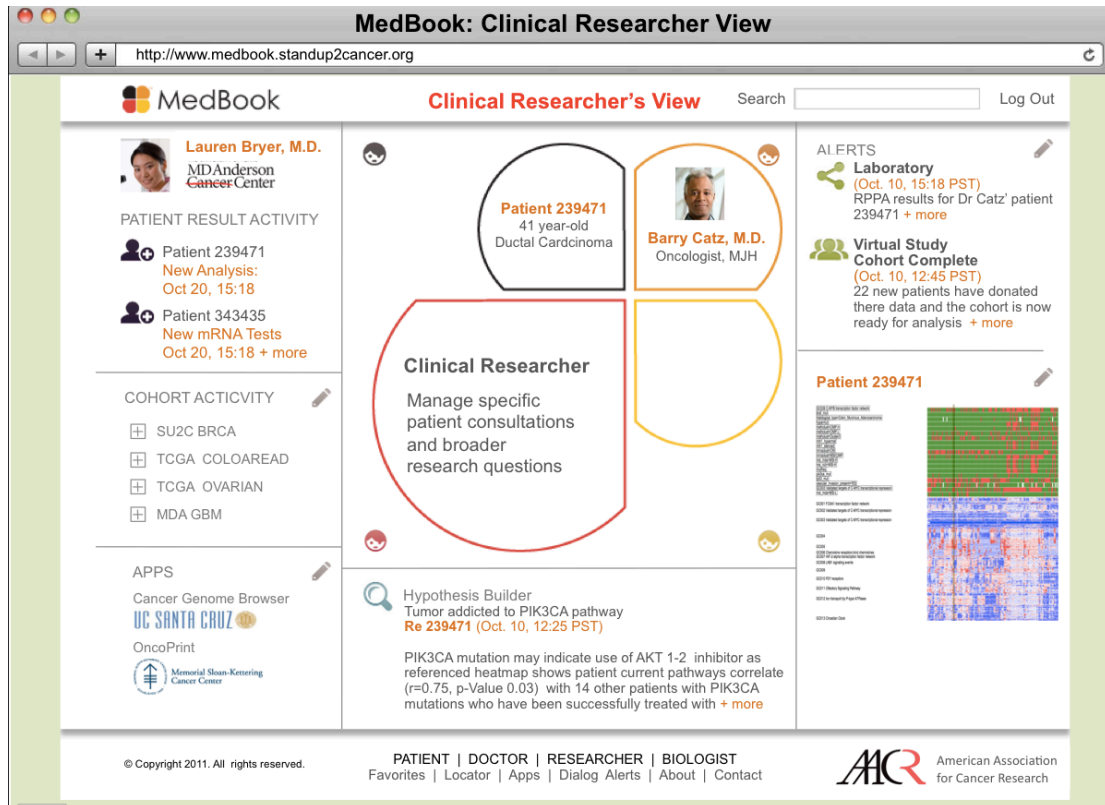


Figure 28 Clinical Researcher view of MedBook. Focus balanced between patient care and research questions, scientific data views are in context of patient information presentations.

4.3.3 Researcher Discovery Use Case

Another researcher is interested in whether a particular fusion gene is predictive of disease progression. One interesting result has been shown in the FOXM1 pathway and P53 pathways transcriptional regulators implicated in multiple cancers, but never observed in ovarian cancers. See Figure 29. Further analysis is needed to support this breakpoint and potential gene fusion, beginning by determining if the breakpoint maintains the reading frame across the exons of the fused genes.

Determining if the fusion in a tumor sample is somatic or a germline variant requires that the researcher analyze both the tumor and matched normal data. The researcher has contracted with a commercial high-throughput sequencing company to obtain shotgun reads for DNA and RNA sequencing for a patient cohort and control group.

The researcher had already enabled the BamBam mutation calling app to automatically trigger on upload events and perform somatic SNP-calling, germline and somatic inter- and intra-chromosomal rearrangements, estimates of relative copy number, and inference of loss of heterozygosity (LOH).

To address this issue, the BamBam application simultaneously dispatches every pair of sample sets to MedBook's cluster for processing. BamBam summarizes the results and ranks them by significance.

The researcher uses the NIH dbGap and the UCSC database of reconstructed cancer genomes and transcriptomes from other studies to determine what other samples exhibit these gene fusions and summarizes those studies. The researcher opens the Script Library and Authoring application, finds a previous statistical analysis. The researcher checks the reputation rating on the author to ensure that they have received positive feedback. The researcher makes some minor modifications to the script to determine drug response associations with this gene fusion.

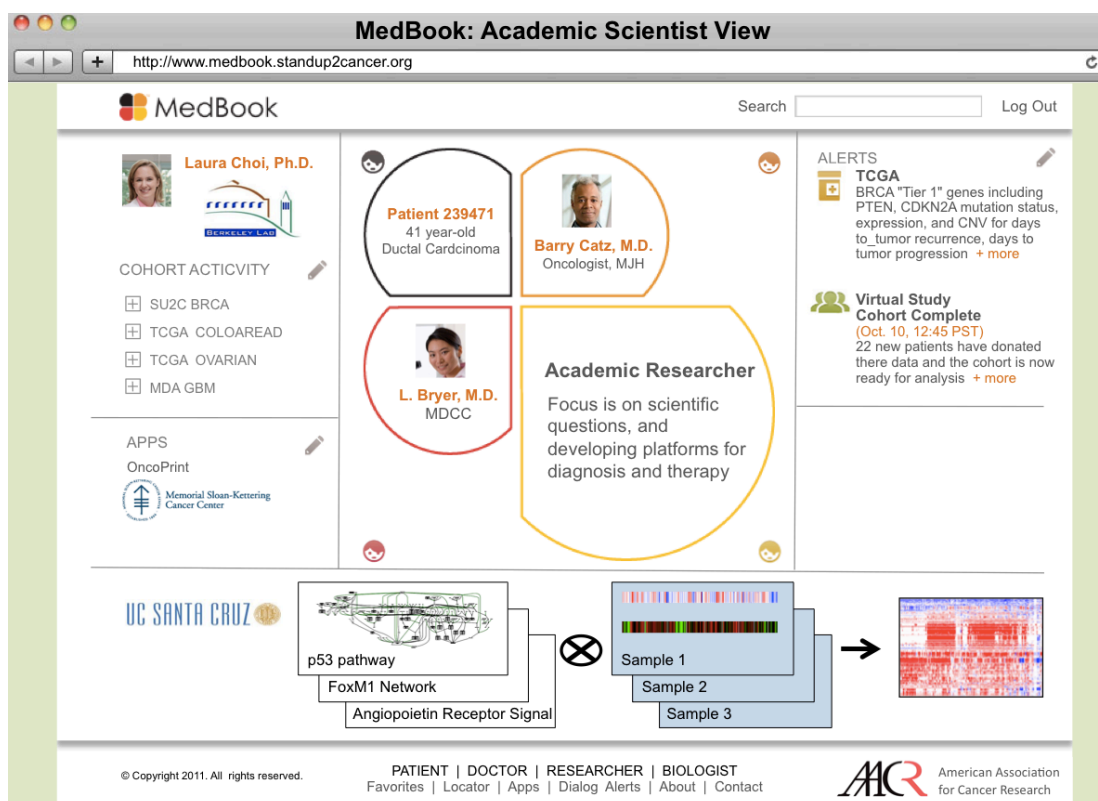


Figure 29 Academic Researcher view of MedBook. Patient names are anonymized and the tools favor research questions. Scientific data views mapped to analysis are foremost.

4.3.4 Virtual Tumor Board

Tumor boards are a key component of complex patient care, teaching and scientific discovery. Yet, bioinformatics plays little role in the process. MedBook will support a virtual tumor board app that supports the complex dialogue between surgeons, radiologists, pathologists, and therapeutic oncologists.

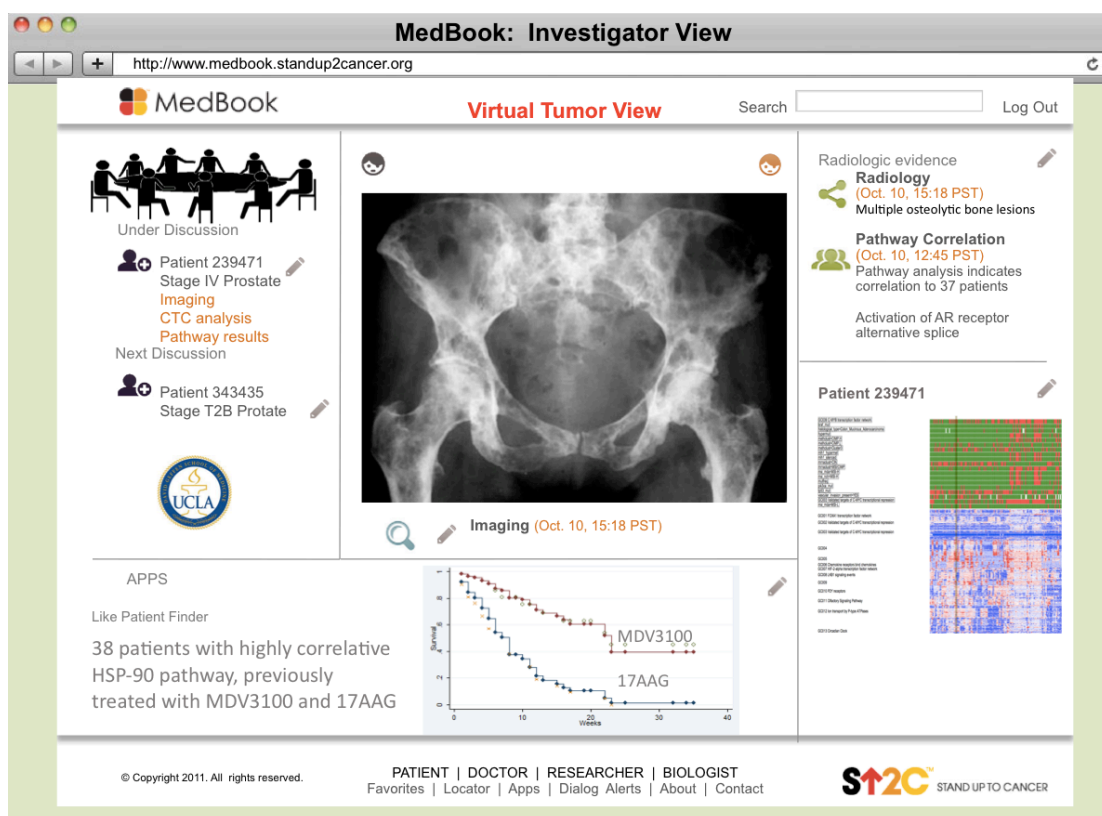


Figure 30 Virtual Tumor Board View of MedBook, which extends the conversation to include drug genomic profiling in the conversation.

4.4 Current MedBook Status

MedBook is being built from the bottom up. The first users are bioinformaticians and clinical and laboratory research investigators on the SU2C West Coast Dream Team. Histology images from the first seven patients have been loaded into MedBook as their tumors are being sequenced. Under UCSF Internal Review Board auspices, de-identified data from the UCSF clinical records system will soon be imported into MedBook. Eventually MedBook should be useful to the treating physicians, the entire clinical staff and patients and their family. Fair arbiters of rules (umpires) about use of data will be necessary. The pyramid diagram shown in Figure 31 is annotated with the current state of completion:

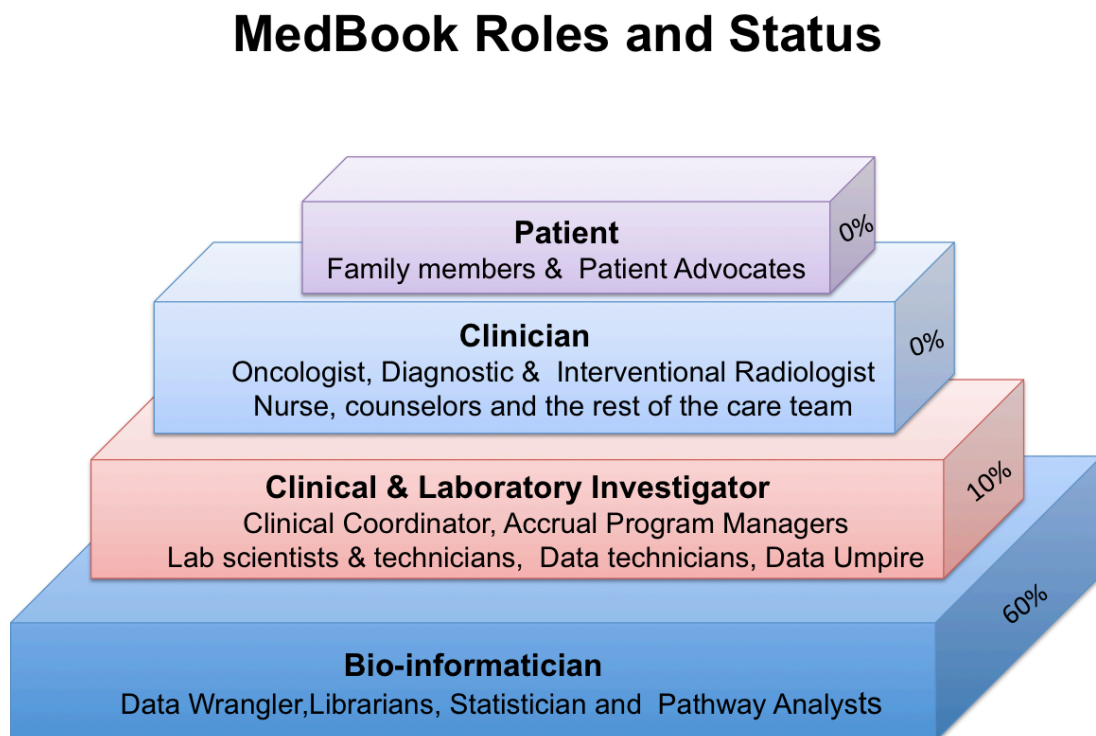


Figure 31 User community and current state of MedBook completion.

My experience in building complex systems tells me that it is necessary to rapidly build skeletal elements at each phase of a complex hierarchy. For example, at Apple, the iPod, iTunes, and iTunes Music products and store started very simply. They evolved into the sophisticated environment of iPhone, iPad, and apps supported by a sophisticated developer community over multiple years and driven by feedback.

4.4.1 Integrating High-throughput Genomic Sequencing into a Medical Context

MedBook has begun as a tool in a clinical genomic pipeline that is integrated into the electronic health record system, the clinical trial management system, and patient workflow of UCSF Hospital and the other SU2C West Coast Dream Team health care organizations. An important distinction should be made between raw genomic data analysis and medical data analysis. Raw data analysis consists of high-volume machine assay interpretation of genomic molecular features and is automated by existing batch data processing. Medical knowledge data analysis consists more of knowledge systems drawn from cellular pathways and systems biology. The underlying MedBook/Galaxy pipeline provides raw data analysis and sequence management and medical data analysis as part of a comprehensive bench to bedside loop integrated. This is illustrated below in Figure 32.

Clinical Genome Pipeline

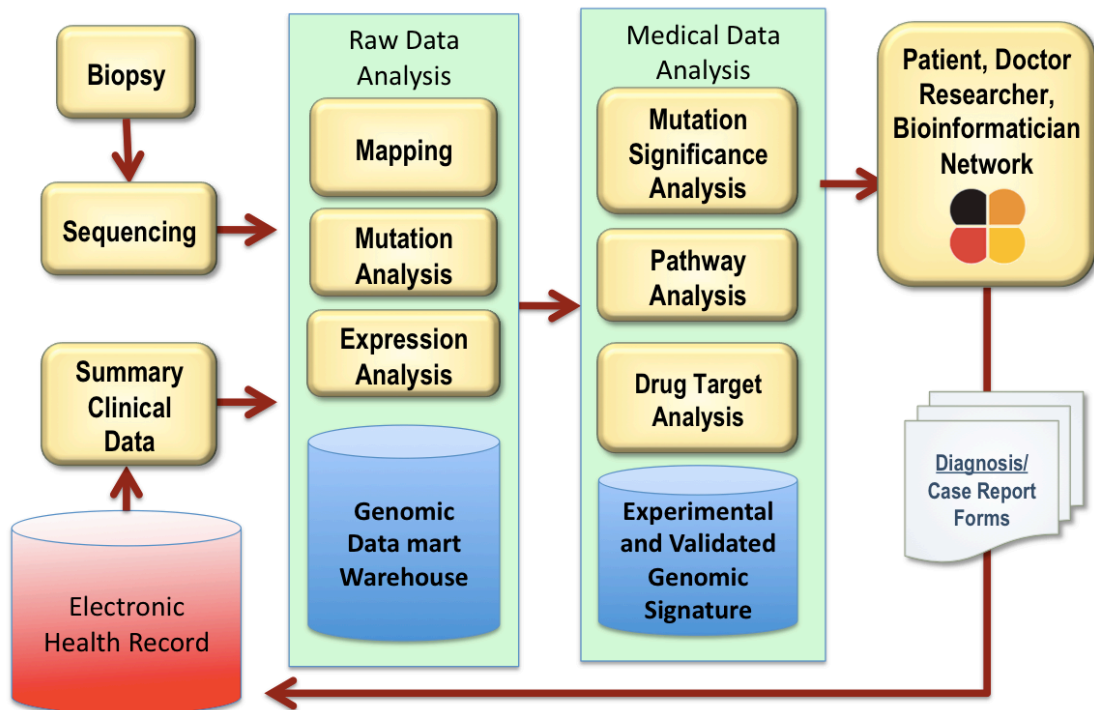


Figure 32 Clinical Genomic Pipeline segregates mechanical raw data analysis associated with genomic data from medical knowledge based analysis.

Medical findings from CLIA clinical laboratory test results data once cleared will be returned and stored into the electronic health records, clinical trial systems, and other systems of record permanent storage.

Data privacy is critical and defended by anonymization and data segregation. This part of the MedBook architecture is depicted in Figure 33 and shows the network segregation model. The red zone is operated in a hospital or other clinical environment and contains the electronic health record system, the imaging database, the clinical trial management system, and the patient sample tracking system.

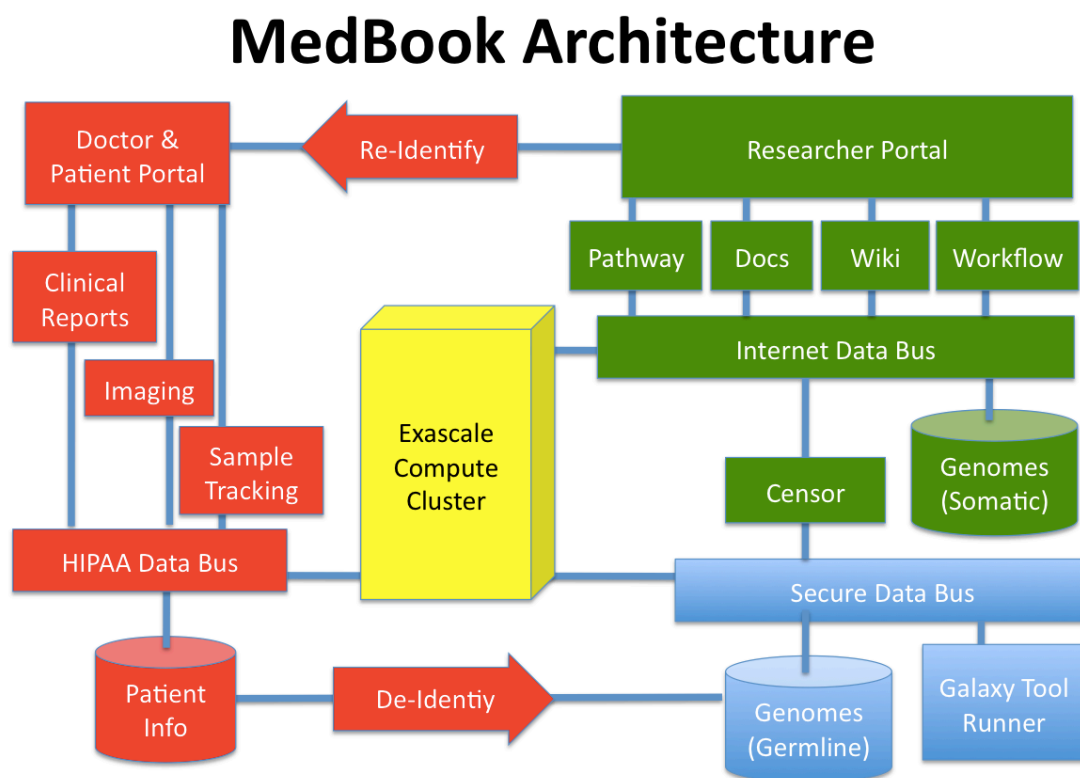


Figure 33 MedBook architecture consists of three zones: Red zone (HIPAA), Blue zone (genomic data), Green zone (anonymized data) and access to exascale computers.

Patient identity of medical information can be restricted to medical professionals in the red zone. The blue zone contains the high-volume genomic data

analysis. Blue zone genomic sequence data contains information about germline mutations. All zones need access to a robust exascale computing service. Lawrence Livermore National Laboratories has built such networks, which support shared access to super computers among systems that have different security levels, so it is possible to create this type of segregated data environment.

4.4.2 Data Analysis Workflows

Galaxy was chosen as the foundation for MedBook because Galaxy has a well-designed workflow framework. Several hundred bioinformatic tools have been developed by universities and industrial groups and adapted to work in the Galaxy system. As well, it is very straightforward to take an existing command line tool and wrap it with an XML description to make a Galaxy tool. Tools may be interactively executed or grouped together for sequential and parallel execution.

Workflows can be shared among users at multiple locations. Even in MedBook's current early stage of development, it is already common for our team to conduct dynamic collaborations where individuals are observing the progress of an investigation at multiple locations. Workflows create a history of immutable steps that can be replayed. Both the workflow and the history can be journaled for a persistent record of the data analysis. This is critical for scientific knowledge reconstruction. Reliable science requires that data can be reanalyzed multiple times.

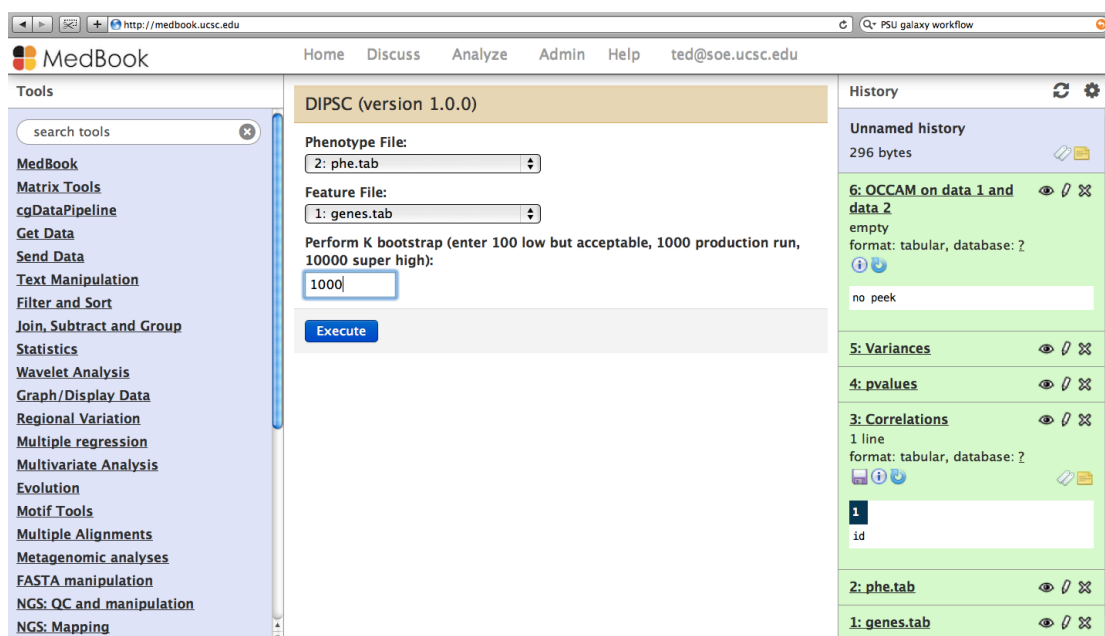


Figure 34 DIPSC tool in the workflow in MedBook

4.4.3 Visualization of Data by Dynamic Heatmaps

Data about cancer cohorts is traditionally visualized by the examination of heatmaps. MedBook heatmaps are dynamic HTML5 web pages that incorporate multiple data elements into the workflow. These multi-faceted web pages can be interactively explored by way of a point and click interface. Observations are captured directly by way of interactive comments and researchers at many different locations will see annotations that are dynamically created.

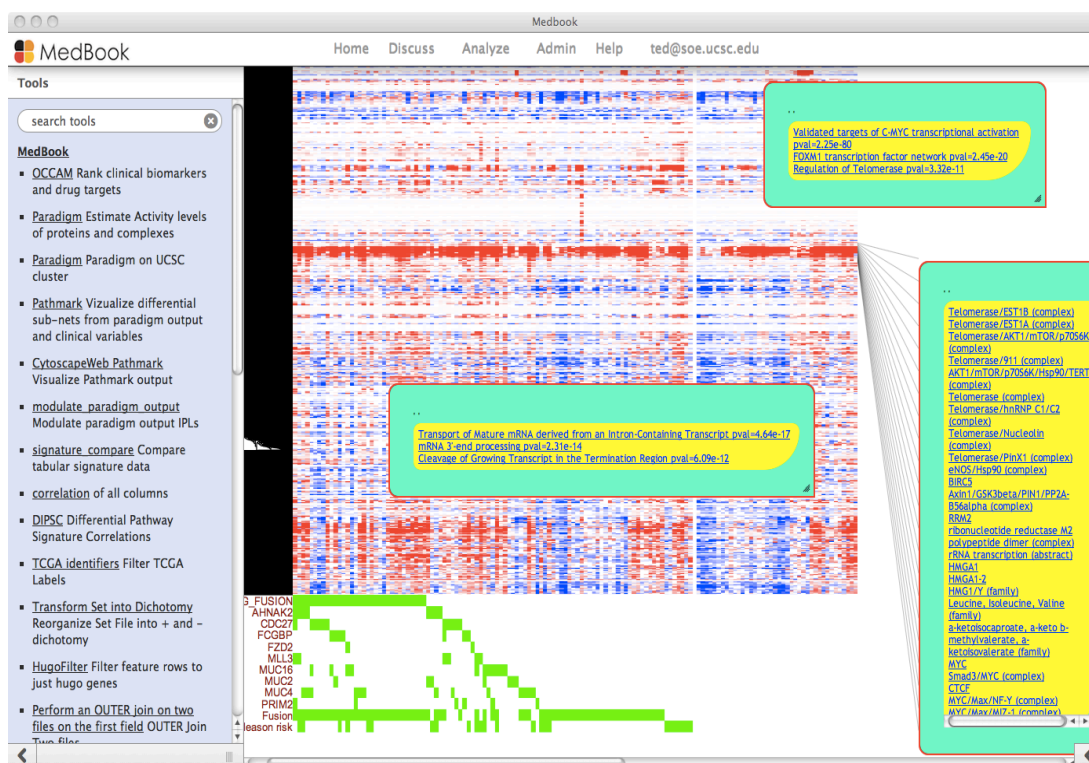


Figure 35 Dynamic Heatmap of patients by gene expression with annotations

Once annotated, a page can be published to the MedBook knowledge web and shared with one teammate, many teammates, a large collection of collaborators, or the world at large.

4.4.4 OCCAM and DIPSC Tools

The OCCAM and DIPSC tools are characteristic of simple, easy to use tools.

A few simple parameters define the input, which is a set of phenotypes and IPL features (see Figure 36). The phenotype file is either a dichotomized file or a file that the automatic dichotomizer can handle. Phenotype labels are listed across the columns (top) and samples are listed along the side (the first column). The automatic dichotomizer splits numeric columns based on the mean. Enumerations (for example: *LumA*, *LumB*, *Her2*, *Basal*) are broken into four columns: (*LumA* vs. *not*), (*LumB* vs. *not*), (*Her2* vs. *not*) (*Basal* vs. *not*). The dichotomized results can also be run separately and processed using other tools.

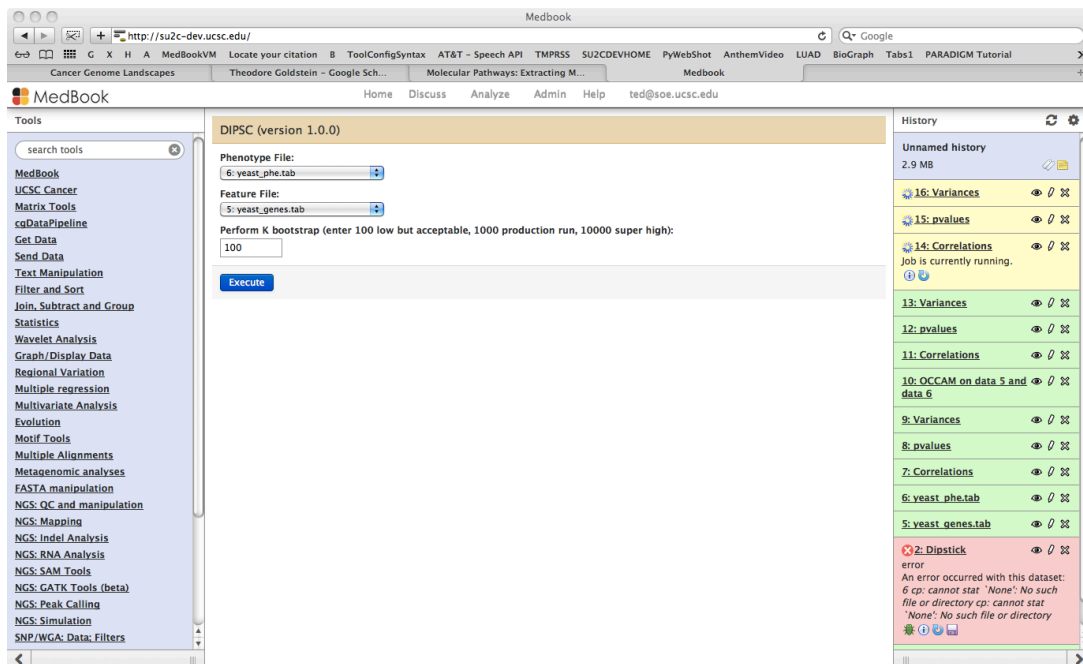


Figure 36 MedBook DIPSC tool requires setting the phenotype and clinical feature file and produces correlations, p-values and variance data sets.

4.4.5 DipStick Tool

The DipStick tool is an intuitive visualization tool that provides a readable alternative to heatmaps. The columns of a heatmap correlation matrix are sorted in descending order and mapped along the vertical access of the red/blue gradient. The DipStick tool uses the correlation, variance, and p-values associated with each pair of correlations. These data are easily produced by the previously described DIPSC tool.

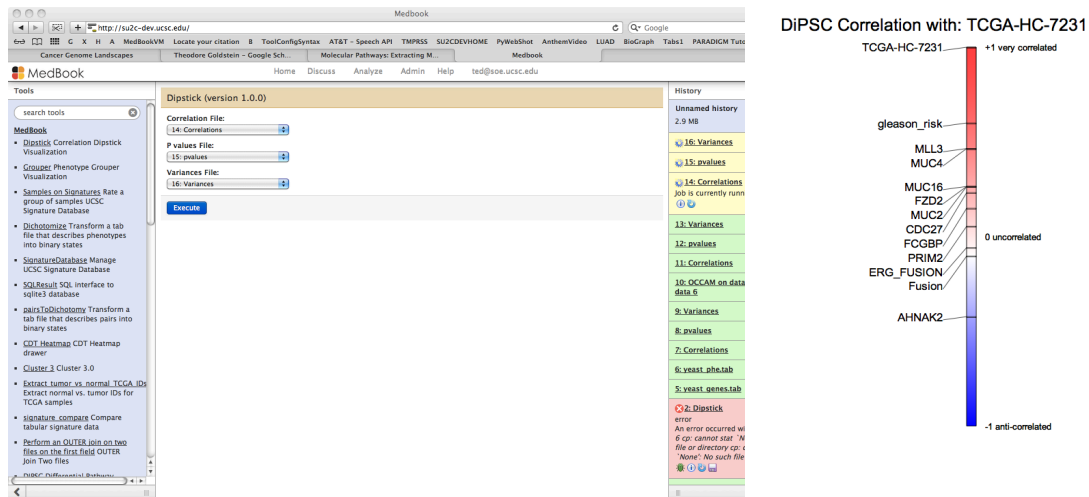
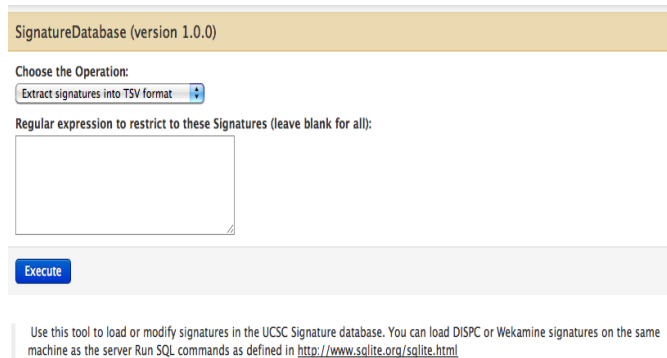


Figure 37 DipStick tool (left) and DipStick figure (right).

4.4.6 UCSC Activitome Signature Database



The screenshot shows a web interface for the UCSC Activitome Signature Database. At the top, a yellow header bar contains the text "SignatureDatabase (version 1.0.0)". Below this, the section "Choose the Operation:" features a dropdown menu with "Extract signatures into TSV format" selected. Underneath, a label reads "Regular expression to restrict to these Signatures (leave blank for all):" followed by a large, empty text input box. A blue "Execute" button is positioned below the input box. At the bottom, a small text block provides instructions: "Use this tool to load or modify signatures in the UCSC Signature database. You can load DISPC or Wekamine signatures on the same machine as the server Run SQL commands as defined in <http://www.sqlite.org/sqlite.html>".

Figure 38 UCSC Activitome Signature Database Tool in MedBook

The UCSC Activitome Signature Database shown in Figure 38 and Figure 39 is based on phenotypes and activitome data. The first signatures are derived from the DIPSC method, but it has also been extended to include graduate student James Durbin's signatures derived from the WEKAmine machine learning infrastructure. UCSF research oncologists want to use this tool to develop a signature to recognize a gene-fusion event with the minimum number of genes without cost prohibitive diagnostic tests. The goal is to subdivide patients into classes that make them good candidates for particular drugs.

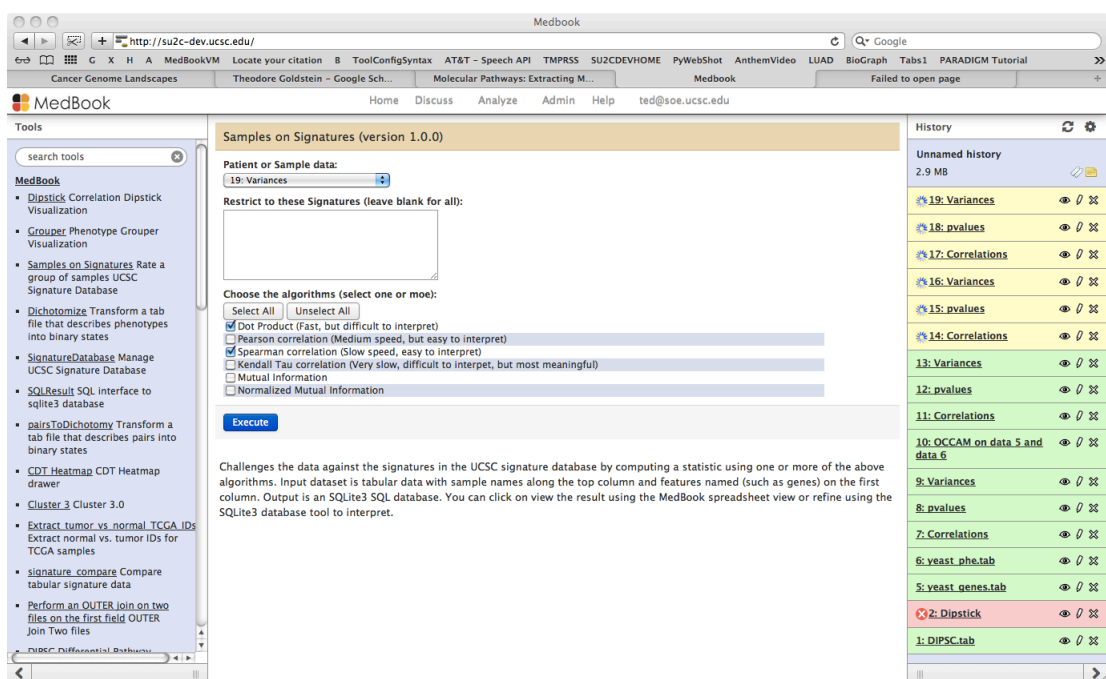


Figure 39 UCSC Activitome Signature Database Differential App allows the user to run various methods comparing signatures to samples.

4.5 Business Model

I propose that MedBook be funded and operated by a network of hospitals, research institutions, pharmaceutical companies, and other related interested entities in a federated network. The network itself might be operated as a nonprofit foundation or as a public benefit corporation along the lines of a VISA or MasterCard institution. It may charge fees, but it returns any residual after expenses back to the entities that incurred the fees according to some equitable formula.

4.5.1 Rules of Engagement

Creating a central medical and research data repository faces the huge challenges of data collection, standardization, quality assurance, security, auditability, and customer (patients, physicians and researchers) acceptance. As well, there are increasing numbers of both white hat and black hat hackers who wish to hijack a system for sport, profit, prestige, politics, or entertainment. As of this writing, Charles Schwab brokerage website was down this evening for a few hours because of a denial of network service attack. All of these reasons reinforce the need for a federated system design that allows for part of the MedBook network of systems to be down without requiring any central repository to be up.

One idea to borrow from the VISA and MasterCard world is *stand in processing* (STIP), which allows one system to temporarily fill in for another system, albeit with some possible loss of function.

The first bottleneck in data analysis is the time it takes to collect the information from diverse sources. The incentive to use applications and cross-compare samples to others in the system will motivate users to asynchronously contribute data. Rather than putting the burden of data input on a single central team, it distributes the work to contributing centers and individuals.

It is essential that the system ensures patient confidentiality and preserves the physician-patient relationship. There is concern that data incorrectly interpreted by patients may cause anxiety and poorly advised decisions (such as medically unwarranted prophylactic mastectomies) Some apps may require that a physician, or

appropriately authorized care-provider consent for a patient to be able to view the results only after consultation. It also, for example, should allow a genetic counselor to consult with a patient first before running a breast-cancer predisposition prediction app.

Patient privacy and anonymity is essential. The system will be built to HIPAA confidentiality standards. Highly sensitive data, such as HIV status, will be encrypted and not browsable.

4.5.2 System Implementation Considerations

OCCAM and the UCSC Cancer Signature database are steps toward a clinical decision support system for oncologists. This is a software architecture that provides a larger context for the framework used in building OCCAM and the UCSC Activitome Signature Database. It is useful to think about the computer software architecture long before software developers know all of the details on how to build it.

Building a system that allows: 1) large central repositories; 2) decentralized hospital-centered databases, and 3) cloud-based solutions for individual clinicians combined with a method to exchange data is crucial to success.

4.6 Conclusion

The rate of scientific discovery is challenging to all of the healthcare community, but especially so to oncology. Part of the solution is to turn oncology into a rapid learning community—both for the potential to improve patient outcomes as well as to accelerate drug development. Collaborative research teams in partnership with major cancer centers, can rapidly prototype such a community and evolve it into a national cancer trials network for the 21st Century. Imagine no boundaries between cancer trials and care. Every patient is an N-of-1 trial, receiving the best possible treatment based on everything that is known everywhere about their disease, while contributing data that advances the standard of care for all. Moreover, by virtue of its scale and informatics leadership, the MedBook community will become the default partner of choice for researchers in academia and pharmaceutical companies and biotech companies seeking to rapidly validate a rational treatment hypothesis or to test an investigational drug.

Our current methods for sharing results are slow, but our current methods of discovery are getting faster. In this thesis, I have shown that genomic big data can be turned into clinically relevant knowledge using my OCCAM and DIPSC tools. I have commenced the construction of the *activitome*, a knowledge base of signatures that characterize cancer's mutations, clinical phenotypes and drug signatures using my tools.

In MedBook, I have shown a vision for how to use these tools as part of a social-network infrastructure to gather the data, process the data into knowledge, and

to use that knowledge to deliver better care. The MedBook system and community will learn from every success and every failure to improve the process for all patients in a continuous cycle of discovery and translation.

The discovery process that brings a finding from one research group to another or to a physician and ultimately to a patient currently takes years to decades. This slow pace has everything to do with the way the information is organized and communicated. MedBook will provide an infrastructure to greatly improve communication and collaboration speeds. Producing an effective web service of this scale requires a large team. Patient advocates, patients, physicians, clinical researchers, web designers, and software experts working together can make this happen today and have already started. In this way, MedBook will create a system that will live well beyond the lifetime of my project.

Glossary

Activitome: the collected knowledge base of cancer's activity signatures that will guide treatment and assist in developing new therapies.

Activitome signature: the mathematical representation (typically a vector of probabilities) of the molecular differential diagnostic effect of a molecular event or clinical phenomenon.

Annotation app: allows physicians and researchers to discuss clinical data.

App: an application that runs in MedBook that stores and retrieves information, analyzes data, generates hypothesis, tracks events, and makes predictions.

Assay: a biological or clinical medical test that measures some physical phenomenon that is of interest to researchers, doctors or patients.

BamBam: rapid mutation analyzer.

Biomarker: an indicator of the underlying biological molecular state of an individual or cell. Biomarkers frequently guide treatment decisions.

Biomedical Evidence Graph (BMEG): The BMEG is a hypergraph (a graph of graphs) whose constituent graphs represent a web of knowledge.

Bootstrap: a method for estimating a statistic from a set of samples by *sub sampling*, a repeated process of randomly selecting a *subset of samples*.

Case Report: allows a physician to describe interesting patients and invite discussion.

Clinical trials matcher: a web-based app that matches the demographic and genomic profile of a patient to the set of available clinical trials.

Commons: a discussion forum assembled within MedBook either explicitly or implicitly based on shared attributes.

Copy number: the number of copies of DNA of a gene in a cell.

Data Wrangler: a researcher or technician who uses the *Data Wrangler App*.

Data Wrangler App: a class of MedBook application that rapidly uploads, annotates, and curates cancer data sets.

Differential Signature App: an app that performs differential analysis using activitome signatures

DIPSC: the Differential Pathway Signature Correlation tool determines the correlation of activitome signatures connected to events (such as mutation) even when the events occur on the same set of samples.

Dip: a single sub-sampling and correlation of a cohort used in DIPSC. Some statistical literature calls this a *sub-sampling replica*.

DipStick: a one-dimensional data visualization tool.

Donate Your Data app: an app that empowers patients to authorize and upload their data for the common benefit and for their own benefit.

Genomic Lesion: a mutation, gene fusion, or other (typically deleterious) alteration in a patient's DNA that causes disease or pathway dysfunction.

Grouper: allows researchers to assemble cohorts of samples or patients by querying a database or establishing predicates that are stored queries.

Heatmap: a two-dimensional data visualization tool.

Hypotheses Structured Discussion: Social media forum for patients with the same molecular subtype or on the same trial or drug, for docs who want to discuss similar cases, for researchers who want to collaboratively discuss molecular disease models, as well as what pathways are involved in each subtype, and how to block them.

MedBook: a web-based social network that unites patients, families, doctors, and researchers to create knowledge to treat and cure disease such as cancer.

Molecular Disease Model Builder: an application which assists a board of experts to create a Molecular Disease Model (MDM) that defines the subtypes based on the original tissue, the patient's demographics and environmental drivers, and the molecular and pathway alterations as well as the cancer's stage.

OCCAM: OmiC data Cancer Analytic Model, a software method and tool that creates activitome signatures, mathematical representations that help to find new biomarkers to guide treatment using existing drugs and to find new molecular targets that show high potential for new drugs.

Ome: a collection of biological knowledge.

OncoPrint: identifies collections of mutations in a cohort of patients that may act synergistically in disease onset or progression.

Outcomes app: genomic and outcomes data is analyzed, and the findings reported, integrated and disseminated. It captures the patient's general health state through treatments.

Outlier app: allows researchers to search for interesting samples and patients across all data sets.

PARADIGM: maps genomic and expression data onto curated pathway databases using a probabilistic graphical model and infers the state of unmeasured components.

Pathway Signatures: the output of PARADIGM

Patient blog: an online discussion forum that allows a patient to describe the effects of the cancer and the treatment.

Publication Builder: allows researchers to create time-driven collaborations that allow for data assembly, document authoring, and publication to major journals.

Prognostic apps: Apps that allow a single patient's data to be compared to existing cohorts to find commonalities.

SAM: Statistical Analysis of Microarray method that uses a modified *t-test* to provide a list of differentially altered genes or features.

Sample and Data Dashboard: database app which shows which labs have what sample; what is the current stage of analysis; provides direct access to data and results.

Script Library: Library of reusable data processing scripts and online authoring tools.

Subtype: a biological grouping of a set of patients with common disease etiology and treatment.

Pathway: a series of biochemical and reactions in a cell or between cells that process information, metabolic material, or both.

Signature: see activitome signature.

SuperPathway: the collection of all pathways in a cell.

t-test: a statistical test that determines whether two groups belong to the same distribution using William Stanley Gossets' *Students t distribution*.

Targeted Therapy: a drug that targets a particular molecular pathway, as opposed to cytotoxic chemotherapy that targets any growing cell.

Therapy Finder App: app that uses the molecular characteristics, the stage, and the site of origin of the tumor to match it to a subtype and a therapy.

Translational Medicine: the practice of taking the lessons from research and applying them to clinical practice.

Volcano plot: a chart that depicts p-value significance on the vertical axis and some other attribute, such as correlation or fold change on the horizontal axis.

References

- [Albert 2006] Albert, J. M. (2006). Targeting the Akt/mammalian target of rapamycin pathway for radiosensitization of breast cancer. *Molecular Cancer Therapeutics*, 5(5), 1183–1189. doi:10.1158/1535-7163.MCT-05-0400.
- [Allison 2008] Allison, M. (2008). Trouble at the office. *Nature Biotechnology*, 26(9), 967–969. doi:10.1038/nbt0908-967.
- [Auber 2002] David Auber. Tulip. GRAPH DRAWING. Lecture Notes in Computer Science, 2002, Volume 2265/2002, 488-491, DOI: 10.1007/3-540-45848-4_34.
- [Azmy 2013] Azmy, A., Abdelwahab, S., & Yassen, M. (2013). Oxaliplatin and Bolus-Modulated 5-Fluorouracil as a Second-Line Treatment for Advanced Pancreatic Cancer: Can Bolus Regimens Replace FOLFOX When Considered for Second Line? *ISRN oncology*, 2013, 358538. doi:10.1155/2013/358538.
- [Baker 2013] Baker, M. (2013, February 28). Big biology: The 'omes puzzle. *Nature*, pp. 416–419. doi:10.1038/494416a.
- [Barker 2009] Barker, A. D., Sigman, C. C., Kelloff, G. J., Hylton, N. M., Berry, D. A., & Esserman, L. J. (2009). I-SPY 2: An Adaptive Breast Cancer Trial Design in the Setting of Neoadjuvant Chemotherapy. *Nature Translational Medicine*.
- [Bell 2011] Bell, D., Berchuck, A., Birrer, M., Chien, J., Cramer, D. W., Dao, F., et al. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353), 609–615. doi:10.1038/nature10166.
- [Benjamini 2001] Benjamini and Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* (2001) vol. 29 (4) pp. 1165-1188.
- [Bender 1999] Ralf Bender, S. L. (1999). Multiple test procedures other than Bonferroni's deserve wider use. *BMJ : British Medical Journal*, 318(7183), 600.
- [Benz 1982] Benz, C. Optimal scheduling of methotrexate and 5-fluorouracil in human breast cancer. *Cancer Res* (1982) vol. 42 pp. 2081-2086.
- [Bielas 2006] Jason H. Bielas, Keith R. Loeb*, Brian P. Rubin, Lawrence D. True, and Lawrence A. Loeb. Human cancers express a mutator phenotype. *Proceedings of the National Academy of Sciences* (2006) vol. 103 (48) pp. 18238.

- [Bielas 2006] Jason H. Bielas, Keith R. Loeb*, Brian P. Rubin, Lawrence D. True, and Lawrence A. Loeb. Human cancers express a mutator phenotype. *Proceedings of the National Academy of Sciences* (2006) vol. 103 (48) pp. 18238.
- [Bocci 2006] Bocci, G., Barbara, C., Vannozzi, F., Di Paolo, A., Melosi, A., Barsanti, G., et al. (2006). A pharmacokinetic-based test to prevent severe 5-fluorouracil toxicity. *Clinical pharmacology and therapeutics*, 80(4), 384–395. doi:10.1016/j.clpt.2006.06.007.
- [Butte 2006] Atul J Butte, Andrea Califano, Stephen Friend, Trey Ideker, Eric Schadt. Integrative Network-based Association Studies: Leveraging cell regulatory models in the post-GWAS era. *Nature* (2011) pp. 1-22
- [Cairns 2006] John Cairns. Mutation selection and the natural history of cancer. *Science's SAGE KE* (2006) vol. 2006 (10).
- [Caponigro 2011] Caponigro, G., & Sellers, W. R. (2011). Advances in the preclinical testing of cancer therapeutic hypotheses. *Nature Reviews*, 10, 179–187. doi:10.1038/35077232.
- [Chen 2007] Chen R, Li L, Butte AJ. AILUN: reannotating gene expression data automatically. *Nature methods* 2007; 4: 879.
- [Ciriello 2012] Ciriello G, Cerami E, Sander C, Schultz N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome research* 2012; 22: 398-406.
- [Collisson 2012] Collisson, E. A., Cho, R. J., & Gray, J. W. (2012). What are we learning from the cancer genome? *Nature Publishing Group*, 9(11), 621–630. doi:10.1038/nrclinonc.2012.159.
- [Crick 1958] On protein synthesis. *Symp Soc Exp Biol.* 1958 - PubMed - NCBI. (1958). http://profiles.nlm.nih.gov/SC/B/B/F/T/_/scbbft.pdf
- [Daley 2003] Daley, G. Q. (2003). Gleevec resistance: lessons for target-directed drug development. *Cell cycle (Georgetown, Tex.)*, 2(3), 190–191.
- [De Benedetti 2001] De Benedetti, V. M., Travis, L. B., Welsh, J. A., van Leeuwen, F. E., Stovall, M., Clarke, E. A., et al. (n.d.). p53 mutations in lung cancer following radiation therapy for Hodgkin's disease. CEBP, AACR, 2001.
- [Dees 2012] Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., et al. (2012). MuSiC: Identifying mutational significance in cancer genomes. *Genome Research*, 22(8), 1589–1598.

- [De Giorgi 2011] De Giorgi, U., Cohen, E. N., Gao, H., Mego, M., Lee, B.-N., Lodhi, A., et al. (2011). Mesenchymal stem cells expressing GD2 and CD271 correlate with breast cancer-initiating cells in bone marrow. *Cancer Biology & Therapy*, 11(9), 812–815. doi:10.4161/cbt.11.9.15178
- [de Hoon 2004] de Hoon, M. J. L. M., Imoto, S. S., Nolan, J. J., & Miyano, S. S. (2004). Open source clustering software. *Bioinformatics (Oxford, England)*, 20(9), 1453–1454.
- [Diehn 2006] Maximilian Diehn, Ramona Bhattacharya, David Botstein, Patrick O Brown. Genome-Scale Identification of Membrane-Associated Human mRNAs. *PLoS Genet* (2006) vol. 2 (1) pp. e11.
- [Dobzhansky 1973] Dobzhansky, T. (1973). Nothing in biology makes sense except in the light of evolution. *American Biology Teacher*.
- [Efron 1990] Efron, B. (1990). Taylor & Francis Online :: More Efficient Bootstrap Computations - Journal of the American Statistical Association - Volume 85, Issue 409. *Journal of the American Statistical Association*.
- [Efron 2007] Bradley Efron and Robert Tibshirani. On testing the significance of sets of genes. *Annals of Applied Statistics* (2007) vol. 1 (1) pp. 107-129.
- [Eisen 1998] Eisen, M. B. M., Spellman, P. T. P., Brown, P. O. P., & Botstein, D. D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(25), 14863–14868.
- [Ellis 2011] Ellis MJ, Suman VJ, Hoog J, Lin L, Snider J, Prat A et al. Randomized phase II neoadjuvant comparison between letrozole, anastrozole, and exemestane for postmenopausal women with estrogen receptor-rich stage 2 to 3 breast cancer: clinical and biomarker outcomes and predictive value of the baseline PAM50-based intrinsic subtype--ACOSOG Z1031. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2011; 29: 2342-2349.
- [Ellis 2012] Ellis MJ, Ding L, Shen D, Luo J, Suman VJ, Wallis JW et al. Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* 2012; 486: 353-360.
- [Engreitz 2010] Jesse M Engreitz, Alexander A Morgan, Joel T Dudley, Rong Chen, Rahul Thathoo, Russ B Altman, Atul J Butte. Content-based microarray search using differential expression profiles. *Bioinformatics* 2010, 11:603.

- [FDA 2013] Downloaded April 13, 2013 from <http://www.fda.gov/drugs/scienceresearch/researchareas/pharmacogenetics/ucm083378.htm>
- [Fisher 1915] Fisher R.A. “Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population”. *Biometrika* 10 (4): 507–521.
- [Fodde 2002] Riccardo Fodde and Ron Smits. Cancer biology. A matter of dosage. Science (New York, NY) (2002) vol. 298 (5594) pp. 761.
- [Franke 2003] Franke, T. F., Hornik, C. P., Segev, L., Shostak, G. A., & Sugimoto, C. (2003). PI3K/Akt and apoptosis: size matters. *Oncogene*, 22(56), 8983–8998.
- [Friend 2013] Friend, S. H., & Norman, T. C. (2013). Metcalfe's law and the biology information commons. *Nature Biotechnology*, 31(4), 297–303.
- [Gagnon-Bartsch 2012] Gagnon-Bartsch, J. A., & Speed, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3), 539–552.
- [Goecks 2010] Goecks, J., Nekrutenko, A., Taylor, J., Galaxy Team. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology*, 11(8), R86.
- [Goetz 2004] Matthew P. Goetz, Matthew M. Ames and Richard M. Weinshilboum. Primer on medical genomics part XII: pharmacogenomics—general principles with cancer as a model. *Mayo Clinic Proceedings* (2004) vol. 79 (3) pp. 376
- [Goldstein 1996] TC Goldstein. The Gateway Security Model in the Java Electronic Commerce Framework, LECTURE NOTES IN COMPUTER SCIENCE FINANCIAL CRYPTOGRAPHY, Springer 1996.
- [Goldstein 2013] Goldstein, T., Paull, E. O., Ellis, M. J., & Stuart, J. M. (2013). Extracting Medical Knowledge from High-throughput Genomic Data. *AACR Clin Cancer Res*. Published Online Feb 21, 2013 [Epub ahead of print].
- [Gray 2003] Allan Balmain, Joe Gray, Bruce Ponder. The genetics and genomics of cancer. *Nat Genet* (2003) vol. 33 (3s) pp. 238-244.
- [Green 2011] Green, E. D., Guyer, M. S., Green, E. D., Guyer, M. S., Manolio, T. A., & Peterson, J. L. (2011). Charting a course for genomic medicine from base pairs to bedside. *Nature*, 470(7333), 204–213. doi:10.1038/nature09764

- [Greenfield 2010] Greenfield, A., Madar, A., Ostrer, H., & Bonneau, R. (2010). DREAM4: Combining genetic and dynamic information to identify biological networks and dynamical models. *PloS One*, 5(10), e13397.
- [Gutschner 2012] Gutschner, T., Hammerle, M., Eissmann, M., Hsu, J., Kim, Y., Hung, G., et al. (2013). The Noncoding RNA MALAT1 Is a Critical Regulator of the Metastasis Phenotype of Lung Cancer Cells. *Cancer research*, 73(3), 1180–1189.
- [Hanahan 2000] Douglas Hanahan and Robert A. Weinberg. The Hallmarks of Cancer Review. *Cell* (2000)
- [Heiser 2011] Laura M. Heiser, Anguraj Sadanandam,, Wen-Lin Kuo, Stephen C. Benz, Theodore C. Goldstein, Sam Ng, William J. Gibb, Nicholas J. Wang, Safiayyah Ziyad, Frances Tong, Nora Bayani, Zhi Hu, Jessica I. Billig, Andrea Dueregger, Sophia Lewis, Lakshmi Jakkula, James E. Korkola, Steffen Durinck, François Pepin, Yinghui Guan, Elizabeth Purdom, Pierre Neuvial, Henrik Bengsston, Kenneth W. Wood, Peter G. Smith, Lyubomir Vassilev, Bryan T. Hennessy, Joel D. Greshock, Kurtis E. Bachman, Mary Ann Hardwicke, John W. Park, Laurence Marton, Denise M. Wolf, Eric A. Collisson, Richard M. Neve, Gordon B. Mills, Terence P. Speed, Heidi S. Feiler, Richard F. Wooster, David Haussler,, Josh Stuart, Joe W. Gray, and Paul T. Spellman. Subtype and pathway specific responses to anti-cancer compounds in breast cancer. Submitted to Proceedings of the National Academy of Sciences (2011)
- [Hemelaar 2006] Hemelaar, J., Gouws, E., Ghys, P. D., & Osmanov, S. (2006). Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. *AIDS*, 20(16), W13–W23. doi:10.1097/01.aids.0000247564.73009.
- [Hochhaus 2002] Hochhaus, A., Kreil, S., Corbin, A. S., La Rosée, P., Müller, M. C., Lahaye, T., et al. (2002). Molecular and chromosomal mechanisms of resistance to imatinib (STI571) therapy. *Leukemia*, 16(11), 2190–2196.
- [Huang 2010] Haiyan Huanga, Chun-Chi Liub, and Xianghong Jasmine Zhou. Bayesian approach to transforming public gene expression repositories into disease diagnosis databases. *Proceedings of the National Academy of Sciences* (2010) vol. 107 (15) pp. 6823
- [Huang 2010] Huang H, Liu CC, Zhou XJ. Bayesian approach to transforming public gene expression repositories into disease diagnosis databases. *Proceedings of the National Academy of Sciences of the United States of America* 2010; 107: 6823- 6828.

- [Hurwitz 2004] Hurwitz, H., Fehrenbacher, L., Novotny, W., Cartwright, T., Hainsworth, J., Heim, W., et al. (2004). Bevacizumab plus irinotecan, fluorouracil, and leucovorin for metastatic colorectal cancer. *New England Journal of Medicine*, 350(23), 2335–2342. doi:10.1056/NEJMoa032691
- [Jamieson 2004] Jamieson, C. H. M., Ailles, L. E., Dylla, S. J., Muijtjens, M., Jones, C., Zehnder, J. L., et al. (2004). Granulocyte–Macrophage Progenitors as Candidate Leukemic Stem Cells in Blast-Crisis CML. *New England Journal of Medicine*, 351(7), 657–667. doi:10.1056/NEJMoa040258
- [Kennecke 2012] Kennecke, H. F., Speers, C. H., Ennis, C. A., Gelmon, K., Olivotto, I. A., & Hayes, M. (2012). Impact of Routine Pathology Review on Treatment for Node-Negative Breast Cancer. *Journal of Clinical Oncology*, 30(18), 2227–2231.
- [Kent 2002] W. James Kent, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, and David Haussler. The human genome browser at UCSC. *Genome research* (2002) vol. 12 (6) pp. 996
- [Kent 2003] W. James Kent, Robert Baertsch, Angie Hinrichs, Webb Miller, and David Haussler. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences of the United States of America* (2003) vol. 100 (20) pp. 11484
- [Kim 2008] KB Kim. Phase II trial of imatinib mesylate in patients with metastatic melanoma. *British Journal of Cancer* (2008) vol. 99 pp. 734-740
- [Knudson 1971] Alfred G. Knudson, Jr. Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences of the United States of America* (1971) vol. 68 (4) pp. 820
- [Kobayashi 2005] Kobayashi, S., Boggon, T. J., Dayaram, T., Jänne, P. A., Kocher, O., Meyerson, M., et al. (2005). EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. *New England Journal of Medicine*, 352(8), 786–792. doi:10.1056/NEJMoa044238
- [Kohane 2011] Zachary Kohane. The McDonald's Rx: How Computers Can And Should Change Doctoring. <http://commonhealth.wbur.org/2011/02/isaac-kohane-mcdonalds-doctors/> downloaded and verified 19 March 2011
- [Kops 2005] Geert J. P. L. Kops, Beth A. A. Weaver and Don W. Cleveland. On the road to cancer: aneuploidy and the mitotic checkpoint. *Nat Rev Cancer* (2005) vol. 5 (10) pp. 773-785

- [Kschischang 2001] Kschischang, Frey, Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory* 2001; 47.
- [Kobayashi 2005] Kobayashi, S., Boggon, T. J., Dayaram, T., Jänne, P. A., Kocher, O., Meyerson, M., et al. (2005). EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. *New England Journal of Medicine*, 352(8), 786–792. doi:10.1056/NEJMoa044238
- [Kucherlapati 2012] Comprehensive molecular characterization of human colon and rectal cancer. (2012). Comprehensive molecular characterization of human colon and rectal cancer., 487(7407), 330–337.
- [Langland 2013] Langlands, F. E., Horgan, K., Dodwell, D. D., & Smith, L. (2013). Breast cancer subtypes: response to radiotherapy and potential radiosensitisation. *British Journal of Radiology*, 86(1023), 20120601–20120601.
- [Lingle 1978] Wilma L. Lingle, Susan L. Barrett, Vivian C. Negron, Antonino B. D'Assoro, Kelly Boeneman, Wanguo Liu, Clark M. Whitehead, Carol Reynolds, and Jeffrey L. Salisbury. Centrosome amplification drives chromosomal instability in breast tumor development. *Proceedings of the National Academy of Sciences of the United States of America* (2002) vol. 99 (4) pp. 1978
- [Lobo 2007] Lobo, N. A., Shimono, Y., Qian, D., & Clarke, M. F. (2007). The Biology of Cancer Stem Cells. *Annual Review of Cell and Developmental Biology*, 23(1), 675–699.
- [Lopes 2010] Christian T Lopes, Max Franz, Farzana Kazi, Sylva L Donaldson, Quaid Morris, Gary D Bader, Cytoscape Web: an interactive web-based network browser. *Bioinformatics* (2010) vol. 26 (18) pp. 2347
- [Loscalzo 2007] Joseph Loscalzo, Isaac Kohane & Albert-Laszlo Barabasi. Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. *Mol Syst Biol* 3: 124.
- [Maley 2010] Carlo Maley. Video lecture at Cancer Speaks. <http://cancer-insights.asu.edu/2010/04/video-carlo-maley-phd-speaks-at-asu-seminar/>
- [Manning 2002] G. Manning, D. B. Whyte, R. Martinez, T. Hunter and S. Sudarsanam. The protein kinase complement of the human genome. *Science* (2002)

- [McFarland 2013] Impact of deleterious passenger mutations on cancer progression. (2013). Impact of deleterious passenger mutations on cancer progression., *110*(8), 2910–2915.
- [Mills 2001] Gordon B. Mills, Yiling Lu, and Elise C. Kohn. Linking molecular therapeutics to molecular diagnostics: Inhibition of the FRAP/RAFT/TOR component of the PI3K pathway preferentially blocks PTEN mutant cells in vitro and in vivo. *Proceedings of the National Academy of Sciences* (2001) vol. 98 (18) pp. 10031-10033
- [Mocellin 2010] Simone Mocellin, Jeff Shrager, Richard Scolyer, Sandro Pasquali, Daunia Verdi, Francesco M. Marincola, Marta Briarava, Randy Gobbel, Carlo Rossi, Donato Nitti. Targeted Therapy Database (TTD): A Model to Match Patient's Molecular Profile with Current Knowledge on Cancer Biology. *PloS one* (2010) vol. 5 (8) pp. e11965
- [Mooij 2010] Joris M. Mooij. libDAI: A Free and Open Source C++ Library for Discrete Approximate Inference in Graphical Models. *Journal of Machine Learning Research* (2010) vol. 11 pp. 2169-2173
- [Mukherjee 2010] Mukherjee, S . THE EMPEROR OF ALL MALADIES: A BIOGRAPHY OF CANCER, New York: Scribners, 2010.
- [NatBio 2000] Cancer multidrug resistance. Editors.(2000). Cancer multidrug resistance. *Nature Biotechnology*, *18*(Supp), IT18–IT20.
- [NCCN 2013] National Cancer Care Network Breast Cancer Treatment Guidelines Ver 2.2013. http://www.nccn.org/professionals/physician_gls/pdf/breast.pdf
- [NCI]: <http://www.cancer.gov/cancertopics/factsheet/Therapy/targeted> downloaded and verified 19 March 2011
- [Ng 2012] Ng, S., Collisson, E. A., Sokolov, A., Goldstein, T., Gonzalez-Perez, A., Lopez-Bigas, N., et al. (2012). PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. *Bioinformatics (Oxford, England)*, *28*(18), i640–i646.
- [Nicholson 2001] José María, García Buena, Alberto Ocaña, Paola Castro-García, Carmen Gil Gas, Francisco Sánchez-Sánchez, Enrique Poblet, Rosario Serrano, Raúl Calero, Carmen Ramírez-Castillejo. EGFR and cancer prognosis. *European Journal of Cancer* (2001) vol. 37 pp. 9-15
- [Nowak 2006] Martin A. Nowak, Franziska Michor, and Yoh Iwasac. Genetic instability and clonal expansion. *Journal of theoretical biology* (2006) vol. 241 (1) pp. 26-32

- [Nowell 1976] PC Nowell. The clonal evolution of tumor cell populations. *Science* (1976) vol. 194 (4260) pp. 23
- [Pan 2005] Shuang Pan, Qing Dong, Li-Sha Sun and Tie-Jun Li. Mechanisms of Inactivation of PTCH1 Gene in Nevroid Basal Cell Carcinoma Syndrome: Modification of the Two-Hit Hypothesis. *Clinical Cancer Research* (2010) vol. 16 (2) pp. 442-450
- [Parsons 2008] Laura D. Wood, D. Williams Parsons, Siân Jones, Jimmy Lin, Tobias Sjöblom, Rebecca J. Leary, Don Shen, Simina M. Boca, Thomas Barber, Janine Ptak, Natalie Silliman, Steve Szabo, Zoltan Dezso, Vadim Ustyanksky, Tatiana Nikolskaya, Yuri Nikolsky, Rachel Karchin, Paul A. Wilson, Joshua S. Kaminker, Zemin Zhan, Randal Croshaw, Joseph Willis, Dawn Dawson, Michail Shipitsin, James K. V. Willson, Saraswati Sukumar, Kornelia Polyak, Ben Ho Park, Charit L. Pethiyagoda, P. V. Krishna Pant, Dennis G. Ballinger, Andrew B. Sparks, James Hartigan, Douglas R. Smith, Erick Suh, Nickolas Papadopoulos, Phillip Buckhaults, Sanford D. Markowitz, Giovanni Parmigiani, Kenneth W. Kinzler, Victor E. Velculescu, Bert Vogelstein. Genomic Landscapes of Human Breast and Colorectal Cancers. (2008)
- [PathGuide 2011] <http://www.pathguide.org/> downloaded and verified 20 March 2011
- [Pearl 1982] Judea Pearl. Reverend Bayes on inference engines: A distributed hierarchical approach. *Proceedings of the AAAI National Conference on AI* (1982) pp. 133-136
- [Pearson 1900] Pearson, K. “On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling.” via web.
- [Perl 2011] Alexander Perl and Martin Carroll. BCR-ABL kinase is dead; long live the CML stem cell. *J. Clin. Invest.* (2011) vol. 121 (1) pp. 22-25
- [Peters 2002] G.J Peters, H.H.J Backus, S Freemantle, B van Triesta, G Codacci-Pisanelli, C.L van der Wilt, K Smida, J Lune, A.H Calvert, S Marsh, H.L McLeod, E Bloemena, S Meijer, G Jansen, C.J van Groeningen and H.M Pinedo. Induction of thymidylate synthase as a 5-fluorouracil resistance mechanism. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* (2002) vol. 1587 (2-3) pp. 194-205
- [Pettitt 2001] Pettitt, A. R. (2001). p53 dysfunction in B-cell chronic lymphocytic leukemia: inactivation of ATM as an alternative to TP53 mutation. *Blood*, 98(3), 814–822.

- [PHRMA 2011] <http://newmeds.phrma.org/> downloaded and verified 19 March 2011
- [Piccard-Gebhard 2005] Piccart-Gebhart, M. J., Procter, M., Leyland-Jones, B., Goldhirsch, A., Untch, M., Smith, I., et al. (2005). Trastuzumab after Adjuvant Chemotherapy in HER2-Positive Breast Cancer. *New England Journal of Medicine*, 353(16), 1659–1672. .
- [Pierotti 2011] Pierotti, M. A., Tamborini, E., Negri, T., Pricl, S., & Pilotti, S. (2011). Targeted therapy in GIST: in silico modeling for prediction of resistance. *Nature Reviews Clinical Oncology*, 8(3), 161–170.
- [Pullarkat 2001] ST Pullarkat, J Stoehlmacher, V Ghaderi, Y-P Xiong, SA Ingles, A Sherrod, R Warren, D Tsao-Wei, S Groshen, and H-J Lenz. Thymidylate synthase gene polymorphism determines response and toxicity of 5-FU chemotherapy. *Pharmacogenomics journal* (2001) vol. 1 (1) pp. 65-70
- [Robertson 2000] Keith D. Robertson and Peter A. Jones. DNA methylation: past, present and future directions. *Carcinogenesis* (2000) vol. 21 (3) pp. 461
- [Rollinson 1982] Rolinson, G. N. (1982). The history and background of Augmentin. *South African medical journal*, 62(5 Spec No), 3A–4A.
- [Saito 1997] Saito, Y., Milross, C. G., Hittelman, W. N., Li, D., Jibu, T., Peters, L. J., & Milas, L. (1997). Effect of radiation and paclitaxel on p53 expression in murine tumors sensitive or resistant to apoptosis induction. *International journal of radiation oncology, biology, physics*, 38(3), 623.
- [Sakharkar 2002] Meena Kishore Sakharkar, Vincent T.K. Chow, Pandjassame Kanguane. Distributions of exons and introns in the human genome. *In Silico Biol.* 4, 387–393 (2004).
- [Salk 2010] Salk, J. J., Fox, E. J., & Loeb, L. A. (2010). Mutational Heterogeneity in Human Cancers: Origin and Consequences. *Annual Review of Pathology: Mechanisms of Disease*, 5(1), 51–75.
- [Saltz 2008] Saltz, L. B., Clarke, S., Díaz-Rubio, E., Scheithauer, W., Figer, A., Wong, R., et al. (2008). Bevacizumab in Combination With Oxaliplatin-Based Chemotherapy As First-Line Therapy in Metastatic Colorectal Cancer: A Randomized Phase III Study. *Journal of Clinical Oncology*, 28(12), 2013–2019.
- [Scannell 2012] Diagnosing the decline in pharmaceutical R&D efficiency. (2012). Jack W. Scannell, Alex Blanckley, Helen Boldon & Brian Warrington 11(3), 191–200.

- [Schaffter 2011] Schaffter, T., Marbach, D., & Floreano, D. (2011). GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16), 2263–2270.
- [Schoenewolf 2011] Schoenewolf, N. L., Urosevic-Maiwald, M., & Dummer, R. (2011). Tumour heterogeneity of mucosal melanomas during treatment with imatinib. *The British journal of dermatology*, 165(2), 419–424.
- [Schwab 2011] Schwab, M. (2011). *Encyclopedia of Cancer*. Springer Verlag.2
- [Schwender 2003] Schwender, Holger; Krause, Andreas; Ickstadt, Katja “Comparison of the empirical bayes and the significance analysis of microarrays”, Universität Dortmund, SFB 475 Komplexitätsreduktion in Multivariaten Datenstrukturen, Technical Report No. 2003,44.
- [Shah 2009] Nigam H Shah, Clement Jonquet, Annie P Chiang, Atul J Butte, Rong Chen and Mark A Musen. Ontology-driven indexing of public datasets for translational bioinformatics. *BMC Bioinformatics* 2009 10:S1 (2009) vol. 10 (Suppl 2) pp. S1
- [Sheridan 20089] Clare Sheridan, Gabriela Brumatti, and Seamus J. Martin. Oncogenic B-RafV600E Inhibits Apoptosis and Promotes ERK-dependent Inactivation of Bad and Bim. *Journal of Biological Chemistry* (2008) vol. 283 (32) pp. 22128-22135
- [Shrager 2003] Shrager, J. (2003). The fiction of function. *Bioinformatics*. Vol. 19 no. 15 2003, pages 1934–1936 DOI: 10.1093/bioinformatics/btg277
- [Sierra 2010] Sierra, J. R., Cepero, V., & Giordano, S. (2010). Molecular mechanisms of acquired resistance to tyrosine kinase targeted therapy. *Molecular Cancer*, 9(1), 75.
- [Sirota 2011] Sirota, Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci Transl Med*. August 2011
- [Smith 2010] Smith, H. J., & Williams, H. (2010). Smith and Williams' Introduction to the Principles of Drug Design and Action, CRC Press page 423.
- [Stark 1984] George R. Stark and Geoffrey M. Wahl. Gene amplification. *Annual review of biochemistry* (1984) vol. 53 (1) pp. 447-491
- [Storchova 2004] Zuzana Storchova and David Pellman. From polyploidy to aneuploidy, genome instability and cancer. *Nat Rev Mol Cell Biol* (2004) vol. 5 (1) pp. 45-54.

- [Storey 2003] John D. Storey and Robert Tibshirani. Statistical methods for identifying differentially expressed genes in DNA microarrays. *Methods Mol Biol*, 224:149–57.
- [Stuart 2013] Personal communication Professor Josh Stuart, group leader of the PANCAN effort.
- [Talpaz 2002] Moshe Talpaz, Richard T. Silver, Brian J. Druker, John M. Goldman, Carlo Gambacorti-Passerini, Francois Guilhot, Charles A. Schiffer, Thomas Fischer, Michael W. N. Deininger, Anne L. Lennard, Andreas Hochhaus, Oliver G. Ottmann, Alois Gratwohl, Michele Baccarani, Richard Stone, Sante Tura, Francois-Xavier Mahon, Sofia Fernandes-Reese, Insa Gathmann, Renaud Capdeville, Hagop M. Kantarjian, and Charles L. Sawyers. Imatinib induces durable hematologic and cytogenetic responses in patients with accelerated phase chronic myeloid leukemia: results of a phase 2 study. *Blood* (2002) vol. 99 (6) pp. 1928-1937.
- [Tamayo 1999] Pablo Tamayo, Donna Slonim, Jill Mesirov, Qing Zhu, Sutsak Kitareewan, Ethan Dmitrovsky, Eric S. Lander, f and Todd R. Golub. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences of the United States of America* (1999) vol. 96 (6) pp. 2907.
- [TCGA Colorectal Cancer 2012] Comprehensive Molecular Characterization of Human Colon and Rectal Cancer. (TCGA Marker Paper). *Nature* July 2012
- [TCGA Breast Cancer 2012] *Comprehensive molecular portraits of human breast tumours*, *Nature* October 2012
- [TCGA Uterine and Endometrial Cancer 2013] published in *Integrated Genomic Characterization of Endometrial Carcinoma*, *Nature*, May 2013.
- [TCGA Lung Adenocarcinoma Cancer] submitted to *Nature*.
- [TCGA Clear Cell Kidney Renal Cancer] submitted to *Nature*.
- [TCGA Ovarian 2011] Network TCGA. Integrated genomic analyses of ovarian carcinoma. *Nature* 2011; 474: 609-615.
- [Tian 2009] Tian, Z., Hwang, T., & Kuang, R. (2009). A hypergraph-based learning algorithm for classifying gene expression and arrayCGH data with prior knowledge. *Bioinformatics (Oxford, England)*, 25(21), 2831–2838.

- [Tibshirani 2006] Tibshirani. A simple method for assessing sample sizes in microarray experiments. *BMC bioinformatics* (2006) vol. 7 (1) pp. 106.
- [Toulany 2007] Toulany, M., Baumann, M., & Rodemann, H. P. (2007). Stimulated PI3K-AKT Signaling Mediated through Ligand or Radiation-Induced EGFR Depends Indirectly, but not Directly, on Constitutive K-Ras Activity. *Molecular Cancer Research*, 5(8), 863–872.
- [True 2006] LD True and KR Loeb and JH Biela. Human cancers express a mutator phenotype. *PNAS* (2006).
- [Tusher 2001] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* 2001; 98: 5116-5121.
- [Ugander 2011] Ugander, J., Karrer, B., Backstrom, L., & Marlow, C. (2011). The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*.
- [Vandin 2011] Vandin F, Upfal E, Raphael BJ. Algorithms for detecting significantly mutated pathways in cancer. *Journal of computational biology : a journal of computational molecular cell biology* 2011; 18: 507-522.
- [Varadan 2012] Varadan, V., Mittal, P., Vaske, C., & Benz, S. (2012). The Integration of Biological Pathway Knowledge in Cancer Genomics: A review of existing computational approaches. *IEEE Signal Processing Magazine*, 29(1), 35–50.
- [Vaske 2010] Charles J. Vaske, Stephen C. Benz, J. Zachary Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, David Haussler, and Joshua M. Stuart. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* (2010) vol. 26 (12) pp. i237.
- [Watters 2003] James W. Watters, Howard L. McLeod. Cancer pharmacogenomics: current and future applications. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* (2003) vol. 1603 (2) pp. 99-111.
- [Weiss 1961] Weiss, Lionel, Statistical Decision Theory, McGraw-Hill Book Company, Inc., New York, (1961).
- [Winder 2010] Winder, T., & Lenz, H.-J. (2010). Molecular predictive and prognostic markers in colon cancer. *Cancer Treatment Reviews*, 36(7), 550–556. doi:10.1016/j.ctrv.2010.03.005.

- [Witten 2005] Ian H. Witten, Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes, and Sally Jo Cunningham. *Weka: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann (2005).
- [Yap 2010] Yap, T. A., Sandhu, S. K., Workman, P., & de Bono, J. S. (2010). Envisioning the future of early anticancer drug development. *Nature Reviews Cancer*, 10(7), 514–523. doi:10.1038/nrc2870.
- [Ye 2007] C.J. Ye, G. Liu, S.W. Bremer, and H.H.Q. Heng, The dynamics of cancer chromosomes and genomes. *Cytogenet Genome Res* (2007) vol. 118 (2-4) pp. 237-246.
- [Yu 2007] Zhiwei Yu, Titus J. Boggon, Susumu Kobayashi, Cheng Jin, Patrick C. Ma, Afshin Dowlati, Jeffrey A. Kern, Daniel G. Tenen, and Balázs Halmos. Resistance to an Irreversible Epidermal Growth Factor Receptor (EGFR) Inhibitor in EGFR-Mutant Lung Cancer Reveals Novel Treatment Strategies. *Cancer Research* (2007) vol. 67 (21) pp. 10417-10427.
- [Ziogas 2009] Ziogas and Roukos. Genetics and Personal Genomics for Personalized Breast Cancer Surgery: Progress and Challenges in Research and Clinical Practice. *Ann Surg Oncol* (2009) vol. 16 (7) pp. 1771-1782.