2013

Development of the "carcinogenome": genomic signatures of carcinogenicity

BOSTON UNIVERSITY

SCHOOL OF MEDICINE

Thesis

DEVELOPMENT OF THE "CARCINOGENOME": GENOMIC SIGNATURES OF CARCINOGENICITY

by

JESSALYN M. UBELLACKER

M.P.H., University of Kentucky- Lexington, 2011

B.S., University of Kentucky- Lexington, 2010

Submitted in partial fulfillment of the

requirements for the degree of

Master of Arts

2013

Approved by

First Reader _____

David Sherr, Ph.D.
Professor, Department of Environmental Health

Second Reader _____

Gwynneth D. Offner, Ph.D.
Director M.A. Medical Sciences Program
Associate Professor of Medicine

ACKNOWLEDGEMENTS

I wish to thank Dr. David Sherr (Boston University) for serving as an outstanding research mentor and as an inspiration for my further pursuit of a research career. Thanks to Dr. Scott Auerbach (National Toxicology Program at NIEHS) for providing the majority of the carcinogens and non-carcinogens used in the pilot experiments. Thanks to Dr. Stefano Monti (Boston University) for the bioinformatics analysis and interpretation. Dr. Xiaodong Lu, Dr. Aravind Subramanian, and Dr. Rajiv Narayan (The Broad Institute) provided significant support in cell culture preparations for the experiments, and assistance in running the prepared plates on the Broad's L1000 machine. Additional thanks to Abbie Kew (Boston University Core Facilities) for assistance in usage of the core instruments to carry out the pilot experiments.

I appreciate the supportive and entertaining lab environment made complete by Dr. Jennifer Schlezinger, Ashley Parks, Faye Andrews, Olga Novikov, Elizabeth Stanford and James Watt. I'm grateful to have worked with you as colleagues and as friends.

Funding for this project was generously provided by a Boston University School of Public Health pilot project grant.

DEVELOPMENT OF THE "CARCINOGENOME":

GENOMIC SIGNATURES OF CARCINOGENICITY

JESSALYN M. UBELLACKER

Boston University School of Medicine, 2012

Major Professor: David Sherr, Ph.D., Professor, Department of Environmental Health

ABSTRACT

This project presents a new method of developing genomic signatures of carcinogenicity and offers a way of understanding the mechanisms of action for environmental chemicals, and a method for characterizing the carcinogenicity of said chemicals. Specifically, this project is an effort to "use a high-dimensional assay that allows for the genome-wide characterization of the transcriptional response to chemical exposure at a cost unmatched by other technologies, paired with the biological and computational expertise necessary to maximally leverage the information produced by the assay."[11] The primary aim of these preliminary experiments was to determine if the protocol of treating the cells at Boston University under strict IBC-approved conditions and then gathering data at the Broad Institute (Boston, Massachusetts) is a reasonable technique for obtaining high-throughput genomic signatures with chemicals of unknown carcinogenicity status. The results of this experiment demonstrate this methodology for obtaining genomic signatures has the potential to be a replicable, high-throughput procedure to assist in creating a "carcinogenome" database (CGDB).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AhR	Aryl Hydrocarbon Receptor
BMD	Benchmark Dose
BMDL	Benchmark Dose (Lower Confidence Limit)
BSL-2	Biosafety Level 2
BU	Boston University
cDNA	Complementary Deoxyribonucleic acid
CGBD	Carcinogenome Database
CMap	Connectivity Map
CO_2	Carbon Dioxide
DM	Drug Matrix
DMSO	Dimethyl Sulfoxide
EPA	Environmental Protection Agency
FBS	Fetal Bovine Serum
FICZ	Fluorescein isothiocyanate
FTP	File Transfer Protocol
GEO	Gene Expression Omnibus
IARC	International Agency for Research on Cancer
iPSC	inducible pluripotent stem cell
L1000	Landmark 1,000
LINCS	Large Scale Expression Analysis of Cellular States
LOEL	Lowest Observable Effect Level

mL	milliliter
mM	milimolar
mRNA	messenger ribonucleic acid
uL	microliter
MoA	Mechanism of Action
MIT	Massachusetts Institute of Technology
NIEHS	National Institute of Environmental Health Sciences
NIH	National Institute of Health
NOEL	No observed effect level
NTP	National Toxicology Program
ORF	Open Reading Frame
PBS	Phosphate Buffered Saline
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PoC	Pathway of Carcinogenicity
PPARA	Peroxisome Proliferator-activated receptor, alpha
PXR	Pregnane X Receptor
rpm	Rotations per minute
RPMI	Roswell Park Memorial Institute
shRNA	short hairpin Ribonucleic acid
SoC	Signature of Carcinogenicity

INTRODUCTION

Cancer is a disease which continues to threaten the well-being of individuals across the world. In America alone, 21% of the population will die from cancer.[1] Since genetic factors have been found to contribute only as a minor cause of cancer mortality, much focus has been placed on studying environmental influences on the disease.[1-3] The field of environmental chemicals and cancer research is one in which there is little completed research. The 2010 Presidential Cancer Panel recognizes this field as a potential area for understanding cancer causation by environmental chemicals through the use of high-throughput technologies.[2,4]

The Need for Models of Carcinogenicity

Current research on carcinogen testing involves costly and time-consuming rodent studies of individual chemicals. Of the approximately 80,000 chemicals in commercial usage, only 1,500 have been tested using this standard.[5-7] This model is also limited because it is difficult to determine the result of mixtures of chemicals and carcinogenic effect. Previous studies have demonstrated the carcinogenicity of chemicals can be characterized by understanding chemical-induced gene mutations, chromosomal aberration and aneuploidy.[8-10]

This project presents a new method of developing genomic signatures of carcinogenicity and offers a way of understanding the mechanisms of action for environmental chemicals, as well as a way of characterizing the carcinogenicity of said chemicals. Specifically, this project is an effort to "use a high-dimensional assay that

allows for the genome-wide characterization of the transcriptional response to chemical exposure at a cost unmatched by other technologies, paired with the biological and computational expertise necessary to maximally leverage the information produced by the assay."[11]

Proposed Method for a Predictive Model of Carcinogenicity

The ability to create a carcinogenome requires a high-throughput technology such as one developed by the Broad Institute (a collaborative research institution between MIT and Harvard University, Boston, Massachusetts) through the NIH Common Fund's LINCS program. This technology, the "L1000", uses correlations between data obtained with a limited set of targeted genes and the more traditional microarray technology to characterize over 20,000 genes in the human transcriptome into 1,000 gene clusters represented by "landmark" genes.[15] A landmark gene is representative of a specific gene cluster, and 1,000 landmark gene transcripts on Luminex beads (which make up the foundation of the L1000 technology) accurately predicts 80% of the gene connections discovered by traditional, more-expensive, and less high-throughput means (Figure 1). The 1,000 landmark genes were selected on the basis of principal component analyses (PCA) of data from the Gene Expression Omnibus (GEO).[23]

The Broad Institute anticipates that a reduced representation transcription paired with a computational inference model is able to accurately create a genome-wide expression profile. With this technology, the Broad anticipates working toward the creation of "CMaps," or "Connectivity Maps" on a high-throughput scale to develop a

primary screening library of "perturbagens" complete with numerous different cell lines and treatment parameters.[21] The three main applications the Broad foresees for this technology are: 1) the creation of a database with which scientists can develop more educated and directed hypotheses, 2) understanding disease or gene-specific signatures and 3) the creation of a chemical library-scale connectivity map in order to understand gene expression as a sensor of bioactivity.[21]

Whereas the traditional microarray, e.g. an Affymetrix chip, costs approximately $500/sample, the L1000 costs approximately $5/sample and can be run on 384 well plates (with a different chemical in each well) and thereby is able to provide "80% of the information at 1% of the cost". The Broad Institute has the technological capacity to run over 100 plates per week. As of October 2012, the Broad Institute has over 500,000 genomic profiles for 10 different cell types of 4,000 small molecules, 3,000 Open Reading Frames (ORF), and 9,000 shRNAs (3 constructs of each shRNA).

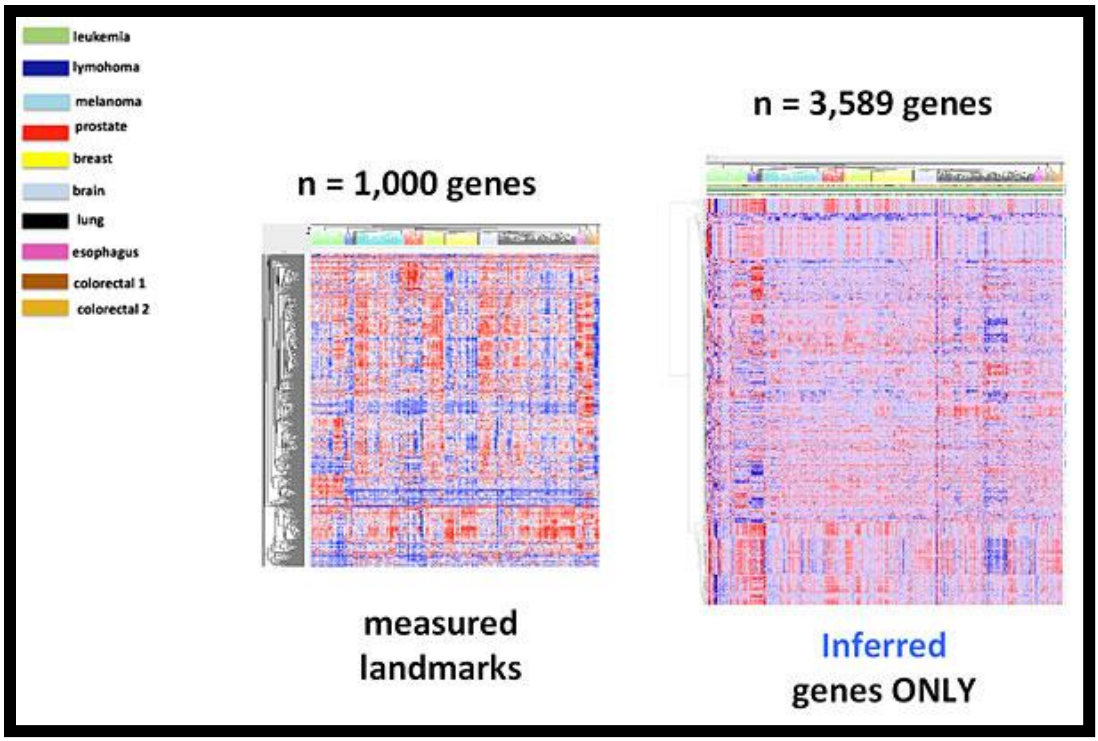


Figure 1. L1000 Computational Inference Potential. Based on the 1,000 Landmark genes, the Broad Institute is able to predict inferred connections. (Image courtesy of Aravind Subramanian, Ph.D.).[23]

Specific Aims

The focus of this research project is to determine the efficacy and efficiency of the development of a predictive genomic signature model for carcinogenicity (a "carcinogenome") which will quantitatively characterize the degree of carcinogenicity of specific environmental pollutants. Specifically, we will develop a computational predictive model of carcinogenicity by analyzing the effect that a specific chemical has on upregulating or downregulating genes. A carcinogenome database (CGDB) could ultimately be used to provide a carcinogen predictive model for any given chemical in the database.

In order to determine the feasibility of this model, we will first study the effects of approximately 100 different chemicals (which will include compounds with a known carcinogenicity status) in collaboration with The Broad Institute (Boston, Massachusetts), on immortalized cell lines. We will test the ability of the expression profile determined from this technology to correctly predict the known carcinogenic status of the chemicals being tested in the pilot experiment by determining the efficacy, sensitivity, and specificity.

Because the Broad Institute does not have the institutional capacity to treat with carcinogenic compounds, the Sherr Laboratory, a BSL2 laboratory with standard operating procedures for handling multiple carcinogens at once, will prepare and dose the cells at Boston University using high through-put robotics. This collaborative effort will be important to insure the reliability of the L1000 in representing outcomes of the entire genome and to demonstrate that the chemical carcinogenicity can be predicted by computational analyses.

For this initial project specific aims will be:

1.) To run quality control checks with the Boston University Core facility instruments to insure a minimal level of error in plating the cells, dosing the compounds, and washing the carcinogen-containing media from the plates

2.) To prepare cell lysate for genomic signatures at Boston University:
    a. for 52 carcinogen and 51 non-carcinogen compounds
    b. in four different cell lines
    c. at two different dosing timepoints (6 hours and 24 hours)
3.) To determine if the protocol for treating the cells at BU and then gathering data at the Broad Institute is a reasonable technique for obtaining high-throughput genomic signatures with chemicals of unknown carcinogenicity.

METHODS

Introduction

Protocols for the pilot experiments were obtained from Xiaodong Lu, Ph.D., at the Broad Institute. Cell culture and treatment work was completed at the Boston University High Throughput Core Facility (Boston, Massachusetts) and L1000 plate reading was completed at the Broad Institute (Boston, Massachusetts) July 2011-February 2013.

Compound Selection

Compounds to be included in this study were prioritized in collaboration with the Environmental Protection Agency (EPA), the National Toxicology Program (NTP), and the National Institute of Environmental Health Sciences (NIEHS), specifically in collaboration with Raymond Tice, Ph.D. and Scott Auerbach, Ph.D. These compounds were drawn from ToxCast (www.epa.gov/ncct/toxcast/chemicals.html) and the International Agency for Research on Cancer (IARC, http://monographs.iarc.fr/ENG/Classification) databases and were selected based upon known carcinogenicity status, mode of action, and degree of presence in the environment.

Table 1. Carcinogenic and non-carcinogenic compounds on pilot plate

Compound Classification	***Number of Compounds***
Non-Carcinogen	51
Carcinogen	52
TOTAL	103

Of the chemicals included in the preliminary compound plate, 51 compounds are known to be non-carcinogenic and 52 chemicals are known to be carcinogenic (Table 1). Each chemical is characterized based upon it's mode of action as a direct genotoxic agent (such as an alkylating agent), an indirect genotoxic agent (such as topoisomerase inhibitors), or as a non-genotoxic/receptor mediated chemical (such as AhR-, PXR-, PPARα-, and oxidative stress-inducing compounds). Compounds which are negative and positive by the Ames test are also included as another means of comparison in carcinogenicity classification.[9,12] Additionally, because the Sherr laboratory has knowledge and experience with Aryl Hydrocarbon Receptor (AhR) ligands, these chemicals are included in the compound listing.[13,14] The AhR plays an important role in chemically-induced cancers. The complete listing of compounds included in this pilot experiment is found in Appendix 1.

Optimal dosage for each compound was determined by published *in vitro* toxicity data, and by preliminary toxicity testing performed in the Sherr Laboratory. Compounds were included in randomized triplicate on the 384 well-plate.

Cell line selection

Cell lines were selected based on lines the Broad Institute has successfully collected genomic signatures of cells exposures to a library of drugs and other compounds. Additionally, cell lines were selected based on applicability to environmental chemical-induced cancer from previous studies. The four cell lines used in this pilot experiment include: MCF7 (human epithelial breast cancer cell line from mammary

gland adenocarcinoma), HA1E (human embryonic kidney cell line), HMELz (human breast epithelial cell line), and PC3 (human epithelial prostate cell line from a grade IV adenocarcinoma). Cells were generously provided by Xiaodong Lu, Ph.D., from The Broad Institute (Boston, Massachusetts).

PROTOCOL

Initial Quality Control Tests

Quality control tests were completed before the initial pilot experiment to ensure the accuracy and precision of the procedure. These tests included:

1.) **Precision of robotic versus hand pipetting of cells.**

1,000 MCF7 cells were plated into 252 wells of a 384 well-plate using a cell plating robot (ThermoScientific, Multidrop Combi Reagent Dispenser, 5840300) or by hand using an electronic, multi-channel pipet (ThermoScientific, FinnPipette Novus Multichannel Pipetter, 14-387-115). Cell counts and distribution were compared using an image-based microplate reader (Cyntellect, Celigo Image-Based Microplate Reader, 77816883)

2.) **Precision of core high-throughput dosing robot (Apricot 550 Liquid Handling System)**.

Fluorescein isothiocyanate (FITC) was plated as a fluorescent dye to confirm even distribution on the compound plate, and was used to dose 2,000 MCF7 cells in five 384 well-plates. The final concentration in the dosed cell plate was determined with a plate reader (Tecan, SpectraFluor Plus) and even distribution of dosing was monitored via analysis of FITC fluorescence in each well.

3.) **PBS wash cell viability test.**

A plate washer (Bio-Tek ELx50 Auto Strip) was used to aspirate media from the wells dosed with carcinogens and the plates were washed three times with PBS before harvesting the RNA from the wells. PBS was added by electronic, multi-channel pipet ThermoScientific, FinnPipette Novus Multichannel Pipetter, 14-387-115). Cells were counted with Cell Titer Glo (Promega, PR-G7570) both before and after the PBS washes to determine viability differences due to the washes.

Cell Preparation

Cells were trypsinized in cell culture dishes prior to being quenched using FBS-containing media. Trypsinized cells were centrifuged at 1,000 rpm for four minutes, the medium was aspirated and 5 mL of fresh media was dispensed onto the cell pellet. 19 milliliters of media was added to the cell pellet mixture and mixed to create a uniform suspension. The resuspended cells were counted and viability was quantified with trypan blue. For the cell plating in the 384-well plate, 3,500 cells were plated per well in 45uL of medium using the Combi Reagent Dispenser (ThermoScientific, Multidrop Combi Reagent Dispenser, 5840300). Cells were maintained at room temperature for 20 minutes, and then placed in stacks of 5 plates in the incubator at 37 degrees Celsius and 5% CO_2. The plates were incubated for 24 hours.

Cell Treatment

Next, the compound plate with carcinogens and non-carcinogens was prepared. Compounds were obtained from the National Toxicology Program (NTP) or from Sigma-Aldrich. Compounds were prepared at 10mM, and 10uL of each compound was placed in triplicate on the plate. DMSO was used as a control for the compounds, as compounds were dissolved in this solvent. The compound plate was covered with an aluminum plate sealer (Bio-Rad, 181-4040) and centrifuged before use. At the time of cell dosing, the compound plate and the plated cell plate were placed on the Apricot 550 Liquid Handling System along with a dilution plate which contained 99uL of 10% RPMI media. The robot transferred 1uL of compound to 99uL of media and mixed the compound with the media three times to create a uniform solution. The dilution plate was then sealed with a foil plate seal (Bio-Rad, 181-4040) and spun down in a centrifuge for one minute at 1,000rpm. Next, the robot transferred 5uL of the compound diluted in the RPMI media onto the plated cells. Finally, the plates were placed in the incubator and sets were harvested at 6 and 24 hour timepoints after the initial dosage.

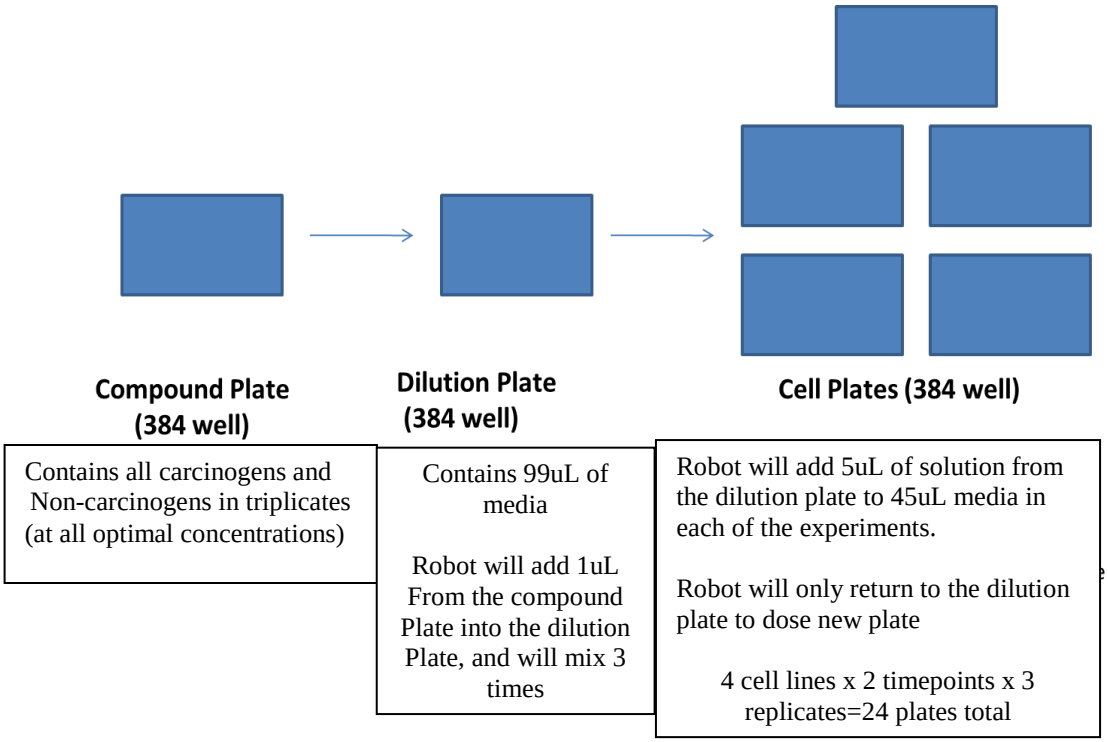


Figure 2. Protocol for cell dosing. Robot will pick up compounds from compound plate, and then transfer to the dilution plate. Next, the dilution/compound mixture will be added to 384-well plates with cells.

Cell Lysate Harvest

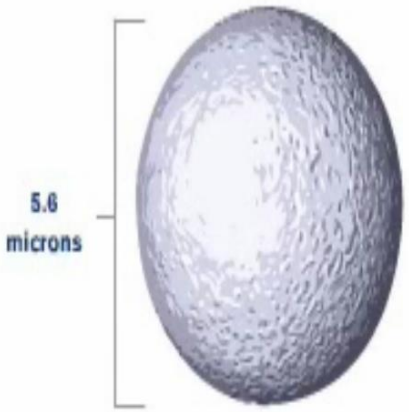
For the harvest, media was removed by a plate washing robot (Bio-Tek Instruments, ELx50 Auto Strip washer) and washed with PBS (Mediatech, MT-20-030-CV) three times. Next, 35uL of TCL buffer (Qiagen, cat#: 72271) was added and cell lysis plates incubated at room temperature for 30 minutes, then stored in the -80^{o} freezer.

Cell Titer Glo was used to check viability of cells at the 24 hour timepoint on a parallel test plate. 10uL of Cell Titer Glo was added to each well, and read on a plate reader after 15 minutes.

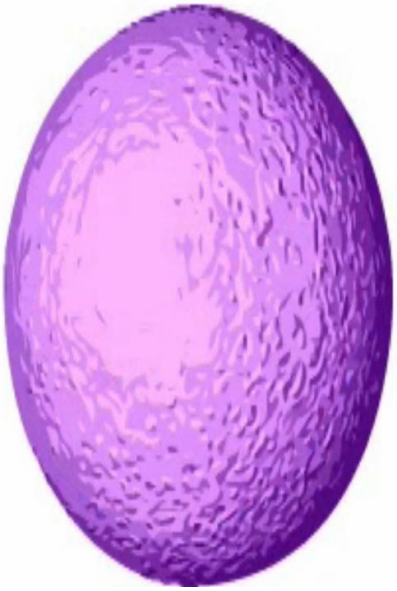
Cell lysate plates were transferred to the Broad Institute. At the Broad Institute, "poly-A mRNAs [were] captured from the lysates on oligo-dT-coated plates, and the 1,000-plex ligation-mediated amplification procedure [was] performed. The samples were detected on Luminex FlexMap-3D instruments and 2 x 500-color Luminex beads, using two different bead concentrations as a means to obtain 1,000 analytes from 500 bead colors."[11,16]

This ligation-mediated amplification process uses the mRNA transcripts from the lysates and amplifies the gene transcripts of interest using probe pairs to detect gene transcript levels by the Luminex Flexmap system (Figure 3).[24] This technology uses microspheres which have different dyes and fluorophores for each genomic transcript of interest.[24]

The Luminex software program detects each of the individual microspheres bound to a specific sample.After standardized quality control measures, the data were transferred to Boston University via file transfer protocol (FTP).

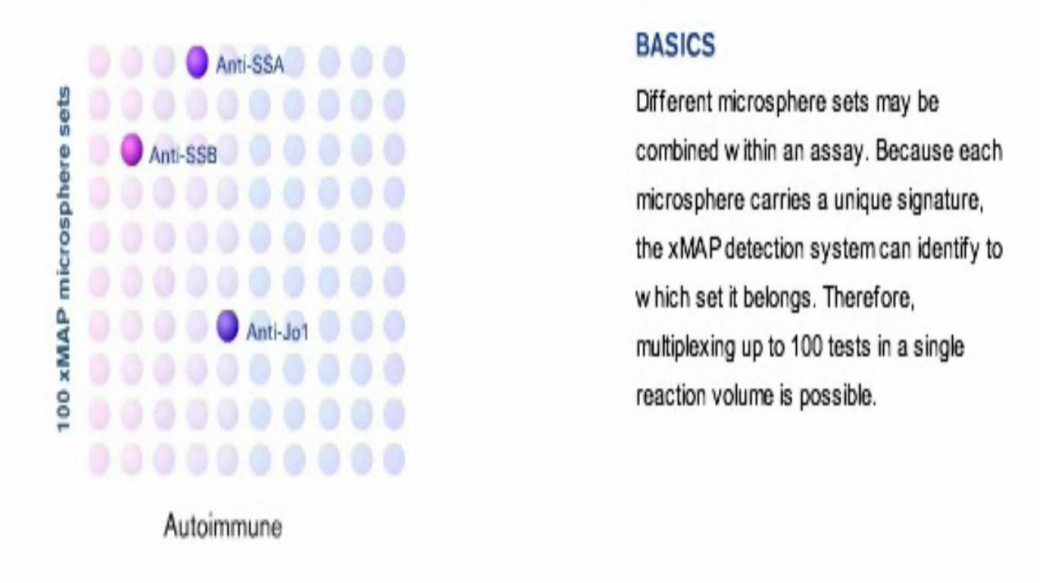


The xMAP technology uses 5.6 micron polystyrene microspheres which are internally dyed with red and infrared fluorophores.



BASICS

Using different intensities of the two dyes for different batches of microspheres, we have created 100 xMAP microsphere sets, each with a unique spectral signature determined by its red/infrared mixture.



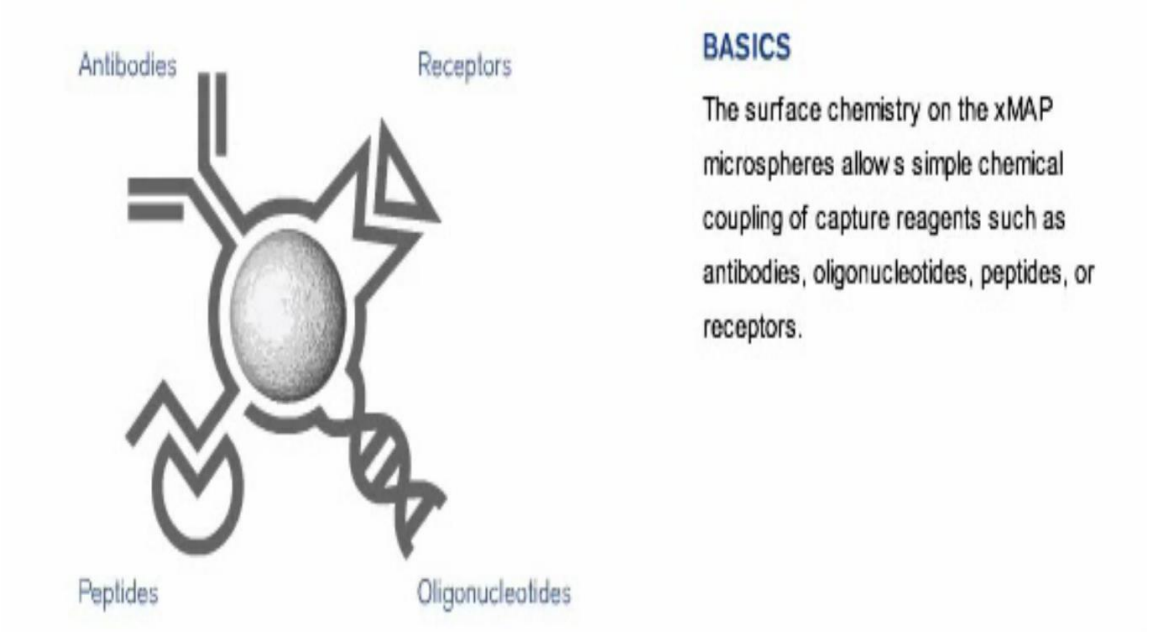


Figure 3. xMAP Technology as described by Luminex[24]

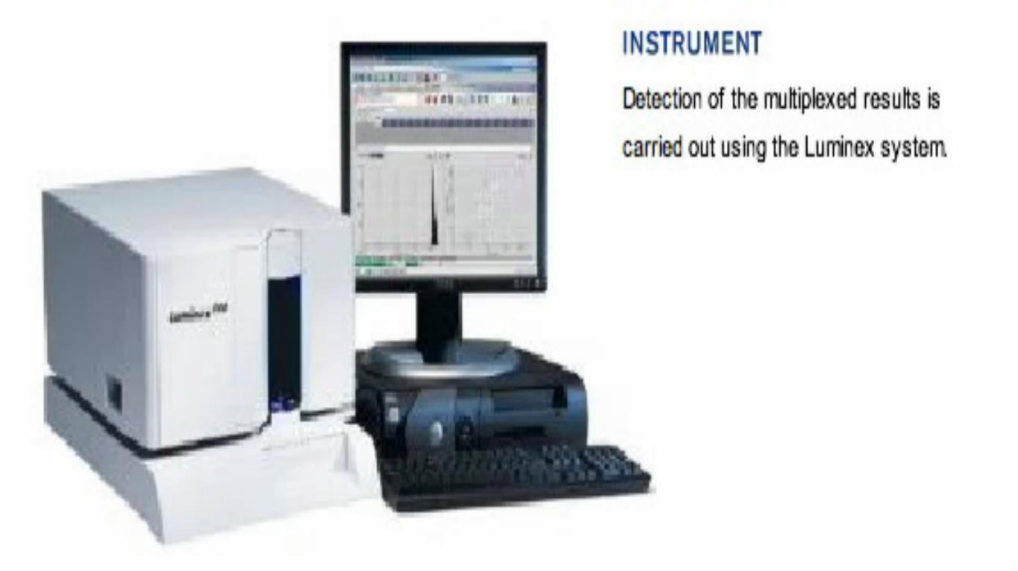
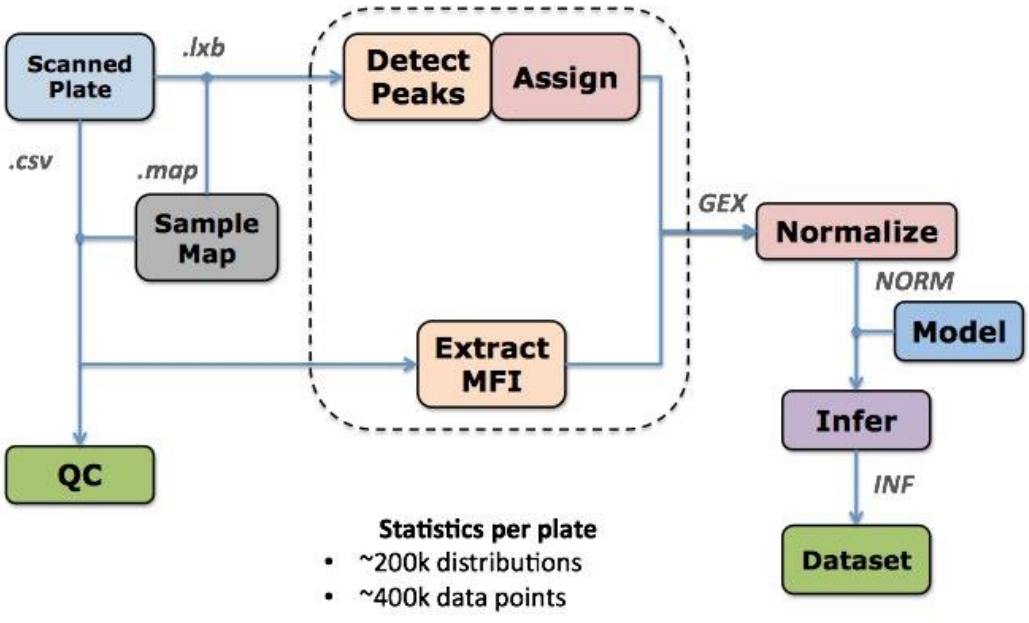


Figure 3. xMAP Technology as described by Luminex[24]



plate[22]

Figure 4. Flow diagram of Broad Institute dataset generation from scanned lysate

Data interpretation

Stefano Monti, Ph.D. (Boston University) analyzed the dataset generated by the Broad Institute with a *[20,000 x N]* data matrix using computational inference models in addition to classification algorithms (*Weighted voting, k-nearest neighbors, Support Vector Machines*, etc).[11,17,18,20] Biological interpretation was based on the Cancer Cell Line Encyclopedia (http://www.broadinstitute.org/ccle) and the Connectivity Map (www.broadinstitute.org/cmap).

RESULTS

Initial Quality Control Test Results

1.) Precision of Robot-assisted versus hand-pipeting of cells.

When the cell plating robot results are compared to that of the electronic multi-channel plating results, the average number of cells plated manually (482.04 cells/well) was significantly less than that plated by the robot (746.33 cells/well), (Table 2). The robot approached the actual counted and calculated number of cells per well (1,000 cells/well). It was decided to plate the cells for the pilot experiments with the mechanical cell plater because of the higher average number of cells distributed in each well. These values only give an estimate of the number of cells in each well, as the image-based microplate reader (Cyntellect Celigo) used to count the cells also contributes a source of error to these presented values.

Table 2. Robot versus hand-pipetting accuracy for plating cells in 384-well plate

	Hand Pipet (cell counts)	***Robot (cell counts)***
Average	482.04	746.33
Minimum	193	202
Maximum	1043	1660
Standard Deviation	161.15	178.22
Number of wells	252	252
Standard Error	10.15	11.21

2.) Precision of Apricot dosing robot.

The FITC plated on compound plate was added to a dilution plated and used to distribute 2,000 MCF7 cells into 5 different 384 well-plates. The final concentration in the dosed cell plate was determined with a plate reader (Tecan, SpectraFluor Plus) and even distribution of dosing was monitored.

Figure 5 displays a visual representation of the dosing of the 5 different 384 well-plates used for this test. If the fluorescence reading of FITC was greater than 650, the well was conditionally formatted as red. If the fluorescence reading of the well was less than 450, the well was conditionally formatted yellow. As demonstrated by the visual representation of the distribution of the higher and lower ranges of the fluorescence readings, the very first plate dosed had the greatest variation. Because of this, for the pilot experiments, a primary "test" plate was dosed, but discarded to eliminate the variation in the first dosed plate. Outer wells of the 384 well-plate were not used for the pilot experiment due to edge effects and concerns regarding evaporation of media around the corners of the plates.

The variations in dosing for the plates was determined to be randomized, and not associated with specific pins on the Apricot robot, and was within a reasonable range of error (10%) for these initial tests. This percent error threshold was based on Apricot company recommendations for high-throughput experiments.

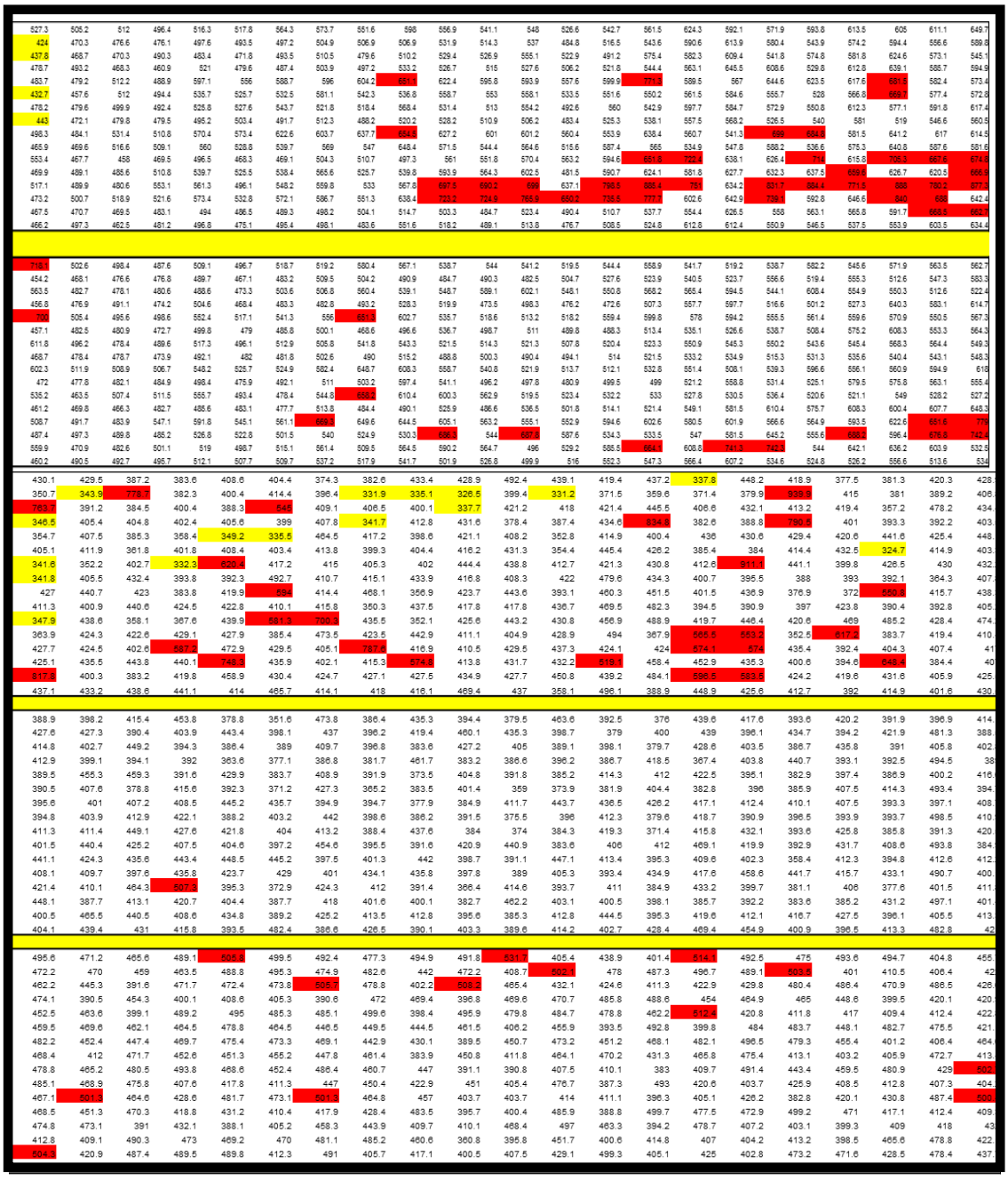


Figure 5. Distribution of even dosing of compounds by Boston University core instruments. Plates 1-5, 384-well plates. Measurement >650=red; cell counts <450=yellow

3.) PBS Wash Viability Tests

The results of the PBS wash cell viability tests demonstrate that PBS washing reduces the number of cells in each well by approximately 25% (Table 3B). This may be due to the mechanical action of the Auto Strip washer (Bio-Tek Instruments, ELx50 Auto Strip washer), and the additional stress of adding PBS to the cells three times. Cell Titer Glo (Promega, PR-G7570) viability tests demonstrate the PBS washes increase variation in cell counts between the different plates (a 7.4% error rate between plates), (Table 3A). This step is necessary because the cell plates must be washed with PBS to rid the plate of the toxic media before transferring the plate to the Broad Institute. Consequently possible error introduced by the additional washing step must be taken into consideration when interpreting the results. In an effort to reduce the error, two PBS wash steps instead of three were used for the initial pilot tests.

Table 3A. PBS washes cell viability Plates 1-5

	PBS Washed Plate 1	***PBS Washed Plate 2***	***PBS Washed Plate 3***	***Non-washed Plate 4***	***Non-washed Plate 5***
Average	2.07	2.62	0.45	3.01	3.18
SD	0.67	0.92	0.17	1.03	1.07
SE	0.03	0.05	0.01	0.05	0.05
Min	0.49	0.66	0.21	0.90	0.91
Max	4.88	7.30	1.44	7.63	7.32

	AVERAGE PBS WASHED PLATES	*AVERAGE NON PBS WASHED PLATES*	*All Plates*
Average	1.71	3.09	2.26
SD	0.58	1.05	0.37
SE	0.03	0.05	0.17
Min	0.21	0.90	0.21
Max	7.30	7.63	7.63

Table 3B. PBS washes versus no PBS washes cell viability comparison table

RESULTS OF INITIAL PILOT EXPERIMENT

General Findings

These preliminary experiments provide data for correlations between the triplicate compounds for plates dosed at the Broad Institute compared to those dosed at Boston University. Additional correlations for comparability were analyzed between replicates for the PBS washed plates versus the non-washed control plates.

In another set of plates, one plate was dosed and harvested as cell lysate, then transferred to the Broad, as indicated in the protocol. A second plate went through the cDNA extraction protocol at BU's core as well, so that the reverse transcriptase step leading up to the PCR reaction was also completed at Boston University. Correlations between these differences in protocol were also compared.

It is evident from these results that the genomic signature strength is not strong enough to record a conclusive reading for the majority of the compounds from these preliminary runs. For both the washed and non-washed plates, those dosed at BU and at those prepared at the Broad Institute, and for the RNA extraction plate compared with the

cDNA prepared plate, approximately 25% of all the compounds had detectable genomic signatures.

In general, there was more correlation between the plates which did not undergo a PBS washing as compared with those that were washed 3 times. There was stronger correlation between the triplicates in cells treated at the Broad Institute as compared with Boston University. Additionally, the correlation was strong in the RNA plates as compared with the cDNA extraction plates, both which were prepared at Boston University.

Boston University-treated versus Broad-treated cell correlations

As indicated in Figures 7 and 8, for the RNA lysates prepared at Boston University, the signature strength is lower than that of the Broad's signature strength. Interestingly, the washed versus non-washed conditions did not make a difference in signature strength. From Figure 8, correlation between replicates (the three randomized wells on the 384-well plate dosed with the same compound) was evident. However, the threshold level was unable to provide a strong genomic signature for the lysed cells for a significant number of the compounds.

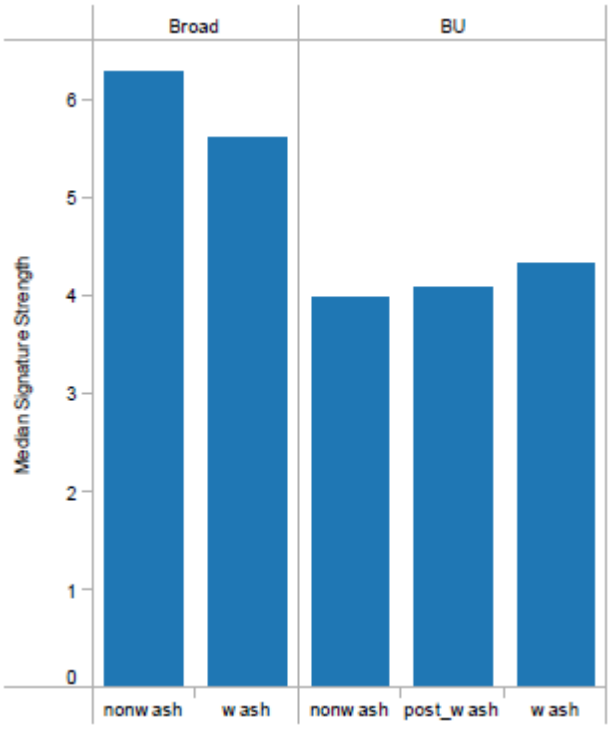


Figure 7. Mean Signature Strength for washed and non-washed compounds at the Broad Institute and Boston University. Comparison of the genomic signature reading levels for plates washed with PBS, and plates which were not washed with PBS.

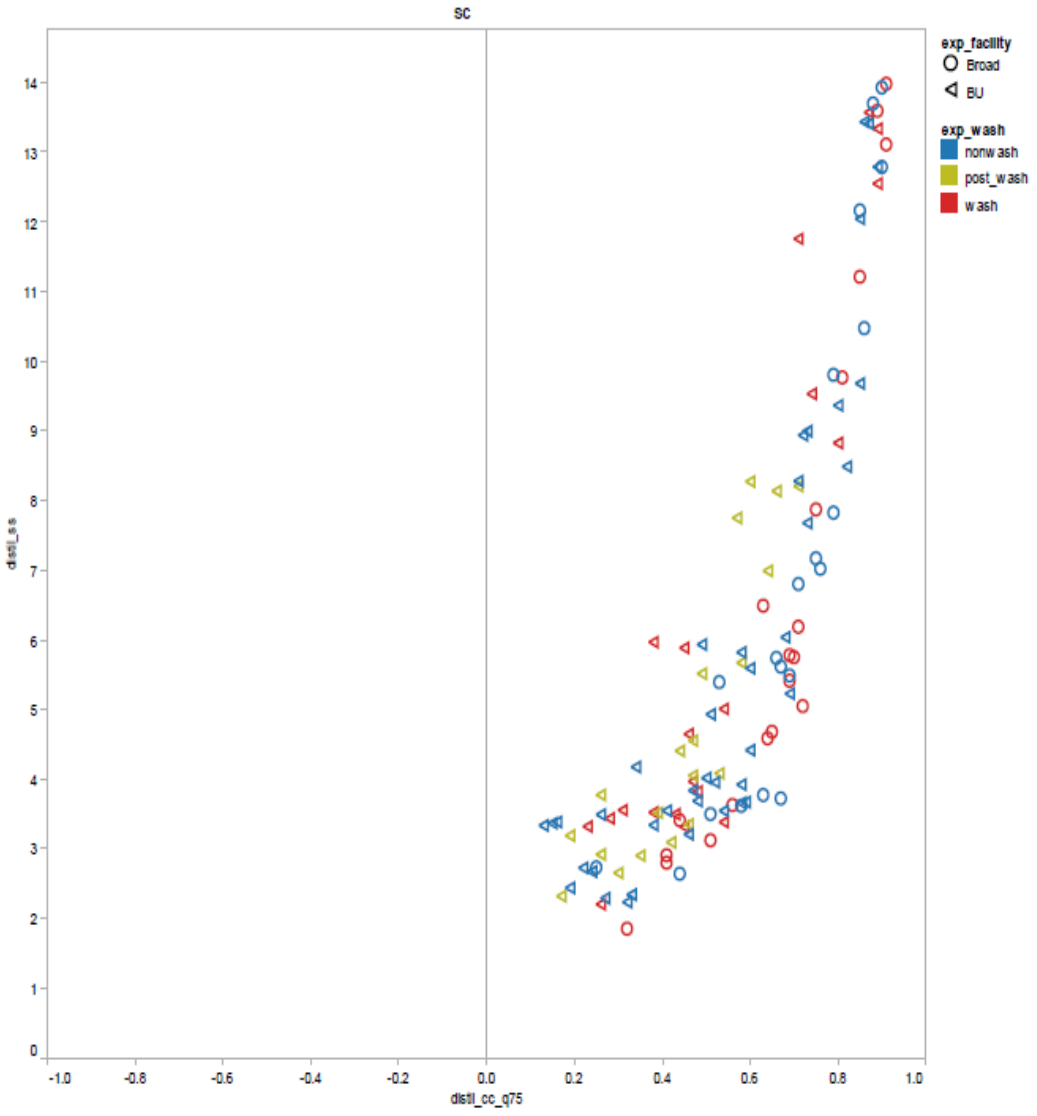


Figure 8. Signature Strength Visualization for different treatment facilities and PBS washing conditions. Comparsion of the genomic signature readout for plates washed with PBS, and not washed with PBS at both The Broad Institute, and Boston University.

cDNA extraction at Boston University versus Broad

As demonstrated from Figure 6, compounds which had the reverse transcriptase cDNA extraction step performed at the Boston University facility (post_wash: gold triangles on the figure) had a statistically significant lower signature signal than compounds which had the cDNA extraction performed at the Broad Institute (demonstrated by the red circles). Additionally, the data presented in Figure 5 demonstrate the methods preformed at the Broad Institute gave a higher signal result than at Boston University.

The differences in signature signals in data generated at the two locations may be attributed to the precision of the high throughput robots performing the cDNA extraction. This source of error would be amplified by using the Boston University Apricot liquid handler high-throughput robot for additional steps.

The treatment of the cells must take place at Boston University because of the high hazard carcinogen work, but the cDNA extraction may be performed at either facility. In the future, cDNA extraction will be performed at The Broad Institute because the data demonstrate stronger signature signals are obtained following extraction at The Broad Institute.

Washed versus non-washed cells correlations

Figures 6 and 7 display the comparison between plates which had been washed three times with PBS versus treated cells with did not receive the wash steps, for which RNA was harvested immediately.

From Figure 6 (data from plates run at The Broad Institute), the median signature strength for the treated cell lysates was 6.2 for plates which were not washed with Phosphate Buffered Saline, and 5.5 for plates which underwent the three time wash step. At Boston University, the median signature strength was 4.3 for the washed plates, and 4.0 for the non-washed plates. These correlations are presented in an alternate form in Figure 7, which includes the range of the different signature strengths for each facility, and for each condition.

The wash step for this high-throughput experiment is unavoidable, because it is necessary to ensure the carcinogens used to treat the cells are removed before harvesting the plate for RNA. The results of the washed and non-washed cells from The Broad Institute demonstrate that the wash step is reducing the strength of the signature. This is likely attributed to each wash lifting some of the treated, adherent cells from the plate. Reduced cell number before RNA preparation could cause a reduced signature capture. In the future, two PBS washes instead of three PBS washes will be used to reduce the reduction in signature strength induced by the wash step in addition to increasing the original number of cells in each well.

BU versus Broad F-Statistic Correlations

Figure 8 demonstrates the correlations between triplicates for cell treatment plates at Boston University. Correlation between triplicates compared with non-triplicates was statistically significant. Figure 9 displays that correlation between replicates as compared to non-replicates for cDNA extraction protocol-treated plates at Boston University was significant. In a comparison of Boston University-treated cell plates with plates which were washed, statistical significance was reached between replicates versus non-replicates (Figure 10).

F-Statistic tests were performed to analyze the degree of variance from normality for the treated cell lysate plates. Greatest deviation was demonstrated with Broad-treated and washed cells (Figure 11). Variation was reduced with washed lysate plates as compared with non-washed lysate plates (Figure 11).

These data demonstrate a reduced degree of error is necessary to see increased deviation from normal distribution, and consequently stronger signatures.

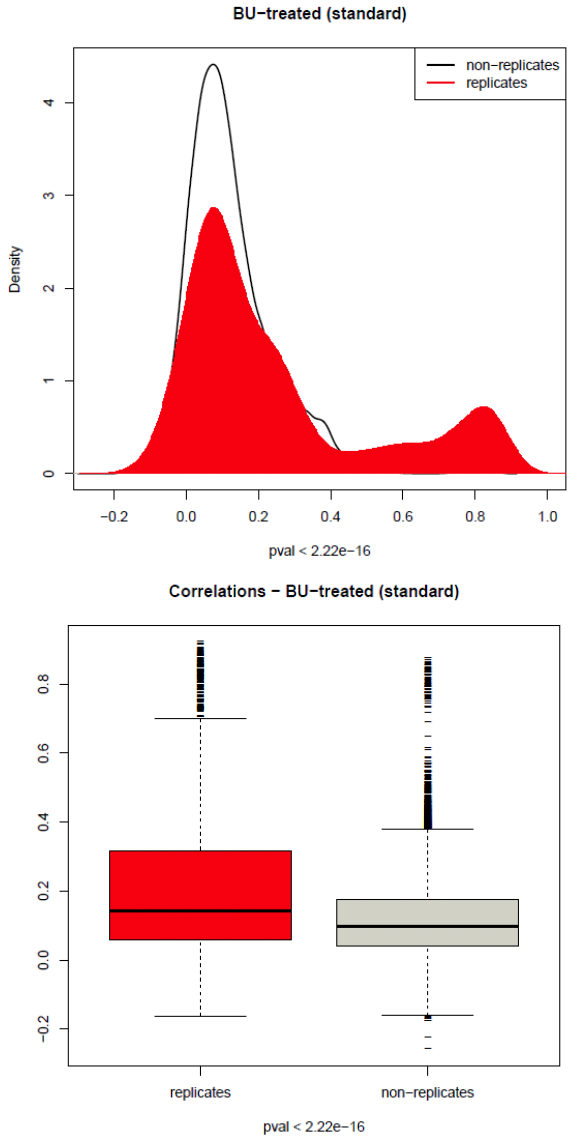


Figure 9. BU cDNA extraction

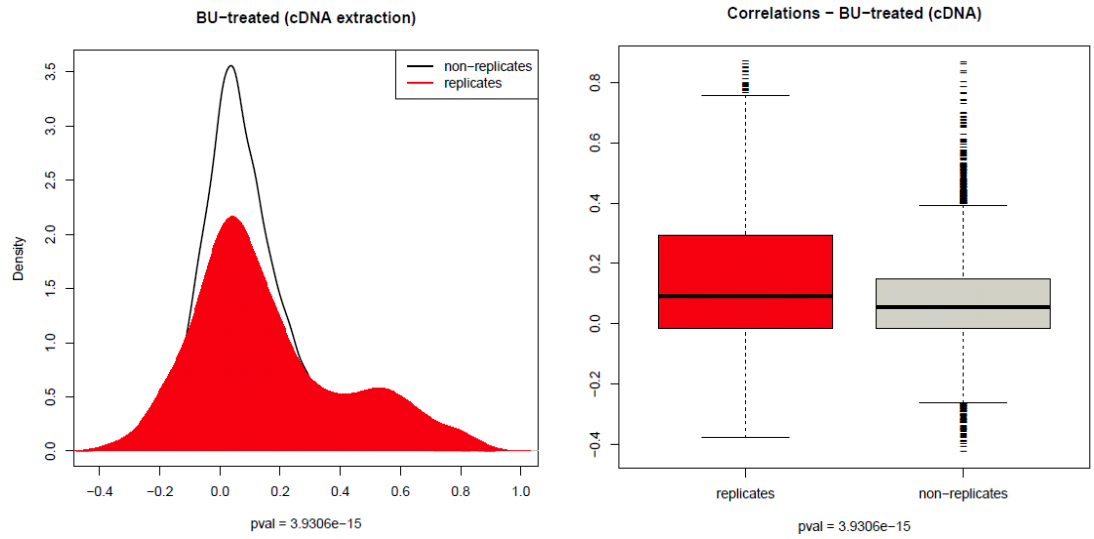


Figure 8. BU treated versus Broad treated cells

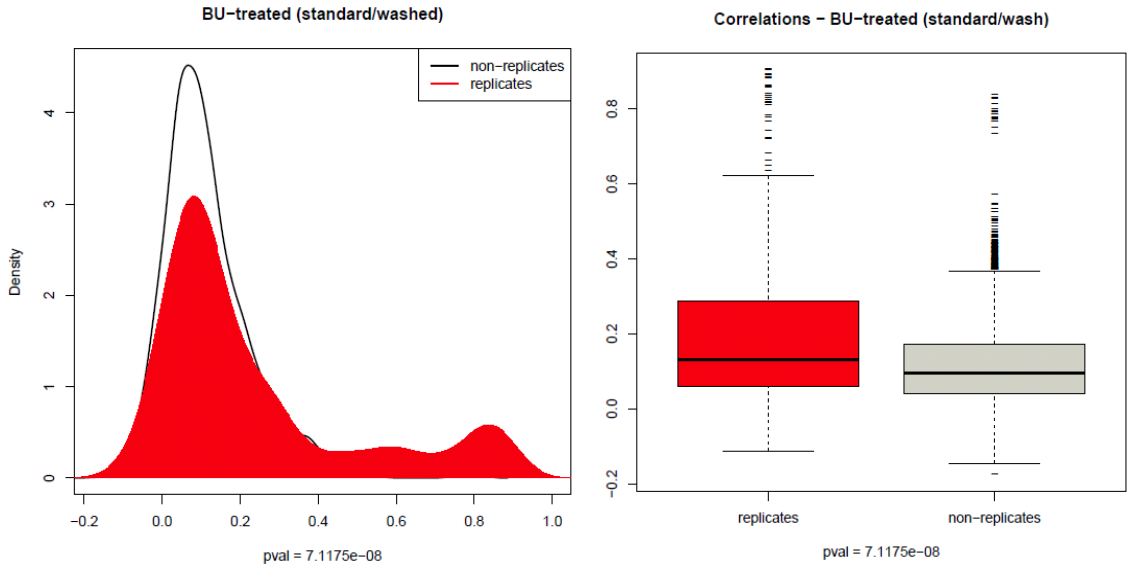


Figure 10. BU-treated compounds washed versus non-washed

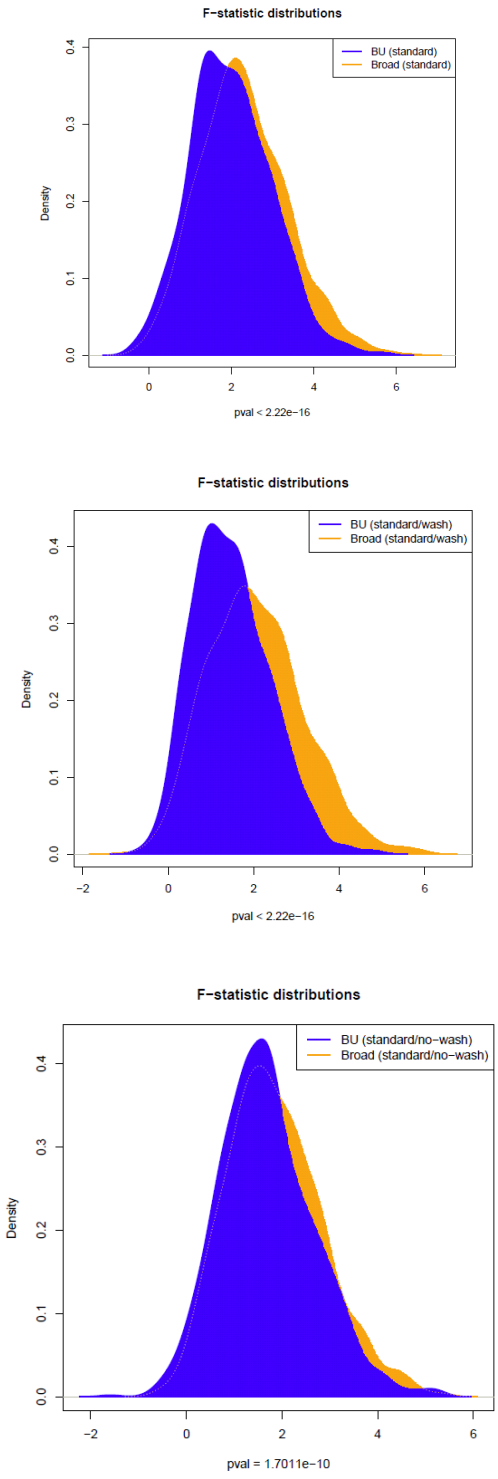


Figure 11. BU versus Broad F-Statistic distributions

DISCUSSION

Specific Aims

These experiments achieved the specific aim of insuring quality control checks within the Boston University Core facility instruments for future genomic screens. From these experiments, we can estimate a total 10% mechanical error from the cell plater, the Apricot dosing error, and from the PBS washes. Care was taken to minimize the error from these three sources and in the future it will be important to seek alternative methods for reducing the error even further. Even though the error is 10% within the core facilities, this error becomes magnified with the L1000 reactions.

The second aim of creating strong genomic signatures through the use of Boston University's core equipment shows promising results. Although there is significant variability between compounds dosed at the Broad Institute versus Boston University and between the PBS washed and non-washed test runs, it is believed these differences would be minimal in capturing a signal if the genomic signature has a significantly strong initial reading.

The third aim of this project, determining if the protocol of treating the cells at BU and then gathering data at the Broad Institute is a reasonable technique for obtaining high-throughput genomic signatures with chemicals of unknown carcinogenicity status, appears promising based on the preliminary data. Although there are significant sources of error with the Boston University core equipment and the added wash step, future minimization of this error may be achieved through improvement of the core facility equipment (updates to robotic components, calibration of instruments more precisely,

purchase of newer equipment). Such changes will likely result in stronger genomic signature readouts from the L1000.

Future Experiments and Impact

The results of these experiments demonstrate that a methodology for obtaining genomic signatures has the potential to be a replicable, high-throughput procedure to assist in creating a "carcinogenome" database (CGDB). The creation of such a database will elucidate environmental drivers for cancer and could serve as the basis for exposure risk models for cancer development. Additionally, such a reference would provide the exposure mechanisms of action (MoA), pathways of carcinogenicity (PoC), and signatures of carcinogenicity (SoC), which further assist in developing drug targets and potential cancer therapeutics.

A CGDB would provide a prediction model for labeling a chemical as a "carcinogen" or "non-carcinogen" and would serve as a measurement for the safety level of the over 80,000 chemicals in human use which have not been tested for carcinogenic potential.

Potential translational implications include a method for determining bioactivity of compounds in relation to carcinogenicity. Additionally, relative risk of exposure to mixtures could be determined via a developed CGDB which could assist in the development of cancer prevention guidelines and determining what level of chemical exposure is safe.

Future steps in increasing the breadth of impact for this method of developing carcinogenic signatures include:

1) addition of multiple cell lines (including inducible pluripotent stem cells, iPSCs),

2) the inclusion of carcinogen/carcinogen, carcinogen/non-carcinogen, and non-carcinogen/non-carcinogen mixtures in the carcinogenome, and

3) study of the different concentrations of chemical and changes in the genomic signature strength

Studying different cell lines will allow the carcinogenome database to have broad applicability to a range of diseases and conditions. Using stem cells as a base cell line will assist in the development of patient-specific monitoring of gene regulation through the L1000 genomic signature readout. Such applications depend on the ability of the L1000 technology to precisely and accurately predict genome-wide changes across multiple cell lines.

The second objective of including carcinogen/carcinogen, carcinogen/non-carcinogen, and non-carcinogen/non-carcinogen mixtures in the carcinogenome will further extend the applicability of this technology. This will provide an extensive database to which scientists and physicians may turn to understand the effects of the previously 80,000 uncharacterized environmental chemicals to which humans are frequently exposed.

Potential correlations between different concentrations of chemical and changes in the genomic signature strength will play a critical role in increasing the strength of the captured genomic signatures. Although the dose concentrations used for these preliminary experiments were based on Pubmed searches of *in vitro* toxicity data for the individual compounds, it is likely that the optimal dose for a genomic signature will differ from the highest non-toxic dose. This signature-level genomic benchmark dose was introduced by Scott Auerbach, Ph.D., through his work on DrugMatrix (DM) and its application to cancer hazard characterization at the National Toxicology Program at NIEHS. Figure 12 demonstrates Auerbach's findings of increased signature strength of a transcript with increasing dose concentration of a compound.

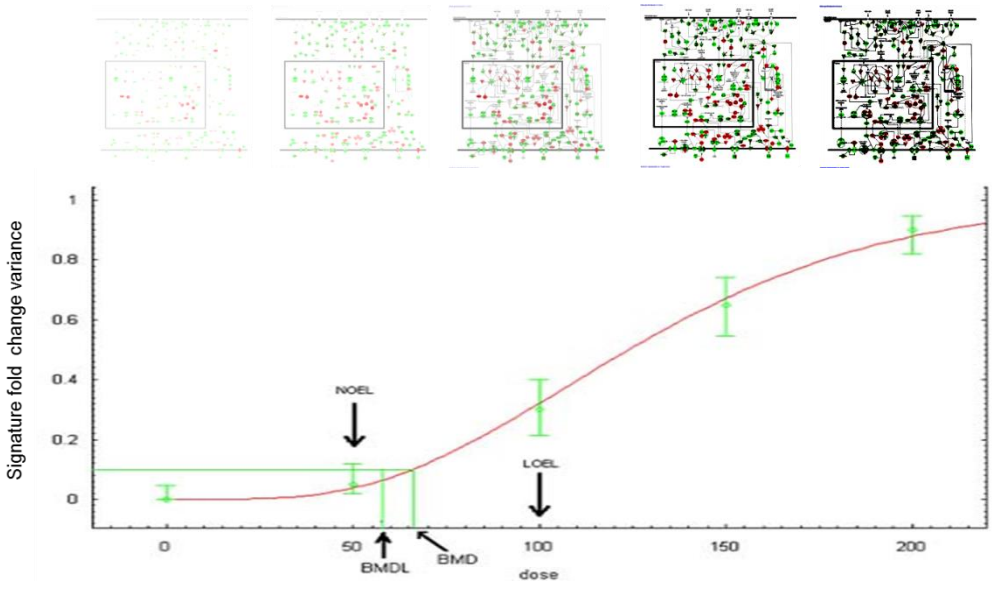


Figure 12. Computational Genomic Models of Environmental and Chemical Carcinogenicity (as presented at Boston University October 25, 2012).

In the future, these considerations will be incorporated into additional plate preparations. Further cell plates will be run in the near future to minimize the variability between triplicates, and in the additional PBS wash steps, through the methods discussed above. Then, this process will be used to capture genomic signatures in a replicable, high-throughput procedure and will assist in the creation of a “carcinogenome” database (CGDB).

APPENDIX

Appendix A: List of compounds included on Preliminary Plate

1,2,3-propanetricarboxylic acid, 2-hydroxy-3
1,2-propylene glycol
1,4-dioxane
10-undecenal
17 beta-estradiol
17-beta-trenbolone
2,3-benzofuran
2,6-octadien-1-ol, 3,7-dimethyl-,(2E)-3
2-acetylaminofluorene
2-nitrofluorene
5-fluorouracil
6-mercaptopurine monohydrate
acetaldehyde, neat
acetamide
actinomycin D
adriamycin, hydrochloride
aflatoxin B1
ampicillin trihydrate
anethole
azathioprine
Benzo [e] pyrene B[e]P
benzofuran
benzoic acid
B-estradiol
beta testosterone
bisphenol A
cadmium chloride
calcidefiol
calcipotriene
calcitriol
carbaryl
carbon tetrachloride
Proprietary Compound 1
Proprietary Compound 2
CH223191

chlorambucil
chlordecone (kepone)
chloroform
chlorotrianisene
ciglitazone
cis-Dichlorodiamine platinum
cyclophosphamide monohydrate
dexamethazone
dieldrin
diethylstilbestrol
dimethylnitrosamine
D-mannitol
DMBA
Emodin
estriol
ethinyl estradiol
etoposide
fluoxymesterone
furan
Geldanamycin
Genistein
glycerol
hydrocortisone
indol-3-carbinol (I3C)
isoniazid
isophosphamide
L-ascorbic acid
L-tryptophan
melphalan
methapyrilene hydrochloride
methotrexate
methoxychlor
methyl testosterone
monocrotaline
monoethylhexylphthalate
myleran
n-nitrosodiethylamine
oxymetholone
paciltaxel
pentachlorophenol, purified

podophyllotoxin
SAHA
silybin
streptozotocin
TCDD dioxin
tetrachloroethylene
tetracycline hydrochloride
tranilast
tributyl tin
trichloroethylene
triphenyl tin
Troglitizone
urethane
vinblastine
vincristine
Wortmannin
zearalenone

REFERENCES

1. David, A.R. & Zimmerman, M.R. Cancer: an old disease, a new disease or something in between? *Nature Reviews Cancer* **10**, 728-733 ST- Cancer: an old disease, a new disease (2010).
2. Lee Davis, D., Donovan, M., Herberman, R., Gaynor, M., Axelrod, D., van Larebeke, N., Sasco, A. The Need to Develop Centers for Environmental Oncology. *Biomedicine & Pharmacotherapy* **61**, 614-622(2007).
3. Sorensen, T.I.A., Nielsen, GG, Andersen, PK., Teasdale, TW. Genetic and Environmental Influences on Premature Death in Adult Adoptees. *New England Journal of Medicine* **318**, 727-732(1988).
4. Leffall, L.D. & Kripke, M.L. *President's Cancer Panel: Reducing Environmental Cancer Risk. 2008-2009 Annual Report* (National Cancer Institute: 2010) <http://deainfo.nci.nih.gov/advisory/pcp/annualReports/index.htm>.
5. Bucher, J.R. & Portier, C. Human Carcinogenic Risk Evaluation, Part V: The national toxicology program vision for assessing the human carcinogenic hazard of chemicals. *Toxicological sciences: an official journal of the Society of Toxicology* **82**, 363-6(2004).
6. Huff, J., Jacobson, M.F. & Davis, D.L.D.A.-J.30 2008 D.O.-10 1289/ehp 10716 The Limits of Two-Year Bioassay Exposure Regimens for Identifying Chemical Carcinogens. *Environmental Health Perspectives* **116**, 1439-1442(2008).
7. Gold, L.S. & Zeiger, E., Supplement to the Carcinogenic Potency Database (CPDB): results of animal bioassays published in the general literature through 1997 and by the

National Toxicology Program in 1997-1998. *Toxicological sciences : an official journal of the Society of Toxicology* **85**, 747-808(2005).

8. Jonker, M.J. Bruning O, van Iterson M, Schaap MM, van der Hoeven TV, Vrieling H, Beems RB, de Vries A, van Steeg H, Breit TM, Luijten M. Finding Transcriptomics Biomarkers for in Vivo Identification of (non)Genotoxic Carcinogens using Wild-type and Xpa/p53 Mutant Mouse Models. *Carcinogenesis* **30**, 1805-1812(2009).
9. Waters, M.D., Jackson, M. & Lea, I. Characterizing and Predicting Carcinogenicity and Mode of Action using Conventional and Toxicogenomics methods. *Mutation Research/Reviews in Mutation Research* **705**, 184-200(2010).
10. Kirkland, D., Aardema, M., Henderson, L., and Muller, L. Evaluation of the Ability of a Battery of Three in Vitro Genotoxicity Tests to Discriminate Rodent Carcinogens and Non-carcinogens: I. Sensitivity, Specificity and Relative Predictivity. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis* **584**, 1-256(2005).
11. Sherr, David and Monti, Stefano. Boston University School of Public Health Pilot Grant Proposal. 2011.
12. Ames, B.N., Durston, W.E., Yamasaki, E., and Lee, F.D. Carcinogens are Mutagens: A Simple Test System Combining Liver Homogenates for Activation and Bacteria for Detection. *Proceedings of the National Academy of Sciences* **70**, 2281-2285(1973).
13. Schlezinger, J.J., Bernard, P.L., Haas, A., Gradjean, P., Weihe, P., Sherr, D.H. Direct Assessment of Cumulative Aryl Hydrocarbon Receptor Agonist Activity in Sera from Experimentally Exposed Mice and Environmentally Exposed Humans. *Environmental Health Perspectives* **118**, 693-8(2010).

14. Teague, J.E., Ryu, H.Y., Kirber, M., Sherr, D.H., Schlezinger, J.J. Proximal Events in 7,12-dimethylbenz[a]anthracene-induced, Stromal Cell-dependent Bone Marrow B cell Apoptosis: stromal cell-B cell communication and apoptosis signaling. *Journal of immunology* **185**, 3369-78(2010).

15. Subramanian, A. , The Broad Institute. Manuscript in preparation. (2011).

16. Peck, D., Crawford, E., Ross, K., Stegmaier, K., Golub, T., and Lamb, J. A Method for High-throughput Gene Expression Signature Analysis. *Genome Biology* **7**, R61(2006).

17. Polo, J.M., Juszcynsky P., Monti, S., Cerchietti, L., Ye, K., Greally, J.M., Shipp, M., and Melnick, A. Transcriptional Signature with Differential Expression of BCL6 Target Genes Accurately Identifies BCL6-dependent Diff use Large B cell Lymphomas. *Proceedings of the National Academy of Sciences.* **104**, 3207-3212(2007).

18. Chen, L., Monti ,S., Juszczynsky, P., Daley, J., Chen, W., Witzig, T., Habermann, T., Kutok, J., and M. Shipp. SYK-dependent Tonic B-cell receptor Signaling is a Rational Treatment Target in Diffuse Large B-cell Lymphoma. *Blood* **111**, 2230-2237(2008).

19. Breiman, L. Arcing Classifiers (with Discussion). *The Annals of Statistics* **26**, 801-849 ST - Arcing Classifiers (with Discussion)(1998).

20. Huang, D.W., Sherman, B.T. & Lempicki, R.A. Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. *Nature protocols* **4**, 44-57(2009).

21. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ,

Clemons PA, Wei R, Carr SA, Lander ES, Golub TR. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science (New York, N.Y.)* **313**, 1929-35(2006).

22. The Broad Institute. Library of Integrated Network-based Cellular Signatures (LINCS). http://www.broadinstitute.org/LINCS/dataanalysis.html Accessed 28 October 2012.

23. The New York Academy of Sciences. The Reactivity of the Cellular Transcriptome to Xenobiotic Compound Perturbation. http://www.nyas.org/Publications/EBriefings/Detail.aspx?cid=d4bb2b89-e5db-43fc-92af-6ed1b14cff6e. Accessed 15 February 2013.

24. Luminex Company Website. http://www.luminexcorp.com/TechnologiesScience/xMAPTechnology/. Accessed 15 February 2013.

[redacted]

[redacted]

[redacted]

[redacted]