Genomic Signatures of Disease and Environmental Exposure in the Peripheral Blood

by

Heather Garren LaBreche

University Program in Genetics and Genomics
Duke University

Date:_____
Approved:

_____
Joseph R. Nevins, Supervisor

_____
Sally Kornbluth

_____
Mariano Garcia-Blanco

_____
Jen-Tsan Ashley Chi

_____
Xiao-Fan Wang

A dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor of Philosophy in the
University Program in Genetics and Genomics in the Graduate School
of Duke University

2011

ABSTRACT

Genomic Signatures of Disease and Environmental Exposure in the Peripheral Blood

by

Heather Garren LaBreche

University Program in Genetics and Genomics
Duke University

Date:_____
Approved:

_____
Joseph R. Nevins, Supervisor

_____
Sally Kornbluth

_____
Mariano Garcia-Blanco

_____
Jen-Tsan Ashley Chi

_____
Xiao-Fan Wang

A dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor of Philosophy in the University Program in
Genetics and Genomic in the Graduate School
of Duke University

2011

Copyright by
Heather Garren LaBreche
2011

Abstract

My thesis research has centered on the concept of the peripheral blood cell (PBC) as an indicator of disease and environmental exposure. Historically, PBCs have served as indicators of health based on count, morphology or the expression of particular cell surface markers. Now PBCs can be characterized as a function of their gene expression profiles in response to disease and toxicant exposure, creating a molecular phenotype, or signature. These signatures serve as biological markers (biomarkers) that can indicate the presence of disease or aid in monitoring the response to treatment while providing insight into the underlying biological mechanisms at work. This paradigm has been applied in a number of contexts, including toxicant exposure and solid tumors. In the work presented here, we describe signatures of lead (Pb) exposure and breast cancer based on peripheral blood gene expression.

Our objective in generating a blood-based signature of lead exposure was to develop a potential predictor of past and present exposure. This is particularly relevant because of continued widespread Pb exposure through both environmental and occupational sources. Pb causes significant toxicities in a number of different organ systems and is considered a potential carcinogen. Recent work also indicates that even low-level Pb exposure can impact health, especially in children, who are more susceptible to these negative health consequences. We used a mouse model of *per os* Pb exposure to generate signatures corresponding to two different doses of Pb, reflecting either high-level exposure or low-level exposure. We also analyzed the gene expression changes following removal of the Pb source. We were able to generate robust, dose-specific signatures of Pb exposure, supporting the growing body of evidence that even

low levels of Pb exposure can have biological effects, and that there is likely no safe level of exposure. We also utilized a collection of pathway signatures to identify those pathways that were activated or repressed in response to Pb exposure compared to controls. We observed an increase in interferon-gamma pathway activity in response to low-level Pb exposure and an increase in E2F1 pathway activity in response to high-level Pb exposure. These results support previous findings that low-level Pb exposure can increase interferon-gamma production, whereas high-level Pb has been shown to increase DNA synthesis.

We also describe a predictor of human breast cancer based on peripheral blood gene expression. The objective of this study was to identify and characterize PBC gene expression patterns associated with the presence of a breast tumor. A blood-based breast tumor predictor would potentially offer an easy and noninvasive means of detecting primary breast cancer as well as monitoring patients for recurrences or metastases. In addition, the concept of using a blood-based biomarker for cancer detection would have positive implications for other types of cancer for which there are no effective screening methods, such as ovarian cancer. We were able to successfully identify a robust predictor of human breast tumors based on PBC gene expression. This predictor exhibits a high level of sensitivity and specificity in distinguishing breast cancer patients and controls in an independent validation cohort. However, the true novelty in this study is that it integrates a factor modeling approach and a transgenic mouse model of breast cancer to identify biologically meaningful gene expression changes in the mouse PBC transcriptome. These genes were then used as the starting point for developing a human breast cancer predictor. This establishes an experimental system in which we can address questions that are inherently difficult to answer in human studies, such as

whether this predictor is useful in detecting breast tumors early or in monitoring patients for recurrence or metastasis. In fact, our work suggests that tumor-associated gene expression changes in the PBCs can be detected in asymptomatic mice. Our results support those of previous studies, while improving upon the sensitivity and specificity of the predictions. This may suggest that our strategy of using a mouse model to first identify informative genes allowed us to focus on those genes most relevant to the presence of a breast tumor and overcome the influence of the high degree of variation in blood gene expression in our human population. In order to be clinically useful, the predictor we report here would need to be tested in additional, large validation sets to establish its utility in an early detection setting and its specificity in distinguishing breast cancer from other cancer types as well as other potentially confounding conditions such as infection and inflammation. We describe some preliminary experiments in the mouse model intended to address these important questions.

Dedication

This work is dedicated to my precious daughter, Samantha, and my sweet baby boy, Alexander.

Table of Contents

Abstract iv

List of Tables xii

List of Figures xiii

Acknowledgements xv

1. Introduction 1

1.1. The Era of Personalized Medicine 1

1.2. Using Genomic Analysis to Dissect Complex Phenotypes 3

1.2.1. Toxicogenomics 3

1.2.2. Cancer Genomics 7

1.3 Gene Expression Signatures as Molecular Phenotypes 9

1.4 Peripheral Blood Cells as Natural Biosensors 10

1.4.1 Peripheral Blood Cells 10

1.4.2 Peripheral Blood Cells as Indicators of Disease and Environmental Exposure 11

2. Peripheral Blood Gene Expression Signatures of Lead Exposure 14

2.1 Introduction 16

2.2 Materials and Methods 17

2.2.1 Murine Pb Exposure Study 17

2.2.2 RNA Preparation and Microarray Analysis 20

2.2.3 Statistical analysis 20

2.2.4 Gene Set Enrichment Analysis 20

2.2.5 Signature Annotation 21

2.2.6 Pathway Analysis 22

2.3 Results 22

2.3.1 Generation of Dose-specific Signatures of Lead Exposure 22

2.3.2 Differential expression of IFN-gamma and E2F1 pathways.....31

2.4 Discussion 33

4. Development of a Peripheral Blood Predictor of Breast Tumors 39

4.1 Background: Peripheral Blood Gene Expression Profiling in the Context of Breast Cancer 39

4.1.1 Diagnosis..... 39

4.1.2 Prognosis 41

4.1.3 Patient Surveillance and Monitoring Response to Therapy 43

4.1.4 Rationale for a Peripheral Blood Predictor of Human Breast Tumors..... 45

4.2 Generation and Characterization of a Mouse Mammary Tumor Predictor Based on Peripheral Blood Gene Expression 51

4.2.1 Introduction 51

4.2.2 Materials and Methods 52

4.2.2.1 Mouse housing and maintenance. 52

4.2.2.2 Mouse blood collection, leukocyte isolation and RNA extraction..... 52

4.2.2.3 Complete blood count and differential 53

4.2.2.4 Quantitative RT-PCR..... 53

4.2.2.5 Comparison of MYC transcript levels from microarray analysis..... 54

4.2.2.6 Generation of a mouse mammary tumor predictor 54

4.2.3 Results..... 55

4.2.3.1 Development of a peripheral blood predictor of mouse mammary tumors 55

4.2.3.2 Probing the biological basis of a mouse mammary tumor predictor 58

4.2.4 Discussion 62

4.3 Characterization of Signature Sensitivity 64

4.3.1 Introduction 64

4.3.2 Materials and Methods 64

4.3.3 Results 65

4.3.4 Discussion 71

4.4 Development of a Peripheral Blood Signature that Predicts the Presence of Breast Cancer in Humans 72

4.4.1 Introduction 72

4.4.2 Materials and Methods 73

4.4.2.1 Patient samples 73

4.4.2.2 Human Blood Collection, PBMC Isolation and RNA Extraction 73

4.4.2.3 Development of a human breast tumor predictor 74

4.4.2.4 Functional Annotation 76

4.4.2.5 Gene Set Enrichment Analysis 77

4.4.2.6 QUADRA 77

4.4.3 Results 78

4.4.3.1 Human patient characteristics 78

4.4.3.2 Development of a peripheral blood predictor of human breast tumors .. 83

4.4.3.3 A use case for employing open source distributed version control and web-hosting tools 90

4.4.3.4 Functional annotation of the genes comprising the breast tumor predictor 94

4.4.4 Discussion 98

4.5 Characterization of Signature Specificity 101

4.5.1 Introduction 101

4.5.2 Materials and Methods 101

4.5.3 Results 102

4.5.4 Discussion 105

5. Conclusions 107

6. Perspectives 108

References 111

Biography 127

Publications 127

Presentations 127

List of Tables

Table 1. Top gene ontology categories associated with upregulated and downregulated genes in each Pb signature. 28

Table 2. TRANSFAC annotations for those genes that are upregulated in response to low-level Pb exposure, but downregulated in response to high-level Pb exposure. 30

Table 3. qRT-PCR primers for *MYC* amplification. 54

Table 4. Patient characteristics by training and validation data set. 80

Table 5. Patient characteristics by phenotype 82

List of Figures

Figure 1. Schematic describing the role of the peripheral blood cell as an indicator of various physiologic and pathologic conditions. 12

Figure 2. Experimental design for Pb exposure study. 19

Figure 3. Dose-specific signatures of Pb exposure. 24

Figure 4. Transcriptional changes following Pb exposure and recovery. 25

Figure 5. Overlap between upregulated and downregulated genes. 27

Figure 6. Differential pathway activation. 32

Figure 7. Experimental design. 48

Figure 8. Generation and validation of the mouse mammary tumor signature. 57

Figure 9. Expression levels of Myc transcript in peripheral blood. 59

Figure 10. Analysis of leukocyte subpopulations in mouse peripheral blood. 61

Figure 11. Sensitivity of the signature in detecting mammary tumors prior to palpation. 67

Figure 12. Signature accurately detects tumors up to 120 days prior to palpation. 68

Figure 13. Crossvalidation results for the refined mouse mammary tumor signature. ... 69

Figure 14. Relationship between age and tumor signature strength. 70

Figure 15. Inclusion probabilities of factors in the predictive model. 85

Figure 16. Gene expression signature predicts human breast cancer. 86

Figure 17. Distribution of test characteristics from data set permutation testing. 88

Figure 18. Distribution of test characteristics from phenotype permutation testing. 89

Figure 19. Screenshot of http://quadra.genome.duke.edu/mouse-2-human-mammary-tumor-pbmc-project. 91

Figure 20. Example of a published script from QUADRA in an HTML-formatted webpage. 92

Figure 21. Factor coherence between the training and validation sets. 93

Figure 22. Inclusion probabilities of factors in the swapped model. 97

Figure 23. Specificity of the breast tumor predictor.....103

Acknowledgements

This project has been interdisciplinary from its inception. The work described here could not have been completed without the help of many, many people. First and foremost, I would like to recognize my mentor and thesis advisor, Joe Nevins, for believing in me and believing in this project. He has helped shape me as a scientist by giving me enough independence to formulate questions and explore their answers on my own, while always providing the guidance needed to keep me from going too far astray. While he has always provided sound scientific advice, he has also known when to wear the hat of supportive friend. Graduate school does not occur in a vacuum. The relentless pursuit of scientific progress must sometimes relent and give way for the events that make up a life. Throughout these events, Joe has remained a source of encouragement and support. Although I might have muddled through without this support, it certainly made reaching my goal a lot easier. Perhaps it is a reflection on him, that his lab members are equally supportive while always providing an intellectually challenging and stimulating atmosphere. I am so grateful to have worked with and learned from the many members of the Nevins lab, past and present. Although I cannot adequately thank them all, I would like to mention a few in particular who contributed to my success. Andrea Bild mentored me as a rotation student in the lab and set a high bar for scientific rigor and work ethic. She continues to be a good friend and colleague, equally likely to give me advice on parenting as on science. Rachel Rempel and Eran Andrechek introduced me to mouse model work for the first time and helped me with the design and implementation of many of the projects described herein. Eran also spent many hours teaching me mouse surgical techniques, and his company in the “mouse house” always

made the work go by much faster. Kaye Culler, Dawn Chasse and Laszlo Jakoi make the Nevins lab run. Without these individuals, productivity would drop dramatically. I cannot thank them enough for helping me navigate everything from interdepartmental red tape to ordering reagents and designing primers. Kaye has helped me with preparing numerous fellowship applications, scheduling committee meetings, reserving rooms and submitting publications. I am especially appreciative of Laszlo's assistance in lobbying for automating our RNA isolation process. This device made my life so much easier. I am also particularly thankful to Dawn for helping me work through so many RNA preparations, not to mention the daily support – both scientific and personal.

This project would have been fairly one-dimensional without the help of Jeff Marks. I am indebted to him and his staff for helping me to obtain all of the human blood samples used in this study. Jeff always showed interest in my project, but had a healthy dose of skepticism as well, which kept me motivated to generate some convincing data. Nancy Glover was instrumental in locating and correctly annotating all of the samples, as well as helping me to learn how to use Microsoft Access to organize the clinical data. Meanwhile, these samples were actually collected in the clinic by nurse Liz Wilderman. I enjoyed meeting her at our halfway point to pick up samples and chat for a moment before getting back to work. Obviously, these samples would not have existed if not for the consent of the patients, eager to participate in our research study. Liz often commented on the cheery willingness of these women to give an extra tube of blood just to help with our research and I would like to thank each and every one of them for their contribution.

My collaboration with Erich Huang breathed a new life into my project at a time when it definitely needed resuscitation. I am thankful to Erich in general for teaching me

how to ask a really good question and use a well-defined system to answer it. More specifically, he has helped me to identify a methodology that is more appropriate for analyzing the data obtained in this study and has allowed us to link our mouse model system to human disease. This is a novel development in the field and has the potential to impact the way peripheral blood biomarkers are identified and investigated. In addition, it has been a rewarding experience to work with and learn from such a devoted scientist, physician and father.

I would also like to thank John Chute and his lab manager and technician, Sarah Meadows, for their collaboration on the lead exposure studies. The Chute lab's work in radiation exposure paved the way for our subsequent lead exposure research and set a valuable precedent for gene expression signatures of toxicant exposure. I am especially grateful for the work done by Sarah Meadows in designing and carrying out the lead exposure experiments.

I would like to acknowledge my family for their love and support. My mom and dad have encouraged me from the beginning and were my first science teachers. I am so grateful that I grew up seeing the world through their eyes. My mom is a tireless educator and a kid at heart. This combination makes her the most imaginative teacher and parent I know of, creating learning opportunities out of everyday experiences. Although I took these experiences for granted at the time, they helped shape me into the inquisitive person I am today. My dad is an outdoorsman. He has passed along to all his daughters a love of the natural world. Those early hikes and camping trips fostered an appreciation for the biology that surrounds us daily and a curiosity about its inner workings. Both of my parents worked very hard to provide me with opportunities to further my education and pursue my interest in science. They have provided emotional

support throughout graduate school and have had confidence in my abilities even when I have not. They have nursed my bruised ego when I have needed it, but they have also provided a good dose of perspective when necessary. My sisters have also provided a network of support throughout graduate school, not to mention throughout my entire lifetime. Each has served as a sounding board, therapist or even editor as needed. Although this is second nature to them, I am always pleasantly surprised by their willingness to drop everything and help out.

Last, but certainly not least, I would like to thank my husband, Matthew. He has believed in and supported me throughout graduate school. He has been by my side at each turn, helping me weather numerous rejections and setbacks, but also helping me celebrate the milestones and achievements. At times my hours and obligations have not been convenient, but he has always shown understanding and an appreciation for my work. He has patiently listened to practice talks, read manuscripts and grant proposals and accompanied me to lab in the middle of the night so that I could collect cells at exactly the right time. Graduate school has also been a special time in our lives, as we have grown into a small family during these years. The addition of Samantha to our lives in 2009 signified the happiest moment of our lives together and she continues to add joy to each day. We are looking forward to the addition of our son to this family very soon. I hope to share with my children my love of science and my passion for research.

1. Introduction

1.1. The Era of Personalized Medicine

February 2011 marks the 10-year anniversary of the completion of the first two drafts of the human genome (*1*). This monumental task required billions of dollars, the efforts of scientists across the world and over a decade to complete. Many touted the completion of the Human Genome Project as the beginning of personalized medicine and the key to curing disease. Indeed, we have already witnessed some of the fruits of this labor as evidenced by breakthroughs in cancer genomics, discoveries about the molecular basis of inherited diseases and a greater understanding of the contribution of genomic structural variation to human health (*2-4*). Such successes have prompted many in the popular press to herald the arrival of the personal genomics era. This type of optimism is evident in a statement from then Senator Barack Obama on the Genomics and Personalized Medicine Act:

> Over the past decade, we have unlocked many of the mysteries about DNA and RNA. ...This knowledge isn't just sitting in books on the shelf nor is it confined to the workbenches of laboratories. We have used these research findings to pinpoint the causes of many diseases. ... Moreover, scientists have translated this genetic knowledge into several treatments and therapies prompting a bridge between the laboratory bench and the patient's bedside.
>
> Sen. Barack Obama (S.976), March 23, 2007[1]

However, it is clear that there are many hurdles to realizing the promises of the Human Genome Project. An equally tremendous task remains in front of us; namely, translating this wealth of genomic information into meaningful insights for the

---

[1] S.976: Genomics and Personalized Medicine Act of 2007 (http://www.govtrack.us/congress/billtext.xpd?bill=s110-976)

prevention, diagnosis and treatment of disease. The key to this next step lies in the basic dogma taught in genetics courses – the relationship between genotype and phenotype. Since Mendel's discoveries, geneticists have been fascinated with understanding the link between the "molecule of life," DNA, and the variation it gives rise to. Investigating this phenomenon may involve observing a phenotype of interest and attempting to track dawn the causal genetic element(s) or it may mean introducing experimental changes to the genome and learning from the resulting phenotype. No matter what the approach, the concept of linking genotype with phenotype is especially important in human health and disease where we can apply this knowledge to improving health care, extending longevity and providing a high quality of life throughout each stage of human development. Since the first draft of the human genome, DNA sequencing technology has advanced at such a rate over the past decade that we can now sequence an individual's entire genome in a single day for less than $6000 (*5*). In essence, we can have access to a person's complete genotype. Although genome sequencing can currently provide information about Mendelian diseases and limited estimates of the risk of developing certain complex diseases, we are a long way from being able to provide comprehensive diagnosis of disease, prediction of future risk or an assessment of the most appropriate preventive strategies and therapeutics. In 1902, when Archibald Garrod described an inherited disease called an inborn error of metabolism, he also described a concept that he called "chemical individuality."

> ...the factors which confer upon us our predisposition and immunities from disease are inherent in our very chemical structure, and even in the molecular groupings which went to the making of the chromosomes from which we sprang.
>
> Archibald Garrod, 1902(*6*)

Garrod could only imagine a time when this chemical individuality could be known, catalogued and linked to human disease. The tools simply did not exist in his

lifetime. In fact, it would be half a century until Watson and Crick would publish the structure of the double helix[2]. Unlike Archibald Garrod, however, we now have the tools to dissect our very chemical structure. In fact, the age of genomics has yielded a plethora of tools for analyzing the genome, including DNA, RNA, proteins and metabolites. The challenge now is to integrate these with existing clinical tools, such as family health histories and health risk assessments in order to gain an understanding of the genomic causes of disease and apply this knowledge to medical decision-making (7).

1.2. Using Genomic Analysis to Dissect Complex Phenotypes

Researchers and physicians now have a variety of tools at their disposal for assessing variation at a global level: genome and transcriptome sequencing; analysis of genes, proteins and metabolites; and assessment of chromosomal aberrations. These tools give us the ability to establish potential biomarkers of disease and environmental exposure that can be used in the clinic and the laboratory. In the clinic, these biomarkers can be applied to improve the accuracy of diagnosis and inform the selection of effective therapies. In addition, these tools aid scientific researchers in their goal to better understand and dissect complex biological phenotypes by revealing underlying mechanisms of disease.

1.2.1. Toxicogenomics

Advances in genomic technologies have revolutionized the field of toxicology. This field is dedicated to identifying environmental factors that pose risks to human health and assessing or quantifying that risk. Toxicologists are facing the issue of how to

---

[2] On April 25, 1953 James Watson and Francis Crick first describes the double helical structure of DNA. Watson J.D. and Crick F.H.C. (1953) A Structure for Deoxyribose Nucleic Acid. Nature 171, 737-738.

best dissect the complex biological phenotypes associated with the myriad of toxins we are exposed to every day. Complicating this is the fact that the health problems caused by environmental exposures can manifest themselves as vague and common symptoms that can be caused by a number of different diseases and/or toxins and affect a wide range of tissues and organ systems. Dissecting the complex phenotypes associated with environmental exposures is made even more complicated because of the contribution of natural human genetic variation, developmental stage at the time of exposure, gender differences, level and duration of exposure and the presence other disease states. In addition, environmental and occupational exposures can often co-occur, meaning that individuals are often exposed to multiple toxins simultaneously thus making diagnosis rather difficult.

Toxicologists have traditionally used a number of *in vivo* and *in vitro* systems to study the relationship between an exposure and a well-defined biological endpoint. However, even the most comprehensive studies carried out in this manner struggle to address all of the complicating and confounding factors listed above. New genomic tools, which include genome-wide analyses of susceptibility genes, gene expression, protein expression, and epigenetic modifications, offer new methods of characterizing the effects of an exposure and establishing its relationship to the biological endpoint of interest (called phenotypic anchoring) (*8-10*). These genomic profiles offer a common language for comparing exposures under a wide variety of conditions (dose, type, etc.). These new approaches to understanding the impact of environmental exposure on human health have given rise to the hybrid field of toxicogenomics, which focuses on the application of molecular tools to the study of toxicant effects at the whole-genome level. The ultimate goal of this new field is to provide an understanding of the complex biological responses

to toxic exposure by generating comprehensive genomic profiles of exposures (*11-16*). Toxicogenomics may also allow researchers to characterize dose response at low levels of exposure for which no adverse effects are readily observed, and even predict future health outcomes based on assessments of exposure (*17, 18*).

There are many highly collaborative efforts underway to develop standardized methods of collecting and comparing this type of data as well as efforts to create an easily accessible repository for the data. Some of the existing public databases with this so-called chemogenomic information include the Chemical Effects in Biological Systems (CEBS) knowledgebase (*19, 20*), Public Expression Profiling Resources (PEPR) web database[3], ArrayExpress genomic repository[4] genomic repository (*21*), the Gene Expression Omnibus (GEO) repository[5], and the Environment, Drugs, and Gene Expression database (EDGE) (*22*).

An example of the successful use of toxicogenomic tools in improving human health can be seen in work done by Dressman, et al., which successfully used gene expression profiling of peripheral blood mononuclear cells (PBMCs) to identify mice and human patients exposed to specific doses of radiation (*23*). Additional investigation showed that these signatures could accurately predict radiation exposure in another mouse strain and up to seven days after the exposure (*24*). The group also identified gene expression signatures of partial body irradiation that showed specificity based on the area of the body exposed and the dosage received (*25*).

Similarly, in the field of benezene exposure, researchers have taken full advantage of genomic tools to generate a comprehensive genomic profile of human

---

[3] http://pepr.cnmcresearch.org/jsp/background.jsp
[4] http://www.ebi.ac.uk/arrayexpress
[5] http://www.ncbi.nlm.nih.gov/geo/

benzene exposure. Zhang et al. (*26*) combine a variety of whole-genome approaches to reveal novel biology associated with human benzene exposure. They integrate gene expression profiling, proteomic analysis, sequencing of susceptibility genes and functional targeting of those genes using RNAi techniques to better understand how benzene causes hemtatotoxicity and ultimately leads to leukemia. Although this work is ongoing, the group have already identified genes and pathways associated with hemtatotoxicity (*27*) and some associated chromosomal abnormalities (*27-29*). Other groups in the benzene exposure field have described characteristic hypomethylation profiles (*30*) and gene expression patterns (*31, 32*). Genomic studies have also helped strengthen the case for benzene as a carcinogen. Smith, et al. (*33*) have studied the benzene metabolome and the leukemic "exposome" as a novel approach to understanding how low-level benzene exposure is related to carcinogenesis. Additional work by Heijne et al. (*31*) and Hendriksen et al. (*32*) demonstrate the use of transcriptomics in dissecting concomitant exposures that include benzene.

Many of the same toxicogenomic principles that have been successful in benzene research are being applied to studies of Pb exposure. For example, Ruden et al. used expression quantitative trait locus mapping techniques to identify genomic regions containing putative "master-regulators" of response to Pb exposure in fruit flies (*34*). Bouten et al. (*35*) characterized gene expression in immortalized astrocytes, thus confirming several genes previously reported to be induced by lead and also identifying some novel ones. This work prompted a follow up study, which investigated the mechanisms by which Pb induced vascular endothelial growth factor (VEGF) expression in astrocytes, demonstrating the potential for microarray studies to reveal novel mechanisms of toxicity (*36*). Finally, Kasten-Jolly et al. have described the effects of

developmental Pb exposure on gene expression in the spleen (*37*). The results of this study provided insight into the role of Pb exposure on inhibition of heme biosynthesis and the generation of peptides that could contribute to autoimmune syndromes. It is vital to understand the effects of Pb exposure on target tissues and organs, and the gene expression patterns described in these studies serve as biomarkers of effect. However, spleen and brain tissues are not easily accessible and thus do not represent a means to assess exposure in a noninvasive manner. Our study, which presents the first microarray analysis of the effects of Pb exposure on PBCs, highlights the potential to use blood-based biomarkers to identify low- and high-level exposures and gain a better understanding of the underlying mechanisms of Pb toxicity.

1.2.2. Cancer Genomics

Perhaps no field offers better examples of the application of genomic technologies than that of cancer genomics, which has made significant strides in successfully applying the tools of genomics to dissect the complexity and heterogeneity of cancer. The ultimate goal in this field is to classify cancers based on a comprehensive understanding of the cancer genome, and then apply this knowledge to the treatment of disease. Approaches include analyzing tumor gene expression (*38-41*), characterizing the loss and amplification of genes (*42-46*), sequencing the entire cancer genome or specific subsets of genes already known to play a role in oncogenesis (*43, 44, 47*). Other more functional approaches have included systematic targeting of oncogenes with RNA interference (RNAi) to discover potential therapeutic targets (*48, 49*) or screening the effects of large libraries of compounds on gene expression in cancer cells to better understand cancer biology and identify drugs with potential anti-cancer activity (*50, 51*).

Many recent studies take advantage of the increasing accessibility and affordability of genomic technologies by taking a multifaceted approach to understanding the cancer genome. The most prominent example of this is The Cancer Genome Atlas (TCGA)[6]. The TCGA, a program piloted by the National Institutes of Health (NIH) in 2006, used this strategy of layering genomic information to better understand heterogeneity within cancer. Since 2009, the TCGA has sought to establish comprehensive genomic maps of 20 different cancer types. While this ambitious program has already had significant payoffs in the field of glioblastoma (*52*), the work is ongoing. The long-term goal of the project is to revolutionize cancer treatment. However, the short-term rewards come from the wealth of data made available to researchers who are studying combination therapies and targeted therapies. The genomic data provided through the TCGA will allow them to categorize patients based on gene expression, chromosomal aberrations, sequence polymorphisms, gene variants or some combination of these and then test whether these categories respond differently to treatment (*45*).

Regardless of the approach, it is now abundantly clear that cancers of the same tissue of origin, which were once thought of and treated as a single disease, can be stratified based on genomic information. Our laboratory, among others, has focused on gene expression as one means to better classify tumors in a way that not only instructs us about the biology of cancer, but also provides meaningful information about those genes and pathways that can be targeted therapeutically. This stratification often transcends traditional classification based on stage and other clinical factors (*53-55*). Most importantly, these subtypes are associated with significant differences in prognosis (*53-56*), response to chemotherapy (*56*), metastatic potential (*57, 58*), chromosomal

---

[6] http://www.genome.gov/17516564

aberrations (*59*) and the activity of important pathways in oncogenesis (*59-61*). Each of the studies listed here centers around the concept of using gene expression data as a surrogate molecular phenotype, or signature. This approach allows us to identify new cancer subtypes and predict important clinical features. However, the power of this approach is the ability of these molecular phenotypes to serve as a common currency between different experimental systems (e.g. primary cell culture systems, cancer cell lines, murine models and human clinical samples). This aspect is the key to our own studies, which rely on mouse models to generate robust gene expression signatures, which can then be used to better understand the underlying mechanisms of biological and pathological processes. These gene expression signatures can also serve as the basis for developing biomarkers of human disease.

1.3 Gene Expression Signatures as Molecular Phenotypes

Previous work in our laboratory has led to the development of a standard framework for developing gene expression signatures that reflect a particular biological state (*62*). The strength of these signatures is driven by the ability to model two distinct biological states in a defined and controlled system and generate a molecular phenotype based on differential patterns of gene expression. These signatures represent the extremes along a phenotypic continuum, but have been used to predict the status of samples that exist at an unknown point along the continuum. In each of these cases, the signature is generated directly from the tumor tissue or cells that have been perturbed in some way. Now, we have adapted this approach to develop gene expression signatures that reflect the peripheral blood cell response to an internal physiological change or an external environmental exposure. In this way, peripheral blood gene expression patterns

serve as molecular phenotypes of disease and exposure. These molecular phenotypes offer the advantage of providing not only a convenient way to monitor exposure or disease state through blood sampling, but also a wealth of information about the biological effects on the transcriptome. This type of data can provide insight into the underlying biological mechanisms and pathways that play a role in the response to an environmental exposure, or the role of the peripheral blood cells in diseases, such as cancer.

1.4 Peripheral Blood Cells as Natural Biosensors

1.4.1 Peripheral Blood Cells

Peripheral blood is an extremely heterogeneous tissue composed of a number of different cell types that have a wide array of functions. The cellular components of blood (referred to as peripheral blood cells or PBCs) include platelets, red blood cells, and white blood cells. Platelets (or thrombocytes) come from the fragmentation of the precursor megakaryocyte and do not possess a nucleus. They play an important role in hemostasis and the formation of blood clots. They are also a source of growth factors, including platelet-derived growth factor and transforming growth factor beta. Red blood cells (also referred to as erythrocytes) are the predominant type of cell in the peripheral blood and have the important task of delivering oxygen to the tissues of the body via the circulatory system. These cells bind oxygen with an iron-containing molecule called hemoglobin that is abundant in the cytoplasm, giving the cell its red color. White blood cells (or leukocytes) are immune cells and defend the body against foreign pathogens and infections.

There are five different types of leukocytes (neutrophils, eosinophils, basophils, lymphocytes and monocytes), all of which are derived from a multipotent hematopoietic

stem cell. Leukocytes can also be grouped into two categories based on the presence or absence of granules in the cytoplasm: granulocytes and agranulocytes. Granulocytes (polymorphonuclear leukocytes) have granules in their cytoplasm that can be viewed under the light microscope because of differential staining. These include neutrophils, basophils and eosinophils. Neutrophils are involved in the immune response to bacterial or fungal infection and function by phagocytosing pathogens. Basophils are an important component of allergic responses as they release histamine and initiate inflammation. Eosinophils are involved in immune response to parasitic infections and are also predominant in allergic responses. Agranulocytes (mononuclear leukocytes) lack apparent granules in their cytoplasm and include lymphocytes, monocytes and macrophages. The blood typically has three types of lymphocytes: natural killer cells, T cells and B cells. Natural killer cells are a type of cytotoxic lymphocyte that is able to kill cancerous cells as well as cells infected by viruses by releasing cytoplasmic granules (perforin and granzyme) that cause apoptosis. T cells (or T lymphocytes) play an integral role in cell-mediated immunity and are distinguished by the presence of cell surface receptors called T cell receptors. There a number of specialized T cells that perform very specific functions. These include T helper cells, cytotoxic T cells, memory T cells, regulatory T cells, natural killer T cells and gamma delta T cells. B cells are the central players in the body's humoral response and are responsible for making antibodies against antigens, as well as antigen presentation and immune memory (*63*).

1.4.2 Peripheral Blood Cells as Indicators of Disease and Environmental Exposure

The peripheral blood is abundant and easily accessible by routine venipuncture. This makes the blood an attractive surrogate tissue for assessing the presence of disease or the level of toxicant exposure (*64*). Furthermore, peripheral blood cells have a long

history as an indicator of disease and exposure. Physicians commonly use white blood cell counts and differentials as tools to investigate allergies, infection and even leukemia. These cells are a well-documented indicator of disease based on morphology or expression of particular cell surface markers. Recent studies have also shown that PBC gene expression profiles reflect a variety of disease phenotypes and environmental exposures (Figure 1). For example, Dressman et al. successfully used gene expression profiling of peripheral blood mononuclear cells (PBMCs) to identify patients exposed to varying doses of radiation. Gene expression analysis of PBMC transcript levels was used to identify signatures that reflected prior radiation exposure in mice, with specificity in terms of dosage. Furthermore, these signatures were used to predict radiation exposure in samples from human patients (*23*). Additional investigation showed that these signatures could accurately predict radiation exposure in another mouse strain and up to seven days after the exposure (*24*).

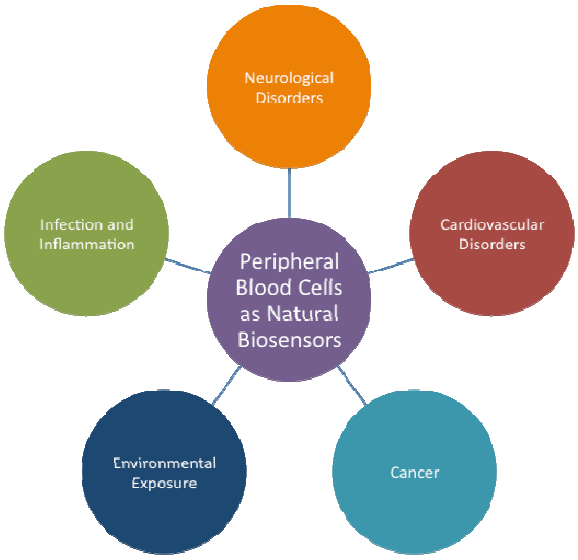


Figure 1. Schematic describing the role of the peripheral blood cell as an indicator of various physiologic and pathologic conditions.

Other researchers have identified patterns of gene expression that distinguish between phenotypes in viral and bacterial infections (*65-67*), inflammatory diseases (*68*), leukemia and lymphoma (*40, 41*), and even neurological disorders (*69*). Moreover, researchers have identified surrogate transcriptional markers that reflect the presence of solid tumors, such as renal cell carcinoma (*70*), melanoma (*71*) and even breast cancer (*72, 73*). Twine et al. identified PBMC transcripts that were differentially expressed between patients with advanced renal cell carcinoma and normal individuals. Additionally, they identified small sets of genes that were able to accurately discriminate between PBMCs from cancer patients and healthy controls (*70*). Furthermore, these gene sets have some prognostic value and specificity of tumor type (*74*). A similar predictive gene set was identified in melanoma patients. This 10-gene signature, which was identified in a purified population of CD8+ T cells, was shown to accurately discriminate between T cells from melanoma patients and healthy controls with a p-value of ≤ 0.05 (*71*). Sharma et al. identified a set of 37 genes, which discriminated between blood samples from breast cancer patients versus healthy subjects with 82% accuracy (*73*). In a 2010 follow-up study, this group performed a similar analysis using a larger sample set and a commercially available microarray platform. They were able to generate a 738-gene expression signature that discriminated between breast cancer patients and controls with 79.5% accuracy (*72*). In a study of patients with colorectal cancer, Han, et al. found a 5-gene classifier that was able to identify samples with 88% sensitivity and 64% specificity (*75*). In non-small cell lung cancer, researchers have reported a gene expression profile capable of distinguishing patients with disease from healthy controls with 78% accuracy. Furthermore, they reported that this expression signature diminished or disappeared altogether in patients following surgery (*76*).

Together these findings support the concept that solid tumors can elicit specific changes in gene expression in circulating blood cells. Additionally some of these reports suggest that peripheral blood gene expression signatures may be useful in determining prognosis (*74*) or predicting response to therapy (*77, 78*).

2. Peripheral Blood Gene Expression Signatures of Lead Exposure

The development of human cancers is clearly a complex process reflecting the acquisition of large numbers of genome alterations. For example, DNA sequence analysis of the human breast cancer genome has revealed that there are, on average, 39 to 193 mutations that alter amino acid composition in a typical breast tumor; it is estimated that an average of 14 mutations per tumor are responsible for driving tumorigenesis. While substantial progress has been made in identifying the alterations in cancer genomes, much less is understood of the contributions of environmental agents in defining the course of cancer development.

Mounting evidence indicates a role for Pb in carcinogenesis. Epidemiological studies show that environmental and occupational Pb exposures increase cancer risk, particularly lung and stomach cancer (*79-82*). There is also ample evidence from animal studies that Pb is carcinogenic, causing lung, brain, hematopoietic and kidney tumors (*83-85*). Furthermore, recent studies have revealed several biological mechanisms that might contribute to the carcinogenic effects of Pb, including the inhibition of DNA synthesis and repair, oxidative damage, interaction with DNA-binding proteins and tumor suppressor proteins and alterations to gene transcription (*86, 87*). Pb has also been shown to damage chromosomes by causing micronuclei formation, chromosomal aberrations and sister chromatid exchanges. (*85, 88*). Finally, Pb enhances the

carcinogenicity of other genotoxic and mutagenic substances such as UV-radiation, oxygen radicals and various chemical mutagens in the context of co-exposure (*89-93*). Therefore Pb exposure can cause the deregulation of several cellular pathways that are critical for normal cell division, DNA synthesis and damage repair and gene transcription, setting the stage for tumorigenesis.

An ability to detect and quantify the effects of past and present lead exposure on specific cellular pathways might facilitate the monitoring of individuals for increased cancer risk and also gain an understanding of how Pb exposure contributes to carcinogenesis. Current research emphasizes the need for research to evaluate the mechanisms of lead-induced carcinogenesis, particularly oxidative stress/apoptosis and the roles of cellular defense mechanisms, signaling pathways, and intracellular lead-binding patterns.

Despite nationwide efforts to eliminate Pb exposure in children, the Center for Disease Control and Prevention (CDC) estimates that over 300,000 U.S. children (ages 1-5 years) have elevated blood lead levels (BLLs). Although Pb has been banned from gasoline, residential paint and solder used for food cans and water pipes, it continues to pose a risk to children, who are especially vulnerable to its effects, even at low levels. Exposure is most likely to occur through inhalation or ingestion of Pb dust or through exposure to soil or water contaminated with Pb from industrial and manufacturing sources. Pb poisoning in children can cause brain damage, behavioral problems, growth delays, hearing problems, headaches and in rare cases seizures, coma and even death (*85, 94*). The CDC has set a goal to eliminate BLLs ≥10μg/dL in young children by 2010. However, many studies indicate that even BLLs below 10μg/dL can cause negative health

effects (*95-97*) and the CDC has stated that there is "no safe blood lead level" in children (*85*).

Accumulating evidence indicates that low-level Pb exposure can have a detrimental effect on health, especially that of children. In addition, many studies point to a role for Pb in facilitating carcinogenesis. Therefore, sensitive tools are needed to be able to detect past and present Pb exposures, even at low levels. In addition to assessing individual exposure history, it is imperative to be able to understand the biological mechanisms through which Pb contributes to negative health outcomes, including increased cancer risks. Therefore, two major challenges in the field of Pb toxicology are to 1) evaluate the health impact of low-level exposures that do not cause any observable health effects and 2) understand the mechanisms by which Pb affects normal cellular pathways and how this may contribute to future health risks.

2.1 Introduction

Previous studies have highlighted the opportunity to use gene expression patterns in peripheral blood cells (PBCs) as a basis for measuring exposure to environmental agents as well as a way to probe the underlying mechanisms of toxicity. These studies capitalize on the natural facility of PBCs as indicators of environmental exposure combined with the power of global gene expression analysis. For example, molecular signatures generated to diagnose radiation exposure (*23-25*) also have the potential to identify those genes that are affected by radiation exposure and may have implications in the treatment of radiation injury. Many other groups have described blood-based gene expression signatures of occupational and environmental exposures (*8-11*). We use a similar approach to develop signatures of Pb exposure based on

peripheral blood cell gene expression. However, our approach is unique in that we have made use of a collection of signatures that represent the activation of various cell-signaling pathways to identify the pathways that are either activated or repressed in response to Pb exposure. We describe dose-specific signatures of Pb exposure in a mouse model and also identify pathways that are perturbed in a dose-specific manner. These findings suggest that the response to Pb exposure may be quite distinct depending on the level of exposure. In addition, our results confirm previous studies that suggest there is no safe level of Pb exposure. Even low levels of Pb exposure alter normal gene expression patterns in peripheral blood cells and result in the aberrant activation of cellular signaling pathways involved in cell cycle regulation and immunity, which could contribute to the carcinogenic properties of Pb.

2.2 Materials and Methods

2.2.1 Murine Pb Exposure Study

Animal use and husbandry was conducted humanely and with regard to alleviation of suffering in accordance with institutional and federal guidelines. Nine-week old female Balb/c mice (Jackson Laboratories, Bar Harbor, ME) were divided into three groups of fifteen mice each. Each group received drinking water with different concentrations of Pb in the form of lead acetate trihydrate (Sigma Aldrich, St. Louis, MO) for a period of two weeks (Figure 2). The control group received 0μg/mL; the low exposure group received 5μg/mL; and the high exposure group received 50μg/mL. We chose these concentrations of Pb based on previous studies by Iavicoli et al. (*98*), who found that the relationship between the amount of Pb administered in the drinking water and the resulting BLL after two weeks of exposure could be represented by the following formula:

$$PbB=0.328\ PbW+1.617$$

In this formula PbB is blood-lead concentration and PbW is water-lead concentration. The levels of Pb exposure used in this study were chosen to reflect a PbB <10μg/dL and PbB >10μg/dL based on the CDC's definition of an elevated blood lead level (EBLL) in children <72months old (≥10μg/dL). Based on the above formula, the final PbB in the mice is predicted to be either significantly below (3.3μg/dL) or above (18μg/uL) 10μg/dL. After two weeks of exposure, peripheral blood was collected by submandibular bleed from half of the control mice (n=7) and all of the exposed mice (n=15 per group). All mice were then switched to Pb-free drinking water for two additional weeks. A second blood sample was collected after this recovery period (controls, n=7; low Pb, n=15; high Pb, n=14). All blood samples (73 total blood samples) were collected using GoldenRod animal lancets (MEDIpoint, Inc., Mineola, NY) and stabilized using RNAProtect Animal Blood Tubes and total RNA was extracted using the RNEasy Protect Animal Blood Kit (Qiagen, Valencia, CA). Globin transcript was removed using GLOBINclear for mice (Ambion, Austin, TX).

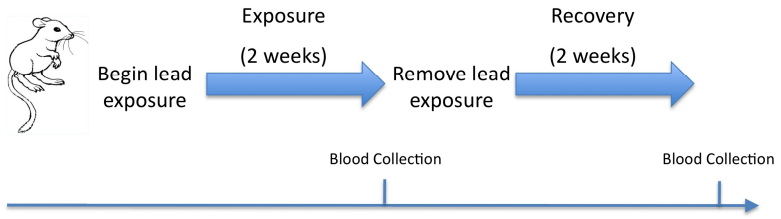


Figure 2. Experimental design for Pb exposure study.

Schematic of experimental design, in which mice were exposed to Pb via drinking water over the course of two weeks at which time a blood sample was collected. Then the lead source was removed and the mice had a two-week recovery period before a second blood collection.

2.2.2 RNA Preparation and Microarray Analysis

Total RNA quality was assessed using an Agilent 2100 Bioanalyzer (Agilent, Palo Alto, CA). Each sample (150ng) was amplified using the MessageAmp Premier RNA Amplification Kit (Ambion, Austin, TX) and analyzed using the Mouse Genome 430 2.0 Array (Affymetrix, Santa Clara, CA).

2.2.3 Statistical analysis

We compared gene expression profiles from whole blood samples isolated from mice exposed to lead acetate through drinking water or control mice. We generated a Pb exposure-associated gene expression signature composed of a subset of genes most correlated to Pb exposure by applying established methods of supervised classification and binary regression (*99*) to data normalized using the Robust Multi-array (RMA) method. The data presented here are based on signatures composed of 250 Affymetrix probes for the purposes of comparing gene lists of equivalent size. Optimal signatures for each dose (meaning the signature used to achieve maximum accuracy in a leave-one-out crossvalidation) may vary from those reported here. We calculated the average probe expression value for each group of mice at each timepoint (following exposure and recovery) and calculated the level of gene expression as the difference between Pb-exposed mice and controls for each timepoint.

2.2.4 Gene Set Enrichment Analysis

Gene Set Enrichment Analysis (GSEA, version 2.0.5) was used as previously described (*100, 101*). Briefly, we used the gene set annotation feature of GSEA[1] to identify those gene sets that overlap with the Pb signatures. These gene sets offer

---

[1] http://www.broadinstitute.org/gsea/msigdb/annotate.jsp

insights into the functional annotation of our experimentally derived signatures and have the potential to reveal the underlying biology of each signature. First, we copied the list of the Affymetrix probe identifiers comprising each signature generated by the binary regression algorithm (genecoefficients.txt) and pasted it into the browser's query field. After selecting the appropriate identifier platform (Affymetrix M430 2.0), we chose to compute overlaps between each signature and the following: C2, or curated gene sets, which is a collection of 1,892 curated gene sets; TFT, or transcription factor targets, which includes 500 gene sets that contain genes that share a transcription factor binding site defined in the TRANSFAC[2] (version 7.4) database; and C5, which are gene sets that are named by GO term and contain genes that are annotated by that term. We focused on the top 20 gene sets in our efforts to annotate the biological function of each signature.

2.2.5 Signature Annotation

In order to better understand the underlying biological response represented by each Pb exposure signature, we used publicly available tools to annotate each signature. Briefly, each signature can be represented by a list of Affymetrix (M430 2.0 Array) probe identifiers. This list was used as the input into GATHER[3] to generate the following annotations: gene ontology (GO); KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways; and enrichment for transcription factor binding sites (TRANSFAC v8.2 Professional). We also used the Affymetrix NetAffx[4] online batch query tool to translate each factor from its list of M430 2.0 Array probe identifiers to gene symbols and descriptions.

---

[2] http://www.gene-regulation.com/
[3] http://gather.genome.duke.edu/
[4] http://www.affymetrix.com/analysis/index.affx

2.2.6 Pathway Analysis

We analyzed the activity of several experimentally derived pathway signatures as they related to our Pb exposure dataset. Briefly, we used the ScoreSignatures module found on GenePattern[5] to score a series of signatures on our gene expression dataset. First, the RMA-normalized data were translated from Affymetrix M430 2.0 probe identifiers to HU133 Plus 2.0 using the FileMerger[6] tool and a bridging file from ChipComparer[7]. We averaged gene expression values for duplicated probes using an R script (by Jeff Chang, see Supplementary File 1). The dataset file was then saved in .gct file format and uploaded to the GenePattern server. After running the ScoreSignatures module, we plotted the predicted pathway probabilities (probabilities.txt) for each pathway in GraphPad Prism version 4.0b (GraphPad Software, Inc., La Jolla, CA). We tested whether the predicted probabilities for each class were significantly different using a One-way ANOVA, with Bonferroni's Multiple Comparison Test to compare the classes in a pairwise fashion.

2.3 Results

2.3.1 Generation of Dose-specific Signatures of Lead Exposure

We generated blood-based gene expression signatures that reflect changes in the transcriptome of peripheral blood cells. As shown in Figure 3, we were able to generate robust signatures for each dose of Pb exposure. Each signature demonstrated a strong capacity to distinguish Pb-exposed mice from controls in a leave-one-out crossvalidation. The low Pb signature (Figure 3A-B) showed a sensitivity of 93% and specificity of 100%

---

[5] https://cagt.igsp.duke.edu/gp/
[6] http://filemerger.genome.duke.edu/
[7] http://chipcomparer.genome.duke.edu/

based on a receiver-operator characteristic curve (AUC = 0.9619, p-value = 0.0006340). The high Pb signature (Figure 3C-D) showed 100% sensitivity and specificity in classifying samples based on exposure (AUC = 1.000, p-value=0.0002173). It was also evident from this analysis that the signatures were dose-specific. In other words, the high dose signature could not accurately distinguish low Pb exposure from control and vice versa (Figure 3B, D). Consistent with this specificity, the 250-probe signatures only have 28 probes in common. (Supplementary File 2 lists the probes in each signature along with the corresponding gene symbol and description.)

One objective of developing gene expression signatures of Pb exposure was to explore the potential that this might provide a method sensitive enough to reveal a history of exposure. To evaluate the extent to which the signature persists following Pb exposure, we analyzed the expression of the signature genes following a two-week absence of Pb in the drinking water. As shown in Figure 4, the effect was transient, with the expression of the signature genes returning to the level of background after two weeks of Pb-free drinking water. As such, it would appear that the distinct expression signature resulting from Pb exposure reflects the acute and ongoing exposure to Pb rather than a lasting effect on gene transcription patterns in the peripheral blood cells.

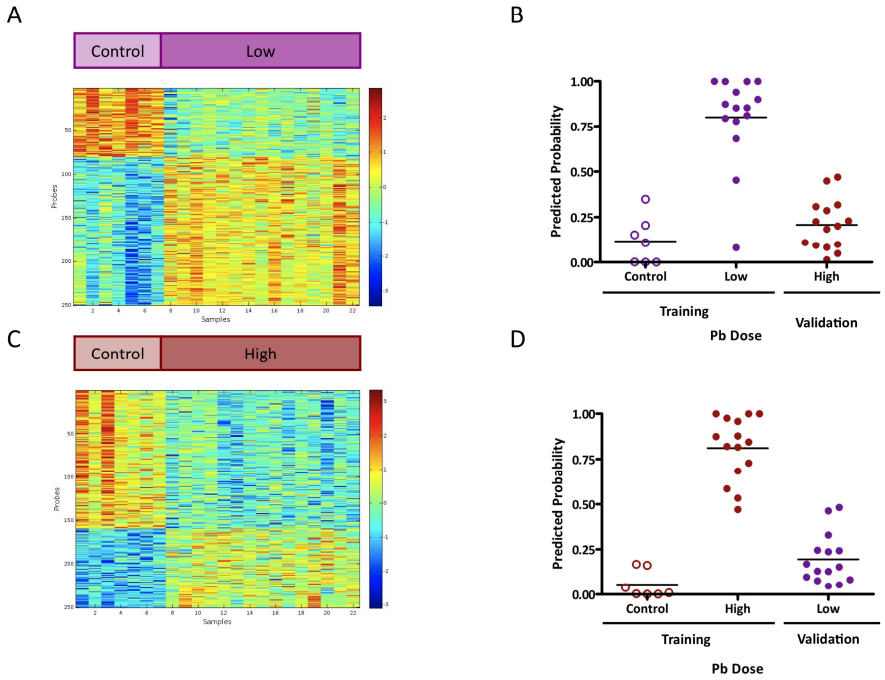


Figure 3. Dose-specific signatures of Pb exposure.

We generated signatures of lead exposure based on gene expression analysis of whole blood from exposed mice and control mice. A) A 250-probe signature of low-dose lead exposure is shown as a heatmap, where each column represents an individual biological sample and each row represents an Affymetrix M430 2.0 probe identifier. Red = high expression; blue = low expression. B) The signature is able to accurately distinguish samples from control mice from those exposed to low levels of lead as shown in the leave-one-out crossvalidation. However, this signature is specific to low level lead exposure, as there is no significant difference in predicted probabilities of samples from control mice and those exposed to high-dose lead. C) A 250-probe signature of high-dose lead exposure is shown as a heatmap of gene expression. D) While the signature is able to accurately distinguish between samples from control mice and mice exposed to high levels of lead, it is unable to classify samples from mice exposed to low lead levels, demonstrating its dose specificity. Open purple circles = control mice; close purple circles = mice exposed to low-level lead; open red circles = control mice; close red circles = mice exposed to high-dose lead.

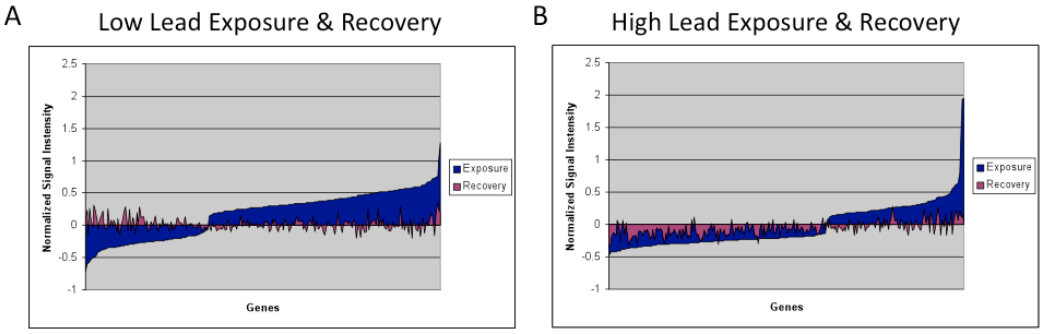


Figure 4. Transcriptional changes following Pb exposure and recovery.

A) The x-axis represents the 250 probes in the low-dose Pb signature, ordered from lowest to highest average expression value relative to controls. The y-axis represents the normalized average signal intensity of the probe across the samples. The blue field represents the mean signal intensity of the signature probes in mice exposed to low-dose Pb, normalized by subtracting the mean values in the control mice. The magenta field represents the mean signal intensity of the same signature probes after the two-week recovery period and normalized against controls. B) A comparison of the gene expression signature in mice following two weeks of exposure to high-dose lead (blue) and then two weeks of recovery (magenta).

In order to better understand the similarities and differences between the two signatures, we broke each signature into two components – genes that were either upregulated or downregulated in response to Pb exposure – and compared these for each dose. The high dose Pb signature was composed of 91 upregulated probes and 159 downregulated probes. The low dose Pb signature was composed of 170 upregulated probes and 80 downregulated probes. A comparison of the upregulated probes revealed that the signatures shared only 10 in common; a comparison of the downregulated probes revealed that the signatures only shared 7 in common (Figure 5). Table 1 shows the top gene ontology categories for the genes that are upregulated and downregulated in response to each level of Pb exposure. An analysis of the gene ontology categories that were significantly enriched in the upregulated sets of genes revealed that although the genes induced by the low and high doses of Pb did not overlap extensively, they reflected some of the same, or similar, biological processes, namely the cellular response to unfolded proteins. However, the high-dose Pb signature also includes several upregulated genes related to the processes of angiogenesis and blood vessel formation. A similar analysis of the gene ontology categories enriched in the downregulated genes revealed that the high dose signature genes were related to the cellular response to heat, and G-protein coupled receptor signaling. The downregulated genes within the low dose signature represented anti-apoptotic genes and genes involved in antigen presentation. A particularly interesting finding came from a comparison of those genes downregulated in response to high levels of Pb, but upregulated in response to low doses. This particular subset of genes was enriched for binding sites for various hormone receptors, such as androgen receptor, progesterone receptor and glucocorticoid receptor (Supplementary Table 2).

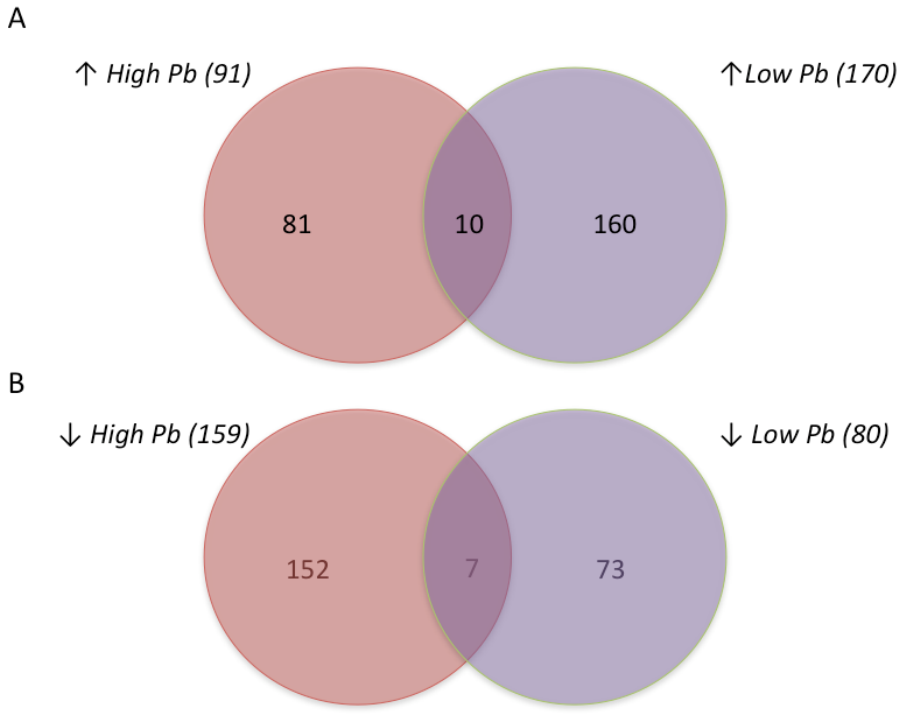


Figure 5. Overlap between upregulated and downregulated genes.

Each signature was broken down into genes either upregulated or downregulated in response to Pb exposure compared to controls. A) A Venn diagram comparing the upregulated gene lists from each signature (high-dose Pb, n=91; low-dose Pb, n=170). B) Venn diagram comparing genes downregulated in response to each dose of lead (high-dose lead, n=159; low-dose lead, n=80).

Table 1. Top gene ontology categories associated with upregulated and downregulated genes in each Pb signature.

Annotation	Count	Bayes Factor	neg ln(p value)	FE: neg ln(FDR)	Genes
Up in High Pb					
GO:0006986 [5]: response to unfolded protein	4	6.56	10.41	4.87	Dnajb1 Ern1 Hsf1 Hspa
GO:0001568 [4]: blood vessel development	4	3.84	7.28	2.8	Adra2b Edg1 Gna13 Pp
GO:0006457 [7]: protein folding	5	3.15	6.65	2.47	Dnaja1 Dnajb1 Fkbp4 H Hspa1a
GO:0001525 [6]: angiogenesis	3	2	5.33	1.98	Adra2b Edg1 Gna13
GO:0048514 [5]: blood vessel morphogenesis	3	1.67	4.96	1.98	Adra2b Edg1 Gna13
Up in Low Pb					
GO:0006986 [5]: response to unfolded protein	5	6.15	9.72	4.85	Atf6 Ern1 Hsp105 Hspc Hspd1
GO:0050877 [4]: neurophysiological process	1	5.44	9.16	0	Rcvrn
GO:0009408 [5]: response to heat	4	5.12	8.8	4.53	Dnaja1 Hsp105 Hspcb H
GO:0007186 [6]: G-protein coupled receptor protein signaling pathway	3	4.72	8.33	0	Cx3cr1 Emr4 P2ry6
GO:0009266 [6]: response to temperature	4	4.46	7.85	4.27	Dnaja1 Hsp105 Hspcb H
Down in High Pb					
GO:0009408 [5]: response to heat	4	5.64	9.31	4.29	Dnaja1 Hsp105 Hspa1a
GO:0007186 [6]: G-protein coupled receptor protein signaling pathway	2	5.16	8.91	0	P2y5 Rgs19
GO:0009266 [6]: response to temperature	4	4.98	8.8	4.29	Dnaja1 Hsp105 Hspa1a

Table continued on next page.

(Table 1 continued from previous page.)

GO:0044237 [4]: cellular metabolism	53	3.73	7.23	3.17	2610101N10Rik Calr Cb Cebpa Cept1 Chd3 Csf1 Cyp4f16 Dmrt2 Dnaja1 Eif4a1 Fos Grcc2f Hecw Hif3a Hprt1 Hsp105 Hs Hspca Jmy Lbh Magoh Mcm2 Ndufa7 Ndufab1 Nr2f2 Nrip1 Parp1 Per1 Plk4 Plod2 Ppm1g Prdx Psmb7 Ptpn2 Pxmp3 Rp Sdhb Sfrs2 Sfrs6 Slc5a8 Spcs2 Ssbp1 Suclg1 Tht Tra1 Ubc Uble1b Uhmk
GO:0006465 [6]: signal peptide processing	2	3.66	7.21	3.41	Spcs1 Spcs2
Down in Low Pb					
GO:0006916 [8]: anti-apoptosis	3	2.68	6.1	1.97	Birc1e Birc3 Cebpb
GO:0019882 [5]: antigen presentation	3	2.41	5.77	1.97	Fcgrt H2-Oa Hfe
GO:0009605 [4]: response to external stimulus	2	2.09	5.38	0	Cxcr4 Hspa8
GO:0043066 [7]: negative regulation of apoptosis	3	1.91	5.19	1.97	Birc1e Birc3 Cebpb
GO:0043069 [6]: negative regulation of programmed cell death	3	1.86	5.16	1.97	Birc1e Birc3 Cebpb

Table 2. TRANSFAC annotations for those genes that are upregulated in response to low-level Pb exposure, but downregulated in response to high-level Pb exposure.

Annotation	Count	Bayes Factor	neg ln(p value)	FE: neg ln(FDR)	Genes
V$AR_03: is recognized by androgen receptor	3	6.24	10.32	2.89	Dsip1 Tbc1d16 Zgpat
V$GR_01: high affinity binding site for glucocorticoid receptor	4	4.94	8.91	2.25	Dok2 Dsip1 Tbc1d16 Zgpat
V$PR_02: is recognized by progesterone receptor	3	3.41	7.27	1.12	Dsip1 Tbc1d16 Zgpat
V$PR_01: high affinity binding sites for progesterone receptor	2	3.06	6.84	1.07	Dsip1 Zgpat
V$BARBIE_01: barbiturate-inducible element	7	2.75	6.56	1.03	Dok2 Dsip1 Hsp105 Spcs1 Tbc1d16 Tra1 Zgpat
V$MYCMAX_B: c-Myc:Max binding sites	6	2.22	5.93	0.59	Dsip1 Hsp105 Spcs1 Tbc1d16 Tra1 Zgpat
V$E2F1_Q4: E2F-1	5	1.81	5.5	0.55	Dsip1 Hsp105 Spcs1 Tbc1d16 Tra1
V$E2F1DP1_01: E2F-1:DP-1 heterodimer	5	1.81	5.5	0.55	Dsip1 Hsp105 Tbc1d16 Tra1 Zgpat
V$AR_02: high affinity binding sites for androgen receptor	1	1.67	5.36	0.55	Zgpat
V$GRE_C: Glucocorticoid response element	2	1.65	5.3	0.55	Tbc1d16 Zgpat
V$CETS168_Q6	7	1.27	4.91	0.33	Dok2 Dsip1 Hsp105 Spcs1 Tbc1d16 Tra1 Zgpat
V$OSF2_Q6	2	1.18	4.79	0.01	Dok2 Dsip1
V$ER_Q6: estrogen receptor	1	1.08	4.68	0.01	Hsp105
V$CETS1P54_01: c-Ets-1(p54)	7	0.93	4.51	0.14	Dok2 Dsip1 Hsp105 Spcs1 Tbc1d16 Tra1 Zgpat
V$AR_Q2: androgen receptor	2	0.89	4.46	0.14	Dsip1 Zgpat
V$NFKAPPAB_01: NF-kappaB	3	0.83	4.4	0.14	Dok2 Hsp105 Tbc1d16
V$CDX2_Q5	2	0.83	4.4	0.01	Dsip1 Hsp105

2.3.2 Differential expression of IFN-gamma and E2F1 pathways

In light of the differences we observed between the low-dose and high-dose Pb signatures, we hypothesized that Pb exposure could result in the activation or repression of varying pathways based on the level of exposure. To test this theory, we took advantage of a collection of gene expression signatures representing the activity of particular cellular pathways (*59, 60*). We applied this set of signatures to the gene expression data from Pb-exposed mice and identified those pathways that were differentially activated between Pb-exposed mice and controls based predicted probability. As shown in Figure 6, the E2F1 pathway was found to be upregulated in the blood of mice exposed to high doses of Pb relative to the low dose (p-value = 0.0249). We also observed a trend in which the interferon-gamma (INFγ) pathway was upregulated in the blood of mice exposed to low doses of Pb as compared to the high dose (p-value = 0.0840).

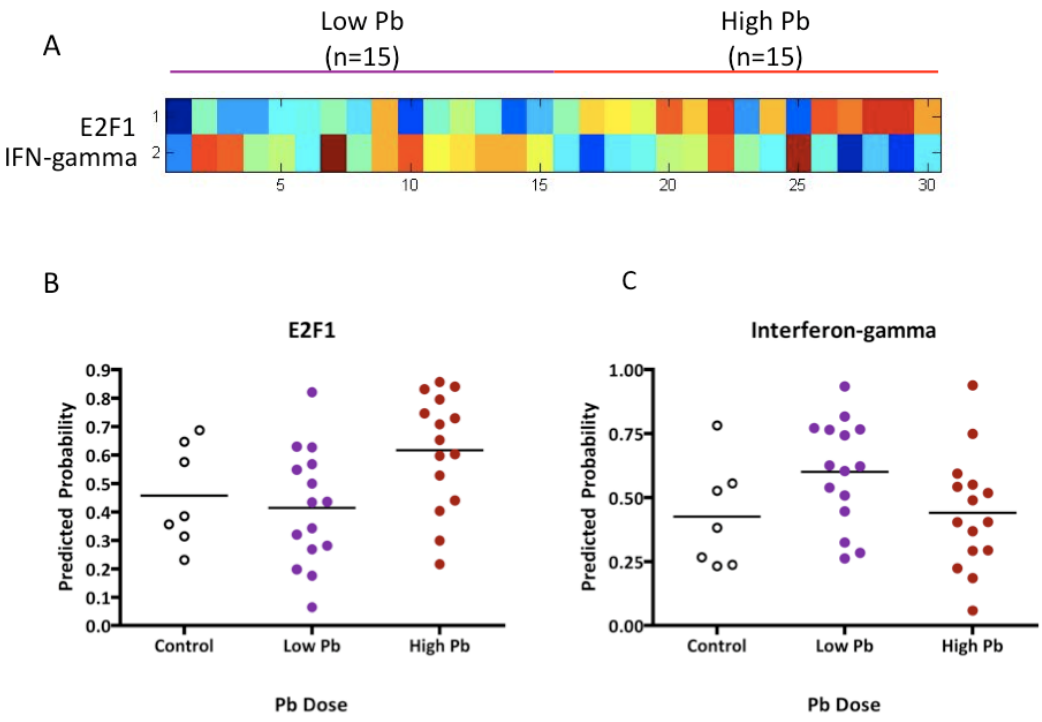


Figure 6. Differential pathway activation.

A) A heatmap displaying the predictions of pathway activity shows the predicted probability of activation of a particular pathway (E2F1, IFN-gamma) across samples from mice exposed to different levels of lead. Red = high probability; blue = low probability. B) Scatter plot of predicted E2F1 pathway activity for samples from mice exposed to low-dose lead (purple circles), high-dose lead (red circles) or no lead (control, open circles). The x-axis represents the different exposure groups and the y-axis represents the probability that the E2F1 pathway is activated in an individual sample. C) Predicted probabilities for IFN-gamma pathway activity in mouse blood samples.

2.4 Discussion

The concept of using the blood to assess Pb exposure is well established in the medical and public health fields. Measuring BLLs directly is the most common method of determining blood Pb concentration. BLL is a measure of circulating Pb, but it does not measure total Pb stored in the body. Nor does it measure the effects of current or cumulative Pb exposure. An alternative method to screen for Pb exposure is to measure the level of erythrocyte protoporphyrin (EPP) or zinc protoporphyrin (ZPP), which is an early and reliable indicator of impaired heme biosynthesis, a process known to be impaired by Pb. This method is not sensitive enough to be used to screen children, but can be used in monitoring occupationally exposed adults and can reflect average Pb exposure over a period of a few weeks. Our approach, as outlined in Figure 2, was to measure global gene expression in the blood of mice exposed to lead acetate via their drinking water and to generate a signature based on the genes most highly correlated with the exposure. Our hypotheses were that 1) PBC gene expression signatures would be good indicators of low-level Pb exposure which is not associated with observable effects, but contributes to negative health outcomes, and 2) PBC gene expression signatures may reflect past exposures, representing an opportunity to characterize an individual's exposure history and its relationship to current health issues and future health risks.

We report robust and dose-specific blood-based signatures of Pb exposure. The low-level Pb signature is particularly promising because it represents an estimated BLL of 3.3μg/dL, which is well below the 10μg/dL level for children set by the CDC. It is also below the lower limits of detection for the EPP and ZPP tests. The ability to detect such low-level Pb exposure could allow researchers to better monitor low-level exposures and associated side effects (both short-term and long-term).

Unfortunately, the signatures reported here do not seem to be indicators of past exposure. However, the gene expression data generated provide a wealth of molecular information that is useful in understanding the biological processes and signaling pathways underlying the toxic effects of Pb at different doses. Many toxicogenomic studies have analyzed transcriptome or gene expression data using various annotation or pathway building tools. We have used similar approaches to confirm some important genes and pathways that have previously been implicated in Pb exposure, such as the unfolded protein response, cell death processes, angiogenesis and endocrine disruption. We observed the upregulation of genes related to the unfolded protein response in both the low and high Pb signatures. The unfolded protein response is involves the transcription of chaperones used to aid in proper protein folding and is known to be triggered by some toxins, including Pb. Pb is thought to trigger the unfolded protein response through a mechanism that involves inhibiting protein folding by substituting for zinc or calcium ions (*102*). Apoptotic cell death is known to play a role in lead-induced neurotoxicity (*103-105*), and it is feasible that it occurs in the peripheral blood even at low levels of exposure. Additional anti-apoptosis genes are actually downregulated in the low Pb signature strengthening the argument that this low level of exposure induces cell death in the PBCs. The high Pb signature also included some upregulated genes related to angiogenesis and blood vessel formation. This is consistent with work mentioned previously, in which the mechanism of Pb-induced VEGF expression was elucidated in astrocytes (*35, 36*). Interestingly, a subset of genes that were upregulated in response to low doses of Pb and downregulated in response to high doses of Pb was significantly enriched for binding sites for various hormone receptors, such as androgen, progesterone and glucocorticoid receptors. Although Pb is a known

endocrine disruptor, leading to reproductive impairment in vertebrates, the mechanisms for this are not entirely clear and the effects of low-level exposure are not well understood (*106-109*).

In addition to traditional gene annotation approaches, we also use a novel approach to assess pathway activity by applying a panel of gene expression signatures generated by specific perturbations in controlled in vitro experiments to our dataset. Although use of this panel of signatures is not new, its application to evaluating toxicant exposure is. This approach allows us to assess the relative activity of a given pathway across samples. This approach has been successful in identifying the deregulation of oncogenic pathways in either cancer cell lines or patient tumor samples and predicting which targeted therapies might be most effective (*59, 60*). In the context of environmental exposure to toxic metals, this approach could be used as a tool for discovery to identify new pathways that are affected by a particular toxin or as a means of studying how pathway activity varies depending on factors, such as level of exposure, age, sex and co-exposure with other toxins.

At low levels of exposure, the IFNγ pathway is upregulated. This supports previous findings that Pb inhibits IFNγ production in a dose dependent fashion (*110*) enhancing development of Th2 (Type 2 T helper cell) responses versus Th1 (Type 1 T helper cell) responses (*37, 108*). Mechanistic studies suggest that Pb selectively inhibits translation of IFNγ (*111*). IFNγ plays a major role immune system function by regulating macrophage activation, differentiation of progenitor helper T cells and enhancement of the major histocompatibility complex molecule expression. INFγ is known to have some immunomodulatory effects and may be responsible for the disruption to the various endocrine pathways observed in response to low-dose Pb exposure. It is responsible for

regulating cell-mediated immune responses to infectious pathogens. It is also reported to have antitumor activity. Even transient disruption of this pathway in response to low levels of Pb exposure may have significant consequences in terms of susceptibility to disease, including infection, allergy and even cancer.

High Pb levels were found to enhance E2F1 pathway activity. This is also supported by previous findings that suggest that Pb exposure affects cell cycle and can actually increase DNA synthesis. This has been observed in a variety of contexts. Pb stimulated DNA synthesis in vascular smooth muscle cells (*112*) and rat kidney cells (*113*). Razani-Boroujerdi et al. (*114*) demonstrated that Pb enhanced proliferation of rat splenic lymphocytes in vitro. Furthermore, Lu et al. (*115*) has shown that Pb stimulates DNA synthesis in human astrocytoma cells. Increases in cell proliferation have also been observed in vivo in both rat kidney (*116*) and rat liver (*89, 117*). This cumulative evidence for the role of Pb in enhancing cell proliferation and DNA synthesis may have relevance to the potential role of Pb in carcinogenesis.

Our work has demonstrated the ability to detect the effects of Pb exposure on PBC gene expression, even at low levels of exposure. Furthermore, our results suggest that distinct biological processes are involved in the response to Pb depending on the level of exposure. For instance, we show that the IFNγ and E2F1 pathways are aberrantly regulated in response to low-level and high-level Pb exposure respectively. Unlike the current methods used to assess Pb exposure, our approach of assessing the activity of defined pathways in the peripheral blood, not only provides a dose-specific manner indicator of exposure, but it also provides information about the effects of the exposure on cellular pathways and processes. These results underscore the importance of understanding the effects of Pb exposure on critical cellular pathways. By identifying the

pathways that are impacted by Pb, we can better understand the mechanisms underlying the negative health consequences of Pb exposure and gain insight into the role of environmental and occupational exposures in contributing to overall health.

An alternative explanation for the signatures we have observed is possible contamination of the Pb source with endotoxin, which activates some of the same pathways as Pb. Through our analysis, we clearly identify many pathways previously reported in the literature as related to Pb exposure. However, to address this issue we have utilized a mouse model of endotoxin (lipopolysaccharide, or LPS) exposure. We exposed mice to various doses of endotoxin by intraperitoneal injection and collected blood samples prior to exposure, and at 6, 12 and 24 after exposure. We plan to apply our Pb signatures to gene expression data from the endotoxin-exposed mice as a validation of the specificity of the Pb signatures. In addition, future studies will include preliminary testing of the Pb source for endotoxin to rule out this possibility. A second alternative explanation for the signatures we report here is that the observed changes in gene expression are due to a Pb-induced difference in the proportions of PBC subsets. We have proposed a follow-up study to address this issue, by analyzing complete blood cell counts prior to exposure, immediately following the 2-week exposure and following the 2-week recovery period. Although Pb is known to cause anemia, it is not clear whether it affects white blood cell counts.

In addition to addressing the issues described above, additional studies are needed to truly determine whether the signatures described here are indicators of the effects of Pb or only an indicator of exposure. While an indicator of exposure (particularly very low-level exposure) is useful, there is a significant opportunity to take advantage of the gene expression data as a means to understand mechanisms of Pb

toxicity and how they relate to human health problems, including cancer. Establishing that these signatures are related to the effects of Pb would require demonstrating that the dose-specific signatures are correlated with actual blood lead levels in the mice. This would relate our signatures to established biological markers of Pb exposure. The next step would be to functionally confirm the differential regulation of some of the genes and pathways identified in our study. For instance, we are interested in using biochemical and genetic assays to confirm the activation of IFN-gamma and E2F1 in the peripheral blood. Finally, we would confirm the downstream effects of activation of these pathways, by assessing the effects on T helper cell populations in the case of IFN-gamma and assessing affects on proliferation in the case of E2F1. Both of these pathways are critical to normal cell function and have been implicated in cancer. Understanding the systemic effects of Pb on these pathways would contribute to the evolving understanding of Pb as a carcinogen.

4. Development of a Peripheral Blood Predictor of Breast Tumors

4.1 Background: Peripheral Blood Gene Expression Profiling in the Context of Breast Cancer

4.1.1 Diagnosis

Despite advances in diagnosis and treatment of breast cancer, early detection remains the key to successful treatment. Mammography, one of the most common breast cancer screening tools, aims to detect tumors when they are smallest and most treatable. In fact, the increased detection of small tumors by mammography led to a rapid increase in breast cancer incidence rates between 1980 and 1987 (*118, 119*). However, this screening tool has several limitations. There is evidence to suggest that mammograms fail to detect anywhere from 6-46% of invasive cancers, particularly in young women, women with dense breasts or in mucinous, lobular or fast growing cancers (*120, 121*). Moreover, an estimated 33% of the positive diagnoses by mammogram are the result of overdiagnosis and lead to overtreatment (*122*). Finally, it is not clear whether mammograms provide an actual decrease in mortality rates (*123, 124*). Perhaps the most important statistic is that despite the widespread use of mammography, over 200,000 cases of invasive breast cancer will be diagnosed in the United States this year (*125*). These statistics underscore the need for improved methods of early breast cancer detection.

There is a great deal of research focused on improved methods of early detection. These methods include advanced high-resolution imaging techniques, assessment of tumor cells shed into the blood, testing for predisposing variants of genes and detection of protein biomarkers in the blood. Unfortunately most of these approaches are used to

further evaluate a lesion that has been identified by imaging or biopsy, rather than as a stand-alone diagnostic or screening tool. Also none of these approaches has been shown to be more effective at early detection than mammography. One such technology is ultrasonography, which was found by the European Group for Breast Cancer Screening to be associated with "unacceptably high rates of both false positive and false negative outcomes" in population screening of asymptomatic women (*126*). There are similar caveats to using magnetic resonance imaging (MRI) as a screening modality. MRI is currently used to evaluate existing masses or differentiate between cancer and scar tissue. However, it has a high degree of false positives and studies estimate specificity from 37% to 97% (*127, 128*). Assessment of circulating tumor cells has demonstrated some utility in assessing prognosis of patients diagnosed with primary (*129*) and metastatic breast cancer (*130-132*). This type of test is currently commercially available[1] and is being administered in the clinic. However, it is typically used only in the metastatic setting as a prognostic marker and is used in conjunction with other methods of monitoring, such as imaging, physical exams and serum tumor marker tests. Genetic testing can identify patients with an increased risk of developing breast cancer. These patients can then receive heightened screening, which increases the chances their cancer will be detected early. For example, BRCA-1 and BRCA-2 mutations serve as markers of increased risk, but are informative primarily in determining prognosis (*133*). Recent studies have also characterized potential protein biomarkers of breast cancer, such as cancer antigen 15-3 (CA15-3), in the blood. Although an increase in CA15-3 can be indicative of recurrence after primary treatment, the levels of CA15-3 have not been shown to have the sensitivity and specificity required for its use in screening (*133-137*).

---

[1] CellSearch® by Veridex; LLC http://www.veridex.com/CellSearch/CellSearchHCP.aspx

Many of the early detection methods described above lack the necessary sensitivity or specificity required in a screening modality. Furthermore, there is an inherent difficulty in studying breast cancer in humans. By the time a patient exhibits symptoms of disease complex genetic and genomic changes have already taken place. We plan to overcome these limitations by using a transgenic mouse mammary tumor model combined with analysis of blood samples from human breast cancer patients. We believe that this approach has the potential to reveal PBC gene expression patterns that are predictive of the presence of a breast tumor. Such a predictor could potentially be used in the diagnosis of or screening for breast tumors. The use of the mouse model will also allow us to characterize the sensitivity and specificity of potential biomarkers in a controlled system.

4.1.2 Prognosis

Traditionally, physicians have relied on gross tumor pathology, histology and cell surface markers to determine breast cancer stage, prognosis and treatment. However, gene expression profiling of tumor tissue has shown the potential to revolutionize this process by providing a detailed molecular phenotype of the tumor. This approach provides a more precise classification of tumor types and subtypes as well as information about the molecular mechanisms underlying breast cancer etiology and progression. Our lab has successfully used this approach to create predictive models of tumor phenotype based on gene expression data and Bayesian regression models. We have shown that this methodology can be used predict clinical outcomes, such as metastasis and recurrence (*53-55, 138*), activation status of specific oncogenic pathways and response to treatment (*57, 59, 60, 139*). The ability of this methodology to predict breast cancer recurrence is illustrated by a case study of primary breast cancer recurrence,

which demonstrates that the combination of gene expression information with traditional clinical risk factors yields predictions that are more accurate than those made by using a single genomic predictor or clinical data alone (*53*). Another study demonstrated that a gene expression signature reflecting anchorage-independence, a hallmark of cancer, was able to predict metastatic potential in human tumors (*138*).

Additionally, we have successfully demonstrated that gene expression patterns, or signatures, predict activation of specific oncogenic pathways in tumors and subsequently predict response to targeted therapeutics (*54, 57, 59, 60, 139*). In work by Bild et al. (*60*), several large collections of human cancers were classified based on the activity of experimentally derived oncogenic pathway signatures. Tumors were clustered based on pathway signatures to further define prognosis of specific patient subsets, demonstrating that signatures of oncogenic pathway deregulation reflect the biology and outcome of specific cancers. Predictions of pathway activation in cancer cell lines were used to successfully predict sensitivity to therapeutic agents targeted to that pathway. This suggests that oncogenic pathway signatures not only provide information about the pathway status of tumors, but they also show potential application in guiding the selection of therapeutic strategies in the clinic. Gatza et al. (*59*)

Gene expression profiling of tumor tissue has allowed researchers to better characterize existing tumors and improve therapeutic strategies. This work has helped us better understand which tumors are likely to progress or respond to targeted therapy. Surrogate tissues, such as PBCs, would be ideal prognostic biomarkers of solid tumors due to their accessibility. In fact, peripheral blood gene expression is an established prognostic biomarker in leukemias and lymphomas (*40*). In addition, Burczynski et al. (*74*) reported that pretreatment PBC gene expression in renal cell carcinoma patients

was predictive of patient outcome. These studies suggest that PBCs may not only serve a diagnostic role, but also a prognostic one. Although this is not something we have addressed in the work presented here, we recognize the importance for incorporating transcriptome analysis of peripheral blood in prospective breast cancer clinical trials. This could reveal potential blood-based biomarkers of prognosis or even response to treatment.

4.1.3 Patient Surveillance and Monitoring Response to Therapy

In addition to regular physical exams and mammography, breast cancer patients may receive bone scans, liver sonography, chest x-rays, and blood tests of liver function in an effort to detect recurrence or metastasis. However, evidence from several randomized trials shows that this level of follow-up does not improve survival or quality of life when compared to routine physical examinations (*140-142*). Even when these tests permit earlier detection of recurrent disease, patient survival is unaffected (*142*). These data suggest that a physical examination and annual mammography may constitute sufficient follow-up for asymptomatic patients. However, this remains controversial. Recent research on improved methods of post-treatment monitoring focuses on high-resolution imaging, protein biomarkers in the serum, and the detection circulating tumor cells, or their transcripts. These approaches to post-treatment monitoring have many of the same limitations as new technologies for early detection. For example, new methods of high-resolution imaging have demonstrated earlier detection of metastases, but have not been shown to yield any actual survival benefit (*142, 143*). Protein biomarkers detected in patient serum, such as CA15-3, lack the sensitivity and specificity to be used in patient monitoring. Guidelines published by the American Society of Clinical Oncology state that although rising levels of CA15-3 may detect recurrence, this

screening method has not demonstrated clinical benefit in terms of improved patient outcome (*144*). In addition, CA15-3 is not specific for breast cancer. This serum marker is found in increased concentrations in some healthy individuals, patients with certain benign diseases (e.g. liver disease) and in patients with other types of adenocarcinomas (*134, 136, 145, 146*). The detection of particular transcripts from circulating cancer cells has shown potential prognostic and monitoring value. For instance, the detection of cytokeratins in peripheral blood of breast cancer patients can aid in prognosis (*147*). Serial monitoring of cytokeratin levels of an individual patient may also help monitor response to treatment (*148*). However, these and other transcriptional markers have not demonstrated the specificity and sensitivity needed for widespread clinical implementation. In one example of the application of blood gene expression profiling to predicted response to treatment, DePrimo et al. found that expression levels of only four transcripts in PBMCs of patients with advanced colorectal cancer could predict which treatment arm a patient was in but not response to therapy (*77*).

These methods of post-treatment monitoring have a common issue. Although they may detect metastases or recurrence earlier, they do not translate into a survival benefit. A possible explanation is that these tests may rely on a relatively high tumor burden required for detection. Once particular proteins or cancer cell transcripts reach detectable levels in the blood, the cancer may have progressed past the point at which intervention would yield a survival benefit. These potential monitoring methods also lack the sensitivity and specificity needed for widespread clinical application. We believe our approach begins to address these issues by 1) not directly relying the tumor or its by-product for detection and 2) using an animal model to characterize the signature's specificity and sensitivity.

4.1.4 Rationale for a Peripheral Blood Predictor of Human Breast Tumors

The concept of using peripheral blood cells (PBCs) as the source of information to predict the presence of cancer relies on the natural role of these cells in responding to physiologic and pathologic changes coupled with the potential of genome-scale transcriptome data to identify patterns of expression that reflect a given phenotype. Using this approach, researchers have described gene expression signatures that distinguish cancer patients from healthy individuals in a number of different types of cancer (*70-73, 75, 76, 149*). Twine et al. (*70*) generated a peripheral blood gene expression signature that could distinguish renal cell carcinoma patients from healthy volunteers. Likewise, Osman et al. (*149*) reported a whole-blood gene expression signature that could distinguish bladder cancer patients from controls. Similar studies have also been conducted in the contexts of colorectal cancer (*75*), melanoma (*71*), non-small cell lung cancer (*76*) and even breast cancer (*72, 73*). A few of these studies have also suggested that blood gene expression profiling could be used to not only distinguish cancer patients from normal, healthy volunteers, but also differentiate among tumor types (*70, 149*). Taken together, these studies highlight the potential of peripheral blood signatures as cancer biomarkers.

Our previous work has made use of expression data to develop predictors of tumor outcome, activation of various cell signaling pathways, and response to targeted therapies. In each instance, the ability to accurately define the extremes of particular phenotypes (short versus long survival, pathway off versus pathway on) has been useful in the development of informative expression signatures. These extreme cases represent an opportunity to generate gene expression data, which can be used to train a prediction

model. This prediction model can then be applied to a wide spectrum of phenotypes (*54, 60, 150*). Using a similar strategy in the context of signatures from peripheral blood cells, we hypothesized that a mouse model of breast cancer might offer the best opportunity to maximize the distinction between animals with and without tumor. As such, we have made use of a transgenic mouse model of breast cancer, which gives rise to spontaneous mammary tumors with approximately 100% penetrance in multiparous mice. Our approach was to develop latent factor models in the mouse PBC transcriptome that distinguish controls from mice with advanced mammary tumors, identify orthologous transcripts on human microarrays and use this information as the starting point for a predictive model of human breast tumors (Figure 7).

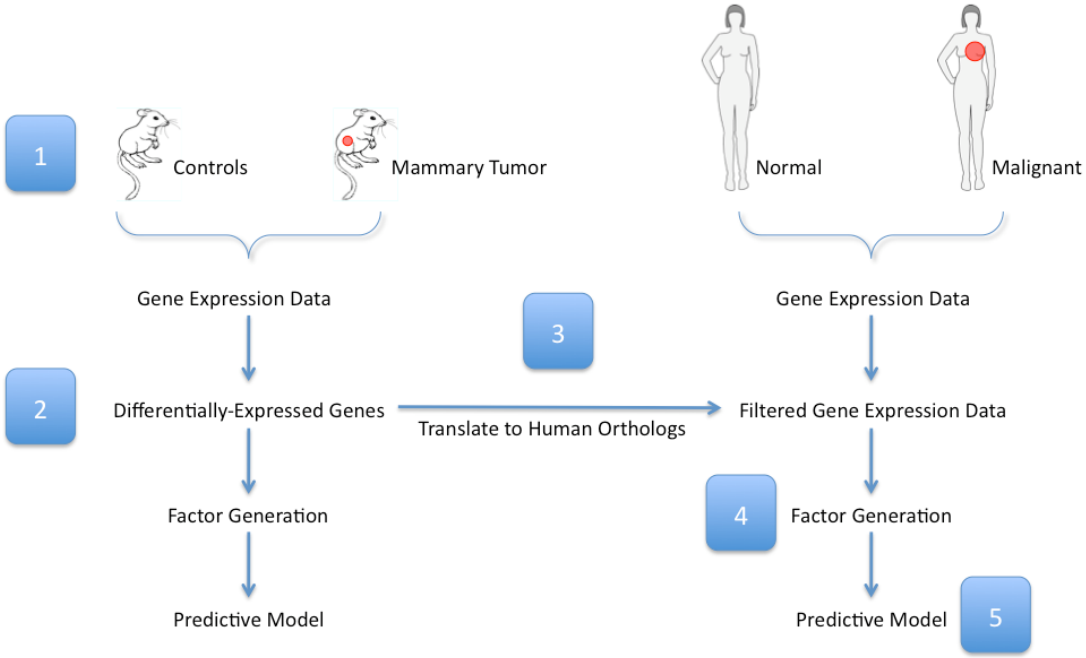

1
Controls
Mammary Tumor
Normal
Malignant
Gene Expression Data
Gene Expression Data
3
2
Differentially-Expressed Genes
Translate to Human Orthologs
Filtered Gene Expression Data
Factor Generation
4
Factor Generation
Predictive Model
Predictive Model
5

Figure 7. Experimental design.

(See opposite page.) To generate a human breast tumor predictor we took the following steps: (1) compared gene expression in PBC samples collected from tumor-bearing transgenic mice versus tumor-free controls that are matched for age and parity; (2) used a sparse ANOVA to identify 4,276 mouse Affymetrix probes with at least 0.99 posterior probability (these probes were used to independently generate a set of 49 factors that were used to develop a predictive model based on the mouse gene expression data); (3) translated these into 2,595 orthologous human probes; (4) applied BFRM to this subset of 2,595 probes to yield 26factors from a training set of 30 human PBMC samples; (5) used SSS to build 5000 possible predictive models from the training set, then projected these into a separate validation set of samples.

This strategy overcomes many of the limitations of earlier studies by using the model system to reduce noise and identify transcripts associated with the presence of a mammary tumor over other potentially confounding factors. Our results serve as a proof-of-concept in using an animal model to develop a blood-based diagnostic and establish an experimental framework for identifying possible breast cancer signatures (factor models).

There is no doubt that the computational approaches applied in this study are challenging. We therefore made use of a recently initiated pilot revision control system entitled QUADRA (Quality-Auditable Data Repository and Analysis) for storing an easily accessible version history of the work described. QUADRA was developed and implemented by Dr. Erich Huang. The guiding principle behind QUADRA is that the robust methods for maintaining, sharing and collaborating on computer programming projects espoused in the open source software movement are equally applicable to computational genomics. We believe using such a system is essential for practicing transparent and reproducible genomic science.

Breast cancer is the most common cancer for women across the world. According to (*151*), breast cancer is responsible for an estimated 1,300,000 new cases and 465,000 deaths each year. The most common method of breast cancer screening is mammography, which aims to identify breast tumors before they become symptomatic. Breast tumors identified during mammographic screening tend to be smaller and are less likely to have spread to distant sites within the body. This has important implications for a patient's prognosis. Relative survival rates for women in the United States diagnosed with localized cancer are over 98%, but for women whose cancer has metastasized, it is as low as 23% (*118*), emphasizing the importance of detecting tumors at an early stage so

that treatment can be initiated. Despite the success of mammography in reducing breast cancer mortality, there are several scenarios that illustrate the need for new and innovative methods of breast cancer detection. The growth rate of breast tumors varies significantly from patient to patient (*152*), meaning that even a biannual mammography screening program will miss some fast-growing cancers (*143, 153*). Additionally, mammographic sensitivity is lower in young women and women with dense breasts (*120, 121, 154*). Sensitivity is also lower in women who are at higher risk of developing breast cancer because of a genetic predisposition or family history (*133, 155*).

4.2 Generation and Characterization of a Mouse Mammary Tumor Predictor Based on Peripheral Blood Gene Expression

4.2.1 Introduction

While no single mouse model can fully represent the complexity of human breast cancer, genetically engineered mice (GEM) allow researchers to study specific signaling pathways, physiological events and the contribution of hormones and therapeutics. The use of transgenic and knockout mouse models has accelerated the pace of breast cancer research and has made it possible to identify and understand genetic pathways relevant to human cancer. With recent advances in microarray technology, these mouse models provide a way to explore the more complex transcriptional changes that take place during tumor formation and progression. In this study, we use transgenic mice that overexpress the c-*MYC* proto-oncogene. This mouse model shares some important features with human breast cancers. For instance, c-*MYC* is amplified in 15-20% of human breast cancers and overexpressed in 70% of breast cancers. Much like human cancer, c-*MYC* overexpression alone is not enough to transform murine mammary epithelial cells and not all cells overexpress the oncogene (*156*). Other genetic lesions are required for transformation, such as the Kras2 mutations, which are found in 44% of MMTV/c-*MYC* tumors (*157*). These tumors also share histological features with human breast cancers (*58, 158*) and were found by Desai et al. to have changes in gene expression pattern similar to human breast cancer. Desai, et al. also found this model to have changes in tumor gene expression patterns similar to other transgenic tumor mouse models, with a small subset of genes showing oncogene specific expression (*159*). Taken together, these characterizations of the MMTV/c-*MYC* mouse model indicate that it is a good model of human breast cancer based on an analysis of genetic events that

contribute to tumor initiation and maintenance, histological analysis of tumor type and gene expression analysis of the tumor tissue.

4.2.2 Materials and Methods

4.2.2.1 Mouse housing and maintenance.

Animal use and husbandry was in accordance with institutional and federal guidelines. These strains were derived by Dr. Eran Andrechek (*58*) based on the original model described by Leder, et al. (50). Female MMTV/c-*MYC* transgenic mice expressed the c-*MYC* proto-oncogene or a more stable point mutation variant (T58A) of the gene under the control of the hormone-responsive MMTV long terminal repeat (LTR) in an FVB/NJ background (Jackson Laboratories, Bar Harbor, ME). The hormones released during pregnancy and lactation have been shown to enhance expression of the oncogene. Thus, the mice were maintained in a continuous breeding program. Mice were monitored twice weekly for tumor development by palpation and tumors were measured twice weekly. Once the tumors reached 2cm^3 (approximately 12 months) the animals were sacrificed and tissue was obtained to confirm the tumors by histological analysis. As a control, female mice of the same age and background strain were maintained in the same facility and under the same breeding conditions as their transgenic counterparts.

4.2.2.2 Mouse blood collection, leukocyte isolation and RNA extraction.

Blood (50-250 µL, based on weight of the mouse) was collected from MMTV/c-MYC female mice and controls at regular intervals (approximately once per month) and again prior to euthanization (average age 431 days). Samples were collected using the submandibular (cheekpouch) method. Briefly, the vein that drains the face and cheek area was punctured by a lancet (GoldenRod animal lancet, MEDIpoint, Inc., Mineola, NY) and blood was captured in BD Microtainer™ tubes with potassium-EDTA

anticoagulant (Becton Dickinson, Franklin Lakes, NJ). Tubes were immediately inverted 10-12 times and placed on ice. Following erythrocyte lysis, total RNA was isolated from the leukocyte pellet using the QIAamp RNA Blood Mini Kit (Qiagen, Valencia, CA). Alternatively, the leukocyte pellet was immediately lysed, homogenized and stored at -80°C. RNA was isolated at a later date using the protocol described above or the adapted protocol associated with the QIAamp RNA Blood Mini Kit for use in the QIAcube (Qiagen, Valencia, CA). The quantity of RNA was assessed by absorbance at 260 nm by the NanoDrop ND-1000 (Thermo Fisher Scientific, Wilmington, DE). RNA quality was assessed by Agilent 2100 Bioanalyzer RNA 6000 PicoChip (Agilent, Palo Alto, CA). Purified total RNA (200 ng) was amplified using NuGEN Ovation™ RNA Amplification System (NuGEN Technologies, Inc., San Carlos, CA). Amplified RNA was labeled and hybridized onto Affymetrix M420 2.0 GeneChip oligonucleotide arrays (Affymetrix, Santa Clara, CA) according to the manufacturer's instructions.

4.2.2.3 Complete blood count and differential

Whole blood was collected from mice immediately following euthanization. The samples were collected from the hepatic portal vein using a 3 mL syringe and a 19-gauge needle. The sample was immediately transferred to a collection tube containing K2EDTA, inverted 10-12 times and placed on ice. Samples (130 µL minimum) were delivered to the Duke University Medical Center Veterinary Diagnostic Laboratory and were analyzed within 8 hours of collection using a Cell-Dyn 3700 Hematology Analyzer. Manual differentials were performed as necessary.

4.2.2.4 Quantitative RT-PCR

RNA was isolated from mouse PBCs as previously described. We analyzed PBC RNA from each of the 5 transgenic lines and wildtype FVB/N females in triplicate. We

also analyzed mammary tumor RNA from a transgenic mouse and mammary gland RNA from a wildtype lactating female. We used the following primer pairs to amplify the c-*MYC* transgene and total *MYC* (transgenic and endogenous).

Table 3. qRT-PCR primers for *MYC* amplification.

	Primer 1 (5' -> 3')	Primer 2 (5' -> 3')
c-MYC transgene	CTGTCCATTCAAGCAGACGA	GTATGGGTACCCTGCACCAG
Total MYC	GCCATAATTTAACTGCCTCAAA	CCTATTTACATGGGAAAATTGGA

All reported values represent the threshold change as compared to the housekeeping gene beta-actin and are relative to *MYC* expression in the wildtype lactating mammary gland.

4.2.2.5 Comparison of MYC transcript levels from microarray analysis

We identified probes from the M430 2.0 Affymetrix GeneChip related to the myelocytomatosis oncogene (avian v-myc myelocytomatosis viral oncogene homolog) by using the NetAffx query tool[4]. We then compared the average signal intensity of these probes across all 4 groups of mice based on MAS5 normalized data: 1) FVB virgin controls; 2) FVB age-matched controls; 3) MMTV/c-MYC pre-tumor and 4) MMTV/c-MYC post-tumor.

4.2.2.6 Generation of a mouse mammary tumor predictor

The entire mouse gene expression dataset was standardized employing principal components of the Affymetrix control probe sets (*160*). We then performed a sparse ANOVA on the entire mouse gene expression data set (45,101 Affymetrix probes) to identify 4,276 genes that were significantly differentially expressed between the control

and tumor-bearing mice (based on a posterior probability of ≥ 0.99). Next we randomly divided the samples into roughly equal-sized training (n = 46) and validation cohorts (n=47). We used BFRM (*161*) to identify 49 factors, or linear combinations of genes, in our training dataset. We utilized the Shotgun Stochastic Search (SSS) (*162*) method to identify small subsets of the factors that predicted tumor status within the training set, which included samples from 14 control mice and 32 transgenic tumor-bearing mice. Using our training set, we generated 5000 possible predictive models and used model averaging (based on median posterior marginal probability) of the top 200 predictive models. Model averaging has been shown to perform better than algorithms that use the single best model for prediction because it gives a truer estimation of uncertainty (*163*). Then the resulting fitted regression models were used to predict tumor status in a separate validation set of mouse PBC samples from 14 control mice and 33 transgenic tumor-bearing mice. (For fully versioned and annotated source code and data objects, please see the following QUADRA repositories: "mouse-data-prep" and "mouse-classification-model" at http://quadra.genome.duke.edu/mouse-2-human-mammary-tumor-pbmc-project.)

4.2.3 Results

4.2.3.1 Development of a peripheral blood predictor of mouse mammary tumors

The overall modeling strategy we employed involves using the mouse model to first identify those genes that are most relevant to the PBC response to a mammary tumor and then using these genes as the basis for developing a human breast tumor predictor. This strategy is outlined in Figure 7. Based on our initial findings that these transcripts were sufficient to develop a robust and accurate predictor of mouse mammary tumors, we felt it reasonable to pursue this approach. To generate a predictive

model based on peripheral blood gene expression in the mouse, we first randomly divided samples into roughly equal training and validation cohorts. We strictly adhered to maintaining the independence of the training and validation cohorts by performing all model-building steps in the training set. Following this principle, we next used a sparse ANOVA method to identify 4,276 differentially expressed transcripts between transgenic tumor-bearing mice and age-matched controls in the training cohort and subsequently used BFRM as previously described here (*161*) to generate a collection of 49 factors in the these samples. These factors were used as variables for SSS as previously described here (*162*), to generate a predictive model, which was based solely on the training data and then validated in our independent validation cohort. The predictive model performed with 100% sensitivity and specificity in the training set. More importantly, it performed equally well in the validation set, distinguishing samples based on tumor status with 100% sensitivity and specificity (Figure 8).

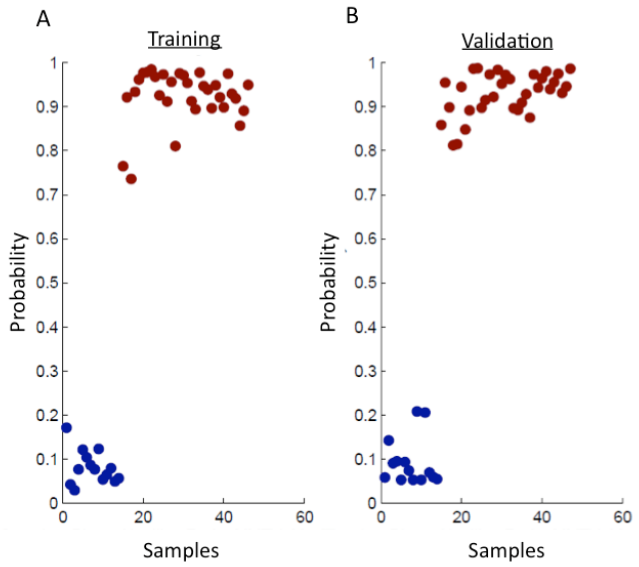


Figure 8. Generation and validation of the mouse mammary tumor signature.

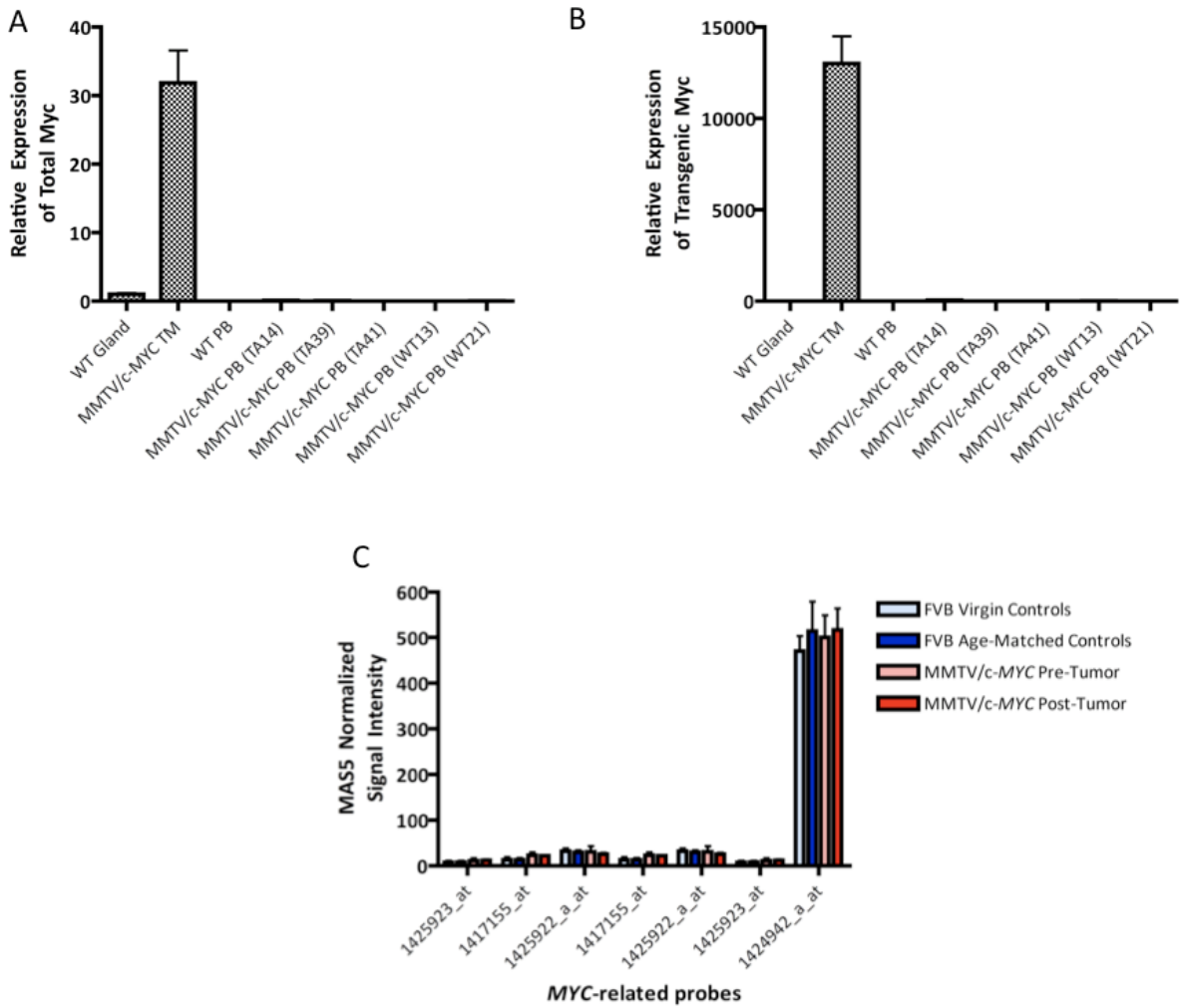
A mouse mammary tumor predictor based on gene expression of PBCs from a training set of tumor•bearing transgenic mice (red circles, n=32) and nontransgenic tumor•free mice (blue circles, n=14). A) This signature is capable of distinguishing the two classes accurately within the training data set as shown in the model fit diagram. B) Furthermore, this signature was applied to an independent set of PBC samples from transgenic tumor mice (n= 33) and nontransgenic controls (n=14) to predict the tumor status of each sample. It demonstrated 100% sensitivity and 100% specificity in predicting tumor status, using the optimal threshold of 0.8118 as calculated using Youden's J-statistic.

4.2.3.2 Probing the biological basis of a mouse mammary tumor predictor

Although the MMTV promoter is known to control c-*MYC* expression in a mammary-specific manner, we examined c-*MYC* expression in the blood cells of transgenic mice in order to confirm that the peripheral blood signature we observed was not the result of aberrant c-*MYC* expression. We used quantitative RT-PCR to confirm that the transgenic mice used in these experiments have little to no expression of the c-MYC mRNA transcript in the peripheral blood cells relative to the high levels observed in the mammary tumors from the transgenic mice. In a second approach, we compared the levels of MYC expression among the different groups of mice based on the signal intensity of the Affymetrix probe sets. Our results (Figure 9) suggest that the peripheral blood signature we observe is not a result of c-*MYC* expression in the peripheral blood cells. In addition, we show that the transgenic and wildtype mice used as the basis for the breast tumor predictor have similar leukocyte profiles, indicating that the observed gene expression changes are not due to a difference in the proportions of the various leukocyte subpopulations (Figure 10).

Figure 9. Expression levels of Myc transcript in peripheral blood.

(See opposite page.) The levels of Myc mRNA transcript in the peripheral blood were measured by quantitative RT-PCR in MMTV/c-myc mice of each of the five transgenic lines derived by Eran Andrechek. As a negative control, we analyzed expression levels in the peripheral blood and mammary gland of a wildtype lactating mouse. As a positive control, we analyzed expression levels in the tumor tissue from a transgenic mouse. Expression was normalized according to expression of the housekeeping gene beta-actin and set relative to the wildtype lactating mammary gland. A) Total Myc levels were measured, including both endogenous and transgenic transcripts. B) Levels of the Myc transgene alone were also measured, using primers specific to the transgenic construct. C) Additionally, peripheral blood Myc transcript levels were calculated based on signal intensity of the Affymetrix probesets for the myelocytomatosis oncogene (1425923_at, 1417155_at,1425922_a_at, 1417155_at, 1425922_a_at, 1425923_at, and 1424942_a_at). There were no significant differences in expression levels for any of the three probes across all three classes of mouse. Light blue = wildtype FVB virgin mice; dark blue = wildtype FVB age-matched controls; pink = MMTV/c-myc transgenic mice prior to tumor palpation; red = MMTV/c-myc transgenic tumor-bearing mice.


A
Relative Expression of Total Myc
40
30
20
10
0
WT Gland
MMTV/c-MYC TM
WT PB
MMTV/c-MYC PB (TA14)
MMTV/c-MYC PB (TA39)
MMTV/c-MYC PB (TA41)
MMTV/c-MYC PB (WT13)
MMTV/c-MYC PB (WT21)
B
Relative Expression of Transgenic Myc
15000
10000
5000
0
WT Gland
MMTV/c-MYC TM
WT PB
MMTV/c-MYC PB (TA14)
MMTV/c-MYC PB (TA39)
MMTV/c-MYC PB (TA41)
MMTV/c-MYC PB (WT13)
MMTV/c-MYC PB (WT21)
C
MAS5 Normalized Signal Intensity
600
500
400
300
200
100
0
1425923_at
1417155_at
1425922_a_at
1417155_at
1425922_a_at
1425923_at
1424942_a_at
MYC-related probes
FVB Virgin Controls
FVB Age-Matched Controls
MMTV/c-MYC Pre-Tumor
MMTV/c-MYC Post-Tumor

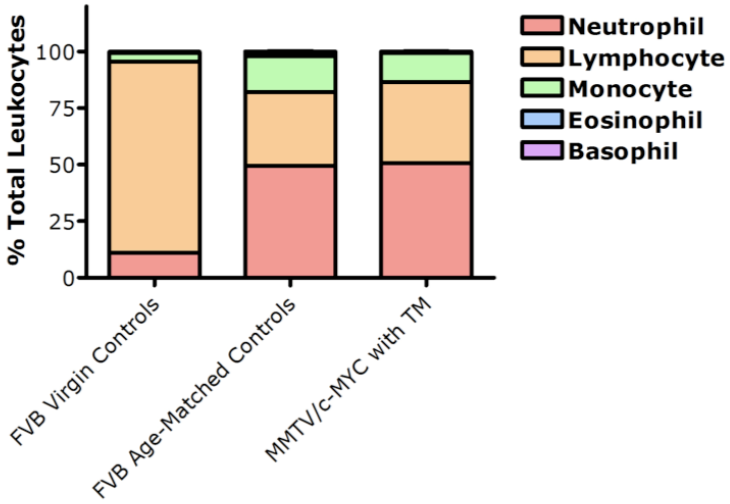


Figure 10. Analysis of leukocyte subpopulations in mouse peripheral blood.

Samples were collected from mice by venipuncture from the hepatic portal vein following euthanization in BD Microtainer™ tubes with potassium-EDTA anticoagulant, placed on ice and analyzed within 8 hours. The Duke University Medical Center Veterinary Diagnostic Laboratory analyzed samples using a CELL-DYN 3700 Hematology Analyzer. Leukocyte subpopulations were counted and calculated as a percentage of total leukocytes in each cohort of mice: virgin control mice (n=4); controls matched for age and parity (n=15); transgenic mice with advanced tumors (n=28); and wildtype mice with MMTV/c-myc tumor implants that have reached approximately 1cm in diameter (n=5). The Virgin Control mice, which were considered to be immunologically naïve, have a distinctly different distribution of leukocytes subgroups. However, all other groups of mice show similar leukocyte profiles, indicating that any gene expression differences observed are likely a result of the presence of the tumor, rather than differences in the proportion of a particular cell type.

4.2.4 Discussion

Our analysis of *MYC* transcript levels shows that the transgene is expressed at similar levels in transgenic mice and controls, indicating that the gene expression patterns we observed are not due to *MYC* expression in the PBCs. However, in order to address the potential concern that the reported signature is due to constitutive MYC expression throughout development in the transgenic mouse, we have performed similar experiments in a conditional MYC mouse model, in which MYC is turned on and off by administration of doxycyline (*157, 164*). Using this system, we hope to demonstrate that our signature is capable of accurately predicting tumor status in an independent mouse model without constitutive transgene expression. This mouse model also exhibits tumor regression upon *MYC* inactivation in a subset of mice. This would present an interesting opportunity to test whether our signature can distinguish between mice whose tumors regress compared to those whose tumors do not. This would have implications in human disease, as it would indicate that our signature was capable of monitoring patients following treatment.

Our approach to developing a predictor of human breast cancer makes use of the power of a transgenic mouse model system to first identify those genes most relevant to the presence of a tumor. The model organism provides many advantages in generating a blood-based predictor of the presence of mammary tumors. This system eliminates many of the confounding variables present in human studies, such as heterogeneous genetic backgrounds, varied environmental exposures and a wide array of co-morbid conditions and drug regimens. All of these factors contribute to a high degree of experimental noise. By using the transgenic mouse model, we are able to control for these factors and focus on the relevant phenotype – the presence of a mammary tumor. Using this approach, we

were able to identify a robust signature that was capable of predicting tumor status in an independent set of mouse peripheral blood samples with 100% sensitivity and specificity. These initial experiments will serve as the foundation for the development of a blood-based predictor of human breast tumors and will allow us to address other questions that are not as tractable in human studies, such as whether peripheral blood biomarkers are detectable in the early stages of tumorigenesis, and whether they can be used to detect cancer recurrence or metastasis.

4.3 Characterization of Signature Sensitivity

4.3.1 Introduction

Although there is a great emphasis on improving methods of early detection of breast cancer, it is extremely difficult to study human breast cancer in its early stages. By the time a patient presents with symptoms, many genetic and genomic changes have already taken place. To help circumvent these obstacles, we chose a transgenic mouse model of mammary carcinogenesis with a defined genetic background and a homogeneous natural history of disease. This allowed us to not only to establish a robust gene expression profile associated with the presence of an advanced mammary tumor, but it also allowed us study peripheral transcriptional markers at very early stages of disease in a controlled system as well as test their predictive capacity in mice at early stages of tumor development (prior to tumor palpation). We hypothesized that the robust changes in peripheral blood gene expression observed in the mouse might serve as a basis for detecting tumors at an early stage. The original mouse model provided gene expression data from two extreme phenotypes, which we used to build a predictor of the presence of a tumor. We then applied this predictor to samples collected from the same transgenic mouse cohort throughout the process of tumorigenesis. Our goal was to assess the sensitivity of our predictor as a function of tumor progression and establish a threshold of detection.

4.3.2 Materials and Methods

We compared gene expression profiles of PBCs isolated from tumor-bearing transgenic mice and nontransgenic tumor-free controls (n=15 per group). We generated a tumor-associated gene expression signature composed of a subset of genes most correlated to tumor development by applying established methods of supervised

classification and binary regression developed by West, et al. to MAS5 normalized data (*99*). Then, we applied this signature to a separate validation set of samples not used to build the signature, which included 50 samples from tumor-bearing mice and 13 controls. In order to test the sensitivity of the signature as a function of tumor progression, we used samples collected from transgenic mice prior to the time a tumor was first palpated (n=3 for each timepoint) as well as control samples (n=11) that reflected the same range of ages of the transgenic mice. We analyzed gene expression as described above and MAS5 normalized these data together with that used to generate the original signature. We also applied ComBat[1] software developed by Johnson, et al. (165) and written in R to account for batch effects. We re-trained the model using the entire mouse gene expression data set generated from tumor-bearing mice (n=65) and tumor-free controls (n=28), with the addition of half of the control samples collected over a range of ages (n=6). In this way we used the control samples to account for changes in blood gene expression patterns due to age and refine the gene expression pattern originally observed in the transgenic mouse model. We used this refined signature to predict the tumor status of the remaining control samples as well as the samples from transgenic mice using the methods described above.

4.3.3 Results

Past experience has demonstrated the feasibility of using gene expression signatures generated from two distinct biological states to accurately predict the status of samples representing a range of intermediate phenotypes. Therefore, we hypothesized that the mouse mammary tumor signature could be used to detect tumors early based on blood samples collected from our transgenic mice prior to tumor formation. To test this

[1] http://statistics.byu.edu/johnson/ComBat

hypothesis, we used blood samples collected from our transgenic mouse cohort throughout the course of tumorigenesis as well as samples from the control mice over the same time period. First, we used some of these control samples to refine our original signature in order to reflect the impact of age on the peripheral blood gene expression profile. Next, we applied this refined signature to samples collected from transgenic mice over the course of tumor development to predict whether any given sample came from an animal with a tumor present. We observed a general increase in predicted tumor probability as the tumor progressed (Figure 11). Especially noteworthy is that samples collected up to 120 days prior to tumor palpation showed a marked increase in tumor probability compared to those collected earlier than 120 days or controls (Figure 12). This difference in predicted probability is significant as determined by receiver-operator characteristic (ROC) curve comparing all samples collected more than 120 days prior to tumor palpation to those collected at 120 or less (AUC = 0.8074, p-value = 0.01337) as shown in Figure 12B. The results of the leave-one-out crossvalidation for the samples in the training data set are shown in Figure 13. We confirmed that the increase in predicted probability was not due to the potentially confounding factor of age by plotting predicted probability as a function of age of the mouse. Although probability of a tumor is positively correlated with age in the transgenic mouse, we observed no such correlation in the tumor-free controls, suggesting that the effect is due primarily to the presence of the tumor (Figure 14).

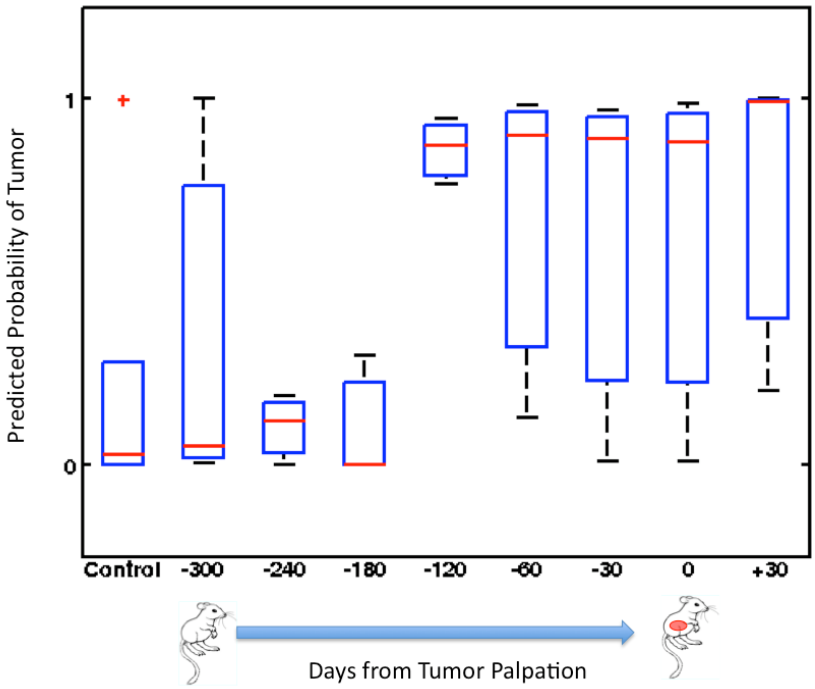


Figure 11. Sensitivity of the signature in detecting mammary tumors prior to palpation.

Boxplot of predicted tumor probability of blood samples collected from MMTV/c•myc mice at timepoints ranging from 300 days before tumor palpation until 30 days after tumor palpation. Each timepoint includes 3 separate biological samples. The control group includes 5 separate biological samples. The median value is represented by a red line; the interquartile range is shown as a blue box; the whiskers extend to data points within 1.5 times the interquartile range; and outliers are depicted as red crosses.

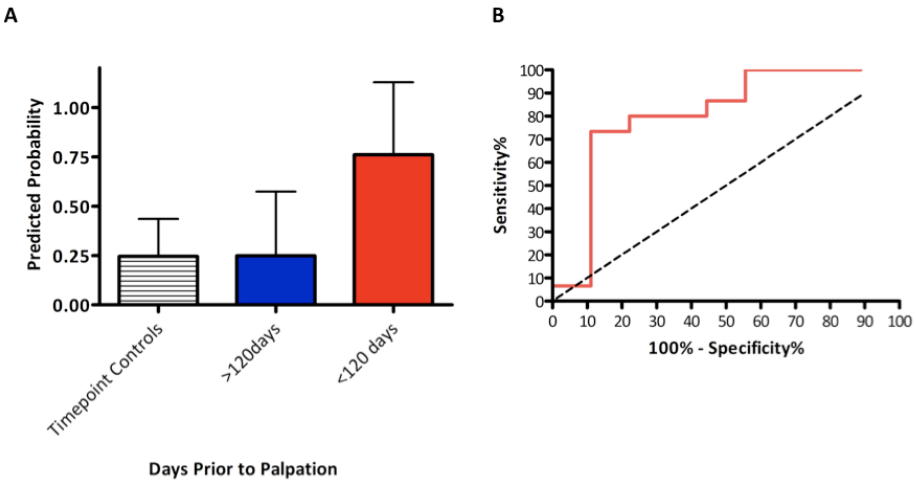


Figure 12. Signature accurately detects tumors up to 120 days prior to palpation.

The samples were grouped into two categories – those collected more than 120 days before tumor palpation (>120) and those collected 120 days or less before tumor palpation (≤120). A) The mean predicted probabilities were 0.25 and 0.76 respectively, which are significantly different with a p-value of 0.0023 using an unpaired T-test. B) We used a ROC curve to analyze the ability of the model to accurately classify samples as either positive or negative for a tumor based on an optimal cutoff of 0.3448. This model had an overall accuracy of 81%, calculated as the area under the curve (p-value = 0.01337).

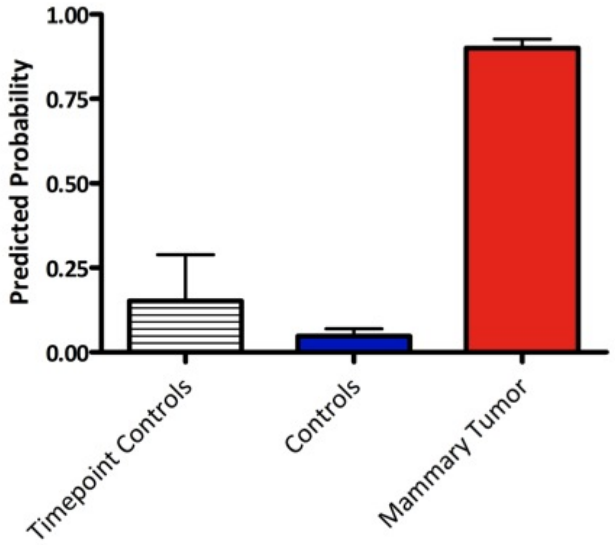


Figure 13. Crossvalidation results for the refined mouse mammary tumor signature.

Predicted probabilities are plotted for the samples used to generate the refined signature. Horizontal lines = timepoint controls chosen to match the age range of the transgenic mice (n=5); blue = controls matched for age and parity (n=28); and red = MMTV/c-myc transgenic mice with advanced tumors (n=65).

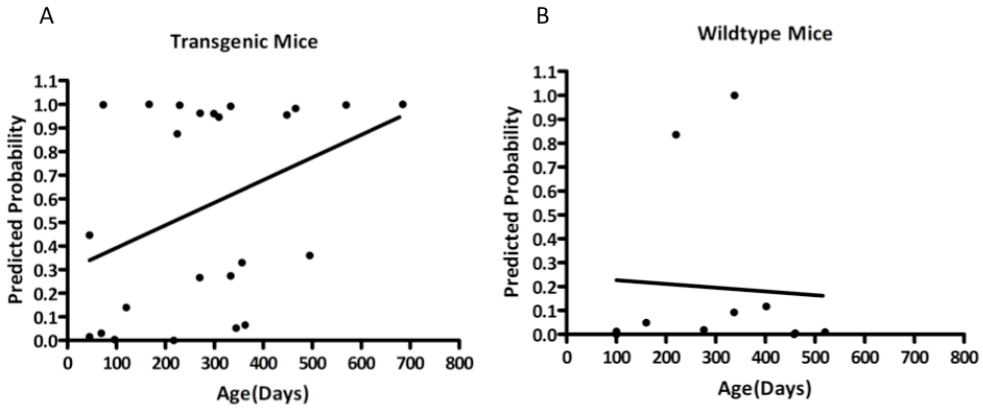


Figure 14. Relationship between age and tumor signature strength.

A) Data indicate a strong positive correlation between age of the mouse and predicted tumor probability (r^2=0.1438). However, the observed peripheral blood signature is most likely a function of the presence and progression of the mammary tumor rather than the age of the mouse. B) This is supported by the fact that we observed no such correlation in the tumor-free controls (r^2=0.004103).

4.3.4 Discussion

The use of the mouse model system allows us to identify robust gene expression signatures that can be then be applied to heterogeneous clinical samples to develop blood-based biomarkers of breast cancer. Furthermore, the mouse model facilitates exploration of the presence of a tumor-associated signature in mice prior to evidence of a palpable tumor. Our results suggest that the mouse mammary tumor signature may be sensitive enough to detect tumors at a very early stage of development. These findings would be further strengthened by additional research to correlate changes in gene expression with more precise measures of tumor burden based on *in vivo* imaging or histological staining of the mammary tissue. Although we do not fully know how the timeline of tumor progression in the mouse translates into the natural progression of human breast cancer, the results of this study show promise in using a peripheral blood biomarker for the purposes of early cancer detection. In human breast cancer, the ability to identify disease at an early stage, prior to the spread of micrometastases, presents a significant opportunity for improving outcome in breast cancer. Unfortunately, in many instances patients present with advanced disease. This approach could prove to be valuable in other contexts as well. A key example can be seen with ovarian cancer where the vast majority of patients are diagnosed with stage IV disease. Given that patients diagnosed with such advanced disease face a poor prognosis, the development of methods for detection at an early stage of the disease process would have a substantial impact on outcome.

4.4 Development of a Peripheral Blood Signature that Predicts the Presence of Breast Cancer in Humans

4.4.1 Introduction

Based on our success in developing a robust predictor of mammary tumors in the mouse model using a factor modeling approach, we sought to develop an analogous predictor of human breast tumors. As previously described, our strategy was to utilize the mouse model as a way to focus on those genes that were clearly related to the subtle, but characteristic PBC response to a tumor. We did this by using the significantly differentially genes identified in the mouse as the starting point for developing a factor model to predict the presence human breast tumors. One potential alternative strategy was to apply the factor modeling approach directly to human PBMC gene expression data to generate a possible predictive model. We tested this strategy as well, and it is discussed below; however, it does not take into account confounding variables related to the collection of the PBMC samples that could potentially result in overfitting of the model to this particular data set. Namely, samples collected from healthy cancer-free women were collected prior to a screening mammogram. However, in order to enrich for patients with a diagnosis of either benign breast abnormalities or breast cancer, samples were also collected from women referred for a diagnostic biopsy. This means that the samples were collected in two different locations and under different circumstances. This could conceivably have an effect on peripheral blood gene expression. However, we have controlled for this variable by first identifying significantly differentially expressed transcripts in the mouse model, allowing us to focus on gene expression changes related to the presence of the tumor rather than other factors. This follows the strategy our laboratory has successfully applied in other contexts by generating a gene expression

signature in a well-defined experimental system, so that it can then be easily applied to more complex and heterogeneous sample sets such as human clinical samples.

4.4.2 Materials and Methods

4.4.2.1 Patient samples

Peripheral blood mononuclear cell (PBMC) samples were collected from women with a suspect initial mammogram prior to undergoing a diagnostic biopsy procedure to determine whether the detected abnormality was benign or malignant. In total, we collected blood from 57 women with a diagnosis of breast cancer and 37 with a benign diagnosis. We also collected blood from 31 women with normal initial mammograms as negative controls and 15 breast cancer patients following surgery. All breast cancer patient samples were collected at the Duke University Medical Center under an institutional review board (IRB)-approved protocol[1] after obtaining informed consent and were provided by Dr. Jeffrey Marks. PBMC samples from patients with various gastrointestinal cancers (n=15) were collected and stored at Duke University Medical Center under IRB-approved protocols[2] and were provided by Dr. Jeffrey Marks. Peripheral blood leukocyte samples from patients with brain tumors (n=7) were provided by Dr. John Sampson and were collected by leukapheresis as part of IRB-approved protocols[3].

4.4.2.2 Human Blood Collection, PBMC Isolation and RNA Extraction

Blood samples (8 mL) were drawn into BD Vacutainer™ CPT™ Cell Preparation tubes with sodium citrate (Becton Dickinson, Franklin Lakes, NJ) by routine venipuncture, inverted 8-10 times to mix, labeled with a patient identification number

---

[1] Duke eIRB#12025
[2] Duke eIRB#12010 and 12025
[3] Duke eIRB#00003877 and #00009403

and transported to the laboratory at room temperature for processing within 2 hours. Tubes were mixed again immediately prior to processing. Tubes were centrifuged for 30 minutes at 2,500 rpms in a centrifuge with a swinging bucket rotor. The plasma layer was removed, and then the remaining buffy coat was poured into a 15 mL conical tube. Next, 5 mL of chilled phosphate buffered saline (PBS) with 2% fetal bovine serum (FBS) was added to the CPT tube, which was capped and inverted to mix. The contents were poured into the same 15 mL conical tube, which was then centrifuged for 10 minutes at 1200 rpms at room temperature. The supernatant was then aspirated and the pellet was resuspended in 1 mL of freezing media (RPMI with 20% DMSO and 20% FBS). The cells were then transferred to a cryovial (Nunc, Roskilde, Denmark) and placed in a freezer box at -80°C. After 24 hours samples were transferred to liquid nitrogen storage. Total RNA from peripheral blood mononuclear cells was isolated using the Ambion Ribopure™-Blood kit (Ambion, Austin, TX). The quantity of RNA was assessed by absorbance at 260 nm by the NanoDrop ND-1000 (Thermo Fisher Scientific, Wilmington, DE). RNA quality was assessed by Agilent 2100 Bioanalyzer RNA 6000 NanoChip (Agilent, Palo Alto, CA). Purified total RNA (2.5 μg) was labeled and hybridized onto Affymetrix U133 Plus 2.0 GeneChip oligonucleotide arrays (Affymetrix, Santa Clara, CA) according to the manufacturer's IVT 1-cycle processing protocol.

4.4.2.3 Development of a human breast tumor predictor

In order to generate a human breast tumor predictor, we utilized the list of significantly differentially expressed genes identified previously from the mouse gene expression data. This list of 4,276 Affymetrix Mouse Genome 430 2.0 probe identifiers was translated into 2,595 orthologous human probes (Affymetrix HU133 Plus 2) and used to filter our human dataset, which was normalized employing principal components

of the Affymetrix control probe sets. We divided our samples into completely independent training and validation sets. The training set consisted of 30 PBMC samples (10 breast cancer patients, 10 patients with benign breast abnormalities and 10 healthy individuals). To minimize potential batch effects, these samples were hybridized in 4 batches that each included samples from all three phenotypic categories: healthy individuals, patients with benign breast abnormalities and patients with breast tumor. We used BFRM to identify 26 factors, or linear combinations of genes, in the training dataset. The 26 factors represent common patterns of expression within the set of 2,595 genes. We utilized the SSS method to identify small subsets of the factors that were predictive of breast tumor status within the training set. Using our training set, we generated 5000 possible predictive models and used model averaging (based on median posterior marginal probability) of the top 200 predictive models. We identified those factors contributing significantly to the model average based on median posterior marginal probability (Figure 15). Then the resulting fitted regression models were used to predict breast tumor status in a separate validation set of PBMC samples. A schematic of our experimental design can be found in Figure 7.

To evaluate the reproducibility of the predictive model, we performed permutation testing and a swap experiment. The permutation test used the same factors generated from our human PBMC training set. However, we randomly permuted (x100) which samples were assigned to the training and validation sets. The "swap experiment" is based on the recommendations of the MicroArray Quality Control (MAQC)-II study (166) and consisted of swapping the training and validation sets, generating new factors from the original validation set and then validating these new factors in the original training set. Next, we assessed the biological relevance of the factors to the phenotype in

question (presence of a breast tumor) by performing two negative controls. First, we used the original factors, but scrambled the phenotypic labels (Normal and Malignant) iteratively, evaluated the predictive models and looked at the distributions of the test characteristics. We performed 200 iterations. Second, we generated a set of mock factors from a publicly available insulin/muscle-biopsy dataset (GSE7146) and projected them into the original training and validation sets. We then compared this predictive model to the human breast tumor predictor generated from the mouse ortholog data. As a validation of our overall approach, we also tested the utility of generating a breast tumor predictor from the human PBMC gene expression data using the same parameters applied to the mouse gene expression data[4].

4.4.2.4 Functional Annotation

To explore the underlying biology of a breast tumor predictor based on peripheral blood gene expression, we analyzed the genes included in the predictor using publicly available annotation tools. To simplify our analysis, we included all 115 unique Affymetrix probes (3 probes were represented in 2 factors each) that comprise the 3 top-performing factors in the predictive model (Figure 15). We used the Affymetrix NetAffx Batch Query tool to translate the 115 HU133 Plus 2.0 Array probe identifiers into 115 gene symbols and descriptions. We used GATHER[3] to input the lists of 83 upregulated and 32 downregulated probe identifiers and generate annotations for gene ontology categories, KEGG pathways, protein binding partners and transcription factor binding sites.

---

[4] For fully versioned and annotated source code and data objects, please see the corresponding QUADRA repositories by visiting http://quadra.genome.duke.edu/mouse-2-human-mammary-tumor-pbmc-project.

4.4.2.5 Gene Set Enrichment Analysis

GSEA (version 2.0.5) was used as previously described. Briefly, we used the gene set annotation feature of GSEA[1] to identify those gene sets that overlap with the upregulated and downregulated genes in the breast tumor predictor. These gene sets offer insights into the functional annotation of our experimentally derived factors and have the potential to reveal the underlying biology of each factor. We used the list of Affymetrix probe identifiers as the input into the browser's query field. After selecting the appropriate identifier platform (Affymetrix HU133 Plus 2.0), we chose to compute the top 10 overlaps with the each of the following categories of gene sets: C1, which represents 326 positional gene sets; C2, which is a collection of 1,892 curated gene sets; C3, which contains 836 gene sets with a shared, putative cis-regulatory element that is conserved across species; C4, which contains 881 computational gene sets that are defined by mining large collections of cancer-oriented microarray data and C5, which contains 1,454 gene sets annotated by a common GO term. In our efforts to annotate the biological function of each factor, we focused on those gene sets that incorporated a large proportion of the genes within the predictor and had the most statistically significant overlap.

4.4.2.6 QUADRA

QUADRA was developed and launched by collaborator and co-author, Dr. Erich Huang. It encompasses both an open source distributed version control system 'Git' (Version 1.7.2.2[5]), an adaptation of the open source Ruby on Rails web hosting platform 'Gitorious[6]' as well as 'good operating principles' in genomic computation. The basic

---

[5] available at www.git-scm.org
[6] www.gitorious.com

principles of QUADRA encompass the following: (1) all data manipulations start with raw data (as obtained from a core facility or GEO), and all manipulations/interactions with the data are performed and versioned within a scientific computing environment, in this case MATLAB (Version 7.9.0.529, R2009b, Mathworks, Natick, MA), (2) the entire analytic flow is recorded and versioned in source code; therefore, all aspects of factor and model generation are performed programmatically and can be read as MATLAB source code, (3) once an analysis is finalized, its "publish" run is recorded as an HTML file so that the scientific community can reconcile the actual analysis with what is reported and the source code; and (4) all source code and data are available to the public[7].

4.4.3 Results

4.4.3.1 Human patient characteristics

The training set contained samples from women diagnosed with invasive breast cancer or benign breast abnormalities, as well as samples from women with no evidence of breast cancer. The entire training set (including Benign, Normal and Malignant samples) was used to generate factors. The predictive model was built based on a comparison of the Normal and Malignant samples. The validation set contained samples from women with normal, benign or malignant diagnoses, as well as samples from patients with a variety of gastrointestinal cancers or brain cancer. The clinical and demographic characteristics of the training and validation patient cohorts are broken down into phenotypic subgroups (Normal and Malignant) and reported in Table 4. These two cohorts did not show any statistically significant differences in terms of age, race,

[7] For the work described here, the code repository is available at http://quadra.genome.duke.edu/mouse-2-human-mammary-tumor-pbmc-project.

gender, diagnosis, BI-RADS score, hormone receptor status or lymph node status. The training cohort consisted of 10 samples from healthy, control individuals (average age of 56 years; range 44-75) and 10 samples from breast cancer patients (average age of 61 years; range 43-85); the validation cohort consisted of 21 samples from healthy, control individuals (average age of 58 years; range 46-80) and 47 samples from breast cancer patients (average age of 59 years; range 32-85). A comparison of patient characteristics based on phenotypic group can be found in Table 5. Notably, we ensured that the patient samples in the training cohort were all processed together to minimize technical batch effects.

Table 4. Patient characteristics by training and validation data set.

Normal (n=31)				
Category		**Training (n=10)**	**Validation (n=21)**	**p-value**
Age (y)				0.5400*
	Mean	55.80	57.81	
	Median	55.00	55.00	
	Max	75.00	80.00	
	Min	44.00	46.00	
Gender				N/A
	Female	20	68	
	Male	0	0	
Race				0.2216†
	White	9	14	
	Other	1	7	
BI-RADS Score				0.2540‡
	0	0	2	
	1	9	13	
	2	1	6	
	3	0	0	
	4	0	0	
	5	0	0	
	6	0	0	
	Incomplete/ Unavailable	0	0	
Malignant (n=57)				
Category		**Training (n=10)**	**Validation (n=47)**	**p-value**
Age (y)				0.6444*
	Mean	61.40	58.68	
	Median	56.00	61.00	
	Max	85.00	85.00	
	Min	43.00	32.00	
Gender				N/A
	Female	10	47	
	Male	0	0	

(Table 4. Continued from previous page.)

Race				1.0000†
	White	9	14	
	Other	1	7	
BI-RADS Score				0.8695‡
	0	0	2	
	1	0	2	
	2	1	2	
	3	0	0	
	4	3	16	
	5	4	12	
	6	0	2	
	Incomplete/ Unavailable	2	11	
HER2				0.2253‡
	Positive	3	5	
	Negative	4	18	
	Unknown	3	24	
ER				0.1036‡
	0 - 2	3	15	
	3 - 8	4	6	
	Unknown	3	26	
PR				0.2844‡
	0 - 2	4	12	
	3 - 8	3	8	
	Unknown	3	27	
Positive Lymph Nodes				0.0685‡
	No LN Sampling	1	23	
	0	5	18	
	1	1	4	
	2	1	1	
	3	0	1	
	9	1	0	

Table 5. Patient characteristics by phenotype

Training (n=20)				
Category	-	**Normal (n=10)**	**Malignant (n=10)**	**p-value**
Age (y)				0.5288*
	Mean	55.80	61.40	
	Median	55.00	56.00	
	Max	75.00	85.00	
	Min	44.00	43.00	
Gender				N/A
	Female	10	10	
	Male	0	0	
Race				0.3049†
	White	9	10	
	Other	1	0	
Validation (n=68)				
Category		**Normal (n=21)**	**Malignant (n=47)**	**p-value**
Age (y)				0.7804*
	Mean	57.81	58.68	
	Median	55.00	61.00	
	Max	80.00	85.00	
	Min	46.00	32.00	
Gender				N/A
	Female	21	47	
	Male	0	0	
Race				0.0028†
	White	14	45	
	Other	7	2	

4.4.3.2 Development of a peripheral blood predictor of human breast tumors

Based on the performance of the mouse mammary tumor predictor, we sought to use these studies as a foundation for developing a predictor of human breast tumors. To develop this predictor, we started with the list of differentially expressed genes derived from the mouse model and translated them into their human gene orthologs. We used this gene list to filter our human gene expression data. Next, we used the same modeling pipeline employed for the mouse model to generate sparse factors and train a predictive model using shotgun stochastic search. In building the models, we again randomly divided the dataset into independent training and validation cohorts, as previously described, and ensured that factors and the predictive model built from them were generated from the training set alone and completely insulated from the validation set.

The performance of the derived predictive model in distinguishing controls from cancer patients is shown in Figure 16 (A-C). The predictor showed robust capacity to distinguish between the two groups in the training set. Moreover, we were able to predict tumor status in the validation set with 89% sensitivity and 100% specificity (as calculated using Youden's J-statistic to obtain a threshold of 0.3760). The diagnostic characteristics of this predictive model were tested by a receiver-operator characteristic (ROC) curve, which demonstrated an area under the curve (AUC) of 0.9696. We obtained similar results from the swap experiment in which we swapped the training and validation sets and completely regenerated the factors and the predictive model (again maintaining strict independence of the switched training and validation sets), indicating the validity of the modeling approach based on the orthologs identified in mouse (Figure 16 D-F). The results of the mock factor analysis (Figure 16 G-I) confirm that a model based on a set of unrelated, biologically orthogonal factors is not sufficient to build a

predictive model of breast tumor status. The mock factors, which were generated from an insulin/muscle-biopsy gene expression dataset (GSE 7146), were unable to distinguish between samples based on breast tumor status in the independent validation set. In addition to the swap experiment and mock factor analysis, we also performed permutation testing to evaluate the reproducibility of the predictive model. We used the same factors generated in the original training set, but randomly assigned samples to either the training or validation sets (100 permutations). The distribution of test characteristics from this analysis are shown in Figure 17 and indicate that we are able to generate a robust predictive model regardless of which samples were assigned to the training and validation sets. Figure 18 shows the results (overlaid with the results in Figure 17 for comparison) of a negative control experiment in which we scrambled the phenotypic labels of the samples but used the same factors generated from the original training set to generate a potential predictive model. The distribution of test characteristics from these analyses show that the factors generated from the original training set is related to phenotype. Finally, to evaluate the possibility of generating a predictor derived purely from human peripheral blood mononuclear cell (PBMC) samples, we performed a sparse ANOVA (using identical parameters as those used with the mouse data) on a training cohort of patient samples with no malignant or benign breast lesions versus patients with malignancy. This yielded a small subset of 81 genes. Employing BFRM as previously described to identify potentially predictive factors, we identified a single factor, which performed no better than the mouse-derived factors.

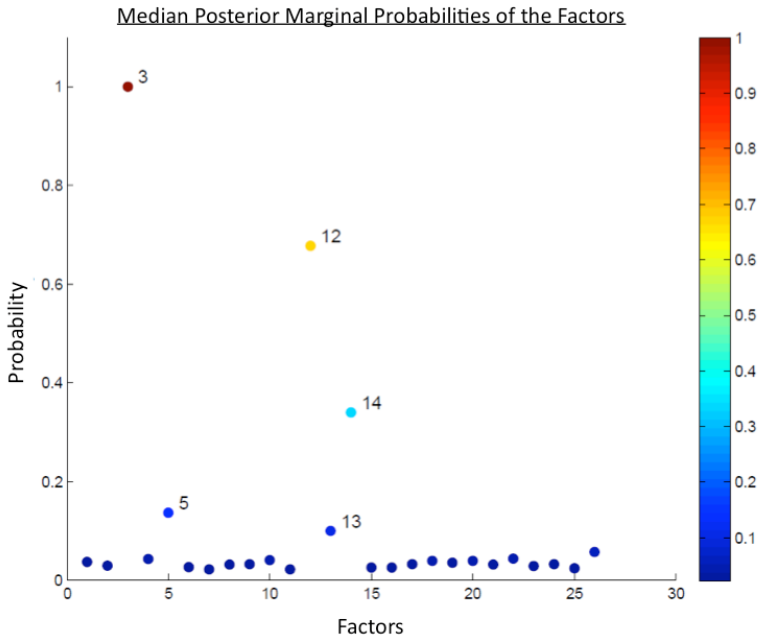
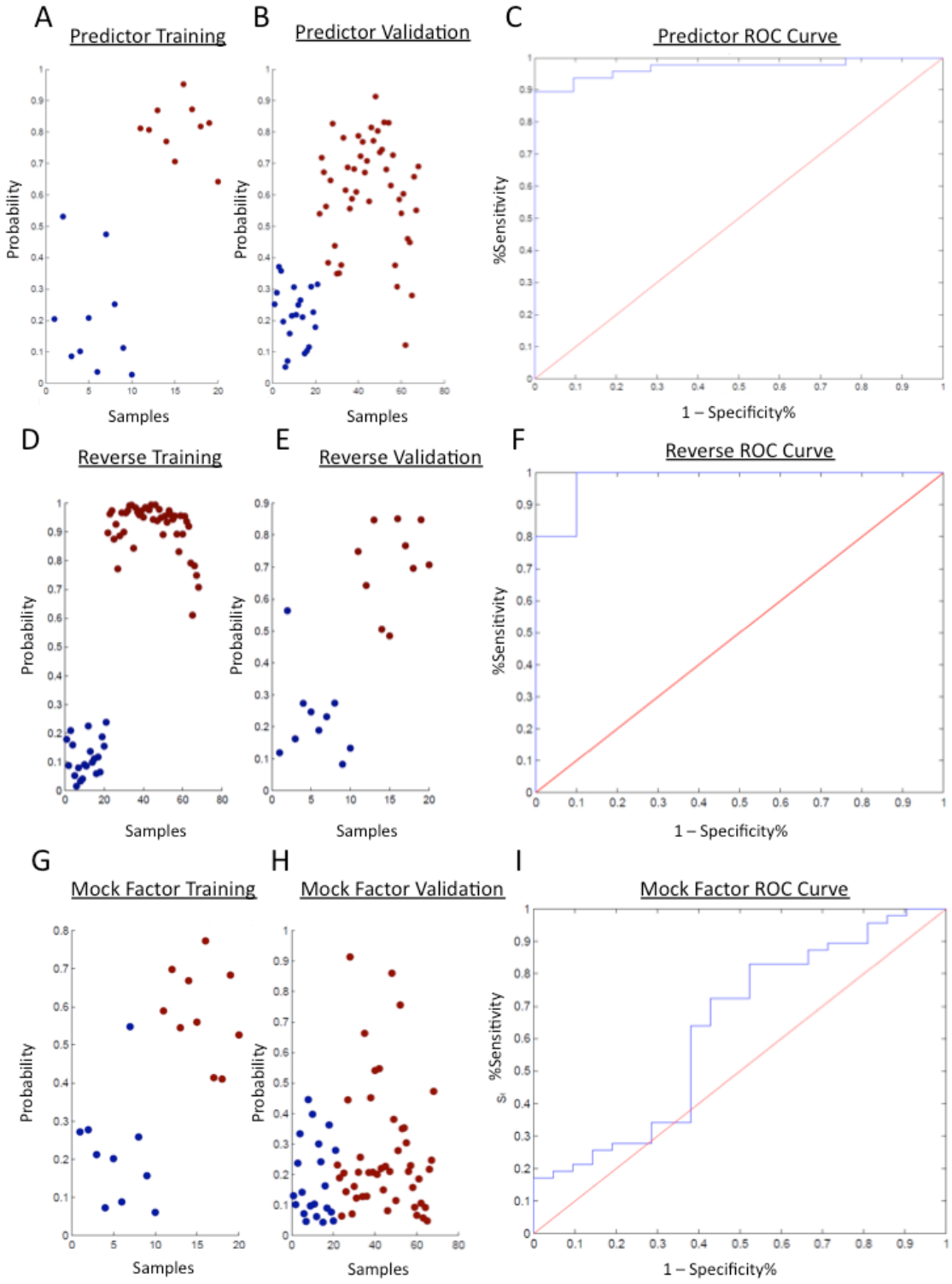


Figure 15. Inclusion probabilities of factors in the predictive model.

We generated a collection of 26 factors from the human PBMC training set using the methods described previously and used SSS to put these together in various combinations to form predictive models (5000 iterations), which were validated in an independent sample set. We calculated the top performing factors based on their inclusion probability (median posterior marginal probability) in the top 200 models. These 26 factors are plotted along the x-axis and the median posterior marginal probabilities are plotted on the y-axis. The top 3 factors with inclusion probabilities significantly above the background noise are 3, 12 and 14.

Figure 16. Gene expression signature predicts human breast cancer.

(See next page.) We generated a human breast cancer predictor from human PBMC samples by using a mouse model of breast cancer to first identify the most informative probes to use in a subsequent factor modeling approach. A) We generated the predictive model from a training set of healthy cancer-free individuals (n = 10) and patients with invasive breast cancer (n = 10). Blue = normal; red = malignant. An assessment of the model fit shows that this predictor has a robust capacity to discriminate among samples based on breast tumor status with 100% sensitivity and specificity. B) We then used this predictive model to evaluate an independent set of samples (n = 162) for the capacity to distinguish controls from patients with a diagnosis of malignant breast cancer. This represents an external validation using samples not used in either the factor generation or the model building process. C) We were able to predict breast cancer status with a sensitivity of 89% and specificity of 100% (AUC = 0.97) as shown in the ROC curve. The optimal threshold was calculated as 0.3760 based on Youden's J-statistic. D) We then tested the validity of our modeling strategy by swapping the training and validation sets. New factors were generated based on the original validation set and new models were generated. The model fit diagram shows the ability to generate a robust model from the original validation set. E) This new model was validated in the original training set. F) It demonstrated a sensitivity of 100% and a specificity of 90% (AUC = 0.98). As a negative control, we generated mock factors from a publicly available dataset that was biologically unrelated to breast tumor status. We then projected these factors into the training (G) and validation (H) sets. I) The sensitivity was 83% and specificity was 48% (AUC = 0.63).


A
Predictor Training
Probability
Samples
B
Predictor Validation
Probability
Samples
C
Predictor ROC Curve
%Sensitivity
1 – Specificity%
D
Reverse Training
Probability
Samples
E
Reverse Validation
Probability
Samples
F
Reverse ROC Curve
%Sensitivity
1 – Specificity%
G
Mock Factor Training
Probability
Samples
H
Mock Factor Validation
Probability
Samples
I
Mock Factor ROC Curve
%Sensitivity
1 – Specificity%

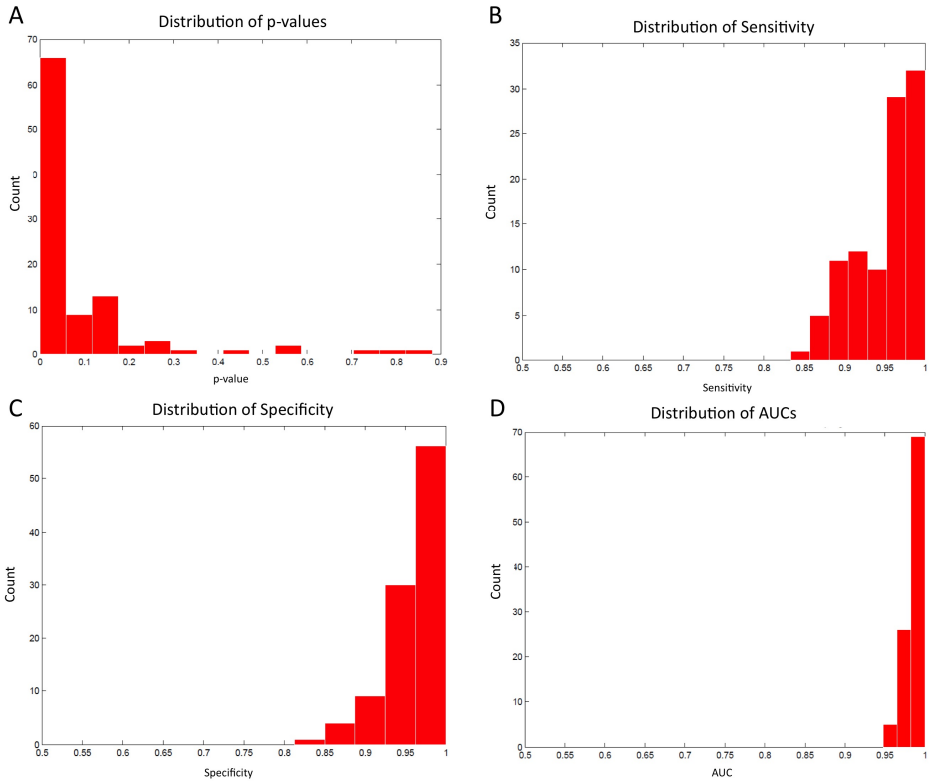


Figure 17. Distribution of test characteristics from data set permutation testing.

We used the factors generated from the original training set, but randomly assigned samples to either the training or validation set (100 permutations) and plotted the distribution of the following test characteristics: p-value (A), sensitivity (B), specificity (C) and AUC (D).

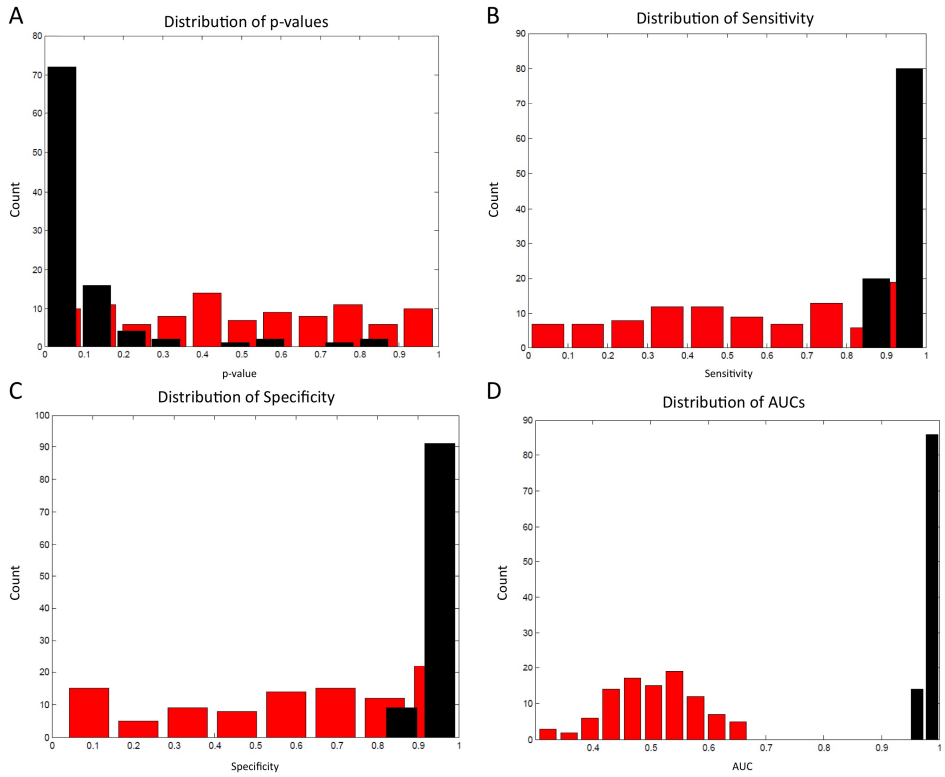


Figure 18. Distribution of test characteristics from phenotype permutation testing.

We used the factors generated from the original training set, but phenotypic labels of the samples were randomly permuted (200 iterations) and plotted the distribution of the following test characteristics: p-value (A), sensitivity (B), specificity (C) and AUC (D). Black = original phenotypes and red = scrambled phenotypes.

4.4.3.3 A use case for employing open source distributed version control and web-hosting tools

With the goal of ensuring that all the methods and conclusions for this study are reproducible, we employed a pilot version control system entitled QUADRA (Figure 19). This system is built on the open source tools "Git" and "Gitorious". The paradigm of QUADRA is that virtually all the computation described here, the intensive factor generation and model building steps, are performed programmatically, and every step of these analyses are securely stored in the system. Therefore, the entire process of generating all the factor models and intermediate data objects (including their version histories) is entered into Git repositories and is "pushed" to a publicly accessible web repository[7], facilitating inspection and downloading of the source code. Further, the final "publication run" of all the scripts is captured in HTML-formatted webpages that are also versioned and can be downloaded from the QUADRA web-host (Figure 20).

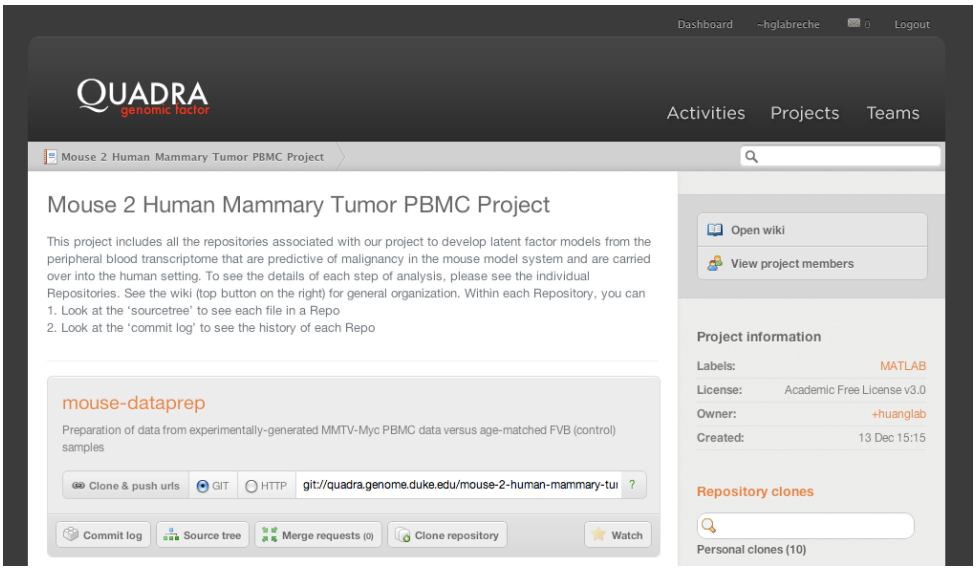


Figure 19. Screenshot of http://quadra.genome.duke.edu/mouse-2-human-mammary-tumor-pbmc-project.

Mouse_Classification_Model.m

Taking the latent factors generated from the training dataset and developing a classification model distinguishing FVB mice from MMTV-Myc mice

Contents

- Authors:
- Files:
- Notes:
- Confirm data object fingerprints
- Load in the data
- Build a model on the training data
- 'Project' factors defined in the TRAINSET into the independent VALIDSET
- Validation of the Shotgun Stochastic Search classification model
- Calculate some test characteristics from the model

Authors:

- Erich S. Huang, MD, PhD
- Heather Labreche, BS
- Institute for Genome Sciences & Policy
- Duke University
- Durham, NC

Files:

1. mF.txt
2. TRAINPHENO.txt
3. TRAINSET.txt
4. VALIDPHENO.txt
5. VALIDSET.txt
6. probeids.txt

```
% These files are carried over from the previous latent factor script
% (mF.txt), or the dataprep script
```

Notes:

This script starts by taking the factors generated in the TRAINSET and creating a SSS classification model based on these factors and validates this on the VALIDSET.

Figure 20. Example of a published script from QUADRA in an HTML-formatted webpage.

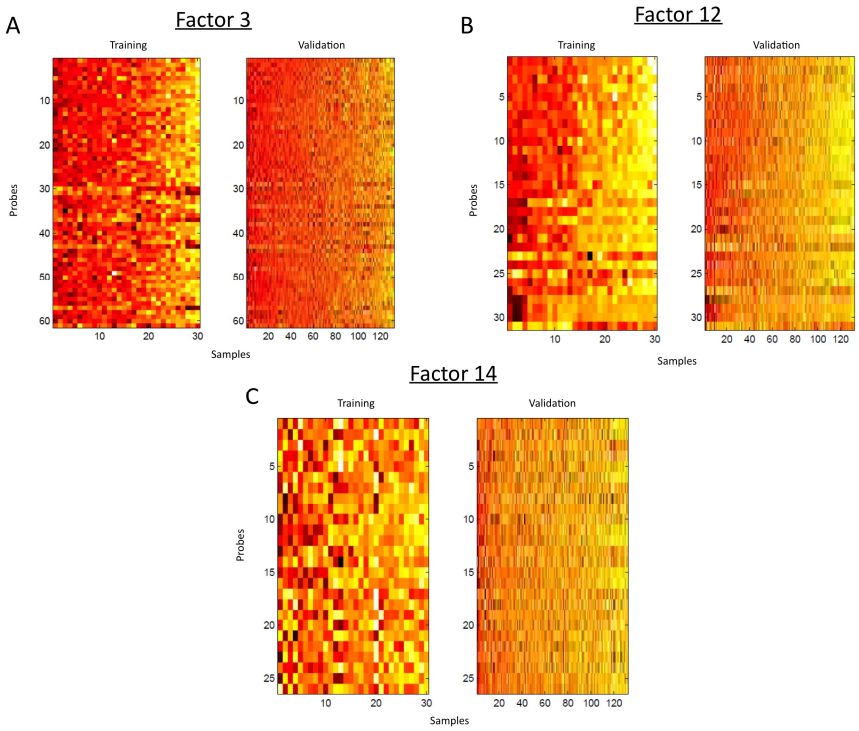


Figure 21. Factor coherence between the training and validation sets.

Each of the top 3 factors that compose the human breast cancer predictor (3, 12 and 14) exhibits coordinated gene expression across the training set (A, C and E). Furthermore, this coordinate expression is recapitulated in the validation set (B, D and F). Each column represents a human PBMC sample. Samples are ordered left to right in descending order of their loading on the 1st principal component. Each row is a gene (probe set) in descending order of correlation. Red = high expression and yellow = low expression.

4.4.3.4 Functional annotation of the genes comprising the breast tumor predictor

The predictor described in this study reflects changes in the transcriptional program of the peripheral blood cells in response to the presence of a breast tumor. However, the mechanism(s) underlying this response are unknown and could be due to either a direct interaction between tumor and immune cells or indirect communication via signaling molecules. In order to better understand the biological underpinnings of this PBC response, we focused on the 3 top-performing factors based on their median posterior marginal probabilities, which were significantly above the level of background noise (Figure 15). These factors, designated Factor 3, Factor 12 and Factor 14, exhibit coordinated expression across the samples in the training set. In addition, the relationship of the genes in the factors is largely conserved in the validation set (Figure 21). These factors are described in terms of their constituent Affymetrix probe identifiers (Supplementary File 3). However, in order to simplify the presentation of these data, we have combined the gene lists of the three top factors in the predictor (3, 12 and 14) for a total of 115 unique probe identifiers (Supplementary File 4). Next, we analyzed this combined gene list based on the subsets of genes that were either upregulated (83 probe identifiers) or downregulated (32 probe identifiers) in the PBMCs of breast cancer patients compared to healthy controls using the various functions on GATHER and the GSEA Molecular Signatures Database. The results of these analyses are included in the Supplementary Files section.

Using the GO function of GATHER, we identified the following gene ontology categories that were significantly enriched in the upregulated subset of genes (p-value≤0.003): I-kappaB kinase/nuclear factor (NF)-kappaB cascade (GO:0007249); protein kinase cascade (GO:0007243); intracellular signaling cascade (GO:0007242);

positive regulation of I-kappaB kinase/ NF-kappaB cascade (GO:0043123); regulation of I-kappaB kinase/NF-kappaB cascade (GO:0043122); pentose-phosphate shunt (GO:0006098); NADPH regeneration (GO:0006740); positive regulation of signal transduction (GO:0009967); NADPH metabolism (GO:0006739). We also identified similar gene ontology categories by calculating the overlap between the upregulated subset of genes within the predictor and the GSEA GO gene sets. Furthermore, the KEGG analysis revealed that this subset of upregulated genes contained a significant number of genes involved in the canonical MAPK (mitogen-activated protein kinase) signaling pathway, which plays a role in NF-kappaB pathway signaling. In addition to the MAPK pathway, KEGG analysis also implicated the pentose phosphate pathway and the Jak-STAT (Janus kinase and signal transducer and activator of transcription) signaling pathway. The pentose phosphate pathway is involved in NADPH production and metabolism. This finding coincides with the initial GATHER GO analysis, which identified NADPH regeneration and metabolism categories as being significantly enriched within the upregulated genes in the predictor. The Jak-STAT signaling pathway is activated in response to interferon, interleukins and cytokines and results in the transcription of genes responsive to a particular STAT. In fact, the Protein Binding function of GATHER identified a significant enrichment of STAT3 binding partners within the upregulated set of genes and GSEA analysis reported significant overlap between the upregulated genes and gene sets related to IL-6 (interleukin 6), which is an activator of STAT3.

Using the TRANSFAC® function of GATHER, we identified several transcription factor binding sites that were significantly enriched in the downregulated subset of genes (p-value≤0.003). In particular, 3 of these transcription factors had putative binding sites

in almost all of the downregulated genes in the predictor: 1) V$CETS1P54_01: c-Ets-1 (p54); 2) V$KROX_Q6; and 3) V$E2F_Q6: E2F. This subset of downregulated genes also contained 5 genes that overlapped with the peripheral blood signature of bladder cancer described by Osman et al. (*149*) (p-value=0.000228).

We also compared the predictive models generated from the original experiment and the swap experiment based on factor annotation. The top factors in the predictive model from the swap experiment consisted of 457 unique Affymetrix probe sets (Figure 22 and Supplementary File 28). Of these, 24 overlapped with the original model (115 probes). These 24 overlapping probe sets, or genes, are implicated in positive regulation of the I-kappaB kinase/NF-kappaB cascade, which strengthens the identification of this particular pathway in the original predictive model.

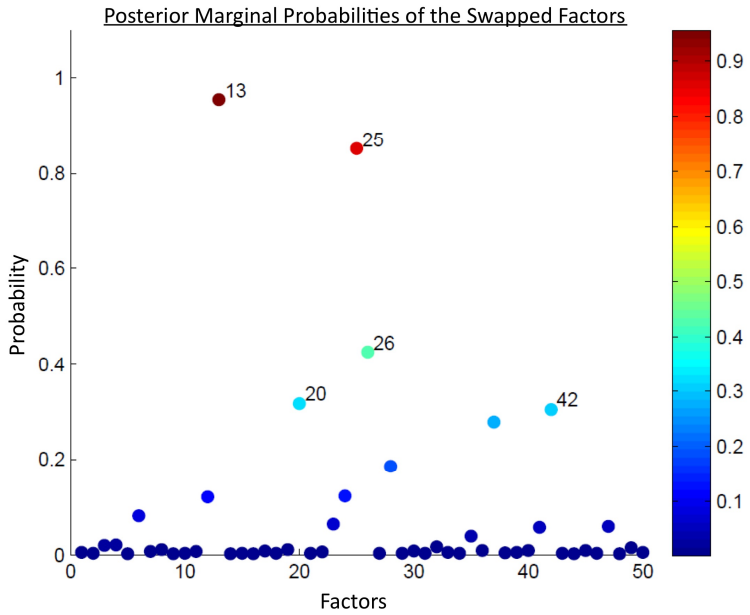


Figure 22. Inclusion probabilities of factors in the swapped model.

We generated a collection of 50 factors from the human PBMC training set using the methods described previously and used SSS to put these together in various combinations to form predictive models (5000 iterations), which were validated in an independent sample set. We calculated the top performing factors based on their inclusion probability (posterior marginal probability) in the top 200 models. These 50 factors are plotted along the x-axis and the median posterior marginal probabilities are plotted on the y-axis. The top 5 factors with inclusion probabilities significantly above the background noise are 13, 25, 26, 20 and 42.

4.4.4 Discussion

Previous work demonstrates the potential for measuring gene expression within peripheral blood as a means of assessing ongoing disease states. This includes the development of genomic analyses used to categorize and treat stroke (167, 168), elucidation of mechanisms of autoimmune syndromes such as lupus (169) and rheumatoid arthritis (*68*) and, more recently, as a measure of viral infection, with the goal of early detection and better management of disease (*67*). Several studies have extended the concept to cancer (*70-73, 75, 76, 149*). Our approach to developing a predictor of human breast cancer makes use of a transgenic mouse model system to first identify those genes most relevant to the presence of a tumor. As an experimental system, the mouse offers several advantages over human studies. It provides a homogeneous genetic background and the ability to control a number of variables, including environmental exposures and the presence of co-morbid conditions. Using this approach, we were able to identify a robust predictor that was capable of predicting tumor status in an independent validation set of mouse peripheral blood samples. We also used the gene expression data generated through these mouse experiments as the starting point for building a predictor of human breast tumors. This predictor proved to be capable of distinguishing breast cancer patients from healthy individuals based on transcriptome patterns in the peripheral blood cells with a sensitivity of 89% and specificity of 100% in an independent validation cohort. This supports earlier findings by other groups (*72, 73*) who described gene expression profiles capable of accurately classifying patient samples based on breast tumor status although it does not overlap with their reported signature (with the exception of a single probe identifier, 221476_s_at, which corresponds to ribosomal protein L15). Notably, our predictor

exhibits a greater sensitivity and specificity than those previously reported and is the first to successfully use a transgenic mouse model to develop a human cancer biomarker. Most importantly, our work represents a conceptual advancement in that we were able to link a mouse model of breast cancer with human breast cancer by applying a factor modeling approach to peripheral blood gene expression data. This suggests that the biological clarity of the mouse model system improves our ability to identify a common underlying biological response to the formation of the tumor. It also provides an experimental framework that can be used to pose new questions that cannot be easily addressed in human studies.

While the studies we report here do appear promising, we recognize that these represent first steps towards the goal of developing a blood-based predictor of human breast cancer. More importantly, we believe the approach described here offers a strategy to address important questions in the development of blood-based cancer diagnostics that are not easily addressed in human studies alone. The use of the mouse model system allows us to develop predictive models that often cannot be derived directly from heterogeneous clinical sample data and could potentially facilitate exploration of the presence of a tumor-associated complex biomarker in mice prior to evidence of a tumor. This approach could prove to be valuable in other contexts as well. A key example can be seen with ovarian cancer where the vast majority of patients are diagnosed with stage IV disease. Given that patients diagnosed with such advanced disease face a poor prognosis, the development of methods for detection at an early stage of the disease process would have a substantial impact on outcome.

Finally, with the complexity of the analytic process described here, we felt that it was essential that all the code and data be made available in a distributed version control

repository that is accessible via a web interface and easily provides for online inspection and downloading of code and data. This online repository is also of significant utility during the development of an analytic process because it affords the freedom to try different approaches to a problem with the knowledge that the code is securely versioned while providing a means to work collaboratively on a complex project. Ultimately, only by building our analysis with this type of infrastructure can we ensure that our work will be useful to the scientific community, as others can learn from, experiment with or modify our source code.

4.5 Characterization of Signature Specificity

4.5.1 Introduction

In order to be useful as a clinical tool, a breast tumor predictor would have to exhibit a certain level of specificity. The exact level of specificity required may be different depending on the application. For instance, if the predictor is used as a screening tool, a lower specificity may be acceptable in order to increase sensitivity and detect as many true positives as possible. Unfortunately, this would result in a higher number of false positives that would have to be ruled out with additional procedures such as imaging or biopsy. On the other hand, if the predictor is used as a diagnostic tool, it is necessary for it to have a very high degree of specificity in order to provide physicians with accurate information on which to base treatment decisions. In this scenario, the blood-based biomarker may even used in conjunction with diagnostic imaging or a biopsy to increase sensitivity. We have already established that peripheral blood gene expression data contain sufficient information to accurately predict the presence of a breast tumor. However, it is not clear whether the transcriptional changes in the PBCs represent a breast tumor-specific signature or a response to a tumor or lesion in general. We hypothesized that the factor modeling approach used in this study may be capable of identifying breast tumor-specific factors, and we sought to test this by applying our predictive model to samples collected from patients with benign breast abnormalities or other types of cancer, such as gastrointestinal and brain cancer.

4.5.2 Materials and Methods

In order to explore the specificity of the signature in terms of cancer type, we used PBMC samples collected (as previously described) from patients with a variety of other cancer types (these included 15 gastrointestinal cancers and 8 brain tumors,

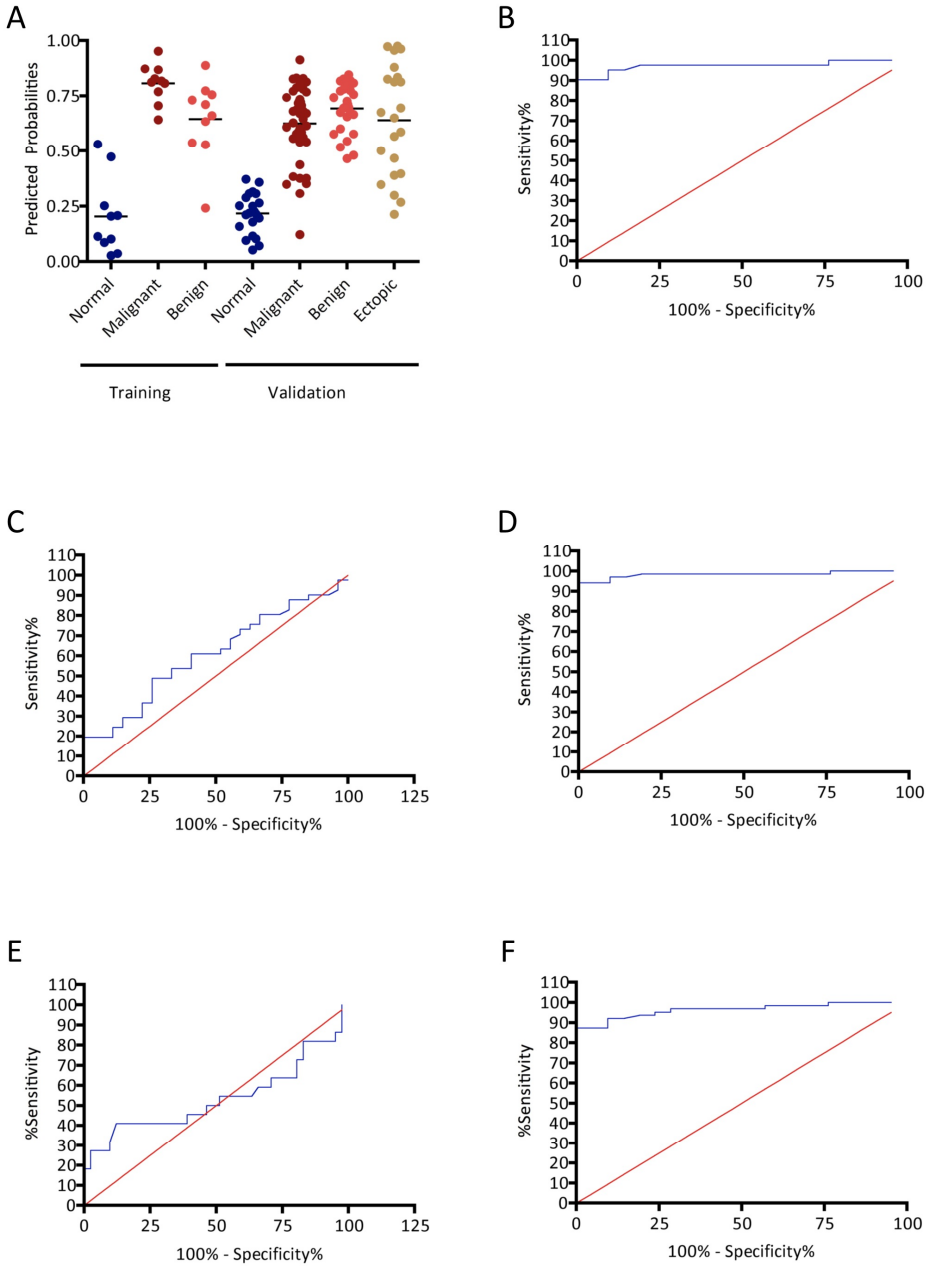
referred to as “Ectopic”). We isolated RNA, then normalized and analyzed gene expression as previously described. It is important to note that these Ectopic samples were not used in the factor generation or model building processes. They remained part of the validation data set only. We applied the breast cancer predictor to these samples as part of the validation set to assess its specificity in terms of cancer type.

4.5.3 Results

As shown in Figure 23C, the breast cancer predictor was unable to distinguish between samples from patients with benign disease and those with invasive breast cancer (AUC = 0.6057). In addition, it did not exhibit specificity in terms of cancer type, as shown in Figure 23E (AUC = 0.5349). The predictor was unable to distinguish between breast cancer and the other cancer types represented in our validation data set (e.g. gastrointestinal and brain). However when evaluated as a predictor of breast lesions (Figure 23D) or cancer (Figure 23F) in general, it performed well. The predictor distinguished between controls and all breast lesions with a sensitivity of 94.12% and a specificity of 100% (AUC = 0.9835); it was able to distinguish between controls and all cancers with a sensitivity of 87.3% and a specificity of 100% (AUC = 0.9633).

Figure 23. Specificity of the breast tumor predictor.

(See opposite page.) A) Predicted probabilities of the training and validation data sets (blue = Normal; red = Malignant; pink = Benign; brown = Ectopic). B) ROC curve showing the sensitivity and specificity of the predictions of Normal and Malignant samples in the validation set. C) ROC curve showing the ability of the predictor to distinguish between Malignant and Benign samples. D) ROC curve showing the ability of the predictor to distinguish between Normal samples and all samples from patients with breast lesions (Malignant and Benign). E) ROC curve showing the ability of the predictor to distinguish between Malignant and Ectopic samples. F) ROC curve showing the ability of the predictor to distinguish between Normal samples and all samples from patients with a diagnosis of cancer (Malignant and Ectopic).


A
Predicted Probabilities
1.00
0.75
0.50
0.25
0.00
Normal
Malignant
Benign
Normal
Malignant
Benign
Ectopic
Training
Validation
B
Sensitivity%
100% - Specificity%
C
Sensitivity%
100% - Specificity%
D
Sensitivity%
100% - Specificity%
E
%Sensitivity
100% - Specificity%
F
%Sensitivity
100% - Specificity%

4.5.4 Discussion

Although the predictor we describe here is capable of distinguishing breast cancer patients from healthy individuals based PBC gene expression patterns, it has proved more difficult to develop a predictor that can discriminate between patients with breast cancer and benign breast abnormalities or between cancer types.

We hypothesize that there may be a number of plausible explanations for our predictor does not discriminate between breast cancer and other types of lesions, including the possibility that these growths all elicit a common transcriptional response in the blood cells. This reduced accuracy could also reflect a possible limitation of the mouse model used to generate our original list of informative genes. Whereas the MMTV/c-MYC mouse model was ideal for generating two extreme phenotypes (advanced mammary cancer compared to the complete absence of disease), it did not allow us to model benign conditions (e.g., cysts or fibroadenomas) or other tumor types. Therefore, future studies should include a wide variety of mouse tumor models and models of benign or "premalignant" lesions, allowing us to capture the full heterogeneity of PBC gene expression and identify those transcript that are most relevant to a specific cancer phenotype. These data could then serve as the starting point to generate factor models in a variety of human cancers. This would require that future studies include clinical samples from many cancer types. These could be used in generating a comprehensive catalogue of factors that reflect the overlapping and unique pathways and genes involved in various cancer types and would allow us to potentially generate cancer-specific predictors. The ability to not only detect cancer, but to also diagnose cancer type based on a peripheral blood sample seems like an idea from science fiction. However, the many examples of blood-based cancer biomarkers discussed here indicate that may not

be so far-fetched. Already, it is possible to distinguish patients with a variety of solid tumors from cancer-free individuals. The next steps will require large collaborative efforts to define blood-based biomarkers for specific cancer types and the ability to share and merge huge datasets.

To begin to address the question of tumor-type specificity, we have initiated a study in conjunction with Dr. John Sampson to collect peripheral blood leukocytes from patients with brain cancer as well as from individuals undergoing microvascular surgery of the brain for benign conditions. These samples can be used in conjunction with animal models of brain cancer to generate a brain tumor predictor similar to that described here for breast cancer. In addition these samples will serve as an independent sample set in which we can validate our breast tumor predictor's specificity and vice versa. The methodology presented here and the infrastructure used to manage and analyze our data make this type of collaborative effort possible.

5. Conclusions

This thesis describes peripheral blood gene expression signatures of Pb exposure and breast cancer. First, we describe robust and dose-specific signatures of Pb exposure in a mouse model. We also use experimentally defined pathway signatures to confirm the dose-specific activation of two pathways previously reported in the literature to respond to Pb exposure. Our results also indicate the ability to detect gene expression changes in the blood at low levels of exposure. These findings will have an important impact on the detection of subclinical Pb exposure, and improve the understanding of the health consequences of low-level exposure. Furthermore, our strategy represents an opportunity and a framework for applying a collection of experimentally derived pathway signatures to toxicogenomic studies as a type of phenotypic anchor. Second, we describe a peripheral blood gene expression signature of human breast tumors. This signature was generated using a well-defined transgenic mammary tumor model as the starting point for identifying gene expression changes associated with the presence of a tumor. This signature accurately predicts the presence of breast tumors in human samples and shows some promise of being used to detect tumors in their early stages in a mouse model. The signature reported here is significant within the field in two ways: first, it has demonstrated higher sensitivity and specificity than previously reported breast cancer signatures; second, it serves as a proof-of-concept for using mouse models of human cancer in the identification of potential clinical biomarkers. These two projects are connected by more than simply methodology and approach. They are united by the concept of the PBC as an indicator of human disease and toxicant exposure. In each project, our goal has been to combine global gene expression profiling with clearly defined experimental systems as a way to link complex phenotypes, such as human

cancer and toxicant exposure, to a molecular phenotype of the PBC response. The major challenges in this approach revolve around questions of sensitivity and specificity. The ability to detect disease and toxicant exposure at levels that are amenable to treatment or intervention represent an opportunity to successfully treat existing conditions and prevent future health problems. This will require fine mapping of gene expression patterns to very subtle phenotypes. Likewise, the ability to successfully diagnose a disease or exposure based on the PBC transcriptome, which is capable of responding to a huge variety of environmental and pathological insults, will require researchers to parse out those responses that are specific to a particular phenotype of interest. This will demand an extensive database of gene expression signatures reflecting the many biological and chemical perturbations that affect the blood.

We believe that the approaches described here represent an experimental framework for understanding the complexity of peripheral blood gene expression data by first defining the phenotype of interest in a controlled experimental system that can be used to better understand the heterogeneity of the peripheral blood cell response to human disease and environmental exposure.

6. Perspectives

We are in the early days of personalized medicine. At this time the promise and hype often seem to outweigh actual results and clinical applications. However, there are potential success stories in the fields of both cancer genomics and toxicogenomics as researchers thoughtfully and creatively go about the painstaking process of understanding human health and disease at the genomic level. It is now possible to envision the world that Garrod describes in which we understand those factors that

contribute to our "predisposition to and immunity from disease." There is no doubt that these factors include not only our genetic makeup, but also the environment in which we live. It is truly this gene-environment interaction that poses the next great obstacle to medical and scientific advancement. Although the projects described in this thesis do not go so far as to elucidate the connection between environmental exposure and cancer, they highlight the approaches and methodologies that will be required to do so.

Genomics provides a common language that can be used by researchers across disciplines to discover overlapping biological mechanisms at work in disease, drug response and toxicant exposure. Canonical cell signaling pathways will have to be reconsidered as we discover their roles in new processes and pathologies. Gene annotations will have to be rewritten to accommodate new knowledge about their functions in normal and diseased cells. To reach the next phase of personalized medicine, scientists and physicians must begin now by building the infrastructure to collect, annotate and process samples, as well as store, analyze and distribute the overwhelming amount of data generated from such studies. The very nature of this enterprise must be collaborative and open-source to enable participation of researchers worldwide. This endeavor will also require a new training paradigm in undergraduate and graduate education. Scientists who hope to be successful in this environment will have to have experience in computational biology and bioinformatics in order to make the most of their data.

This project does not represent a definitive treatise on peripheral blood gene expression profiles of disease and environmental exposure. Rather, it represents the integration of several important tools available to scientists at this time: traditional tools, such as cell culture systems or animal models that aid us in creating clearly defined

phenotypes; genomic tools used to query DNA, RNA, proteins and metabolites at the genome level; and bioinformatics tools that provide a way to find meaningful patterns in this deluge of data, thus creating a new way of understanding human health and disease.

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ. Initial sequencing and analysis of the human genome. Nature 2001 Feb 15;409(6822):860-921.

2. Golub T. Counterpoint: Data first. Nature 2010 Apr 1;464(7289):679.

3. Weinberg R. Point: Hypotheses first. Nature 2010 Apr 1;464(7289):678.

4. Lander ES. Initial impact of the sequencing of the human genome. Nature 2011 Feb 10;470(7333):187-97.

5. Venter JC. Multiple personal genomes await. Nature 2010 Apr 1;464(7289):676-7.

6. Nussbaum RL. Genetics in Medicine. In: McInnes RR, editor. 7 ed. Philadelphia, PA: Saunders Elsevier; 2007.

7. Ginsburg GS, Willard HF. Genomic and personalized medicine: foundations and applications. Transl Res 2009 Dec;154(6):277-87.

8. Jayapal M, Bhattacharjee RN, Melendez AJ, Hande MP. Environmental toxicogenomics: a post-genomic approach to analysing biological responses to environmental toxins. Int J Biochem Cell Biol 2010 Feb;42(2):230-40.

9. Orphanides G. Toxicogenomics: challenges and opportunities. Toxicol Lett 2003 Apr 11;140-141:145-8.

10. Waters MD, Olden K, Tennant RW. Toxicogenomic approach for assessing toxicant-related disease. Mutat Res 2003 Nov;544(2-3):415-24.

11. Afshari CA, Hamadeh HK, Bushel PR. The evolution of bioinformatics in toxicology: advancing toxicogenomics. Toxicol Sci 2011 Mar;120 Suppl 1:S225-37.

12. Kiyosawa N, Manabe S, Sanbuissho A, Yamoto T. Gene set-level network analysis using a toxicogenomics database. Genomics 2010 Jul;96(1):39-49.

13. Kiyosawa N, Manabe S, Yamoto T, Sanbuissho A. Practical application of toxicogenomics for profiling toxicant-induced biological perturbations. Int J Mol Sci 2010;11(9):3397-412.

14. McHale CM, Zhang L, Hubbard AE, Smith MT. Toxicogenomic profiling of chemically exposed humans in risk assessment. Mutat Res 2010 Dec;705(3):172-83.

15. Van Hummelen P, Sasaki J. State-of-the-art genomics approaches in toxicology. Mutat Res 2010 Dec;705(3):165-71.

16. Cui Y, Paules RS. Use of transcriptomics in understanding mechanisms of drug-induced toxicity. Pharmacogenomics 2010 Apr;11(4):573-85.

17. Daston GP, Naciff JM. Predicting developmental toxicity through toxicogenomics. Birth Defects Res C Embryo Today 2010 Jun;90(2):110-7.

18. Zarbl H, Gallo MA, Glick J, Yeung KY, Vouros P. The vanishing zero revisited: thresholds in the age of genomics. Chem Biol Interact 2010 Mar 19;184(1-2):273-8.

19. Waters M, Boorman G, Bushel P, Cunningham M, Irwin R, Merrick A, Olden K, Paules R, Selkirk J, Stasiewicz S, Weis B, Van Houten B, Walker N, Tennant R.

Systems toxicology and the Chemical Effects in Biological Systems (CEBS) knowledge base. EHP Toxicogenomics 2003 Jan;111(1T):15-28.

20. Waters M, Stasiewicz S, Merrick BA, Tomer K, Bushel P, Paules R, Stegman N, Nehls G, Yost KJ, Johnson CH, Gustafson SF, Xirasagar S, Xiao N, Huang CC, Boyer P, Chan DD, Pan Q, Gong H, Taylor J, Choi D, Rashid A, Ahmed A, Howle R, Selkirk J, Tennant R, Fostel J. CEBS--Chemical Effects in Biological Systems: a public data repository integrating study design and toxicity data with microarray and proteomics data. Nucleic Acids Res 2008 Jan;36(Database issue):D892-900.

21. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A, Rocca-Serra P, Sansone SA. ArrayExpress--a public repository for microarray gene expression data at the EBI. Nucleic Acids Res 2003 Jan 1;31(1):68-71.

22. Hayes KR, Vollrath AL, Zastrow GM, McMillan BJ, Craven M, Jovanovich S, Rank DR, Penn S, Walisser JA, Reddy JK, Thomas RS, Bradfield CA. EDGE: a centralized resource for the comparison, analysis, and distribution of toxicogenomic information. Mol Pharmacol 2005 Apr;67(4):1360-8.

23. Dressman HK, Muramoto GG, Chao NJ, Meadows S, Marshall D, Ginsburg GS, Nevins JR, Chute JP. Gene expression signatures that predict radiation exposure in mice and humans. PLoS Med 2007 Apr;4(4):e106.

24. Meadows SK, Dressman HK, Muramoto GG, Himburg H, Salter A, Wei Z, Ginsburg GS, Chao NJ, Nevins JR, Chute JP. Gene expression signatures of radiation response are specific, durable and accurate in mice and humans. PLoS One 2008;3(4):e1912.

25. Meadows SK, Dressman HK, Daher P, Himburg H, Russell JL, Doan P, Chao NJ, Lucas J, Nevins JR, Chute JP. Diagnosis of partial body radiation exposure in mice using peripheral blood gene expression profiles. PLoS One 2010;5(7):e11535.

26. Zhang L, McHale CM, Rothman N, Li G, Ji Z, Vermeulen R, Hubbard AE, Ren X, Shen M, Rappaport SM, North M, Skibola CF, Yin S, Vulpe C, Chanock SJ, Smith MT, Lan Q. Systems biology of human benzene exposure. Chem Biol Interact 2010 Mar 19;184(1-2):86-93.

27. McHale CM, Lan Q, Corso C, Li G, Zhang L, Vermeulen R, Curry JD, Shen M, Turakulov R, Higuchi R, Germer S, Yin S, Rothman N, Smith MT. Chromosome translocations in workers exposed to benzene. J Natl Cancer Inst Monogr 2008(39):74-7.

28. McHale CM, Zhang L, Lan Q, Li G, Hubbard AE, Forrest MS, Vermeulen R, Chen J, Shen M, Rappaport SM, Yin S, Smith MT, Rothman N. Changes in the peripheral blood transcriptome associated with occupational benzene exposure

identified by cross-comparison on two microarray platforms. Genomics 2009 Apr;93(4):343-9.

29. Zhang L, Lan Q, Guo W, Hubbard AE, Li G, Rappaport SM, McHale CM, Shen M, Ji Z, Vermeulen R, Yin S, Rothman N, Smith MT. Chromosome-Wide Aneuploidy Study (CWAS) in Workers Exposed to an Established Leukemogen, Benzene. Carcinogenesis 2011 Jan 7.

30. Ji Z, Zhang L, Peng V, Ren X, McHale CM, Smith MT. A comparison of the cytogenetic alterations and global DNA hypomethylation induced by the benzene metabolite, hydroquinone, with those induced by melphalan and etoposide. Leukemia 2010 May;24(5):986-91.

31. Heijne WH, Jonker D, Stierum RH, van Ommen B, Groten JP. Toxicogenomic analysis of gene expression changes in rat liver after a 28-day oral benzene exposure. Mutat Res 2005 Aug 4;575(1-2):85-101.

32. Hendriksen PJ, Freidig AP, Jonker D, Thissen U, Bogaards JJ, Mumtaz MM, Groten JP, Stierum RH. Transcriptomics analysis of interactive effects of benzene, trichloroethylene and methyl mercury within binary and ternary mixtures on the liver and kidney following subchronic exposure in the rat. Toxicol Appl Pharmacol 2007 Dec 1;225(2):171-88.

33. Smith MT, Zhang L, McHale CM, Skibola CF, Rappaport SM. Benzene, the exposome and future investigations of leukemia etiology. Chem Biol Interact 2011 Feb 17.

34. Ruden DM, Chen L, Possidente D, Possidente B, Rasouli P, Wang L, Lu X, Garfinkel MD, Hirsch HV, Page GP. Genetical toxicogenomics in Drosophila identifies master-modulatory loci that are regulated by developmental exposure to lead. Neurotoxicology 2009 Nov;30(6):898-914.

35. Bouton CM, Hossain MA, Frelin LP, Laterra J, Pevsner J. Microarray analysis of differential gene expression in lead-exposed astrocytes. Toxicol Appl Pharmacol 2001 Oct 1;176(1):34-53.

36. Hossain MA, Bouton CM, Pevsner J, Laterra J. Induction of vascular endothelial growth factor in human astrocytes by lead. Involvement of a protein kinase C/activator protein-1 complex-dependent and hypoxia-inducible factor 1-independent signaling pathway. J Biol Chem 2000 Sep 8;275(36):27874-82.

37. Kasten-Jolly J, Heo Y, Lawrence DA. Impact of developmental lead exposure on splenic factors. Toxicol Appl Pharmacol 2010 Sep 1;247(2):105-15.

38. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D. Molecular portraits of human breast tumours. Nature 2000 Aug 17;406(6797):747-52.

39. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P, Borresen-Dale AL. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A 2001 Sep 11;98(19):10869-74.

40. Rosenwald A, Staudt LM. Clinical translation of gene expression profiling in lymphomas and leukemias. Semin Oncol 2002 Jun;29(3):258-63.

41. Staudt LM, Dave S. The biology of human lymphoid malignancies revealed by gene expression profiling. Adv Immunol 2005;87:163-208.

42. Taylor BS, Barretina J, Socci ND, Decarolis P, Ladanyi M, Meyerson M, Singer S, Sander C. Functional copy-number alterations in cancer. PLoS One 2008;3(9):e3179.

43. Berger MF, Levin JZ, Vijayendran K, Sivachenko A, Adiconis X, Maguire J, Johnson LA, Robinson J, Verhaak RG, Sougnez C, Onofrio RC, Ziaugra L, Cibulskis K, Laine E, Barretina J, Winckler W, Fisher DE, Getz G, Meyerson M, Jaffe DB, Gabriel SB, Lander ES, Dummer R, Gnirke A, Nusbaum C, Garraway LA. Integrative analysis of the melanoma transcriptome. Genome Res 2010 Apr;20(4):413-27.

44. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M, Mc Henry KT, Pinchback RM, Ligon AH, Cho YJ, Haery L, Greulich H, Reich M, Winckler W, Lawrence MS, Weir BA, Tanaka KE, Chiang DY, Bass AJ, Loo A, Hoffman C, Prensner J, Liefeld T, Gao Q, Yecies D, Signoretti S, Maher E, Kaye FJ, Sasaki H, Tepper JE, Fletcher JA, Tabernero J, Baselga J, Tsao MS, Demichelis F, Rubin MA, Janne PA, Daly MJ, Nucera C, Levine RL, Ebert BL, Gabriel S, Rustgi AK, Antonescu CR, Ladanyi M, Letai A, Garraway LA, Loda M, Beer DG, True LD, Okamoto A, Pomeroy SL, Singer S, Golub TR, Lander ES, Getz G, Sellers WR, Meyerson M. The landscape of somatic copy-number alteration across human cancers. Nature 2010 Feb 18;463(7283):899-905.

45. Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP, Alexe G, Lawrence M, O'Kelly M, Tamayo P, Weir BA, Gabriel S, Winckler W, Gupta S, Jakkula L, Feiler HS, Hodgson JG, James CD, Sarkaria JN, Brennan C, Kahn A, Spellman PT, Wilson RK, Speed TP, Gray JW, Meyerson M, Getz G, Perou CM, Hayes DN. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 2010 Jan 19;17(1):98-110.

46. Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, Sboner A, Esgueva R, Pflueger D, Sougnez C, Onofrio R, Carter SL, Park K, Habegger L, Ambrogio L, Fennell T, Parkin M, Saksena G, Voet D, Ramos AH, Pugh TJ, Wilkinson J, Fisher S, Winckler W, Mahan S, Ardlie K, Baldwin J, Simons JW, Kitabayashi N, MacDonald TY, Kantoff PW, Chin L, Gabriel SB,

Gerstein MB, Golub TR, Meyerson M, Tewari A, Lander ES, Getz G, Rubin MA, Garraway LA. The genomic complexity of primary human prostate cancer. Nature 2011 Feb 10;470(7333):214-20.

47. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 2010 Oct;11(10):685-96.

48. Root DE, Hacohen N, Hahn WC, Lander ES, Sabatini DM. Genome-scale loss-of-function screening with a lentiviral RNAi library. Nat Methods 2006 Sep;3(9):715-9.

49. Hahn WC, Dunn IF, Kim SY, Schinzel AC, Firestein R, Guney I, Boehm JS. Integrative genomic approaches to understanding cancer. Biochim Biophys Acta 2009 Jun;1790(6):478-84.

50. Lamb J. The Connectivity Map: a new tool for biomedical research. Nat Rev Cancer 2007 Jan;7(1):54-60.

51. McArt DG, Zhang SD. Identification of candidate small-molecule therapeutics to cancer by gene-signature perturbation in connectivity mapping. PLoS One 2011;6(1):e16382.

52. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 2008 Oct 23;455(7216):1061-8.

53. Pittman J, Huang E, Dressman H, Horng CF, Cheng SH, Tsou MH, Chen CM, Bild A, Iversen ES, Huang AT, Nevins JR, West M. Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. Proc Natl Acad Sci U S A 2004 Jun 1;101(22):8431-6.

54. Huang E, Ishida S, Pittman J, Dressman H, Bild A, Kloos M, D'Amico M, Pestell RG, West M, Nevins JR. Gene expression phenotypic models that predict the activity of oncogenic pathways. Nat Genet 2003 Jun;34(2):226-30.

55. Huang E, West M, Nevins JR. Gene expression profiling for prediction of clinical characteristics of breast cancer. Recent Prog Horm Res 2003;58:55-73.

56. Friedman DR, Weinberg JB, Barry WT, Goodman BK, Volkheimer AD, Bond KM, Chen Y, Jiang N, Moore JO, Gockerman JP, Diehl LF, Decastro CM, Potti A, Nevins JR. A genomic approach to improve prognosis and predict therapeutic response in chronic lymphocytic leukemia. Clin Cancer Res 2009 Nov 15;15(22):6947-55.

57. Mori S, Chang JT, Andrechek ER, Potti A, Nevins JR. Utilization of genomic signatures to identify phenotype-specific drugs. PLoS One 2009;4(8):e6772.

58. Andrechek ER, Cardiff RD, Chang JT, Gatza ML, Acharya CR, Potti A, Nevins JR. Genetic heterogeneity of Myc-induced mammary tumors reflecting diverse

phenotypes including metastatic potential. Proc Natl Acad Sci U S A 2009 Sep 22;106(38):16387-92.

59. Gatza ML, Lucas JE, Barry WT, Kim JW, Wang Q, Crawford MD, Datto MB, Kelley M, Mathey-Prevot B, Potti A, Nevins JR. A pathway-based classification of human breast cancer. Proc Natl Acad Sci U S A 2010 Apr 13;107(15):6994-9.

60. Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, Joshi MB, Harpole D, Lancaster JM, Berchuck A, Olson JA, Jr., Marks JR, Dressman HK, West M, Nevins JR. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 2006 Jan 19;439(7074):353-7.

61. Bild AH, Parker JS, Gustafson AM, Acharya CR, Hoadley KA, Anders C, Marcom PK, Carey LA, Potti A, Nevins JR, Perou CM. An integration of complementary strategies for gene-expression analysis to reveal novel therapeutic opportunities for breast cancer. Breast Cancer Res 2009;11(4):R55.

62. Nevins JR, Potti A. Mining gene expression profiles: expression signatures as cancer phenotypes. Nat Rev Genet 2007 Aug;8(8):601-9.

63. Zucker-Franklin D. Atlas of blood cells : function and pathology. 2nd ed. Milano Philadelphia: Edi. Ermes; Lea & Febiger; 1988.

64. Rockett JC, Burczynski ME, Fornace AJ, Herrmann PC, Krawetz SA, Dix DJ. Surrogate tissue analysis: monitoring toxicant exposure and health status of inaccessible tissues through the analysis of accessible tissues and cells. Toxicol Appl Pharmacol 2004 Jan 15;194(2):189-99.

65. Boldrick JC, Alizadeh AA, Diehn M, Dudoit S, Liu CL, Belcher CE, Botstein D, Staudt LM, Brown PO, Relman DA. Stereotyped and specific gene expression programs in human innate immune responses to bacteria. Proc Natl Acad Sci U S A 2002 Jan 22;99(2):972-7.

66. Rubins KH, Hensley LE, Jahrling PB, Whitney AR, Geisbert TW, Huggins JW, Owen A, Leduc JW, Brown PO, Relman DA. The host response to smallpox: analysis of the gene expression program in peripheral blood cells in a nonhuman primate model. Proc Natl Acad Sci U S A 2004 Oct 19;101(42):15190-5.

67. Zaas AK, Chen M, Varkey J, Veldman T, Hero AO, 3rd, Lucas J, Huang Y, Turner R, Gilbert A, Lambkin-Williams R, Oien NC, Nicholson B, Kingsmore S, Carin L, Woods CW, Ginsburg GS. Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans. Cell Host Microbe 2009 Sep 17;6(3):207-17.

68. Batliwalla FM, Baechler EC, Xiao X, Li W, Balasubramanian S, Khalili H, Damle A, Ortmann WA, Perrone A, Kantor AB, Gulko PS, Kern M, Furie R, Behrens TW, Gregersen PK. Peripheral blood gene expression profiling in rheumatoid arthritis. Genes Immun 2005 Aug;6(5):388-97.

69. Borovecki F, Lovrecic L, Zhou J, Jeong H, Then F, Rosas HD, Hersch SM, Hogarth P, Bouzou B, Jensen RV, Krainc D. Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease. Proc Natl Acad Sci U S A 2005 Aug 2;102(31):11023-8.

70. Twine NC, Stover JA, Marshall B, Dukart G, Hidalgo M, Stadler W, Logan T, Dutcher J, Hudes G, Dorner AJ, Slonim DK, Trepicchio WL, Burczynski ME. Disease-associated expression profiles in peripheral blood mononuclear cells from patients with advanced renal cell carcinoma. Cancer Res 2003 Sep 15;63(18):6069-75.

71. Xu T, Shu CT, Purdom E, Dang D, Ilsley D, Guo Y, Weber J, Holmes SP, Lee PP. Microarray analysis reveals differences in gene expression of circulating CD8(+) T cells in melanoma patients and healthy donors. Cancer Res 2004 May 15;64(10):3661-7.

72. Aaroe J, Lindahl T, Dumeaux V, Saebo S, Tobin D, Hagen N, Skaane P, Lonneborg A, Sharma P, Borresen-Dale AL. Gene expression profiling of peripheral blood cells for early detection of breast cancer. Breast Cancer Res 2010;12(1):R7.

73. Sharma P, Sahni NS, Tibshirani R, Skaane P, Urdal P, Berghagen H, Jensen M, Kristiansen L, Moen C, Zaka A, Arnes J, Sauer T, Akslen LA, Schlichting E, Borresen-Dale AL, Lonneborg A. Early detection of breast cancer based on gene-expression patterns in peripheral blood cells. Breast Cancer Res 2005;7(5):R634-44.

74. Burczynski ME, Twine NC, Dukart G, Marshall B, Hidalgo M, Stadler WM, Logan T, Dutcher J, Hudes G, Trepicchio WL, Strahs A, Immermann F, Slonim DK, Dorner AJ. Transcriptional profiles in peripheral blood mononuclear cells prognostic of clinical outcomes in patients with advanced renal cell carcinoma. Clin Cancer Res 2005 Feb 1;11(3):1181-9.

75. Han M, Liew CT, Zhang HW, Chao S, Zheng R, Yip KT, Song ZY, Li HM, Geng XP, Zhu LX, Lin JJ, Marshall KW, Liew CC. Novel blood-based, five-gene biomarker set for the detection of colorectal cancer. Clin Cancer Res 2008 Jan 15;14(2):455-60.

76. Showe MK, Vachani A, Kossenkov AV, Yousef M, Nichols C, Nikonova EV, Chang C, Kucharczuk J, Tran B, Wakeam E, Yie TA, Speicher D, Rom WN, Albelda S, Showe LC. Gene expression profiles in peripheral blood mononuclear cells can distinguish patients with non-small cell lung cancer from patients with nonmalignant lung disease. Cancer Res 2009 Dec 15;69(24):9202-10.

77. DePrimo SE, Wong LM, Khatry DB, Nicholas SL, Manning WC, Smolich BD, O'Farrell AM, Cherrington JM. Expression profiling of blood samples from an SU5416 Phase III metastatic colorectal cancer clinical trial: a novel strategy for biomarker identification. BMC Cancer 2003 Feb 7;3:3.

78. Critchley-Thorne RJ, Yan N, Nacu S, Weber J, Holmes SP, Lee PP. Down-regulation of the interferon signaling pathway in T lymphocytes from patients with metastatic melanoma. PLoS Med 2007 May;4(5):e176.

79. Fu H, Boffetta P. Cancer and occupational exposure to inorganic lead compounds: a meta-analysis of published data. Occup Environ Med 1995 Feb;52(2):73-81.

80. Steenland K, Boffetta P. Lead and cancer in humans: where are we now? Am J Ind Med 2000 Sep;38(3):295-9.

81. Silbergeld EK, Waalkes M, Rice JM. Lead as a carcinogen: experimental evidence and mechanisms of action. Am J Ind Med 2000 Sep;38(3):316-23.

82. Lustberg M, Silbergeld E. Blood lead levels and mortality. Arch Intern Med 2002 Nov 25;162(21):2443-9.

83. Overall evaluations of carcinogenicity: an updating of IARC Monographs volumes 1 to 42. IARC Monogr Eval Carcinog Risks Hum Suppl 1987;7:1-440.

84. Inorganic and organic lead compounds. IARC Monogr Eval Carcinog Risks Hum 2006;87:1-471.

85. Toxicological Profile for Lead. In: Registry AfTSaD, editor. Atlanta, GA: U.S. Department of health and Human Services, ATSDR; 2007.

86. Garcia-Leston J, Mendez J, Pasaro E, Laffon B. Genotoxic effects of lead: an updated review. Environ Int 2010 Aug;36(6):623-36.

87. Johnson FM. The genetic effects of environmental lead. Mutat Res 1998 Apr;410(2):123-40.

88. Report on Carcinogens, Eleventh Edition. In: U.S. Department of Health and Human Services PHS, National Toxicology Program, editor.2003.

89. Ledda-Columbano GM, Columbano A, Pani P. Lead and liver cell proliferation. Effect of repeated administrations. Am J Pathol 1983 Dec;113(3):315-20.

90. Columbano A, Ledda-Columbano GM, Ennas MG, Curto M, Chelo A, Pani P. Cell proliferation and promotion of rat liver carcinogenesis: different effect of hepatic regeneration and mitogen induced hyperplasia on the development of enzyme-altered foci. Carcinogenesis 1990 May;11(5):771-6.

91. Roy NK, Rossman TG. Mutagenesis and comutagenesis by lead compounds. Mutat Res 1992 Dec;298(2):97-103.

92. Hartwig A. Role of DNA repair inhibition in lead- and cadmium-induced genotoxicity: a review. Environ Health Perspect 1994 Sep;102 Suppl 3:45-50.

93. Fischer AB, Skreb Y. In vitro toxicology of heavy metals using mammalian cells: an overview of collaborative research data. Arh Hig Rada Toksikol 2001 Nov;52(3):333-54.

94. Interpreting and Managing Blood Lead Levels <10 µg/dL in Children and Reducing Childhood Exposures to Lead: Recommendations of CDC's Advisory Committee on Childhood Lead Poisoning Prevention. In: Prevention CfDCa, editor. Atlanta, GA: U.S. Department of Health and Human Services, CDC; 2007.

95. Barbosa F, Jr., Tanus-Santos JE, Gerlach RF, Parsons PJ. A critical review of biomarkers used for monitoring human exposure to lead: advantages, limitations, and future needs. Environ Health Perspect 2005 Dec;113(12):1669-74.

96. Bellinger D, Dietrich KN. Low-level lead exposure and cognitive function in children. Pediatr Ann 1994 Nov;23(11):600-5.

97. Bellinger DC. Very low lead exposures and children's neurodevelopment. Curr Opin Pediatr 2008 Apr;20(2):172-7.

98. Iavicoli I, Carelli G, Stanek EJ, 3rd, Castellino N, Calabrese EJ. Effects of per os lead acetate administration on mouse hepatocyte survival. Toxicol Lett 2002 Mar 24;129(1-2):143-9.

99. West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JA, Jr., Marks JR, Nevins JR. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci U S A 2001 Sep 25;98(20):11462-7.

100. Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP. GSEA-P: a desktop application for Gene Set Enrichment Analysis. Bioinformatics 2007 Dec 1;23(23):3251-3.

101. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005 Oct 25;102(43):15545-50.

102. Qian Y, Tiffany-Castiglioni E. Lead-induced endoplasmic reticulum (ER) stress responses in the nervous system. Neurochem Res 2003 Jan;28(1):153-62.

103. Loikkanen J, Chvalova K, Naarala J, Vahakangas KH, Savolainen KM. Pb2+-induced toxicity is associated with p53-independent apoptosis and enhanced by glutamate in GT1-7 neurons. Toxicol Lett 2003 Sep 30;144(2):235-46.

104. Lidsky TI, Schneider JS. Lead neurotoxicity in children: basic mechanisms and clinical correlates. Brain 2003 Jan;126(Pt 1):5-19.

105. Oberto A, Marks N, Evans HL, Guidotti A. Lead (Pb+2) promotes apoptosis in newborn rat cerebellar neurons: pathological implications. J Pharmacol Exp Ther 1996 Oct;279(1):435-42.

106. Damstra T. Toxicological properties of lead. Environ Health Perspect 1977 Aug;19:297-307.

107. Hirsch HV, Possidente D, Possidente B. Pb2+: an endocrine disruptor in Drosophila? Physiol Behav 2010 Feb 9;99(2):254-9.

108. Iavicoli I, Fontana L, Bergamaschi A. The effects of metals as endocrine disruptors. J Toxicol Environ Health B Crit Rev 2009 Mar;12(3):206-23.

109. Lockitch G. Perspectives on lead toxicity. Clin Biochem 1993 Oct;26(5):371-81.

110. Iavicoli I, Carelli G, Stanek EJ, 3rd, Castellino N, Calabrese EJ. Below background levels of blood lead impact cytokine levels in male and female mice. Toxicol Appl Pharmacol 2006 Jan 1;210(1-2):94-9.

111. Heo Y, Mondal TK, Gao D, Kasten-Jolly J, Kishikawa H, Lawrence DA. Posttranscriptional inhibition of interferon-gamma production by lead. Toxicol Sci 2007 Mar;96(1):92-100.

112. Fujiwara Y, Kaji T, Yamamoto C, Sakamoto M, Kozuka H. Stimulatory effect of lead on the proliferation of cultured vascular smooth-muscle cells. Toxicology 1995 Apr 12;98(1-3):105-10.

113. Hitzfeld B, Planas-Bohne F, Taylor D. The effect of lead on protein and DNA metabolism of normal and lead-adapted rat kidney cells in culture. Biol Trace Elem Res 1989 Jul-Sep;21:87-95.

114. Razani-Boroujerdi S, Edwards B, Sopori ML. Lead stimulates lymphocyte proliferation through enhanced T cell-B cell interaction. J Pharmacol Exp Ther 1999 Feb;288(2):714-9.

115. Lu H, Guizzetti M, Costa LG. Inorganic lead stimulates DNA synthesis in human astrocytoma cells: role of protein kinase Calpha. J Neurochem 2001 Aug;78(3):590-9.

116. Choie DD, Richter GW. Cell proliferation in mouse kidney induced by lead. I. Synthesis of deoxyribonucleic acid. Lab Invest 1974 May;30(5):647-51.

117. Liu JY, Lin JK, Liu CC, Chen WK, Liu CP, Wang CJ, Yen CC, Hsieh YS. Augmentation of protein kinase C activity and liver cell proliferation in lead nitrate-treated rats. Biochem Mol Biol Int 1997 Oct;43(2):355-64.

118. Altekruse SF KC, Krapcho M, Neyman N, Aminou R, Waldron W, Ruhl J, Howlader N, Tatalovich Z, Cho H, Mariotto A, Eisner MP, Lewis DR, Cronin K, Chen HS, Feuer EJ, Stinchcomb DG, Edwards BK. SEER Cancer Statistics

Review, 1975-2007. Bethesda, MD2010 [cited 2010 June 15]; based on November 2009 SEER data submission, posted to the SEER website]. Available from: http://seer.cancer.gov/csr/1975-2006/.

119. Garfinkel L, Boring CC, Heath CW, Jr. Changing trends. An overview of breast cancer incidence and mortality. Cancer 1994 Jul 1;74(1 Suppl):222-7.

120. Kerlikowske K, Grady D, Barclay J, Sickles EA, Ernster V. Likelihood ratios for modern screening mammography. Risk of breast cancer based on age and mammographic interpretation. JAMA 1996 Jul 3;276(1):39-43.

121. Rosenberg RD, Hunt WC, Williamson MR, Gilliland FD, Wiest PW, Kelsey CA, Key CR, Linver MN. Effects of age, breast density, ethnicity, and estrogen replacement therapy on screening mammographic sensitivity and cancer stage at diagnosis: review of 183,134 screening mammograms in Albuquerque, New Mexico. Radiology 1998 Nov;209(2):511-8.

122. Zahl PH, Strand BH, Maehlen J. Incidence of breast cancer in Norway and Sweden during introduction of nationwide screening: prospective cohort study. BMJ 2004 Apr 17;328(7445):921-4.

123. Duffy SW, Tabar L, Smith RA. Screening for breast cancer with mammography. Lancet 2001 Dec 22-29;358(9299):2166; author reply 7-8.

124. Olsen O, Gotzsche PC. Cochrane review on screening for breast cancer with mammography. Lancet 2001 Oct 20;358(9290):1340-2.

125. Jemal A, Siegel R, Ward E, Murray T, Xu J, Smigal C, Thun MJ. Cancer statistics, 2006. CA Cancer J Clin 2006 Mar-Apr;56(2):106-30.

126. Teh W, Wilson AR. The role of ultrasound in breast cancer screening. A consensus statement by the European Group for Breast Cancer Screening. Eur J Cancer 1998 Mar;34(4):449-50.

127. Lawrence WF, Liang W, Mandelblatt JS, Gold KF, Freedman M, Ascher SM, Trock BJ, Chang P. Serendipity in diagnostic imaging: magnetic resonance imaging of the breast. J Natl Cancer Inst 1998 Dec 2;90(23):1792-800.

128. Weinreb JC, Newstead G. MR imaging of the breast. Radiology 1995 Sep;196(3):593-610.

129. Zhao S, Liu Y, Zhang Q, Li H, Zhang M, Ma W, Zhao W, Wang J, Yang M. The prognostic role of circulating tumor cells (CTCs) detected by RT-PCR in breast cancer: a meta-analysis of published literature. Breast Cancer Res Treat 2011 Feb 11.

130. Cristofanilli M, Budd GT, Ellis MJ, Stopeck A, Matera J, Miller MC, Reuben JM, Doyle GV, Allard WJ, Terstappen LW, Hayes DF. Circulating tumor cells, disease

progression, and survival in metastatic breast cancer. N Engl J Med 2004 Aug 19;351(8):781-91.

131. Cristofanilli M, Hayes DF, Budd GT, Ellis MJ, Stopeck A, Reuben JM, Doyle GV, Matera J, Allard WJ, Miller MC, Fritsche HA, Hortobagyi GN, Terstappen LW. Circulating tumor cells: a novel prognostic factor for newly diagnosed metastatic breast cancer. J Clin Oncol 2005 Mar 1;23(7):1420-30.

132. Budd GT, Cristofanilli M, Ellis MJ, Stopeck A, Borden E, Miller MC, Matera J, Repollet M, Doyle GV, Terstappen LW, Hayes DF. Circulating tumor cells versus imaging--predicting overall survival in metastatic breast cancer. Clin Cancer Res 2006 Nov 1;12(21):6403-9.

133. Warner E, Plewes DB, Hill KA, Causer PA, Zubovits JT, Jong RA, Cutrara MR, DeBoer G, Yaffe MJ, Messner SJ, Meschino WS, Piron CA, Narod SA. Surveillance of BRCA1 and BRCA2 mutation carriers with magnetic resonance imaging, ultrasound, mammography, and clinical breast examination. JAMA 2004 Sep 15;292(11):1317-25.

134. Duffy MJ. CA 15-3 and related mucins as circulating markers in breast cancer. Ann Clin Biochem 1999 Sep;36 (Pt 5):579-86.

135. Duffy MJ. Serum tumor markers in breast cancer: are they of clinical value? Clin Chem 2006 Mar;52(3):345-51.

136. Duffy MJ, Shering S, Sherry F, McDermott E, O'Higgins N. CA 15-3: a prognostic marker in breast cancer. Int J Biol Markers 2000 Oct-Dec;15(4):330-3.

137. Sidransky D. Emerging molecular markers of cancer. Nat Rev Cancer 2002 Mar;2(3):210-9.

138. Mori S, Chang JT, Andrechek ER, Matsumura N, Baba T, Yao G, Kim JW, Gatza M, Murphy S, Nevins JR. Anchorage-independent cell growth signature identifies tumors with metastatic potential. Oncogene 2009 Aug 6;28(31):2796-805.

139. Shats I, Gatza ML, Chang JT, Mori S, Wang J, Rich J, Nevins JR. Using a stem cell-based signature to guide therapeutic selection in cancer. Cancer Res 2011 Mar 1;71(5):1772-80.

140. Impact of follow-up testing on survival and health-related quality of life in breast cancer patients. A multicenter randomized controlled trial. The GIVIO Investigators. JAMA 1994 May 25;271(20):1587-92.

141. Liberati A. The GIVIO trial on the impact of follow-up care on survival and quality of life in breast cancer patients. Interdisciplinary Group for Cancer Care Evaluation. Ann Oncol 1995;6 Suppl 2:41-6.

142. Rosselli Del Turco M, Palli D, Cariddi A, Ciatto S, Pacini P, Distante V. Intensive diagnostic follow-up after treatment of primary breast cancer. A randomized

trial. National Research Council Project on Breast Cancer follow-up. JAMA 1994 May 25;271(20):1593-7.

143. Elmore JG, Armstrong K, Lehman CD, Fletcher SW. Screening for breast cancer. JAMA 2005 Mar 9;293(10):1245-56.

144. Bast RC, Jr., Ravdin P, Hayes DF, Bates S, Fritsche H, Jr., Jessup JM, Kemeny N, Locker GY, Mennel RG, Somerfield MR. 2000 update of recommendations for the use of tumor markers in breast and colorectal cancer: clinical practice guidelines of the American Society of Clinical Oncology. J Clin Oncol 2001 Mar 15;19(6):1865-78.

145. Cheung KL, Graves CR, Robertson JF. Tumour marker measurements in the diagnosis and monitoring of breast cancer. Cancer Treat Rev 2000 Apr;26(2):91-102.

146. Nicolini A, Carpi A. Postoperative follow-up of breast cancer patients: overview and progress in the use of tumor markers. Tumour Biol 2000 Jul-Aug;21(4):235-48.

147. Stathopoulou A, Vlachonikolis I, Mavroudis D, Perraki M, Kouroussis C, Apostolaki S, Malamos N, Kakolyris S, Kotsakis A, Xenidis N, Reppa D, Georgoulias V. Molecular detection of cytokeratin-19-positive cells in the peripheral blood of patients with operable breast cancer: evaluation of their prognostic significance. J Clin Oncol 2002 Aug 15;20(16):3404-12.

148. Slade MJ, Smith BM, Sinnett HD, Cross NC, Coombes RC. Quantitative polymerase chain reaction for the detection of micrometastases in patients with breast cancer. J Clin Oncol 1999 Mar;17(3):870-9.

149. Osman I, Bajorin DF, Sun TT, Zhong H, Douglas D, Scattergood J, Zheng R, Han M, Marshall KW, Liew CC. Novel blood biomarkers of human urinary bladder cancer. Clin Cancer Res 2006 Jun 1;12(11 Pt 1):3374-80.

150. Huang E, Cheng SH, Dressman H, Pittman J, Tsou MH, Horng CF, Bild A, Iversen ES, Liao M, Chen CM, West M, Nevins JR, Huang AT. Gene expression predictors of breast cancer outcomes. Lancet 2003 May 10;361(9369):1590-6.

151. Kamangar F, Dores GM, Anderson WF. Patterns of cancer incidence, mortality, and prevalence across five continents: defining priorities to reduce cancer disparities in different geographic regions of the world. J Clin Oncol 2006 May 10;24(14):2137-50.

152. Weedon-Fekjaer H, Lindqvist BH, Vatten LJ, Aalen OO, Tretli S. Breast cancer tumor growth estimated through mammography screening data. Breast Cancer Res 2008;10(3):R41.

153. Porter PL, El-Bastawissi AY, Mandelson MT, Lin MG, Khalid N, Watney EA, Cousens L, White D, Taplin S, White E. Breast tumor characteristics as predictors

of mammographic detection: comparison of interval- and screen-detected cancers. J Natl Cancer Inst 1999 Dec 1;91(23):2020-8.

154. Kolb TM, Lichy J, Newhouse JH. Comparison of the performance of screening mammography, physical examination, and breast US and evaluation of factors that influence them: an analysis of 27,825 patient evaluations. Radiology 2002 Oct;225(1):165-75.

155. Amir E, Freedman OC, Seruga B, Evans DG. Assessing women at high risk of breast cancer: a review of risk assessment models. J Natl Cancer Inst 2010 May 19;102(10):680-91.

156. Shen Q, Brown PH. Transgenic mouse models for the prevention of breast cancer. Mutat Res 2005 Aug 25;576(1-2):93-110.

157. D'Cruz CM, Gunther EJ, Boxer RB, Hartman JL, Sintasath L, Moody SE, Cox JD, Ha SI, Belka GK, Golant A, Cardiff RD, Chodosh LA. c-MYC induces mammary tumorigenesis by means of a preferred pathway involving spontaneous Kras2 mutations. Nat Med 2001 Feb;7(2):235-9.

158. Cardiff RD, Anver MR, Gusterson BA, Hennighausen L, Jensen RA, Merino MJ, Rehm S, Russo J, Tavassoli FA, Wakefield LM, Ward JM, Green JE. The mammary pathology of genetically engineered mice: the consensus report and recommendations from the Annapolis meeting. Oncogene 2000 Feb 21;19(8):968-88.

159. Desai KV, Xiao N, Wang W, Gangi L, Greene J, Powell JI, Dickson R, Furth P, Hunter K, Kucherlapati R, Simon R, Liu ET, Green JE. Initiating oncogenic event determines gene-expression patterns of human breast cancer models. Proc Natl Acad Sci U S A 2002 May 14;99(10):6967-72.

160. Lucas J, Carvalho C, West M. A bayesian analysis strategy for cross-study translation of gene expression biomarkers. Stat Appl Genet Mol Biol 2009 Jan;8(1):Article 11.

161. Carvalho CM, Lucas, J.E., Wang, Q., Chang, J., Nevins, J.R., West, M. High-dimensional sparse factor modelling - Applicaitons in gene expression genomics. Journal of the American Statistical Association 2008 2008;103(484):1438-56.

162. Hans M, Luvisetto S, Williams ME, Spagnolo M, Urrutia A, Tottene A, Brust PF, Johnson EC, Harpold MM, Stauderman KA, Pietrobon D. Functional consequences of mutations in the human alpha1A calcium channel subunit linked to familial hemiplegic migraine. J Neurosci 1999 Mar 1;19(5):1610-9.

163. Hoeting JA. Bayesian Model Averaging: A Tutorial. Statistical Science 1999;14(4):382-417.

164. Boxer RB, Jang JW, Sintasath L, Chodosh LA. Lack of sustained regression of c-MYC-induced mammary adenocarcinomas following brief or prolonged MYC inactivation. Cancer Cell 2004 Dec;6(6):577-86.

165. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 2007 Jan;8(1):118-27.

166. Shi L, Campbell G, Jones WD, Campagne F, Wen Z, Walker SJ, Su Z, Chu TM, Goodsaid FM, Pusztai L, Shaughnessy JD, Jr., Oberthuer A, Thomas RS, Paules RS, Fielden M, Barlogie B, Chen W, Du P, Fischer M, Furlanello C, Gallas BD, Ge X, Megherbi DB, Symmans WF, Wang MD, Zhang J, Bitter H, Brors B, Bushel PR, Bylesjo M, Chen M, Cheng J, Chou J, Davison TS, Delorenzi M, Deng Y, Devanarayan V, Dix DJ, Dopazo J, Dorff KC, Elloumi F, Fan J, Fan S, Fan X, Fang H, Gonzaludo N, Hess KR, Hong H, Huan J, Irizarry RA, Judson R, Juraeva D, Lababidi S, Lambert CG, Li L, Li Y, Li Z, Lin SM, Liu G, Lobenhofer EK, Luo J, Luo W, McCall MN, Nikolsky Y, Pennello GA, Perkins RG, Philip R, Popovici V, Price ND, Qian F, Scherer A, Shi T, Shi W, Sung J, Thierry-Mieg D, Thierry-Mieg J, Thodima V, Trygg J, Vishnuvajjala L, Wang SJ, Wu J, Wu Y, Xie Q, Yousef WA, Zhang L, Zhang X, Zhong S, Zhou Y, Zhu S, Arasappan D, Bao W, Lucas AB, Berthold F, Brennan RJ, Buness A, Catalano JG, Chang C, Chen R, Cheng Y, Cui J, Czika W, Demichelis F, Deng X, Dosymbekov D, Eils R, Feng Y, Fostel J, Fulmer-Smentek S, Fuscoe JC, Gatto L, Ge W, Goldstein DR, Guo L, Halbert DN, Han J, Harris SC, Hatzis C, Herman D, Huang J, Jensen RV, Jiang R, Johnson CD, Jurman G, Kahlert Y, Khuder SA, Kohl M, Li J, Li M, Li QZ, Li S, Liu J, Liu Y, Liu Z, Meng L, Madera M, Martinez-Murillo F, Medina I, Meehan J, Miclaus K, Moffitt RA, Montaner D, Mukherjee P, Mulligan GJ, Neville P, Nikolskaya T, Ning B, Page GP, Parker J, Parry RM, Peng X, Peterson RL, Phan JH, Quanz B, Ren Y, Riccadonna S, Roter AH, Samuelson FW, Schumacher MM, Shambaugh JD, Shi Q, Shippy R, Si S, Smalter A, Sotiriou C, Soukup M, Staedtler F, Steiner G, Stokes TH, Sun Q, Tan PY, Tang R, Tezak Z, Thorn B, Tsyganova M, Turpaz Y, Vega SC, Visintainer R, von Frese J, Wang C, Wang E, Wang J, Wang W, Westermann F, Willey JC, Woods M, Wu S, Xiao N, Xu J, Xu L, Yang L, Zeng X, Zhang M, Zhao C, Puri RK, Scherf U, Tong W, Wolfinger RD. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat Biotechnol 2010 Aug;28(8):827-38.

167. Baird AE. Blood genomics in human stroke. Stroke 2007 Feb;38(2 Suppl):694-8.

168. Xu H. Gene expression in peripheral blood differs after cardioembolic compared with large-vessel atherosclerotic stroke: biomarkers for the etiology of ischemic stroke. Journal of Cerebral Blood Flow and Metabolism 2008;28(7):1320-8.

169. Bennett L, Palucka AK, Arce E, Cantrell V, Borvak J, Banchereau J, Pascual V. Interferon and granulopoiesis signatures in systemic lupus erythematosus blood. J Exp Med 2003 Mar 17;197(6):711-23.

Biography

Heather Michelle Garren was born on April 30, 1981 in Rutherfordton, North Carolina to Jim and Wanda Garren. She attended the University of North Carolina at Chapel Hill, where she earned a Bachelors of Science degree in Biology and graduated with honors in 2003. She matriculated to Duke University through the Program in Genetics and Genomics in 2004 and graduated with a Doctor of Philosophy degree in 2011. She was awarded the AACR Scholar-in-Training Award in 2008. She is married to Matthew Thomas LaBreche and resides with her husband and her two children, Samantha and Alexander, in Cary, North Carolina.

Publications

Garren, HM. (2003). Senior Honors Thesis: Human Fucosyltransferases and Cancer Cells. University of North Carolina at Chapel Hill.

Woo, CY, Osada, T, Yang, XY, Lei, G, McKinney, M, LaBreche, H, Hartman, Z, Amalfitano, A, Morse,MA, Lyerly, HK and Clay, TM. (2006). Induction of Wilms' tumor protein (WT1)-specific immunity using a WT1-expressing adenovirus vaccine. Proceedings of the American Association of Cancer Research, 47.

LaBreche, HG, Nevins, JR, Huang, ES. (2011). Integrating Factor Analysis and a Transgenic Mouse Model to Reveal a Peripheral Blood Predictor of Breast Tumors. (manuscript submitted)

LaBreche, HG, Meadows, SK, Nevins, JR, Chute, JP. (2011) Peripheral Blood Signatures of Lead Exposure. (manuscript under review).

Presentations

LaBreche, HG, Andrechek, E, Potti, A, Nevins, JR. (2006). "Early detection of breast cancer by gene expression profiling of peripheral blood cells." Genomes at Grandover: First Annual Retreat of the Institute for Genome Sciences and Policy (ISGP), Greensboro, NC.

LaBreche, HG, Andrechek, E, Marks, J, Nevins JR. (2007). "Characterization of peripheral blood cell signatures as novel breast cancer biomarkers." Genomes at Grandover: Second Annual Retreat of the Institute for Genome Sciences and Policy (IGSP), Greensboro, NC.

LaBreche, HG. (2007). "Peripheral blood cell signatures for early detection of breast cancer." Presented at the Center for RNA Biology Meeting, Duke University.

LaBreche, HG, Andrecheck, E, Marks, J, Nevins JR. (2008). "Peripheral blood cell signatures as novel breast cancer biomarkers." AACR Translational Cancer Medicine 2008: Cancer Clinical Trials and Personalized Medicine, Monterey, CA.

LaBreche, HG and Reeves, J. (2010). Cancer Genomics: Gene Expression Signatures as Cancer Phenotypes. Presented at the Institute for Genome Sciences and Policy (IGSP) house course entitled "From Science to Society: Decoding your Genome" on February 17, 2010.

LaBreche, HG. (2010). Transcriptional Control of Gene Expression. Presented at North Carolina Central University Cell Biology (BIOG 5310) course on September 14, 2010.

LaBreche, HG. (2011) Personalized Medicine. Lecture series presented at North Carolina Central University Genetics (BIOG5120) course on April 4, 6 and 8, 2011.