The Wistar Institute of Anatomy and Biology

Annual Progress Report: 2011 Nonformula Grant

Reporting Period

July 1, 2013 – June 30, 2014

Nonformula Grant Overview

The Wistar Institute of Anatomy and Biology received $991,900 in nonformula funds for the grant award period June 1, 2012 through August 29, 2014. Accomplishments for the reporting period are described below.

Research Project: Project Title and Purpose

Diagnostic Markers for Early-stage Lung Cancer in PAXgene Blood Samples – This proposal will verify the hypothesis based on preliminary results (published and unpublished) that tumor-specific signatures induced by a cancer in the lung are present in the peripheral blood and can be identified by global gene expression analyses. The commercialization of this technology requires that the samples be collected in a manner that stabilizes the nucleic acids in the samples at the time of the blood draw. A commercially available blood collection system, the PAXgene tube developed by PreAnalytX, meets this requirement. The purpose of the present study is to confirm the presence of the tumor-specific signatures in blood samples collected using PAXgene samples. This simplification of the blood collection will facilitate commercialization of our blood-based diagnostic platforms.

Anticipated Duration of Project

6/1/2012 – 8/29/2014

Project Overview

Our novel finding is that mononuclear white blood cells (PBMC) from lung cancer patients contain a tumor-relevant gene signature that accurately predicts early-stage lung cancer. Using microarray data from 267 individuals, we developed: 1) a 29-gene signature that predicts the presence of lung cancer with 86% accuracy when compared to a wide variety of controls with non-malignant, smoking-related lung disease, and 2) a 24-gene signature that distinguishes control nodules from cancers with an apparent specificity of 80% using a subset of the controls (41 controls) that included only those with various types of lung nodules.

The product we propose to develop is a gene-expression-based signature derived from an easily obtained peripheral blood sample. In the first application, this blood test would be used to provide additional information as to whether a suspicious lung nodule identified by a spiral

computerized tomography (CT) scan is malignant. Because CT scans cannot identify whether a nodule is malignant or benign, the genomic information provided by this test would help to reduce unnecessary surgery or biopsies and their associated expense and risks. We have already demonstrated that our diagnostic system works accurately using ribonucleic acid (RNA) from PBMC collected under research conditions under peer review.

The next step, the development of a clinically useful test, requires the simplification of the sample collection to one that is appropriate to a variety of clinical settings including clinicians' offices and more remote clinical settings. The commercially available blood collection tubes (PAXgene) from PreAnalytiX perfectly meet this requirement. These tubes provide rapid stabilization of RNA in whole blood at the moment of the blood draw so that a reproducible and highly standardized blood collection system can be established. Samples are stable for several days at room temperature and for more than a year at -20 to -80°C. Our preliminary results from microarray studies on 48 cancer and control samples in PAXgene support the viability of this approach.

These results support the feasibility of moving our PBMC results to a commercially viable platform. The validation of our preliminary studies using the PAXgene collection system is the object of this proposal. We have proposed to analyze 600 PAXgene samples to test and validate a blood-based approach to develop an accurate lung cancer diagnostic.

Specific Aims:
Aim 1: Collect patient and control samples in PAXgene from two collection sites.
Aim 2: Process 600 PAXgene RNA samples (~300 from cancer patients) and analyze gene expression profiles.
Aim 3: Identify the minimal number of genes and micro-RNAs (miRNAs) that provide the most accurate classification of cancers.
Aim 4: Assess new and old IP with Business Development.

Principal Investigator

Louise C. Showe, PhD
Professor
The Wistar Institute of Anatomy and Biology
3601 Spruce Street
Philadelphia, PA 19104-4265

Other Participating Researchers

Michael K. Showe, PhD; Andrei V. Kossenkov, PhD; Qin Lui, MD, PhD; George D. Hobbs, JD– employed by The Wistar Institute of Anatomy and Biology
Steven M. Albelda, MD – employed by the University of Pennsylvania School of Medicine
Gerard J. Criner, MD – employed by Temple University
Thomas L. Bauer, MD – employed by the Helen F. Graham Cancer Center at Christiana Care, Newark, DE

Expected Research Outcomes and Benefits

Lung cancer remains the primary cause of cancer-related deaths. In part, this is due to the lack of viable early detection protocols and the fact that common early symptoms are easily ignored by the individual. We know that a very large (>95%) percentage of lung cancers are the result of smoking, with additional risk factors that include exposure to asbestos, radon, and second-hand smoke. Thus, there is a large and identifiable population at risk for developing lung cancer. Deaths from lung cancer are higher for minorities, higher in men than women, and higher in underserved populations. In a ranking of lung cancer incidence by state into four classes, with the fourth class being the highest incidence class, Pennsylvania's lung cancer rate is in the third class (70 cases/100,000, with a range from 22-98/100,000). Neighboring Delaware, where many Pennsylvania workers live, is in the fourth or highest incidence class. In the rankings for lung cancer deaths, Pennsylvania is in the second class and Delaware is in the fourth. These numbers are striking given the fact that both of these states are located in a corridor of prestigious academic, medical, and biotech organizations.

Although the use of CT scans is projected to increase significantly for lung cancer detection, access to this technology will be much less available in rural settings and less likely to be available, and also perhaps less desirable, to underserved and ethnically diverse communities. Blood tests, by contrast, are routine and well-accepted occurrences associated with a visit to a doctor's office. The collection system we propose could also be easily used in community settings where, for example, flu shots are given or blood donations are taken. Such a test can be an early warning for imaging tests that would not otherwise be warranted; or it can be used in combination with imaging results that are inconclusive, essentially providing a second opinion based on an independent assessment.

It is clear from years of lung cancer statistics that early detection of lung cancer is the most viable way to prevent lung-cancer-related deaths and to minimize the costs of treatment, the costs to employers, and the tolls on families of the affected individuals. Recent advances in sample collection and storage and more sensitive and reliable platforms for testing multi-gene diagnostic panels suggest that the development of such platforms is now feasible.

Summary of Research Completed

<u>Aim 1: Collect New Samples for Analysis in PAXgene Tubes:</u> We proposed to collect 600 samples over a 2-year period. Four hundred thirty samples were to be collected from the Helen Graham Cancer Center (HGCC) and Temple University (TU) over the two years. As of June 30, 2014, we have received 261 samples from HGCC, exceeding our original goal and making up the shortfall in samples collected at TU. Only 12 samples were collected at TU in the first year. This problem was described in the 2012 progress report. To assure collection of adequate samples, we added Fox Chase Cancer Center (FCCC) as a second PA collection site, bringing Dr. Margie Clapper on as an additional collaborator to make up this shortfall. IRB approval by the FCCC Research Review Committee (RRC) and the Institutional Review Board (IRB) was received in March 2013. Additional samples from FCCC include 25 cancers and 25 controls. Although we have a large number of samples from Dr. Rom (now 1,080) collected as part of the

National Lung Screening Trial at New York University (NYU), most of these represent smoking controls with and without lung nodules in line with expected cancer incidence among smokers. However, an additional 34 cancer samples from the screening were received from Dr. Rom in the past year. In addition, Dr. Harvey Pass, also at NYU, has now collected an additional 75 cancers and 75 controls specifically for these studies. This brings the total number of samples collected to 679 [261 HGCC + 15 TU + 133 Penn + 70 (35 cancers + 35 controls) NYU + 150 Harvey Pass + 50 FCCC] exceeding our goal of 600 and providing a reasonable margin of error for failed samples. This does not include the large numbers of control samples available from the National Lung Screening Trial at NYU. All of the samples and associated information are being entered into a secure database.

Aim 2: Process 600 PAXgene RNA Samples for Gene and miRNA Expression: We proposed to process 300-400 samples for gene and miRNA expression in year 1. As of 7/1/2014, we had processed 523 samples on whole genome arrays and 513 on our micro-RNA platform leaving 156 samples to complete processing.

In addition to the samples described being processed above, we carried out a pilot project with Dr. Harvey Pass (NYU) to compare arterial and venous blood gene expression profiles for 5 lung cancer patients. Although Dr. Pass is not funded by the CURE funds, he is listed as a collaborator in the original proposal. All of the work was done with appropriate IRB approval from NYU and the Wistar Institute (IRB protocol # 2403205-4b). Although not anticipated in the original proposal, after discussion with our clinical collaborators it was felt to be an important control for the studies supported by this proposal and would be a necessary part of any forthcoming publications. No CURE funds were used to collect the samples, but since the data from this study will also be integrated into the larger analyses, which are the main aims of the proposal, the arrays were processed with CURE funds. It should also be clear that this is a different study than the pilot project that was reported in the 2012 CURE Program progress report. That study was a preliminary analysis of a subset of the total samples collected at that time and the pilot analysis was built into the original data analysis plan. The reason for doing the venous arterial comparison is because while samples from our cancer patients and control with suspicious nodules undergoing biopsy are routinely taken in the Operating Room arterially, control samples including post-surgery samples are taken by vena puncture. Although we had examined this problem briefly previously for our Peripheral blood mononuclear cell (PBMC) study, we did not have the kind of matched samples that Dr. Pass agreed to provide for this test set. The patients are summarized in Table 1. Arterial and venous samples were acquired at the same time from each patient. Samples were stored at -80C until all were collected and then transferred to the Wistar genomics facility for processing. RNA from one of the samples, the venous draw from patient 1715, had an RNA Integrity Number (RIN) number of 6.3 which was below our RIN 7.5 cut-off. We included this sample in the array study anyway. Looking at a Principal Component Analysis (PCA) using the 20,057 probes that passed quality control indicated that this sample was slightly distant from the other samples (Figure 1). To address this issue, we carried out two levels of analysis, one using all 5 patients and a second with patient 1715 removed from the study. We found minor differences between the 2 studies, but they did not change the overall conclusions and we will focus on the analysis using all samples. We first applied a paired t-test to compare the matched arterial and venous blood using a p-value cutoff of

0.05 and identifying 722 significant genes that are differentially expressed between the venous and arterial samples with a very high false discovery rate **(FDR) of 96%.** Furthermore, only 125/722 genes (17%) had a p-value < 0.01, indicating that the differences arterial/venous samples are not highly significant (Figure 2). In addition, none of the genes had a fold change greater than 1.5 (Figure 3) and most were less than 1.2. Additionally, a heatmap of the top 100 genes (Figure 4) did not reveal any clear separation between the arterial and venous samples. The samples did not cluster according to blood type and no patterns in the levels of gene expression associated with sample type are evident.

In conclusion, we found the differences between the samples obtained from arterial and venous blood were minimal and small in magnitude (small fold changes). Hierarchical clustering of our 4 sample analysis provided a clearer separation of arterial and venous samples when using 4 patients instead of 5, but the false discovery rate associated with that comparison was 100%, indicating that no major differences exist between samples obtained from either blood source. It may be possible to achieve smaller p-values with a larger sample size; however, the magnitude of the fold change is likely to remain small. This issue was examined in our previous PBMC study to see if there was any over representation of one sample type or the other in the misclassifications, but none were found. However, those samples were not PAXgene samples. The present study is a much better test and results support the hypothesis that differences due to the presence of a cancer are much more significant than differences between the 2 sample types.

Aim 3: Identify minimal classifiers for cancers vs. all non-cancers and cancer vs. non-malignant nodules: We have accumulated sufficient samples to surpass the 550 samples calculated to be needed to achieve 90% accuracies of classification. The analysis is in progress but will not be completed until the final samples for 2 independent locations are processed.

Aim 4: Patent Application Progress: No changes to this patent application or development of new applications have been taken at this time.

Commercialization Progress: We have negotiated a 2 year subsidized research agreement with Oncocyte. In addition, we have met with Dr. Greg Gonye, technical representative of Nanostring, to discuss the possibilities of using their FDA approved platform for simultaneously assaying a biomarker panel that includes both genes and miRNAs.

Figures And Data For Aim 2 Pilot Project: Comparison Of Gene Expression In Arterial And Venous Blood Paxgene Samples From Patient With Lung Cancer.

Table 1: Aim 2 Pilot Project: Demographics Of Lung Cancer Patients With Matched Venous And Arterial Blood Samples And RNA Quality Scores.

Patient ID	Collection Date*	Diagnosis**	Sex	Race	Age	RIN # Arterial	RIN # Venous
NYU1625	8/21/2013	Metastatic Adenocarcinoma	Female	White	60	8.4	8.3
NYU1672	8/5/2013	Mesothelioma	Male	White	75	8.0	7.9
NYU1675	8/7/2013	Adenocarcinoma	Male	White	74	7.8	8.6
NYU1690	8/14/2013	Adenocarcinoma	Male	White	70	8.8	8.2
NYU1715	10/11/2013	Adenocarcinoma	Female	White	68	8.8	**6.3**

Table 1: *****All samples were collected at NYU with appropriate patient consent and IRB approval. Samples were all whole blood and were collected in PAXgene tubes as described in the proposal. Samples were shipped overnight on dry ice and RNA purified in the Wistar Genomics Facility as described in the original proposal. ******All cancers were primary site lung cancers.

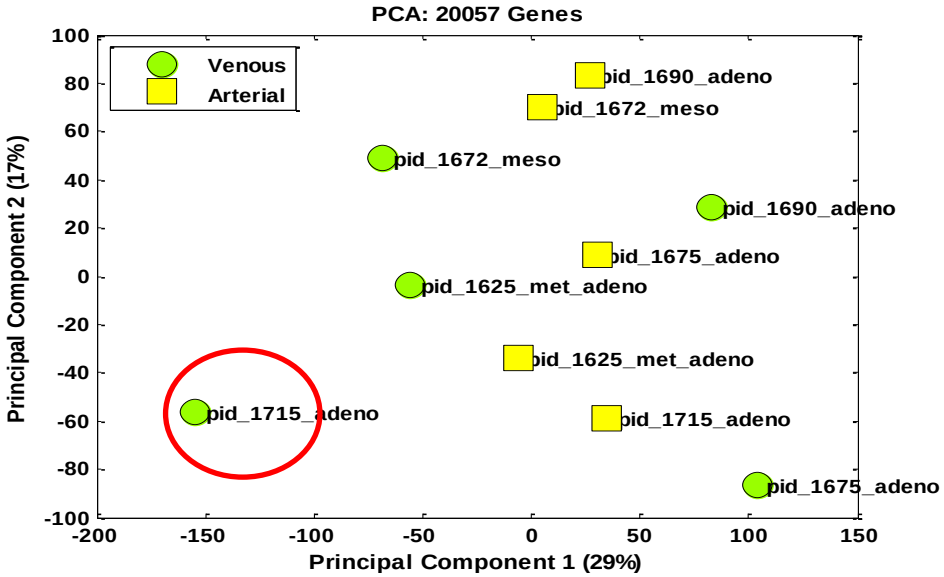

Figure 1: Aim 2 Pilot Project: Principal Component Analysis (PCA) Of Gene Expression Data To Assess Relative Differences And Similarities Between Venous and Arterial Samples From Lung Cancer Patients Listed In Table 1. The PCA plot was generated using all genes on the microarrays that passed quality control. Arterial samples are shown in yellow (■) and venous samples in green (●). The circle indicates the sample with the 6.3 RIN number which is slightly skewed from the rest of the samples. pid= patient numerical ID from Table 1, meso=mesothelioma, adeno=adenocarcinoma, met= metastasis. The plot demonstrates there are no clear separations between the 2 sample types.

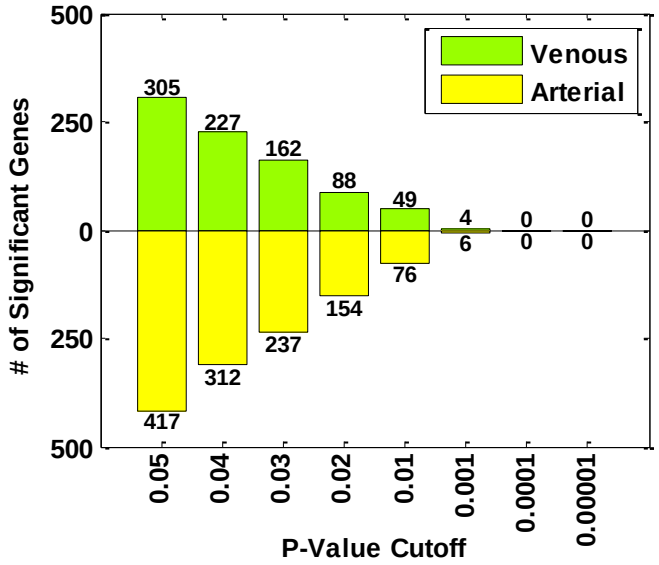

Figure 2: <u>Aim 2 Pilot Project:</u> Number Of Differentially Expressed Genes Between Venous And Arterial Samples Of Patients In Table 1 At Decreasing p-Values. The histogram shows how many of the differentially expressed genes are detected at different p-value cut-offs in venous blood (■) and arterial blood (■). Approximately 58% of the 722 genes were more highly expressed in arterial blood, but only 125 genes (17%) actually had a p-value < 0.01, indicating that the differences in expression between the 2 sample types are not highly significant.

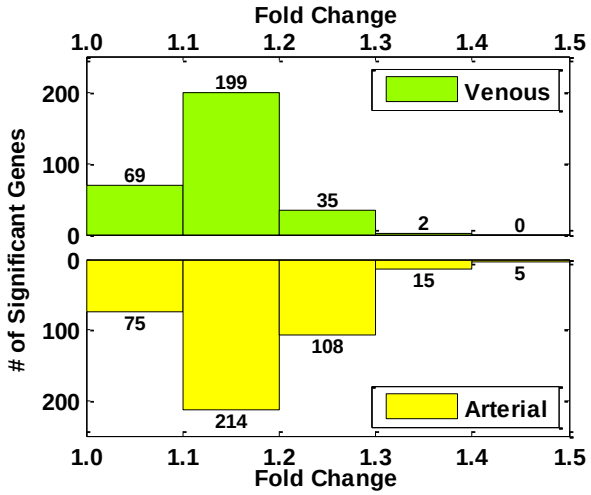

Figure 3: <u>Aim 2 Pilot Project</u>. A Summary of The Fold Change Distribution Of Genes Differentially Expressed Between Venous and Arterial Samples Of Patients In Table 1. Bar chart depicts the number of significantly expressed genes in venous (■) and arterial (■) blood at different fold change intervals. Only 5 genes have a fold change greater than 1.4 (but less than 1.5). All 5 genes (PELI1, LOC30820, FPR2, EIF4E3 and NAMPT) were more highly expressed in arterial blood than venous blood. Meanwhile, 557 genes (77%) have a less than 20% fold change in expression between venous and arterial samples.

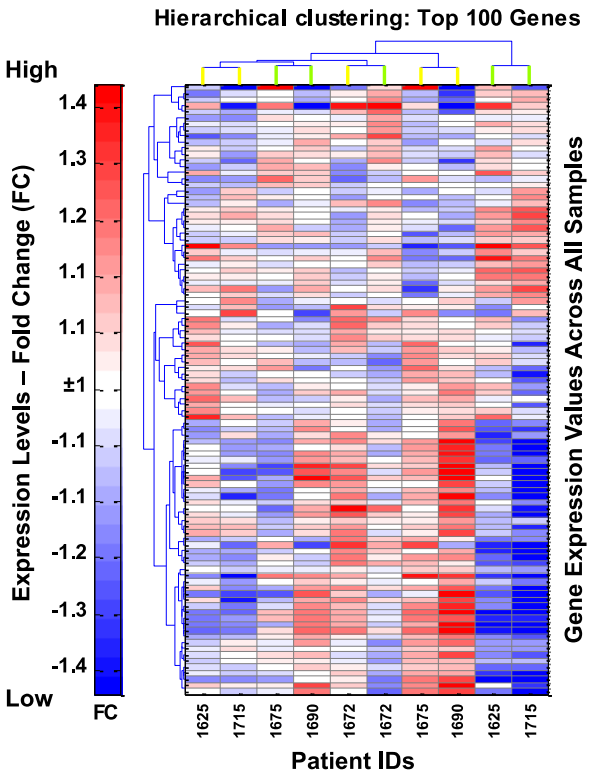

Figure 4: <u>Aim 2 Pilot Project</u>. Heatmap Of Top 100 Genes Differentially Expressed between Venous And Arterial Samples. A heatmap showing the relative levels of gene expression in matched arterial ( ) and venous ( ) blood samples. Each column represents a single PAXgene sample with patient IDs being listed on the X axis. Each row shows the expression values for one gene across all samples with the expression organizational tree on the Y axis indicating clusters of genes related by their expression patterns across the samples. The color bar at the left indicates the expression level differences with blue indicating lower expression and red higher levels of expression. It should be noted that the samples did not cluster according to blood sample type as indicated by the relational tree at the top of the heat map, and no patterns in the levels of gene expression characteristic of a specific sample type are evident. These results further support the conclusion that there are very few gene expression differences between the venous and arterial samples from the same individual.