Final Progress Report for Research Projects Funded by Health Research Grants

Instructions: Please complete all of the items as instructed. Do not delete instructions. Do not leave any items blank; responses must be provided for all items. If your response to an item is "None", please specify "None" as your response. "Not applicable" is not an acceptable response for any of the items. There is no limit to the length of your response to any question. Responses should be single-spaced, no smaller than 12-point type. The report **must be completed using MS Word**. Submitted reports must be Word documents; they should not be converted to pdf format. Questions? Contact Health Research Program staff at 717-783-2548.

1. **Grantee Institution:** The Wistar Institute

2. **Reporting Period (start and end date of grant award period):** 06/1/2012 – 08/29/2014

3. **Grant Contact Person (First Name, M.I., Last Name, Degrees):** Russel E. Kaufman, M.D.

4. **Grant Contact Person's Telephone Number:** 215-898-3926

5. **Grant SAP Number:** 4100059200

6. **Project Number and Title of Research Project:** 01 - Diagnostic Markers for Early-Stage Lung Cancer in PAXgene Blood Samples

7. **Start and End Date of Research Project:** 06/1/2012-08/29/2014

8. **Name of Principal Investigator for the Research Project:** Louise C. Showe, Ph.D.

9. **Research Project Expenses**.

   9(A) Please provide the total amount of health research grant funds spent on this project for the entire duration of the grant, including indirect costs and any interest earned that was spent:

   $ $1,004,637.07

   9(B) Provide the last names (include first initial if multiple individuals with the same last name are listed) of **all** persons who worked on this research project and were supported with health research funds. Include position titles (Principal Investigator, Graduate Assistant, Post-doctoral Fellow, etc.), percent of effort on project and total health research funds expended for the position. For multiple year projects, if percent of effort varied from year to year, report in the % of Effort column the effort by year 1, 2, 3, etc. of the project (x% Yr 1; z% Yr 2-3).

Last Name, First Name	Position Title	% of Effort on Project	Cost
Showe, L.	Professor and PI	19% Yr1; 22% Yr2	108,389.90
Showe, M.	Senior Scientist	25% Yr1; 50% Yr2	108,512.94
Liu, Q.	Assoc. Professor	20%	75,329.48
Majumdar, R.	Wistar Research Assist.	20% Yr1; 54.7% Yr2	53,246.79
Dawany, N.	Staff Scientist	8.9% Yr2	6,620.52
Albelda, S.	Sub-site PI	5%	26,760.00
Bauer, T.	Sub-site PI	2.52% Yr1;5.72% Yr2	19,429.63
Swanson, P.	Research Nurse Supervisor	3.86% Yr1;10.6% Yr2	12,870.97
Hamm, D.	Research Nurse Coordinator	30.2% Yr1; 0.9% Yr2	22,226.65
Rubin, L.	Clinical Research Assoc.	4.76% Yr1;17.55% Yr2	20,533.52
Johnson, M.	Research Nurse Coordinator	6.8% Yr1; 4.78% Yr2	11,371.33
Czerwinski, C.	Research Nurse Coordinator	0.1% Yr1; 0.2% Yr2	192.45
Bruegel, C.	Research Nurse Coordinator	0.31%	329.90
Clapper, M.	Sub-site PI	6%	10,947.83
Vanderveer, L.	Scientific Associate	21%	17,356.46
Criner, G.	Sub-site PI	1%	5,194.72
Bedenko, E.	Research Coordinator	2.4%	1,533.85
Bolla, S.	Postdoctoral Fellow	18%	15,226.37

9(C) Provide the names of **all** persons who worked on this research project, but who *were not* supported with health research funds. Include position titles (Research Assistant, Administrative Assistant, etc.) and percent of effort on project. For multiple year projects, if percent of effort varied from year to year, report in the % of Effort column the effort by year 1, 2, 3, etc. of the project (x% Yr 1; z% Yr 2-3).

Last Name, First Name	Position Title	% of Effort on Project
Combs, K.	Research Manager	0.4% Yr1; 0.2% Yr2
Huelsenbeck-Dill, L.	Technician	0.4% Yr1; 0.17% Yr2

9(D) Provide a list of **all** scientific equipment purchased as part of this research grant, a short description of the value (benefit) derived by the institution from this equipment, and the cost of the equipment.

Type of Scientific Equipment	Value Derived	Cost
None		

10. Co-funding of Research Project during Health Research Grant Award Period. Did this research project receive funding from any other source during the project period when it was supported by the health research grant?

Yes____X___ No_____

If yes, please indicate the source and amount of other funds:

We were co-investigators on an R21 CA156087-02 (Dr. A. Vachani), "Peripheral Blood Gene Expression for the Diagnosis of Indeterminate Lung Nodules," with a budget from 8/1/2012 to 7/31/2013 of $44,665 to carry out gene expression on Paxgene samples which were to be included in this larger study.

We received $100,000 in direct research support from Oncocyte, a subsidiary of BioTime Inc., 1301 Harbor Bay Parkway, Alameda, CA 94502, as they were interested in licensing the technology if successful.

11. Leveraging of Additional Funds

11(A) As a result of the health research funds provided for this research project, were you able to apply for and/or obtain funding from other sources to continue or expand the research?

Yes ____X_____ No_____

If yes, please list the applications submitted (column A), the funding agency (National Institutes of Health—NIH, or other source in column B), the month and year when the application was submitted (column C), and the amount of funds requested (column D). If you have received a notice that the grant will be funded, please indicate the amount of funds to be awarded (column E). If the grant was not funded, insert "not funded" in column E.

Do not include funding from your own institution or from CURE (tobacco settlement funds). Do not include grants submitted prior to the start date of the grant as shown in Question 2. If you list grants submitted within 1-6 months of the start date of this grant, add a statement below the table indicating how the data/results from this project were used to secure that grant.

A. Title of research project on grant application	B. Funding agency (check those that apply)	C. Month and Year Submitted	D. Amount of funds requested:	E. Amount of funds to be awarded:
Sponsored Research Agreement to study Sp100 Isoforms	☐NIH ☐ Other federal (specify:_____) X Nonfederal source (specify: Oncocyte_Corp.)	January 2013	$100,000	$100,000

We had preliminary results from the samples collected under R21 CA156087-02. These were described in the original application. Analyses of these samples were carried out in the early months of this CURE grant.

11(B) Are you planning to apply for additional funding in the future to continue or expand the research?

Yes___X_____ No_____

If yes, please describe your plans:

We have procured an additional $100,000 from Oncocyte that will allow us to analyze additional samples if needed.

12. Future of Research Project. What are the future plans for this research project?

We plan to test our most accurate gene and microRNA panels on a more clinically appropriate platform. We have been in discussions with Nanostring Technologies, Seattle, Washington, as their nCounter® technology which measures mRNA or miRNA levels in many types of samples is FDA approved for clinical use. Oncocyte Corporation (Oncocyte), the company with whom we have been in discussions, expects to purchase two of the Nanostring machines for testing their own Biomarker panels using the nCounter® technology and to develop a CLIA lab for this purpose. (The Clinical Laboratory Improvement Amendments of 1988 Congressional legislation requires that quality assurance practices be implemented in labs carrying out clinical tests. CLIA requires labs to measure performance at each step of the testing process from the beginning to the end-point in order to meet the CLIA requirements). The Wistar Institute (Wistar) is also considering the purchase of one of the Nanostring nCounter® machines as part of a strategic plan for developments in personalized medicine. Further development of our diagnostic signature will be carried out at Wistar or in collaboration with Oncocyte.

13. New Investigator Training and Development. Did students participate in project supported internships or graduate or post-graduate training for at least one semester or one summer?

Yes___X_____ No_____

If yes, how many students? Please specify in the tables below:

	Undergraduate	Masters	Pre-doc	Post-doc
Male		2		
Female	1			1
Unknown				
Total	**1**	**2**		**1**

	Undergraduate	Masters	Pre-doc	Post-doc
Hispanic				
Non-Hispanic	1	2		1
Unknown				
Total	**1**	**2**		**1**

	Undergraduate	Masters	Pre-doc	Post-doc
White		2		1
Black				
Asian	1			
Other				
Unknown				
Total	**1**	**2**		**1**

14. **Recruitment of Out-of–State Researchers**. Did you bring researchers into Pennsylvania to carry out this research project?

    Yes_____ No ____X_____

    If yes, please list the name and degree of each researcher and his/her previous affiliation:

15. **Impact on Research Capacity and Quality**. Did the health research project enhance the quality and/or capacity of research at your institution?

    Yes ___X_____ No_____

    If yes, describe how improvements in infrastructure, the addition of new investigators, and other resources have led to more and better research.

    We developed a structure for the collection of human samples that is now being used by several Wistar investigators. A collaboration with the Helen F. Graham Cancer Center (HFGCC), which was initiated with this proposal, has been expanded to include investigators working on several different cancers, including melanoma and ovarian cancer. Participants from both institutions meet quarterly.

16. **Collaboration, business and community involvement.**

    16(A) Did the health research funds lead to collaboration with research partners outside of your institution (e.g., entire university, entire hospital system)?

    Yes____X_____ No_____

    If yes, please describe the collaborations:

Samples for these studies were collected at HFGCC (DE), New York University (NYU), Temple University (TU), University of Pennsylvania (Penn), Roswell Park (RP) and Fox Chase Cancer Center (FCCC). Each of these institutions provided samples that were used in the supported studies.

16(B) Did the research project result in commercial development of any research products?

Yes_____ No ___X_____

If yes, please describe commercial development activities that resulted from the research project:

16(C) Did the research lead to new involvement with the community?

Yes _____X___ No_____

If yes, please describe involvement with community groups that resulted from the research project:

Pennsylvania lung cancer survivors were hosted at Wistar and informed of ongoing efforts to develop a blood-based detection platform for diagnosis of early stage lung cancer.

17. Progress in Achieving Research Goals, Objectives and Aims.

List the project goals, objectives and specific aims (as contained in the grant agreement). Summarize the progress made in achieving these goals, objectives and aims for the period that the project was funded (i.e., from project start date through end date). Indicate whether or not each goal/objective/aim was achieved; if something was not achieved, note the reasons why. Describe the methods used. If changes were made to the research goals/objectives/aims, methods, design or timeline since the original grant application was submitted, please describe the changes. Provide detailed results of the project. Include evidence of the data that was generated and analyzed, and provide tables, graphs, and figures of the data. List published abstracts, poster presentations and scientific meeting presentations at the end of the summary of progress; peer-reviewed publications should be listed under item 20.

This response should be a DETAILED report of the methods and findings. It is not sufficient to state that the work was completed. Insufficient information may result in an unfavorable performance review, which may jeopardize future funding. If research findings are pending publication you must still include enough detail for the expert peer reviewers to evaluate the progress during the course of the project.

Health research grants funded under the Tobacco Settlement Act will be evaluated via a performance review by an expert panel of researchers and clinicians who will assess project

work using this Final Progress Report, all project Annual Reports and the project's strategic plan. After the final performance review of each project is complete, approximately 12-16 months after the end of the grant, this Final Progress Report, as well as the Final Performance Review Report containing the comments of the expert review panel, and the grantee's written response to the Final Performance Review Report, will be posted on the CURE Web site.

There is no limit to the length of your response. Responses must be single-spaced below, no smaller than 12-point type. If you cut and paste text from a publication, be sure symbols print properly, e.g., the Greek symbol for alpha (α) and beta (ß) should not print as boxes (□) and include the appropriate citation(s). DO NOT DELETE THESE INSTRUCTIONS.

Specific Aims: (From the original proposal)
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 miRNAs that provide the most accurate classification of lung cancers.
Aim 4: Assess new and old IP with Business Development.

Specific Aims and Research Completed:

Aim 1: Collect patient and control samples in PAXgene tubes from two collection sites:
We proposed to collect 300 blood samples in PAXgene tubes from patients with non-small cell lung cancer (NSCLC) and at risk controls with smoking related lung disease including benign nodules. Samples were to be collected over an 18-24 month period from the Lung Center at Temple University (TU) and the Helen F. Graham Cancer Center at Christiana Care (HFGCC). In year 2, to meet our projected accrual, we found it necessary to identify another Pennsylvania collection site in addition to TU. Dr. Marjorie Clapper, Professor and Co-leader of the Cancer Prevention and Control Program at Fox Chase Cancer Center (FCCC), now a part of TU, was recruited to provide an additional collection site. Because the goal of 600 samples outlined in the original proposal would have been difficult in the two-year time frame, we proposed to use PAXgene samples which were already collected from collaborators at two additional collection sites under an R21 grant on which Dr. Showe is a Co-Investigator. These samples were to be available for analysis in year one of the project. This group included in total 174 usable samples, 114 from the Hospital of the University of Pennsylvania (HUP) and 60 from Roswell Park (RP). The HUP samples also included a subset of samples collected ~ 6-12 months post lung resection. Additional samples needed to meet the goal of 600 were collected at NYU as part of their extensive lung screening program. (TABLE 1).

Aim 2: Process 600 PAXgene RNA Samples for Gene and microRNA Expression: By the end of year 2, we collected 656 PAXgene samples for processing on Illumina HT 12v4 gene expression arrays and Applied Biosystems' OpenArray miRNA detection platform. Very few samples were eliminated based on Ribonucleic Acid (RNA) quality based on their RNA integrity number (RIN). Only samples with RINs >7.5 were used for the analysis. Unfortunately, the 60 samples

from RP collected as part of the R21 grant had to be eliminated from the analysis. The data from these samples could not be integrated with the data for the HUP samples from the same study despite efforts to correct for the differences. This left us with 596 samples from the five remaining locations which were collected and assayed over different time periods. All samples, except 50 from HUP and 60 from RP, were processed by the Wistar genomics facility using highly standardized methods for RNA purification and microarray processing. RNA was prepared from the PAXgene tubes as recommended by the manufacturer (Pre-Analytx, Qiagen) to isolate both mRNA and miRNA species. The collection process is described in Figure 1.

Aim 3: Identify minimal classifiers for cancers vs. all non-cancers and cancer vs. non-malignant nodules:

Messenger Ribonucleic Acid (MRNA) And MicroRibonucleic Acid (MiRNA) Data Preprocessing:

In order to eliminate technical variability between samples, the mRNA expression data was first quantile normalized. The data was then inspected using Principal Component Analysis for artifacts and batch effects. The data was checked for several batch effects, including both batch effect related to collection site and batch effect associated with sample preparation/amplification/hybridization. Indeed, since the samples in this study came from four different collection sites and were prepared/assayed over 2 + years in 17 batches using different lots of reagents and arrays, we observed both batch effects present in the data (Figure 2, left). Distance Weighted Discrimination (DWD) method [3, 6, 10] was then applied to the data to correct for *batch effects* associated with the collection sites and/or correlated with time of amplification. Figure 2 demonstrates PCA plots for the gene expression data before and after application of DWD and clearly shows that the technical batch effects could be corrected. Obvious outlier samples were removed from the analysis before the DWD adjustment. In several cases, outliers were reanalyzed on arrays to eliminate the possibility that the effect was not array related but in all cases this was sample (collection) related, but included <2% of the samples. The only samples that could not be DWD corrected, and remained a problem even in the array reruns, were the 60 samples from RP. As described in an earlier report comparing these samples with the HUP samples prepared at the same time, we concluded there was some basic problem with the RP sample collection. These samples had to be eliminated from further study. Since we had built in a 10% failure rate in the study, it did not affect the sample number projected for analysis.

The OpenArray miRNA PCR array data was first filtered to remove any miRNAs with more than 10% missing values across samples, which resulted in a list of 202 significantly detected miRNAs. This number is consistent with our previous results and those of others using peripheral blood samples. Missing values for the 202 miRNAs were imputed using nearest neighbor method [9], and the data was then quantile normalized. DWD-corrected mRNA and miRNA data were used for all further analyses.

Datasets Analyses:

Results for this report were generated using *data only on the 415 samples* that were completed on both platforms by early August 2014. Analyses with additional samples will be integrated into ongoing studies.

Two types of comparisons were carried out during the preliminary analysis. The first included all of the pre-treatment/pre-surgery lung cancers (Cancers, LC) vs only those control patients with benign lung nodules (nodule controls, NOD) identified primarily by Computerized Tomography (CT) scans. The second analysis included pre-treatment/pre-surgery lung cancers vs all Controls, including NODs (benign nodules) plus all controls with non-malignant smoking related disease (smoking controls, SC).

As reported previously, the data was analyzed using a Support Vector Machine (SVM) [4, 5, 8]. Sample classes, analysis strategy and numbers of samples and their subtypes are summarized in Table 1. Of the 415 total samples in analysis, 345 samples had unambiguously assigned Cancer (LC) or Control (NOD or SC) labels (set A) and were used for training and testing purposes. The remaining 70 samples included samples with indistinct phenotypes (set B): post lung resection samples and samples from nodule patients who later developed LC and were used for further classification by the classifier developed on the 345 unambiguously assigned samples (clinically confirmed as case or control but not including post resection samples). Samples from both sets were randomly split into 70% for the training set (242 samples for Set A) and a set aside 30% for the testing set (103 samples for Set A).

The training set was used to find the best classifier by SVM with a10-fold cross-validation routine using Radial Basis Function (RBF) kernel and forward feature selection (FFS) that at each step picked one best feature (gene or miRNA) which improved overall training accuracy. Alternatively, we tried using linear kernel and Recursive Feature Elimination (RFE), which we used successfully in the past [8], but forward feature selection with RBF kernel gave better accuracy on the preliminary training set. A classifier built for the number of features that provided the best training accuracy was then selected as a final classifier and applied to the independent set-aside testing set to estimate its unbiased accuracy.

Using the described classifier development process, we used three data sets to create three different classifiers for comparison: (1) using only mRNA data; (2) using only miRNA expression data, and (3) analyzing the combined mRNA and miRNA data. Each dataset/classification analysis resulted in a report based on the testing set performance and included accuracy, sensitivity, specificity and area under ROC-curve (AUC). The results are listed in Table 2. According to the table, the best accuracy was achieved by general Cancer vs all Controls classifier (83% accuracy) that used both mRNA and miRNA data at the same time (145 total features), which demonstrates advantage of using both platforms in the same classification The individual scores for each sample from the independent testing set assigned by the classifier are shown in Figure 3, where each sample received a score assigned by the SVM classifier with positive scores indicating cancer and negative scores a control, and which can be interpreted as a measure of the strength or the reliability of the classification.

Ongoing Analyses:

Although the accuracies listed in Table 2 do not yet meet our identified goal of 90%, the number of samples used in this initial analysis are below the number indicated by our Power calculation included in the 2013 report (Figure 4), which projected that 550 samples equally divided between patients and nodules would be needed to achieve the 90% goal. In fact, based on 83% accuracy achieved by using 242 samples as a training set, the observed performance (green

square) is right on the predicted line. So, by incorporating data from all samples, we are still on track to achieve our 90% goal. In addition, the number of NOD samples in these initial studies is below what we predicted would be needed to achieve our 90% accuracy goal. The data from these additional samples have been collected and will be incorporated in ongoing analyses. Most of the samples have been processed, but there has been some delay in completing the OpenArray miRNA studies, as the last batch of arrays purchased had some technical problems and we are waiting for replacements from Life Sciences.

In addition to increasing the size of the training sets with new data, we plan to explore a number of additional approaches to the preprocessing and analysis of the data. Two of our sites are continuing sample collection, and this will provide us with the additional flexibility to analyze and explore how different subsets of either patients or controls relate to each other. Also, modifying the batch correction routine can provide an additional increase in classifier performance. In addition to using DWD to combine all samples in a single batch, we will try using the samples that we have replicated between batches to correct for array batch effects. These replicates were prepared to specifically deal with this potential problem. Also, several other available batch-correction methods will be applied to the data, including mean-centering (PAMR), surrogate variable analysis (SVA), genometric ratio-based method (Ratio_G) and others reviewed in *Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods*[1] . In case of poor performance of all proposed batch correction methods, we will attempt developing classifier with multiple linear regression methods [2] that allows including batch information as additional variables in the classification model. Finally, when data for the full set of samples is available, we will re-try using SVM with linear kernel again, separately or in combination with RFE as a feature selection routine.

Aim 4: Patent Application Progress: The Peripheral Blood Mononuclear Cell (PBMC)/blood gene expression test system is the subject of PCT/US2008/013450. No changes to this application or development of new applications have been taken at this time, but will be discussed as the majority of our specimens are analyzed in the next months.

PCT/US2009/063602 is a patent application owned by Wistar and directed to the use of a miRNA expression level or an expression level profile of multiple miRNAs to access lung disease in a patient, including lung cancer. No licenses or options to this patent application have been granted to third parties at this time.

Commercialization:
We have established a subsidized research agreement with Joseph Wagner, the CEO of Oncocyte, which extends to August 2015 with options to license the technology depending on outcome. We have also established collaboration with Nanostring to test some of our samples with their system. The Nanostring nCounter® technology platform is FDA approved and sample requirements are flexible. In addition Oncocyte is planning to use this technology to implement other diagnostic signatures that they are developing. We will also continue to explore the OpenArray PCR Array platform, as we have quite successfully used it in the past to move a microarray signature to the more clinically friendly PCR platform[7].

References:

- 1 Chao Chen, Kay Grennan, Judith Badner, Dandan Zhang, Elliot Gershon, Li Jin, and Chunyu Liu, 'Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods', *PLoS ONE,* 6 (2011), e17238.
- 2 C.C. Huang, S.H. Tu, H.H. Lien, J.Y. Jeng, J.S. Liu, C.S. Huang, L.C. Lai, and E.Y. Chuang, 'Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. ', *J Data Min Bioinform,* 9 (2014), 149-71.
- 3 Hanwen Huang, Xiaosun Lu, Yufeng Liu, Perry Haaland, and J.S. Marron, 'R/Dwd: Distance-Weighted Discrimination for Classification, Visualization and Batch Adjustment', *Bioinformatics,* 28 (2012), 1182-83.
- 4 A. V. Kossenkov, A. Vachani, C. Chang, C. Nichols, S. Billouin, W. Horng, W. N. Rom, S. M. Albelda, M. K. Showe, and L. C. Showe, 'Resection of Non-Small Cell Lung Cancers Reverses Tumor-Induced Gene Expression Changes in the Peripheral Immune System', *Clin Cancer Res,* 17 (2011), 5867-77.
- 5 Andrew V. Kossenkov, Noor Dawany, Tracey L. Evans, John C. Kucharczuk, Steven M. Albelda, Louise C. Showe, Michael K. Showe, and Anil Vachani, 'Peripheral Immune Cell Gene Expression Predicts Survival of Patients with Non-Small Cell Lung Cancer', *PLoS ONE,* 7 (2012), e34392.
- 6 JS Marron, and MJ Todd, 'Distance Weighted Discrimination ', in *School of Operations Research and Industrial Engineering,Cornell University* (2003).
- 7 Sharmistha Pal, Yingtao Bi, Luke Macyszyn, Louise C. Showe, Donald M. O'Rourke, and Ramana V. Davuluri, 'Isoform-Level Gene Signature Improves Prognostic Stratification and Accurately Classifies Glioblastoma Subtypes', *Nucleic Acids Research* (2014).
- 8 M. K. Showe, A. Vachani, A. V. Kossenkov, M. Yousef, C. Nichols, E. V. Nikonova, C. Chang, J. Kucharczuk, B. Tran, E. Wakeam, T. A. Yie, D. Speicher, W. N. Rom, S. Albelda, and L. C. Showe, '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,* 69 (2009), 9202-10.
- 9 O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, 'Missing Value Estimation Methods for DNA Microarrays', *Bioinformatics,* 17 (2001), 520-5.
- 10 A. Vachani, M. Nebozhyn, S. Singhal, L. Alila, E. Wakeam, R. Muschel, C. A. Powell, P. Gaffney, B. Singh, M. S. Brose, L. A. Litzky, J. Kucharczuk, L. R. Kaiser, J. S. Marron, M. K. Showe, S. M. Albelda, and L. C. Showe, 'A 10-Gene Classifier for Distinguishing Head and Neck Squamous Cell Carcinoma and Lung Squamous Cell Carcinoma', *Clin Cancer Res,* 13 (2007), 2905-15.

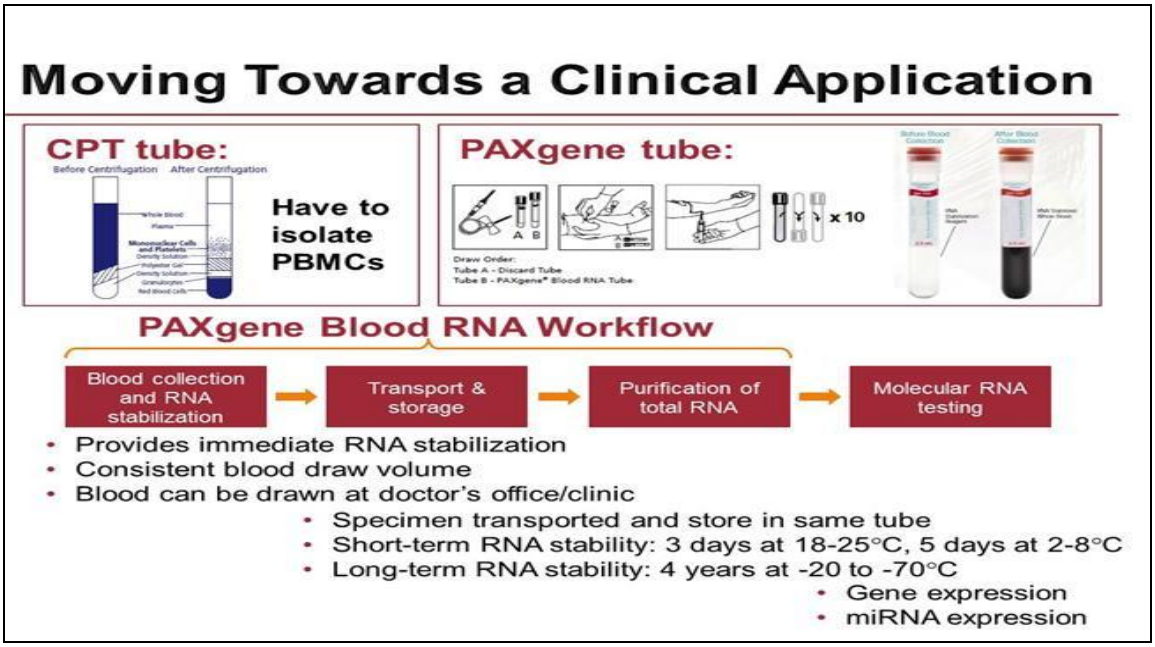

Figure 1: Diagram showing the workflow for processing PAXgene samples for RNA. 2.5 mls are collected and processed for mRNA and miRNA studies. The RNA isolations are carried out under standardized conditions in the Wistar Genomics facility. The CPT system on the left was used to isolate PBMC for our earlier study[8] . It requires 6 mls of blood, and samples must be processed within 1-2 hrs of collection. The stabilities of the RNA in the PAXgene system over time are shown in the figure and make it more amenable to the collection of clinical samples.

Table 1: Summary Of The Number Of Samples Used For The Various Analyses**.**

Number of Samples Analyzed					
		set A*			set B**
Comparison	Class	total	training	testing	total
LC vs. NOD	LC	181	127	54	
	NOD	99	69	30	
	total	280	196	84	70+65
LC vs. NOD+SC	LC	181	127	54	
	NOD+SC	164	115	49	
	total	345	242	103	70

*(Set A) 345 samples unambiguously assigned as Cancer (LC) or Control (NOD or SC) were used for training and testing. **(Set B) 70 samples with indistinct phenotypes. These 70 samples include post lung resection samples and samples from nodule patients who later developed LC, so the status of the cancer signature was essentially unknown. The LC vs. NOD comparison also included 65 SC samples that were not used in training-testing, but were available for classification.

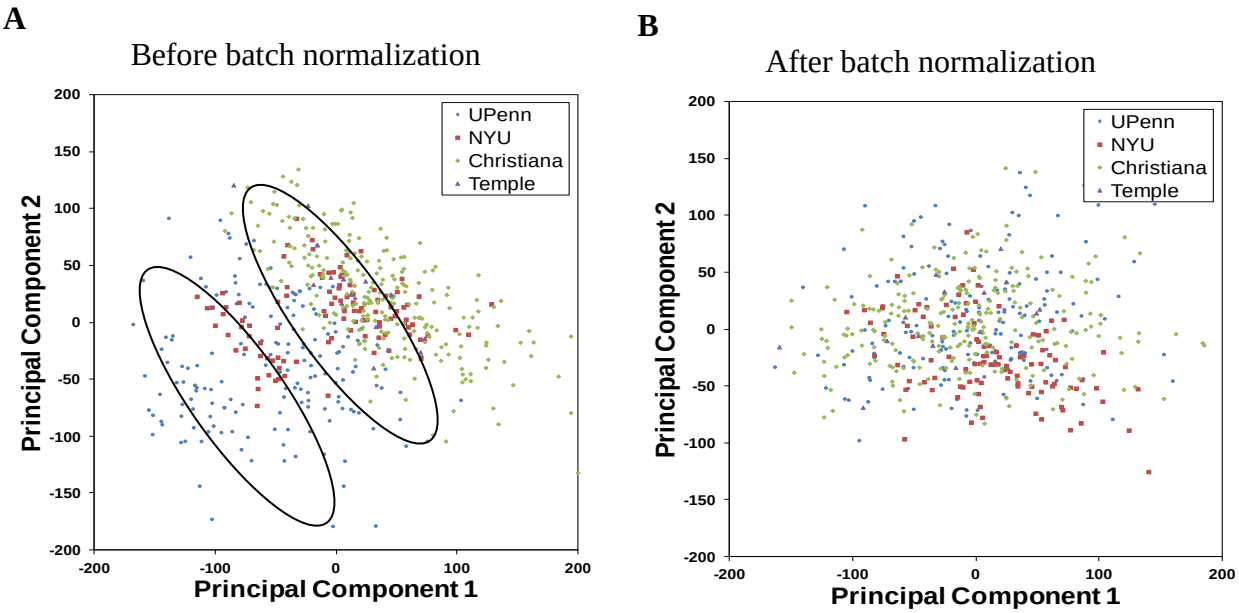

Figure 2. Batch Effect Correction Using DWDB. Principal Component Analysis (PCA) of the 536 microarray samples (595 minus the 60 from RP) before (A) and after (B) the application of DWD. The Batch effect correction is most noticeable for the NYU samples, which cluster in two groups as circled in A, and which are reorganized into one group in B after the application of DWD.

Table 2: Preliminary Accuracies, Sensitivities and Specificities in Distinguishing patients with lung cancer (LC) from patients with benign nodules (NOD) and smoking controls without nodules (SC).*

Comparison	Data Type	Feature #	Accuracy	Sensitivity	Specificity	AUC
LC vs. NOD	mRNA	161	81%	92%	60%	0.86
	miRNA	5	75%	83%	60%	0.75
	Both	**147**	**79%**	**87%**	**67%**	**0.87**
LC vs. NOD+SC	mRNA	151	79%	78%	80%	0.88
	miRNA	26	71%	69%	73%	0.77
	Both	**145**	**83%**	**81%**	**84%**	**0.88**

*Data is presented for the analyses using only gene expression (mRNA), only miRNA expression and mRNA+miRNA expression(Both). NOD=nodules, SC=Smoking controls without nodules.

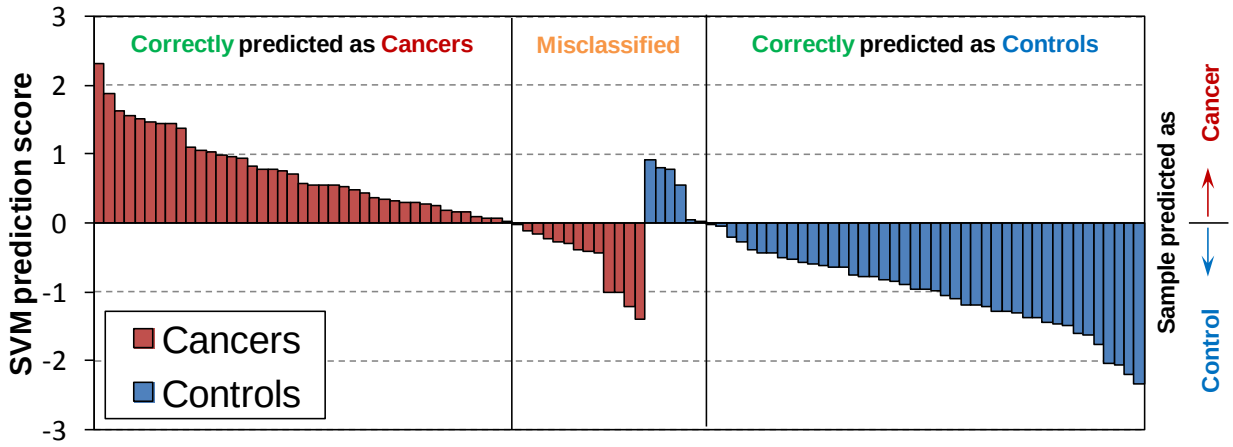

Figure 3. Combining mRNA and miRNA signatures. The figure shows the classification accuracy using expression data from 125 gene probes and 20 miRNA on *the independent test* set of 84 samples(see Table 1, LC vs.NOD, Testing). The AUC was 0.87.

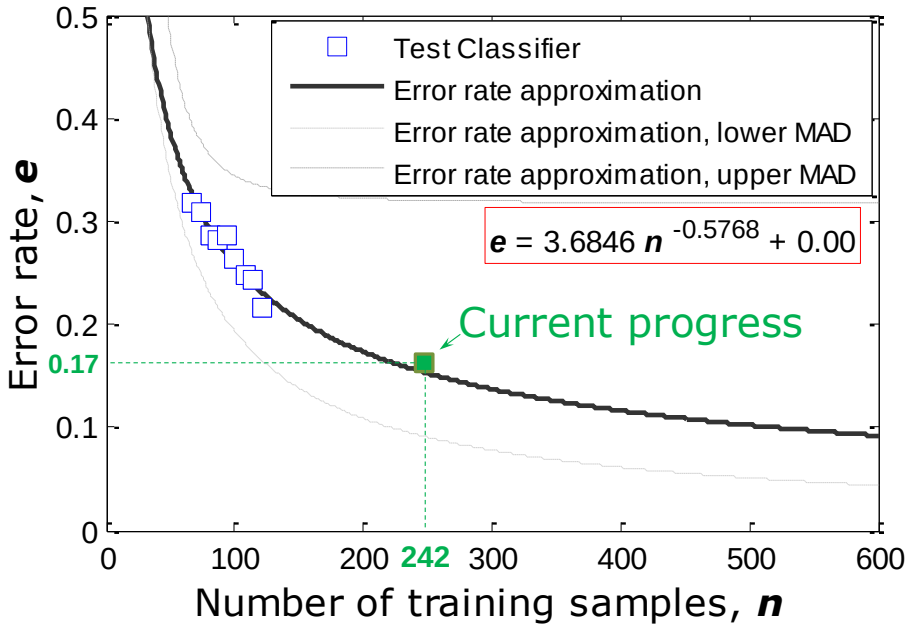

Figure 4: Progress Error Rate in the Context of Our Previous Power Calculation for Accuracy of Classification. The Power curve was developed on our preliminary studies of 10X samples described in the 2013 Progress Report. The power function was fit by selecting different training set sizes from the overall data and plotting it against the corresponding error rate of the classification for that data. The relationship between the number of samples used for training and the error rate shows that, by increasing the training set size, we can achieve higher accuracies in the classification of NSCLC versus controls with and without nodules. 90% classification accuracy can be achieved by using a training set containing approximately 550 samples. The results for the 242 samples used for the training in this report are indicated in green on the curve; the error rate of this analysis is 0.17 and is right on the target with our earlier prediction. MAD: median absolute deviation across 50 re-samplings.

18. Extent of Clinical Activities Initiated and Completed. Items 18(A) and 18(B) should be completed for all research projects. If the project was restricted to secondary analysis of clinical data or data analysis of clinical research, then responses to 18(A) and 18(B) should be "No."

18(A) Did you initiate a study that involved the testing of treatment, prevention or diagnostic procedures on human subjects?

_____Yes
__X_ No

18(B) Did you complete a study that involved the testing of treatment, prevention or diagnostic procedures on human subjects?

____ Yes
__X_ No

If "Yes" to either 18(A) or 18(B), items 18(C) – (F) must also be completed. (Do NOT complete 18(C-F) if 18(A) and 18(B) are both "No.")

18(C) How many hospital and health care professionals were involved in the research project?

_____Number of hospital and health care professionals involved in the research project

18(D) How many subjects were included in the study compared to targeted goals?

_____Number of subjects originally targeted to be included in the study
_____Number of subjects enrolled in the study

Note: Studies that fall dramatically short on recruitment are encouraged to provide the details of their recruitment efforts in Item 17, Progress in Achieving Research Goals, Objectives and Aims. For example, the number of eligible subjects approached the number that refused to participate and the reasons for refusal. Without this information it is difficult to discern whether eligibility criteria were too restrictive or the study simply did not appeal to subjects.

18(E) How many subjects were enrolled in the study by gender, ethnicity and race?

Gender:
_____Males
_____Females
_____ Unknown

Ethnicity:
_____Latinos or Hispanics
_____Not Latinos or Hispanics
_____Unknown

Race:
_____American Indian or Alaska Native
_____Asian
_____Blacks or African American
_____Native Hawaiian or Other Pacific Islander
_____White
_____Other, specify:_____
_____Unknown

18(F) Where was the research study conducted? (List the county where the research study was conducted. If the treatment, prevention and diagnostic tests were offered in more than one county, list all of the counties where the research study was conducted.)

19. Human Embryonic Stem Cell Research. Item 19(A) should be completed for all research projects. If the research project involved human embryonic stem cells, items 19(B) and 19(C) must also be completed.

19(A) Did this project involve, in any capacity, human embryonic stem cells?
_____Yes
___X__No

19(B) Were these stem cell lines NIH-approved lines that were derived outside of Pennsylvania?
_____Yes
_____ No

19(C) Please describe how this project involved human embryonic stem cells:

20. Articles Submitted to Peer-Reviewed Publications.

20(A) Identify all publications that resulted from the research performed during the funding period and that have been submitted to peer-reviewed publications. Do not list journal abstracts or presentations at professional meetings; abstract and meeting presentations should be listed at the end of item 17. **Include only those publications that acknowledge the Pennsylvania Department of Health as a funding source** (as required in the grant agreement). List the title of the journal article, the authors, the name of the peer-reviewed publication, the month and year when it was submitted, and the status of publication (submitted for publication, accepted for publication or published.). Submit an electronic copy of each publication or paper submitted for publication, listed in the table, in a PDF version 5.0.5 (or greater) format, 1,200 dpi. Filenames for each publication should include the number of the research project, the last name of the PI, and an abbreviated title of the publication. For example, if you submit two publications for Smith (PI for Project 01), one publication for Zhang (PI for Project 03), and one publication for Bates (PI for Project 04), the filenames would be:

Project 01 – Smith – Three cases of isolated
Project 01 – Smith – Investigation of NEB1 deletions
Project 03 – Zhang – Molecular profiling of aromatase
Project 04 – Bates – Neonatal intensive care

If the publication is not available electronically, provide 5 paper copies of the publication.

Note: The grant agreement requires that recipients acknowledge the Pennsylvania Department of Health funding in all publications. Please ensure that all publications listed acknowledge the Department of Health funding. If a publication does not acknowledge the funding from the Commonwealth, do not list the publication.

Title of Journal Article:	Authors:	Name of Peer-reviewed Publication:	Month and Year Submitted:	Publication Status (check appropriate box below):
1. None				☐Submitted ☐Accepted ☐Published

20(B) Based on this project, are you planning to submit articles to peer-reviewed publications in the future?

Yes___X____ No_____

If yes, please describe your plans:

The following paper has been submitted and is under review at Cancer Research:

K. Gumireddy, A. Li, D.H. Chang, Q. Liu, J. Yan, R. J. Korst, A. Patel, H.Xu, L. Zhang, G.A.P. Ganepola, L.C. Showe, Q. Huang. AKAP4 Is A Circulating Biomarker For Non-Small Cell Lung Cancer. Clinical Cancer Research, 2015. Submitted.
(Article is not listed in table above because results are confidential until publication.)

Once the analyses of all the samples are finalized, the results will be prepared for review by our collaborators and will be submitted for peer review to additional journals.

21. Changes in Outcome, Impact and Effectiveness Attributable to the Research Project. Describe the outcome, impact, and effectiveness of the research project by summarizing its impact on the incidence of disease, death from disease, stage of disease at time of diagnosis, or other relevant measures of outcome, impact or effectiveness of the research project. If there were no changes, insert "None"; do not use "Not applicable." Responses must be single-spaced below, and no smaller than 12-point type. DO NOT DELETE THESE INSTRUCTIONS. There is no limit to the length of your response.

NONE

22. Major Discoveries, New Drugs, and New Approaches for Prevention Diagnosis and Treatment. Describe major discoveries, new drugs, and new approaches for prevention, diagnosis and treatment that are attributable to the completed research project. If there were no major discoveries, drugs or approaches, insert "None"; do not use "Not applicable." Responses must be single-spaced below, and no smaller than 12-point type. DO NOT DELETE THESE INSTRUCTIONS. There is no limit to the length of your response.

This is still a work in progress. The development of a biomarker panel that is 90% accurate in identifying early stage NSCLC from a blood sample could save lives through early

detection. A relatively non-invasive blood test could simplify or act as an adjunct to repetitive testing by CT for recurrences after lung resection and the way lung suspicious lung nodules are repetitively monitored and potentially reduce costs and radiation exposure.

23. Inventions, Patents and Commercial Development Opportunities.

23(A) Were any inventions, which may be patentable or otherwise protectable under Title 35 of the United States Code, conceived or first actually reduced to practice in the performance of work under this health research grant? Yes_____ No____X_____

If "Yes" to 23(A), complete items a – g below for each invention. (Do NOT complete items a - g if 23(A) is "No.")

a. Title of Invention:

b. Name of Inventor(s):

c. Technical Description of Invention (describe nature, purpose, operation and physical, chemical, biological or electrical characteristics of the invention):

d. Was a patent filed for the invention conceived or first actually reduced to practice in the performance of work under this health research grant?
Yes_____ No____

If yes, indicate date patent was filed:

e. Was a patent issued for the invention conceived or first actually reduced to practice in the performance of work under this health research grant?
Yes_____ No____
If yes, indicate number of patent, title and date issued:
Patent number:
Title of patent:
Date issued:

f. Were any licenses granted for the patent obtained as a result of work performed under this health research grant? Yes_____ No____

If yes, how many licenses were granted?_____

g. Were any commercial development activities taken to develop the invention into a commercial product or service for manufacture or sale? Yes___ No____

If yes, describe the commercial development activities:

23(B) Based on the results of this project, are you planning to file for any licenses or patents, or undertake any commercial development opportunities in the future?

Yes____X_____ No_____

If yes, please describe your plans:

We hold two patents related to these studies. We will seek a partner for commercial development of the technology.

Method for Diagnosing Lung Cancers Using Gene Expression Profiles in Peripheral Blood Mononuclear Cells, US 20140005065

Biomarkers in Peripheral Blood Mononuclear Cells for Diagnosing or Detecting Lung Cancers, US 20110251098

24. Key Investigator Qualifications. Briefly describe the education, research interests and experience and professional commitments of the Principal Investigator and all other key investigators. In place of narrative you may insert the NIH biosketch form here; however, please limit each biosketch to 1-2 pages. *For Nonformula grants only – include information for only those key investigators whose biosketches were not included in the original grant application.*

Biosketches were included in the original application**.**