AD______________

**AWARD NUMBER: W81XWH-14-1-0231**

**TITLE: A Simple System for the Early Detection of Breast Cancer**

**PRINCIPAL INVESTIGATOR:** Stephen Albert Johnston

**CONTRACTING ORGANIZATION:** Arizona State University
Tempe, AZ 85287-6011

**REPORT DATE:** July 2015

**TYPE OF REPORT:** Annual

**PREPARED FOR:** U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

**DISTRIBUTION STATEMENT:** Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-0188*

**1. REPORT DATE**
**July 2015**

**2. REPORT TYPE**
**Annual**

**3. DATES COVERED**
**1Jul2014 - 30Jun2015**

**4. TITLE AND SUBTITLE**

**A Simple System for the Early Detection of Breast Cancer**

**5a. CONTRACT NUMBER**
**W81XWH-14-1-0231**

**5b. GRANT NUMBER**

**5c. PROGRAM ELEMENT NUMBER**

**6. AUTHOR(S)**

**Stephen Albert Johnston**

E-Mail: stephen.johnston@asu.edu

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**

**Arizona State University**
**Tempe, AZ 85287-6011**

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)**

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

**10. SPONSOR/MONITOR'S ACRONYM(S)**

**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

**12. DISTRIBUTION / AVAILABILITY STATEMENT**

Approved for Public Release; Distribution Unlimited

**13. SUPPLEMENTARY NOTES**

**14. ABSTRACT The immunosignature (IMS) technology is a new approach to diagnosis. Because it measures the profile of antibodies in response to disease, it is particularly suited to detect disease early. The goal of this project is to determine if this technology can be useful in the early detection of breast cancer (BC), and if so, what the optimal protocol for its application is. The first Aim is to determine how early BC can be detected by immunosignatures. Our goal was to use a sample size of case and controls that was large enough to overcome the overfitting problems common with biomarker discovery. 240 sera samples from women up to one year before they were diagnosed with stage I cancer were assayed and compared to 500 samples from non-cancer women in the same PLCO cohort. These samples are from 12 different sites. We have run all these samples 2-3 times. Using leave-portions out of the case and controls, we demonstrated that the IMS is capable of ~89% accuracy in diagnosis. We did encounter problems with array consistency and had to exclude ~20% defective ones. We also did not have enough arrays to implement the original design and so the distribution of blinded samples was biased. Unfortunately, these problems lead to a bias in the ims of the unknowns. We will need to execute this experiment again with the improved arrays. In the process we have also developed a simple array based diagnostic based on frame-shift neo-antigens which looks promising. After repeating the year 0-1 experiment, if the results merit, we will proceed in analysis of samples taken 1-4 years before Stage 1 diagnosis, using a commercial source of arrays with less variability. For Aim 2 will we determine if there is a signature distinguishing benign, non-cancer and invasive. We have obtained most of the samples for this study from Duke University and started preliminary analysis. Finally, in Aim 3 we will follow the IMS of the same women over time before they were diagnosed with stage I cancer. The baseline developed from such a time series should greatly help isolate the BC-specific signature.**

**15. SUBJECT TERMS**

**Nothing Listed**

| | | | | | |
|---|---|---|---|---|---|
| **16. SECURITY CLASSIFICATION OF:** | | | **17. LIMITATION OF ABSTRACT** | **18. NUMBER OF PAGES** | **19a. NAME OF RESPONSIBLE PERSON**<br>USAMRMC |
| **a. REPORT**<br>Unclassified | **b. ABSTRACT**<br>Unclassified | **c. THIS PAGE**<br>Unclassified | Unclassified | **25** | **19b. TELEPHONE NUMBER** *(include area code)* |

## TABLE OF CONTENTS

## ABSTRACT:

The immunosignature (IMS) technology is a new approach to diagnosis. Because it measures the profile of antibodies in response to disease, it is particularly suited to detect disease early. The goal of this project is to determine if this technology can be useful in the early detection of breast cancer (BC), and if so, what the optimal protocol for its application is. The first Aim is to determine how early BC can be detected by immunosignatures. Our goal was to use a sample size of case and controls that was large enough to overcome the overfitting problems common with biomarker discovery. 240 sera samples from women up to one year before they were diagnosed with stage I cancer were assayed and compared to 500 samples from non-cancer women in the same PLCO cohort. These samples are from 12 different sites. We have run all these samples 2-3 times. Using leave-portions out of the case and controls, we demonstrated that the IMS is capable of ~89% accuracy in diagnosis. We did encounter problems with array consistency and had to exclude ~20% defective ones. We also did not have enough arrays to implement the original design and so the distribution of blinded samples was biased. Unfortunately, these problems lead to a bias in the ims of the unknowns. We will need to execute this experiment again with the improved arrays. In the process we have also developed a simple array based diagnostic based on frame-shift neo-antigens which looks promising. After repeating the year 0-1 experiment, if the results merit, we will proceed in analysis of samples taken 1-4 years before Stage 1 diagnosis, using a commercial source of arrays with less variability. For Aim 2 will we determine if there is a signature distinguishing benign, non-cancer and invasive. We have obtained most of the samples for this study from Duke University and started preliminary analysis. Finally, in Aim 3 we will follow the IMS of the same women over time before they were diagnosed with stage I cancer. The baseline developed from such a time series should greatly help isolate the BC-specific signature. This project is a by-product of the Aim 1 assays. We have not initiated this Aim. The conclusion of this work should determine the usefulness of IMS for early diagnosis and create a potentially very rich database on the immune responses to cancer and in non-cancer individuals.

## 1. INTRODUCTION:

The purpose is to determine if the immunosignature diagnostic platform can be used for the early detection of breast cancer. One project will use sera samples collected 1-4 years before the diagnosis of stage I breast cancer to determine how early and how accurately breast cancer can be detected. A subset of these samples that were collected in the same woman over time will be examined to determine the value of baseline diagnostics. Finally, benign, invasive and non-cancer samples will be compared for distinctive signatures.

## 2. KEYWORDS:

Immunosignature, diagnosis of breast cancer, early diagnosis, serology, informatics, immunology, benign cancer, prognostic studies, baseline diagnostics, heatmaps, ROCurves

## 3. ACCOMPLISHMENTS:

### What were the major goals of the project?

1. Determine how early breast cancer can be reliably detected by immunosignature.
   (0-15m)(40%)
2. Determine if immunosignatures can distinguish benign growths from invasive tumors and distinguish classes of benign growths.
   (12-20m) (10%)
3. Determine the value of personal baseline immunosignatures in detecting cancer early.
   (20-24m) (0%)

### What was accomplished under these goals?

**Major Activities**

1. Production of IMS arrays: We originally planned on purchasing the arrays for these studies from HealthTell. They ran into production problems and after much delay, produced arrays that were not satisfactory and had to be rejected. Therefore CIM had to gear up to try to produce the 2100 arrays needed to assay the first cohort (740) of the PLCO samples. This occupied a substantial period of time and we only had 6 of the 8 arrays made by April. This was a substantial delay. However, in the process we did learn to produce arrays of better performance than previously.
2. Obtaining PLCO samples: Procuring the PLCO samples is through a regular grant process. We submitted a grant proposal which went to the PLCO study section. Our initial application was not supported so we had to submit an appeal for re-review. Eventually we were allowed to receive the samples. The process took ~ 10 months.
3. IRB approval: We were delayed ~1-2months in obtaining DoD IRB-exempt status to allow us to initiate the experiments. Tragically, the administrator handling our application died and the case was mislaid.
4. Sample assays: All 740 samples (240 case/500 controls) were assayed 2-4 times. The unknowns for testing (50 case/50 controls) were included in this number and assayed at the same time. The unknowns were replicated up to 4 times. Because we had fewer wafers than anticipated and were time constrained, there was a bias in the distribution of samples across wafers. This was the largest single set of samples we have assayed. In this assay, all sera were diluted 1500x in PBST buffer. They were then applied to an array and incubated for 1hr at RT. The arrays were then washed in buffer and then fixed in IPA and dried. The arrays were then incubated for 1hr in buffer with 5nM labeled secondary antibody. They were then washed, dried and scanned in an Innoscan laser scanner. The image files were aligned with the gal-file of the features with peptide identity. This created the raw data files of florescence for each peptide. These files were the starting point for statistical analysis.
5. Analysis: For immunosignatures the diagnostic is defined by looking for a consistent pattern differences between the cases and controls.

**Specific Objectives**:

We aimed in this period to have developed a signature(s) for breast cancer occurring up to one year before diagnosis at stage I. In the process we would learn how to handle such a large and varied data set to create the signature with future sets. Once the signature was established we would use it to call the 100 unknown samples we were provided. These calls would be sent to PLCO and the key revealed. By comparing our calls to the key would establish the statistical characteristics of our diagnostic on these samples.

**Significant Results**:

We have now compiled the raw florescent data on all 740 samples 2-4 times. Since there are 330,000 peptide features per array, this is approximately $7x10^8$ data points. This is a substantial and potentially very valuable data set for other comparisons, particularly the 500 non-cancer cohort that could be analyzed for other characteristics or compared to other populations. This group will also be used as controls for the other cancer year cohorts.

The result of classifying case and controls is presented in Figure 1 as a heatmap and PCA graph. For this presentation, peptides with significant t-test p-values were used to group case and controls with LOOV reiterated. The statistics of classification are based on these tests. The clinical data for these samples are presented below the heatmap. There may be sub-groups in both the case and controls which we will examine in the future. A preliminary examination indicates that a signature distinguishing samples near the time of diagnosis from later can be seen.

Of note, in the PLCO study women were not regularly screened for breast cancer as part of the study. Samples were taken when they were screened for lung, colon or ovarian cancer. The diagnosis of breast cancer was based on individually scheduled screening. So even though there is an immunosignature before these women were diagnosed, they may have been able to be diagnosed if screened by mammography at that time.

In Figure 2 we represent a more rigorous evaluation. 161 case and 436 controls were used (some arrays did not met QC so some case/controls were eliminated). 100 of each were randomly chosen as the training set to create the IMS. This signature was used to call the remaining case and controls. This was repeated 100 times to create the ROC. We felt this would be a good indicator of the performance to expect on the blinded samples. At 95% specificity, sensitivity was ~80%, accuracy 89%, PPV 68%. This is considerably better than most published results on early detection. When the sample names were scrambled the accuracy was 50% (red bar)

We point out that, quite remarkably if this analysis holds, we can identify a diagnostic signature that spans presumably multiple types of breast cancer before diagnosis. We did not group the analysis by any clinical (ER, Her, etc.) data. Going into this study we had two major concerns. One was that there would be multiple, distinct signatures for breast cancer at this point and not an overarching one. This would pose a formidable informatic analysis challenge. The second was that the variability in the non-cancer group would be so large a breast cancer signature would not be detectable. It appears both problems may be overcome.

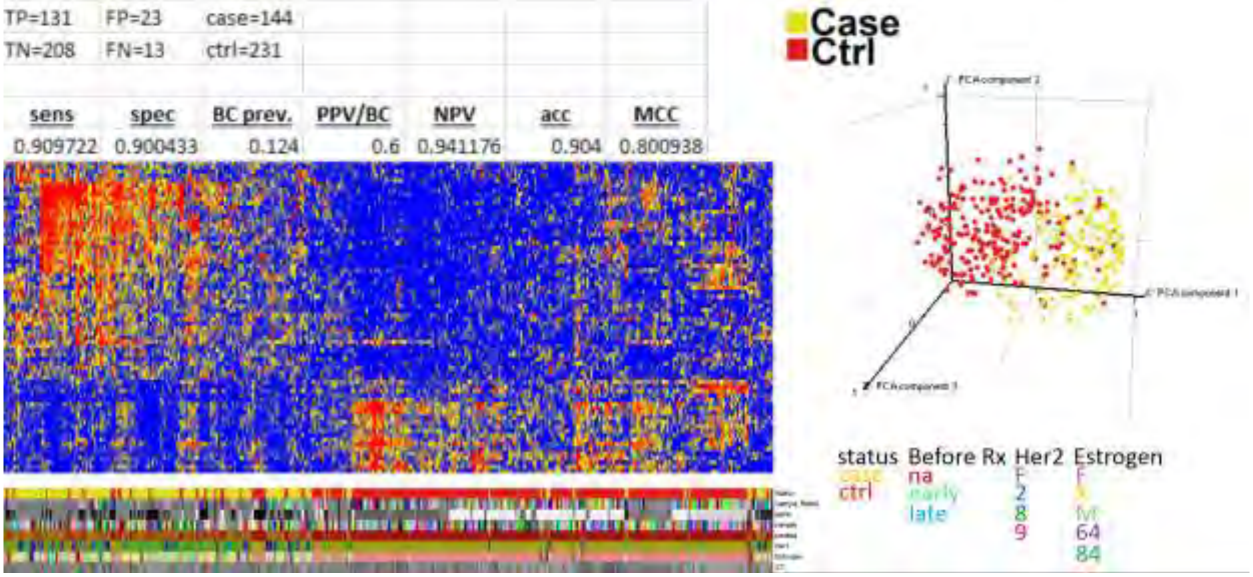
The sequences of the peptides are chosen from random sequence space so cannot be directly aligned to the human proteome. However, we can use the signature peptides in some cases to identify possible motifs that the antibodies recognize. In Figure 3 we show one such alignment. It is possible that these alignments could identify possible vaccine candidates.

When we went to classify the 100 blinded samples we noted a disturbing trend. The unknown samples had a distinctive signature from both the cases and controls. As can be seen in Figure 3, the unknowns are readily distinguished from case+controls. We think the basis of this bias is that because of the shortage of arrays we ran the unknown samples in a biased fashion across the 6 available wafers. Given the wafer variability the classification was biased by wafer rather than disease. We have now produced more stable wafers internally and will have access to commercially produced wafers from HealthTell. We plan to repeat this experiment on these wafers.

In the process of this analysis we invented a new approach to classification, in part born of the concern that the complexity of the signatures, as noted above, would preclude accurate classification. We had previously printed arrays that contained 800 frame-shift peptides that would be generated by trans-splicing or microsatellite insertion/deletions. These were frame-shifts that we had observed or predicted in human and dog tumors.

As shown in Figure 4, 23 controls and 24 cases were analyzed. We could identify 14 peptides that were only positive in controls (ie people with cancer did not react to them) and 22 peptides that were predominately positive in cases (ie people with cancer were more likely to react to these peptides). The accuracy in classification of these samples is encouraging. This is a very small data set and a test set has not been analyzed. We believe it is worth pursuing this type of array in parallel as it is an independent assessment and has several appealing technical aspects.

We were to initiate Aim 2 in year 2. We have received most of the samples for this project from Dr. Marks. We used these experiments in a pilot experiment to determine if benign, malignant and non-cancer sera samples were distinguishable. 30 samples were used of each class. Note, all these samples were run in duplicate from arrays from one wafer. This wafer is from the most recent process.



| TP=131 | FP=23 | case=144 | | | | |
|---|---|---|---|---|---|---|
| TN=208 | FN=13 | ctrl=231 | | | | |
| sens | spec | BC prev. | PPV/BC | NPV | acc | MCC |
| 0.909722 | 0.900433 | 0.124 | 0.6 | 0.941176 | 0.904 | 0.800938 |

Figure 1: *In assaying the PLCO samples on the ASU 330K arrays, it was noted that intra-wafer variability influenced the precision with which the samples were classified. Technical replicates were processed on alternating wafers (each wafer provides 288 individual 330K arrays, or twelve 24-up slides). In order to accommodate the wafer variance, a quality control program was used (appendix X lists the 'R' code). The program lists wafers from best to worst, using data distributions, correlations across samples, finally ranked by a root-mean-squared measure of these two QC terms. This figure uses 5 of the 6 wafers on which ASU processed samples. 375 total samples were used, 231 control and 144 case, as selected by the ASU QC program (see*

*Appendix). Control samples are colored red in the heatmap legend, case samples are colored yellow in the legend. A t-test was used to obtain 150 peptides that distinguished case from control, with the minimum p-value of 6.21x10$^{-14}$ and the maximum p-value of 1.51x10$^{-8}$. These peptides were used to cross-validate the samples by leaving one sample from each of the two classes out, and predicting the class of the remaining samples. This was done until every sample had been left out at least once, yielding the performance shown in the figure. 23 controls were called case (false positive) and 13 cases were called control (false negative), yielding a sensitivity of 0.91, a specificity of 0.90 and a cumulative accuracy of 0.90. MCC is the Matthews correlation coefficient, used to conservatively estimate binary classification performance of unbalanced cohorts yielding a 0.80 score. Prevalence of breast cancer is 12.4%, and was used to estimate the positive predictive value of 0.60 in a population.*

*Heatmap displays 375 samples, with case shown in yellow, control in red. Each peptide is colored by intensity, with blue corresponding to low intensity, red corresponding to high intensity. The Principal Components map to the right shows every sample colored identically as the heatmap. The colored bars below the heatmap show the status (case vs. control), unique sample name, wafer number, replicated sample (when possible), time from diagnosis broken into early (200 or more days before diagnosis) or late (199 or fewer days before diagnosis), Her 2 status, estrogen status, and QC rank.*

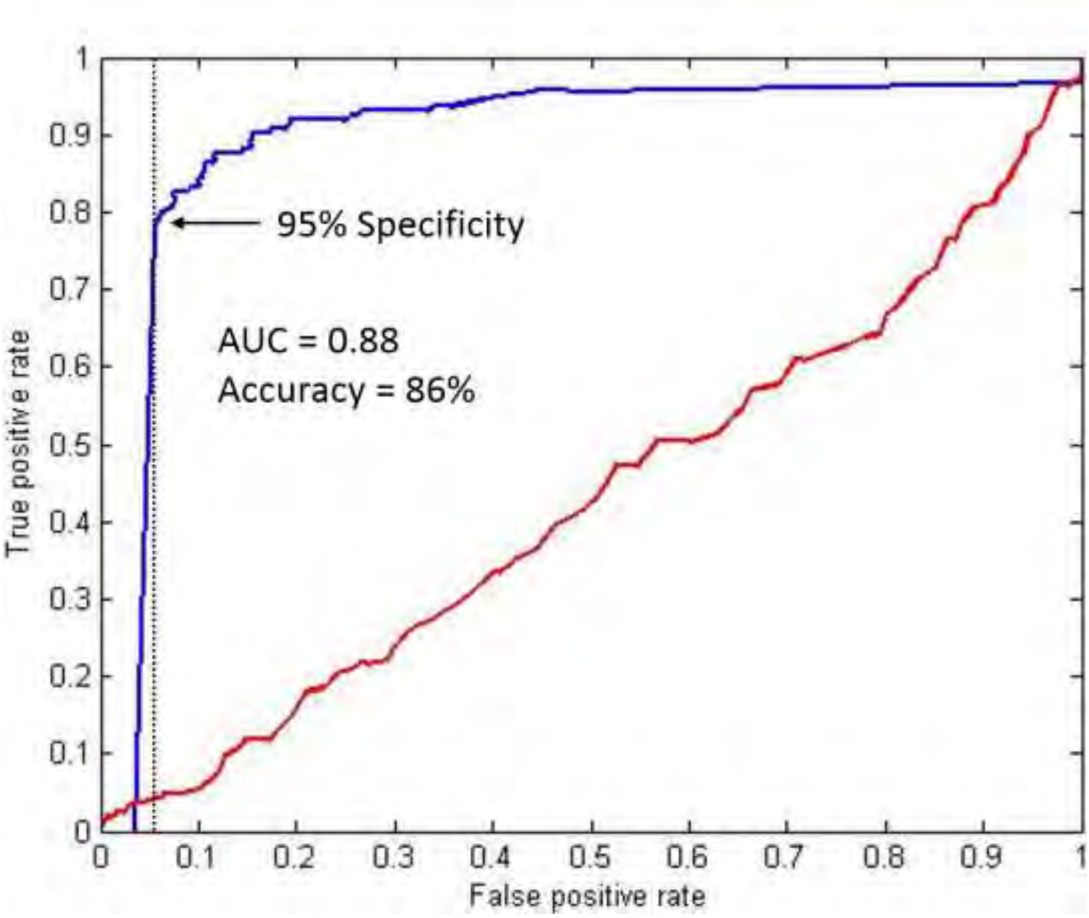
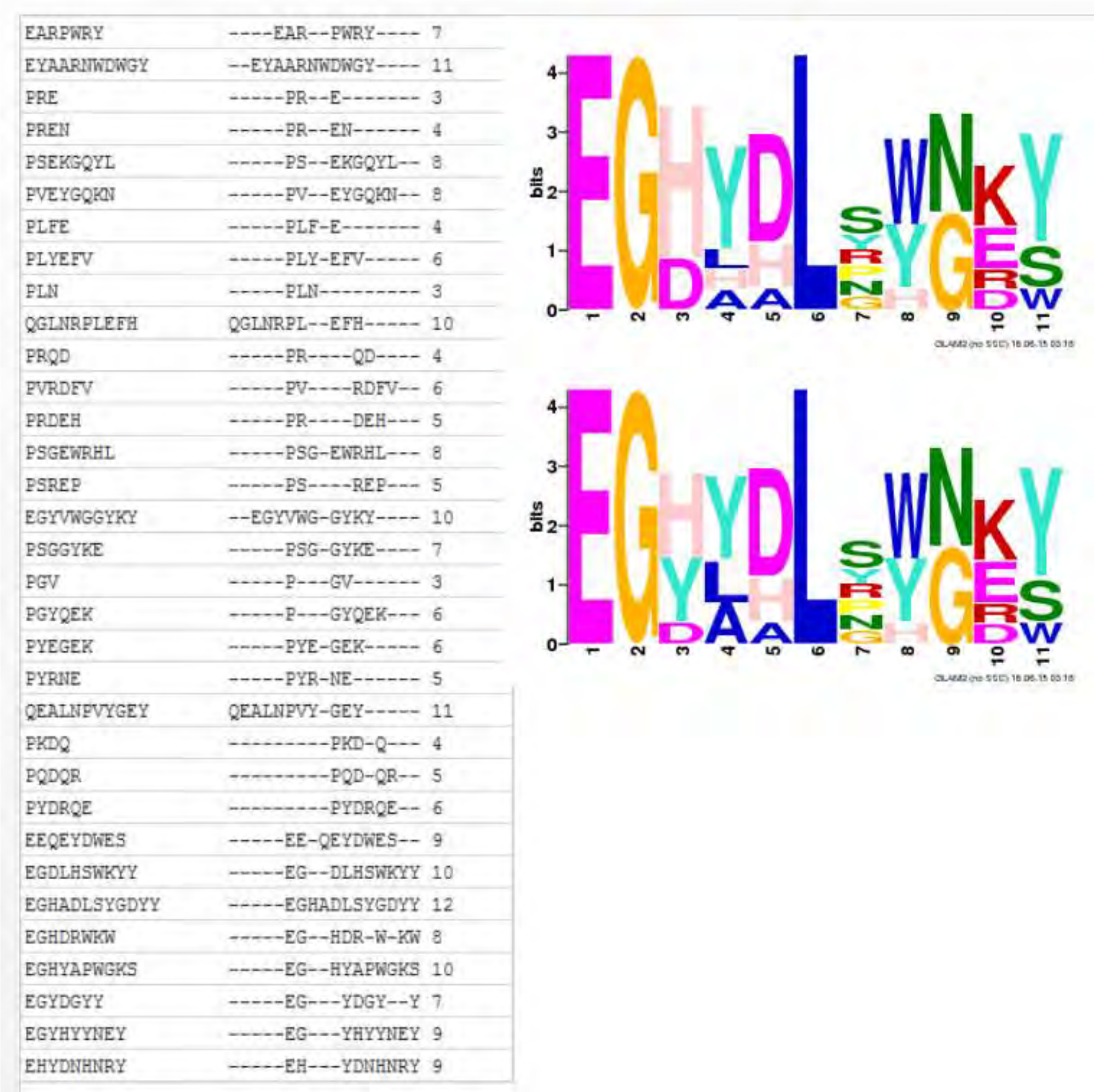


Figure 2: *Receiver-Operator Curve (ROCurve) of 640 samples consisting of 190 case and 450 controls. These samples were assayed on arrays containing 330,000 peptides. Many of the*

*samples were assayed with technical replicates. ~20% of the arrays did not pass QC (an unbiased combination of mean correlation across all arrays and mean Kolmogorov-Smirnov test across all arrays – see Appendix) and the information from those arrays was not used. The remaining arrays represent 161 unique case and 436 unique control samples. When duplicates were present, the better of the two was used based on the QC score. 100 case and 100 control samples were then randomly selected and used as the training set. 100 features (peptides) were then selected via a 2-sided t-test using the training set. Then, an SVM classifier was built using these features and the training set. The remaining 61 case and 336 control samples which were not used in the feature selection (blinded) were then called using this classifier. The process was then repeated 300 times, each time selecting another 100 case and 100 control samples randomly to re-train a new classifier. The result was a score for each sample based on the fraction of the time that it was called correctly. This score was used to build this ROCurve. The blue curve was calculated for the samples with labels intact. The red curve was calculated after scrambling the labels. Note that the blue curve does not go through (0,0) or (1,1) as one might expect for a standard biomarker. This is because there were a few cases that were scored as controls 100% of the time and a few controls that were scored as cases 100% of the time, perhaps due to a biological rationale rather than a mathematical rationale.*



```
EARPWRY         ----EAR--PWRY---- 7
EYAARNWDWGY     --EYAARNWDWGY---- 11
PRE             -----PR--E------- 3
PREN            -----PR--EN------ 4
PSEKGQYL        -----PS--EKGQYL-- 8
PVEYGQKN        -----PV--EYGQKN-- 8
PLFE            -----PLF-E------- 4
PLYEFV          -----PLY-EFV----- 6
PLN             -----PLN--------- 3
QGLNRPLEFH      QGLNRPL--EFH----- 10
PRQD            -----PR----QD---- 4
PVRDFV          -----PV----RDFV-- 6
PRDEH           -----PR----DEH--- 5
PSGEWRHL        -----PSG-EWRHL--- 8
PSREP           -----PS----REP--- 5
EGYVWGGYKY      --EGYVWG-GYKY---- 10
PSGGYKE         -----PSG-GYKE---- 7
PGV             -----P---GV------ 3
PGYQEK          -----P---GYQEK--- 6
PYEGEK          -----PYE-GEK----- 6
PYRNE           -----PYR-NE------ 5
QEALNPVYGEY     QEALNPVY-GEY----- 11
PKDQ            ---------PKD-Q--- 4
PQDQR           ---------PQD-QR-- 5
PYDRQE          ---------PYDRQE-- 6
EEQEYDWES       -----EE-QEYDWES-- 9
EGDLHSWKYY      -----EG--DLHSWKYY 10
EGHADLSYGDYY    -----EGHADLSYGDYY 12
EGHDRWKW        -----EG--HDR-W-KW 8
EGHYAPWGKS      -----EG--HYAPWGKS 10
EGYDGYY         -----EG---YDGY--Y 7
EGYHYYNEY       -----EG---YHYYNEY 9
EHYDNHNRY       -----EH---YDNHNRY 9
```

Figure 3: *Peptides from Figure 1 were used in CLUSTALW to identify subgroups of similar peptides, which were then examined by GLAM2 (Gapped Local Alignment Method) to produce a set of possible antigen motifs. If a linear peptide acts as a tumor antigen, this method may find a motif of the eliciting tumor (manuscript in press for Cancer Informatics). The left column above is the CLUSTAL W result, the right image was produced by GLAM2, and represents the consensus motif likely to be found in a linear tumor antigen. EG and L are absolutely required at their specific positions, the other positions indicate a degree of flexibility by the antibody*

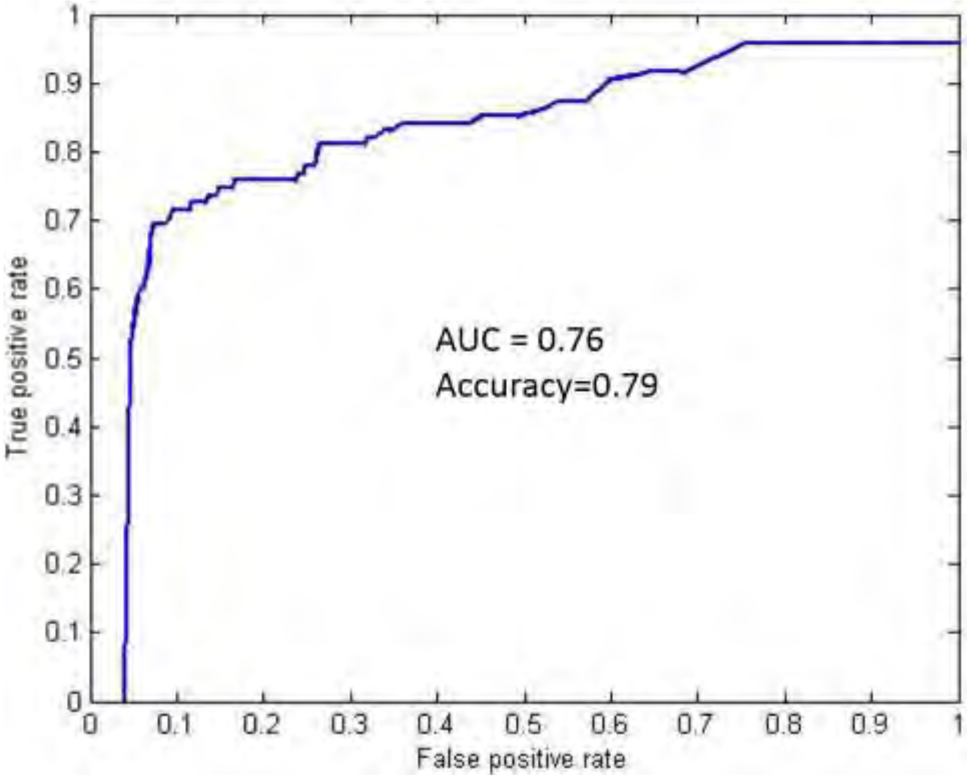
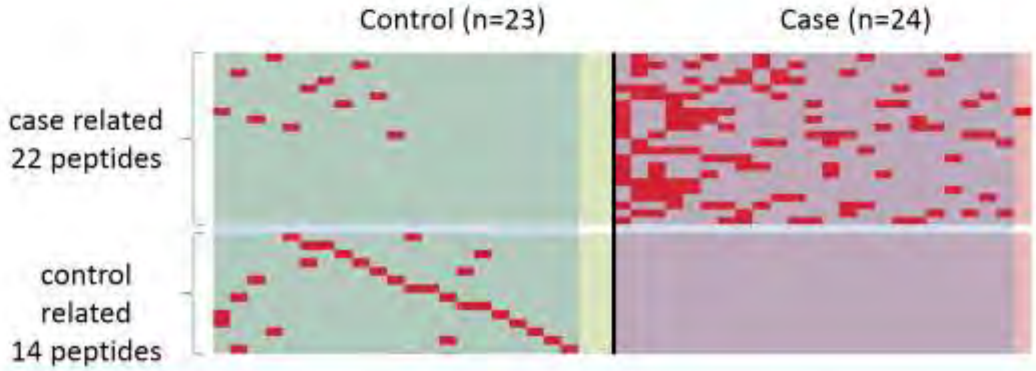


Figure 4: *This ROC represents the difference between 'blinded' and all other training samples. Here, 95 blinded samples are called 'case' and the 597 unblinded cancer and non-cancer training samples were called 'control'. 75 'case' and 'control' samples were subsequently used to train, then predict the classes. As before, these training samples were selected randomly and independently 300 times. The ROCurve was calculated by changing the cutoff for the fraction each sample must be scored as 'case' in order to be considered case. Surprisingly, the blinded samples themselves form a strong and independent group from the known unblinded case and controls, suggesting a sample bias (blinded samples were not randomly chosen from a composite set of case and controls, samples were handled differently) or a statistical anomaly forced the blinded samples to behave differently as a group from the training samples.*

## Frame-shift Peptide Diagnostic Arrays Applied to BC Samples



| | peptide set | criterion | sensitivity (%) | specificity (%) | % of total call |
|---|---|---|---|---|---|
| to call case | case related | ≥2 positive peptide | 95.8 | 100 | 93.6 |
| to call control | control related | ≥1 positive peptide | 91.3 | 100 | |

Figure 5: *Arrays of 800 frame-shift peptides spotted on NSB surfaces were used to screen 23 control and 24 case sera samples. NSB Pos-tech creates a dendrimer on which peptides are attached. This dendrimer ensures a precise 3nm spacing between peptides. Peptides were chosen with a cut-off of mean +2 fold of the STDEV of the control and is required to be >10% positive in all tested cases. Note that these are selected peptides and should be independently tested before these can be called 'diagnostic'.*

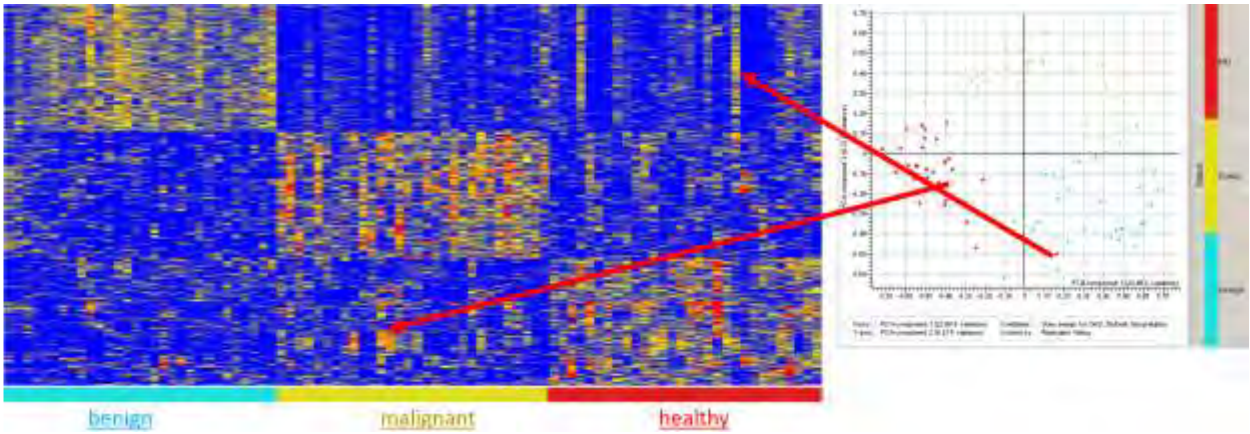


Figure 6: *Samples from Duke University (Dr. Jeffrey Marks) were processed on 330K arrays using clinical classification of "benign" (signifying a positive mammography, but no cancer detected), "malignant" (signifying a positive mammography and an aggressive form of breast cancer detected), and "healthy" (signifying no positive mammography, no other disease). 150 peptides were selected that discriminated the 3 classes. On the left a heatmap shows the relative*

*intensity of each peptide (row) for each patient (column). The benign group, although distinct from healthy in this analysis, can also be used in a two-class feature selection (data not shown) vs. malignant indicating a greater overlap in antibodies between healthy and benign than between healthy and malignant or benign and malignant. The right figure is a PCA showing where individual outliers fall relative to the heatmap. Two of the most obvious outliers are shown. Year 2 samples are currently at ASU awaiting processing.*

**What opportunities for training and professional development has the project provided?**
*If the project was not intended to provide training and professional development opportunities or there is nothing significant to report during this reporting period, state "Nothing to Report."*

*Describe opportunities for training and professional development provided to anyone who worked on the project or anyone who was involved in the activities supported by the project. "Training" activities are those in which individuals with advanced professional skills and experience assist others in attaining greater proficiency. Training activities may include, for example, courses or one-on-one work with a mentor. "Professional development" activities result in increased knowledge or skill in one's area of expertise and may include workshops, conferences, seminars, study groups, and individual study. Include participation in conferences, workshops, and seminars not listed under major activities.*

Approximately 7 undergraduate or recent graduates work in the core on the array assays and data analysis. The dataset is rich and will be used to introduce new students to early detection of disease. Also, the complexity of the data, the opportunity to mine cancer-specific epitopes, and the opportunities to recruit off-site collaborators will be great once this dataset is made public.

**How were the results disseminated to communities of interest?**
*If there is nothing significant to report during this reporting period, state "Nothing to Report."*

*Describe how the results were disseminated to communities of interest. Include any outreach activities that were undertaken to reach members of communities who are not usually aware of these project activities, for the purpose of enhancing public understanding and increasing interest in learning and careers in science, technology, and the humanities.*

The data will be made public once validated, and publications in multidisciplinary journals is a high priority. Students working on the project are encouraged to present at scientific meetings.

**What do you plan to do during the next reporting period to accomplish the goals?**
*If this is the final report, state "Nothing to Report."*

*Describe briefly what you plan to do during the next reporting period to accomplish the goals and objectives.*

For Aim1 we will first solve the problem of the unknown samples' unusual behavior, or in liue we would actively reconstruct the blinded test from the PLCO for the year 0-1 cohort and re-assay. This data will be sent to PLCO to support the release of the year 1-4 samples. We will test the new HealthTell arrays and if they are as good as we expect from their current data, we will analyze the year 0-4 samples on these arrays. This data would be reported to PLCO and published to complete Aim 1. For Aim 2 we have already received the samples from Duke and have the regulatory approvals. We have initiated the analysis of these samples and will complete within two months. These samples include 63 atypical hyperplasia, 43 DCIS, 48 invasive cancer, 40 normals, 83 proliferative benign conditions, and 46 proliferative benign conditions. For Aim 3 the samples are imbedded in the PLCO samples so the completion is an informatics analysis.

**4. IMPACT:** Describe distinctive contributions, major accomplishments, innovations, successes, or any change in practice or behavior that has come about as a result of the project relative to:

**What was the impact on the development of the principal discipline(s) of the project?**
*If there is nothing significant to report during this reporting period, state "Nothing to Report."*

*Describe how findings, results, techniques that were developed or extended, or other products from the project made an impact or are likely to make an impact on the base of knowledge, theory, and research in the principal disciplinary field(s) of the project. Summarize using language that an intelligent lay audience can understand (Scientific American style).*

1. We demonstrated that the immunosignature diagnostic platform could obtain exceptional specificity, sensitivity and accuracy on pre-diagnostic samples by blinding case and control samples.
2. We demonstrated that a pan-breast cancer signature is obtainable.
3. We created a new assay procedure that optimizes our ability to control the primary antibody binding physically distinct from the secondary antibody detection, enabling optimal conditions for primary binding (patient sera) and secondary detection, separately. This protocol also speeds the assay process enabling higher throughput.
4. We invented a new type of signature approach, frameshift signatures that could complement the immunosignature diagnostic.

**What was the impact on other disciplines?**

Epitope mining of frameshift peptides produced by breast cancer tumors is enhanced by the large dataset provided. Bioinformatic procedures for finding these epitopes from random peptide sequences is in its infancy but is developing quickly, due in part to the diverse and large sample set provided by PLCO.

*If there is nothing significant to report during this reporting period, state "Nothing to Report."*

*Describe ways in which the project made an impact, or is likely to make an impact, on commercial technology or public use, including:*

- *transfer of results to entities in government or industry;*
- *instances where the research has led to the initiation of a start-up company; or*
- *adoption of new practices.*

Invention disclosures were filed on the two inventions noted above.
If this approach is successful we envision the company HealthTell will incorporate it into its CLIA laboratory assays for the early detection of breast cancer.

**What was the impact on society beyond science and technology?**
*If there is nothing significant to report during this reporting period, state "Nothing to Report."*

*Describe how results from the project made an impact, or are likely to make an impact, beyond the bounds of science, engineering, and the academic world on areas such as:*

- *improving public knowledge, attitudes, skills, and abilities;*
- *changing behavior, practices, decision making, policies (including regulatory policies), or social actions; or*
- *improving social, economic, civic, or environmental conditions.*

If successful the concept of early detection and treatment of cancer would have wide reaching implications.

5. **CHANGES/PROBLEMS:** The PD/PI is reminded that the recipient organization is required to obtain prior written approval from the awarding agency grants official whenever there are significant changes in the project or its direction. If not previously reported in writing, provide the following additional information or state, "Nothing to Report," if applicable:

**Changes in approach and reasons for change**
*Describe any changes in approach during the reporting period and reasons for these changes. Remember that significant changes in objectives and scope require prior approval of the agency.*

We did not change any of goals of the proposal. The lack of arrays from HealthTell was a significant delay but our proposal had listed a contingency plan that we would make the arrays.

**Actual or anticipated problems or delays and actions or plans to resolve them**

*Describe problems or delays encountered during the reporting period and actions or plans to resolve them.*

One delay was production of the arrays. This was solved by setting up and producing the arrays in CIM. This was the rate-limiting step initially. Going forward HealthTell will produce more consistent arrays.

The second delay was in getting DoD IRB approval. This took longer than expected due to administrative problems at DoD. We have the approval so this will not be a problem going forward.

One other aspect that took longer than expected was approvals and receipt of samples from PLCO. We now know better how this system works and this should not be a problem in the future.

Finally, the data analysis was more complicated and time-consuming than anticipated. We will allow more time for this in the future and build on the experience we are gaining.

**Changes that had a significant impact on expenditures**

*Describe changes during the reporting period that may have had a significant impact on expenditures, for example, delays in hiring staff or favorable developments that enable meeting objectives at less cost than anticipated.*

Nothing to report.

**Significant changes in use or care of human subjects, vertebrate animals, biohazards, and/or select agents**

*Describe significant deviations, unexpected outcomes, or changes in approved protocols for the use or care of human subjects, vertebrate animals, biohazards, and/or select agents during the reporting period. If required, were these changes approved by the applicable institution committee (or equivalent) and reported to the agency? Also specify the applicable Institutional Review Board/Institutional Animal Care and Use Committee approval dates.*

**Significant changes in use or care of human subjects**

Nothing to report.

**Significant changes in use or care of vertebrate animals**

Nothing to report

**Significant changes in use of biohazards and/or select agents**

Nothing to report.

6. **PRODUCTS:** List any products resulting from the project during the reporting period. If there is nothing to report under a particular item, state "Nothing to Report."

- **Publications, conference papers, and presentations**
  Report only the major publication(s) resulting from the work under this award.

  **Journal publications.** *List peer-reviewed articles or papers appearing in scientific, technical, or professional journals. Identify for each publication: Author(s); title; journal; volume: year; page numbers; status of publication (published; accepted, awaiting publication; submitted, under review; other); acknowledgement of federal support (yes/no).*

  Nothing to report

  **Books or other non-periodical, one-time publications.** *Report any book, monograph, dissertation, abstract, or the like published as or in a separate publication, rather than a periodical or series. Include any significant publication in the proceedings of a one-time conference or in the report of a one-time study, commission, or the like. Identify for each one-time publication: author(s); title; editor; title of collection, if applicable; bibliographic information; year; type of publication (e.g., book, thesis or dissertation); status of publication (published; accepted, awaiting publication; submitted, under review; other); acknowledgement of federal support (yes/no).*

  Nothing to report

  **Other publications, conference papers and presentations**. *Identify any other publications, conference papers and/or presentations not reported above. Specify the status of the publication as noted above. List presentations made during the last year (international, national, local societies, military meetings, etc.). Use an asterisk (*) if presentation produced a manuscript.*

  Nothing to report

  *activities. A short description of each site should be provided. It is not necessary to include the publications already specified above in this section.*

  Nothing to report

- **Technologies or techniques**
  *Identify technologies or techniques that resulted from the research activities. Describe the technologies or techniques were shared.*

Invented new assay procedure as described above. Disclosure filed and used in this project. Not published yet.

Invented new diagnostic arrays (Frameshift signatures). Disclosed and using in this project. Not published yet.

- **Inventions, patent applications, and/or licenses**
  *Identify inventions, patent applications with date, and/or licenses that have resulted from the research. Submission of this information as part of an interim research performance progress report is not a substitute for any other invention reporting required under the terms and conditions of an award.*

Nothing to report

- **Other Products**
  *Identify any other reportable outcomes that were developed under this project. Reportable outcomes are defined as a research result that is or relates to a product, scientific advance, or research tool that makes a meaningful contribution toward the understanding, prevention, diagnosis, prognosis, treatment and /or rehabilitation of a disease, injury or condition, or to improve the quality of life. Examples include:*
  - *data or databases;*
  - *physical collections;*
  - *educational aids or curricula;*
  - *instruments or equipment;*
  - *research material (e.g., Germplasm; cell lines, DNA probes, animal models);*
  - *clinical interventions;*
  - *new business creation; and*
  - *other.*

A large database on the immunosignatures of 240 cases (up to one year before stage I diagnosis) and 500 non-cancer controls has been created.

## 7. PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS

### What individuals have worked on the project?

*Provide the following information for: (1) PDs/PIs; and (2) each person who has worked at least one person month per year on the project during the reporting period, regardless of the source of compensation (a person month equals approximately 160 hours of effort). If information is unchanged from a previous submission, provide the name only and indicate "no change".*

Name: Stephen Johnston
Project Role: Principle Investigator
Researcher Identifier (e.g. ORCID ID):
Nearest person month worked: 3

Contribution to Project: Responsible for the implementation of the research and interactions between ASU, DUMC and PLCO. Has participated in all phases of the work and is responsible for presentation of work in publications and milestone meetings.
Funding Support:

Name: Neal Woodbury
Project Role: Co-Investigator
Researcher Identifier (e.g. ORCID ID):
Nearest person month worked: 1.2

Contribution to Project: Responsible for modeling and analytical development for the immunosignaturing process. Involved in the data analysis and optimization of immunosignatures.
Funding Support:

Name: Phillip Stafford
Project Role: Co-Investigator
Researcher Identifier (e.g. ORCID ID):
Nearest person month worked: 3

Contribution to Project: Responsible for processing of samples through the Peptide Array Core and the bioinformatics analysis of the data.
Funding Support:

Name: Jeffrey Marks
Project Role: PI of Subcontract to Duke University
Researcher Identifier (e.g. ORCID ID):
Nearest person month worked: 1.2

Contribution to Project: Responsible for assembly of samples and subject data to be analyzed by ASU.

Funding Support:

Name: H. Kim Lyerly
Project Role: Investigator on the subcontract to Duke University
Researcher Identifier (e.g. ORCID ID):
Nearest person month worked: .6

Contribution to Project: Participation in the design and interpretation of study.
Funding Support:

**Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?**
*If there is nothing significant to report during this reporting period, state "Nothing to Report."*

*If the active support has changed for the PD/PI(s) or senior/key personnel, then describe what the change has been. Changes may occur, for example, if a previously active grant has closed and/or if a previously pending grant is now active. Annotate this information so it is clear what has changed from the previous submission. Submission of other support information is not necessary for pending changes or for changes in the level of effort for active support reported previously. The awarding agency may require prior written approval if a change in active other support significantly impacts the effort on the project that is the subject of the project report.*

DHS Contract HSHQDC-15-C-B0008 Beginning 5/1/2015 Ending 6/30/2016
Johnston 3.25 months, Woodbury 3.25 months, Stafford 6.5 months

NSF MCB-1518528 Beginning 12/15/2014 Ending 11/30/2015
Woodbury .12 months

*Describe partner organizations – academic institutions, other nonprofits, industrial or commercial firms, state or local governments, schools or school systems, or other organizations (foreign or domestic) – that were involved with the project. Partner organizations may have provided financial or in-kind support, supplied facilities or equipment, collaborated in the research, exchanged personnel, or otherwise contributed.*

*Provide the following information for each partnership:*
*Organization Name:*
*Location of Organization: (if foreign location list country)*
*Partner's contribution to the project (identify one or more)*

- *Financial support;*
- *In-kind support (e.g., partner makes software, computers, equipment, etc., available to project staff);*
- *Facilities (e.g., project staff use the partner's facilities for project activities);*
- *Collaboration (e.g., partner's staff work with project staff on the project);*

- *Personnel exchanges (e.g., project staff and/or partner's staff use each other's facilities, work at each other's site); and*
- *Other.*

PLCO, EEMS
Claire Zhu, Program Director
Early Detection Research Group
Division of Cancer Prevention
National Cancer Institute
9609 Medical Center Drive
Room 5E106
Rockville, MD  20850

Provided specimens for testing.

## 8. SPECIAL REPORTING REQUIREMENTS

**COLLABORATIVE AWARDS:** For collaborative awards, independent reports are required from BOTH the Initiating Principal Investigator (PI) and the Collaborating/Partnering PI. A duplicative report is acceptable; however, tasks shall be clearly marked with the responsible PI and research site. A report shall be submitted to **https://ers.amedd.army.mil** for each unique award.

**QUAD CHARTS:** If applicable, the Quad Chart (available on https://www.usamraa.army.mil) should be updated and submitted with attachments.

## 9. APPENDICES: Attach all appendices that contain information that supplements, clarifies or supports the text. Examples include original copies of journal articles, reprints of manuscripts and abstracts, a curriculum vitae, patent applications, study questionnaires, and surveys, etc.

**Appendix:**

In order to complete an analysis of the ASU dataset exactly as ASU did, the following code should be used to quality-filter the arrays. The cutoff of 375 arrays was used for Figure 1, but this is not likely to be optimal. The following code should be run in 'R' using the complete dataset of 1440 arrays from ASU. Rank is supplied by the program as an integer with 1 being the best, 1440 being the worst.

```
sortArray_ASU_QC <-
# sortArray_ASU_QC() rank orders immunosignaturing arrays by similarity
# QC <- sortArray_ASU_QC(FG,arraylimit=10,pepsmpl=20,cor.scale=0.2,ks.scale,
#   logFG=FALSE,method="pearson")
#
# ARGUMENTS:
#    FG: numeric matrix of foreground spot intensities from an immunosignaturing
#       array, with peptide features in rows and arrays in columns.  It is
#       assumed that the columns have names related to the arrays, though this
#       is not strictly required.  You may wish to remove particular sequences
#       (for example, control peptides) from FG.
#    arraylimit: integer; minimum number of arrays to compare-- the stopping
#       point for the algorithm.
#    pepsmpl: integer; fraction of peptides to sample, where 20 means every
#       20th; pepsmpl=1 will use the entire set.
#    cor.scale: number; correlation scaling factor for calculating distance;
#       default value is taken from the ASU Matlab implementation.  Set to NA
#       to use method analogous to the default ks.scale approach (see below).
#    ks.scale: number; scaling factor to apply to mean KS p-values when
#       calculating the QC distance.  If omitted, it will be calculated as in
#       the ASU implementation: max(mean.ks) - mean(mean.ks), 1st iteration.
#       Set to Inf to use correlation only; note that the KS test takes up 99%
#       of the run time.
#    logFG: logical; if TRUE, log-transform the FG matrix after normalization
#    method: character; method argument passed to the cor() function.  Use
#       "spearman" for rank correlation.
#
# VALUE; returns a data.frame with rownames taken from the column names of FG,
#    ordered by the rank column, and columns:
#      rank: rank order of the arrays, from best (1) to worst (2)
#      mean.cor: mean Pearson correlation to arrays with lower rank values
#      mean.ks: mean log10(p-value), KS test of distribution, to arrays with
#         lower rank values
#      qc.dist: the distance calculated from scaled mean.cor and mean.ks
#
# - The algorithm iteratively compares the average "distances" between arrays
#   and removes the array farthest from the others in succession, stopping when
#   only arraylimit arrays remain.  The distance is calculated as a Euclidean
#   distance based on the mean Pearson correlation between each array and all
#   the others, and the mean log10(p-value) for the similarity of distributions
#   between each array and all others by 2-sample Kolmogorov-Smirnov test.
#   Each of these factors is scaled by cor.scale and ks.scale, respectively.
# - By default, only every 20th peptide is used in the distance calculations.
#   Calculation time scales ~linearly with 1/pepsmpl, but may give different
#   results.  However, it may be possible to use a large pepsmpl to quickly
#   throw out the worst arrays, then rerun using all or most features
# - Each array's intensities are normalized by their median value.  It appears
#   that intensities are not log- or otherwise transformed before being passed
#   in.
#
# Developed by Phillip Stafford and Neil Woodbury, 2013
# - added logFG and method arguments; fixed inconsistency in argument ordering
#   between the documentation and implementation
# - fixed normalization bug
# - moved pairwise correlation and KS p-value calculations out of the loop to
#   vastly improve performance
# - added waitbar to track progress on the KS bit
function(FG,arraylimit=10,pepsmpl=20,cor.scale=20,ks.scale,logFG=FALSE,
   method="pearson"){
   # median-normalize the intensities
```

```
n.arrays <- ncol(FG)
# normalize, and log-transform the data if logFG is TRUE
meds <- apply(FG,2,median,na.rm=T)
FG <- t(t(FG)/meds)
if(logFG) FG <- log(FG+1)
sset <- seq(1,nrow(FG),pepsmpl)  # sample of peptides to use
# calculate pairwise correlation and KS p-values ONCE
cors <- cor(FG[sset,],use="pair",method=method)
cors[seq(1,length(cors),nrow(cors)+1)] <- NA
p.ks <- matrix(NA,n.arrays,n.arrays)
skip.KS <- ifelse(missing(ks.scale),FALSE,is.infinite(ks.scale))
if(skip.KS){
  # skip the KS test
  p.ks[T] <- 1
}else{
  # set up progress bar
  n.todo <- (n.arrays-1)*n.arrays/2
  n.done <- 0
  success <- require("tcltk")
  if(success){
    waitbar <- tkProgressBar("Comparing distributions...",label="",0)
    on.exit(close(waitbar))
  }
  t0 <- Sys.time()
  # calculate KS-GOF p-value for each pair of arrays
  for(Ri in 1:(n.arrays-1)){
    for(Ci in (Ri+1):n.arrays){
      p.ks[Ri,Ci] <- p.ks[Ci,Ri] <- suppressWarnings(
        log10(ks.test(FG[sset,Ri],FG[sset,Ci])$p.val+1e-320))
      n.done <- n.done + 1
      if(success & n.done%%floor(n.todo/20)==0){
        # update the progress bar
        now <- Sys.time()
        elapsed <- round(difftime(now,t0,units="auto"),1)
        left <- round(elapsed * (n.todo - n.done)/n.done,1)
        msg <- paste(n.done,sep="","/",n.todo," in ",elapsed," ",
          attr(elapsed,"units")," (~",left," ",attr(elapsed,"units"),
          " left)")
        setTkProgressBar(waitbar,n.done/n.todo,label=msg)
      }
    }
  }
}
# set the scaling parameters for the QC distance metric
if(is.na(cor.scale)){
  mean.cors <- apply(cors,2,mean,na.rm=T)
  max.cor <- max(mean.cors)
  cor.scale <- max.cor - mean(mean.cors)
}
if(missing(ks.scale)){
  mean.ks <- apply(p.ks,2,mean,na.rm=T)
```

```
    max.ks <- max(mean.ks)
    ks.scale <- max.ks - mean(mean.ks)
  }
  # data.frame to keep track of order in which arrays were dropped, stats
  QC <- data.frame(rank=rep(NA,n.arrays),mean.cor=NA,mean.ks=NA,qc.dist=NA,
    row.names=colnames(FG))
  # iteratively find the array that differs most from the others and remove
  while(sum(is.na(QC$rank))>=arraylimit){
    # get subset of arrays remaining (most similar)
    ndxs <- which(is.na(QC$rank))
    nleft <- length(ndxs)
    # calculate mean correlation, KS p-value to remaining arrays
    mean.cors <- apply(cors[ndxs,ndxs],2,mean,na.rm=T)
    max.cor <- max(mean.cors,na.rm=T)
    mean.ks <- apply(p.ks[ndxs,ndxs],2,mean,na.rm=T)
    max.ks <- max(mean.ks)
    # calculate the distance metric
    qc.dist <- sqrt(((max.ks - mean.ks)/ks.scale)^2
        + ((max.cor - mean.cors)/cor.scale)^2)
    if(nleft>arraylimit){
      # drop the array with the greatest distance from the others
      far.out <- which.max(qc.dist)
      QC[ndxs[far.out],] <- list(rank=nleft,mean.cor=mean.cors[far.out],
        mean.ks=mean.ks[far.out],qc.dist=qc.dist[far.out])
    }else{
      # use the current stats to rank the remaining arrays
      far.out <- order(qc.dist)
      QC[ndxs[far.out],] <- list(rank=1:nleft,mean.cor=mean.cors[far.out],
        mean.ks=mean.ks[far.out],qc.dist=qc.dist[far.out])
    }
  }
  # return the summary of array stats, ordered from best to worst (1st dropped)
  QC <- QC[order(QC$rank),]
  return(QC)
}
```