AD____________________

Award Number: W81XWH-06-1-0100

TITLE: Affinity-Based Serum Proteomics for Ovarian Cancer Early Diagnosis

PRINCIPAL INVESTIGATOR: Martin W. McIntosh, Ph.D.

CONTRACTING ORGANIZATION: Fred Hutchinson Cancer Research Center
Seattle, Washington 98109-1024

REPORT DATE: December 2006

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 *(DD-MM-YYYY)* 01-12-2006	**2. REPORT TYPE** Annual	**3. DATES COVERED** *(From - To)* 15 Nov 2005 – 14 Nov 2006

4. TITLE AND SUBTITLE

Affinity-Based Serum Proteomics for Ovarian Cancer Early Diagnosis

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-06-1-0100

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Martin W. McIntosh, Ph.D.

E-Mail: mmcintos@fhcrc.org

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

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

Fred Hutchinson Cancer Research Center
Seattle, Washington 98109-1024

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

Our research project is intended to exploit unique characteristics of phage and yeast recombinant antibodies as the basis for a serum biomarker discovery platform for ovarian cancer. In brief, we select from large recombinant libraries those binding sequences which bind to cancer related material but not to control serum, then we evaluate these sub libraries in high throughput using novel recombinant antibody arrays probed with serum from our serum repository. At present, we are on track based on our initial proposal. We have (1) selected a well-balanced group of cases (serum and proximal fluid) and controls for our initial discovery, (2) identified thousands of unique binding sequences that bind to the cases and not controls, (3) printed over 1,700 recombinant antibodies on high density arrays and (4) probed those arrays with individual sera from 50 cases (including early and late stage, and high and average risk women) and 50 asymptomatic controls. In addition to these tasks, we have also undertaken several research tasks to further optimize our experimental protocols. These include a series of shotgun proteomics experiments used to characterize the protein constituents of the clinical materials used in our selection, an evaluation of multiple array normalization and processing protocols to tailor data analysis to our array platform, and improved methods for high throughput shuffling (yeast library only) and purification of antibodies. At present, materials from our project include libraries of binding agents and data, including microarrays profiling dozens of specimens and mass spectrometry data characterizing the constituents of ovary tumor proximal fluid. To date, the major findings of our proposal include the proof of principle that (based on our data analysis) the panning and array procedures are capable of evaluating thousands of unique antibodies, and that (based on the proteomics measurements) the selection material is rich in putative biomarkers.

15. SUBJECT TERMS
No subject terms provided.

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** USAMRMC
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	12	**19b. TELEPHONE NUMBER** *(include area code)*

Table of Contents

Affinity-Based Serum Proteomics for Ovarian Cancer Early Detection
Martin McIntosh, PhD, Principal Investigator

INTRODUCTION:

Our research project is intended to exploit unique characteristics of phage and yeast recombinant antibodies as the basis for a serum biomarker discovery platform for ovarian cancer. Essentially, we select from large recombinant libraries those binding sequences which bind to cancer related material (serum or tumor proximal fluid) but not to control serum, then we evaluate these sub libraries in high throughput using novel recombinant antibody arrays probed with serum from our valuable serum repository. At the present time we are on track based on our initial proposal. We have (1) selected a well-balanced group of cases (serum and proximal fluid) and controls for our initial discovery, (2) identified thousands of unique binding sequences that bind to the cases and not controls, (3) have printed over 1,700 recombinant antibodies on high density arrays and (4) probed those arrays with individual sera from 50 cases (including early stage and late stage, and high and average risk women) and 50 asymptomatic controls.

BODY:

We first relate our progress directly to the stated aims of our proposal, as a large part of our progress followed the specific tasks and timeline initially proposed. In addition, we have made advancements in our protocols and procedures to make our experimental platform more efficient. We describe those below. Our initial aims are in italics, and our response is in normal text:

Aim 1 Tasks: (Months 1 to 6): To identify thousands of candidate ovarian cancer biomarkers from phage and yeast recombinant libraries by selecting sub-libraries against reactivity to common abundant proteins and a heterogeneous pool of control sera and for reactivity to a biological material (sera and/or tumor constituent proteins) from a heterogeneous pool of cancer sera.

1. Select cases and controls to be pooled and used for biomarker discovery panning.
2. Perform panning to enrich for scFv that bind constituents in cancer material

We chose sera from 12 cases and 12 controls for our initial library selection. Cases and controls were tightly matched on age, menopausal status and cancer risk (a more detailed document on selection criteria and samples are available on request). In brief, all of the background variables were controlled in the selection by the use of Propensity Score statistical matching procedures. Pools from the cases and controls were used for several rounds of positive (cases) and negative (controls) selection, respectively, using scFv expressed on the surface of both phage and yeast. In addition, we have utilized proximal fluid (ascites and cystic fluid) from ovarian cancer cases (serous only, late stage) to perform a second series of positive selections to enrich the library with additional potential biomarkers. We believe that these proximal fluids, collected immediately adjacent to the tumor may be enriched for cancer biomarkers. This is confirmed by the summary of mass spectrometry profiling below. For both yeast and phage scFv we have selected nearly 5,000 (each) scFv which appear to bind to proximal fluid but do not bind to normal control sera, and a sum of 1700 were shuffled and purified and used in our array fabrication described below. We continue to purify the remainder of the library for future evaluation.

3. Perform shuffling of the yeast scFv sub-libraries and purify selected phage scFv.

Evaluating the scFv from steps 1 and 2 requires further biochemical processing before being printed on the arrays, including: (a) Shuffling (yeast only): Plasmid shuffling is required in order to express the yeast scFv in their secreted form and (b) Purification (both phage and yeast): Once selected, each scFv must be expressed in large volumes and then purified. Our initial efforts at both of these steps turned out to be somewhat less efficient that we originally planned. Therefore, we have spent significant effort to streamline our purification effort. We have developed novel approaches to shuffle the yeast libraries in a high throughput manner (protocol available on request), and have developed procedures for purifying the phage scFv in high volumes in 96-well plates. For the yeast we can now shuffle and purify over 100 scFv each week (based on 1 full time effort), and can purify 96-phage scFv each day. We have applied these strategies to the selection described in steps 1 and 2 above, and as of September 15 (the date of our most recent array fabrication) a total of 1012 phage and 776-yeast scFv specific for proximal fluid have been purified and were available for printing and high throughput evaluation in arrays (step 4).

4. Construct and print scFv microarrays and compare their utility to ELISA methods.

On September 15, 2006 we printed 200 antibody arrays containing scFv from proximal fluid, 338 full-length antibodies, and many control antibodies all in triplicate, of which 144 have been probed with case and control sera. We currently use Nexterion H substrate slides onto which the antibodies are printed using an Omnigrid robotic arrayer (Gene Machines). Coupling of the antibody is complete in 10 min, and the slides are then blocked with 1% BSA for 30 min. After washing twice, the Cy3/5 labeled plasma protein mixture (i.e., case vs. reference or control vs. reference) is applied immediately to the wet slide followed by addition of a coverslip for a 2 hour incubation. The slides are then washed 3 times in PBS for 10 min each. Finally, centrifuging at 500-x g for 5 min dries the slides. The slides are scanned using a GenePix 4000B (Axon Instruments) scanner. The images are processed into numerical data via Genepix Pro 6.0.

In addition to printing our scFv, we also have added to these arrays 338 full-length antibodies (most from commercial sources). These full length antibodies include known ovary tumor markers (e.g., HE4, SMR, CA 125), and also potential novel tumor markers. As a first pass at evaluating the potential of our platform we evaluated the ability to classify cases versus controls (described in Aim 2). We established that the known full-length antibodies that are from known markers classify as predicted, and the other antibodies are enriched for biomarkers (e.g., overall they tend to be elevated in controls). Overall, based on false discovery rate calculations, we estimate a total of 18 differential markers from among these full length antibodies.

Aim 2 Tasks (months 6 to 18): To profile serum from a heterogeneous set of 75 ovarian cancer cases and 75 matched controls using antibody microarrays containing the recombinant antibody sub-libraries.

1. Select cases and controls for profiling as described in Methods section.
2. Perform case-control analysis of the 75 cases and controls including quality assurance tests (by profiling arrays using the serum from the pools used to produce the sub-libraries).

Table 1 Summary of design for first round of profiling

Day1	**Day2: cancer vs healthy** **(all average risk and matching on age)**	**Day3: cancer vs surgical** **(all average risk and matching on age)**	**Day4: cancer vs benign** **(all average risk, matching on age)**	**Day5: average vs high risk** **(match on histology, stage, age)**	**Day6: early vs healthy** **(matching on risk and age)**
12 old case	10 late stage, cancer serous	10 cancer serous (1 early and 9 late)	10 late stage, cancer serous	10 average risk (5 healthy, 1 early and 4 late stage cancer serous)	10 early stage, 3 high and 7 average risk with mixed histology: 3 mucinous, 3 clear cell, 2 serous, 2 endometrio)
12 old control	10 healthy	10 surgical normal	10 benign serous	10 high risk (5 healthy, 1 early and 4 late stage cancer serous)	10 healthy: 3 high and 7 average risk
	4 QC	4 QC	4 QC	4 QC	4 QC

We are now in the process of profiling large libraries of ovarian cancer case and control sera using the proximally selected scFv described above in Table 1. A total of six days of profiling allowed us to interrogate 100 cases and controls, plus we also interrogated the initial 24 cases and controls used to define our sub-library. On each day a total of 4 arrays are used for quality control purposes. We are in the process now of evaluating these arrays to identify a classifier set of scFv, and will then profile another large set of specimens to add to our data and to also validate this classifier. This will complete the work defined in Aims 1 and 2.

<u>Additional work not explicit in aims</u>

Pilot data characterizing the quality of the scFv sub-library: Although we take steps to ensure that the scFv sub-library remains diverse (e.g., result in unique binding sequences) during the selection process, we have implemented some quality control steps to ensure that this remains the case. We have selected what we currently define as our top 20 candidate scFv and sequenced their DNA and protein. We found that all of the sequences of the variable binding regions were different in all cases at the DNA level and 11/12 were different at the protein level. This result indicates that diversity of the scFv sub-library has been maintained at least at the DNA level. We intend to continue this level of QA for all of our selections for this research proposal.

Pilot LC-MS/MS experiments to characterize the quality of the library and protein identifications. We have undertaken a series of experiments to characterize the protein constituents of the proximal fluid we used for selecting the sub libraries and also to evaluate the targets of the scFv library. The intent is to confirm that sufficient overlap between the proximal fluid and serum biomarker candidates exist in order to justify proceeding with our strategy to begin with proximal fluid. Following several pilot studies which were intended to evaluate the best sample preparation scheme, we present two sets of experiments:

1) Raw proximal fluid shotgun sequence: After using a MARS column to immunodeplete abundant proteins from proximal fluid, we submitted the remaining material to interrogation by LC-MS/MS (in triplicate).
2) Serum pull-down shotgun sequence: We utilized the proximal fluid specific sub-library of ~2000 scFv to immunoprecipitate antigens from ovarian cancer case sera, and then submitted the anitgens to shotgun sequencing by LC-MS/MS.

The protein identifications from these experiments were submitted to database searching and all protein identifications with confidence exceeding 0.95 were selected. We identified 145 unique proteins from proximal fluid and 131 from the serum pull down (see Figure 1). A list of these proteins can be provided on request. In addition, we find that a total of 59 of these proteins are in common between the two experiments.

We also evaluated these protein identifications in order to identify the quality of their identifications with respect to interesting cancer biomarkers. Based on literature research, we have established that 28 of the proteins (see Table 2) have been previously described as having a role in one or more cancers, suggesting that this material may be promising for generating candidate biomarkers.

Based on these findings we are currently expanding these proteomic interrogations to include a larger number of fractions in order to increase the overall protein coverage.

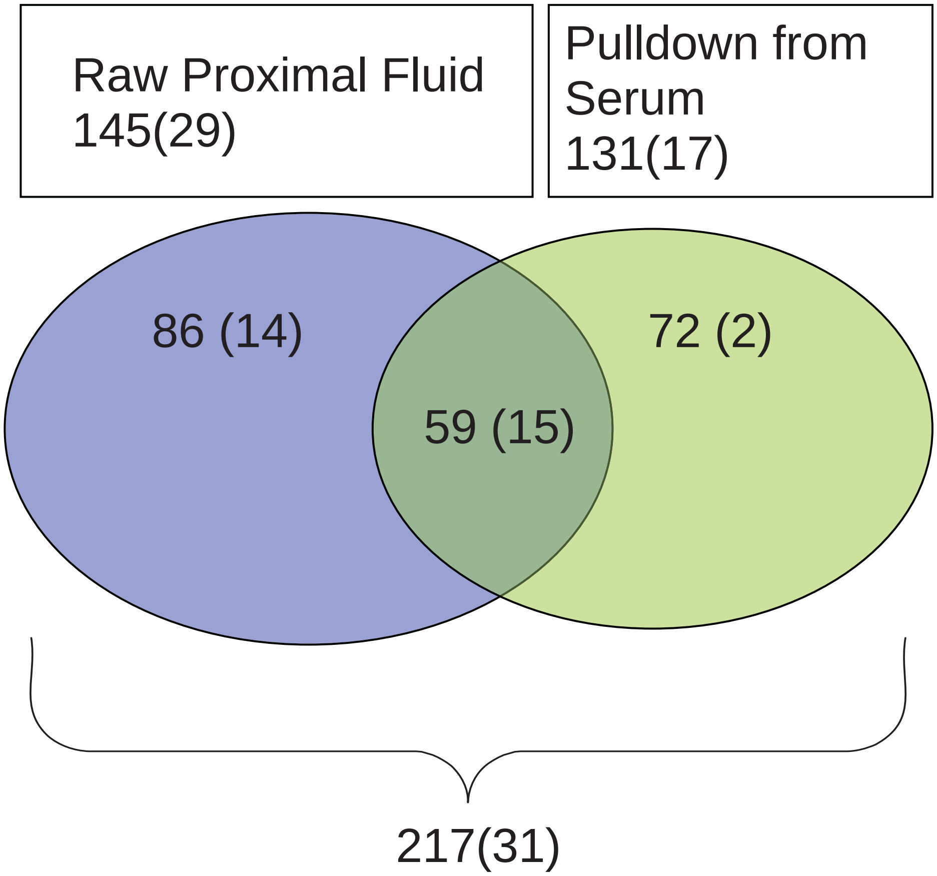

Figure 1 Summary of overlap of protein identifications between proximal fluid shotgun identifications and pulldown experiments in serum

Table 2 Summary of protein identifications from proximal fluid identified by LC-MS/MS which have been previously implicated in cancer

Protein ID	Description	Gene Name	Peptides	No. Cancers implicated
IPI00021841	Apolipoprotein A-I precursor	APOA1	16	6
IPI00022432	Transthyretin precursor	TTR	5	6
IPI00025426	Pregnancy zone protein precursor	PZP	5	9
IPI00006662	Apolipoprotein D precursor	APOD	3	6
IPI00022213	Gastricsin precursor	PGC	1	12
IPI00478493	Haptoglobin precursor	HP	1	11
IPI00017601	Ceruloplasmin precursor	CP	33	7
IPI00022434	ALB protein		25	6
IPI00553177	Alpha-1-antitrypsin precursor	SERPINA1	14	7
IPI00298497	Fibrinogen beta chain precursor	FGB	10	7
IPI00022429	Alpha-1-acid glycoprotein 1 precursor	ORM1	8	5
IPI00021842	Apolipoprotein E precursor	APOE	7	6
IPI00022463	Serotransferrin precursor	TF	7	7
IPI00022463	Serotransferrin precursor	TF	7	7
IPI00166729	Alpha-2-glycoprotein 1, zinc	AZGP1	6	7
IPI00291262	Clusterin precursor	CLU	6	7
IPI00291262	Clusterin precursor	CLU	6	7
IPI00023673	Galectin-3 binding protein precursor	LGALS3BP	4	9
IPI00021854	Apolipoprotein A-II precursor	APOA2	4	7
IPI00019568	Prothrombin precursor	F2	3	7
IPI00027350	Peroxiredoxin 2	PRDX2	2	5
IPI00026314	Gelsolin precursor	GSN	1	6
IPI00007047	Calgranulin A	S100A8	1	5
IPI00019580	Plasminogen precursor	PLG	1	16
IPI00022431	Alpha-2-HS-glycoprotein precursor	AHSG	1	6
IPI00021857	Apolipoprotein C-III precursor	APOC3	1	6
IPI00022368	Serum amyloid A protein precursor	SAA2	1	7
IPI00022420	Plasma retinol-binding protein precursor	RBP4	1	5

Methods and procedures to normalize and evaluate antibody micro-arrays

The antibody arrays have many characteristics in common with two-channel oligonucleotide (genomic) arrays, but spot morphology and background characteristics are different and so adapting standard array processing methods may not be optimal. We have undertaken a systematic study of multiple processing and normalization techniques to assure our processing is appropriate. Specifically, our main components of processing data as we have determined are as follows:

(1) Correct background. We adopted the “normexp” method developed by Smyth (2005). It’s a convolution of normal and exponential distributions fitting to the foreground intensities using the background intensities as a covariate, and the expected signal given the observed foreground becoming the corrected intensity. This results in a smooth monotonic transformation of the background subtracted intensities such that all the corrected intensities are positive.

(2) Detect and filter out poor quality antibodies on a slide using measurements from multiple spots (Tseng GC (2001)). Figure 2 shows the CV (coefficient of variation of three replicates) versus mean intensity (average of cy3 and cy5 signals). We mark all antibodies having CV values larger than a threshold as poor quality data by a windowing procedure. For

each antibody we construct a windowing subset by selecting 50 antibodies whose mean intensities are closest to this antibody. If the CV of this antibody is within the top 10% among antibodies in its windowing subset then we regard the data on this antibody as unreliable. These antibodies will be filtered out and have no influence on normalization processing and statistical analysis.

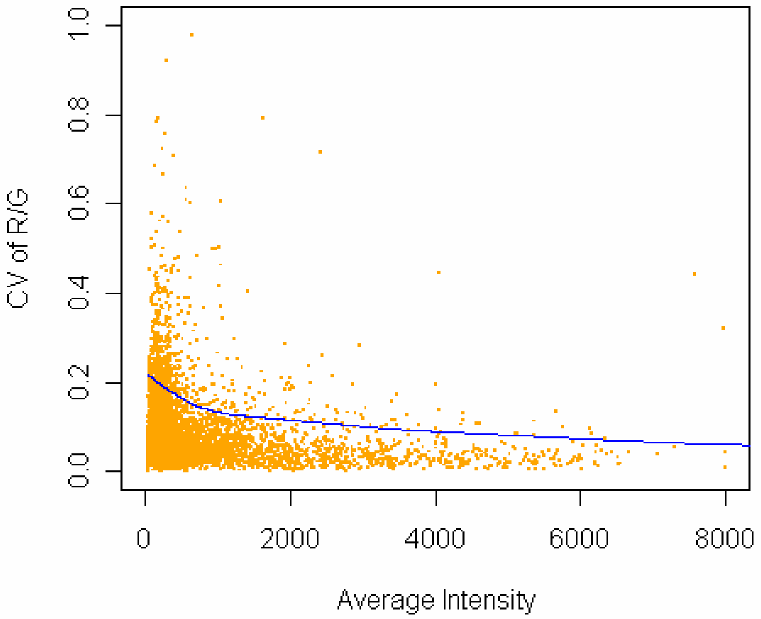

Figure 2: Coefficient of variation (CV) of replicates versus average intensity (Cy5+Cy3)/2 in the array. The curve indicates the 10th upper percentile in the moving window containing the 50 nearest antibodies. Antibodies with a CV larger than this curve will be filtered out. Only antibodies with a low CV have high agreement in replicate spots hence representing high experiment quality.

(3) Perform slide-dependent non-linear normalization using only good quality antibodies. Three different methods were tested and compared. Method1: Print-tip loess developed by Yang YH (2002). This method used lowess regression for each print tip group to produce robust location estimates of the intensity log ratios M ($\log_2R$-$\log_2G$) for various intensity levels A ($1/2(\log_2R+\log_2G)$) and to adjust each antibody with a different normalization value depending on its overall intensity. This method adjusts the systematic bias introduced by print-tip and the antibody affinity. Method2: Plate loess and adjusted by print-tip mean. By processing the control arrays (same serum sample co-hybridized to two channels), we found the major bias was not introduced by print tip but by antibody producing batches (plates). So, we modified Yang's print-tip loess method in which we did lowess location normalization within plate group and then we had a constant adjustment within each print-tip group to force the distribution of the log ratios to a median zero. The assumptions under the first two methods are: first, only a relatively small proportion of the protein varies significantly between the two co-hybridized serum samples; and second, the distribution of up/down-regulated protein levels is symmetric. Given the pre-selection of ScFv antibodies, these two assumptions may or may not fail depending how well our pre-selection worked. Method 3: Global loess for housekeeping proteins recommended by Tseng GC (2001). This approach is based on the assumption that if

a protein is up regulated its intensity rank in one channel, say Cy5, should be significantly higher than the rank in the other, and vice verse. The ranks of Cy3 and Cy5 intensities of each antibody on the slide are separately computed. For a given antibody if the ranks of Cy3 and Cy5 intensities different by less than a threshold value *d* and also the rank of the averaged intensity are not among the highest *l* ranks or lowest *l* ranks, this protein is classified as a non-differentially expressed protein. A threshold value of 20% of the data for both *d* and *l* was used for now. However, this percentage can be determined by a more sophisticated iterative selection scheme (Schadt EE 2001). After selecting non-differentially expressed proteins, global lowess regression for three types of antibodies (phage ScFv, yeast ScFv and full-length) were done separately to capture their different operating characteristics. Ideally, method 2 (plate loess method) should be applied for these housekeeping proteins. But because we keep less than 20% data, the lowess regression by plate group won't be very robust.

After comparing the three methods, we are certain that method 1 is not appropriate to our data. However, we haven't settled on the second or the third method yet. Method 2 (using for now) seemed worked better in terms of producing reproducible signatures across multiple days. This is demonstrated in Figure 3 below. Each figure plots the total intensity A value (horizontal axis) versus log red-to-green ratio M value (vertical axis). The reason for this may be because method 2 uses all good quality spots (average 80%) in normalization whereas method 3 used less than 20% housekeeping proteins.

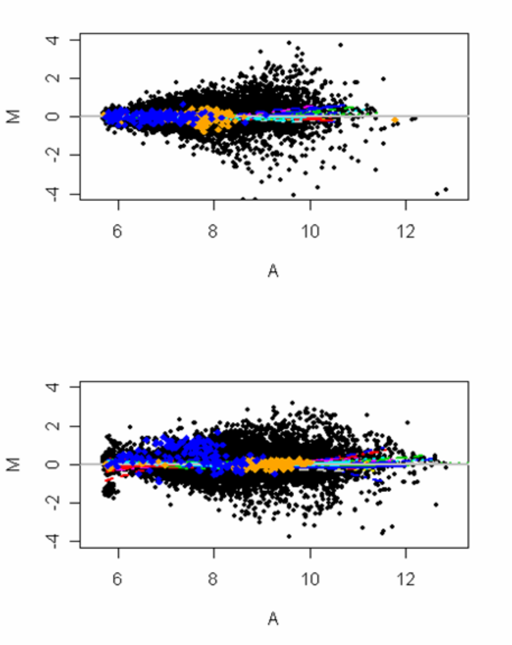

Figure 3 MA-plot of two arrays hybridized using same serum sample but done on different days. Orange spots are IL1B positive controls and blue spots are buffer negative controls.

KEY RESEARCH ACCOMPLISHMENTS:

The key goals of our first year of funding included the establishment of optimal selection materials, generation of a biomarker array pipeline, and establishment of a proof of principle of our platform. All these goals have been reached.

REPORTABLE OUTCOMES:

Here we summarize reportable outcomes, summarized from the description above, in three areas, including primary findings, Data, Affinity agents and Presentations.

Data and analysis methods: We have the following data resources generated from the current project.

- Tandem MS data: We have generated several data sets which summarize the protein identifications from the scFv and as a library profiling the ovarian cancer proximal fluid. A total of 25 LC-MS/MS experiments were performed and the data are available for mining.

- Array data: We have profiles of over 200 ovarian cancer cases and controls of various histology's and risk groups. These data are presently available for mining.

Affinity agents and arrays: The following affinity agents have been obtained and are stored in our freezers.

- We have selected over 5,000 scFv by panning which may be putative biomarkers for ovarian cancer.
- We have shuffled and purified over 1,000 of these scFv and are available for high throughput evaluation.
- Putative binding sequences (biomarkers): We have identified a subset of 20 scFv which are putative binding sequences which may be useful biomarkers for cancer. We are in the process of confirming these markers and, once confirmed, their antigens (their protein sequence) will be identified using LC-MS/MS.

Presentations and posters:

Arturo Ramirez presented "Ovarian Cancer Biomarker Discovery using single chain antibody arrays" at the 6th Biennial Marsha Rivkin Ovarian Cancer Symposium in September 2006 in Seattle, WA.

Christian Loch presented "Recombinant antibody (rscFv) arrays for colon cancer screening" at The American Cancer Society Postdoctoral Fellows Symposium, Santa Ana Pueblo, NM, November, 2006 and at the Canary Foundation Annual Stakeholders Symposium, San Jose, CA, May, 2006.

CONCLUSION:

The most significant finding of our work thus are (1) that our overall platform proof of principle has been accomplished and we are confident that we can rapidly profile ovarian cancer related proteomic profiles with our high throughput technology, and (2) that selection of the scFv sublibraries using proximal fluid may be superior to primary selection using serum or plasma.

REFERENCES:

Smyth, GK (2005). Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions using R and Bioconductor, Gentleman R, Carey V, Huber W, Irizarry RA, Dudoit S (eds.), Springer, New York, pages 397-420.

Tseng GC, Oh MK, Rohlin L, Liao JC and Wong WH (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. Nucleic Acids Research Vol 29(12): 2549-2557

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J and Speed TP (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Research 30(4):e15

Schadt EE, Li C, Ellis B and Wong WH (2001) Feature extraction and normalization algorithm for high-density oligonucleotide gene expression array data. Preprints 303, Department of Statistics, UCLA, Los Angeles, CA.

APPENDICES:

None.

SUPPORTING DATA:

Supporting data are provided in tables above. Raw and processed MS/MS data are available on request.