AD_____

Award Number: DAMD17-98-1-8649

TITLE: Use of Novel Technologies to Identify and Investigate Molecular Markers for Ovarian Cancer Screening and Prevention

PRINCIPAL INVESTIGATOR: Nicole Urban, Sc.D.

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

REPORT DATE: October 2000

TYPE OF REPORT: Annual

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

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 2000	Annual (1 Oct 99 - 30 Sep 00)

4. TITLE AND SUBTITLE
Use of Novel Technologies to Identify and Investigate Molecular Markers for Ovarian Cancer Screening and Prevention

5. FUNDING NUMBERS
DAMD17-98-1-8649

6. AUTHOR(S)
Nicole Urban, Sc.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Fred Hutchinson Cancer Research Center

Seattle, Washington 98109-1024

E-MAIL:
nurban@fhcrc.org

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. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

Report contains color graphics.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; Distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

The purpose of this study is to identify novel genes that encode proteins that can be used to detect ovarian cancer before it spreads outside the ovary and becomes incurable. The goal is to assemble a panel of known and novel ovarian tumor markers that may form the basis of a cost-effective, serologic screening test for early stage ovarian tumors. The research encompasses the use of two novel technologies to identify such genes. In Project 1 we use HDAH to identify genes that are over-expressed in ovarian cancer tissue. In Project 2 novel ovarian tumor antigens are being identified by SEREX. We have identified a large number of genes that are over-expressed in ovarian cancer tissue relative to the ovarian tissue obtained from women without cancer or ovarian pathology. We have also identified several oncogenic proteins that elicit antibodies detectable in the blood of some ovarian cancer patients. These discoveries are providing the foundation for ongoing work in early detection of ovarian cancer, funded by the NCI as part of a SPORE in ovarian cancer.

14. SUBJECT TERMS
Ovarian Cancer Screening, Gene Expression, Serum Antibody, High Density Array Hybridization (HDAH), SEREX

15. NUMBER OF PAGES
160

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Annual Report for Grant DAMD17-98-1-8649

October 1, 1999 – September 30, 2000

Use of Novel Technologies to Identify and Investigate Molecular Markers for Ovarian Cancer Screening and Prevention

Nicole Urban, ScD
Principal Investigator

TABLE OF CONTENTS

Annual Report for Grant DAMD17-98-1-8649

October 1, 1999 – September 30, 2000
Year 02

Use of Novel Technologies to Identify and Investigate Molecular Markers for Ovarian Cancer Screening and Prevention

Nicole Urban, ScD
Principal Investigator

INTRODUCTION

Novel genes associated with ovarian cancer. We were awarded funding for a study entitled "Use of Novel Technologies to Identify and Investigate Molecular Markers for Ovarian Cancer Screening and Prevention" (DAMD 17-98-1-8649), to conduct a systematic search for novel genes and gene products associated with ovarian cancer. This interdisciplinary effort, which will require 3 years of work including a no-cost extension that has already been approved (10/98-9/01), addresses gene discovery as it relates to risk-assessment and early detection. Our purpose is to identify novel genes that encode proteins that can potentially be used to detect ovarian cancer before it spreads outside the ovary and becomes incurable. The goal is to assemble a panel of known and novel ovarian tumor markers that may form the basis of a cost-effective, serologic screening test for early stage ovarian tumors. The scope of the research encompasses the use of two novel technologies to identify such genes. It includes two research projects, described below.

HDAH. In Project 1, entitled "Characterization of Genes Overexpressed in Malignant Ovarian Neoplasia by High Density Array Hybridization" (DAMD 17-98-1-8649), we use high density array hybridization (HDAH) to identify genes that are over-expressed in ovarian cancer tissue. Drs. Leroy Hood, Nancy Kiviat and Michel Schummer are using high-density cDNA array hybridization (HDAH) to compare the expression of genes in normal and in neoplastic ovarian tissue. Genes that are highly expressed in malignant tissue, but expressed at low levels in benign and normal tissue, are potential candidates for development as diagnostic markers. We are building our own libraries from unique ovarian tissues for the hybridization work to ensure that we will discover *novel* genes. We have found many over-expressed genes using HDAH, from which the most promising have been selected for further work-up.

For example, a top candidate for development is HE4, an epididymal gene that maps to a region of the genome that is a hot spot for changes in ovarian and other cancers. This region is found to be amplified in ovarian and breast cancer, as well in some glioblastomas. It is possible that this amplification of 20q12-13.1 in ovarian cancers causes HE4 to be over-expressed. Dr. Schummer is developing an assay for HE4 in collaboration with Drs. Ingegerd and Karl-Erik Hellstrom. Another is mesothelin, a 40-kDa glycoprotein present on the surface of many different malignancies including the majority of mesotheliomas and ovarian cancers. Drs. Ingegerd and Karl-Erik Hellstrom have recently developed an assay to detect mesothelin in serum. These genes will be evaluated for their contribution to a panel of markers to detect ovarian cancer.

SEREX. In addition to HDAH, we have used SEREX, a novel serological method that identifies immunogenic gene products for which antibodies are present in the sera of women with ovarian cancer but not in those of controls. In the project entitled "Antibody Immunity to Cancer Related Proteins as a Serologic Marker for Ovarian Cancer" (DAMD 17-98-1-8649) Drs. Brad Nelson and Mary L. Disis are using SEREX to identify antibodies to novel cancer-associated proteins. This new technology involves (1) construction of a bacterial cDNA expression library from a pooled tissue sample representing the tumors of selected ovarian cancer patients, (2) probing of the library by immunoblot with serum from both cancer

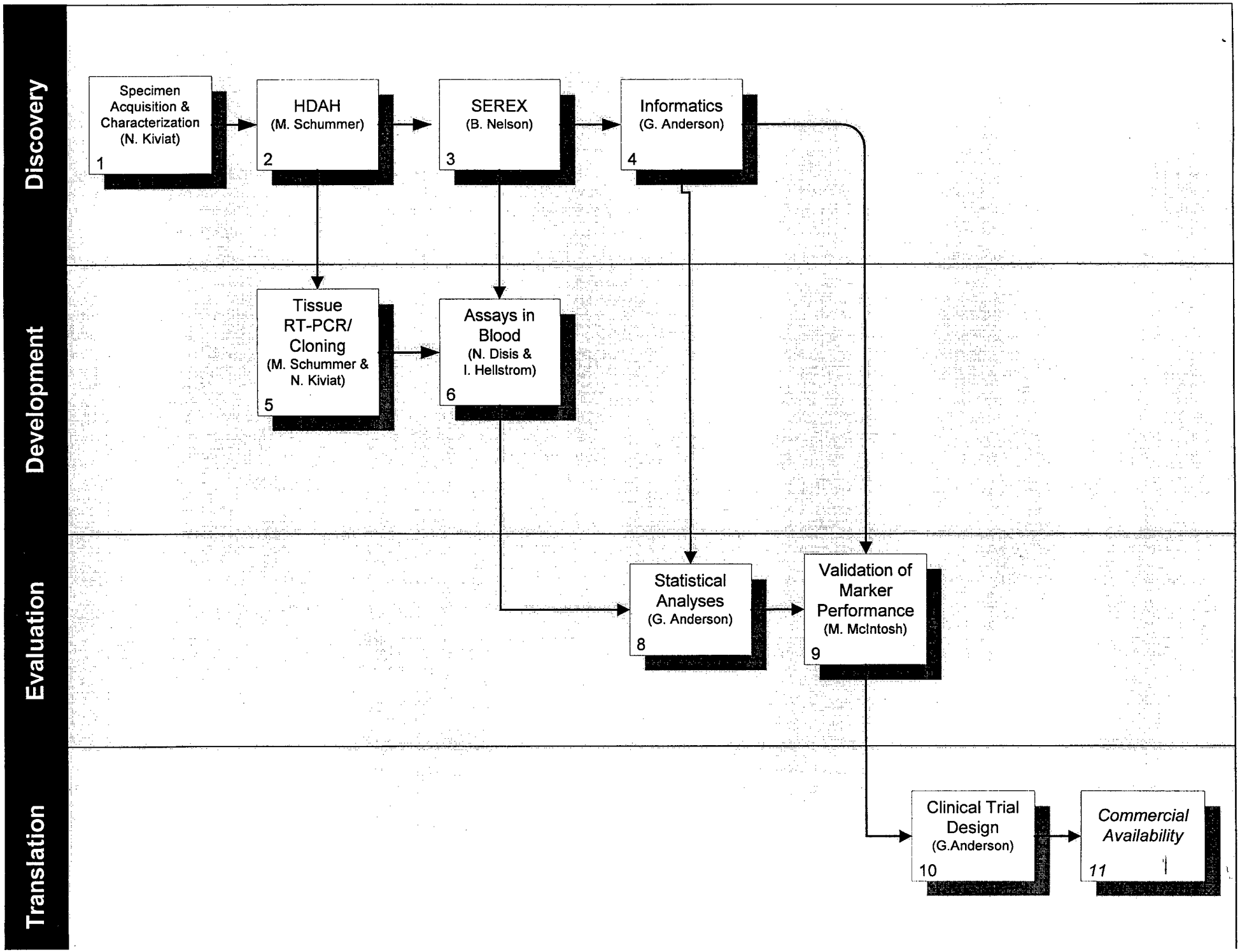
patients and control individuals, and (3) identification of bacterial colonies that are recognized by serum antibodies from cancer patients but not normals. Antibodies found only in the sera of cancer patients are candidates for development, validation, and evaluation for inclusion in a set of markers to be used as a first-line screen in the early detection of ovarian cancer. In addition, this technology defines genes by the immunogenic proteins that they encode. These genes would be possible candidates for DNA vaccines. We have identified several oncogenic proteins using SEREX and are evaluating the most promising for use as markers of ovarian cancer risk.

Core. Novel genes identified by HDAH have been evaluated for development as tumor markers. Similarly, antigens identified by SEREX are being further evaluated using purified proteins and larger numbers of normal, benign and cancer serum in an ELISA format. Statistical analyses are being employed to determine the sensitivity and specificity of HDAH- and SEREX-defined tumor markers for the detection of early-stage ovarian cancer. Towards this goal, serum antibody responses to the known tumor antigens p53 and HER2 are being evaluated by ELISA in patients with ovarian cancer versus normal controls.

Specialized Program of Research Excellence (SPORE) in Ovarian Cancer. Our DOD-funded work has led directly to funding by the NCI of a SPORE in ovarian cancer. Our goal is to improve ovarian cancer outcomes, including morbidity and quality of life (QOL) as well as survival, incidence and mortality. Each of the 5 projects in the SPORE is designed to improve ovarian cancer outcomes. Projects 1 and 2 are designed to provide information that will enable us eventually to reduce cancer mortality and incidence through treatment and prevention interventions respectively; Project 1 uses the HDAH technology to identify genes associated with resistance to chemotherapy. Project 3 provides the methods that will be needed to conduct screening and prevention trials when appropriate screening and prevention interventions have been identified. Project 4 provides the intervention for use in a screening trial. Projects 3 and 4 make use of the markers that we have found, in combination with previously known markers, in panels designed to measure risk of ovarian cancer.

An example of marker discovery, validaton and application is depicted below.

Pacific Ovarian Cancer Research Consortium



Separate tasks were identified for each project and core as being imperative to the successful completion of this project.

In the original statement of work, 5 major tasks were identified for Project 1, 5 major tasks were identified for Project 2, and major tasks were identified for the Statistical, Clinical and Laboratory Coordinating Core. These tasks are listed in a table included in Appendix A. Included as Appendix B is a timeline detailing project progress during Year 02 and plans for Year 03.

Project 1

Identification of Potential Markers for Population Based Screening for Ovarian Cancer: Characterization of Differential Gene Expression in Malignant Neoplasia by Use of High Density Array Hybridization (HDAH).

Nicole Urban, ScD, Michel Schummer, Ph.D., Nancy Kiviat, MD

INTRODUCTION

It is well established that the set of genes expressed in tumor cells differ from that expressed their normal counterparts in both a qualitative (different genes expressed) and quantitative fashion. These differences in gene expression, and specifically overexpression are *exceedingly common in cancers* at the level of mRNA and provide a logical basis for cancer screening assays. We are proposing a rapid and accurate approach to identification of genes which are overexpressed in ovarian cancers and which are likely to be of interest for use in ovarian cancer screening assays. We will use multiple rounds of cDNA array hybridization to identify a subset of a few dozen genes which are overexpressed in a high percentage of early and late stage ovarian cancers but not in normal tissues. Once such genes were identified by array hybridization, they were sequenced and by comparing sequences to described sequences on public databases, we were able to target those which appear to code for secreted and/or for transmembrane proteins for further characterization a) by quantitative RealTime PCR on tissues and circulating cells from peritoneal washes, and b) by ELISA on patient sera after the generation of monoclonal antibodies to the newly found proteins.

BODY:

Work proposed:

Task 1. Generation of representative cDNA arrays:

- Three cDNA libraries will be generated from normal, metastatic and late stage neoplastic ovarian tissues.
- These libraries will then be used to construct first generation solid phase membrane arrays containing 100,000 clones.

Task 2. Primary Characterization of Normal and Neoplastic Ovarian Tissue:

- Hybridization of the first generation membranes with cDNA probes derived from 12 normal (pre and post menopausal ovarian tissue, 3 peripheral blood samples, peripheral blood cell culture and 1 liver tissue, 4 2 benign cystadenomas, 1 early stage and 12 late stage ovarian serous adenocarcinomas
- Evaluation of hybridization results and selection of 2,000-3,000 genes overexpressed in malignant tissues.
- These clones will be used to construct second generation cDNA arrays.

Task 3. Further Characterization of Gene Expression in Normal and Neoplastic Ovarian Tissue:

- Hybridization of the second generation arrays with cDNA from tissues used in Task 2, plus 29 additional normal tissues (20 ovarian and 9 skeletal muscle controls), 15 cystadeomas, 20 additional early and 20 late stage ovarian serous adenocarcinomas.
- Evaluation of hybridization results and selection of ~400 genes that show a high degree of overexpression in at least 75% of tumors examined.

Task 4. Characterization of highly expressed genes associated with cancer:

- Sequence determination of the ~400 overexpressed ovarian cancer-associated genes identified in Task 3.
- Confirmation of tissue specific expression using RT-PCR and Northern blot techniques.
- Selection of clones with overexpression in ovarian cancers negative for overexpression of p53, Her2/neu and c-myc transcripts for further analysis by serum-based detection technologies in Project 2.

Task 5. Final analyses and report writing:

- Final analyses of serum-based patient screening assays will be performed.
- A final report and initial manuscripts will be prepared

All tables and figures included in the progress report of this Project are also included in Appendix C.

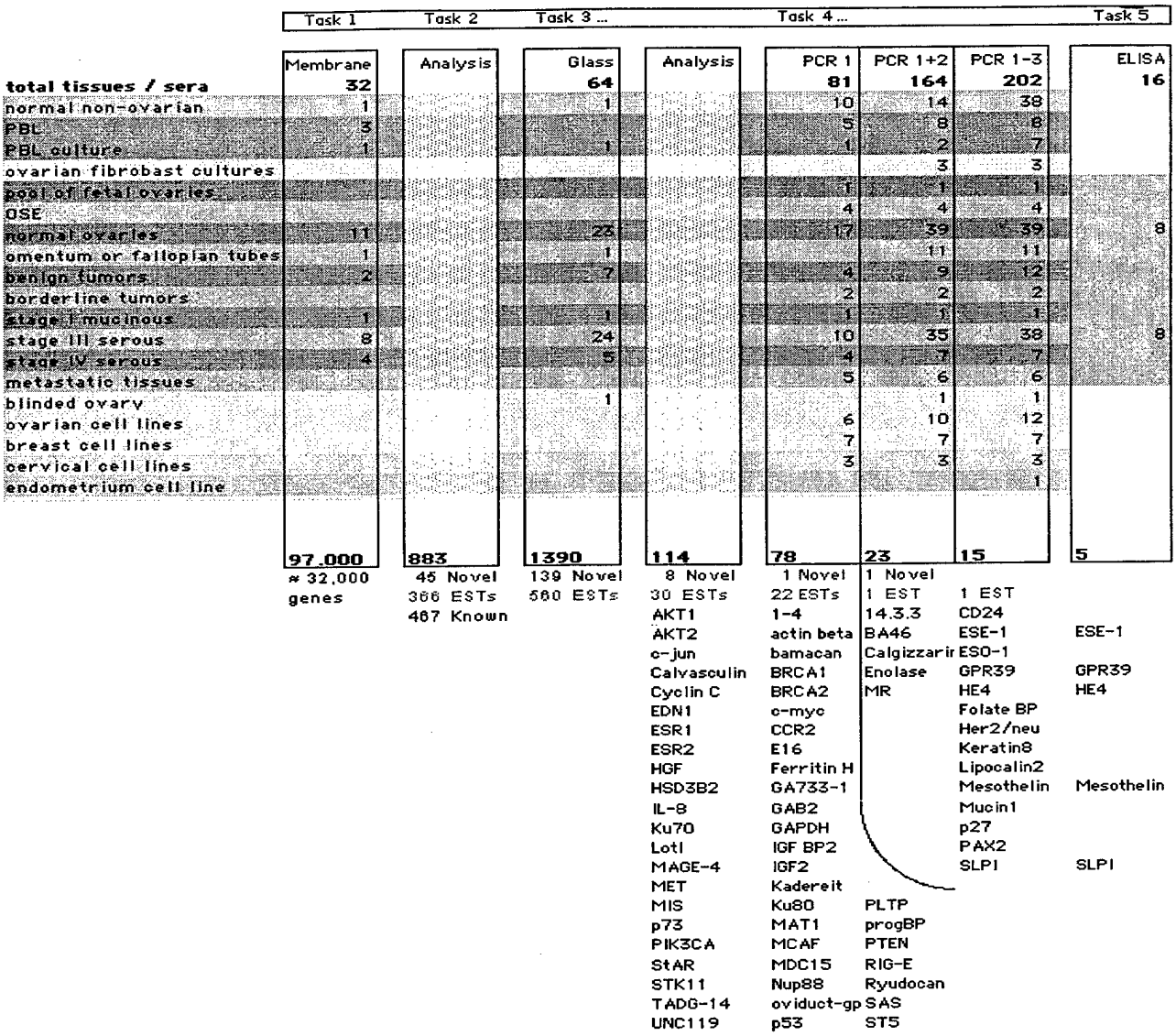


Table 1 - Work Flow

Starting from an array of 100,000 clones hybridized with probes from 32 tissues on the left, we selected 883 genes to be potential marker genes and combined them with 507 genes compiled from previous membrane arrays and other sources, to form a glass array (next panel to the right). This glass array was interrogated with probes from 64 tissues upon which 114 genes were selected as potential marker genes. Of these, 78 genes were selected for expression validation by RealTime PCR. During the latter, the number of potential marker genes was reduced on the first panel of 81 tissues, and 23 genes were passed on to the second panel which added 83 tissues. Likewise, 15 genes still displayed a stronger expression in the ovarian tumors compared to the normal tissues and they were tested on an panel of additional 38 tissues. Of these, 5 were selected for further characterization by analysis of protein expression in tissues and patient sera (ELISA). For this, monoclonal antibodies are being generated against fusion proteins. In two cases, antibody assays are already in place and patient sera are currently being screened for the presence of Mesothelin and SLPI. The tasks related to this project are displayed above the panels. The genes that were selected to be

passed on to the next higher level of scrutiny are displayed below the panels in a way that does not display gene names if they are listed in the panel to the right of it.

Work accomplished:

Task 1: Generation of representative cDNA arrays from early and late stage ovarian carcinomas

We created four cDNA libraries from pooled tissues (20 pooled fetal ovaries, 4 benign ovarian cystadenomas, 3 normal ovaries, and 4 late stage serous ovarian cystadenomas) plus an additional library made from 6 metastatic ovarian carcinomas. For quality control, from each library, 96 clones were randomly chosen. The clones were sequenced and analyzed by similarity analysis against the non-redundant and EST database. The criteria for a satisfactory cDNA library were an average insert size around 1 kb, a low number of mitochondrial and ribosomal sequences, a limited number of clones with no insert, and significant cDNA diversity (Nelson et al., 1998). Three out of the five libraries fulfilled these criteria. For the fetal and benign libraries, the average insert sizes were considerately below 1 kb. In addition, the number of clones without insert plus the ones with repeats or genomic fragments exceeded one third, a number far to high for consideration to array. Diversity of clones with homology to known genes was similar in all cases. The titer of the three remaining libraries was small which reflects the fact that we chose not to amplify the libraries in order to have a better representation of the lowly expressed clones. We selected 102,680 clones from the three libraries (9,216 from the normal, 17,664 from the late stage and 83,712 from the metastatic library each) and arrayed the colonies onto 32 sets of 5 nylon membranes, each holding 20,536 colonies. The colonies were lysed and the DNA was fixated onto the membranes using a modified Southern blot protocol (after the membranes were placed on filter paper trenched in denaturing and neutralizing solutions, they were dried and subsequently submerged in fat-free milk with 0.5% SDS and 2 x SSC for 2 hours, upon which they were dried and stored at 4°C until usage). As a result of vigorous testing of lysis protocols, this protocol provided the best signal-to-noise ratio for a colony-based membrane hybridization. One set of membranes was hybridized with a probe recognizing the vector portion of each clone. The resulting hybridization pattern revealed that out of the 120,680 colonies that were arrayed, 97,803 actually grew on the membranes. Figure 1 shows a close view on one such membrane where more than 95% of the colonies give a positive signal with the vector probe.

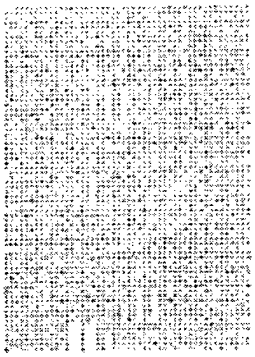


Figure 1 - Sample hybridization

Close view on 1/6 of a membrane containing 3456 colonies that was hybridized with a probe recognizing the vector portion of the cDNA. Where there is no signal, no colony grew. Overall, the number of colonies that did grow reaches 95%

Task 2: Primary characterization of normal and neoplastic tissues using these arrays

The membrane-based cDNA arrays were interrogated with 33P-labeled first-strand cDNA probes which were reverse transcribed using an oligo-$dT_{19}V$ primer from 100 µg of total RNA generated from the tissues listed in Table 2. These tissues had been accrued through the Tissue Collection Core. For most patients and controls, several tissue blocks were generated and some remained in the depository at the Marsha Rivkin Center for later use. Likewise, blood was drawn from each patient and questionnaire data was generated. For details refer to the Core section. The sera will be used at the end for the validation of potential serum markers found in the course of this project (Task 5).

Tissue Type	Description
liver	liver from Clontech
pbl male	white blood cells from 140 ml blood
pbl mix	pooled RNA from 3 controls (male and female)
pbl female	white blood cells from 160 ml blood
pbl culture	lymphocyte culture
normal cyst	paraovarian cyst, 0.95 g
normal ovary	normal ovar./tube tissue, 1.05 g
normal ovary	normal ovar./tube tissue, 0.15 g
normal ovary	normal ovar./tube tissue, right ovary, 1.06 g
normal ovary	normal ovar./tube tissue, left ovary, 270 mg
normal cyst	right ovary, Paraovarian cyst, 720 mg
normal ovary	normal ovar./tube tissue, right ovary, 250 mg
normal ovary	normal ovar./tube tissue, left ovary, 550 mg
normal ovary	normal ovar./tube tissue, right ovary, 520 mg
normal ovary	normal ovar./tube tissue, 0.47 g
normal ovary	normal ovar./tube tissue, left ovary, 0.80 g
normal ovary	fallopian tube from patient with tumor t037
benign ovarian tumor	serous cystadenoma, 0.476 g
benign ovarian tumor	serous cystadenoma, 0.9 g
mucinous stage I	mucinous carcinoma, grade A, stage Ia, 1.6 g
serous stage III	serous carcinoma, grade C, stage IIIc, 0.5g
serous stage III	serous carcinoma, grade B, stage IIIC
serous stage III	serous carcinoma, grade C, stage IIIc, 1.05 g
serous stage III	serous carcinoma, grade C, stage IIIc, 1.54 g
serous stage III	serous carcinoma, 0.92 g
serous stage III	serous carcinoma, grade B, stage IIIc, 0.37 g
undiff. stage III	undifferenciated carcinoma, grade C, stage IIIc, 0.37 g
serous stage III	Serous carcinoma, grade C, stage IIIc,0.71 g.
serous stage IV	adenocarcinoma, NOS, grade C, stage IVb, 4.3g
serous stage IV	adenocarcinoma, NOS, grade C, stage IVa 1.4 g
serous stage IV	serous carcinoma, grade C, stage IVa 1.25 g
serous stage IV	adenocarcinoma, NOS, grade C, stage IVb,0.61 g

Table 2 - Tissues used for interrogation of the 100,000 clones membrane array
We used normal, non-ovarian tissues (blue), normal ovarian tissues (green), benign ovarian tumors and invasive ovarian carcinomas (red)

The probe preparation, hybridization of the membranes and extraction of the hybridization intensities was performed as described earlier (Schummer et al., 1999). The intensity value for each cDNA hybridized with one of the 30 tissues was stored in a database. This database thus contains the entries from 102,680 cDNAs and 45 hybridization events (30 tissues of which 5 had been hybridized 3 times and 2 twice, plus the hybridization with the vector probe and a "junk" probe recognizing housekeeping genes that are commonly overexpressed in tumors but have no relevance as markers). In addition, the database contains the patient information gathered during tissue accrual by the patient questionnaire, as well as the marker status for Her2/neu and p53 from the marker tests performed in the core laboratory. All in all the database contains more than 4.8 million entries. Of the 103,680 clones that could have been present on the membranes, 97,802 grew as colonies. This is the number we will further refer to a total number of clones. Once in a digital format, the data was analyzed using the most recently developed algorithms for expression analysis.

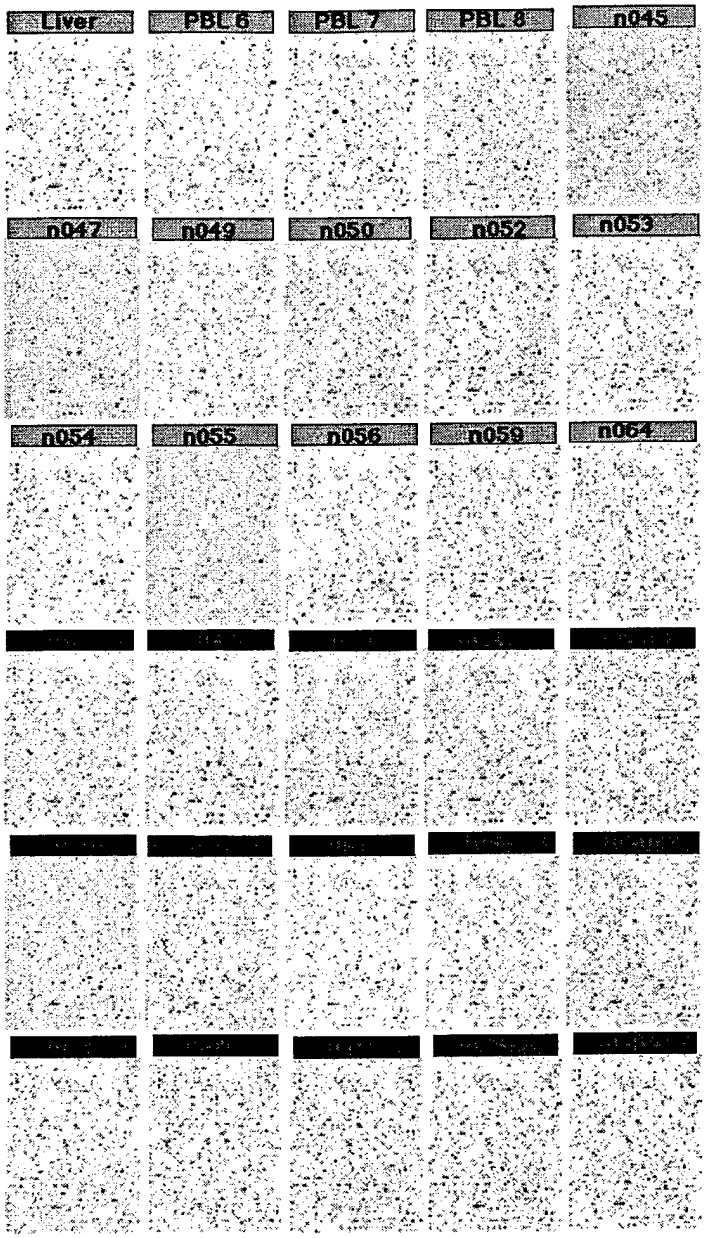


Figure 2 - Hybridization results

Displayed is one field containing 3456 colonies, replicated 30 times and hybridized with probes from 30 different tissues as indicated by the color. Although it may be possible to spot the most obvious differences and similarities in the hybridization pattern by eye, a computer-guided image processing is necessary to detect more subtle changes in expression.

The first task was to identify and exclude from further analysis the clones that code for genes previously known to be overexpressed in cancers due to their higher metabolic rate. We found earlier that these clones are often expressed at high levels (Schummer et al., 1999). During analysis, their high values would bias the dataset. We designed a probe composed of 41 housekeeping genes (beta actin, comtase, elongation factor 1 alpha, elongation factor 1 gamma, mito-atp6, mito-co1, mito-co2, mito-co3, mito-cyb, mito-nd1, mito-nd2, mito-nd4,

mito-nd5, mito-nd6, oviduct-gp, ribosomal protein L18, ribosomal protein L27, ribosomal protein L3, ribosomal protein L30, ribosomal protein L5, ribosomal protein L6, ribosomal protein L7, ribosomal protein L7a, ribosomal protein L9, ribosomal protein P0, ribosomal protein S11, ribosomal protein S12, ribosomal protein S13, ribosomal protein S14, ribosomal protein S16, ribosomal protein S17, ribosomal protein S18, ribosomal protein S21, ribosomal protein S24, ribosomal protein S25, ribosomal protein S28, ribosomal protein S3a, ribosomal protein S4, ribosomal protein S6, 18 S rRNA, 28 S rRNA) and hybridized a membrane set with it. This probe will be further referred to as the "junk" probe. The hybridization pattern clearly identified three categories of positive clones with strong, medium-strong and weakly strong signals. We sequenced 30 clones from each category. Only the clones from the two high-expression categories were entirely homologous to the 41 genes in the pool. Therefore we selected only those 10,716 (11%) clones for exclusion from further analysis.

The second step was to reduce the number of clones from the remaining 87,086 clones to 2000-3000, the number that will be arrayed on the second generation cDNA array on glass. Since the goal of our project was to discover genes with potential as markers, preferentially serum-based ones, we focused on the genes with overexpression in the tumors versus the normal tissues. In collaboration with Dr Andy Siegel, who is an adjunct professor of Statistics at the University of Washington we employed statistical measurements to reduce the number of clones in this immense dataset from 87,068 to 2,651. The selected clones exhibit a tendency to a higher expression in the tumor tissues. Table 3 lists the statistical algorithms that were employed for the reduction of the dataset.

		#accumulate to
zScore > 10.09 in >1 of all tumors	1192	1192
t Statistic> 4.00	300	1476
avg(Tumor) > 2.5* avg(NormalOvary)	277	1661
avg(Tumor) > 2.5* avg(NormalOvary,PBL,liver)	624	2181
avg(zScore) > 1.4	1439	2949
minus "junk"	298	**2651**

Table 3 - Clone selection by statistics
Statistical analysis that led to the 2651 selected clones. Each statistical method selected a certain number of clones that added up to 2949. The "junk" probe was a probe consisting of 41 housekeeping genes (ribosomal proteins, mitochondrial genes, elongation factors) that were previously found to have elevated expression in carcinomas presumably due to the elevated metabolism. It reduced the number of clones by 298 to 1651.

We sequenced all 2,651 clones on their 5' ends and the results submitted to homology search in the *nr* and *estdb* databases. The result of this homology search is summarized in Table 4. We did not intend to spend too many of our resources on the sequencing. Therefore we opted for a single amplification, single pass sequencing approach. A clone that fails to PCR amplify or that fails to produce a satisfactory sequence would therefore be labeled as "currently

unsequenceable", to be attempted to sequence at a later stage. Of our 2,651 clones, 2,061 generated sequences that could be submitted to database homology search. This excludes the clones labeled as "uninformative" in Table 4. Of the remaining clones, 519 were grouped in a class termed "uninteresting", meaning that these genes are known to be expressed at higher levels in cancers because they are either linked to the metabolism (mitochondrial and ribosomal proteins) or expressed in tumor infiltrating lymphocytes (MHC, immunoglobulins). The remaining 1542 informative clones were grouped into those who matched with more than 80% homology to sequences in the *nr* database ("known"), those who only matched only to sequences in the *esdtb* database ("EST") and those who match to neither of the two ("Novel"). In the cases of a hit to only the EST database, we would note how many ESTs our clone was homologous to and the tissues those ESTs were derived from (data not shown). This would indicate whether our clone represented a frequently expressed gene (many hits in the EST database) or, which is more desired, a rare gene, and whether it is found in many tissues or rather in just the ovary.

	#	%	Comment
Total selected clones	2651		
Bad PCR	302		
Sequenced	**2349**		
Bad sequence	88	4%	
Short sequence	18	0.8%	
Vector	38	2%	
PolyA	107	5%	
Repeat	37	2%	SINE and LINE, genomic, simple repeats
All uninformative	**288**	**12%**	
Mitochondrial	203	9%	
Ribosomal protein	19	1%	
Immunoglobulin	310	13%	
MHC	104	4%	
All uninteresting	**636**	**27%**	
Novel	**45**	**2%**	all unique
EST	298	13%	all unique except for 7 contigs containing 17 clones
Full length EST	68	3%	
All ESTs	**366**	**16%**	
GAPDH	142	6%	
Ferritin H	84	4%	
IGF-2	63	3%	
collagen 1A1	32	1.4%	
SLPI	30	1.3%	
S100A6	18	0.8%	
HE4	17	0.7%	

S100A11	10	0.4%	
Others	618	26%	
All known genes	**1014**	**43%**	excluding Ig, MHC, Mito, repeats, vector
All unique genes	**883**	**38%**	comprising 467 Known genes, 366 ESTs and 45 Novels

Table 4 - Identity of potential marker genes

Explanation of the terms used: "bad PCR" means that the PCR amplification of the clone resulted no band, multiple bands or a smear, we did not attempt to repeat the reaction; "bad sequence" refers to a sequence with an unreadable chromatogram, the sequencing reaction was not repeated; "EST" refers to clones that have no hit in the nr *database at GenBank but one or several hits in the* estdb*; "full length EST" refers to the clones resulting from the full-length sequencing projects (KIAA, DKFZ, FLJ etc.); "known genes" refers to clones that have a hit of 80% or more in the* nr *database at GenBank.*

Our libraries were all oligo-dT primed and hence the clones should all represent the 3' ends of transcripts. All of the novel genes represent unique sequences which means that we have identified 45 novel sequences with potential elevated expression in the ovarian carcinomas. Of the ESTs, 17 clones could be matched to 7 contigs. Therefore the 366 ESTs correspond to a maximum of 356 genes.

Of the clones matching to known genes, the eight genes with the most clones representing them are listed in Table 4. They are discussed briefly.

GAPDH, or glyceraldehyde-3-phosphate dehydrogenase, was the most abundant one with 6% of all sequenced clones. The increased expression of GAPDH in cancers was reported earlier (Kim et al., 1998, Desprez et al., 1992, Chang et al., 1998, Persons et al., 1989, Schek et al., 1988, Tokunaga et al., 1987, Finnegan et al., 1993). In the past, GAPDH has been used for normalization of Northern blots which speaks for its ubiquitous expression.

Ferritin H transcript was found to be elevated in ovarian tumors (Tripathi and Chatterjee, 1996). Ferritin serum levels have been reported to be elevated in ovarian cancer patients (Lahousen et al., 1989, Pinto et al., 1997, Yuan et al., 1988). However, due to its low specificity, Ferritin is not suited as a diagnostic marker (Pinto et al., 1997).

IGF-2, or insulin-like growth factor 2, was reported to play a role in ovarian cancer. In some carcinomas, IGF-2 loses its chromosomal imprinting resulting in an overexpression (Chen et al., 2000, Yun et al., 1996). Antisense oligonucleotides against IGF-2 inhibited cell proliferation and induced apoptosis in human ovarian cancer AO cells (Yin et al., 1998).

Collagens of classes 1 and 3 were reported to be actively produced both locally in the ovary as well as more remotely in the peritoneal cavity (Kauppila et al., 1996). Collagens and procollagen serum levels may be indicative of ovarian cancer disease outcome (Santala et al., 1999).

HE4 is a secreted protease inhibitor previously found by us to display elevated transcript levels in ovarian carcinomas (Schummer et al., 1999). It is regarded as our gold standard, meaning that the selection of potential marker genes should contain this gene, otherwise our selection criteria may need revision.

SLPI is also a secreted protease inhibitor, albeit with a different sequence. It is expressed by several glandular epithelial cells and it is thought to have anti-bacterial anti-HIV properties (Wingens et al., 1998). There is a SLPI ELISA test commercially available which will be described further below. SLPI has not been implicated in any cancer.
S100A6, or prolactin receptor-associated protein PRA, or calcyclin, binds GAPDH (Filipek et al., 1995). Its protein is overexpressed in a variety of tumors including colorectal adenocarcinomas (Komatsu et al., 2000).
S100A11, or calgizzarin, or S100 calcium-binding protein A11, is expressed in colorectal carcinomas (Tanaka et al., 1995) and may be involved in the regulation of cell transformation and/or differentiation (Moog-Lutz et al., 1995). Genes of the S100 family are implicated in a variety of cancers, among them melanoma (Van Ginkel et al., 1998) and breast (Pedrocchi et al., 1994).

The aforementioned 1542 informative clones correspond to 883 unique genes. They have a high potential to be marker genes, but as pointed out earlier, there is a significant error associated with them and a large portion of them may have been selected as false positives. Before we can validate the expression of these genes by a method different from array hybridization (namely by quantitative RealTime PCR), we will narrow down their number. This will be achieved in Task 3.

Task 2 (addendum): Cluster analysis of 2651 clones

The accrual of ovarian cancer tissues, especially the early stage serous carcinomas) was less effective than originally anticipated. With the end of Task 2, the few early stage cancer tissues that the tissue collection core was able to accumulate were of non-serous histology. Therefore we postponed the beginning of Task 3 (further characterization of gene expression by glass array) and instead attempted to analyze the data generated by the membrane array. This data was never meant to be analyzed in depth due to the high variability associated with it. We reasoned that in spite of the shortcomings, the genes with the most striking tumor-typical expression would stick out. The facts that we based our assumptions on are as follows.
1. The array contains genes from ovarian tumor libraries, hence a gene with high expression in the tumor should be present with multiple copies. Chances are that only a few copies of a gene show a suboptimal hybridization result and that the other copies can be used for proper analysis.
2. For some genes that were represented by multiple clones (such as for HE4, SLPI, GAPDH etc.), we calculated the standard of means as an assessment of expression variation between these clones. When calculated for each tissue separately, the standard of means averaged at 75%, ranging from 20% to 190%. When we averaged for each clone its expression values across the tumors and, separately, across the normal tissues, followed by a calculation of the two standard of means, they averaged at 23%, ranging from 10% to 37%. This is a significant reduction in variability which leads us to the next conclusion.
3. Knowing that each individual clone is associated with high error, rather than looking at the expression of one gene across all tissues we would use an analysis tool which takes into account the expression of all genes in all tissues simultaneously. This software, Bioclust™

(Ben-Dor et al., 2000b), was designed for large expression data sets like the present one. In its initial stage it was capable of analyzing datasets of maximal 5000 clones within reasonable time limits. The updated version that is available today has no such limitations.

The dataset generated in Task 2 consisted of 2651 clones assayed on 30 tissues (or hybridizations). In order to provide an assessment of the variability between hybridizations, we selected 7 tissues and repeated their hybridizations twice more (on 5 tissues, resulting in triplicated hybridization) or once more (on 2 tissues, resulting in duplicated hybridization). This resulted in an extended dataset with the same number of clones but 16 more hybridizations.
In order to determine the degree of consistency between hybridizations of the same probe, the standard of means of each clone across the two or three repeated hybridizations was calculated and averaged. The average standard of means for the replicate hybridizations is 61%, ranging from 6% to 161%. For comparison, the same value for 8 hybridizations with 8 probes from 8 different tissues averaged at 71%, ranging from 10% to 210%. Viewed in the context of the vast majority of the clones (~95%) expressing uniformly across all tissues (Schummer et al., 1999), this shift in the standard of means is significant. In addition, as will be shown below, when treated as hybridization probes coming from separate tissues, the replicates display a higher tendency to cluster together than the unrelated tissues.

We performed two separate clusterings, the ones of the tissues and the ones of the clones (Figure 3). In both cases, the algorithm was performed several times with varying starting parameters and varying fractions of the dataset. Please refer to our recent publication for details (Ben-Dor et al., 2000a).

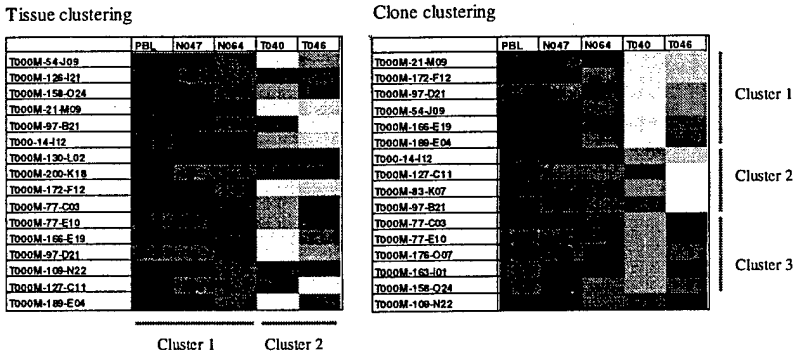


Figure 3 - Schematic explanation of the clustering

For better visual impression, the dataset is represented as a table and the values have been replaced by greyscale where white stands for high expression. Shown are 16 clones out of the 2651 (in the rows) and 5 hybridizations out of the 46 (in the columns): PBL (peripheral blood lymphocytes), two normal ovaries (N...) and two ovarian tumors (T...). In the left panel the tissues were clustered into two groups, one consisting of the normal ovaries and the PBL, the other consisting of the tumors. In order to select potential marker genes, the same clustering algorithm was repeated with a decreasing number of clones that would sort the tissues as nicely as displayed. The minimal number of clones that achieve this grouping are regarded as potential markers. In the right panel the clones were clustered into three groups. It is

conceivable that members of a group are either clones representing the same gene or gene family or genes that share similar function or similar pathways. A clone that consistently clusters with a known tumor gene would be regarded as a potential marker gene. The small example shown here was applied to the full dataset as shown in Figure 4.

The clustering of the clones grouped certain clones together that turned out to be either mostly copies of the same gene or genes with similar behavior (Figure 4). One such group contains the "gold standard" HE4 that was found earlier to be a potential marker gene for ovarian cancer (Schummer et al., 1999), together with other genes, among them SLPI, a secreted protease inhibitor just like HE4. Other clones from this group show no match to any known gene and may be potential novel marker genes.

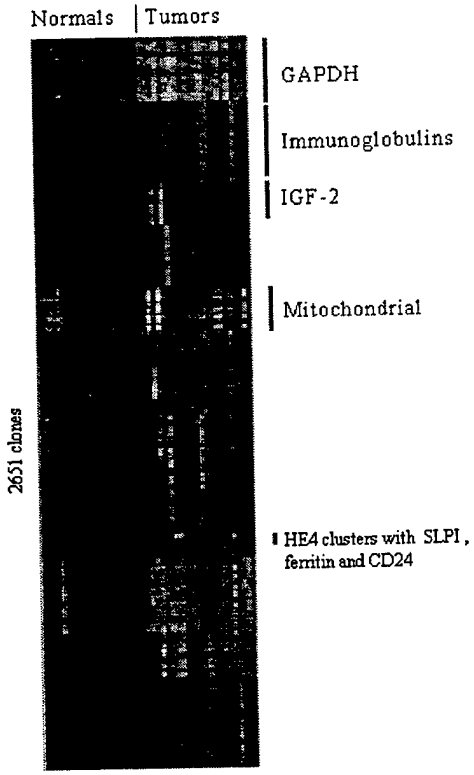


Figure 4 - Clone clustering on full dataset
Clone clustering performed on the full dataset of 2651 clones. The expression values are displayed as greyscale with white standing for high expression and black for a low one. The normal tissues (liver, PBL, normal ovaries) are shown on the left, the ovarian tumors on the right. Overall the expression of the normal tissues is lower than that of the tumors which reflects the selection criteria of these 2651 clones (low expression in normal tissues, high in tumors). In the present example the clones were clustered into 75 groups of varying size. The biggest groups consist to more than 80% of clones matching to GAPDH, immunoglobulins, IGF-2 and mitochondrial genes. Some of the smaller groups contain known tumor genes (such as CD24, ferritin and HE4) together with genes that were previously not known to be

associated with tumors (such as SLPI and clones that do not match known sequences in the public databases). These clones were regarded as potential marker genes.

The clustering of the tissues was initially performed on the full dataset. The specific clustering experiment that was performed was of the leave-one-out nature. Briefly, all tissues were labeled as either normal, malignant tumor or neither (primarily benign tumors). Bioclust was used for clustering to determine which clones were particularly useful to achieve the best separation between the "tumor" and "normal" groups. The clustering was performed using all tissues but one; the left-out tissue was then introduced into the analysis and recorded as having been classified correctly (e.g. a tumor classifying as a tumor) or incorrectly (e.g. a normal classifying as a tumor). This experiment was performed repeatedly until all tissues had been left out once, using several different starting parameters. An example of a typical set of clones that resulted from this analysis is shown in Table 5. An example of an optimal tissue clustering result is given in Figure 5.

#	Gene name	Comment
62	GAPDH	
6	ferritin H	ovarian cancer
4	collagen 1A1	
3	Immunoglobulin	TIL
2	EST	
2	HE4	ovarian cancer
2	Keratin 18	breast cancer
2	MHC	TIL
2	Mitochondrial	metabolism
2	Novel	
2	SLPI	
1	TMPO	cell proliferation
1	SSR4	
1	PKM 2	hepatoma
1	Lactate dehydrogenase	
1	COX7b	
94	**TOTAL**	

Table 5 - Example of clones that were able to group the tissues into tumors and normals
Typical result from the tissue clustering. The first column lists the number of clones found per gene. When performed using different starting parameters, Bioclust™ would come up with a similar set of genes, differing in the genes that appear only once. GAPDH, ferritin H, the immunoglobulins, collagen 1A1, the Major Histocompatibility Complex genes (MHC), and the mitochondrial genes, were present in all of those sets. TIL: tumor infiltrating lymphocyte.

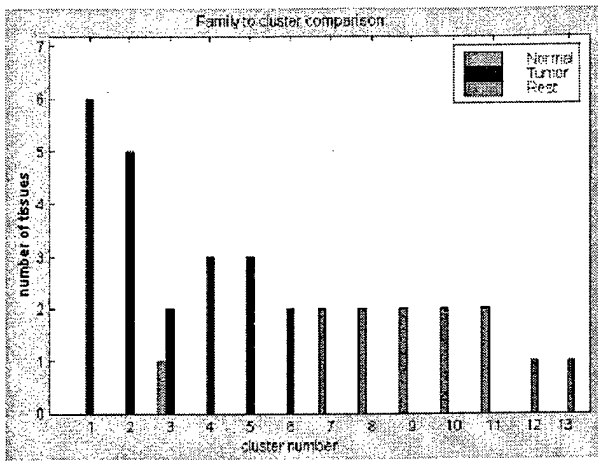


Figure 5 - Example of a tissue clustering result on the entire dataset

Displayed is a typical result for the leave-one-out tissue clustering analysis. The software generated 6 groups which - with the exception of one normal tissue - consist of tumors and five groups that contain only normal tissues. The duplicate and triplicate hybridizations of one tissue were treated as if they had been derived from separate tissues. As a result they either cluster in separate groups, which would be an indicator of low similarity, or they cluster in the same groups, indicating that they are indeed very similar to each other. Of the 7 tissues with repeated hybridizations, 5 have their replicates cluster in the same groups, one has two replicates in a "tumor" group and another replicates in a neighboring "tumor" group, and one has two replicates in a "normal" group and a single replicate in a "tumor" group. The groups 1-13 are formed from the following tissues: 1: hwbc3, t037, t051, t051a, t040, t065; 2: t025, t060, t066, t044a, t044b; 3: n050a, t048, t044; 4: t046, t046a, t046b; 5: t063, t048a, t048b; 6: n039a, t043; 7: hpbl7, hpbl8; 8: n047a, n047b; 9: n050, n050b; 10: hliv2, hpbl6; 11: n056, n064; 12: t062; 13: t058. An "a" or a "b" behind the tissue name refers to the duplicate and triplicate hybridization.

Each of the clustering experiments resulted in a list of genes (See legend to Table 5) of which many were found to be the same in different experiments. These clones included such metabolism-related genes as the mitochondrial genes, ribosomal proteins, elongation factors and GAPDH. The non-meatbolism genes which were picked up by all clustering experiments are listed in Table 6. HE4, SLPI and S100A11 have been discussed above with respect to their tumor-relatedness. Beta-actin transcript was reported to be expressed at higher levels in colorectal neoplasia (Naylor et al., 1992). CD24 is a known marker for breast cancer (Fogel et al., 1999). ESE-1, or ELF3, is an epithelial-specific transcription factor (Oettgen et al., 1997) related to the *ets* family and is expressed in lung carcinomas (Tymms et al., 1997). Folate Binding Protein was previously reported to be overexpressed in ovarian carcinomas (Toffoli et al., 1997). GPR39 is a G-protein coupled receptor (McKee et al., 1997). Keratin 8 is an epithelial gene that was reported to be expressed in a variety of tumors. It may be of diagnostic value in cervical cancer (Martens et al., 1999). Pax 2 is expressed in Wilm's tumors (Davies et al., 1999). It encodes a DNA binding, transcription factor whose expression is essential for the development of the renal epithelium (Dressler and Woolf, 1999).

This selection of genes show that the cluster analysis was capable of detecting among our 2651 clones a large number of cancer-related genes. It was therefore our primary interest to a) confirm their expression by a method other than array hybridization and to focus on the genes and clones with no previous cancer role ascribed to them, such as SLPI and GPR39 and the sequences with no match to the known gene databases. The validation of gene expression is described in Task 4.

Gene name	GenBank Accession Number
5 ESTs	AA522512, AI271417, AA131674, AW300236, AL080004
2 novel sequences	
beta-actin	NM_001101
CD24	L33930
ESE-1	U73844
Folate BP	X69516
GPR39	AF034633
HE4	X63187
Keratin 8	G4504918
PAX 2	AH006910
S100A11	D38583
SLPI	NM_003064

Table 6 - Selection of genes that were found by the cluster analysis of the membrane data based on the expression of 2651 clones in 30 tissues
After performing several rounds of cluster analysis both of the clones and the tissues, we found more than 100 clones that were positive in all experiments. Of these, most coded for metabolism-related genes, including GAPDH, with low marker potential. The other genes are listed here.

Task 3: Further characterization of gene expression in normal and neoplastic ovarian tissue

The 883 genes identified in Task 2 should ideally display an expression pattern that is higher in the tumors than in the normal ovarian tissues. There are several reasons why this observed behavior may not coincide with the actual one. Firstly, there were only 32 tissues used and we don't know how a gene would fare in other tissues. Secondly, there is heterogeneity in cell composition between tissues of the same kind and within the same tissues. Thirdly, the array consisted of single spotted colonies, and commonly triplicate spotting and above is regarded as statistically relevant (Ichikawa et al., 2000, Geiss et al., 2000). Fourthly, the method of hybridization, and image processing adds a certain variation to the values. As a consequence, we did not regard the 883 genes as the final set of cancer genes and rearrayed them on glass for interrogation with more tissues. Glass arrays, if processed properly, have lower signal-to-noise ratio than membrane arrays and due to double spotting combined with double

hybridization, each value is more dependable. But even here some of the aforementioned factors apply.

For each of the 883 genes we selected the longest clones and the ones with the best sequence. Our cDNA glass array could hold as much as 1536 genes or clones. Some of these positions are reserved by controls such as RNA, polyA, non-human clones (Arabidopsis), vector sequence and repeats. These controls amounted to 23 positions, leaving us with 1513 positions to fill. Form earlier cDNA expression arrays, we had accumulated clones with potential as markers for ovarian cancer, one of them being the dataset published in *Gene* (Schummer et al., 1999). These genes together with our 883 genes were PCR amplified using vector-specific primers (mapping 150 bp upstream and downstream from the multiple cloning site) in 100 µl reactions. The PCR products were concentrated to 10 µl, and 10 µl of 100% DMSO was added to prevent evaporation during the arraying process. We used a Generation II arrayer from Molecular Dynamics to array the cDNAs onto Type 7 "mirrored" slides from Amersham. These slides contain an aluminum coating rendering them reflective, thus maximizing the photon yield. Each cDNA was spotted as duplicate. We generated twice as many slides as tissues to hybridize which enabled us to perform duplicate hybridizations, thus lowering the experimental error. The arrays were hybridized with first-strand cDNA probes generated from the tissues listed in Table 7. The hybridization and data extraction was performed according to the conditions described earlier (Geiss et al., 2000) with the exception of the reference probe which was generated from a pool of RNAs from all 64 tissues. Glass arrays allow for cohybridization of two probes, one labeled with Cy3 and one labeled with the Cy5 dye (further referred to as green and red dye). One color is used on the tissue to be interrogated, the other on the reference. Finally, two separate hybridizations to two identical glass arrays were performed for each tissue to be interrogated, one with the tissue cDNA labeled in green and the reference in red, and the second hybridization with the colors swapped. The color swapping compensates for differences in labeling efficiency and light emission of the dyes. The software developed by Roger Bumgarner at the University of Washington merges the four values (one from each duplicate spot times 2 for the duplicate slides) and writes one averaged value into a database together with an assessment of the quality of the individual hybridization. As a result, for each gene on the array and for each tissue, the database hosts a record of the expression relative to the reference.

	Tissue Type		Tissue Type
*	Liver		
*	white blood cell culture		
*	normal ovary	*	stage I ovarian carcinoma
*	normal ovary	*	stage III ovarian carcinoma
*	normal ovary	*	stage III ovarian carcinoma
*	normal ovary	*	stage III ovarian carcinoma
*	normal ovary	*	stage III ovarian carcinoma
*	normal ovary	*	stage III ovarian carcinoma
*	normal ovary	*	stage III ovarian carcinoma
*	normal ovary	*	stage III ovarian carcinoma
*	normal ovary		stage III ovarian carcinoma
	normal ovary		stage III ovarian carcinoma

	normal ovary		stage III ovarian carcinoma
	normal ovary		stage III ovarian carcinoma
	normal ovary		stage III ovarian carcinoma
	normal ovary		stage III ovarian carcinoma
	normal ovary		stage III ovarian carcinoma
	normal ovary		stage III ovarian carcinoma
	normal ovary		stage III ovarian carcinoma
	normal ovary		stage III ovarian carcinoma
	normal ovary		stage III ovarian carcinoma
	normal ovary		stage III ovarian carcinoma
	normal ovary		stage III ovarian carcinoma
	normal ovary		stage III ovarian carcinoma
	normal ovary		stage III ovarian carcinoma
	fallopian tube from ovarian cancer patient		stage III ovarian carcinoma
*	benign ovarian cystadenoma		stage III ovarian carcinoma
*	benign ovarian cystadenoma		stage III ovarian carcinoma
	benign ovarian cystadenoma	*	stage IV ovarian carcinoma
	benign ovarian cystadenoma	*	stage IV ovarian carcinoma
	benign ovarian cystadenoma	*	stage IV ovarian carcinoma
	benign ovarian cystadenoma	*	stage IV ovarian carcinoma
	benign ovarian cystadenoma		stage IV ovarian carcinoma

Table 7 - Tissues used for interrogation of the 1536 clones glass array
The 64 tissues are color coded. Blue stands for normal, non-ovarian tissues, green for normal ovarian tissues, orange for benign ovarian tumors and red for invasive ovarian carcinomas. The asterisk in the first column marks the 25 tissues that were previously used for the interrogation of the membrane array. For the remaining 5 tissues from the membrane array, the tissue RNA was used up for the interrogation of the membrane array and the experiment could not be repeated on glass.

The result from the hybridization was a dataset consisting of a matrix of 64 tissues and 1536 clones (equaling genes). This matrix can be plotted as a spreadsheet with the tissues representing the columns and the clones representing the rows.
During the last 4 years, algorithms were developed for the analysis of complex biological datasets. Some were originally designed to sort and understand other scientific datasets, such as those generated in epidemiology, others were tailored to the array data. We used the clustering approach to extract from our data the genes with potential for markers. These genes needed to show higher expression in the ovarian tumors (invasive, benign or both) than in the normal ovaries and in the liver. The algorithm we employed was written for array data and optimized during our month-long analysis.
We performed clustering analysis using Bioclust™ (Ben-Dor et al., 2000a), clustering both tissues and genes. The clustering experiments were performed as described above under "Task 2 addendum". We found 126 genes with elevated expression patterns in the tumors, among them 8 novel genes and 30 ESTs. These genes are listed in Table 8. It is remarkable that this

list encompasses the genes found earlier when analyzing the membrane data (see Table 6), which is another proof of the quality of our analysis tools given the high degree of error associated with the membrane data. It is also remarkable that all 7 marker genes found earlier to be overexpressed in ovarian tumors by screening an array of 21,000 cDNAs with probes from five ovarian tumors and OSE are contained in this dataset (Schummer et al., 1999). These genes are *14.3.3*, *BA46*, *E16*, *HE4*, *mucin 1*, the *putative progesterone binding protein* and *ryudocan*.

Gene name	GenBank accession
14.3.3	x56468
actin, beta	NM_001101
adenocarcinoma-associated antigen (KSA)	X14758
amyloid protein homologue	L09209
argininosuccinate synthetase (ASS)	NM_000050
BA46	U58516
bamacan	AF067163
bikunin	U78095
c-jun	NM_002228
c-myc	X00364
Calvasculin	NM_002961
CCR2	D29984
CD24 signal transducer	L33930
CD9 antigen (p24)	NM_001769
CDC28 protein kinase 2 (CKS2)	NM_001827
CGGBP, trinucleotide repeat DNA BP p20-CGGBP	AF094481
chaperonin	X74801
CHI3L1, chitinase 3-like 1 (cartilage glycoprotein-39)	NM_001276
collagen 11A1	NM_001854
CRIP1, cysteine-rich protein 1 (intestinal)	NM_001311
cyclin-selective ubiquitin carrier protein	U73379
cytosolic malate dehydrogenase	D55654
DAP-1 (ST kinase)	X76105
density-regulated protein (DRP)	NM_003677
E16	M80244
EDN1	S56805
Efs1 or Efs2	AB001466
Enolase	NM_001428
ESE-1	U73844
FACL3, fatty-acid-Coenzyme A ligase, long-chain 3	NM_004457
Ferritin H	L20941
Folate BP	X69516
GA733-1	NM_002353
GAB2	AB018413
GAPDH	M33197
glia maturation factor-gamma (GMF-GAMMA)	NM_004877
GPR39	AF034633
gpx1, gluthatione peroxidase	X13709
haptoglobin	NM_005143
HE4	X63187
HSPD1, heat shock 60kD protein 1	NM_002156
Her2/neu	M11730
HGF	X16323

Gene name	GenBank accession
IFI27, interferon-induced protein 27	NM_005532
IGF BP2	X16302
IGF2	X07868
Kadereit	L22343
Keratin 18	NM_000224
Keratin 7	NM_005556
Keratin 8	G4504918
KIAA0762	AB018305
LDHA, lactate dehydrogenase A	NM_005566
LGALS1, lectin, galactoside-binding, soluble	NM_002305
lipocalin 2 (oncogene 24p3) (LCN2)	NM_005564
Lipocalin2	NM_005564
MAGE-4	D32075
MAGOH, mago nashi homolog	AF035940
MAT1	L37385
MCAF	M24545
MDC15	U46005
Mesothelin	AF180951
MET	NM_000245
MIS	K03474
Mucin1	X52228
NME4	NM_005009
OGP, oviductal glycoprotein exon 11	U58010
Osteopontin	D14813
oviductal glycoprotein	U09550
p27 alpha-inducible protein 27 (IFI27)	X67325
p73	NM_005427
p76, endosomal, multispanning membrane prt.	U81006
Pax2	AH006910
PLTP	NM_006227
progesterone binding protein	Y12711
pyruvate kinase, muscle (PKM2)	NM_002654
RIG-E	Z68179
Ryudocan	D13292
S100A11	D38583
SAS	U01160
SCNN1A, sodium channel, nonvoltage-gated 1 alpha	NM_001038
SLPIa	NM_003064
ST5	NM_005418
STK11	AF035625
TPI1 triosephosphate isomerase	M10036
tra1, homolog of murine tumor rejection antigen gp96	X15187
TAGLN2, transgelin 2	NM_003564

HSD3B2 M77144 .UNC119 AF125997

Table 8 - Genes that were found by cluster analysis of the glass data based on the expression of 1380 genes in 64 tissues
List of 88 genes (excluding the 8 novel genes and the 30 ESTs) found as a result of the cluster analysis of the glass array data set. All genes are listed with their GenBank accession number for easy identification.

Array hybridization is ideal for the determination of the expression of thousands of genes in dozens of tissues. The array method, however, even the glass-based method using duplicate spotting, has several instances where error is introduced. Firstly, the RNA quantification by spectrophotometry is inaccurate, with standard of means of 30-50% (unpublished results). Secondly, first strand cDNA generation and probe hybridization do not always use highly reproducible results, even with the highest care taken. Thirdly, the spotting of the DNA onto the glass or the membrane can result in differences of amount of DNA that actually remains on the surface. Fourthly, the method used for spot detection and intensity integration adds a minor but detectable variability to the numbers. Taken altogether, for a single gene, we estimate the average error to be as high as 50% of its measured intensity. For this reason, we will validate the expression of the 126 genes by a method that is a) more accurate than array hybridization, b) capable of processing this large number of clones within reasonable time, and c) more sensitive than array hybridization. The method of choice is quantitative Real-Time PCR and will be described in Task 4.

Task 4: Characterization of highly expressed genes associated with cancer

Originally proposed was to sequence the cancer-related clones coming from the glass array analysis in Task 3. Since we have sequenced all clones in Task 2, this is no longer necessary. Task 4 left us with 124 genes to be further characterized. In addition, after discussion with colleagues, we decided to validate 6 more genes with known or suspected involvement in ovarian cancer (*BRCA1*, AF005068; *BRCA2*, U43746; *ESR1*, X03635; *ESR2*, AB006589; *p53*, NM_000546; *StAR*, U17280) and 11 genes our collaborators within the ovarian SPORES were working on (*AKT1*, M63167; *AKT2*, M77198; *Cyclin C*, M74091; *IL-8*, M17017; *Ku70*, J04607; *Ku80*, M30938; *LotI*, U72621; MR, NM_013404; NY-ESO-1, U87459; PIK3CA, Z29090; PTEN, U93051). This increased the number of genes to characterize to 141.

As pointed out above, the first step in the characterization of our potential marker genes was the expression validation by means of RealTime quantitative PCR. This method requires the design of two primers per gene, spaced by ~500 bp. The primers need to have similar melting temperatures (T_m) and should all be of 20-23 nucleotides in length. The primers need to be tested on cDNA that was reverse transcribed from pooled RNA from a number of ovarian cancer tissues (to minimize chances of a negative result due to absence of the transcript in a given tissue).

The primer pair will then be used in a conventional PCR supplemented with a fluorescent dye (SYBR green). This dye emits light upon UV excitation in the presence of double-stranded and single-stranded DNA, the latter with less efficiency. The PCR is performed in a 96-well plate (60 s at 94°, 40 cycles of 25 s at 94°, 25 s at 60°, 45 s at 72° using 1 U/µl of Biolase enzyme made by Bioline and 0.12 mM dNTPs, 0.12 mM of each primer, 1.5 mM $MgCl_2$ and the supplied buffer) in an ABI7700 RealTime PCR machine in which the SYBR green emission is recorded several times during each cycle, thus monitoring in real time the built-up of newly synthesized DNA molecules. The PCR machine comes with software that uses a standard on each 96-well plate to determine the DNA concentration. This standard consists of a twofold serial dilution of cDNA made from a white blood cell RNA preparation that is amplified using the primers for S31iii125 (GenBank accession number U61734), a gene which we find to be expressed in all tissues tested so far.
Since SYBR green cannot distinguish between the actual PCR product and DNA molecules that are made at random (artifacts), the PCR was run on a 1% agarose gel for determination of the quality of the PCR band. In the case of the absence of a band on the gel but the presence of a SYBR signal, we would set the resulting DNA concentration to 0. A typical RealTime PCR result is shown in Figure 6.

Figure 6 - RealTime quantitative PCR result of HE4 on 82 tissues
The tissue names are listed on the bottom. Brown stands for normal non -ovarian tissues, blue for peripheral blood lymphocytes, green for normal ovaries, orange for benign ovarian tumors, red for ovarian carcinomas of increasing stage, and the leftmost 17 entries are ovarian, breast and cervical cell lines. The y-axis shows expression of HE4 relative to the S31iii125 standard. These are arbitrary values that can nevertheless be used for comparison of the degree of expression of different genes. Beta actin, a medium high expressed gene, would show numbers in the 400 range, a lowly expressed gene would show numbers in the 0.1 range. HE4 transcript expression is, with the exception of placenta and lung, clearly restricted to the ovarian tumors. This pattern shows HE4 as a marker gene with high specificity and sensitivity.

Key to the success of a PCR is the proper design of the primer pair which requires, amongst others, an error-free DNA sequence. While this requirement is met by none of the sequences we have produced (single pass sequencing rarely results in an error-free sequence), we can, in the case of the clones that match to known genes, use the published sequences as template for

primer design. However there are caveats that prohibit the generation of functioning primer pairs. In these cases, we generated up to two primer pairs (resulting in four possible PCR products) for one gene before abandoning the primer generation altogether.
In the case of the 30 clones that match only to ESTs (that are also derived by single pass sequencing) the databases gave us several homologous sequences which we compared against each other, and we would design the primers in regions that were 100% identical. However, one shortcoming of the ESTs is their length. As pointed out above, our PCR products are typically ~500 bp long, but we can handle lengths of 350 and below. The EST sequences are often derived from oligo-dT primed cDNA libraries in which case they cover the 3' untranslated region of their gene. This region often contains repeats such as LINEs and SINES (50-150 bp length) which are unsuitable for primer placement. In an average 400 bp EST, this may leave less than 300 nucleotides for the placement of the primers, and combined with the possibility of inaccuracy of some base readings, it will be difficult to generate a PCR product. We were therefore unable to generate PCR primers for 8 of the 30 ESTs.
In the case of the 8 clones that do not match to any published sequence all of their sequences were between 250 and 450 bp long with the number of undecided base pairs growing at the 3' end. We were therefore unable to generate PCR primers in 7 cases.
In the case of the 105 known genes (88 from the glass array plus 17 genes suggested by our collaborators)the chances of generating functional primer pairs are very high. This work is still ongoing and we were so far able to generate primer pairs for 55 of the 105 known genes. In summary, we could generate functioning PCR primers on 78 genes (see Figure 1).

The RealTime quantitative PCR was performed on three separate 96-well plates containing the cDNA templates. Each cDNA was reverse transcribed from 10 µl of total RNA using the Superscript system (BRL) and after completion each cDNA preparation was filled up with water to 500 µl. The cDNAs from the template tissues (normal, non-ovarian tissues; normal ovaries; ovarian surface epithelium primary cultures; benign ovarian tumors; borderline ovarian tumors; stage I, II and IV and metastatic ovarian carcinomas; ovarian, breast, cervical cell lines) were transferred into the wells of three 96-well plates.
Plate 1 contained samples from all tissue classes and served as a prescreen. Genes that showed high expression in the normal ovaries or in the normal, non-ovarian tissues were eliminated from further validation (Table 1). This reduced the number of genes by 55. The 23 genes that showed a tendency to express higher in the ovarian tumors than in the normal tissues were passed on to Plate 2 which contained more tissues from each class, especially more ovarian tissues (normal and cancers). Again, genes that failed to show a overexpression in the ovarian carcinomas, and be it only in a few, were eliminated. This reduced the number of genes by 8, leaving 15 to be assayed on Plate 3 which contained a large number of normal-non ovarian tissues and which was therefore used as a screen for genes that do not significantly express in these tissues.
Of these 15 genes, 5 displayed a clear ovarian cancer-related expression. These genes were *ESE-1*, *GPR39*, *HE4*, *Mesothelin* and *SLPI*. Figure 7 shows their expression across all 202 tissues assayed.

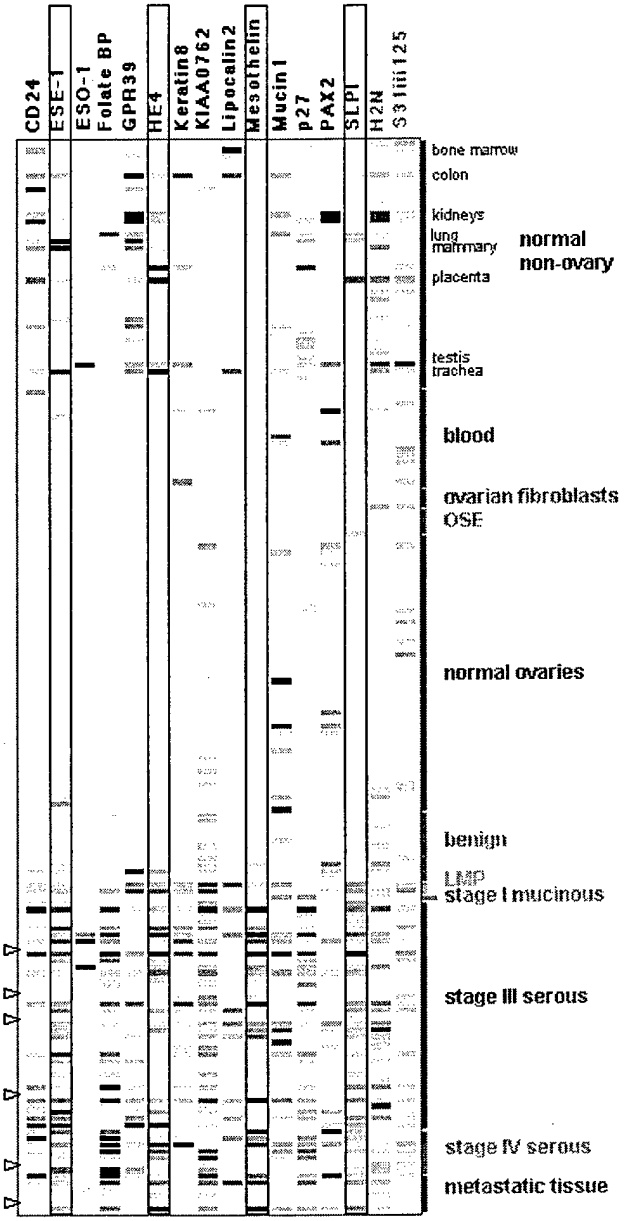


Figure 7 - RealTime data focusing on the expression of the marker genes in all tissues
The expression of 15 genes in 202 tissues was determined by RealTime quantitative PCR. Listed on the right are the tissues using the same colors employed throughout the report. The names of the genes are listed at the top. The expression values are expressed as greyscale bands with black standing for high expression and white for low. The four best performing genes are highlighted. The values are not normalized since normalization requires a gene or a group of genes with prior knowledge of their unchanged expression in the tissues tested. Since this is impossible, we have included in this panel the gene S31iii125 *which is expressed in all tissues shown, albeit with some variation. We would like to point out that had we normalized by the values of this gene, the overall expression pattern would still look the same with some bands being darker or lighter than otherwise. The open triangles on the left side mark tissues that show no elevated expression for either of the marker genes. H2N stands for Her2/neu.*

All of the 15 genes (with the exception of Her2/neu) can discriminate between ovarian cancers and normal ovaries, some better, some worse, and they are thus potential markers for ovarian cancer. One word of caution: the evaluation of these markers requires the availability of tissue which forfeits a role in early detection screening. Early detection screening requires a non-invasive test which by all standards means that the protein be present in the blood (see below). Nevertheless, the marker genes we have found so far may be useful for staging and prognosis. This needs to be evaluated on a larger set of cancer tissues, not only of the serous histology but also of mucinous and endometrioid. We will then be able to answer questions about the discrimination between benign, borderline and invasive tumors. With a larger number of normal tissues we will be able to discriminate between the causes for the oophorectomies that gave us the tissues (such as having had breast cancer, breast and ovarian cancer running in the family and ovarian cysts). Figure 8 shows that some of these new markers are indeed able to complement CA-125 (black dot behind tissue name). It is noteworthy that none of the potential marker genes was able to complement CA125 in all cases and that only the combination of several such markers proved successful. Similarly, out of the 5 patients who had normal ovaries but showed CA-125 levels above 30 U/ml, the newly found markers were low in 4 cases. The fifth one, labeled as n088, shows elevated transcript levels for 3 genes but low levels of all other genes. This hints at a rather large number of markers that need to be combined for high sensitivity and specificity.

CA125	p53	Her2/neu	CD24	ESE-1	ESO-1	GPR39	HE4	Kera-tin8	Lipo-calin	Meso-thelin	Mucin1	p27	PAX2	SLPI	Tissue
U/ml	*	*	**	**	**	**	**	**	**	**	**	**	**	**	
21	0	0	0.0	0.0	0.0	0.1	0.0	0.0	0.3	0.0	0.0	0.8	0.0	0.0	n022
17	3		0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.3	0.0	0.0	n029
7	0	0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	n033
7	3	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	n035
43	1	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.0	n041
9	1	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	n045
24	0	0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	n047
	0	0	0.0	0.1	0.0	0.3	0.0	0.1	0.3	0.0	0.2	0.1	0.0	0.0	n049
20	5	3	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.0	n050
5	1	1	0.0	0.0	0.4	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.0	n052
	0	0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	n053
5	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	n054
14	0	0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.0	n055
7	0	0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	n056
31	1	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	n057
	8	8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.4	0.1	0.0	n059
10	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	n064
	5	1	0.0	0.2	0.0	0.2	0.0	0.0	0.0	0.0	58.8	0.3	2.4	0.0	n082
24	0	0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	n083
9	5	3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	n084
8	0	0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	n085
14	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.1	n087
8	0	0	0.3	0.0	0.0	0.2	0.0	0.1	0.0	0.0	0.0	0.1	7.9	0.2	n088
8	0	0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	1.0	0.0	n089
12	0	0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	4.9	0.0	3.0	0.0	n090
14	0	0	0.0	0.1	0.0	0.0	0.0	0.1	1.1	0.0	1.2	0.1	5.0	0.0	n092
11	0	0	0.0	0.2	0.0	1.8	0.3	0.2	0.0	0.0	1.6	0.0	0.0	0.1	n093
	5	3	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0	0.0	0.1	n094
17	0	0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.1	4.9	0.1	0.0	0.1	n095
17	0	0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.3	0.0	0.1	n096
	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.1	0.0	0.0	n097
64	0	0	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.7	0.2	n100
	1	1	0.0	0.2	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0	27.5	0.1	n102
10	0	0	0.0	0.1	0.0	0.2	0.0	0.1	0.0	0.0	0.0	0.3	0.6	0.0	n103
	5	3	0.0	0.2	0.0	0.0	0.0	0.1	0.4	0.0	0.0	0.1	0.0	0.0	n105
	0	0	0.5	0.5	0.0	0.0	0.0	0.8	0.0	0.0	1.2	0.1	0.0	0.1	n106
7	1	1	0.7	2.1	0.2	0.4	1.7	1.8	0.5	0.9	0.2	0.7	0.0	0.6	t018b ●
61	0	0	0.2	0.1	0.0	0.6	0.2	0.3	1.0	0.1	0.2	0.1	0.0	0.3	t027b
8			0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.3	0.1	0.0	t034b ●
7	0	0	0.2	0.1	0.0	0.2	0.4	0.1	0.0	0.1	0.0	0.0	0.1	0.5	t042b ●
22	0	0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	6.9	0.0	0.0	0.0	t091b ●
4	0	0	0.0	0.1	0.0	1.0	0.0	0.1	0.0	0.5	0.4	0.0	0.7	0.3	t099b ●
	1	1	0.0	0.1	0.0	0.1	0.0	0.1	0.7	0.0	0.0	0.0	0.0	0.2	t115b
81	5	5	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.3	0.2	0.4	0.0	t125b
	0	0	0.4	1.2	0.0	0.9	0.7	0.2	0.1	1.4	0.8	1.0	5.9	0.3	t202b
	1	1	2.0	1.3	0.3	15.0	3.1	0.7	0.4	0.8	0.3	1.3	4.5	1.3	t203b
24	0	0	0.4	0.1	0.0	16.5	0.0	1.6	1.1	0.0	0.0	0.2	23.1	0.0	t204b ●
202	1	1	0.5	0.4	0.0	2.8	1.3	1.0	0.0	0.1	0.1	0.1	0.4	0.8	t028l
174	1	3	0.0	0.0	0.0	0.0	0.1	0.2	0.0	0.2	0.0	0.3	0.2	0.7	t030l
310	5	3	0.0	0.1	0.0	0.5	0.0	0.4	0.0	0.0	0.8	1.1	0.2	0.3	t037
	5	1	0.3	0.4	0.6	0.3	1.7	0.8	1.8	1.7	0.4	1.8	0.0	1.3	t019
170	1	1	0.8	2.2	3.8	0.3	0.6	4.1	0.1	1.6	0.8	0.6	1.9	0.7	t021
	5	3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	t025
9	1	3	0.2	0.1	0.0	0.8	1.6	1.1	0.5	0.5	0.1	1.3	0.1	1.0	t031 ●
			0.2	0.0	0.1	0.0	0.0	1.3	0.0	0.1	0.0	1.0	0.0	0.7	t038m
	8	8	0.5	0.2	0.0	0.3	1.4	0.5	0.0	2.1	0.6	0.4	0.0	0.6	t043
342			0.4	1.2	0.0	0.0	1.0	0.4	0.3	0.0	0.2	1.1	0.0	0.5	t048
			0.0	0.0	15.7	0.0	0.3	0.1	0.1	0.3	0.0	0.3	0.0	0.2	t051
	5	3	0.3	0.2	0.0	0.3	0.7	0.2	0.3	0.3	0.4	0.4	0.3	0.4	t060
14	1	3	0.1	0.2	0.0	0.0	1.7	0.9	0.0	0.9	0.2	0.5	0.0	0.4	t061 ●
18	1	1	1.1	10.5	0.0	0.9	2.2	0.3	0.1	0.4	0.1	2.5	2.6	0.7	t063 ●
18	0	0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.5	0.0	0.5	t066 ●
91	0	0	0.5	0.4	0.0	0.8	0.1	0.4	0.5	0.6	0.0	0.4	0.8	0.7	t086
	0	0	1.0	1.1	0.0	3.5	0.4	5.6	0.0	3.9	0.0	2.7	0.0	1.4	t098
36	0	0	0.0	1.5	0.0	0.0	1.0	1.2	2.6	0.0	0.0	0.6	0.0	1.1	t101
51	5	3	3.1	2.3	0.0	7.3	3.7	0.2	0.2	0.0	0.1	0.1	0.7	1.3	t104
			0.9	2.3	0.0	0.0	1.8	3.6	8.5	2.7	2.5	1.6	4.8	2.0	t107
	5	3	0.1	2.0	0.0	1.4	3.2	2.4	1.4	3.2	4.9	0.2	1.9	2.4	t108
344	0	0	0.1	4.6	0.0	4.1	0.8	3.2	8.0	6.5	0.5	0.4	1.8	4.2	t109
	5	5	0.0	0.6	0.0	0.6	0.2	0.4	0.0	0.7	11.4	0.6	0.0	1.1	t110
	5	3	0.6	0.5	0.0	0.6	0.7	1.8	1.2	0.3	0.0	0.3	1.5	0.3	t111
	0	0	0.9	2.6	0.0	0.5	0.6	1.4	0.0	0.0	1.0	1.9	0.0	0.7	t112
	5	1	0.3	1.6	0.0	3.1	0.6	0.6	0.9	0.0	0.3	1.4	0.0	0.3	t113
306	8	8	0.2	1.1	0.0	0.7	0.5	0.7	0.0	0.0	0.0	0.1	0.0	1.5	t114
	1	5	0.2	1.1	0.0	0.8	0.7	0.6	0.5	0.0	0.6	0.0	0.0	1.8	t116
51	5	5	0.0	0.2	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.1	0.0	0.1	t117
	5	3	1.8	0.8	0.0	1.4	0.3	1.3	0.2	0.2	0.0	0.8	0.0	2.1	t118
	1	3	0.0	0.1	0.0	0.0	0.1	0.1	0.0	0.1	0.6	0.1	0.0	0.1	t120
	1	1	1.4	1.5	0.0	1.4	0.7	1.9	2.2	3.2	1.2	1.0	0.0	1.5	t122
	5	3	2.5	0.5	0.0	1.0	1.5	0.8	0.0	0.1	0.0	0.1	0.0	1.0	t124
8	1	1	2.9	2.9	0.0	3.7	0.3	0.3	2.5	0.0	0.7	1.6	0.4	2.5	t206 ●
			1.7	0.9	0.0	1.9	0.5	0.9	2.3	0.9	0.0	0.1	0.9	2.2	x119
	0	0	0.6	0.9	0.0	0.5	1.6	5.0	0.4	1.0	0.5	0.8	1.1	1.1	t040
	1	1	0.5	0.4	0.0	0.2	1.3	0.3	0.6	0.0	0.5	1.6	0.0	0.6	t044
63	1	1	0.1	0.4	0.1	0.2	1.0	0.2	0.0	0.0	0.4	2.1	0.0	0.1	t046
			0.0	0.0	0.0	0.2	0.0	0.1	0.1	0.0	0.0	0.1	0.0	0.0	t065
	0	0	1.0	1.4	0.0	1.6	0.5	0.5	2.1	0.4	0.3	0.2	0.0	1.6	t123

Figure 8 - Combined protein/transcript data focusing on the tissues of which there is CA125 information available

CA-125 serum levels of the ovarian cancer patients and the controls paired with the tissue protein levels of p53 and Her2/neu, listed side-by-side with the transcript levels of selected potential marker genes found in this study.

*The patient diagnosis / tissue type is listed in the rightmost column (colors are the same as used in Figure 7). Values are overlaid with color for easier identification: CA-125: 0-29 U/ml (turquoise), 30-99 U/ml (faint red), 100-399 U/ml (red), over 400 U/ml (dark red), white: not done. * p53 and Her2/neu: 0, assay not run;1, no overexpression (turquoise); 3, uninterpretable (light red); 5, intermediate overexpression (red); 6, high overexpression (dark red); 8, assay will not be run. The RealTime quantitative PCR values were normalized by the average expression of each gene in all tissue in order to have the values in each column on the same scale. ** 0-0.1, no expression (white), 0.2-0.9 weak expression (light red), >1.0, high expression (red). Ovarian cancer patients with CA-125 levels below 30 U/ml that have high levels of one or more of the newly found markers are labeled with a black dot after the tissue name.*

Task 5: Final Analyses

The 15 potential marker genes show a great marker potential but so far they require to be tested on tissues which require an invasive procedure to obtain. Biopsies may be acceptable procedure in high risk populations but they are not acceptable for the screening of a general population. Due to the relatively low incidence of ovarian cancer (25,000 new cases every year in the US (American Cancer Society, 1998)) early detection can only be achieved using an inexpensive test with high specificity (Urban, 1999) and sensitivity that uses body fluids such as blood, saliva or urine. The most commonly employed body fluid-based tests are ELISA which detects proteins via an antibody, Western blot, using an antibody too, and quantitative PCR which detects transcripts in circulating cells. The latter can be done using the information and the resources gathered so far, the protein detection assays require that we express the protein and raise monoclonal antibodies to it.

It has to be pointed out that the transcript level of a gene does not always correlate with the amount of protein that is made (Anderson and Seilhamer, 1997). And even if the elevated transcript level were translated into elevated protein level, we estimate the odds to find the protein in the sera of patients to be less than 50%.

We have therefore decided to follow a two-pronged approach, attempting to detect transcripts in circulating cells in blood and peritoneal washes while expressing fusion proteins of the genes and raising monoclonal antibodies against them with the goal of developing a sandwich-ELISA.

As shown in Figure 1, we have chosen 5 genes for the antibody generation: *ESE-1*, *GPR39*, *HE4*, *Mesothelin* and *SLPI*. *Mesothelin* and *SLPI* show the highest degree of specificity and sensitivity on the transcript level. By sheer coincidence, ELISA tests are available for both proteins. For the other genes, we are currently expressing fusion proteins (in collaboration

with Drs Ingegerd and Karl Erik Hellström and Dr Jeff Ledbetter at the Pacific Northwest Research Institute in Seattle) for the immunization of mice which will ultimately lead to the generation of monoclonal antibodies. In the case of HE4, the mice have been immunized twice and the harvesting of the B-cells in to begin.

SLPI protein in serum

SLPI codes for a secreted protease inhibitor that is expressed in mast cells where it may inhibit a mast cell chymase (Westin et al., 1999). SLPI inhibits leukocyte-derived proteinases, has anti-HIV-1, antibacterial, and antifungal properties, and interferes with the induction of synthesis of proinflammatory mediators in monocytes and macrophages (Wingens et al., 1998). Our RealTime PCR results suggest that SLPI is expressed in the salivary gland, in the mammary gland, in the lung, testis, spinal chord, bone marrow, colon, kidney and uterus. SLPI expression was significantly higher in the ovarian cancers which led us to believe that the protein levels may be elevated as well. SLPI in mucosal fluids inhibits HIV-I (Wahl et al., 1997) which is why a Dutch company (Hbt HyCult biotechnology, Uden, Netherlands) developed an ELISA to assay SLPI levels in saliva and possibly correlate them with protection against HIV infection. We have used this ELISA kit on serum samples from 10 ovarian cancer patients whose tissues showed high levels of SLPI transcript expression. We paired these results with sera from 10 normal individuals. The assay was performed in the laboratory of Drs Ingegerd and Karl Erik Hellström at the Pacific Northwest Research Institute in Seattle. Figure 9 shows that there is no difference in serum SLPI levels between these two groups. This disappointing finding correlated well with the fact that SLPI is present in a lot of tissues that could potentially contaminate the blood. The much higher transcript levels in the ovarian tumors do obviously not translate into higher serum levels of the protein.

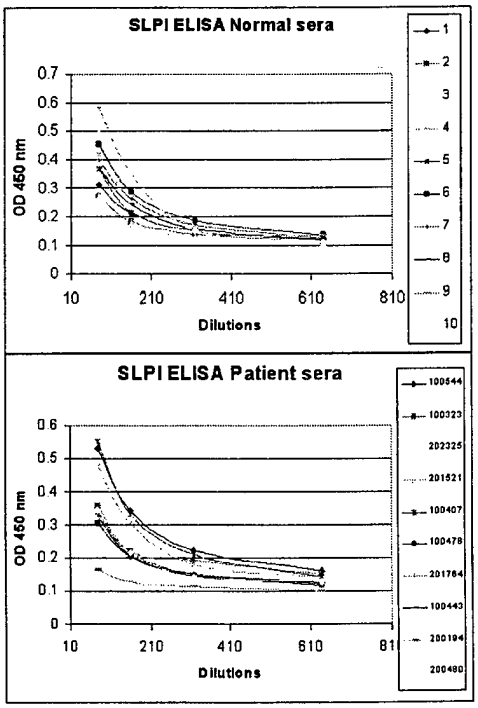


Figure 9 - SLPI ELISA on sera from 10 ovarian cancer patients and 10 controls
Top: sera from 10 normal controls, bottom: sera from 10 ovarian cancer patients whose tissues showed overexpression of SLPI message as assayed by RealTime quantitative PCR (see Figure 7). There is no difference in protein levels between the two groups.

SLPI and HE4 in cells from peritoneal washes

Although SLPI protein was not found to be elevated in patient sera, there is a remote possibility that it is elevated in epithelial cells found in peritoneal fluids. While such as test cannot be regarded as non-invasive, it may be useful in the determination of cancer risk in the case of high-risk women. In collaboration with Dr Nancy Kiviat at Harborview Hospital in Seattle we amplified the transcripts of SLPI and HE4 from epithelial cells gathered from peritoneal washes by means of RT-PCR. Peritoneal washes are carried out only in cases of disease which is why there are no true normal, disease free controls. We compared three groups of patients, those with ovarian cancer, those with other malignancies and those with benign disease and we could not find any difference in neither SLPI nor HE4 expression between these three groups (Table 9). It has to be pointed out that we only assayed for presence or absence of the message but not for the amount of it. It therefore needs to be determined whether SLPI and HE4 are expressed at higher levels in the ovarian cancers. We are currently collecting more peritoneal washes to extract cells for RealTime quantitative PCR and to perform a SLPI ELISA.

		Cytology			SLPI PCR			HE4 PCR		
Histology Diagnosis	# of cases	pos	neg	ND*	pos	neg	ND	pos	neg	ND
ovarian cancer	23	14	9	0	15	4	4	12	3	8
other adenocarcinoma	12	1	11	0	11	1	0	10	2	0
other cancer	13	4	9	0	8	3	2	8	3	2
benign	17	0	17	0	13	3	1	13	2	2
missing	3	0	3	0	1	2	0	3	0	0
TOTAL	68	19	49	0	48	13	7	46	10	12

Table 9 - Cytology and RT-PCR amplification of cDNA from ascites fluid and peritoneal washes by histology diagnosis
68 ascites fluids and peritoneal washes were tested by RT-PCR for the presence of SLPI and HE4 messages. The cytology records whether malignant cells were found in the fluid or not. SLPI mRNA was detected by RT-PCR in 13/16 (81%) benign fluids, 19/23 (83%) malignant fluids from cases other than ovarian cancer, and 15/19 (79%) malignant fluids from ovarian cancer cases. HE4 mRNA was detected by RT-PCR in 13/15 (87%) benign fluids, 18/23 (78%) malignant fluids from cases other than ovarian cancer, and 12/15 (80%) malignant fluids from ovarian cancer cases.

Mesothelin serum assay

The laboratory of Drs Hellström at the Pacific Northwest Research Institute have developed an antibody against Mesothelin which was used to test for the presence of Mesothelin protein in the sera of patients with mesotheliomas. In collaboration with the Hellström laboratory, we are currently performing ELISA tests to determine the expression levels of mesothelin in the serum of ovarian cancer patients. One obstacle on the way is that mesothelin is a member of a family of related proteins. The mesothelin that was assayed in the RealTime quantitative PCR was the membrane-bound form. Whereas the antibody recognizes the same N-terminal amino acid sequence as the membrane-bound portion of mesothelin, it also binds a novel, soluble form of this family which has an 82-bp insert in the membrane-associated part, leading to a frameshift of 212 bp (Scholler et al., 1999). It may therefore be necessary to generate an antibody against the part of the membrane-bound mesothelin that is not shared.

Outlook

We have applied for funding to resequence ~400 clones that either failed to produce a single PCR band or that did not deliver a satisfying sequence. This includes the 7 novel genes which we failed to generate PCR primers for. We have also received funding to generate monoclonal antibodies against three proteins: HE4, ESE-1 and GPR39. Within the next month we expect to have one of the two monoclonal antibodies needed for a Sandwich-ELISA. The other antibodies will follow shortly.

CONCLUSIONS:

We have identified several potential ovarian cancer marker genes, some of them previously known to be related to cancer (HE4, Folate binding protein, CD24, Keratin 8, Mucin 1, Lipocalin, S100A11, S100A6, p27, Her2/neu), some with no previous role in cancer (ESE1, GPR39, SLPI, Mesothelin), some matching to ESTs (22) and some being novel genes (8). These genes and their proteins are currently being evaluated by research laboratories other than ours. The antibody generation is carried out at the Pacific Northwest Research Institute in Seattle I. and KE Hellström); in situ hybridization on tissue sections and transcript quantitation on cells from peritoneal washes are conducted at Harborview Hospital in Seattle (N. Kiviat). Markers from this study are going to be used in a new study funded through the Ovarian SPORE where multiple markers, including CA125, will be evaluated together.

There is little doubt that a useful marker will be a protein that can be found in the blood. It was only during the last 2 years that we discovered the unexpectedly low concordance between mRNA and protein levels of some genes. If we had to repeat the effort of finding a marker again, given the recent advances in proteomics, we would put a lot more emphasis on the protein side. One proposed approach would make use of the ICAT-based protein labeling method to identify membrane-bound proteins in cancer tissues (Gygi et al., 1999).

REFERENCES:

American Cancer Society (1998) *American Cancer Society - Cancer Facts and Figures,* American Cancer Society Inc, Atlanta, GA.

Anderson, L. and Seilhamer, J. (1997) *Electrophoresis,* 18, 533-537.

Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M. and Yakhini, Z. (2000a) *The Forth Annual International Conference on Computational Molecular Biology - RECOMB'2000,* , 54-64.

Ben-Dor, A., Shamir, R. and Yakhini, Z. (2000b) *J. Comp. Biol.,* in print.

Chang, T. J., Juan, C. C., Yin, P. H., Chi, C. W. and Tsay, H. J. (1998) *Oncol Rep,* 5, 469-471.

Chen, C. L., Ip, S. M., Cheng, D., Wong, L. C. and Ngan, H. Y. (2000) *Clin Cancer Res,* 6, 474-479.

Davies, J. A., Perera, A. D. and Walker, C. L. (1999) *Int J Dev Biol,* 43, 473-478.

Desprez, P. Y., Poujol, D. and Saez, S. (1992) *Cancer Lett,* 64, 219-224.

Dressler, G. R. and Woolf, A. S. (1999) *Int J Dev Biol,* 43, 463-468.

Filipek, A., Wojda, U. and Lesniak, W. (1995) *Int J Biochem Cell Biol,* 27, 1123-1131.

Finnegan, M. C., Goepel, J. R., Hancock, B. W. and Goyns, M. H. (1993) *Leuk Lymphoma,* 10, 387-393.

Fogel, M., Friederichs, J., Zeller, Y., Husar, M., Smirnov, A., Roitman, L., Altevogt, P. and Sthoeger, Z. M. (1999) *Cancer Lett,* 143, 87-94.

Geiss, G. K., Bumgarner, R. E., An, M. C., Agy, M. B., van 't Wout, A. B., Hammersmark, E., Carter, V. S., Upchurch, D., Mullins, J. I. and Katze, M. G. (2000) *Virology,* 266, 8-16.

Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H. and Aebersold, R. (1999) *Nat Biotechnol,* 17, 994-999.

Ichikawa, J. K., Norris, A., Bangera, M. G., Geiss, G. K., van 't Wout, A. B., Bumgarner, R. E. and Lory, S. (2000) *Proc Natl Acad Sci U S A,* 97, 9659-9664.

Kauppila, S., Saarela, J., Stenback, F., Risteli, J., Kauppila, A. and Risteli, L. (1996) *Am J Pathol,* 148, 539-548.

Kim, J. W., Kim, S. J., Han, S. M., Paik, S. Y., Hur, S. Y., Kim, Y. W., Lee, J. M. and Namkoong, S. E. (1998) *Gynecol Oncol,* 71, 266-269.

Komatsu, K., Andoh, A., Ishiguro, S., Suzuki, N., Hunai, H., Kobune-Fujiwara, Y., Kameyama, M., Miyoshi, J., Akedo, H. and Nakamura, H. (2000) *Clin Cancer Res,* 6, 172-177.

Lahousen, M., Stettner, H. and Purstner, P. (1989) *Baillieres Clin Obstet Gynaecol,* 3, 201-208.

Martens, J., Baars, J., Smedts, F., Holterheus, M., Kok, M. J., Vooijs, P. and Ramaekers, F. (1999) *Cancer,* 87, 87-92.

McKee, K. K., Tan, C. P., Palyha, O. C., Liu, J., Feighner, S. D., Hreniuk, D. L., Smith, R. G., Howard, A. D. and Van der Ploeg, L. H. (1997) *Genomics,* 46, 426-434.

Moog-Lutz, C., Bouillet, P., Regnier, C. H., Tomasetto, C., Mattei, M. G., Chenard, M. P., Anglard, P., Rio, M. C. and Basset, P. (1995) *Int J Cancer,* 63, 297-303.

Naylor, M. S., Stamp, G. W. and Balkwill, F. R. (1992) *Epithelial Cell Biol,* 1, 99-104.

Nelson, P. S., Ng, W. L., Schummer, M., True, L. D., Liu, A. Y., Bumgarner, R. E., Ferguson, C., Dimak, A. and Hood, L. (1998) *Genomics,* 47, 12-25.

Oettgen, P., Alani, R. M., Barcinski, M. A., Brown, L., Akbarali, Y., Boltax, J., Kunsch, C., Munger, K. and Libermann, T. A. (1997) *Mol Cell Biol,* 17, 4419-4433.

Pedrocchi, M., Schafer, B. W., Mueller, H., Eppenberger, U. and Heizmann, C. W. (1994) *Int J Cancer,* 57, 684-690.

Persons, D. A., Schek, N., Hall, B. L. and Finn, O. J. (1989) *Mol Carcinog,* 2, 88-94.

Pinto, V., Marinaccio, M., Garofalo, S., Vittoria Larocca, A. M., Geusa, S., Lanzilotti, G. and Orsini, G. (1997) *Tumori,* 83, 927-929.

Santala, M., Simojoki, M., Risteli, J., Risteli, L. and Kauppila, A. (1999) *Clin Cancer Res,* 5, 4091-4096.

Schek, N., Hall, B. L. and Finn, O. J. (1988) *Cancer Res,* 48, 6354-6359.

Scholler, N., Fu, N., Yang, Y., Ye, Z., Goodman, G. E., Hellstrom, K. E. and Hellstrom, I. (1999) *Proc Natl Acad Sci U S A,* 96, 11531-11536.

Schummer, M., Ng, W. V., Bumgarner, R. E., Nelson, P. S., Schummer, B., Bednarski, D. W., Hassell, L., L., B. R., Karlan, B. Y. and Hood, L. (1999) *Gene,* 238, 375-385.

Tanaka, M., Adzuma, K., Iwami, M., Yoshimoto, K., Monden, Y. and Itakura, M. (1995) *Cancer Lett,* 89, 195-200.

Toffoli, G., Cernigoi, C., Russo, A., Gallo, A., Bagnoli, M. and Boiocchi, M. (1997) *Int J Cancer,* 74, 193-198.

Tokunaga, K., Nakamura, Y., Sakata, K., Fujimori, K., Ohkubo, M., Sawada, K. and Sakiyama, S. (1987) *Cancer Res,* 47, 5616-5619.

Tripathi, P. K. and Chatterjee, S. K. (1996) *Cancer Invest,* 14, 518-526.

Tymms, M. J., Ng, A. Y., Thomas, R. S., Schutte, B. C., Zhou, J., Eyre, H. J., Sutherland, G. R., Seth, A., Rosenberg, M., Papas, T., Debouck, C. and Kola, I. (1997) *Oncogene,* 15, 2449-2462.

Urban, N. (1999) *Bmj,* 319, 1317-1318.

Van Ginkel, P. R., Gee, R. L., Walker, T. M., Hu, D. N., Heizmann, C. W. and Polans, A. S. (1998) *Biochim Biophys Acta,* 1448, 290-297.

Wahl, S. M., McNeely, T. B., Janoff, E. N., Shugars, D., Worley, P., Tucker, C. and Orenstein, J. M. (1997) *Oral Dis,* 3 Suppl 1, S64-69.

Westin, U., Polling, A., Ljungkrantz, I. and Ohlsson, K. (1999) *Biol Chem,* 380, 489-493.

Wingens, M., van Bergen, B. H., Hiemstra, P. S., Meis, J. F., van Vlijmen-Willems, I. M., Zeeuwen, P. L., Mulder, J., Kramps, H. A., van Ruissen, F. and Schalkwijk, J. (1998) *J Invest Dermatol,* 111, 996-1002.

Yin, D. L., Pu, L. and Pei, G. (1998) *Cell Res,* 8, 159-165.

Yuan, C. C., Ng, H. T., Yeh, S. H., Chen, S. S., Hsu, D. S. and Ho, C. H. (1988) *J Reprod Med,* 33, 193-195.

Yun, K., Fukumoto, M. and Jinno, Y. (1996) *Am J Pathol,* 148, 1081-1087.

Project 2
Antibody Immunity to Cancer Related Proteins as a Serologic Marker for Ovarian Cancer
Nicole Urban, ScD, Brad Nelson, Ph.D., Mary L. Disis, MD

INTRODUCTION

Early diagnosis is essential to make progress in the treatment of and, ultimately, survival from ovarian cancer. Serologic markers, such as CA-125, can potentially indicate the presence of ovarian cancer. However, like many serum markers, CA-125 is shed from the surface of growing tumor and, in general, is associated with bulky disease. A serologic marker that is prevalent and readily detected in early-stage disease would be an optimal candidate to develop as a screening tool.

The immune system has evolved to detect proteins that are abnormal in terms of primary sequence, overexpression, tissue context or inflammatory context. A large number of tumor proteins are abnormal by these criteria and hence trigger T cell and B cell responses in cancer patients. Examples of tumor antigens that are common to a number of different cancers, including ovarian cancer, are p53, HER2/neu and Myc. Studies in breast cancer have shown that tumor-specific antibody responses to p53 and HER2/neu can occur early during tumorigenesis. Moreover, tumor-specific antibodies can be detected by simple and inexpensive ELISA-based blood tests. For these reasons, we are investigating whether serum antibody responses to ovarian tumor antigens could potentially serve as indicators of early-stage disease.

We hypothesize that women with ovarian cancer will demonstrate serum antibody responses to one or more ovarian tumor antigens, and that such responses will be rare or absent in women with benign ovarian disease and normal controls. This hypothesis is being tested by first analyzing antibody responses to two known tumor antigens (p53 and HER2/neu) in women with malignant and benign ovarian disease and normal controls. Second, we are using an immunoscreening technique known as SEREX to discover new tumor antigens that are recognized by serum antibodies in women with ovarian cancer. We will assess the prevalence of antibody responses to new antigens among cases and controls, as well as the extent of overlap with responses to p53, HER2/neu and Myc. The long-term goal is to assemble a panel of ovarian tumor antigens that constitute a sensitive and specific blood test for early-stage ovarian cancer.

BODY

Task 1: Perform ELISA screens for p53, HER2/neu and Myc
Months 1-24:

A. An ELISA based screen will be used to probe serum from ovarian cancer patients and control individuals for the presence of antibodies against the tumor associated

proteins p53, H2N and Myc. It is anticipated to perform this set of tests on 350 cases per year.

1. To develop reproducible assays for detecting HER2 antibodies. Data presented in the first year's report demonstrated that near CLIA grade assays have been developed for the detection of HER2 specific antibodies based on a capture ELISA format. Table 1 demonstrates long term validation data on both the HER2 and p53 antibody assays. Calculations were made on over 100 plates analyzed over 14 months. As previously reported, peptide assays and recombinant protein assays did not prove superior to the capture ELISA format developed. All blood samples collected through the ORCHID study have been analyzed for HER2 antibodies. In addition, a reference population of 175 volunteer blood donors has been analyzed. Results of the final analysis are described below (Task 5).

2. To develop reproducible assays for detecting p53 antibodies. Data presented in the first year's report demonstrated that near CLIA grade assays have been developed for the detection of p53 specific antibodies based on a capture ELISA format. Table 1 demonstrates long term validation data on both the HER2 and p53 antibody assays. Calculations were made on over 100 plates analyzed over 14 months.

In addition, we synthesized the 2 putative immunodominant B cell epitopes of p53 (see previous report). An indirect ELISA was developed. In 96-well microtiter plates (Dynex Technologies, Inc., Chantilly, VA), columns were coated with the p53 peptide, at a concentration of 20 µg/ml, diluted with carbonate buffer and added at 50 µl per well. Alternating columns were coated with 50 µl/well of carbonate buffer alone. The standard curve column, column 12, was incubated with the purified IgG titrations as above, at 4oC overnight. After overnight incubation, all wells were blocked with 1% casein/PBS, 100 µl/well and incubated at room temperature on a rocker for 1-2 hours. Plates were then washed with a 0.15%casein/1% PBS/0.05% Tween-20 wash buffer 4 times before serum diluted in 10%FCS/PBS/1% BSA/25µg/ml mouse IgG at 1:100, 1:200, 1: 400 and 1:800 dilutions. Plates were incubated for 2 hours at room temperature on a rocker. Plates were then washed 4 times with casein-based wash and incubated for 45 minutes at room temperature on a rocker after addition of 50 µl/well IgG-HRP conjugate diluted 1:10,000 in PBS/BSA buffer. After a final 4 washes with casein-based wash buffer, TMB reagent was added 75 µl/well and color reaction read at 640nm until the well containing the 0.16 µg/ml standard reached an OD of 0.3. Reaction was then stopped with 75 µl/well 1N HCL and read at 450nm. The OD of each serum dilution was calculated as the OD of the peptide-coated wells minus the OD of the buffer-coated wells. Values for delta OD were calculated from the log-log equation of the line for the standard curve on each plate. Samples that returned a positive delta OD for 3 of 4 dilutions were counted, and a positive sample was defined as a µg/ml value greater than the mean of the normal population plus 3sd.

ASSAY VALIDATION

- Normal Range: 50 serum samples from normal donors were assayed by peptide ELISA and a normal range established by determining the mean and standard deviation of all samples and calculating a cut-off value of the mean plus 3 standard deviations, a

confidence interval of approximately 99%. The peptide ELISA returned a normal mean and standard deviation of 0.052+/-0.11ug/ml, giving us a cut-off value of 0.382 ug/ml. We found that 1% of our normal samples resulted positive for any peptide.

- Accuracy: The peptide ELISA returned an average CV of 11%.

- Precision: The peptide ELISA returned an intra-assay precision and interassay precision of 9% and 17%, respectively.

Samples from 40 breast cancer patients (archived) were analyzed for p53 protein by the standard assay and the by the p53 peptide assays as described. 20% of the patients had antibody responses to p53 using the capture ELISA method. 13% had antibodies to one or both of the peptides. 2 of those patients did not have detectable p53 antibodies by protein assay. 3/5 of the peptide specific responses could be validated by Western blot. Therefore, we determined that the p53 antibody assay in the capture ELISA format (Table 1) was a more robust determination of pre-existent antibody immunity to p53. Further studies on the peptides will be undertaken as larger populations of p53 antibody positive patients are identifed in subsequent studies.

All blood samples collected through the ORCHID study have been analyzed for p53 antibodies using the capture ELISA format. In addition, a reference population of 175 volunteer blood donors has been analyzed. Results of the final analysis are described below (Task 5).

Table 1

PARAMETERS EVALUATED	RESEARCH ASSAYS		CLIA-BASED STANDARD
	HER2	P53	
Accuracy	12%	10%	<10%
Precision			
Intra assay	9%	12%	<10%
Inter assay	20%	15%	<10%
Specificity	77%	100%	>80%
Sensitivity	89%	93%	>90%
Linearity	r=0.98	r=0.95	r=0.95

3. To develop reproducible assays for detecting c-myc antibodies. Antibodies against c-myc have been reported in patients with cancer. Critical to the development of antibody assays detecting c-myc is the ability to validate responses by western blot or the availability of monoclonal antibodies to use as coating antibodies in a capture ELISA. To date we have not found an antibody to c-myc that will work reproducibly in ELISA. Likewise, the specimen core has had difficulty analyzing specimens for c-myc expression. Therefore, development of a c-myc antibody assay was abandoned. Work in the last 6 months has instead focused on the assessment of CA-125 as a potential antibody target. Preliminary studies using commercial antibodies and ovarian cancer cell lines expressing the glycoprotein demonstrates the capture ELISA template is feasible and Western blot analysis can be reproducibly performed. Experiments evaluating a reference population of volunteer blood donors are underway.

Task 2: Determine SEREX baseline
Months 1-6:

A. Conduct ten serial absorptions on sera from three normal individuals and three ovarian cancer patients with known reactivities to one or more of the p53, H2N and Myc antigens.

The goal of this task was to optimize the signal-to-noise ratio of the SEREX protocol and to pre-clear serum samples of antibodies to E. coli. Experiments performed in Months 1-6 led to a reliable pre-clearing procedure. Serum samples are first incubated overnight with matrix-immobilized protein lysates from E. coli. Serum is further pre-cleared by two serial incubations with nitrocellulose membranes containing empty lambda phage on a lawn of E. coli. This pre-clearing procedure has successfully reduced background reactivity to acceptable levels in over 60 serum samples from cancer patients and normal controls and has become our laboratory standard.

B. Construct a cDNA expression library from pooled ovarian tumor samples.

Ten stage III/IV serous ovarian tumors were used to construct a cDNA library using the lambda phage vector lambda TriplEx. Library construction proceeded as planned.

C. Assess the quality of the library.

Titration experiments demonstrated that the cDNA library contains >1 X 10^6 primary clones. PCR analysis demonstrated > 95% recombinant phage and an average insert size of 1.7 kb. Partial sequencing of six randomly picked clones showed no evidence of genomic or bacterial DNA contamination. Most important, this library has been used successfully to screen over 4.7 x 10^6 phage plaques with serum from 29 ovarian cancer patients and identify multiple tumor antigens (discussed below).

Task 3: Use SEREX to screen serum from ovarian cancer patients
Months 6-20:

A. Identify novel ovarian tumor antigens.

Serum samples from 29 ovarian cancer patients were used for primary SEREX screening of the ovarian tumor cDNA library. The library was plated on a lawn of Y1090⁻ E. coli cells at a density of 2.5 x 10^3 pfu per 100 mm NZYCM plate (NZ amine + yeast extract + 1% casamino acids + 2% $MgSO_4$ + 1.5% agar; Sigma). When phage plaques first became visible (~3h), plates were overlaid with IPTG-impregnated nitrocellulose membranes and incubated at 37°C. After overnight growth, the membranes were removed, washed in TBST (Tris buffered saline [TBS] + 0.05% Tween-20), and blocked in TBS + 1% BSA (bovine serum albumin). Membranes were then incubated overnight at room temperature with pre-cleared patient serum diluted 1:100 in TBS + 1% BSA. The following day, membranes were washed with TBS, and incubated for 45 minutes with alkaline phosphatase-conjugated goat anti-human antibody (specific for IgG, IgA and IgM; Pierce) diluted 1:7500 in TBS + 1% BSA. Membranes were developed with NBT/BCIP. As shown in Fig. 1, positive phage plaques typically appeared as darkened halos or spots. All presumptive positive phage from the primary screen were picked from the original agar plates and stored at 4°C in SM buffer (150 mM NaCl + 10 mM magnesium). All figures and tables in this project report are included as Appendix E.

Fig 1. Example of an immunoreactive phage plaque from a primary SEREX screen. The arrowhead indicates a single immunoreactive plaque (dark halo) amongst several hundred non-reactive plaques (clear spots

To date, we have performed primary SEREX screening on approximately 4.7 x 10^6 phage plaques with serum from 29 ovarian cancer patients. On average, 1-2 immunoreactive phage are found for every 2,500 phage screened. Upon re-screening with secondary antibody alone, the majority of these phage (~90%) are found to encode IgG molecules, which presumably are derived from B cells in the original tumor samples. To rapidly exclude these unwanted phage clones and identify those rare clones with cancer-specific immunoreactivity, we developed a novel array-based method to perform rapid, reproducible, and well-controlled secondary SEREX screens. Two-dimensional arrays containing up to 100 phage plaques were constructed by placing 1 □l drops of phage suspension in a grid-like pattern onto a lawn of E. coli on a rectangular agar plate. As with primary SEREX screens, plates were then overlaid with IPTG-impregnated nitrocellulose membranes and incubated at 37°C overnight. Under these conditions, the vast majority of phage give rise to a single plaque approximately 0.5 cm in diameter, irrespective of the exact titer of the original phage solution. Each membrane thus contains up to 100 individual phage plaques, each with a defined position on a grid. Negative and positive control phage are included for comparison with test phage. Multiple replicates of these arrays can be rapidly constructed using a multi-channel pipettor. Nitrocellulose

membranes containing such phage arrays were then immunoblotted with serum samples using the same methodology as for primary SEREX screens (above).

Fig. 2 shows a subset of the array results obtained from our ovarian cancer study. The figure shows two replicate arrays containing 42 phage plaques. A non-recombinant (i.e., empty) phage was plated at several positions to serve as a negative control. The leftmost array was immunoblotted with serum from an ovarian cancer patient, whereas the rightmost array was immunoblotted with serum from a normal control (a female over the age of 30 with no personal history of cancer). The arrowheads show 6 phage that were reactive with serum antibodies from the cancer patient, but not the normal control. Anywhere from 1-5% of phage from the primary screen show cancer-specific immunoreactivity such as this across the panel of case and control sera. The 6 phage clones in Fig. 2 were subjected to standard DNA sequencing. As summarized below, 4/6 were found to encode the tumor suppressor p53, and 2/6 encoded a novel zinc finger-containing protein called hZF5.

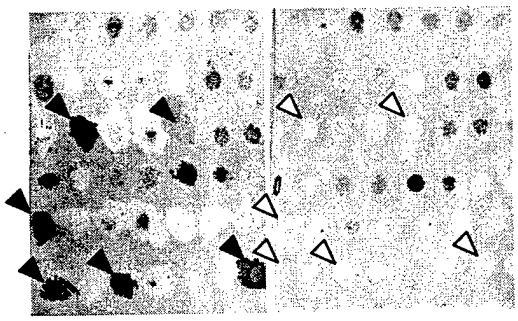


Fig 2. Secondary SEREX screening of ovarian cancer cDNA clones by phage array. The left and right panels show duplicate nitrocellulose membranes containing a 2-D array of recombinant phage clones that were identified in a primary SEREX screen of an ovarian tumor cDNA library. The left panel was immunoblotted with serum from a ovarian cancer patient (stage III , serous) whereas the right panel was immunoblotted with serum from a normal control. Membranes were then probed with a human IgG-specific, AP-conjugated secondary antibody and developed with NBT/BCIP. Immunoreactive phage plaques appear as dark circles, whereas non-reactive phage are clear. The arrows indicate 6 phage that showed a cancer-specific pattern of immunoreactivity with these and other serum samples.

To date, our SEREX screening efforts have identified 15 different gene products that appear to have a cancer-specific pattern of immunoreactivity (when tested in a SEREX array format using sera from 30 cancer cases and 20 normal controls, as described below). We expect additional antigens to be identified with continued screening. The 15 antigens are summarized in Table 2. Among these 15 antigens are the tumor suppressor p53 and the cancer-testes antigen NY-ESO-1, both of which are well-documented tumor antigens. None of the other 13 antigens has previously been shown to be immunogenic in cancer. Among these is a novel member of the MAGE superfamily of tumor antigens that we designate MAGE-E1, as well as a protein called Ubiquilin-1 that has homology to Ubiquitin but has no ascribed function. Intriguingly, one of the antigens we identified by SEREX (IFI27) was also identified by HDAH in Project 1 (see Table 7 in the Project 1 report) due to its overexpression at the mRNA level in ovarian cancer. This suggests that the immunogenicity of IFI27 might be attributable to overexpression by ovarian tumors, leading to broken peripheral tolerance to this self protein.

Current screening efforts are focused on patients who are negative for antibody responses to the 15 antigens identified so far. At this time, we do not know whether these patients are completely deficient in antibody responses to antigens represented in the library, or whether continued screening will reveal novel antigens to which they respond. Our new goal (which was not part of the original proposal) is to screen at least 5 x 10^5 phage clones with serum from each patient in this group. For most patients, only 5-10 x 10^4 clones have been screened so far, therefore our efforts are far from saturated. The longterm goal is to identify additional antigens that allow detection of this patient subset, so as to increase the overall sensitivity of the antigen panel for detecting ovarian cancer.

B. Prioritize the evaluation of novel ovarian tumor antigens. Statistical methods will be applied to identified antigens to determine if the discovery looks promising for translation into a test for use in the general public.

As described above, we are are using SEREX-based arrays for initial prioritization of ovarian tumor antigens discovered through SEREX. All 15 antigens discovered to date have been arrayed and exposed to serum from 30 ovarian cancer patients and 20 normal controls. As shown in Table 2, in this preliminary analysis, some patients showed a response to only 1/15 antigens, whereas others showed responses to several antigens. Likewise, some antigens were recognized by multiple patients, whereas others were recognized by only a single patient. At least 19/30 patients showed an antibody response to at least one antigen in the panel. Based on these results, we have prioritized NY-ESO-1, Ubiquilin-1 and IFI27 for follow-up ELISA studies with larger numbers of case and control sera.

Task 4: Perform ELISA screens for promising candidates
Months 18-24:

A. An ELISA based screen will be used to probe serum for the presence of antibodies against promising candidates that are identified by SEREX technology in Project 2.

We assembled a full-length cDNA clone encoding NY-ESO-1 and produced His-tagged recombinant protein in the mammalian cell line COS-7. The cDNA was inserted into the mammalian expression vector pcDNA4.1/HisMAX (Invitrogen), which fused six histidine residues to the N-terminus of the protein. The resulting plasmid was transiently transfected into COS7 cells using lipofectamine-Plus. Fig. 3 shows an anti-His Western blot of COS-7 lysates containing recombinant histagged NY-ESO-1 and, as controls, histagged p53 and LacZ.

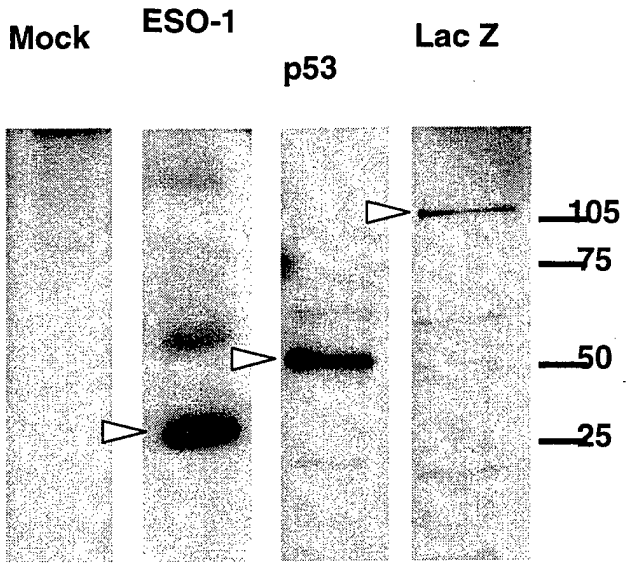
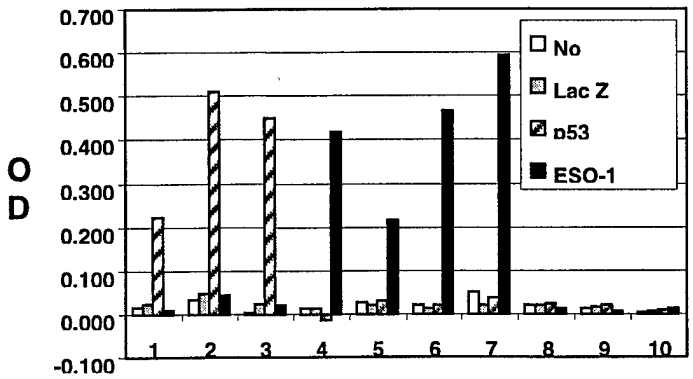


Fig 3. Western blot showing expressing of His-tagged recombinant tumor antigens in mammalian COS7 cells. Cells were transiently transfected with pcDNA3-based expression vectors encoding His-tagged ESO-1, p53 or Lac Z (as a control). Mock transfected cells served as a negative control. Nuclear extracts were prepared, subjected to SDS-PAGE and immunoblotted with a monoclonal antibody to the His tag (Sigma). Antibody detection was by enhanced chemiluminescence. Recombinant proteins are indicated by open arrowheads.

These COS7 cell lysates were used as antigen sources in ELISA to assess serum antibody responses to p53 and NY-ESO-1 in a subset of ovarian cancer patients. Briefly, 96-well nickel-coated ELISA plates (Clontech) were blocked with PBS/1% BSA, washed with PBS/0.5% Tween-20 and then be incubated with lysates (10^8 cells/20 ml PBS) from COS7 cells transfected with plasmids encoding NY-ESO-1, p53 or LacZ, or empty pcDNA4.1/HisMAX to serve as a negative control. After washing, plates were incubated with serum at 1:50 in PBS/1% BSA. After washing, plates were incubated with goat anti-human antibody conjugated to horseradish peroxidase (HRP). Plates were developed with TMB and read at 450 nm. As shown in Fig. 4, in this preliminary experiment, four patients showed antibody responses to NY-ESO-1, and three showed responses to p53. This ELISA method is presently being used by the Disis lab to assess the serum antibody response to NY-ESO-1 in larger groups of ovarian cancer patients and controls, as per Aim 1.

Fig 4. ELISA demonstrating serum antibody responses to p53 and ESO-1 in patients with ovarian cancer. Lysates from COS7 cells (see Fig. 3) expressing His-tagged p53, ESO-1 or, as a negative control, Lac Z were added to nickel-coated ELISA plates. After unbound proteins were washed away, serum from 10 ovarian cancer patients was added at 1:50 dilution, followed by HRP-conjugated goat anti-human IgG secondary antibody. Plates were developed with TMB and read at 450 nm. Patients #1-3 show a serum antibody response to p53, whereas patients #4-7 show a response to ESO-1. Patients #8-10 show no response to either protein.



In addition to NY- ESO-1, we have recently produced recombinant Histidine-tagged Ubiquilin-1. This required use of a prokaryotic expression system, as recombinant Ubiquilin-1 failed to express in COS-7 cells, presumably due to proteolytic degradation common to Ubiquitin class proteins. Once adequate ELISA conditions are established for assessing serum antibody responses to Ubiquilin-1, we will commence large-scale ELISA experiments as per Aim 1.

B. Full length cDNA will be obtained and an ELISA based screen will be developed for at least one of the most promising overexpressed genes discovered through HDAH in Project 1. This task will not be completed by month 24.

Studies are planned to assess the serum antibody response to HE4, which was discovered in Project 1. This work will commence when specific antiserum to HE4 becomes available for use in sandwich ELISA.

All blood samples collected through the ORCHID study have been analyzed for p53 antibodies using the capture ELISA format. In addition, a reference population of 175 volunteer blood donors has been analyzed. Results of the final analysis are described below (Task 5).

Task 5: Pool data for analysis
Months 6-24:

A. All discovery data will be combined with data from the other labs through the coordination core. The Project Director will summarize on a routine basis the results and provide them to the Coordination Core for further interpretation and incorporation.

Data from the ELISA tests for antibodies to p53 and H2N have been merged with core data on patient demographics, clinical characteristics and serum CA-125 levels. Though efforts are continuing to develop a better understanding of the potential for these markers to contribute to screening, some basic summaries of marker performance are available. Serum levels were obtained for several patient groups:

- Blood bank normals: males and females, aged 25-60, anonymous
- ORCHID normals: sera collected from women participating in this project with no evidence of ovarian pathology and no evidence of other cancer
- Benign: sera from ORCHID participants diagnosed benign ovarian pathologies and no evidence of other cancer
- LMP: sera from ORCHID participants diagnosed with tumors of low malignant potential or borderline tumors
- Ovarian Cancer: sera from ORCHID participants with confirmed cancer of the ovary
- Other cancers: sera from ORCHID participants with cancer at a site other than ovary
- The demographic characteristics of these groups are provided in the Core. Table 3 presents the basic distributions of these marker levels by group.

Table 3: Serum CA-125, and H2N and p53 antibody levels by patient group

	Outcome					
	Blood Bank Normals	ORCHID Normal	Benign Ovarian Pathology	LMP/ Borderline	Ovarian Cancer	Other Cancer
N	175	49	28	9	51	18
CA125 mean ±sd	NA	29±60	81±282	93±81	1243±31	31±46
median		13	26	45	297	17
(range)		(3,413)	(4,1517)	(9,203)	(0,14,842)	(8,214)
H2N mean ±sd	.15 ± .49	.47± 1.28	.28 ± .61	.40 ± .51	.85± 1.69	.31 ± .96
median	0	0	0	0	0	0
(range)	(0, 3.74)	(0, 7.26)	(0, 2.25)	(0, 1.27)	(0, 9.19)	(0, 4.01)
p53 mean ±sd	.11 ± .22	.87± 2.02	.22 ± .70	.21±.28	4.02 ± 2.21	.30 ± .54
median	.02	.18	0	0	.12	0
(range)	(0, 1.7)	(0,12.6)	(0, 3.69)	(0, .65)	(0, 56.6)	(0, 1.80)

Mean serum antibody levels for both p53 and H2N are highly elevated in serum from ovarian cancer patients compared to serum levels in normal individuals. In multivariate logistic regression models shown in Table 4, p53 and H2N were both significant predictors of ovarian cancer (as opposed to normals) both individually and combined (Models 1-3). After controlling for age and CA125 levels (Model 4), neither added significantly to the model. ROC curves estimated from these logistic regression models are included in Appendix J to demonstrate simulataneously the true and false positive rates associated with the logistic regression approach to classification with these markers. We note that the area under the curve (AUC) of the ROC graph using only age and log (CA-125) is 0.86 and that by including p53 and H2N (log scale), the AUC increases to 0.89.

Table 4: Logistic regression models of serum markers to predict ovarian cancer

	log(p53+1)		**log (H2N+1)**	
	Odds ratio	p-value	Odds ratio	p-value
Model 1	3.32	0.000		
Model 2			3.62	0.000
Model 3	2.58	0.002	2.44	0.016
Model 4	1.24	0.565	1.68	0.404

As depicted in Figure 1, Appendix J, the use of a convenient sample of normals raises some concern in identifying the true normal range for a general population. The fact that CA125 levels are elevated in the ORCHID normals, women referred to a gynecologic specialty practice, lends credence to our suspicions that these women may not be representative of the general population. The levels determined in sera from blood bank donors may be closer to a screening population. However, these specimens are not well-characterized and they are

known to come from a broader age range and both sexes. Because these samples are anonymous, we cannot investigate the effect of these factors.

Model 4 is based on only ORCHID samples since age and CA-125 levels are not currently available for the blood bank normals. However, we have evidence that the ORCHID normals are not like the blood-bank normals with respect to CA125 levels. We have therefore begun the process of obtaining serum levels from two other well characterized populations with existing stored specimens that are thought to be more representative of the population of interest. The use of additional normal control groups that are well characterized will help us better establish reasonable cutpoints for positivity.

CONCLUSIONS

In addition to the known tumor antigens p53 and HER2/neu that were evaluated in Aim 1, a large number of novel candidate antigens that are immunogenic in ovarian cancer have been identified by SEREX. We are now poised to evaluate serum antibody responses to the combined panel of antigens using large numbers of sera from patients with malignant and benign ovarian disease and normal controls. Moreover, continued SEREX screening is expected to provide additional antigens that may increase the overall sensitivity of the panel. In addition to their potential utility for early detection of ovarian cancer, at least one of these antigens (NY-ESO-1) shows promise as a target for immunotherapy, therefore we have also launched efforts toward this goal. Related studies of breast and colorectal cancer have received funding and been initiated as a result of this work on ovarian cancer

Statistical, Clinical and Laboratory Coordinating Core

Nicole Urban, ScD, Garnet Anderson, Ph.D., Nancy Kiviat, MD, Leona Holmberg, MD, Jane Kuypers, Ph.D., Charles Drescher, MD, Mary Anne Rossing, Ph.D.

INTRODUCTION

The purpose of this shared resource is to support the work of the project investigators by collecting, storing and providing tissue and blood for analyses, as well as statistical analysis of project results. The specific aims of the Statistical, Clinical and Laboratory Coordination Core are:

- To develop a resource of well-characterized women with associated data, blood and tissue specimens that will be used to develop and test new markers for disease as specified in Projects 1 and 2.
- To characterize the blood and tissue from these women with respect to CA-125 levels; expression of p53, HER2/neu, and Myc; and histology.
- To provide statistical design and analysis support for Projects 1 and 2.
- To describe the joint behavior of novel and previously established markers, and to investigate the relationship of these markers to other clinical factors (e.g., stage at diagnosis, history of cancer, sonography findings) and demographic or epidemiologic data (age, menopausal status, number of ovulatory cycles).

The development of this resource of data and specimens has been essential to all phases of the work conducted by the Projects. The Core has assured that data and specimens are collected in a standardized manner, with proper attention paid to the collection protocols to assure the quality of the specimens for the proposed assays. Standardization of specimen collection ensures that all of the specimens analyzed in Projects 1 and 2 are sufficiently and uniformly characterized to allow for the reliable and valid interpretation of project results. In addition, by overlaying the statistical design for the assays of both projects and obtaining the data from the Core laboratory on the same population, we are able study the inter-relationship of both newly identified and established markers. The tasks and status of each task proposed in the original statement of work is included as Appendix A.

Identification and Recruitment of Participants

All patient recruitment for this study occurred at the office of Pacific Gynecology Specialist (PGS), which is located on the campus of Swedish Medical Center (SMC) in Seattle. Approximately 60% of the PGS oncology practice occurs on the campus of SHMC, while the remainder is divided among seven community hospitals in the Puget Sound area. The staff for this study are housed at the Marsha Rivkin Center for Ovarian Cancer Research, which

provides close proximity to and interaction with the clinical community and patient recruitment being conducted for the study.

All potential participants were invited to participate in the study by the attending physician or study nurse at the time of the pre-operative office visit. Interested patients were provided for their review a brochure describing the study, known by its acronym ORCHID – Ovarian Research Collaboration Helping to Improve Detection. The brochure is included as Appendix F. If the patient agreed to participate in the study, the attending physician, nurse or a research study staff member reviewed the consent form and other required enrollment documents with the patient.

Completed enrollment forms were returned to the Marsha Rivkin Center, at which time data from enrollment forms was entered into the study database, with each participant assigned a unique participant number (UPN).

An additional group of women undergoing surgery at other hospitals where tissue collection protocols had not yet been developed were recruited to provide blood during the pre-operative office visit. Collection of blood from this group of women began in 1999 and was not as successful as the tissue donation component of the study. It was originally anticipated that approximately 120 women per year would be recruited for serum only collections. Enrollment in this protocol has been slow, largely due to the lack of time the participant may have available to provide blood during the clinic visit. To date 33 women have been enrolled in this component of the study.

Data Collection, Warehousing and Storage

Development and use of Data Collection Instruments

For this study, data collection instruments were created to ensure that information needed for scientific research as well as that required for human subjects participation was obtained. The data collection instruments utilized by the Coordinating Core can be classified into three basic categories: Enrollment, Specimen Collection and Clinical Data Requirements.

At the time of enrollment, the participant completes a combined enrollment/medical records release form, and an informed consent form and a self-administered 20-minute questionnaire. The self-administered questionnaire to be completed by all study participants was developed by the study epidemiologist, Mary Anne Rossing, Ph.D. The data collected focus on known or suspected epidemiological risk factors for ovarian cancer, including: menstrual and reproductive history; use of exogenous hormones (oral contraceptives and hormone replacement therapy); family and personal history of ovarian, breast and other cancers; sociodemographic factors; history of selected gynecologic surgeries including hysterectomy and tubal ligation; and other relevant medical conditions. The data collected will be sufficient to categorize women according to their menopausal status, estimated total years of ovulation, and prior personal and family history of cancer. Data will also allow an examination of the possible relation of various medical, hormonal, and reproductive factors with levels of various

cancer screening markers of interest. Of the 346 participants enrolled in this study, 303 (87%) fully completed the study questionnaire.

Data collection instruments were also created for all specimen collection, requisition, and specimen. These forms captured data entered into the specimen tracking system, including specimen type, site, processing method and location. A unique 6-digit number was used to label all specimens, with a duplicate of the label attached to the specimen collection form, blood processing form and specimen tracking form. These forms are labeled as Appendix G in the Appendices included with this report. The unique label number for each specimen is linked to each participant's UPN in the specimen inventory database.

Extensive characterization is conducted for all specimens collected for this study. Clinical data collection forms were created to classify the specimens stored in the study repository. The study automatically receives a pathology report and operative notes on patients who provide specimens to this study. Dr. Charles Drescher, a co-investigator on this study conducts the first level of review by assigning a clinical diagnosis to the patient based on the pathology report and operative notes. A second level review is conducted by Dr. Nancy Kiviat at the Core facility whereby all formalin-fixed, paraffin-embedded specimens which correspond directly with the fresh frozen tissues are examined for histology.

In addition to histology classification, the tissue is examined utilizing Immunohistochemistry techniques which is described in detail later in this report. A chart review is also conducted where relevant clinical data is abstracted and entered into the database system. The Data Coordinator reviews the chart to abstract additional data regarding probable diagnosis prior to surgery; type and date of diagnostic tests performed and test results. All of the data captured above is entered into the tracking system and associated at the participant or specimen level. Examples of these forms are included as Appendix H in the Appendices.

Development of the Data Management and Tracking Systems

As part of this study, we have developed a relational database management system programmed in Visual FoxPro and Visual Basic to support all the tracking functions and store all the key scientific data associated with this project. This information system consists of two key components: 1) a participant enrollment database, and 2) a specimen inventory and data tracking system. Enrollment data, including personal contact information, is stored in the participant enrollment database. At the time of enrollment entry, the database generates a unique participant number (UPN) that is used in all subsequent correspondence regarding the study participant, her blood and tissue specimens, and the data associated with these specimens.

The specimen tracking system is a multi-functional application that tracks all specimens collected for the QUEST study as well as for ORCHID and the recently funded ovarian SPORE program. Each individual specimen container or vial is labeled with a unique 6-digit number at the time of collection; these numbers are in turn associated with the UPN of the donating participant and the date of collection at the time of data entry.

Immunohistochemistry data and histology characterizations are also stored in this component. The specimen tracking system also allows for the entry of clinical data relevant to the distribution of tissue and blood specimens, including surgical pathology diagnoses, pre-operative CA 125 results, and other relevant clinical conditions. This system also serves as an inventory database allowing us to track the location of specimens in the freezers and those that have been sent to project investigators.

Both components of the data management system reside on password-protected servers managed by the IS staff of the Cancer Prevention Research Program at Fred Hutchinson Cancer Research Center. In addition to the logging onto the network, a staff member entering the ORCHID system must supply a unique password to run the application.

Collection of epidemiologic and clinical data

During active participant recruitment, epidemiologic and clinical data collection was ongoing in the ORCHID study. Each week, study staff generated a report showing successful collections without other required data (questionnaires, core histologic review, clinical data records etc). For self-administered forms, the project coordinator follows procedures outlined in the Follow Up Protocol described previously. For clinical information, a clinical data follow up reported is created by the study database. A regular chart review of recently enrolled participants was conducted by study staff to obtain detailed information on final diagnosis. A copy of the final pathology report is automatically obtained from Dynacare Laboratory of Pathology and included in the participant's study file. Data from the Core Laboratory is entered at the laboratory and submitted bi-weekly in electronic format for inclusion in the database.

As stated previously, data collected from the questionnaire will be sufficient to categorize women according to their menopausal status, estimated total years of ovulation, and prior personal and family history of cancer. Data will also allow an examination of the possible relation of various medical, hormonal, and reproductive factors with levels of various cancer screening markers of interest. As of October 2000, 303 (87% of total enrollment) questionnaires have been completed and returned to the study office.

As described previously, clinical follow-up data is collected via review of the participant's medical records using standardized forms developed by the Core investigators, and entered into the clinical database by study staff. This data includes selected information on disease characteristics including diagnostic test results, histology, stage, grade, tumor distribution, extent of residual disease and any other standardized data as determined by Core investigators. In addition, the recently funded ovarian SPORE has been the catalyst for implementing quarterly follow-up of participants regarding chemotherapy administered, response to treatment, disease status and survival.

Follow up process for data

With adherence to stringent enrollment procedures, very little participant follow-up for this study was anticipated, however on occasion follow-up was required if a patient had not fully completed enrollment forms or had not returned the study questionnaire. A protocol to address such circumstances was developed. In these situations, a written request, followed by one telephone call, is made by the Study Coordinator. The database system generated a report detailing a list of participants enrolled for at least 30 days, and for which there may be one or more pieces of enrollment information missing. A letter was mailed to the participant if their enrollment materials are incomplete or if the Core has not received their questionnaire within thirty days of their enrollment. After fourteen days, a follow up call is made to the participant if she has not responded to the request.

Collection and Preparation of Tissue and Blood Specimens

Specimen Collection

Recruitment and specimen collection for the ORCHID study is now complete. As of October 19, 2000, 346 women have consented to this study and successful ovarian tissue collections have occurred on 217, of which 58 patients were diagnosed with ovarian cancer, 11 with tumors of Low Malignant Potential, 42 with benign disease, and 92 with no ovarian abnormalities. (Investigators are awaiting pathology reports on 2 collections and 27 collections were conducted with non-ovarian primaries or with no ovaries collected. The cancer cases include 12 patients with early stage disease, of which three are patients diagnosed with early-stage serous tumors.

A dedicated Tissue Collection Specialist is on hand to collect fresh and frozen tissue samples in addition to formalin-fixed and paraffin-embedded specimens. A Specimen, Collection, Processing and Storage protocol detailing collection techniques is included as Appendix Item I. Additionally, up to 50 cc of blood is collected and processed into sera, plasma, and white blood cell pellets. The variety of collection techniques allows the Core to meet the needs of the projects within the program, as well as maximizes use of the same tumors.

QUEST Study Bloods

A portion of the blood specimens to be used as positive controls in the ORCHID study are to be obtained from women enrolled in the QUEST study. A total of 586 women have been randomized to this study, of whom 292 are assigned to the ovarian cancer screening intervention arm. Women in this arm are being consented for additional blood to be drawn for research purposes, including this study. To date, the QUEST study has drawn baseline bloods on approximately 284 women, all of whom will also receive a blood draw in the 2nd year of their participation. The QUEST bloods are inventoried and labeled in the same manner as bloods collected for the ORCHID study, and are stored in the same repository as the ORCHID bloods.

Specimen Allocation Procedures

After characterization in the Laboratory Core, specimens are made available to Project Investigators. After Project needs have been met, specimens may be made available to non-Project Investigators. In such circumstances, the non-Project Investigators will be required to complete a review process for use of said specimens. All specimens transferred to non-Project Investigators must receive approval and/or certification from Study Investigators, and the FHCRC Institutional Review Office (IRO). Specimens provided to commercial entities, or Investigators in collaboration with a commercial entities must also receive approval from the FHCRC Human Specimens Committee.

Specimen Transfer

For all specimen transfers, a report identifying those specimens to be distributed is generated in the specimen inventory database. The investigator's name, laboratory location, and intended use is recorded in the database with the specimens (individually identified) to be sent to the research project. The Tissue Collection Specialist receives a copy of this delivery report, removes the specimens from the repository, and packages the specimens securely for transport to the investigator's laboratory. To ensure the integrity of the specimens, the freezer boxes will not be removed from the freezer for processing until all transport supplies are available for performing the transport procedure. The specimens are packaged in a styrofoam box according to study protocol.

Upon receipt of the delivery, the investigator and Tissue Collection Specialist will review the contents of the delivery and check them against the printed report. Both will sign a transmittal form confirming that the specimens listed were received in full and in satisfactory condition. The completion of this form and confirmation of delivery will be stored in the specimen database and linked to the records of the specimens comprising the delivery. An example of this form is included with the specimen collection and tracking forms in Appendix G.

Histologic Characterization of Tissue and Blood Specimens

The Laboratory Core component of this study conducts a detailed review of all tissue specimens collected during surgery. These reviews allow Investigators to rapidly identify appropriate cases for the projects and perform quality control of the tissue collection and processing.

Dr. Nancy Kiviat is responsible for conducting a pathology review of each tissue specimen collected during surgery for this study. The results of this characterization are coded and associated with each individual tissue specimen in the specimen inventory database.

During months 1 through 24, histological examination was carried out on 209 collections with tissue and classified according to the World Health Organization (WHO) classification of ovarian tumors. Cases identified as tumors (benign or malignant) or other epithelial lesions were further characterized by immunohistochemistry.

In addition, Dr. Irena King of the FHCRC evaluated the analytical performance of CA 125. The Centocor CA 125 II IRMA assay kit was used, which is a one-step heterologous double-determinant solid-phase procedure utilizing the M11 mouse monoclonal antibody as a capture antibody to binds molecules containing OC 125-reactive determinants. These determinants are quantified using radioiodinated OC 125 antibody as tracer. The analytical performance of the CA 125 II IRMA was assessed by evaluating the linearity of the standard curve, the within-run (intra-assay) precision, the between-run (inter-assay) precision, both less than 10% CV. For the validation of accuracy we subscribe to the College of American Pathologists (CAP) Tumor Marker Surveys. Compared to the results provided by CAP, we ranged from +0.8 SDI to –1.3 SDI for all tests for the expected mean values that ranged between 30U/L and 112U/L. With each batch of samples we ran two-level kit controls and a two-level purchased reference standards to monitor the assay stability. Additionally, we are developing lysophosphatidic acid (LPA) procedures which require lipid extraction, thin layer chromatography (TLC) and gas chromatography (GC) separation and quantification to be used on plasma specimens.

Immunohistochemistry & Mutation Analyses

The Core laboratory supporting this study conducts assays for the oncoproteins cerbB-2 and p53 from each malignant tissue and a fraction of all normal tissues. In addition, p53 DNA is isolated and analyzed for mutation in the p53 gene. The results generated by the Core laboratory are compiled and reported monthly to the study investigators. In addition to keeping investigators abreast of ongoing laboratory activity, this report serves as a quality control measure that would reveal problems with screening assays or methods of tissue collection and processing.

Two hundred nine cases were characterized histologically according to the WHO classification of ovarian tumors. Of these, one hundred fifty tissues were characterized by immunohistochemistry. A panel of three antibodies was run on each case. Tissue reactivity was assessed using a monoclonal antibody directed against cytokeratin 8 (Becton Dickenson). To identify p53 overexpression, a monoclonal antibody which reacts with both the wild and mutant form of p53 was used (DAKO Corporation). Tumors that overexpressed the cerbB-2 oncogene product were identified using a polyclonal antibody (DAKO Corporation). The cases that were previously labeled as indeterminate were scored according to a new scoring system developed by Dr. Allen Gown for breast carcinomas. This method gives good correlation between the cerbB-2 results obtained by immunohistochemistry and fluorescent in situ hybridization (FISH) detection of multiple gene copies. The staining intensity is graded on a 0 to 4+ scale. If normal internal or external controls are present, a subtracted score is obtained by subtracting the staining intensity of the normal control from that of the tumor. A tumor would be considered cerbB-2 positive if the subtracted score is greater than or equal to 2 or if the staining intensity of the majority of tumor cells are 3+ or greater. In all situations the staining pattern must be membranous and not cytoplasmic.

Forty cases consisting of normal, benign, and malignant tissues were characterized by a polyclonal antibody directed against the mutant form of EGFR (EGFRvIII). Problems were encountered with non-specific and high background staining with this antibody. It was decided to put this project on hold until a more specific antibody is developed.

The results of the histological and immunohistochemical characterization of the tumors, benign lesions and normal tissues are shown below:

Normal	# of Cases	**P53 +**	p53 -	cerbB-2 +	cerbB-2 -
Normal Appendix	1	0	1	0	1
Normal Cervix	6	0	2	0	2
Normal Colon	1	-	-	-	-
Normal Fallopian Tube	24	0	13	0	13
Normal Myometrium	9	0	1	0	1
Normal Ovarian Tissue	100	0	16	0	16
Normal Uterus	6	0	1	0	1
Corpus Luteum	2	-	-	-	-
Functional Cyst	15	-	-	-	-
Ovarian Fibroma	4	-	-	-	-

Benign Lesions	# of Cases	p53 +	p53 -	cerbB-2 +	cerbB-2 -
Benign Cyst, Not Paraovarian	8	0	8	0	8
Benign Cyst, Paraovarian	1	0	1	0	1
Endometriosis/ Endometriotic Cyst	8	0	4	0	4
Inflammatory Lesions	1	-	-	-	-

Neoplastic Other	# of Cases	p53 +	p53 -	cerbB-2 +	cerbB-2 -
Benign Brenner Tumor, Typical	1	0	1	0	1
Benign Dermoid Cyst	2	0	0	0	0
Thecoma	2	-	-	-	-

Serous Tumors, Benign	# of Cases	p53 +	p53 -	cerbB-2 +	CerbB-2 -
Serous Adenofibroma	2	0	2	0	2
Serous Cysadenofibroma	3	0	3	0	3
Serous Cystadenoma	6	0	6	0	6

Serous Tumors, Low Malignant Potential	# of Cases	p53 +	p53 -	cerbB-2 +	CerbB-2 -
Serous Carcinoma of LMP	6	0	6	0	6

Serous Tumors, Malignant	# of Cases	p53 +	p53 -	cerB-2 +	CerbB-2 -
Serous Carcinoma	42	27	15	7	35

Mucinous Tumors, Benign	# of Cases	p53 +	p53 -	cerbB-2 +	CerbB-2 -
Mucinous Cystadenoma	5	0	5	0	5
Mucinous Cystadenofibroma	2	0	2	0	2

Mucinous Tumors, Low Malignant Potential	# of Cases	p53 +	p53 -	cerbB-2 +	**CERBB-2 -**
Mucinous Carcinoma of LMP	2	0	2	0	2

Mucinous Tumors, Malignant	# of Cases	p53 +	p53 -	cerbB-2 +	CerbB-2 -
Mucinous Carcinoma	3	2	1	2	1

Endometrioid Tumors,	# of	p53 +	p53 -	cerbB-2 +	CerbB-2 -

Malignant	Cases				
endometrioid Carcinoma	4	2	2	1	3

Clear Cell Carcinoma, Malignant	# of Cases	p53 +	p53 -	cerbB-2 +	CerbB-2 -
Clear Cell Carcinoma	4	1	3	3	1

Neoplastic Other, Malignant	# of Cases	p53 +	p53 -	cerbB-2 +	CerbB-2 -
Adenocarcinoma, NOS	13	11	2	6	7
Unclassified Epithelial Tumor	5	3	2	1	4
Colonic Carcinoma	1	1	0	0	1
Cecal Adenocarcinoma	1	0	1	0	1
Endometrial Carcinoma	2	1	1	1	1

In the group of cases where primary and metastatic tissues were collected and characterized by immunohistochemistry, there were no differences in the p53 and cerbB-2 results between the primary and metastatic tumors (data not shown). None of the normal or benign lesions displayed overexpression of either p53 or cerbB-2.

Provision of Specimens to Projects

During the first year, Project 1 was provided with a total of 15 specimens were provided to Project 1 for the construction of cDNA libraries. And 31 specimens were provided to Project 1 to hybridize with the first-generation membranes. This task has been completed and is described in greater detail in the Project 1 section of this report.

During Year two, Project 1 was supplied with a total of 54 specimens to hybridize with second-generation membranes. Originally, 105 specimens were to be provided to this project for second-generation hybridization. This number was revised to 75, and to date 54 specimens have been provided by the Core for this task.

During the first year, Project 2 was provided with provided with tissue specimens from two ORCHID study participants and 8 early stage specimens were obtained from the Gynecologic Oncology Group (GOG). In the second year, Project 2 was provided with 79 serum specimens of (benign and cancer of serous histology, and normal) from the study repository as well as obtained from the GOG. As serous tissue specimen inventory levels were not adequate at the time of allocation and additional 19 late stage snap frozen serous tissue specimens were obtained from the GOG for use in the SEREX analyses.

During Year 2, 179 specimens were provided for Project 2 to Dr. Mary L. Disis to assess antibody response.

Statistical Analyses and Design

Developing statistical algorithms for selecting over-expressed genes for subsequent efforts:

The underlying philosophy for analyzing the quantity of gene expression data on a relatively small sample was to provide a statistical filter that would reduce the number of candidate genes that were under investigation at one stage to a manageable level for the next more intensive level of investigation.

Our primary method ranks the gene candidates based on a modified t-statistic comparing the distribution of expression levels in ovarian cancer tissues to normal ovaries. This approach gives highest priority to genes with an average expression level in ovarian cancer tissues that are highly elevated over the average expression level in normal ovaries, as measured on a scale determined by the variability of expression levels. Initial analyses indicated that for a large proportion of genes of interest, the expression levels in normal ovaries are quite homogeneous. In ovarian cancer tissues, however, the expression values vary greatly. Using a pooled estimated of the variance in the usual t-statistic caused those genes with the greatest overdispersion in the cancers to receive a lower priority ranking, even when the discriminatory power should be very strong based on a visual examination of the distribution. To remedy this, we adopted a modified t-statistic where the estimate of the variance was based on the observed variability in the normal tissues alone. Though this statistic would not be considered

"efficient" in the traditional sense because it does not use all of the information available, it provides a more relevant measure of difference in gene expresion for our purposes. The rankings from the RT-PCR studies of 78 genes using this approach is provided in Appendix J). A separate ranking was also performed based on the comparison to benign tissues but as the number of benign tissues analyzed to date is currently rather small, these values have not been formally incorporated into the selection process.

Statistical analysis of antibody response to Her-2/neu, p53 and c-myc

As described in Project 2, initial analyses of antibody responses to Her-2/neu and p53 levels were evaluated in two sets of normals, women with benign disease, borderline tumors, ovarian cancer and other cancers. Both Her-2/neu and p53 are shown to be useful in discriminating ovarian cancers from normals in logistic regression analysis. After controlling for age and CA125 levels, however, Her-2/neu and p53 did not contribute significantly to the classification. Further efforts to determine a more representative normal control group are underway and will lead to more definitive analyses along the lines presented in Project 2.

Statistical analysis of Her-2/neu, p53 and c-myc expression in tissue:

This task is ongoing based on the data presented above (see *Immunohistochemistry & Mutation Analyses).*

Statistical analyses of select clones with clinical, epidemiologic and other laboratory data:

We have begun to pool all of the clinical epidemiologic and key laboratory data into an analytic data file. Appendix J presents a summary of the key factors by patient group: Noraml, Benign, LMP (borderline) tumors, Ovarian Cancer, and Other Cancers. These groups differ in age, menopausal status and other factors that may be potentially related to biomarker levels. In our basic modelling to identify which marker or panel of markers best discriminate between ovarian cancers and normals, or ovarian cancers and benign disease, we will be cognizant of these differences and incorporate these factors into the models. As mentionned in Project 2, the analysis of Her-2/neu and p53 in conjunction with CA-125 levels suggested that these antibody responses provided only very modest improvent in the accuracy of a screen based on CA-125 and this did not reach statistical significance, despite the fact that the serum levels of these markers are not correlated.

We have identified several aspects that require further examination. In particular, we will obtain one or two other normal control groups from well-characterized cohorts having stored specimens available to us. This will give us an estimate of the distribution of these markers in women more representative of the general population that would be targeted for screening. Second, we will be conducting further analyse the incorporate factors such as tissue expression levels and other clinical features of disease to determine whether these antibody responses are related to specific subtypes of disease.

CONCLUSIONS

The development of a specimen respository with corresponding clinical and epidemiologic data required a substantial effort. This repository, from well characterized patients and with high quality, centralized pathology, and centrally determined laboratory measures provides a valuable resource for stimulating research on many aspects of ovarian cancer. Though initially targeted to early detection, this resource will be useful for testing hypotheses related to disease classification, prognoses, and response to therapy.

Our intial analyses of data from this project has revealed some of the challenges in the design and analyses for these types of biomarker development studies. We more clearly recognize the value of a well-characterized normal control groups that are representative of the population to be screened. We also note the need for further thinking on the appropriate normal controls for gene expression studies (e.g., normal contralateral ovaries, or normal ovaries removed for non-cancer indications), where truly tissue from individuals without any known pathology is very uncommon.

For analyses of gene expression data, statistical methods are still in the developmental stage. There are many groups around the world who are working on the different levels of this problem (e.g., sources of error, spot-finding, normalization). Dr. Schummer has collaborated with many of these to share his data and learn the results of their methods on these data. For purposes of early detection, as is our primary mission, we have assumed that a strong signal to noise ratio is needed in the gene expression level in tissue in order for this signal to be recognizable in a serum based assay. Under this assumption, we have used simple univariate approaches with raw data to identify and rank novel genes that can be investigated further. The success of this approach awaits the outcome of the next phase of this research.

KEY RESEARCH ACCOMPLISHMENTS

Project One

- Construction of three unamplified and non-normalized cDNA libraries from normal ovaries, late stage ovarian carcinomas and metastatic ovarian carcinomas
- Generation of a cDNA membrane array consisting of 97,803 cDNA clones randomly selected from these libraries
- Interrogation of this array with probes from 30 tissues (normal and ovarian cancers) finding 17 genes (2 novel genes, 5 ESTs and 10 known genes) with marker potential
- Generation of a cDNA glass array consisting of 1390 genes selected from the membrane array with potential to code for marker genes
- Interrogation of this array with probes from 64 tissues (normal and ovarian cancer) finding 126 genes (8 novel genes, 30 ESTs and 88 known genes) with marker potential
- Expression validation of 78 genes by RealTime quantitative PCR, finding 15 marker genes
- ELISA test of SLPI on ovarian cancer patient sera reveals no elevated expression of SLPI protein in patient sera
- RT-PCR of SLPI and HE4 finds presence of transcript in epithelial cells from peritoneal washes of ovarian cancer patients but also of patients suffering from other malignancies and benign diseases.
- ELISA test of Mesothelin on ovarian cancer patient sera is in development

Project Two

Task 1

- Fully operational and reproducible assay for detection of HER2 antibodies for use in final analysis.
- Fully operational assay for the detection of p53 antibodies.
- Construction of peptides which are dominant B cell epitopes of p53 and c-myc.
- Fully operational and reproducible assay for detection of HER2 antibodies using recombinant proteins.
- Development of conditions to detect peptide specific antibody responses by ELISA.
- Completed analysis of all ovarian cancer sera collected through the ORCHID study for HER2 and p53 antibodies (ug/ml) as well as analysis of control reference population (n=175) for HER2 and p53 antibodies.

Task 2

- Successful optimization of the signal-to-noise ratio for the SEREX protocol.
- Construction of a serous ovarian tumor cDNA library.
- Development of an array based procedure that allows rapid evaluation of multiple phage clones with multiple serum samples.

Task 3

- Validation of the cDNA library and SEREX immuno-screening procedure by cloning the known ovarian tumor antigens p53 and NY-ESO-1.
- Successful use of the SEREX method to identify 15 candidate ovarian tumor antigens.
- Successful use of SEREX arrays to prioritize the 15 identified antigens on the basis of their immunogenicity across a panel of serum samples from 30 ovarian cancer patients and 20 normal controls.

Task 4

- Development of a reproducible ELISA protocol to assess serum antibody responses to NY-ESO-1 in ovarian cancer patients.
- Production of Histidine-tagged recombinant Ubiquilin for use in ELISA.

Task 5

- A pooled dataset containing participant demographics, clinical characteristics, and selected laboratory values
- Preliminary analyses of the discriminatory power of antibody levels to p53 and H2N for distinguishing ovarian cancers from normal individuals both individually, jointly, and in combination with CA125 levels.

Core

- Development of a recruitment protocol and supporting documents
- Development of a tissue and serum repository containing tissue from 217 women
- Development of a study database that links epidemiological, clinical and laboratory data collected on all women enrolling in this projecct
- Provision of specimens to Projects 1 & 2
- Prioritization of gene expression from RT-PCR data for subsequent develop into protein based serum assays
- Prelimary analyses of antibody responses to p53 and H2N as markers for classifcation purposes

REPORTABLE OUTCOMES

Project One

Publications:

Schummer M, Ng WL, Bumgarner RE, Nelson PS, Schummer B, Hassell L, Rae Baldwin L, Karlan BY, and Hood L (1999) Comparative hybridization of an array of 21,500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas, *Gene* **238**, 375-385

Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M and Yakhini, Z (2000) Tissue Classification with Gene Expression Profiles. The Forth Annual International Conference on Computational Molecular Biology -- RECOMB'2000, pp 54-64

Schummer M, Kiviat N, Bednarski D, Crumb GK, Ben-Dor A, Drescher C and Hood L (2000) Hybridisation of an array of 100,000 cDNAs with 32 tissues finds potential ovarian cancer marker genes, *Int. J. Biol. Markers,* **15 suppl. 1,** 35

Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M and Yakhini, Z (2000) Tissue Classification with Gene Expression Profiles. Journal of Computational Biology, submitted

Databases:

- orchidDB (FileMaker based database holding the 100,000 membrane array clones and their hybridization signals across 57 conditions equaling 40 tissues)
- orchidGlassDB (FileMaker database holding the 1380 glass array genes and their hybridization signals across 64 tissues)

Funding:

- SPORE grant (for the identification of marker genes for chemoresistance in ovarian cancer)
- One pilot grant to the SPORE (for the generation of antibodies)

Project Two

Publications

Stone, B., Schummer, M., Paley, P.J., Crawford, M., Ford, M., and Nelson, B.H. 2000. MAGE-E1, a novel ubiquitously expressed member of the MAGE superfamily identified by SEREX immunoscreening. Submitted.

Presentations:

"Mapping the Immune Response to Ovarian Cancer for Screening and Therapy"

Seminar, Pacific Ovarian Cancer Research Consortium, Fred Hutchinson Cancer Research Center, Seattle WA May 2000 (Brad Nelson, Ph.D.)

Joint meeting, British Columbia Cancer Agency/University of British Columbia/University of Victoria, Victoria BC, June 2000. (Brad Nelson, Ph.D.)

Seminar, Pacific Northwest Research Institute, Seattle WA, September 2000 (Brad Stone, Ph.D.)

Annual Meeting, Society for Biological Therapy, Seattle WA, October 2000 (Brad Nelson, Ph.D.)

Active Funding:

NIH
1 RO1 CA82724-01
B.H. Nelson, P.I.
"Novel Vaccine Targets for Early-Stage Breast Cancer"

03/01/00-02/28/04
Direct total costs: $1,442,014
Direct annual costs: $339,580

The specific aims are:
1. To identify immunogenic proteins in early-stage breast cancer;
2. To select antigens for vaccine development on the basis of humoral and cellular immunogenicity in women with early-stage breast cancer.

NIH
1 R21 CA84359
Brad Nelson, P.I.
"Immunologic Screening for Early-Stage Colorectal Cancer"

8/1/00-7/31/02
Direct total costs: $150,000
Direct annual costs: $75,000

The specific aims are:
1. To classify 20 early-stage colorectal cancer patients as positive or negative with respect to serum antibody responses to a panel of known colorectal tumor antigens.
2. To determine whether patients who lack antibody responses to known tumor antigens instead respond to an undiscovered set of tumor antigens.

Morrison Trust
Brad Nelson, P.I.

1/1/00-12/31/00
Direct total costs: $40,000
Direct annual costs: $40,000

"A Novel Immunologic Blood Test for the Early Detection of Colorectal Cancer"

The specific aims are:
1. To identify a set of tumor proteins that commonly induce an antibody response in patients with early-stage colorectal cancer.
2. To determine the best combination of SEREX-defined tumor antigens to use for the detection of early-stage colorectal cancer.

Core

Development of an ovarian specimen repository housing over 3000 individually identified specimens.

Development of a participant database and specimen inventory tracking system.

Funding of the 1999 ovarian cancer Specialized Program of Research Excellence by the NCI

To continue enrolling women into this research program, to continue building the specimen repository (Charles Drescher, Clinical Core PI)

To develop further statistical methods for using multiple markers for early detection (Martin McIntosh, SPORE Project 4 PI)

To test a panel of markers in an nested case-control design in an existing cohort of post-menopausal women (Garnet Anderson, Project 3 PI)

CONCLUSIONS

We have identified a large number of genes that are over-expressed in ovarian cancer tissue relative to the ovarian tissue obtained from women without cancer or ovarian pathology. In addition we have identified several oncogenic proteins that elicit antibodies detectable in the blood of some ovarian cancer patients. These discoveries are providing the foundation for ongoing work in early detection of ovarian cancer, funded by the NCI as part of a SPORE in ovarian cancer. Specifically, we are developing algorithms for using a panel of markers for ovarian cancer that tailors the use of the markers to the individual woman be accounting for change over time in each of the markers. We are in the process of evaluating the genes and gene products we have found for their likely contribution to the marker panel.

Our discoveries are expected to lead as well to work on the molecular characterization of ovarian cancer and a better understanding of ovarian cancer disease progression and biology. Several of the proteins we have found also have potential for therapeutic or prevention applications. Pilot studies to explore these possibilities are currently underway.

Appendix A
Statement of Work

1. Core Statement of Work Status Table

Major tasks and status listed in Core original Statement of Work

Function Associated with Task	Major Task	Progress
Patient Recruitment	1. Define Data Collection Instruments	Complete, Year 01
Data Warehousing Management	2. Develop data management and tracking systems	Complete, Year 01
Specimen Collection	3. Collect liver and bone marrow specimens.	Accumulated bone marrow only during Year 01
Patient Recruitment	4. Recruit surgery patients	Complete, Year 02
Specimen Collection	5. Collect tissue and blood specimens from surgery patients	Complete, Year 02
Specimen Collection	6. Collect blood specimens from QUEST participants	In process, scheduled for completion, early Year 03
Data management	7. Collect epidemiologic and clinical data	In process, schedule for completion early Year 01
Specimen Characterization	8. Characterize histology specimens	Ongoing. All specimens should be characterized by early Year 03
Specimen Characterization	9. Peform tissue assays for p53, Her2/neu and c-myc	Ongoing. All specimens should be characterized by early Year 03
Specimen Characterization	10. Perform serum assays for CA125	Ongoing. All sera specimens should be characterized for CA125 by early Year 03.
Specimen Distribution	11. Supply Project 1 with 6 specimens to construct cDNA libraries	Complete, Year 01.
Specimen Distribution	12. Supply Project 1 with 30 specimens to hybridize with 1st generation membranes	Complete, Year 02.
Statistical Analyses	13. Develop statistical algorithms for selecting overexpressed genes.	Initiated Year 02. Ongoing.
Specimen Distribution	14. Supply Project 1 with 105 specimens to hybridize with 2nd generation membranes.	Initiated Year 02. In year 02 protocol modified and project needed 75 specimens. 57 specimens provided in Year 02
Specimen Distribution	15. Supply Project 2 with 10 ovarian tumor samples to construct cDNA library	Complete, Year 01.
Specimen Distribution	16. Supply Project 2 with 600 blinded samples for antibody response analyses.	Initiated, Year 02. Will be completed in 1st quarter of Year 03
Specimen Distribution	17. Supply Project 2 with tissue samples from ovarian cancer cases for SEREX analyses.	Complete, Year 02
Statistical Analyses	18. Conduct statistical analyses of antibody response to H2N, p53 and c-myc	Initiated in Year 02, will be completed in Year 03
Statistical Analyses	19. Conduct statistical analyses of	Initiated in Year 02, will be completed

	H2N, p53 and c-myc expression in tissue.	in Year 03.
Statistical Analyses	18. Conduct statistical analyses of select clones with clinical, epidemiological and other lab data.	Initiated, will be completed in Year 03

Appendix B
Project Timeline

1. Project Timeline

				1998									1999												2000																					
				1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
Project	Task	Duration	Status	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
Core	Define Data Collection Instruments	Months 1-3	Complete	x	x	x																																	
Core	Collect Bone Marrow Specimens	Months 1-4	Complete	x	x	x	x	x	x	x	x																												
Core	Develop Data Management and Tracking System	Months 1-9	Complete	x	x	x	x	x	x	x	x	x																											
Core	Recruit Study Participants	Months 1-24	Complete	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x												
Core	Collect Specimens from Participants	Months 1-24	Complete	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x												
Core	Supply Project 1 with 15 specimens to construct cDNA libraries	Months 1-24	Complete	x	x																																		
Project 1	Generation of representative cDNA arrays	Months 1-4	Complete	x	x	x	x																																
Project 2	Perform ELISA Screens for p53, H2N, Myc	Months 1-24	*In process*	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x										
Project 2	Determine SEREX baseline	Months 1-6	Complete	x	x	x	x	x	x	x	x	x	x																										
Core	Supply Project 2 with 10 serous tumor samples for cDNA library construction	Months 1-6	Complete	x	x	x	x	x	x																														
Core	Obtain 2nd set of 10 serous tumor samples for cDNA library construction	Months 1-6	Complete								x																												
Core	Collect Sera from QUEST participants	Months 1-24	Complete	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x												
Core	Collect Epi and Clinical Data on participants	Months 1-24	*In process*	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x										
Core	Characterize Histology of Tissue Specimens	Months 1-24	*In process*	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x										
Core	Perform Tissue Assays for p53, H2N, Myc	Months 3-24	*In process*			x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x										
Core	Perform Serum Assays for CA125	Months 3-24	*In process*			x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x										
Core	Supply Project 2 with 600 blinded samples for AB response analyses	Months 20-24	In process																					x	x	x	x	x	x										
Core	Develop Statistical Algorithms for over-expressed genes selection	Months 4-21	*In process*				x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x											
Core	Supply Project 1 with 31 specimens to hybridize w/1st generation membranes	Month 5	Complete						x	x																													
Project 1	Conduct Primary Characterization of Normal and Neoplastic Ovarian Tissue	Months 6-13	Complete						x	x	x	x	x	x	x	x																							
Project 1	Hybridization of 1st generation membranes with cDNA probes	Months 6-13	Complete						x	x	x	x	x	x	x	x																							
Project 1	Evaluation of hybridization results and selection of 2,000-3,000 overexpressed genes	Months 6-13	Complete						x	x	x	x	x	x	x	x																							
Project 1	Confirmation of tissue specific expression by RT-PCR & Northern blot (~20 genes)	Months 6-13	*In process*						x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x											
Project 1	Construction of second generation cDNA arrays containing 2,000-3,000 clones	Months 6-13	Complete						x	x	x	x	x	x	x	x	x	x	x	x																		
Core	Supply Project 2 with ovca case and control specimens for SEREX analyses	Months 6-12	Complete						x	x	x	x	x	x	x	x																							
Project 2	Use SEREX to screen serum from ovarian cancer patients	Months 6-20	Complete						x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x											
Project 2	Pool Data For Analysis	Months 6-24	*In process*						x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	X	X	X									
Core	Supply Project 2 with 10 mucinous tumor samples for cDNA library construction	Months 8-10									x	x	x																									
Core	Supply Project 1 w/79 tissues for 2nd Generation membranes (n=54)	Months 12-17	Complete												x	x	x	x	x	x	x	x	x	x	x	x												
Project 1	Further characterization of Gene Expression	Months 12-17	Complete												x	x	x	x	x	x																		
Core	Conduct Statistical Analyses of AB response to H2N, p53, Myc	Months 12-24	*In process*												x	x	x	x	x	x	x	x	x	x	x	x	x	x										
Core	Conduct Statistical Analyses of H2N, p53 and Myc expression in tissue	Months 12-24	*In process*												x	x	x	x	x	x	x	x	x	x	x	x	x	x										
Project 1	Characterization of Highly Expressed Genes associated with cancer	Months 17-21	*In process*																	x	x	x	x	x	x	x	x	x										
Project 2	Perform ELISA Screens for Promising Candidates	Months 18-24	*In process*																		x	x	x	x	x	x	x	x										
Core	Conduct Statistical Analyses of Select Clones	Months 21-24	*In process*																					x	x	x	x											
Core	Overall Data Analyses	Months 24-36	*In process*																								x	x	x	x	x	x	x	x	x	x	x	x	x
Core	Manuscript Development	Months 24-36																									x	x	x	x	x	x	x	x	x	x	x	x	x

Key:

x	Activity not yet implemented
x	Activity in process or complete
x	Activity in process and overdue

Appendix C
Project One: Figures

1. Figure 1: Sample Hybridization
2. Figure 2: Hybridization Results
3. Figure 3: Schematic explanation of the clustering
4. Figure 4: Clone clustering on full dataset
5. Figure 5: Example of a tissue clustering result on the entire dataset
6. Figure 6: RealTime quantitative PCR result of HE4 on 82 tissues
7. Figure 7: RealTime data focusing on the expression of the marker genes in all tissues
8. Figure 8: Combined protein/transcript data focusing on the tissues of which there is CA-125 information available
9. Figure 9: SLPI ELISA on sera from 10 ovarian cancer patients and 10 controls

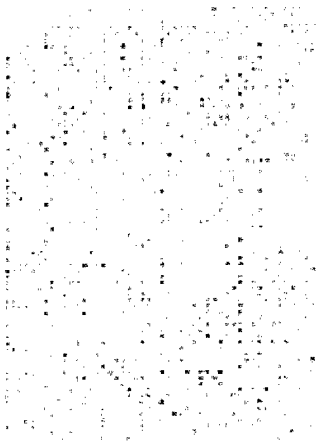


Figure 1 - Sample hybridization
Close view on 1/6 of a membrane containing 3456 colonies that was hybridized with a probe recognizing the vector portion of the cDNA. Where there is no signal, no colony grew. Overall, the number of colonies that did grow reaches 95%.

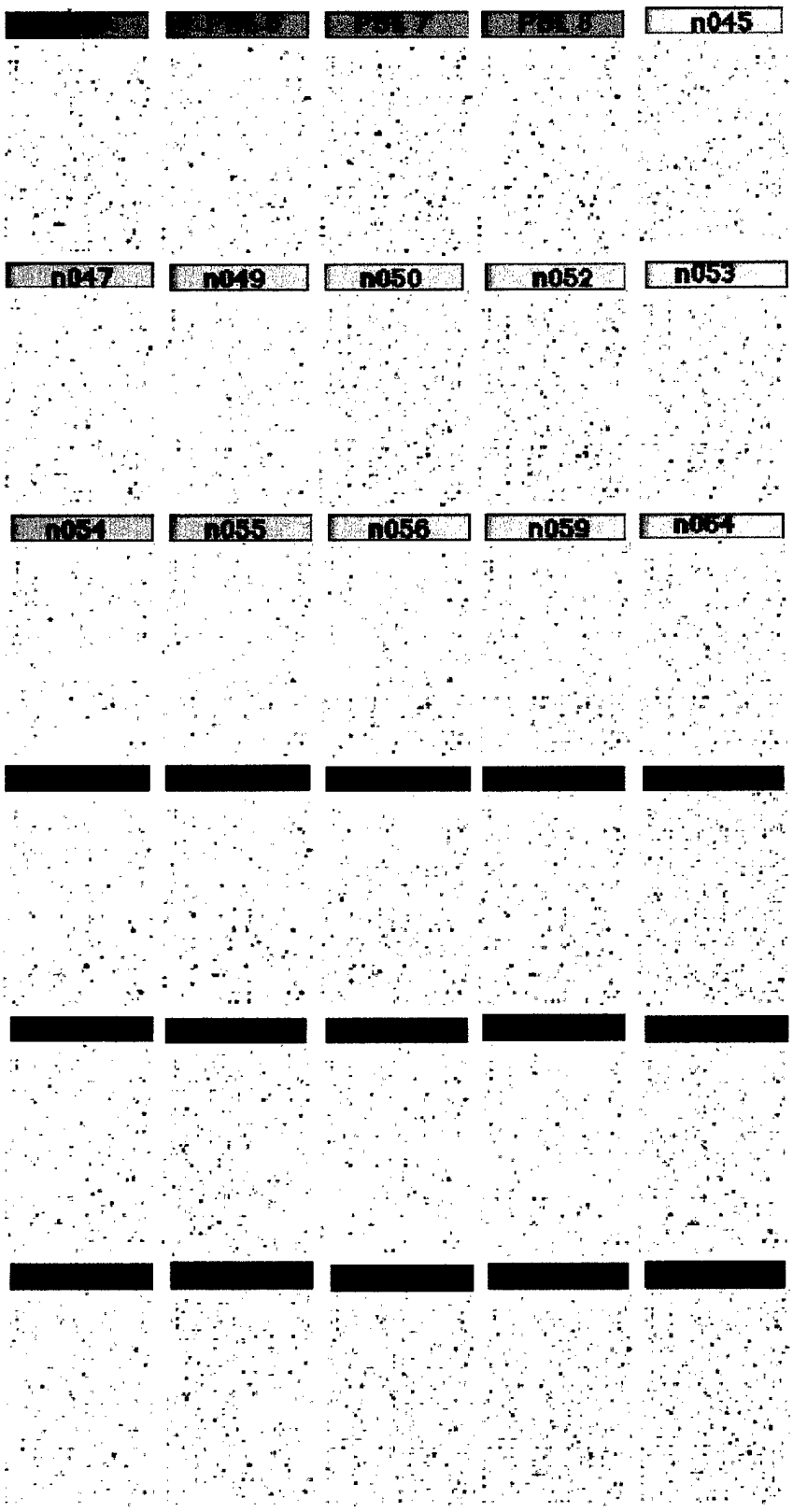


Figure 2 - Hybridization results
Displayed is one field containing 3456 colonies, replicated 30 times and hybridized with probes from 30 different tissues as indicated by the color. Although it may be possible to spot the most obvious differences and similarities in the hybridization pattern by eye, a computer-guided image processing is necessary to detect more subtle changes in expression.

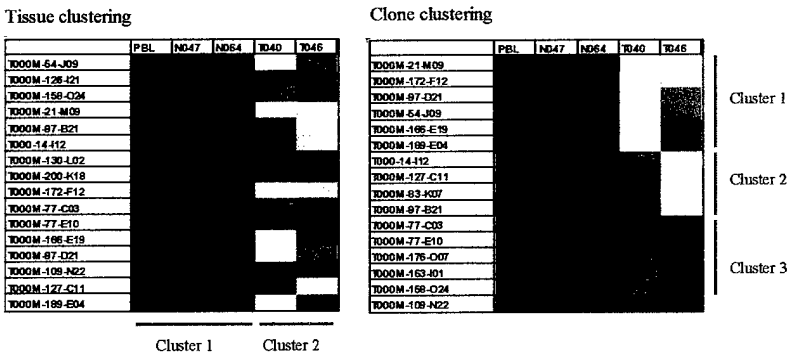


Figure 3 - Schematic explanation of the clustering

For better visual impression, the dataset is represented as a table and the values have been replaced by greyscale where white stands for high expression. Shown are 16 clones out of the 2651 (in the rows) and 5 hybridizations out of the 46 (in the columns): PBL (peripheral blood lymphocytes), two normal ovaries (N...) and two ovarian tumors (T...). In the left panel the tissues were clustered into two groups, one consisting of the normal ovaries and the PBL, the other consisting of the tumors. In order to select potential marker genes, the same clustering algorithm was repeated with a decreasing number of clones that would sort the tissues as nicely as displayed. The minimal number of clones that achieve this grouping are regarded as potential markers. In the right panel the clones were clustered into three groups. It is conceivable that members of a group are either clones representing the same gene or gene family or genes that share similar function or similar pathways. A clone that consistently clusters with a known tumor gene would be regarded as a potential marker gene. The small example shown here was applied to the full dataset as shown in Figure 4.

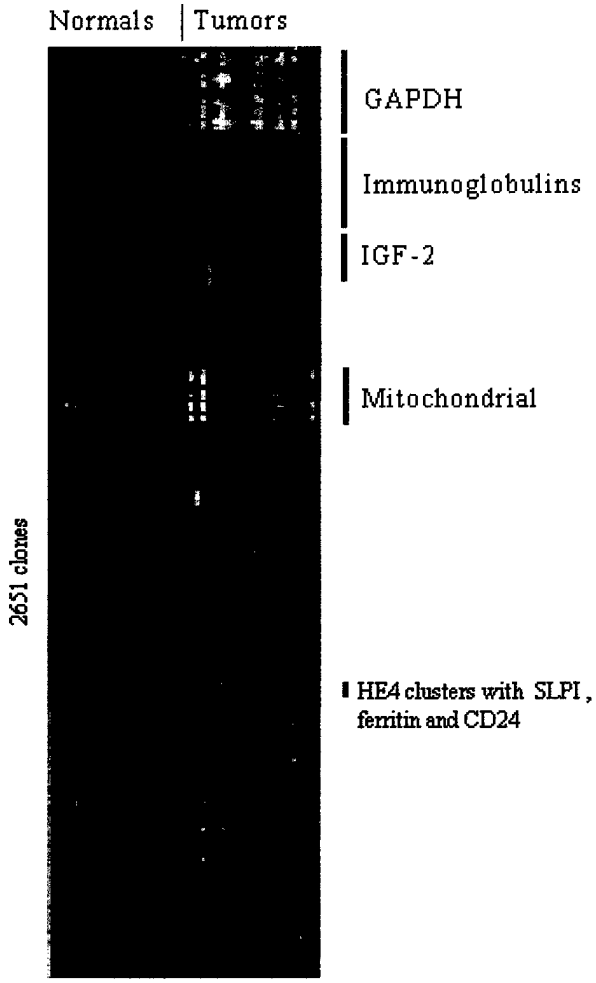


Figure 4 - Clone clustering on full dataset

Clone clustering performed on the full dataset of 2651 clones. The expression values are displayed as greyscale with white standing for high expression and black for a low one. The normal tissues (liver, PBL, normal ovaries) are shown on the left, the ovarian tumors on the right. Overall the expression of the normal tissues is lower than that of the tumors which reflects the selection criteria of these 2651 clones (low expression in normal tissues, high in tumors). In the present example the clones were clustered into 75 groups of varying size. The biggest groups consist to more than 80% of clones matching to GAPDH, immunoglobulins, IGF-2 and mitochondrial genes. Some of the smaller groups contain known tumor genes (such as CD24, ferritin and HE4) together with genes that were previously not known to be associated with tumors (such as SLPI and clones that do not match known sequences in the public databases). These clones were regarded as potential marker genes.

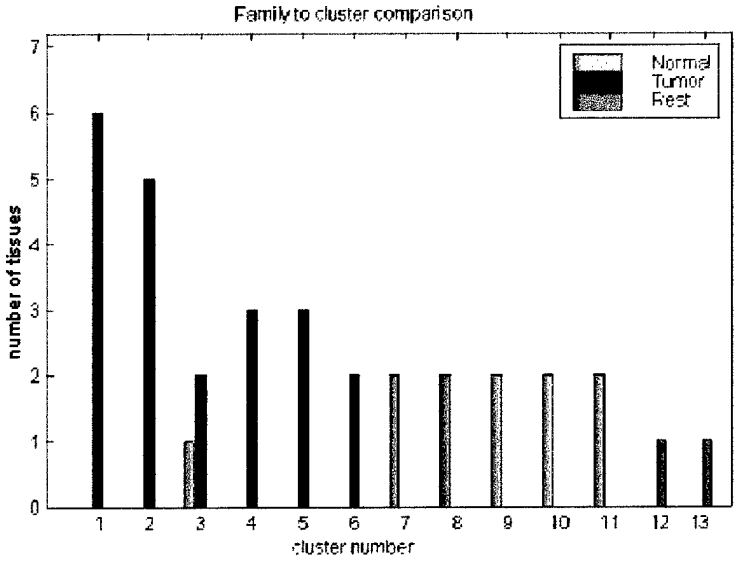


Figure 5 - Example of a tissue clustering result on the entire dataset
Displayed is a typical result for the leave-one-out tissue clustering analysis. The software generated 6 groups which - with the exception of one normal tissue - consist of tumors and five groups that contain only normal tissues. The duplicate and triplicate hybridizations of one tissue were treated as if they had been derived from separate tissues. As a result they either cluster in separate groups, which would be an indicator of low similarity, or they cluster in the same groups, indicating that they are indeed very similar to each other. Of the 7 tissues with repeated hybridizations, 5 have their replicates cluster in the same groups, one has two replicates in a "tumor" group and another replicates in a neighboring "tumor" group, and one has two replicates in a "normal" group and a single replicate in a "tumor" group. The groups 1-13 are formed from the following tissues: 1: hwbc3, t037, t051, t051a, t040, t065; 2: t025, t060, t066, t044a, t044b; 3: n050a, t048, t044; 4: t046, t046a, t046b; 5: t063, t048a, t048b; 6: n039a, t043; 7: hpbl7, hpbl8; 8: n047a, n047b; 9: n050, n050b; 10: hliv2, hpbl6; 11: n056, n064; 12: t062; 13: t058. An "a" or a "b" behind the tissue name refers to the duplicate and triplicate hybridization.

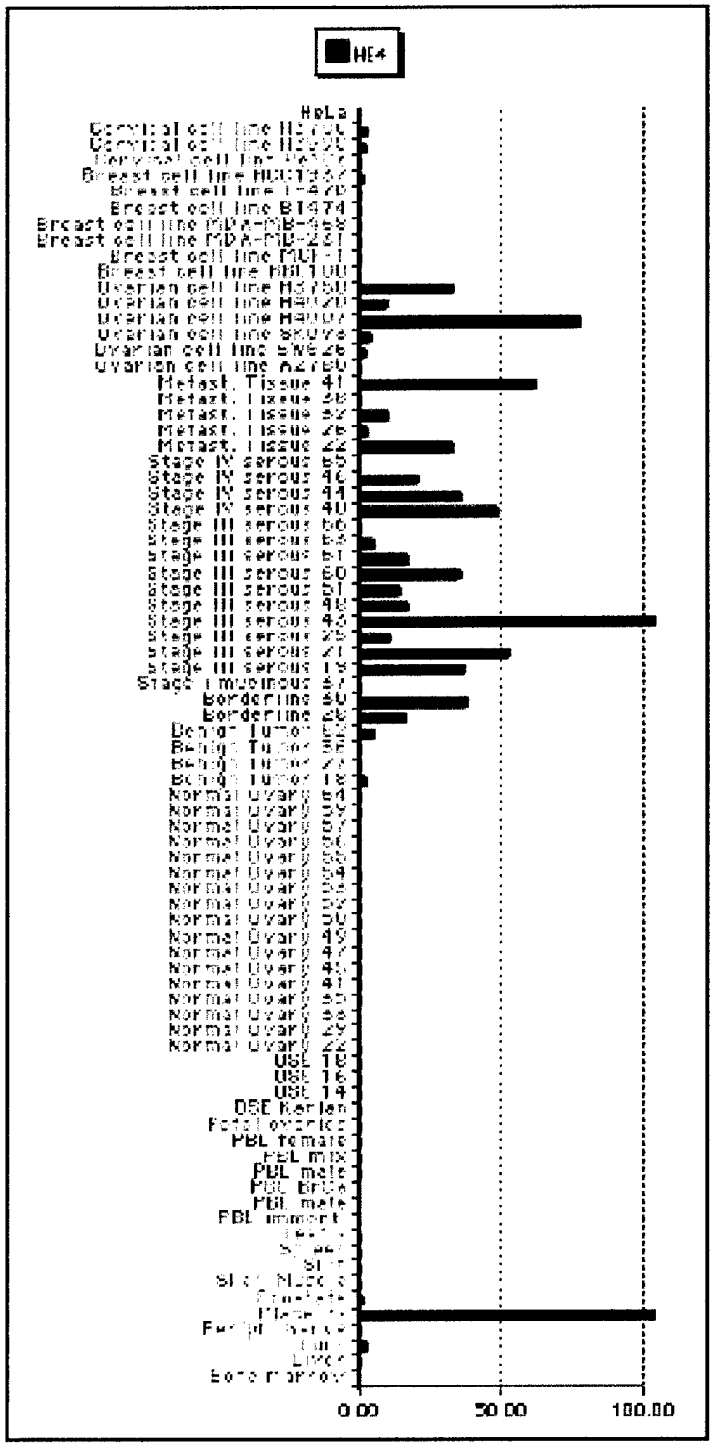


Figure 6 - RealTime quantitative PCR result of HE4 on 82 tissues

The tissue names are listed on the bottom. Brown stands for normal non -ovarian tissues, blue for peripheral blood lymphocytes, green for normal ovaries, orange for benign ovarian tumors, red for ovarian carcinomas of increasing stage, and the leftmost 17 entries are ovarian, breast and cervical cell lines. The y-axis shows expression of HE4 relative to the S31iii125 standard. These are arbitrary values that can nevertheless be used for comparison of the degree of expression of different genes. Beta actin, a medium high expressed gene, would show numbers in the 400 range, a lowly expressed gene would show numbers in the 0.1 range. HE4 transcript expression is, with the exception of placenta and lung, clearly restricted to the ovarian tumors. This pattern shows HE4 as a marker gene with high specificity and sensitivity.

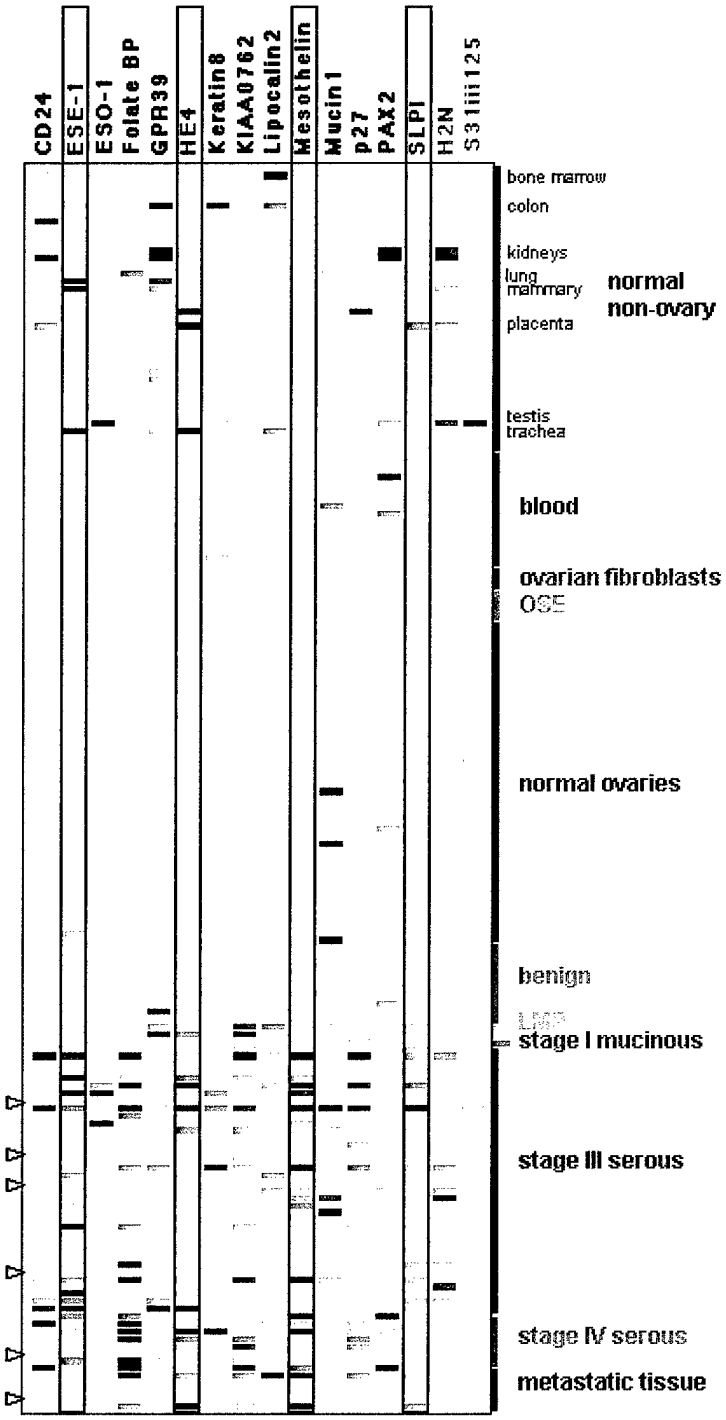


Figure 7 - RealTime data focusing on the expression of the marker genes in all tissues
The expression of 15 genes in 202 tissues was determined by RealTime quantitative PCR. Listed on the right are the tissues using the same colors employed throughout the report. The names of the genes are listed at the top. The expression values are expressed as greyscale bands with black standing for high expression and white for low. The four best performing genes are highlighted. The values are not normalized since normalization requires a gene or a group of genes with prior knowledge of their unchanged expression in the tissues tested. Since this is impossible, we have included in this panel the gene S31iii125 *which is expressed in all tissues shown, albeit with some variation. We would like to point out that had we normalized by the values of this gene, the overall expression pattern would still look the same with some bands being darker or lighter than otherwise. The open triangles on the left side mark tissues that show no elevated expression for either of the marker genes. H2N stands for Her2/neu.*

CA125	p53	Her2/neu	CD24	ESE-1	ESO-1	GPR39	HE4	Kera-tin8	Lipo-calin	Meso-thelin	Mucin1	p27	PAX2	SLPI	Tissue
U/ml	*	*	**	**	**	**	**	**	**	**	**	**	**	**	
21	0	0	0.0	0.0	0.0	0.1	0.0	0.0	0.3	0.0	0.0	0.8	0.0	0.0	n022
17	3		0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.3	0.0	0.0	n029
7	0	0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	n033
7	3	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	n035
43	1	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.0	n041
9	1	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	n045
24	0	0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	n047
	0	0	0.0	0.1	0.0	0.3	0.0	0.1	0.3	0.0	0.2	0.1	0.0	0.0	n049
20		3	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.0	n050
5	1	1	0.0	0.0	0.4	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.0	n052
	0	0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	n053
5	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	n054
14	0	0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.0	n055
7	0	0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	n056
31	1	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	n057
	8	8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.4	0.1	0.0	n059
10	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	n064
		1	0.0	0.2	0.0	0.2	0.0	0.0	0.0	0.0	38.9	0.3	2.4	0.0	n082
24	0	0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	n083
9		3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	n084
8	0	0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	n085
14	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.1	n087
8	0	0	0.3	0.0	0.0	0.2	0.0	0.1	0.0	0.0	0.0	0.1	1.9	0.2	n088
8	0	0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	1.0	0.0	n089
12	0	0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	2.9	0.0	3.0	0.0	n090
14	0	0	0.0	0.1	0.0	0.0	0.0	0.1	0.1	0.0	1.2	0.1	3.0	0.0	n092
11	0	0	0.0	0.2	0.0	1.8	0.3	0.2	0.0	0.0	1.8	0.0	0.0	0.1	n093
		3	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0	0.0	0.1	n094
17	0	0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.1	1.6	0.1	0.0	0.1	n095
17	0	0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.3	0.0	0.1	n096
	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.1	0.0	0.0	n097
64	0	0	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.7	0.2	n100
	1	1	0.0	0.2	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0	27.5	0.1	n102
10	0	0	0.0	0.1	0.0	0.2	0.0	0.1	0.0	0.0	0.0	0.3	0.6	0.0	n103
		3	0.0	0.2	0.0	0.0	0.0	0.1	0.4	0.0	0.0	0.1	0.0	0.0	n105
	0	0	0.5	0.5	0.0	0.0	0.0	0.8	0.0	0.0	1.5	0.1	0.0	0.1	n106
7	1	1	0.7	1.1	0.2	0.4	1.7	1.6	0.5	0.9	0.2	0.7	0.0	0.6	t018b ●
61	0	0	0.2	0.1	0.0	0.6	0.2	0.3	1.0	0.1	0.2	0.1	0.0	0.3	t027b
8			0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.3	0.1	0.0	t036b ●
7	0	0	0.2	0.1	0.0	0.2	0.4	0.1	0.0	0.1	0.0	0.0	0.1	0.5	t062b ●
22	0	0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	5.9	0.0	0.0	0.0	t091b ●
4	0	0	0.0	0.1	0.0	1.8	0.0	0.1	0.0	0.5	0.4	0.0	0.7	0.3	t099b ●
	1	1	0.0	0.1	0.0	0.1	0.0	0.1	0.7	0.0	0.0	0.0	0.0	0.2	t115b
81			0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.3	0.2	0.4	0.0	t125b
	0	0	0.4	1.2	0.0	0.9	0.7	0.2	0.1	1.4	0.8	1.0	5.9	0.3	t202b
	1	1	2.0	1.5	0.3	15.0	5.1	0.7	0.4	0.8	0.3	1.3	4.5	1.5	t203b
24	0	0	0.4	0.1	0.0	16.6	0.0	1.8	1.1	0.0	0.0	0.2	22.1	0.0	t204b ●
	1	1	0.5	0.4	0.0	2.8	1.3	1.0	0.0	0.1	0.1	0.1	0.4	0.8	t020l
	1	3	0.0	0.0	0.0	0.0	0.1	0.2	0.0	0.2	0.0	0.3	0.2	0.7	t020l
		3	0.0	0.1	0.0	0.5	0.0	0.4	0.0	0.0	0.8	1.1	0.2	0.3	t037
		1	0.3	0.4	0.6	0.3	1.7	0.8	1.9	1.7	0.4	1.4	0.0	1.2	t019
	1	1	0.8	2.2	2.3	0.3	0.6	4.1	0.1	1.6	0.8	0.6	1.9	0.7	t021
		3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	t025
9	1	3	0.2	0.1	0.0	0.8	1.6	1.3	0.5	0.5	0.1	1.3	0.1	1.0	t031 ●
			0.2	0.0	0.1	0.0	0.0	1.0	0.0	0.1	0.0	1.0	0.0	0.7	t038m
	8	8	0.5	0.2	0.0	0.3	1.4	0.5	0.0	2.1	0.6	0.4	0.0	0.6	t043
			0.4	1.2	0.0	0.0	1.0	0.4	0.3	0.0	0.2	1.1	0.0	0.5	t048
			0.0	0.0	13.7	0.0	0.3	0.1	0.1	0.3	0.0	0.3	0.0	0.2	t051
		3	0.3	0.2	0.0	0.3	0.7	0.2	0.3	0.3	0.4	0.4	0.3	0.4	t060
14	1	3	0.1	0.2	0.0	0.0	1.3	0.9	0.0	0.9	0.2	0.5	0.0	0.4	t061 ●
18	1	1	2.1	10.6	0.0	0.9	2.2	0.3	0.1	0.4	0.1	2.5	2.5	0.7	t063 ●
18	0	0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.5	0.0	0.5	t066 ●
91	0	0	0.5	0.4	0.0	0.8	0.1	0.4	0.5	0.6	0.0	0.4	0.8	0.7	t096
	0	0	1.0	1.1	0.0	3.6	0.4	3.6	0.0	3.5	0.0	1.7	0.0	1.4	t098
36	0	0	0.0	1.6	0.0	0.0	1.0	1.3	3.6	0.0	0.0	0.6	0.0	1.1	t101
51	3	3	2.1	0.5	0.0	1.6	0.7	0.2	0.2	0.0	0.1	0.1	0.7	1.3	t104
			0.9	2.0	0.0	0.0	2.3	2.0	0.8	2.1	2.5	1.6	4.8	2.0	t107
		3	0.1	0.0	0.0	1.4	1.2	2.0	1.8	3.2	2.0	0.2	1.9	2.4	t108
	0	0	0.1	4.6	0.0	1.0	0.8	4.2	9.0	5.8	0.5	0.4	1.8	4.2	t109
		3	0.0	0.6	0.0	0.6	0.2	0.4	0.0	0.7	1.4	0.6	0.0	1.1	t110
		3	0.6	0.5	0.0	0.6	0.7	1.0	1.2	0.3	0.0	0.3	1.5	0.3	t111
	0	0	0.9	1.3	0.0	0.5	0.6	1.4	0.0	0.0	1.5	1.9	0.0	0.7	t112
		1	0.3	1.5	0.0	1.1	0.6	0.6	0.9	0.0	0.3	1.4	0.0	0.3	t113
	8	8	0.2	1.1	0.0	0.7	0.5	0.7	0.0	0.0	0.0	0.1	0.0	1.5	t114
	1		0.2	1.1	0.0	0.8	0.7	0.6	0.5	0.0	0.6	0.0	0.0	1.8	t116
51		0	0.0	0.2	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.1	0.0	0.1	t117
		3	1.8	0.8	0.0	1.4	0.3	1.5	0.2	0.2	0.0	0.8	0.0	2.1	t118
	1	3	0.0	0.1	0.0	0.0	0.1	0.1	0.0	0.1	0.6	0.1	0.0	0.1	t120
	1	1	1.3	1.5	0.0	1.4	0.7	1.9	2.3	4.2	1.4	1.0	0.0	1.5	t122
		3	2.8	0.5	0.0	1.0	1.5	0.8	0.0	0.1	0.0	0.1	0.0	1.0	t124
8	1	1	2.1	3.2	0.0	2.1	0.3	0.3	2.5	0.0	0.7	1.5	0.4	2.2	t206 ●
			1.1	0.9	0.0	1.3	0.5	0.9	2.5	0.9	0.0	0.1	0.9	2.2	x119
	0	0	0.6	0.9	0.0	0.5	1.6	5.0	0.4	1.0	0.5	0.8	1.1	1.1	t040
	1	1	0.5	0.4	0.0	0.2	1.0	0.3	0.6	0.0	0.5	1.6	0.0	0.6	t044
63	1	1	0.1	0.4	0.1	0.2	1.0	0.2	0.0	0.0	0.4	2.1	0.0	0.1	t046
			0.0	0.0	0.0	0.2	0.0	0.1	0.1	0.0	0.0	0.1	0.0	0.0	t065
	0	0	1.0	1.6	0.0	1.6	0.5	0.5	2.1	0.4	0.3	0.2	0.0	1.6	t123

Figure 8 - Combined protein/transcript data focusing on the tissues of which there is CA125 information available

*CA-125 serum levels of the ovarian cancer patients and the controls paired with the tissue protein levels of p53 and Her2/neu, listed side-by-side with the transcript levels of selected potential marker genes found in this study. The patient diagnosis / tissue type is listed in the rightmost column (colors are the same as used in Figure 7). Values are overlaid with color for easier identification: CA-125: 0-29 U/ml (turquoise), 30-99 U/ml (faint red), 100-399 U/ml (red), over 400 U/ml (dark red), white: not done. * p53 and Her2/neu: 0, assay not run; 1, no overexpression (turquoise); 3, uninterpretable (light red); 5, intermediate overexpression (red); 6, high overexpression (dark red); 8, assay will not be run. The RealTime quantitative PCR values were normalized by the average expression of each gene in all tissue in order to have the values in each column on the same scale. ** 0-0.1, no expression (white), 0.2-0.9 weak expression (light red), >1.0, high expression (red). Ovarian cancer patients with CA-125 levels below 30 U/ml that have high levels of one or more of the newly found markers are labeled with a black dot after the tissue name.*

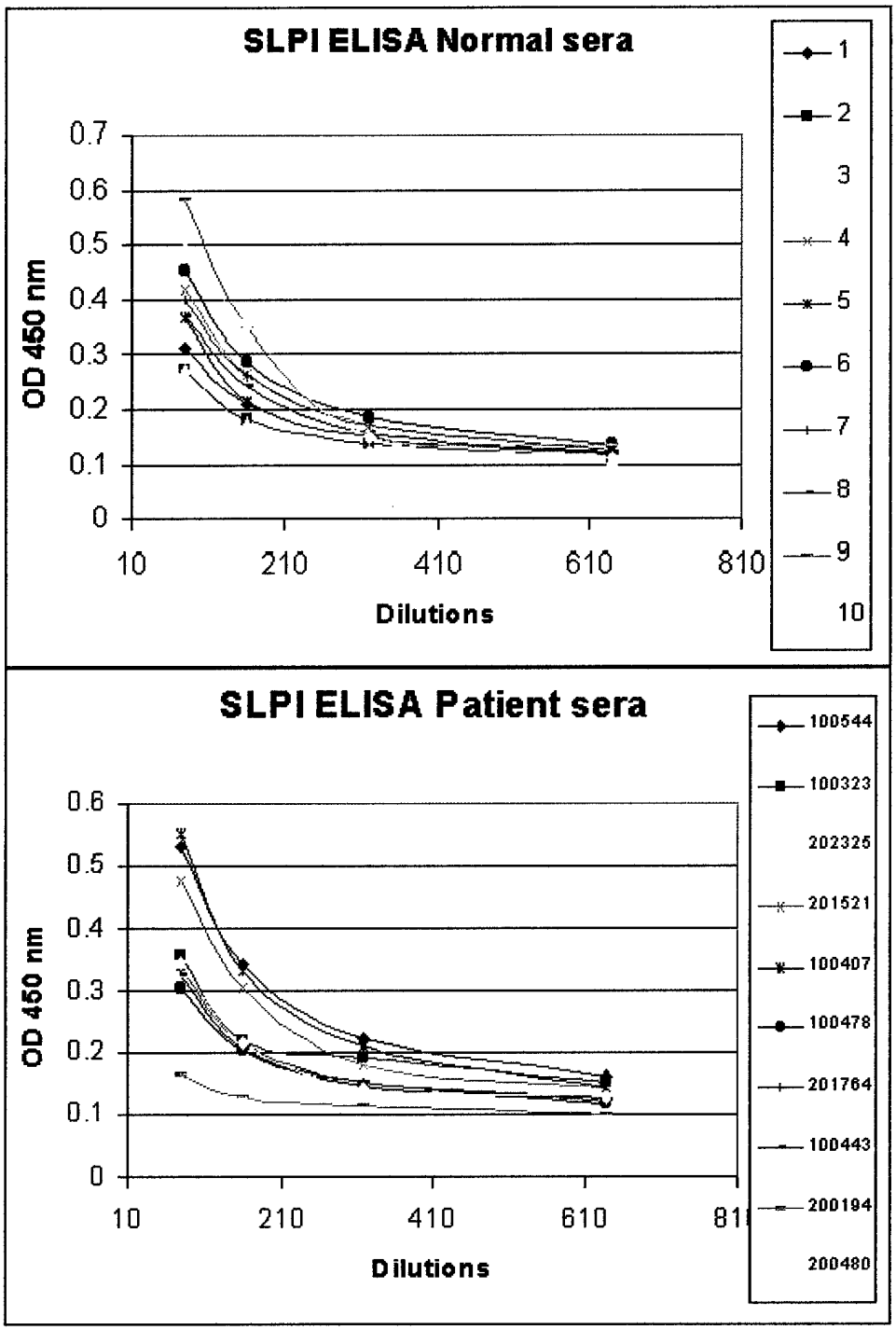


Figure 9 - SLPI ELISA on sera from 10 ovarian cancer patients and 10 controls
Top: sera from 10 normal controls, bottom: sera from 10 ovarian cancer patients whose tissues showed overexpression of SLPI message as assayed by RealTime quantitative PCR (see Figure 7). There is no difference in protein levels between the two groups.

Appendix D
Project 1: Related Publications

1. "Comparative hybridization of an array of 21,500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas
2. "Tissue Classification with Gene Expression Profiles"
3. Abstract: "Hybridisation of an array of 100,000 cDNAs with 32 tissues find potential ovarian cancer marker genes"

ELSEVIER

Gene 238 (1999) 375–385

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Comparative hybridization of an array of 21 500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas

Michèl Schummer [a,*], WaiLap V. Ng [a], Roger E. Bumgarner [a], Peter S. Nelson [a], Bernhard Schummer [b], David W. Bednarski [a], Laurie Hassell [a], Rae Lynn Baldwin [c], Beth Y. Karlan [c], Leroy Hood [a]

[a] *Department of Molecular Biotechnology, University of Washington, Box 357730, Seattle, WA 98195, USA*
[b] *Institut für Pharmakologie und Toxikologie, Fakultät für Klinische Medizin der Universität Heidelberg, Maybachstr. 14-16, 68169 Mannheim, Germany*
[c] *Department of Obstetrics and Gynecology, Cedars-Sinai Medical Center, University of California, Los Angeles, School of Medicine, Los Angeles, CA 90048, USA*

Received 28 January 1999; received in revised form 2 July 1999; accepted 28 July 1999; Received by I. Verma

Abstract

Comparative hybridization of cDNA arrays is a powerful tool for the measurement of differences in gene expression between two or more tissues. We optimized this technique and employed it to discover genes with potential for the diagnosis of ovarian cancer. This cancer is rarely identified in time for a good prognosis after diagnosis. An array of 21 500 unknown ovarian cDNAs was hybridized with labeled first-strand cDNA from 10 ovarian tumors and six normal tissues. One hundred and thirty-four clones are overexpressed in at least five of the 10 tumors. These cDNAs were sequenced and compared to public sequence databases. One of these, the gene *HE4*, was found to be expressed primarily in some ovarian cancers, and is thus a potential marker of ovarian carcinoma.

Keywords: Cancer maker; DNA array; Differential expression; HE4

1. Introduction

Ovarian cancer is the leading cause of gynecological cancer death in the United States. The American Cancer Society estimates that in 1998, some 25 400 women will develop ovarian cancer and 14 500 will die from it (American Cancer Society, 1998). The overall 5 year survival rate is about 46%, and has remained essentially unchanged for 25 years. Ovarian cancer is ranked fifth in cancer mortality among women, and raises concerns both with women and physicians because of its generally poor prognosis. Cancers diagnosed at an early stage have a 5 year survival rate of 92% in contrast to a 25% 5 year survival rate for patients with disseminated disease at diagnosis. Seventy-five per cent of epithelial ovarian cancers are diagnosed at advanced stages. This is in part due to the lack of symptoms early in the disease course, and the absence of a sensitive and specific screening test for early disease detection. Currently available ovarian cancer markers such as CA-125 are neither sensitive nor specific enough for population screening to detect early, treatable ovarian cancers (Jacobs et al., 1993).

We describe the use of 'high-density cDNA array hybridization' (HDAH) to identify transcripts that show high expression levels in ovarian cancer tissues as compared to ovarian surface epithelium (OSE). This technology has been used in a variety of experiments to identify transcripts (Schena et al., 1998), whose expression patterns differ in two tissues (e.g. normal and cancer). Our objective is to find (1) transcripts that are overexpressed in tumor as contrasted with normal ovarian tissue and (2) cDNAs encoding proteins that could be useful diagnostic markers (e.g. secreted or cell-surface pro-

Abbreviations: bp, base pair(s); cDNA, copy DNA; EST, expressed sequence tag; HDAH, high-density array hybridization; HE4, human epididymis gene 4; kb, kilobase(s); nt, nucleotide(s); OSE, ovarian surface epithelium; PBL, peripheral blood lymphocytes; RT-PCR, reverse transcription polymerase chain reaction.

* Corresponding author. Tel.: +1-206-616-5117; fax: +1-206-685-7301.
E-mail address: kikjou@u.washington.edu (M. Schummer)

0378-1119/99/$ – see front matter
PII: S0378-1119(99)00342-X

teins). Two general types of assays are possible: (1) protein assays for secreted proteins or on the surface of cells that metastasize into the circulation, and (2) PCR assays from genes uniquely expressed in blood-borne (or ascites-borne) tumor cells. Hybridizing 21 500 randomly selected cDNAs from normal and neoplastic ovarian tissues with probes from 10 ovarian tumor and six normal tissues, we identified 134 clones with higher expression signals in ovarian tumors as opposed to normal tissues. These clones were sequenced, and in some cases, their expression pattern was confirmed by RT-PCR and Northern blot analysis. The expression pattern of one of these clones, *HE4*, suggests that it may be a potential candidate diagnostic marker for ovarian cancer.

2. Materials and methods

2.1. Tissues and cells

We used the following tissues for our experiments: ovarian surface epithelium short-term culture (Karlan et al., 1995), early passages (OSE); normal ovary consisting of primarily stromal cells (N002, N005, N006, N019 and N035); two benign ovarian tumors (T017B, an endometrioid polyp, and T018B, a serous cystadenoma); one borderline early stage serous carcinoma, LMP (T028L); late-stage, high-grade papillary serous ovarian adenocarcinomas (T001–T006, T008–T011, T014–T016 and T021); two early-stage ovarian adenocarcinomas (one serous: T007 and one mucinous: T037); one late-stage, high-grade serous ovarian adenocarcinoma post-chemotherapy (T012); two late-stage, high-grade serous ovarian adenocarcinoma with massive metastases (T013M and T026M); peripheral blood lymphocytes (PBL1 and PBL2); Fetal ovaries: pool of 25 fetal ovaries (52–103 days); bone marrow, cerebellum, kidney, liver and placenta (Clontech, Palo Alto, CA). In order to minimize the effect of variance in tissue collection on the RNA quality and hence the hybridization patterns, we ensured that tissue collection would adhere to the following guidelines. After surgery, a tissue section was taken for the pathologist's examination and an adjacent section was snap-frozen in liquid nitrogen. All ovarian tumor tissue specimens were examined for their tumor cell content (which was above 80%) and the absence of necrosis. RNA preparations of all tissues or cell cultures were performed using the Trizol method (Life Technologies, Grand Island, NY). Poly(A)$^+$ RNA was prepared using a mRNA purification kit (Stratagene, La Jolla, CA). Tissue samples of 200–400 mg of tumor were used for RNA preparation. We have found that samples of less than 200 mg do not yield sufficient RNA for our analysis. The integrity of total RNA was determined by visual inspection of the 28S and 18S ribosomal bands to ensure that degraded samples that might give a different expression profile than intact RNA were not used.

2.2. Minipreparation of 21 500 ovarian clones

Five cDNA libraries were created from ovarian tissues and cell cultures (OSE, T007, T008, T010 and T012) using the ZAP-cDNA synthesis kit (Stratagene). Examining the cDNA clones using PCR, the insert sizes were found to average between 1.2 and 1.5 kb. From each library, 96 clones were randomly chosen, sequenced and analyzed by similarity analysis against the non-redundant and EST database. The low number of mitochondrial and ribosomal sequences, the limited number of clones with no insert, and the significant cDNA diversity indicated that the libraries were of high quality. Using a 96-deep-well plate-based minipreparation assay (Ng et al., 1996), we picked 21 500 transformants (8600 from the OSE cDNA library and 3225 each from the four tumor cDNA libraries), extracted the cDNAs and transferred them to 384-well microtiter plates.

2.3. Dotting the 21 500 clones onto nylon membranes

Using a hand-held arraying tool with a 384-pin printhead developed in our laboratory (Schummer et al., 1997), we dotted the 21 500 cDNAs onto 16 sets of 14 nylon membranes of 7.5 × 12 cm, which held each of the 1536 clones. The cDNA was denatured and immobilized on the membrane as previously described (Schummer et al., 1997).

2.4. Labeling and hybridization protocol

Each set of membranes was hybridized with a complex probe consisting of ^{32}P-labeled first-strand cDNA. Briefly, 5 μg of poly(A$^+$) RNA or 30 μg of total RNA were reverse-transcribed using Superscript II reverse transcriptase (Life Technologies) and oligo-dT$_{12}$ primers with 30 μCi of alpha-^{32}P-dCTP (3000 Ci/mmol) and unlabeled dATP, dGTP, dTTP at 1 mM each; after 20 min, unlabeled dCTP was added to a final concentration of 1 mM, and the reaction was continued for another 40 min. This unpurified probe was hybridized to 12 membranes under conditions described previously (Schummer et al., 1997). The membranes were washed at increasing stringency (20 min, 2× SSC, 0.5% SDS, RT; 20 min 0.5× SSC, 0.5% SDS, 65°C; 2 × 20 min, 0.2× SSC, 0.5% SDS, 65°C).

2.5. Software for spot detection

After hybridization and washing, the membranes were exposed to a phosphor storage screen, and the hybridization patterns were captured as 16-bit TIFF

images using a PhosphorImager (Molecular Dynamics, Sunnyvale, CA). Nine nylon membranes were imaged simultaneously on a 35 × 45 cm screen. The resulting file was processed using a software package developed in our laboratory. The TIFF image was split into nine smaller images, each representing one of the arrayed membranes. Briefly, the user defined the outer dimensions of each membrane by placing a cursor into each the upper left, upper right and lower right corner of each of the nine array images. Subsequently, the computer superimposed a grid, approximating the positions of the 1536 dots. By five passes of center-of-mass finding, the computer determined the exact center of each of the 1536 dots. It integrated the area of an experimentally determined number of pixels around each center that covered the area of the largest hybridization signal present on the membranes. The intensities of all pixels in the area were integrated. Local background was calculated by choosing one pixel with the lowest intensity out of four pixels situated halfway between one dot and its four diagonal neighbors. Both values were stored in a tab-delimited text file together with the coordinates of the spot on the array.

2.6. Single pass 5′ sequencing, database analysis and sequence comparison

Sequencing was performed on plasmid DNA and PCR products using previously described methods (Ng et al., 1996). The single-pass sequences were edited to remove vector and poly(A) sequences. Edited sequences were compared with those in the EST (dbEST) and non-redundant nucleotide and protein databases (GenBank) at the National Center for Biotechnology Information (NCBI) using the Baylor College of Medicine Search Launcher batch client server 'Search Launcher' (http://www.hgsc.bcm.tmc.edu/Search Launcher/). Nucleotide sequence comparisons were carried out using BLASTN. Comparisons of conceptual protein translations were performed using the program BLASTX with BEAUTY sequence annotation enhancement. Each clone was categorized as to known gene homology, EST homology, or novel.

2.7. RT-PCR

Clones determined by to be differentially expressed by array analysis were confirmed by single tube RT-PCR, which has been shown to be a highly sensitive measure of transcript abundance (Schummer et al., 1998). Two primers, with a base pair length of 20–24 and with T_ms between 64 and 66°C, were designed for each gene. The distance between the primers was 420–660 bp. RT-PCR (Titan®, Boehringer Mannheim, Mannheim, Germany) was performed with 200 ng of total RNA according to the manufacturer, with the following cycles: 30 min at 50°; 2 min at 94°; 10 cycles of 30 s at 94°, 30 s at 60°, 45 s at 68°; 12–25 cycles of 30 s at 94°, 30 s at 60°, 45 s at 68° (with elongation of 5 s for each cycle); 7 min at 68°. For each gene, the logarithmic phase of amplification was determined prior to the Titan®-PCR. The individual reactions were run on a 1% agarose gel stained with SYBR-Green at 500 × diluted concentration for 1 h and scanned on a FluorImager (Molecular Dynamics, Sunnyvale, CA). For each gene and tissue, four identical reactions were performed.

2.8. Northern blot

A *HE4* PCR product of 500 bp was cloned into a pCR2.1 vector using the TA cloning kit (Invitrogen, San Diego, CA). A digoxigenin-labeled riboprobe was prepared from this vector using a Genius RNA DIG labeling kit (Boehringer Mannheim, Germany). The probe was hybridized overnight at 68°C in DIG Easy Hyb buffer and washed in 2 × SSC, 0.1% SDS for 15 min at room temperature; 2 × SSC, 0.1% SDS for 20 min at 68°C; and 0.1 × SSC, 0.1% SDS for 2 × 15 min at 68°C. The hybridized RNA was visualized using the DIG detection kit (Boehringer Mannheim), and the membrane was exposed to X-ray film for 15 min.

3. Results and discussion

3.1. Evaluation of high-density filter hybridization

Tissues comprise many different cell populations. Each type of cell in a tissue exhibits its particular gene expression pattern. Since most ovarian tumors arise from epithelial cells, the comparison of tumors against ovarian surface epithelium should provide a useful comparison. Two qualifications must be made: (1) ovarian surface epithelial cells in a short-term culture will probably have some differences in expression patterns from in-vivo ovarian epithelial cells, and (2) tumors may have intermixed normal cells from the ovary. In order to detect genes that are overexpressed in one cell type or tissue versus another, one needs to know the limitations of the detection system, notably (1) the upper and lower limits of detection (signal-to-noise ratio) which — translated into the number of mRNA molecules detectable per cell — should be suitable for the proposed study, and (2) the measured level of variation in signal intensity on identical membranes interrogated with identical probes. The latter will determine a factor above which overexpression can be regarded as significant.

3.1.1. Determination of detection limit and dynamic range

The sensitivity of the array technology determines the number of detectable mRNA molecules in a cell. In order to determine the mean signal-to-noise ratio, we hybridized 14 identical arrays containing 1536 identical cDNAs coding for the green fluorescent protein (*GFP*) with first-strand cDNA probes made from human liver poly(A)$^+$ RNA in which a *GFP* mRNA was added in decreasing concentrations (14 different concentrations ranging from one transcript in 200 to one in 20 000). As depicted in Fig. 1, the probe with the highest *GFP* concentration yielded a mean value of 8300 ± 416 dpm (decays per minute) per pixel, and the mean background value was determined as 90 ± 18 dpm/pixel. With background subtraction, this represents a dynamic range of 456 (background-subtracted signal divided by background fluctuation: $8210/18 = 456$) or 2.5 orders of magnitude. We established a lower limit of sensitivity of 1 *GFP* RNA in 20 000 liver RNAs, a result similar to those in other studies (Piétu et al., 1996). Based on an estimated 10^5–10^6 transcripts per average eukaryotic cell (Bishop et al., 1974), the membrane-based HDAH can detect a minimum of between five and 50 mRNA molecules in a cell and a maximum of 500–5000. The lower limit falls in the low to medium class of transcripts, and the upper limit lies in the highly expressed gene class (Zhang et al., 1997). This detection range should be sufficient for the identification of overexpressed genes.

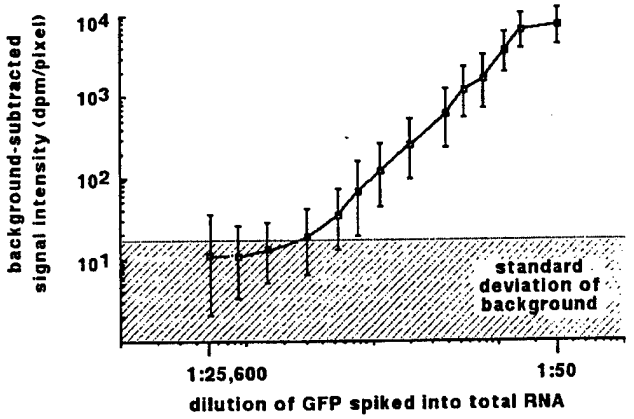


Fig. 1. Determination of the linearity of the hybridization signal. Fourteen replica membranes with 1536 *GFP* cDNAs each were hybridized with first-strand cDNA made from poly(A)$^+$ RNA from human liver with *GFP* mRNA spiked into it in a twofold serial dilution starting with a 1:50 dilution. The signal intensity is measured as the average of all pixels of all 1536 signals. Displayed are the background-subtracted intensities with their standard deviations. The resulting curve is linear over 2.5 orders of magnitude). The standard deviations increase with decreasing signal-to-noise level. The shaded area indicates the standard deviation of the background. The background intensity averaged at 90 ± 18 dpm/pixel, and the highest intensity averaged at 8300 ± 416 dpm/pixel.

3.1.2. Normalization of the hybridization signals

In order to compare hybridization signatures of two identical membranes that have been hybridized with different probes in two separate incubations, one needs to normalize the signals to a standard. Although we adhere to a strict protocol, slight variations can be introduced by minute differences in probe labeling, probe purification, hybridization and wash conditions and exposure time. We normalized the background-subtracted intensities of one membrane by setting the median to 1. Assuming that among the 1536 clones present on one membrane, the majority does not alter its expression (Zhang et al., 1997), we believe that this is justified.

3.1.3. Determination of variation in signal intensity

Two factors influence the accuracy of the hybridization detection for one particular cDNA on a membrane: the amount of cDNA on the membrane (governed by the dotting procedure) and the amount of labeled cDNA that remains bound to the target cDNA on the membrane after hybridization (governed by the efficiency of the probe labeling reaction and the hybridization and washing kinetics). We determined the variation of amounts of DNA spotted by our arraying tool to be $\pm 14\%$ (data not shown). Since the probe consists of a complex mixture of cDNAs, the arrayed DNA is in vast excess of the probe cDNA, and thus the variations caused by the spotted cDNA can be regarded as negligible. In order to assess the probe-to-probe variance, we hybridized four replica membranes containing 1536 ovarian cDNAs with four ^{32}P-labeled first-strand cDNA probes independently generated from one batch of total RNA prepared from liver tissue. We compared the background-subtracted intensities of one cDNA across the four membranes and calculated the standard deviation, thus generating 1536 values. We ranked the clones by their expression and determined three means of standard deviations, one for the upper, the middle and the lower third, corresponding roughly to the high, medium high and low, expression categories of transcripts. The mean of the standard deviations amounted to $\pm 15\%$, $\pm 24\%$ and $\pm 40\%$ respectively, which averages to $\pm 26\%$ for all clones. Using the following equation, we calculated the threshold value for a ratio to be regarded as significant: [1 + standard deviation]/[1 − standard deviation]. In order to be above this threshold of significance, a highly expressed gene needs to display a ratio of 1.35, a medium expressed gene a value of 1.63 and the least expressed gene a value of 2.33. These measurements would suggest that a threshold of significance, which is a function of intensity, should be used and that the threshold will vary from 1.35 for the most highly expressed genes to 2.33 for the least expressed genes. However, the measurements performed here are at best a surrogate system for estimating

error in the tumor data, i.e. the above experiments control for hybridization, filter and analysis variation but do not control for labeling and other sample-handling variation in the tumor samples. With limited tissue available for each tumor, it is not possible to perform replicate measurements on all our samples to generate similar significance curves for the actual data. Hence, we chose to use a ratio of 2.5 or more as the threshold of significance for our tumor data. We recognize that this criterion will result in the exclusion of genes that are differentially regulated at a statistically level. However, given that ourgoal is to develop genes that may serve as serum markers for ovarian cancer, and given the limitations of currently available assay systems for serum marker testing, a factor of 2.5 differential expression is appropriate.

3.2. Screening of 21 500 ovarian clones

An ideal array of cDNAs would contain a single copy of every gene expressed by the tissues to be compared. Since the identification of all human genes is incomplete, we chose to array randomly selected cDNAs derived from a wide spectrum of ovarian tissues including normal ovarian epithelium, early stage ovarian carcinomas, and late-stage pathologically aggressive ovarian carcinomas. We chose to array 8600 clones in form of purified plasmids from an OSE library [short-term culture of ovarian surface epithelial cells (Karlan et al., 1995)], and 3225 each from four ovarian cancer cDNA libraries from increasing malignancy, totaling 21 500 arrayed clones. We created 16 replicate sets of these arrays, each set consisting of 14 membranes of 7×12 cm holding 1536 clones. Each of the membrane sets was hybridized with a ^{32}P-labeled first-strand cDNA probe made from the RNA of an early-stage serous ovarian tumor (T007), eight late-stage serous ovarian tumors (T004, T008, T009, T010, T011, T014, T015, T016), one recurrent ovarian tumor (T012), ovarian surface epithelium (N001S), liver, placenta, bone marrow, cerebellum, and kidney. Two types of comparative experiments were carried out: (1) normal and tumor ovarian tissues were contrasted, and (2) ovarian tissues were compared against a variety of normal tissues. The first comparisons would reveal the tumor-specific cDNAs and the second the ovarian-specific cDNAs (at least with respect to the five different normal tissues). It was not our purpose to analyze early-to-late stage differences or tumor stratification as the limited number of cancerous tissues would not allow this. Our objective was to determine whether it is possible to use this technique to detect genes that are overexpressed in ovarian carcinomas respective to normal ovary and other tissues.

3.3. Differential transcript expression

Using the spot-finding and detection software developed in our laboratory, we determined the hybridization intensities for each clone and calculated their ratios. Comparing the 10 hybridizations with ovarian tumor tissues to those with OSE, the vast majority (>93%) of the clones displayed tumor-to-OSE ratios of less than a factor 2.5, and therefore were considered unchanged; about 7% of the clones exhibited a tumor-to-OSE ratio of more than 2.5, 0.9% a ratio of greater than 5.0, and 0.5% a ratio of greater than 10.0. Thus, most transcripts were expressed at similar levels in normal and tumor tissues, a finding that has been reported in colorectal and pancreatic cancers (Zhang et al., 1997).

No clone exhibited a 2.5-fold difference in expression in more than six of the ovarian tumors relative to OSE. Given the difference in tumor stages (one was an early stage tumor, and one a recurrent late stage tumor, the rest being late-stage ovarian adenocarcinomas) and the fact that the same stages, if they represent different stratified types, do not necessarily reflect high degrees of similarity on the molecular level; given the inter- and intra-tissue heterogeneity (possible proximity of section to areas of necrosis, differences in histology and pathology between tumors and across tumor sample), we did not expect to see a particular clone exhibit high tumor-to-OSE ratios in all tumors.

Sixteen clones showed overexpression in at least six ovarian cancers, but 14 of these 16 were also expressed in at least one non-ovarian tissue. In order to obtain a reasonable number of clones with overexpression in ovarian tumors and not in non-ovary tissues, we chose clones that fulfilled the following criteria: ratios greater than 2.5 in at least five out of the 10 tumors compared to OSE, and ratios below 2.5 in bone marrow, cerebellum, kidney, liver, and placenta compared to OSE. We were able to identify 134 clones that fulfilled these criteria. Sequencing of the partial cDNA clones revealed 60 that matched sequences in the non-redundant (nr) GenBank database. Of these, 17 matched to mitochondrial and ribosomal genes, and 43 matched to 37 other characterized genes (Table 1). Forty-seven clones matched only to sequences in the EST database, and 24 clones did not match any sequence in GenBank and were classified as novel. Three clones of 254, 312 and 323 bp length matched entirely to SINE and LINE sequences and were thus classified as repeats (see Table 1).

The expression patterns of two of these clones, which code for *S-adenosyl homocysteine* hydrolase and *HE4*, are shown in Fig. 2. For both genes, the calculated overexpression by signal intensities in the cancer tissues can be confirmed by visual inspection of the hybridized membranes. It is obvious, however, that by visual inspection alone, these clones would have probably escaped our scrutiny since their expression is rather weak compared to neighboring clones.

The overexpression of the 17 clones with similarity to mitochondrial sequences and ribosomal proteins can

Table 1
Categories of cDNAs present in the 134 clones[a]

Number of sequences	Percentage	Sequence similarity
3	2	Repeats
6	4	Mitochondrial sequences
2	2	Ribosomal RNA
9	7	Ribosomal proteins
24	18	Novel sequences
47	35	ESTs (expressed sequence tags)
43	32	Known genes
134		Total

[a] Novel sequences had less than 60% similarity to either human or non-human sequences. Repeats: genomic, SINE (ALU, MIR) and LINE (LINE1 and LINE2), LTR elements (MaLRs, Retroviral, MER4 group), DNA elements (MER1, MER2, Mariners). GenBank Accession Nos of the clones with similarity to known genes: *14.3.3*, X56468 (2×); *Actin capping protein*, U03269; *alpha-enolase*, M14328; *beta-actin*, M10277; *beta-2 microglobulin*, M17987; *BA46*, U58516; *Catechol-O-methyltransferase*, M65212; *CD44*, L05412; *CLIP/Restin*, M97501/X64838; *E16*, M80244; *Elongation factor 1 beta*, X60489; *Elongation factor 1 gamma*, Z11531 (2×); *Elongation factor 2*, Z11692; *Flightless*, U01184; *HE4*, X63187 (2×); *Initiation factor 4AI*, D13748; *Insulin-like growth factor BP 3 precursor*, M31159; *MDC15*, U46005; *Mucin*, X52229; *Myosin*, M22918; *Oviductal glycoprotein*, U09550 (3×); *p84*, L36529; *Peroxisomal targeting signal receptor I*, U19721; *Phosphatidyl inositol-3-kinase* alpha subunit, M61906; *Poly-A binding protein*, Y00345; *Procollagen alpha COL1A2*, K01078; *putative Progesterone binding protein*, Y12711; *Proteasome subunit HC8*, D00762; *RhoA*, L25080; Ryudocan, D13292; *S-adenosyl-homocysteine hydrolase*, M61831; *Smooth muscle protein*, M95787; *Tenascin precursor*, X56160; *Thiol-specific antioxidant*, Z22548; *Thymosine beta 4* (interferon-inducible), M17733; *Tropomyosin*, M75165; *Ubiquitin*, M10939, X56997 (2×).

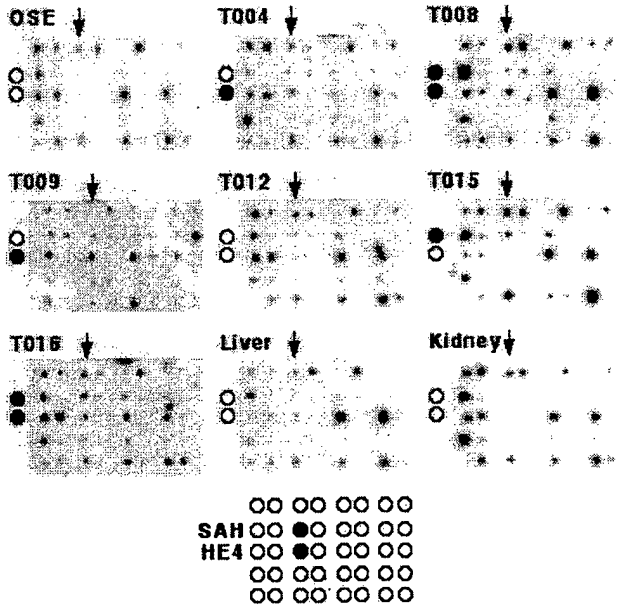


Fig. 2. Visual inspection of the membranes after identification of differentially expressed clones through computing the signal intensities. Displayed are eight columns and five rows of close-ups on nine membranes that have been hybridized with nine different probes. Two clones, coding for *HE4* and *S-adenosyl homocysteine hydrolase* (*SAH*), show high tumor-to-OSE ratios and low normal-to-OSE ratios (see Table 2). The positions of these two clones on the array are marked in the bottom panel. The clone positions on the nine membranes are indicated by an arrow on top of each panel and the circles on the left. An empty circle denotes a weak hybridization signal, and a filled circle denotes a strong signal.

be attributed to the higher metabolic activity of the tumors. Ribosomal protein sequences have been found to be more highly expressed in colon carcinomas (Pogue-Geile et al., 1991). Likewise, five other genes linked to metabolic pathways such as *elongation factor 1 gamma* and *initiation factor 4AI* were overexpressed in ovarian cancer tissues. It is notable that these 22 clones displayed an average tumor-to-OSE ratio of 5.22 ± 2.4, whereas the remaining 38 clones with homology to known genes had a lower average ratio of 4.11 ± 1.8. This underscores the fact that the degree of overexpression alone is not necessarily indicative of a clone that can be used as a marker protein.

In order to estimate the quality of the HDAH in identifying cancer related genes, and since we were realistically capable of processing only a limited number of clones, we focused on the 43 previously characterized clones, as opposed to the 47 clones that match only ESTs or those 24 that do not match any sequence in GenBank. Of the 43 clones with homology to the 37 characterized genes, 10 genes are expressed in epithelial tissues: *14.3.3*, *BA46*, *CD44*, *HE4*, *Mucin1*, *Oviductal glycoprotein*, *Collagen COL1A2*, *Putative progesterone binding protein*, *RhoA*, and *Ryudocan*(GenBank Accession Nos listed in Table 1). This coincides with the fact that the vast majority of ovarian cancers, including all those used for HDAH, arise from the ovarian surface epithelium (Berchuck et al., 1996).

Thirteen of the 37 genes (35%) are known to be overexpressed in various cancers, including lung, breast and colon. Six of these 13 are expressed in ovarian carcinomas, their expression not being restricted to ovarian tissues. The thirteen genes are *14.3.3* [lung cancer (Nakanishi et al., 1997)], *beta-actin* [AML (Blomberg et al., 1987) and colorectal carcinomas (Naylor et al., 1992)], *BA46* [breast cancer (Couto et al., 1996)], *CD44* [ovarian cancer cell lines (Stickeler et al., 1997)], *Clip/Restin* [Hodgkin disease and anaplastic large-cell lymphoma (Delabie et al., 1992)], *Collagen COL1A2* [ovarian cystadenoma (Kauppila et al., 1996)], *E16* [colorectal carcinoma, adenocarcinomas from breast and endometrium (Wolf et al., 1996)], *Insulin-like growth factor BP 3* [breast cancer (Ng et al., 1998)], *Mucin1* [epithelial ovarian cancer (Dong et al., 1997)], *Procollagen-alpha* [ovarian cystadenocarcinoma (Kauppila et al., 1996)], *putative Progesterone binding protein* [ovarian cancer (Isola et al., 1990)], *RhoA proto oncogene* [*ras* activation (Khosravi-Far et al., 1995)], and *MDC15*, a metalloprotease [some metalloproteases are elevated in ovarian tumor cell cultures (Fishman et al., 1997)]. These findings indicate that our approach

is indeed capable of narrowing down the pool of 21 500 randomly selected clones to a few epithelium- and cancer-related genes.

3.4. Confirmation of overexpression of four selected clones by RT-PCR-based transcript quantitation

Any clone with its expression restricted to ovarian carcinomas can be potentially used as a marker without knowing its function. Early detection of ovarian cancer, however, requires that the assay be suitable for routine screening of women, which means that it must be affordable, non-invasive and with a high degree of specificity. Only a serum-based assay can deliver this. Therefore, knowing whether a protein is secreted or membrane-bound maximizes the chance that the protein or its degradation product will be found in the blood either as freely circulating protein or bound to the membrane of a cell that has detached from the tumor. In both cases, an antibody can be used to detect the protein in the blood. A circulating cancer cell can be detected by an RT-PCR assay or fluorescence-activated cell sorting.

In an attempt to find out whether one of the 43 clones that match characterized genes would be a potential candidate for a marker protein in a serum-based assay, we examined which of the clones codes for a cell surface protein such as Her2/neu, used as a target in breast cancer treatment (Baselga et al., 1998) or a secreted protein such as Prostate Specific Antigen (PSA) which is used in prostate cancer diagnosis (Rittenhouse et al., 1998).

From the 43 clones with homology to the 37 known genes, we chose five that are expressed at the cell surface (*progesterone binding protein*, *ryudocan*, *mucin1*, *E16*, *BA46*) and one which is secreted (*HE4*). In addition, we included the gene *14.3.3*, which is expressed in the cytoplasm but which, like *HE4*, appeared twice in our selected clones list. *Beta actin* is often used as a control for quantitative analyses because of its assumed uniformity in expression in a large array of tissues. Our HDAH results suggest, however, that *beta actin* is differentially expressed in some ovarian tumors. We therefore chose to verify *beta actin* expression as well. The characteristics of the eight chosen genes are summarized in Table 2. We used RT-PCR-based transcript quantitation to confirm overexpression in tumors relative to normal tissues.

Due to the small size of our tumor specimens (ranging from 200 to 400 mg per tissue), the RNA preparations used in the array hybridization were exhausted during library construction and probe preparation. Therefore, new ovarian adenocarcinomas matching the stage and grade of the original tumors were used for the RT-PCR analysis. We chose one early-stage, low-grade mucinous ovarian adenocarcinoma (T037) five late-stage, high-grade serous ovarian adenocarcinomas (T001–T006 and T021) and two metastatic ovarian serous adenocarcinomas (T013M and T026M). In order to incorporate different tumor histologies, we included two benign ovarian tissues (T017B and T018B) as well as a borderline ovarian tumor tissue (T028L). In addition, we tested the expression in four normal ovaries (N002, N005, N006 and N019), in a pool of fetal ovaries and in two batches of peripheral blood lymphocytes (PBL1 and PBL2). The reason for analyzing the expression patterns of these genes in peripheral blood lymphocytes is to determine whether they are expressed in blood elements, for if they are, they would not be good candidates for a diagnostic probe in blood samples. The OSE, as well as the liver and placental tissue were the same as used for array hybridization. As a control for the quality of the RNA template, we included a gene that we found to be expressed at high levels in all tissues tested so far, *S31iii125* (GenBank Accession No. U61734, Trower et al., 1996).

Fig. 3 shows the results of the RT-PCR. The quantitated intensities of the PCR bands are summarized in Table 2. While trying to match the tumor tissues in stage and grade, we did not expect an exact reproduction of the ratios from the HDAH analysis. In spite of these shortcomings, we were able to reproduce the tumor-to-OSE ratios observed in the HDAH for seven out of the eight genes, albeit only qualitatively. For the gene *14.3.3*, the tumor-to-OSE ratios were low but still measurable. This discrepancy can be attributed to the difference in tumor samples used or to an erroneous reading of the HDAH signals. For three genes (*BA46*, *E16* and *Ryudocan*), a high placenta-to-OSE ratio stands in discordance with the HDAH results where they had been low. Since the placental RNA used in both cases was the same, and since our quadruple RT-PCR approach is more accurate than the HDAH method, we must conclude that in the HDAH, the placental values must have been misread for these three clones.

14.3.3 shows no tumor-to-OSE ratios above the threshold of significance of 2.5. It displays a mean ratio of 1.5 in four invasive and in one benign ovarian tumor, which does not compare well with the mean ratio of 4.4 determined in the HDAH.

BA46 shows tumor-to-OSE ratios above 2.5 in five tumors but also in one normal ovary and in placenta. In spite of its low expression in PBL (which, as noted in the beginning of this section, is a prerequisite for a serum marker), the relatively low mean ratios in RT-PCR and HDAH of 3.2 make it a second choice marker gene.

Beta actin shows tumor-to-OSE ratios above 2.5 in 10 out of the 12 tumors (a mean of 3.9 compared to 4.4 in the HDAH), but also in some normal tissues, including PBL. Although these numbers do not warrant the consideration as a tumor marker gene, they give cause to question the use of *beta actin* as a normalization standard.

Table 2
HDAH (top) and RT-PCR ratios (bottom) of nine selected genes[a]

Gene name	14.3.3	14.3.3	BA46	β-actin	E16	HE4	HE4	Mucin1	ProgBP	Ryu
Accession No.	X56468	X56468	U58516	X00351	M80244	X63187	X63187	X52229	Y12711	D13292
Protein	Cytopl.	Cytopl.	Membr	Cytopl.	Membr	Secreted	Secreted	Membr.	Membr	Membr
T004				3.1		5.1	3.6		8.9	3.1
T007 early	5.9	3.7			5.5	3.0	2.7	2.6		
T008			4.1			5.1	4.9	2.6	9.9	5.0
T009	2.7			8.5		5.1	5.5		2.8	
T010	5.5	4.3	2.7	2.5						
T011			2.6		3.0		2.7			
T012 recur						2.8		8.0	3.6	
T014	6.8	3.0		4.2	5.2			5.9	7.3	
T015	6.0	2.7	2.7		4.5				2.9	4.1
T016		3.1	4.0	3.6	3.7	2.5	2.6	8.4		4.6
Liver				3.8						2.5
Placenta			2.5	3.4	9.8				1.3	2.1
PBL1	1.3	1.3		3.9					5.8	2.1
PBL2				4.7	3.9			2.0	4.1	2.3
Fetal	2.8	2.8	1.9	1.3	8.9	7.9	7.9	9.7	4.9	1.9
N002			2.1	3.8	2.4			3.7	1.8	1.4
N005			1.3	3.4	6.5	1.2	1.2	2.4	2.1	1.0
N006			2.6	3.5	4.5	2.0	2.0	2.3	4.1	0.6
N019									3.2	0.3
N035			3.1	4.2				2.1		1.3
T017B	1.6	1.6	1.6	2.7	2.0	6.6	6.6	2.3	2.6	5.3
T018B			2.5	3.6	3.3	8.8	8.8	2.8	7.8	
T028L				2.9	1.1	8.2	8.2	3.2	1.3	4.9
T037 early			1.9	3.2	1.4	1.6	1.6	7.9	3.1	4.4
T002			1.0	3.0	1.6	12.0	12.0	2.4	2.7	3.3
T003	1.6	1.6	3.7	3.1	1.5	16.0	16.0	1.9	4.7	3.7
T005			3.0	3.6	9.4	17.0	17.0	2.4	2.0	4.8
T006				2.5		9.7	9.7		2.8	2.1
T001	1.3	1.3	1.1	2.4	2.2	11.4	11.4	2.4		2.7
T021			3.0	5.7	1.2	12.3	12.3	2.5	2.1	3.7
T013M	1.9	1.9		5.2	3.4	14.1	14.1	7.3	9.7	1.4
T026M	1.2	1.2	4.0	5.8	1.9	2.8	2.8	4.5	3.5	

[a] Eight genes out of the 43 clones that match to 37 known genes were validated for their expression by RT-PCR (see Fig. 3). The volumes of the PCR bands were calculated using the software QuanitityOne (BioRad, Hercules, CA). Titan RT-PCR amplifies the template semiquantitatively; therefore, the numbers in this table are merely indicative of a tendency and cannot be translated into copy numbers. The rows show the gene name, GenBank Accession No., protein localization, 10 tumor-to-OSE ratios that were observed in the HDAH (only ratios above 2.5; normal-to-OSE ratios are omitted for they lied all below 2.5), followed by 22 tissue-to-OSE ratios determined in the RT-PCR (for clarity, only ratios above 1 are displayed). The columns are duplicated for *14.3.3.* and *HE4* because two clones were selected for them by HDAH. *Putative Progesterone binding protein*(ProgBP): progesterone binding proteins can be found in low-grade breast cancers and in some ovarian cancer cell lines. The homologous rat sequence has a transmembrane region (Falkenstein et al., 1996), indicating that our clone might also be membrane-bound. *Ryudocan* (abreviated as Ryu.) is a cell-surface proteoglycan with a transmembrane domain; it is expressed in an extensive array of human tissues (Kojima et al., 1993). *HE4* is an epidermal, epididymis-specific protease inhibitor that is thought to be involved in the maturation of spermatozoa (Kirchhoff et al., 1991). The putative *HE4* protein has a leader sequence and it is speculated that it is secreted. *Mucin1* (Dong et al., 1997) is expressed on the cell surface of non-mucinous ovarian tumors with either low malignant or invasive potential. *14.3.3* codes for a cytosolic protein kinase regulator protein that shows elevated expression levels in lung cancer tissues (Nakanishi et al., 1997). *BA46*, also known as lactadherin, is a cell-surface protein expressed in human breast carcinomas. It has been used successfully as a target for experimental breast cancer radioimmunotherapy (Couto et al., 1996). *Beta actin* is a cytoskeletal protein with differential expression in acute myelolytic leukemia (Blomberg et al., 1987) and high expression in colorectal carcinomas (Naylor et al., 1992). *E16* codes for an integral membrane protein that was isolated from peripheral blood lymphocytes (Gaugitsch et al., 1992). It is expressed in colorectal and other human carcinomas (Wolf et al., 1996).

E16 shows tumor-to-OSE ratios above 2.5 in three tumors (with a mean of 5.3 compared to a mean of 4.4 in the HDAH). It also shows high ratios for two normal ovaries and placenta. The low expression in PBL and the high average ratios for the tumors make it a possible marker candidate.

HE4 shows a clear tumor-restricted expression, making its pattern resemble that in the HDAH. Most importantly, the results suggest that it is not expressed in peripheral blood lymphocytes. As noted in the beginning of this section, this accordingly represents a candidate for a serum marker assay. The difference in the mean rates of overexpression measured by RT-PCR (11×) and HDAH (4.1×) can be attributed either to

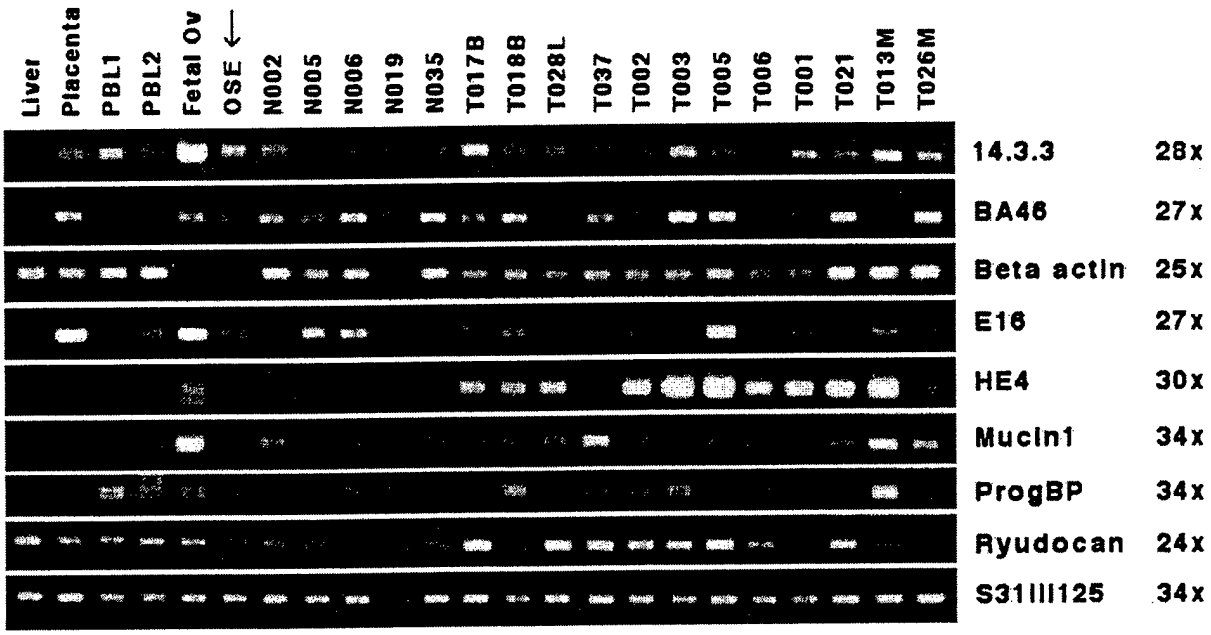


Fig. 3. Expression monitoring by RT-PCR. Eight genes (plus one control) are tested in 23 tissues. Tissue names are on top; OSE is marked with an arrow. Tissues starting with an N are normal ovaries, and those starting with a T are ovarian tumors. The *S31iii125* gene serves as a control. The number of PCR cycles is indicated behind each gene name (on the right). ProgBP stands for 'putative progesterone binding protein'. The PCR bands are in the range of 420–660 bp. All reactions had been performed in four parallel sets with one set shown here.

the better signal-to-noise ratio in the RT-PCR or to the different tumor samples used.

Mucin 1 shows a high RT-PCR value in fetal ovaries, suggesting that this might be a fetal gene that is re-expressed in the tumor. It shows strong bands in three out of the 12 tumors, two of them metastatic and one an early stage tumor, resulting in a mean tumor-to-OSE ratio of 3.0. This result correlates with that of the array hybridization.

The *Putative progesterone binding protein* shows a high tumor-to-OSE ratio for only two tumors, one being similar in stage to a tumor used in the HDAH. All other tumors show medium high ratios but so do the normal tissues, including the PBL. The strong expression in the metastasizing tumor may indicate a role as a marker for tumor staging, prognosis or stratification.

The transcript of *ryudocan* displays a similar pattern of expression as *HE4*, and the mean the tumor-to-OSE ratio of 4.3 are is slightly higher than the one determined by HDAH (where it was 6). The presence of *ryudocan* mRNA in liver, PBL and placenta means that the protein might normally be found in the blood, thus making it a less suitable marker candidate.

3.5. Confirmation of overexpression of HE4 by Northern blot analysis

Of the eight genes tested in the RT-PCR, only *HE4* shows a clear tumor-restricted expression pattern. To further confirm the cancer-restricted expression of *HE4*, we used a Northern blot (Northern Territory™, Invitrogen, San Diego, CA) that contained total RNA from ovaries from four patients who had unilateral ovarian cancer. RNA from both the affected and the unaffected ovary was present on the blot (loaded adjacent to each other). Fig. 4 shows that *HE4* is expressed in two ovarian carcinomas but not in the matching normal ovaries. *HE4* cannot be detected in the tumors nor in the normal ovaries of two other patients. The ratios of *HE4* expression between the unaffected and

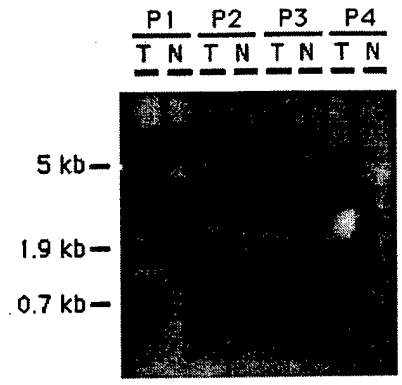


Fig. 4. Northern hybridization of *HE4*. The Northern blot contains RNA from ovarian tumor and matching non-affected ovary from four patients. *HE4* is expressed in two tumors but not in the normal tissue of the same patient. A Digoxigenin-labeled riboprobe was prepared from a 500 bp *HE4* PCR product cloned in a vector. The probe was hybridized over night at 68°C and washed in 2× SSC, 0.1% SDS for 15 min at room temperature; 2× SSC, 0.1% SDS for 20 min at 68°C; 0.1× SSC, 0.1% SDS for 2×15 min at 68°C. The hybridized RNA was visualized using the DIG detection kit (Boehringer Mannheim). The membrane was exposed to X-ray film for 15 min.

the affected ovary was 6.1 for patient 1 and 4.5 for patient 2. Thus, *HE4* is also a candidate for a tumor-staging, prognosis or stratification marker.

3.6. Conclusion

From the 21 500 clones, we chose 43 that were overexpressed in ovarian tumors by HDAH with homology to characterized genes. We chose eight genes for expression validation by RT-PCR. From these eight, seven genes displayed tumor-to-OSE ratios similar to those measured in the HDAH, albeit with different tumor tissues matching grade and stage. Seven of these eight display expression in normal tissues; only *HE4* showed a clear tumor-restricted expression pattern. We conclude that the *HE4* message is significantly overexpressed in a variety of ovarian tumors relative to normal tissues or OSE, thus making it a potential candidate for a marker protein.

The results support the validity of using HDAH combined with a second quantitation method for the identification of genes that are overexpressed in cancers as compared to normal tissues. We are preparing an antibody against *HE4* to further analyze whether it indeed could be a diagnostic marker for ovarian cancer.

Acknowledgements

This work was supported by the Stowers Institute and by grants from the Deutsche Forschungsgemeinschaft, the National Institutes of Health (5RO1HG01713-02) the Marsha Rivkin Center for Ovarian Cancer Research and the National Science Foundation (BIR9214821/9423347). We would like to thank the members of the sequencing group in our laboratory.

References

American Cancer Society, 1998. Cancer Facts and Figures. American Cancer Society, Atlanta, GA.

Baselga, J., Norton, L., Albanell, J., Kim, Y.M., Mendelsohn, J., 1998. Recombinant humanized anti-HER2 antibody (Herceptin) enhances the antitumor activity of paclitaxel and doxorubicin against HER2/neu overexpressing human breast cancer xenografts. Cancer Res. 58, 2825–2831.

Berchuck, A., Kohler, M.F., Bast Jr., R.C., 1996. Molecular genetic features of ovarian cancer. Prog. Clin. Biol. Res. 394, 269–284.

Bishop, J.O., Morton, J.G., Rosbash, M., Richardson, M., 1974. Three abundance classes in HeLa cell messenger RNA. Nature 250, 199–204.

Blomberg, J., Andersson, M., Faldt, R., 1987. Differential pattern of oncogene and beta-actin expression in leukaemic cells from AML patients. Br. J. Haematol. 65, 83–86.

Couto, J.R., Taylor, M.R., Godwin, S.G., Ceriani, R.L., Peterson, J.A., 1996. Cloning and sequence analysis of human breast epithelial antigen BA46 reveals an RGD cell adhesion sequence presented on an epidermal growth factor-like domain. DNA Cell Biol. 15, 281–286.

Delabie, J., Shipman, R., Brüggen, J., De Strooper, B., van Leuven, F., Tarcsay, L., Cerletti, N., Odink, K., Diehl, V., Bilbe, G., et al., 1992. Expression of the novel intermediate filament-associated protein restin in Hodgkin's disease and anaplastic large-cell lymphoma. Blood 80, 2891–2896.

Dong, Y., Walsh, M.D., Cummings, M.C., Wright, R.G., Khoo, S.K., Parsons, P.G., McGuckin, M.A., 1997. Expression of MUC1 and MUC2 mucins in epithelial ovarian tumours. J. Pathol. 183, 311–317.

Falkenstein, E., Meyer, C., Eisen, C., Scriba, P.C., Wehling, M., 1996. Full-length cDNA sequence of a progesterone membrane-binding protein from porcine vascular smooth muscle cells. Biochem. Biophys. Res. Commun. 229, 86–89.

Fishman, D.A., Bafetti, L.M., Banionis, S., Kearns, A.S., Chilukuri, K., Stack, M.S., 1997. Production of extracellular matrix-degrading proteinases by primary cultures of human epithelial ovarian carcinoma cells. Cancer 80, 1457–1463.

Gaugitsch, H.W., Prieschl, E.E., Kalthoff, F., Huber, N.E., Baumruker, T., 1992. A novel transiently expressed, integral membrane protein linked to cell activation. Molecular cloning via the rapid degradation signal AUUUA. J. Biol. Chem. 267, 11267–11273.

Isola, J., Kallioniemi, O.P., Korte, J.M., Wahlstrom, T., Aine, R., Helle, M., Helin, H., 1990. Steroid receptors and Ki-67 reactivity in ovarian cancer and in normal ovary: correlation with DNA flow cytometry, biochemical receptor assay and patient survival. J. Pathol. 162, 295–301.

Jacobs, I., Davies, A.P., Bridges, J., Stabile, I., Fay, T., Lower, A., Grudzinskas, J.G., Oram, D., 1993. Prevalence screening for ovarian cancer in postmenopausal women by CA 125 measurement and ultrasonography [see comments]. Br. Med. J. 306, 1030–1034.

Karlan, B.Y., Jones, J., Greenwald, M., Lagasse, L.D., 1995. Steroid hormone effects on the proliferation of human ovarian surface epithelium in vitro. Am. J. Obstet. Gynecol. 173, 97–104.

Kauppila, S., Saarela, J., Stenback, F., Risteli, J., Kauppila, A., Risteli, L., 1996. Expression of mRNAs for type I and type III procollagens in serous ovarian cystadenomas and cystadenocarcinomas. Am. J. Pathol. 148, 539–548.

Khosravi-Far, R., Solski, P.A., Clark, G.J., Kinch, M.S., Der, C.J., 1995. Activation of Rac1, RhoA, and mitogen-activated protein kinases is required for Ras transformation. Mol. Cell. Biol. 15, 6443–6453.

Kirchhoff, C., Habben, I., Ivell, R., Krull, N., 1991. A major human epididymis-specific cDNA encodes a protein with sequence homology to extracellular proteinase inhibitors. Biol. Reprod. 45, 350–357.

Kojima, T., Inazawa, J., Takamatsu, J., Rosenberg, R.D., Saito, H., 1993. Human ryudocan core protein: molecular cloning and characterization of the cDNA, and chromosomal localization of the gene. Biochem. Biophys. Res. Commun. 190, 814–822.

Nakanishi, K., Hashizume, S., Kato, M., Honjoh, T., Setoguchi, Y., Yasumoto, K., 1997. Elevated expression levels of the 14-3-3 family of proteins in lung cancer tissues. Hum. Antibodies 8, 189–194.

Naylor, M.S., Stamp, G.W., Balkwill, F.R., 1992. Beta actin expression and organization of actin filaments in colorectal neoplasia. Epithelial Cell Biol. 1, 99–104.

Ng, E.H., Ji, C.Y., Tan, P.H., Lin, V., Soo, K.C., Lee, K.O., 1998. Altered serum levels of insulin-like growth-factor binding proteins in breast cancer patients [in process citation]. Ann. Surg. Oncol. 5, 194–201.

Ng, W.L., Schummer, M., Cirisano, F., Baldwin, R.L., Karlan, B.Y., Hood, L., 1996. High-throughput plasmid minipreparations facilitated by micro-mixing. Nucleic Acids Res. 24, 5045–5047.

Piétu, G., Alibert, O., Guichard, V., Lamy, B., Bois, F., Leroy, E.,

Mariage-Samson, R.R.H., Soularue, P., Auffray, C., 1996. Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. Genome Res. 6, 492–503.

Pogue-Geile, K., Geiser, J.R., Shu, M., Miller, C., Wool, I.G., Meisler, A.I., Pipas, J.M., 1991. Ribosomal protein genes are overexpressed in colorectal cancer: isolation of a cDNA clone encoding the human S3 ribosomal protein. Mol. Cell. Biol. 11, 3842–3849.

Rittenhouse, H.G., Finlay, J.A., Mikolajczyk, S.D., Partin, A.W., 1998. Human Kallikrein 2 (hK2) and prostate-specific antigen (PSA): two closely related, but distinct, kallikreins in the prostate [in process citation]. Crit. Rev. Clin. Lab. Sci. 35, 275–368.

Schena, M., Heller, R.A., Theriault, T.P., Konrad, K., Lachenmeier, E., Davis, R.W., 1998. Microarrays: biotechnology's discovery platform for functional genomics. Trends Biotechnol. 16, 301–306.

Schummer, B., Hauptfleisch, S., Siegsmund, M., Schummer, M., Lemmer, B., 1998. Highly accurate quantification of mRNA expression by means of Titan® One Tube RT-PCR and capillary electrophoresis. Biochemica 2, 31–33.

Schummer, M., Ng, W.-L., Nelson, P.S., Bumgarner, R.B., Hood, L., 1997. A simple high-performance DNA arraying device for comparative expression analysis of a large number of genes. BioTechniques 23, 1087–1092.

Stickeler, E., Runnebaum, I.B., Möbus, V.J., Kieback, D.G., Kreienberg, R., 1997. Expression of CD44 standard and variant isoforms v5, v6 and v7 in human ovarian cancer cell lines. Anticancer Res. 17, 1871–1876.

Trower, M.K., Orton, S.M., Purvis, I.J., Sanseau, P., Riley, J., Christodoulou, C., Burt, D., See, C.G., Elgar, G., Sherrington, R., Rogaev, E.I., St George-Hyslop, P., Brenner, S., Dykes, C.W., 1996. Conservation of synteny between the genome of the pufferfish (*Fugu rubripes*) and the region on human chromosome 14 (14q24.3) associated with familial Alzheimer disease. Proc. Natl. Acad. Sci .USA 93, 1366–1369.

Wolf, D.A., Wang, S., Panzica, M.A., Bassily, N.H., Thompson, N.L., 1996. Expression of a highly conserved oncofetal gene, TA1/E16, in human colon carcinoma and other primary cancers: homology to *Schistosoma mansoni* amino acid permease and *Caenorhabditis elegans* gene products. Cancer Res. 56, 5012–5022.

Zhang, L., Zhou, W., Velculescu, V.E., Kern, S.E., Hruban, R.H., Hamilton, S.R., Vogelstein, B., Kinzler, K.W., 1997. Gene expression profiles in normal and cancer cells. Science 276, 1268–1272.

Tissue Classification with Gene Expression Profiles

Amir Ben-Dor *
U. Washington

Laurakay Bruhn
HP Laboratories

Nir Friedman
Hebrew University

Iftach Nachman
Hebrew University

Michèl Schummer
U. Washington

Zohar Yakhini
HP Laboratories

September 30, 1999

Abstract

Constantly improving gene expression profiling technologies are expected to provide understanding and insight into cancer related cellular processes. Gene expression data will also significantly aid in the development of efficient cancer diagnosis and classification platforms. In this work we examine two sets of gene expression data measured across sets of tumor and normal clinical samples. One set consists of 2,000 genes, measured in 62 epithelial colon samples [1]. The second consists of $\approx$ 100,000 clones, measured in 32 ovarian samples [24, 25].

We examine the use of scoring methods, measuring separation of tumors from normals using individual gene expression levels. These are then coupled with high dimensional classification methods to assess the classification power of complete expression profiles. We present results of performing *leave-one-out cross validation* (LOOCV) experiments on the two data sets, employing SVM [6], AdaBoost [12] and novel clustering based classification techniques. As tumor samples can differ from normal samples in their cell-type composition we also perform LOOCV experiments using appropriately modified sets of genes, eliminating the resulting bias.

We demonstrate success rate of at least 90% in tumor vs normal classification, using sets of selected genes, with as well as without cellular contamination related members. These results are insensitive to the exact selection mechanism, over a certain range.

*Contact author. Email: amirbd@cs.washington.edu.

1 Introduction

The process by which the approximately 100,000 genes encoded by the human genome are expressed as proteins involves two steps. First, DNA sequences are transcribed into mRNA sequences which in turn are translated into the amino acid sequences of the proteins that perform various cellular functions. A crucial aspect of proper cell function is that the gene expression process is regulated such that different cell types express different subsets of genes. Measuring mRNA levels can provide a detailed molecular view of the subset of genes expressed in different cell types. Recently, array-based methods have been developed that enable simultaneous measurements of the expression levels of thousands of genes. These measurements are made by quantitating the hybridization (detected for example, by fluorescence) of a cellular mRNA mixture to an array of defined cDNA or oligonucleotide sequences immobilized on a solid substrate. Array methodologies have led to a tremendous acceleration in the rate at which gene expression pattern information is accumulated [14, 17, 7, 28, 15]. Measuring gene expression levels under different conditions is important for expanding our understanding of gene function, how various gene products interact, and how experimental treatments can affect cellular function.

One of the promising usages of gene expression measurements, is the understanding of cancer. Normal cells can evolve into malignant cancer cells through a series of mutations in genes that control the cell cycle, apoptosis, and genome integrity, to name only a few. As determination of cancer type and stage is often crucial to the assignment of appropriate treatment [10], a central goal is the identification of sets of genes that can serve, via expression profiling assays, as classification or diagnosis platforms.

Another important application of these tools, is the understanding of cellular responses to drug treatment. Expression profiling assays performed before, during and after treatment, are aimed at identifying drug responsive genes, indications of treatment outcomes, and at identifying potential drug targets [5]. More generally, complete profiles can be considered as a potential basis for classification of treatment progression or other trends in the evolution of the treated cells.

Data obtained from such studies typically consists of expression level measurements of thousands of genes. This complexity calls for data analysis methodologies that will efficiently aid in extracting relevant biological information. Previous gene expression analysis work emphasizes clustering techniques, which aim at partitioning the set of genes into subsets that are expressed similarly across different conditions. Indeed, such clustering has been demonstrated to identify functionally related families of genes [2, 7, 4, 13, 28, 9]. Similarly, clustering methods can be used to divide a set of cell samples into clusters based on their expression profile. In [1] this approach was applied, and a set of colon samples was divided into two groups, one containing mostly tumor samples, and the other containing mostly normal tissue samples.

Clustering methods, however, do not use any tissue annotation (e.g., tumor vs.normal) in the partitioning step. This information is used only afterward, to asses the success of the method. Such methods are often referred to as *unsupervised*. In contrast, *supervised* methods, attempt to predict the classification of new tissues, based on their gene expression profiles after training on examples that have been classified by an external "supervisor".

The purpose of this work is to rigorously assess the potential of classification approaches on gene expression data. We present a novel clustering based classification methodology, and apply it together with two other recently developed classification approaches, *Boosting* and *Support Vector Machines* to two data sets. Both sets involve corresponding tissue samples from tumor and normal biopsies. The first is the data set of colon cancer [1], and the other is a data set of ovarian cancer [24]. We use established statistical tools to evaluate the predictive power of these methods in the data sets. For this purpose we use *leave one out cross validation* (LOOCV), a well known method for estimating classification accuracy.

One of the major challenges of gene expression data is the large number of genes in the data sets. For example, one of our data sets includes over 97,800 clones. Many of these clones are not relevant to the distinction between cancer and tumor and introduce noise in the classification process. Moreover, for diagnostic purposes it is important to find small sets of genes that are sufficiently informative to distinguish between tumors and normal cells. To this end we suggest a simple combinatorial error rate score for each gene, and use this method to select informative genes. As we show, selecting relatively small subsets of genes can drastically improve the performance. Moreover, this selection process also isolates genes that are potentially intimately related to the tumor makeup.

A major challenge in a realistic assessment of the performance of such methods, is *sample contamination*. Tumor and normal samples may dramatically differ in terms of their cell-type composition. For example, in the colon cancer data [1], the authors observed that the normal colon biopsy also included smooth muscle

tissue from the colon walls. As a result, smooth muscle related genes showed high expression levels in the normal samples compared to the tumor samples. This artifact, if consistent, could contribute to success in classification. To eliminate this effect we remove the muscle specific genes and observe the effect on the success rate of the process.

Very recently, Lander et al. [10] examine gene expression profile differences in AML and ALL (two types of leukemia) biopsies. They employ scoring methods to select informative genes and perform LOOCV experiments to test voting based classification approaches.

The rest of the paper is organized as follows. In Section 2, we describe the principle classification methods we use in this study. These include two state of the art methods from machine learning, and a novel approach based on clustering algorithm of [2]. In Section 3, we describe the two data sets, the LOOCV evaluation method, and evaluate the classification methods on the two data sets. In Section 4 we address the problem of gene selection. We propose a simple method for selecting informative genes and evaluate the effect of gene selection on the classification methods. In Section 5, we examine the effect of sample contamination on possible classification. We conclude in Section 6 with a discussion of related works and future directions.

2 Classification Methods

In this section, we describe the main classification methods that we will be using in this paper. We start by formally defining the classification problem. Assume that we are given a *training set* D, consisting of pairs $\langle x_i, l_i \rangle$, for $i = 1, \ldots, m$. Each *sample* x_i is a vector in $\mathbf{R}^N$ that describes expression values of N genes/clones. The *label* l_i associated with x_i is either -1 or $+1$ (for simplicity, we will concentrate on two-label classification problems). A classification algorithm is a function f that depends on two arguments, training set D, and a query $x \in \mathbf{R}^N$, and returns a predicted label $\hat{l} = f_D(x)$. Our aim in building good classification procedures is that the predicted labels will match the "true" label of the query.

2.1 Nearest Neighbor Classifier

One of the of the simplest classification algorithms is the *nearest neighbor* classifier [8]. The intuition is simple. To classify a query x, find the most similar example in D and predict that x has the same label as that example. To carry out this algorithm we need to define a similarity measure $s(x, y)$ on expression patterns. In our experiments, we use the Pearson correlation as a measure of similarity. Formally, the classification of the nearest neighbor procedure is by the rule

$$nn_D(x) = l_i \text{ s.t. } s(x, x_i) = max_j s(x, x_j)$$

(in situations where there are several nearest neighbors, we choose one of them arbitrarily).

This simple non-parametric classification method does not take any global properties of the training set into consideration. However, it is surprisingly effective in many types of classification problems. We use it in our analysis as a strawman, to which we compare the more sophisticated classification approaches.

2.2 Using Clustering for Classification

Recall, that clustering algorithms, when applied to expression patterns, attempt to partition the set of examples into clusters of patterns, so that all the patterns within a cluster are similar to each other, and different than patterns in other clusters. This suggests that if the labeling of patterns is correlated with the patterns, then the unsupervised clustering of the data (that does not take labels into account) would cluster patterns with the same label together and separate patterns with different labels.

Indeed, such a phenomenon is noted by Alon et al. [1] in their analysis of colon cancer. Their experiment (which we describe in more detail in Section 3), involves gene expression patterns from colon samples that include both tumors and normal tissues. They clustered patterns using a hierarchical clustering procedure (which is quite different from the one we discuss below). They note that the topmost division in the dendrogram they construct divides samples into two groups, one containing mostly tumor samples, and the other containing mostly normal tissue samples.

This suggests that for some types of classification problems, such as tumor vs. normal, clustering can distinguish among labels. Following this intuition, we build a classifier around a clustering algorithm. We first describe the clustering algorithm we use. Than, we present our clustering based classifier.

2.2.1 The clustering algorithm The BioClust algorithm [2], takes as input a threshold parameter t, which controls the granularity of the resulting clusters, and a similarity measure between the tissues[1]. We say that a tissue v has *high similarity* to a set of tissues C, if the average similarity between the v and the tissues in C is at least t. Otherwise, if the average similarity is below t, we say that v has *low similarity* to C.

BioClust constructs the clusters one at a time, and halts when all tissues are assigned to clusters. Intuitively, the algorithm alternates between adding high similarity tissues to C, and removing low similarity tissues from it. Eventually, all the tissues in C have high similarity to C, while all the tissues outside of C have low similarity to C. At this stage the cluster C is closed, and a new cluster is started (See [2] for complete description of the algorithm).

Clearly, the threshold value t, has great effect on the resulting clustering. As t increases, the clusters would get smaller. At the extreme case, if t is high enough, each tissue would form a different cluster. Similarly, as t decreases, the clusters tend to get larger. If t is low enough, all tissues would be assigned to the same cluster.

2.2.2 Clustering based classifier Applying clustering algorithms for classification raises two problems. First, how do we use clustering on training data to classify a new query and, second, how do we decide which "granularity" of clustering to use? We start with the second question, and then return to the first one.

As described above, the BioClust procedure has an input parameter that determines the confidence threshold in construction of clusters. By changing this parameter, we can get different numbers of clusters and different divisions into clusters. A similar situation occurs in other clustering algorithms. For example, in hierarchical clustering algorithms (e.g., [1, 9]) we can choose different numbers of clusters by selecting a "level" of the tree. In either clustering algorithms, it is clear that attempting to partition the data to exactly two clusters, will not be the optimal choice for predicting labels. For example, if the tumor class consists of several types of tumors, then the most noticeable division into two clusters might separate "extreme" tumors from the milder ones and the normal tissues, and only further division will separate the normals from the milder tissues.

To address this question, we propose a measure of cluster *compatibility* with a given labeling. The intuition is simple: On the one hand, we want all the samples in the same cluster to have the same labels. Thus, we penalize pairs of samples that are within the same cluster but have different labels. On the other hand, we do not want to create unnecessary partitions. Thus, we also penalize pairs of samples that have the same label, but are not within the same cluster.

Formally, we define the *compatibility* score of a clusters with the training set as the sum of two terms. The first is the number of tissues pairs (v, u) such that v and u have the same label, and are assigned to the same cluster. The second term is the number of (v, u) pairs that have different labels, and are assigned to different clusters. This score is also called the *matching coefficient* in the literature [11].

It is easy to see that the two terms in this definition tradeoff the requirement that clusters should be as homogeneous as possible, and the requirement that clusters should not create small partitions. It is also important to note that we can evaluate cluster compatibility with a labeling, even when some of the patterns are not assigned a label. We simply restrict the comparison to counting pairs of examples for which we have a label.

Using this definition, we can optimize, using binary search, the choice of clustering parameters to find the most compatible clustering. That is, we consider different threshold values, t, use BioClust to cluster the tissues, and measure the compatibility of the resulting clusters with the given labels.

Finally, we choose the clustering that has maximal compatibility score to the given labeling. Thus, although the clustering algorithm is *unsupervised*, in the sense that it does not take into account the labels, we use a supervised procedure for choosing the clustering threshold. We also stress, that this general idea can be applied to other clustering methods, and is not restricted to our particular choice.

We now return to the question of prediction using clustering algorithms. To return a prediction, we examine the labels of all the patterns in the same cluster as the query. The intuition is that the query's label should agree with the labels of most of these patterns. Thus, we can use a simple majority rule to decide on the label. If the cluster contains exactly the same number of tumor and normal tissues, than the classifier does not give a prediction for the query.

[1] In this work we use the Pearson correlation between gene expression profile as the similarity measure. However, any similarity measure can be used.

2.3 Large-Margin Classifiers

The cluster-based approach we discussed in the previous section attempts to find inherent structure in the data (i.e., clusters of samples) and uses this structure for prediction. We can also use *direct* methods that attempt to learn a *decision surface* that separates the positive labeled samples from the negatively labeled samples.

The literature of supervised learning discusses a large number of methods that learn decision surfaces. These methods can be described by two aspects. First, the class of surfaces from which one is selected. This question is often closely related to the *representation* of the learned surface. Examples include linear separation (which we discuss in more detail below), decision-tree representations, and two-layer artificial neural networks. Second, the learning rule that is being used. For example, one of the simplest learning rules attempts to minimize the number of errors on the training set.

Application of direct methods in our domain can suffer from a serious problem. In gene-expression data we expect N, the number of measured genes, to be significantly larger than m, the number of samples. Thus, due to the large number of dimensions there are many simple decision surfaces that can separate the positive examples from the negative ones. This means that counting the number of training set errors is not restrictive enough to distinguish good decision surfaces from bad ones (in terms of their performance on examples not in the training set).

In this paper, we use two methods that received much recent attention in the machine learning literature. Both methods attempt to follow the intuition that classification of examples depends not only on the region they are in, but also on a notion of *margin*: how close are they to the decision surface. Classification of examples with small margins is not as confidant as classification of examples with large margins. (We can think of the learned decision surface as an estimate, and thus given slightly different data we might move it a bit.) Thus, the reasoning suggests that we should select a decision surface that classifies correctly with large margin all the training examples. This basic intuition is developed in quite different manner in these two approaches. Below we discuss the intuition for both approaches, and defer additional details to the appendices.

2.3.1 Support Vector Machines *Support vector machines* (SVM) were developed in [6, 27]. A tutorial on SVMs can be found in [3]. The intuition for support vector machines is best understood in the example of linear decision rules. A linear decision rule can be represented by a hyperplane in R^N such that all examples on the one side of the hyperplane are labeled positive and all the examples on the other side are labeled negative. Of course, in sufficiently high-dimensional data we can find many linear decision rules that separate the examples. Thus, we want to find a hyperplane that is as far away as possible from all the examples. More precisely, we want to find a hyperplane that separates the positive examples from the negative ones, and also maximizes the minimum distance of the closest points to the hyperplane. This question can be posed as a quadratic program (see Appendix A), and can be solved efficiently. The resulting hyperplane can be written as weighted sum of the training examples, x_i, and the classification of a new example x can be calculated using dot products with the example vectors, $x \cdot x_i$. This treatment can be generalized to deal with training sets that are not linearly separable. We refer the reader to [3] for details.

It is clear that linear hyperplanes are a restricted form of decision surfaces. One method of learning more expressive separating surfaces is to project the training examples (and later on queries) into a higher-dimensional space, and learn a linear separator in that space. For example, if our training examples are in R^1, we can project input values x to the vector $(1, x, x^2)$. A linear separator in the projected space is equivalent to learning an interval in the original representation of the training examples.

Thus, we can fix a projection $\Phi : \mathbf{R}^N \mapsto \mathbf{R}^M$ to higher dimensional space, and get more expressive decision surfaces. In this case, the classification rule for x will be composed of the inner products $\langle \Phi(x), \Phi(x_i) \rangle$. Moreover, for many projections there are *kernel* functions that compute the result of the inner product. A kernel function k for a projection Φ satisfies $k(x, y) = \langle \Phi(x), \Phi(y) \rangle$. Given a legal kernel function, we can use it without knowing the actual mapping Φ.

To summarize, if we want to learn expressive decision surfaces, we can choose a kernel function, and use it instead of inner-product in the execution of the SVM optimization. This is equivalent to learning a linear hyperplane in the projected space.

In this work we consider two kernel functions:

- The linear kernel $k_1(x, y) = \langle x, y \rangle$.
- The quadratic kernel $k_2(x, y) = (\langle x, y \rangle + 1)^2$.

The rational for using these simple kernels, is that since our input space is high dimensional, we can hope to find a simple separation rule in that space. We therefore test the linear separator, and the next order separator as a comparison to check if higher order kernels can yield better results.

2.3.2 Boosting Boosting was initially developed as a method for constructing good classifiers by repeated calls to "weak" learning procedure [20, 12]. The assumption is that we have access to a "weak learner". Such an algorithm constructs a function $f_D(x)$ for each training set. The learner is weak in the sense that the *generalization error* of $f_D(x)$ is only slightly better than that of random guess. Formally, we assume that $f_D(x)$ classifies at least $1/2 + 1/poly(n)$ of the input space correctly.

In this paper, we use a fairly simple weak learner, that finds a simple rule of the form:

$$f(x, j, t_j, d) = \begin{cases} d & x[j] > t_j \\ -d & x[j] < t_j \end{cases}$$

where j is an index of a gene, $x[j]$ is the expression value of the j'th gene in the vector x, t_j is a threshold corresponding to gene j, and $d \in \{+1, -1\}$ is a direction parameters. Such a classifier is called a *decision stump*. We learn decision stumps from data by exhaustively searching all genes, and for each gene search over all thresholds and directions, and finally return the combination that has the smallest number of errors.[2]

Boosting uses the weak learning procedure (e.g., the decision stump learner in our case) to construct a sequence of classifiers $f_1, \ldots, f_k$, and then uses a weighted vote among these classifiers. Thus, the prediction made by the boosting algorithm has the form : $\text{sign}(\sum_j w_j f_j(x))$, where w_i are the weights assigned to the classifiers.

The crux of the algorithm is the construction of the sequence of classifiers. The intuition is simple. Suppose that we train the weak learner on the original training data D to get a classifier $f_1(x)$. Then, we can find the examples in D that are classified incorrectly by f_1. We then want to force the learning algorithm to give these examples special attention. This is done by constructing a new training data set in which these examples are given more weight. Boosting then invokes the weak learner on the reweighted training set and obtains a new classifier. Examples are then reweighted, and the process is iterated. Thus, boosting adaptively reweights training examples to focus on the "hard" ones.[3] In this paper, we use the AdaBoost algorithm of Freund and Schapire [12]. See Appendix B for the details of the algorithm. In practice boosting is an efficient learning procedure that usually has small number of errors on test sets. The theoretical understanding of this phenomenon uses a notion of margin that is quite similar to the one defined for SVMs. Recall, that boosting classification is made by averaging the "votes" of many classifiers. Define the margin of example x_i to be

$$m_i = l_i \sum_j w_j f_j(x_i).$$

By definition, we have that if $m_i > 0$, then $\text{sign}(\sum_j w_j f_j(x_i)) = l_i$, and thus x_i is classified correctly. However, if m_i is close to 0, then this classification is "barely" made. On the other hand, if m_i is close to 1, then a large majority of the classifiers make the right prediction on x_i. The analysis of Schapire et al. [18, 21] shows that the generalization error of boosting (and other voting schemes) depends on the distribution of margins of training examples. Schapire et al. also show that repeated iterations of AdaBoost continually increase the smallest margin of training examples. This is contrasted with other voting schemes that are not necessarily increasing the margin for the training set examples.

3 Evaluation

In the previous section we discussed several approaches for classification. In this section we describe empirical evaluation of the classification performance of these approaches on gene expression classifications.

[2]Note that for each gene, we need to consider only n rules, since the gene takes at most n different values in the training data. Thus, we can limit our attentions to mid-way points between consecutive values attained by the j'th gene in the training data.

[3]More precisely, boosting distorts the distribution of the input samples. For some weak learners, like the stump classifier, this can be simulated by simply reweighting the samples.

3.1 Data sets

Before we describe the evaluation methods, we describe the two datasets we examined. Both of these data sets involve comparing tumor and normal samples of the same tissue.

Colon cancer data set. This data set is a collection of expression measurements from colon biopsy samples reported by Alon et al. [1]. The data set consists of 62 samples of colon epithelial cells. These samples were collected from colon-cancer patients. The "tumor" biopsies were collected from tumors, and the "normal" biopsies were collected from healthy parts of the colons of the same patients. The final assignments of the status of biopsy samples were made by pathological examination.

Gene expression levels in these 62 samples were measured using high density oligonucleotide microarrays. Of the $\approx$ 6000 genes detected in these microarray, 2000 genes were selected based on the confidence in the measured expression levels. The data set of 62 samples vs. 2000 genes is available at http://www.molbio.princeton.edu/colondata.

Ovarian cancer data set. This data set is a collection of expression measurements from 32 samples[4]: 15 ovary biopsies of three types of ovarian carcinomas (benign (2), mucinous (1), and serous (12)), 12 biopsies of normal ovaries, and 5 samples of other tissues (liver and blood).Gene expression levels in these 32 samples were measured using a membrane-based array with radioactive probes. The array consisted of cDNAs representing approximately 100,000 clones from ovarian clone libraries. For some of the samples, there are two or three repeated hybridizations for error assessments . In these cases, we treated the average of the reported expression levels as the expression levels in the samples.

3.2 Estimating Prediction Errors

When evaluating the prediction accuracy of the classification methods we described above, it is important not to use the *training error*. Most classification methods will perform well on examples they have seen during training. To get a realistic estimate of performance of the classifier, we must test it on examples that did not appear in the training set. Unfortunately, since we have a small number of examples, we cannot remove a portion of the examples from the training set, and use them for testing.

A common method to test accuracy in such situations is *cross-validation*. To apply this method, we partition the data into k sets, $C_1, \ldots, C_k$ of samples (typically, these will be of roughly the same size). Then, we construct a dataset $D_i = D - C_i$, that consists of all the training samples, except these in i'th partition. We test the accuracy of the classifier $f_{D_i}()$ on samples from the partition C_i. These steps are repeated for each of the partitions. We can then estimate the accuracy of the method, by averaging the accuracy in each one of the cross-validation trials.

Cross-validation has several important properties . First, the training set and the test set in each trial are disjoint. Second, the classifier is tested on each sample exactly once. Finally, the training set for each trial is $(k-1)/k$ of the original data set. Thus, we get a less biased estimate of the classifier behavior given a training set of size n.

There are several possible choices of k. A common approach is to set $k = n$. In this case, every trial removes one sample and trains on the rest. This, method is known as *leave one out cross validation* (LOOCV). Another common choice is to set $k = 10$ or $k = 5$. LOOCV has been in use since early days of pattern recognition (e.g., [8]). In some situations, using larger partitions reduces the variance of the estimators (see [16]). In this work, we use LOOCV. However, we are in the process of collecting results using 10 fold cross validation, and will use these to reaffirm the estimates based on LOOCV in the conference version of the report.

Table 1 lists the accuracy estimates for the different methods applied to two datasets.[5] As we can see, the clustering approach performs significantly better than the other approaches on the colon cancer data set.

3.3 ROC Curves

Estimates of classification accuracy give only a partial insight on the performance of a method. In our evaluation, we treated all errors as having equal penalty. In many applications, however, errors have asymmetric weights. To set terminology, we distinguish *false positive* errors, where normal tissues are classified as tumor,

[4]The training set contains 28 smaples labeled as tumor or normal

[5]Some of our methods were not run on the ovarian cancer data set due to technical difficulties with the large number of clones. We are currently working on dealing with these technical issues.

Method	Colon	Ovarian
Nearest Neighbor	80.6% +- 5.0%	71.4% +- 8.5%
Clustering	88.7% +- 4.0%	70.5%[a] +- 8.3%
SVM, linear kernel	80.6% +- 5.0%	–
SVM, quad. kernel	79.0% +- 5.1%	–
Boosting, 100 iterations	77.4% +- 5.3%	–
Boosting, 1000 iterations	74.2% +- 5.6%	–
Boosting, 10,000 iterations	77.4% +- 5.3%	–

[a] based on the 17 predictions that the classifier made

Table 1: Summary of classification accuracy of the methods on the two training sets. Reported accuracies denote average number of correct classifications and std. deviation. Estimates are based on LOOCV estimates.

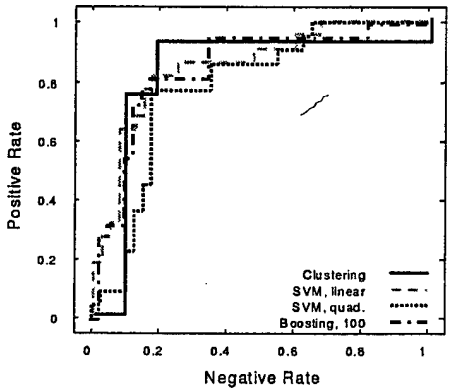


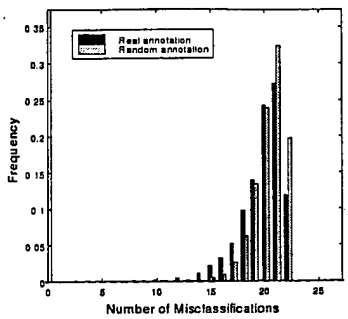
Figure 1: ROC curves for methods applied to colon cancer data set. The x-axis shows percentage of negative examples classified as positives, and y-axis shows percentage of positive examples classified as positive. Each point along the curve corresponds to the percentages achieved by a particular confidence threshold value by the corresponding classification method. Error estimates are based on LOOCV trials.

and *false negative* errors, where tumor tissues are classified as normal. In screening patients, avoiding false negative errors can be crucial, while making false positives might be tolerated (since additional tests will be performed on the patient).

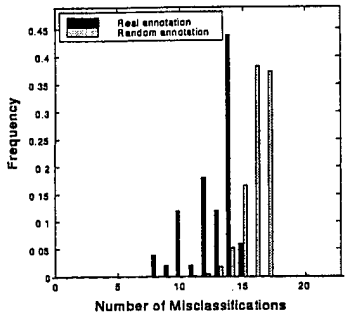
To deal with asymmetric weights for errors, we introduce the *confidence parameter*, τ. In clustering approaches, the modified procedure would predict that a query tissue is tumor, if the cluster containing it has at least a fraction τ of tumors. In a similar manner, we can introduce confidence parameters for SVM and boosting approaches by changing the threshold margin needed for positive classification.

We can evaluate the "power" of a classification method for different asymmetric weights by plotting *ROC curves* (see, for example, [26]). A ROC curve plots the tradeoff between the two types of errors as we change the confidence parameters. Formally, we plot a two dimensional curve. Each point on the curve corresponds to a particular value of the confidence parameter. The (x, y) coordinates of a point specifies the fraction of negative, and positive samples that are classified as positive with this particular confidence parameters. The extreme ends of the curves are the most strict and most permissive confidence values. With the strictest confidence values, the procedure does not classify any example as positive. Thus, this value corresponds to the point $(0, 0)$. On the other hand, with the most permissive confidence value, the procedure will classify each example as positive. Thus, this confidence value corresponds to the point $(1, 1)$. The path between these two extremes shows how useful the classification method is in distinguishing between positive and negative examples. The best case scenario is that the path goes through the point $(0, 1)$. This implies that for some confidence parameter, all positives are classified as positives, and all negatives are classified as negative. The general shape of the curve and in particular the area below the curve, are indicative of the distinguishing power of a classification method.

In Figure 1 we plot the ROC curves for clustering, SVM and boosting on the colon cancer data set. As we can see, there is no clear domination among the methods. (The only exception is SVM with quadratic kernel that is consistently worse than the other methods.) The clustering procedure is clearly dominant in the region where misclassification errors are roughly of the same importance. However, SVM with linear kernel and boosting are preferred to clustering in regions of asymmetric error cost (both ends of the spectrum). We believe that the "weakness" of the clustering in the asymmetric cost regions is due to the fact that the matching coefficient score (see Section 2.2) that determines the cluster granularity treats both types of errors as having equal costs.



(a)



(b)

Figure 2: (a) The distribution of gene scores for the colon cancer data set comparing the scores achieved using the original labels, and the a random labeling. (b) the same histogram for the 50 best scoring genes.

4 Gene Selection

It is clear that the expression levels of many of the genes that are measured in our data sets are irrelevant to the distinction between tumor and normal tissues. Taking such genes into account during classification increases the dimensionality of the classification problem, presents computational difficulties, and introduces unnecessary noise in the process. Another issue with a large number of genes is the *interpretability* of the results. If the "signal" that allows our methods to distinguish tumor from normal tissues is encoded in the expression levels of few genes, then we might be able to understand the biological significance of these genes. Moreover, a major goal for diagnostic research is to develop diagnostic procedures based on inexpensive microarrays that have enough probes to detect diseases. Thus, it is crucial to recognize whether a small number of genes can suffice for good classification.

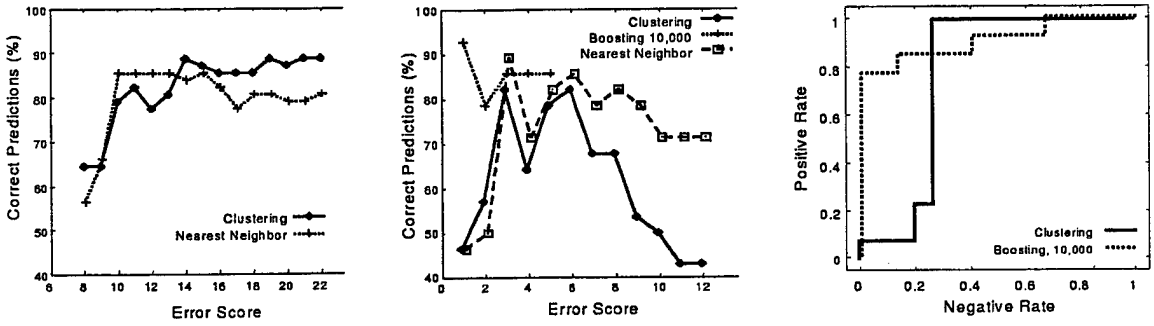
The problem of *feature selection* received a thorough treatment in pattern recognition and machine learning. The gene expression data sets are problematic in that they contain a large number of genes (features) and thus methods that search over subsets of features can be prohibitively expensive. Moreover, these data sets contain only a small number of samples, so the detection of irrelevant genes can suffer from statistical instabilities.

To address these issues, we propose a simple measure of "relevance" of each gene. For each gene, we measure how well we can classify the training examples if allow ourselves to ask one question about the gene's expression level. The intuition is that an informative gene has quite different values in the two classes, and thus we should be able to separate these by a threshold value. Formally, this is equivalent to finding the best decision stump for that gene (as defined in Section 2.3.2), and then measuring how many classification errors this decision stump makes on the training examples. We call this quantity the *error score* of a gene.

An immediate question to ask is whether genes with low error scores are indeed indicative of the classification of expression. In other words, we want to test the significance of the scores of the best scoring genes in our data set. We can measure significance by analyzing the expected behavior of scores if the labeling of samples was independent of gene expression data. We estimate this quantity by creating a random labeling for the gene expression patterns in our data sets. As we can see from Figure 2 the distribution of scores in the randomized dataset is distinctly different than the distribution of scores for the original dataset. In particular, the errors scores achieved by the best scoring genes in the true data are extremely unlikely in random data.

Aside from the statistical significance of the selected genes, we would also like to evaluate their biological significance. To estimate this, we have ordered the genes in both data sets, according to their error score, and examined the genes at the top of the list (those that have low error score). Among the top 100 genes in the colon cancer data set there are a number of genes that are interesting from the perspective of a potential involvement in tumorigenesis including, for example, genes involved in cell cycle regulation and angiogenesis. There were also genes, for example (D63874) HMG-1 (human) and (T55840) tumor-associated antigen L6 (human), that have previously been found to have a particular association with colorectal carcinomas [23, 29].

Among the top scoring 137 clones in the ovarian cancer data, there are 85 clones that match to 8 genes that are cancer related (potential markers or expressed in cancer cells) and one that is related to increased metabolic rate (mitochondrial gene). These genes are keratin 18 (breast cancer), pyruvate kinase muscle 2 (hepatoma), thymopoietin (cell proliferation), HE4 (ovarian cancer), SLPI (many different cancers, among them lung, breast, oropharyngeal, bladder, endometrial, ovarian and colorectal carcinoma), ferritin H (ovarian cancer), collagen 1A1 (ovarian cancer, osteosarcoma, cervical carcinoma), and GAPDH (cancers of lung, cervix and prostate). In addition, 2 clones with no homology to a known gene are found in this selection. Given the high number of cancer related genes in the top 137, it is likely that these novel genes exhibit a



(a) Colon data set (b) Ovarian data set (c) ROC curve

Figure 3: Curves showing how the classification accuracy depends on the threshold for selecting genes. The x-axis shows the error-score threshold for selecting genes. The y-axis shows classification accuracy based on LOOCV. Accuracy curves are (a) for colon data set, and (b) for ovarian data set. (c) shows ROC curves for two methods that are applied to the ovarian data set with error score threshold set to 3.

similar cancer-related behavior. We conducted expression validation for GAPDH, SLPI, HE4 and keratin 18 which confirmed the elevated expression in some ovarian carcinomas compared to normal ovarian tissues.

When using gene selection, we need to pre-process the training data to select genes. Then, the classification procedure is applied to the training data restricted to the subset of selected genes. The gene selection stage is given a parameter k, which determines the largest error-score allowed. It then selects all genes that have a smaller or equal error score on the training data.

To evaluate performance with gene selection, we have to be careful to evaluate together the two stage process of gene selection and classification. Thus, in each cross-validation trial, gene selection is applied based on the training examples in that trial. Note, that since the training examples are different in different cross validation trials, we expect the number of genes with error scores below a given threshold to change between trials.

Figures 3(a) and 3(b) show the classification accuracy for some of the methods we described above when we vary the maximal score distribution of genes we select. We note that SVM and boosting were also run with feature selection. However, due to technical issues, they were run on subsets of fixed sizes. The results for colon cancer show that both achieve approximately 80% accuracy with subsets of size 100 instances and bigger. Note also that for the clustering method this presentation is over pessimistic as it treat unclassified tissues as failures. For example, in the ovarian data set, for error score of 12, the clustering based classifier made 12 correct prediction, 5 wrong predictions, and 11 'unknown' predictions.

Both these graphs show that we can achieve quite a good classification performance with a small number of genes. For example, in the colon cancer data set, feature selection neither helps the clustering approach, nor does it significantly harm its behavior. We see that even for error threshold 10, which corresponds to selecting 10 genes on average, we see good prediction performance. In the ovarian data set, the critical threshold value is 3, which corresponds to selecting, on average, 173 clones.

In the colon data set, gene selection does not lead to significant improvement. On the other hand, in the ovarian data set, gene selection leads to impressive improvement in all methods. All three methods perform well in the region between threshold 3 (avg. 173 clones) to 6 (avg. 4375 clones). Note that Boosting performs well even with fewer clones. Figure 3(c) shows an ROC curve for Boosting and the Clustering approach with threshold of 3. As we can see, although both methods have roughly the same accuracy with this subset of genes, their ROC profile is strikingly different. These curves clearly show that the Clustering approach makes false positive errors, while the boosting approach makes false negative errors.

5 Sample Contamination

Cancer classification based on array-based gene expression profiling may be complicated by the fact that clinical samples, e.g. tumor vs. normal, will likely contain a mixture of different cell types. In addition, the genomic instability inherent in tumor samples may lead to a large degree of random fluctuations in gene expression patterns. Although both the biological and genetic variability in tumor samples have the potential

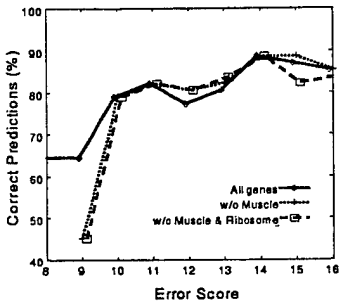


Figure 4: Curves showing the predictive performance of clustering methods in the original Alon et al. data set, and data sets where muscle specific, and ribosomal genes were removed. All estimates are based on LOOCV evaluation. These results show that even without the obvious contaminations, our methods are successful in reliably predicting tissue type.

to lead to confusing and difficult to interpret expression profiles, gene expression profiling does allow us to efficiently distinguish tumor and normal samples, as we have seen in the previous sections. However, the presence of different cell types within and between samples could lead to identification of genes that strongly affect cluster formation but which may have little to do with the process being studied, in this case tumorigenesis. For example, in the case of the colon cancer data set presented above, a large number of muscle-specific genes were identified as being characteristic of normal colon samples both in our clustering results and in the results of Alon et al. [1]. This is most likely due to a higher degree of smooth muscle contamination in the normal versus tumor samples.

This raises the concern that our classification may be biased by the presence of muscle specific genes. To test this hypothesis, we performed the following experiments. We listed the top 200 error-score ranking genes in the colon cancer data set, and identified muscle-specific genes. These include (J02854) myosin regulatory light chain 2, smooth muscle isoform (human); (T60155)actin, aortic smooth muscle (human); and (X12369) tropomyosin alpha chain, smooth muscle (human)that are designated as smooth muscle-specific by Alon et al.'s analysis, and (M63391) desmin (human), complete cds; (D31885) muscle-specific EST (human); and (X7429) alpha 7B integrin (human) which are suspected to be expressed in smooth muscle based on literature searches.

An additional form of "contamination" is due to the high metabolic rate of the tumors. This results in high expression values for ribosomal genes. Although such high expression levels can be indicative of tumors, such a finding does not necessarily provide novel biological insight into the process, nor provide a diagnostic tool since ribosomal activity is present in virtually all tissues. Thus, we also identified ribosomal genes in the top 200 scoring genes.

Figure 4 shows the performance of the clustering approach on three data sets: the full 2000 gene data set, a data set without muscle specific genes, and a data set without both muscle specific and ribosomal genes. As the learning curves show, the removal of genes affects the results only in cases using the smallest sets of genes. From error score threshold of 10 (avg. 9.1 genes) and higher, there is no significant change in performance for the procedure. Thus, although muscle specific genes can be highly indicative, the classification procedure performs well even without relying on these genes.

Although the muscle contamination did not necessarily alter the ability of this gene set to be used to classify tumor vs. normal samples in this case, it will continue to be important to account for possible affects of tissue contamination on clustering and classification results. Experimental designs that include gene expression profiles of tissue and/or cell culture samples representative of types of tissue contaminants known to be isolated along with different types of tumor samples (for example see Perou et al. [19]), can be utilized to help distinguish contaminant gene expression profiles from those actually associated with specific types of tumor cells.

6 Conclusions

In this paper we examined the question of tissue classification based on expression data. Our contribution is three-fold. First, we introduced a new cluster-based approach for classification. This approach builds on the recent development of clustering algorithms that are suitable for gene expression data. Second, we performed rigorous evaluation of this method, and a few known methods from the machine learning literature. These include large margin classification methods (SVM and stump-boosting) and the nearest-neighbor method. Third, we investigated the issue of gene selection in expression data. As our results for the ovarian data set show, a large number of clones can have a negative impact on predictive performance. We showed

that a fairly simple selection procedure can lead to significant improvements in prediction accuracy. In addition, we highlighted the issue of sample contamination and estimated the sensitivity of our approach to such contamination.

One clear future direction is extracting from the learned classifiers the genes that play a dominant role in them (i.e. those genes on which the classification relies the most). This might reveal some previously unknown disease related genes, which might point a direction for biological research.

References

[1] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Nat. Acad. Sci. USA*, 96:6745–6750, 1999.

[2] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 2000. to appear.

[3] C. J. C. Burges. A tutorial on Support Vector Machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.

[4] S. Chu, J. DeRisi, M. Eisen, J. Mullholland, D. Botstein, P. Brown, and I. Herskowitz. The transcriptional program of sporulation in budding yeast. *Science*, 282:699–705, 1998.

[5] P. A. Clarke, M. George, D. Cunningham, I. Swift, and P. Workman. Ananlysis of tumor gene expression following chemotherapeutic treatment of patients with bowel cancer. In *Proc. Nature Genetics Microarray Meeting 99*, page 39, Scottsdale, Arizona, 1999.

[6] C. Cortes and V. Vapnik. Support vector machines. *Machine Learning*, 20:273–297, 1995.

[7] J. DeRisi., V. Iyer, and P. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 282:699–705, 1997.

[8] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, 1973.

[9] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Nat. Acad. Sci. USA*, 95:14863–14868, 1998.

[10] Lander et al. to appear, 1999.

[11] B. Everitt. *Cluster Analysis*. Edward Arnold, London, third edition, 1993.

[12] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Computer and System Sciences*, 55:119–139, 1997.

[13] V.R. Iyer, M.B. Eisen, D.T. Ross, G. Schuler, T. Moore, J.C.F. Lee, J.M. Trent, L.M. Staudt, J. Hudson, M.S. Boguski, D. Lashkari, D. Shalon, D. Botstein, and P.O. Brown. The transcriptional program in the response of human fibroblasts to serum. *Science*, 283:83–87, 1999.

[14] Kim lab home page. http://cmgm.stanford.edu/ kimlab/.

[15] J. Khan, R. Simon, M. Bittner, Y. Chen, S. B. Leighton, T. Pohida, P. D. Smith, Y. Jiang, G. C. Gooden, J. M. Trent, and P. S. Meltzer. Gene expression profiling of *Alveolar rhabdomyosarcoma* with cDNA microarrays. *Cancer Reasearch*, 1998.

[16] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proc. Fourteenth International Joint Conference on Artificial Intelligence (IJCAI '95)*, pages 1137–1143. Morgan Kaufmann, San Francisco, Calif., 1995.

[17] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Want, M. Kobayashi, H. Horton, and E. L. Brown. DNA expression monitoring by hybridization of high density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–1680, 1996.

[18] L. Mason, P. Bartlett, and J. Baxter. Direct optimization of margins improves generalization in combined classifiers. In *Advances in Neural Information Processing Systems 11*. MIT Press, Cambridge, Mass., 1999.

[19] C. M. Perou, S. S. Jeffrey, M. v de Rijn, C. A. Rees, M. B. Eisen, D. T. Ross, A. Pergamenschikov, C. F. Williams, S. X. Zhu, J. C. F. Lee, D. Lashkari, D. Shalon, P. O. Brown, and Botstein D. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Nat. Acad. Sci. USA*, 96:9212–9217, 1999.

[20] R. E. Schapire. The strength of weak learnability. *Machine Learning*, 5:197–227, 1990.

[21] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of Statistics*, 26:1651–1686, 1998.

[22] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 1999. to appear.

[23] TH. Schiedeck, S. Christoph, M. Duchrow, and H.P. Bruch. Detection of hl6-mrna: new possibilities in serologic tumor diagnosis of colorectal carcinomas. *Zentralbl Chir*, 123(2):159–162, 1998.

[24] M. Schummer. in preperation, 1999.

[25] M. Schummer, W. NG, Bumgarner R., Nelson P., B. Schummer, L. Hassell, L. Rae Baldwin, B. Karlan, and L. Hood. Comperative hybridization of an array of 21,500 overian cDNAs for the discovery of genes overexpressed in overian carcinomas. *Gene*, 1999. in print.

[26] J. Swets. Measuring the accuracy of diagnostic systems. *Science*, 240:1285–1293, 1988.

[27] V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, 1999.

[28] X. Wen, S. Furhmann, G. S. Micheals, D. B. Carr, S. Smith, J. L. Barker, and R. Somogyi. Large-scale temporal gene expression mapping of central nervous system development. *Proc. Nat. Acad. Sci. USA*, 95:334–339, 1998.

[29] Xiang YY, Wang DY, Tanaka M, Suzuki M, Kiyokawa E, Igarashi H, Naito Y, Shen Q, and Sugimura H. Expression of high-mobility group-1 mrna in human gastrointestinal adenocarcinoma and corresponding non-cancerous mucosa. *Int J. Cancer*, 74(1):1–6, Feb 1997.

A Support Vector Machines

A linear decision rule can be represented by a hyperplane in R^N such that all examples on the one side of the hyperplane are labeled positive and all the examples on the other side are labeled as negative. Such a rule can be represented by a vector $w \in R^N$ and a scalar b that together specify the hyperplane $w \cdot x + b = 0$. Classification for a new example x is performed by computing $\text{sign}(w \cdot x + b)$. Recall that $|\frac{x \cdot w + b}{||w||}|$ is the distance from x to the line $x \cdot w + b = 0$. Thus, if all points in the training data satisfy

$$l_i(x_i \cdot w + b) \geq 1 \tag{1}$$

then we know that they are all correctly classified, and all of them have a distance of at least $1/||w||$ from the hyperplane. We can find the hyperplane that maximizes the margin of error by solving the following quadratic program:

Minimize $||w||^2$
Subject to $l_i(x_i \cdot w + b) \geq 1$ for $i = 1, \ldots, m$.

Input:

- A data set of N labeled examples $\{(x_1, l_1), \ldots, (x_N, l_N)\}$
- A weak learning algorithm L.

Initialize the distribution over the data set: $D_1(x_i) = 1/N$
For $t = 1, 2, ..., T$

- Call L with distribution D_t; Get back a hypothesis h_t.
- Calculate the error of h_t: $\epsilon_t = \sum_{i=1}^{N} D(x_i) 1(l_i \neq h(x_i))$
- Set $\alpha_t = \frac{1}{2} log \frac{\epsilon_t}{1-\epsilon_t}$
- Set the new distribution to be:

$$D_{t+1}(x_i) = \frac{D_t exp(-\alpha_t 1(l_i \neq h(x_i)))}{Z_t}$$

Where Z_t is a normalization factor, chosen so that D_{t+1} will sum to 1.
Output: The final hypothesis $h(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$

Figure 5: The AdaBoost algorithm.

Such quadratic programs can be solved in the dual form. This dual form can be posed in terms of auxiliary variables α_i. The solution has the property that

$$w = \sum_i \alpha_i l_i x_i,$$

and thus, we can classify a new example x by evaluating

$$\text{sign}(\sum_i \alpha_i l_i \langle x_i, x \rangle + b) \tag{2}$$

In practice there is a range of optimization methods that can be used for solving the dual optimization problem. See [3] for more details.

The SVM dual optimization problem and its solution have several attractive properties. First, only a subset of the training examples determine the position of the hyperplane. Intuitively, these are exactly these samples that are at the distance $1/||w||$ from the hyperplane. It turns out that the dual problem solution assigns $\alpha_i = 0$ to all examples that are not "supporting" the hyperplane. Thus, we only need to store the *support vectors* x_i for which $\alpha_i > 0$. (Hence the name of the technique.)

Second, the dual form of the quadratic optimization problem involves only cross-products of vectors in R^N. In other words, vectors x_i do not appear outside the scope of a cross-product operation. Similarly, the classification rule (2) only examines vectors in R^N inside the cross-product operation. Thus, if want to consider any projection $\Phi : \mathbf{R}^N \mapsto \mathbf{R}^M$, then we can find an optimal separating hyperplane in the projected space, by solving the quadratic problem with cross-products $\langle \Phi(x_i), \Phi(x_j) \rangle$.

In many cases, we can perform the optimization in high-dimensional spaces, by efficient computation of the cross-product in these spaces. A function $k(x, y) = \langle \Phi(x), \Phi(y) \rangle$ is called a *kernel function*. For many projections, the kernel function can be computed in time that is linear in N, regardless of the dimension M.

B AdaBoost

AdaBoost algorithm was introduced in[12]. This algorithm is shown in Figure 5. See also [22] for a justification of the particular reweighting scheme used by Freund and Schapire.

HYBRIDISATION OF AN ARRAY OF 100,000 cDNAs WITH 32 TISSUES FINDS POTENTIAL OVARIAN CANCER MARKER GENES

SCHUMMER M[1], KIVIAT N[1], BEDNARSKI D[1], CRUMB GK[1], BEN-DOR A[1], DRESCHER C[2], HOOD L[1]
[1]University of Washington, Seattle; [2]Swedish Medical Center, Seattle, USA

Ovarian cancer mortality could be largely reduced by early detection through a sensitive, specific and inexpensive serum assay. In order to find new markers, we used array technology to screen for genes with overexpression in the carcinomas vs. the normal tissues. An array of 102,680 clones, randomly selected from 3 unamplified ovarian cDNA libraries, was interrogated with probes from 32 well characterised tissues (normal ovaries, ovarian carcinomas, blood and liver). The hybridisation patterns were anlysed with algorithms specifically created for such analysis. We found 2650 clones representing 883 genes with stronger expression in the tumours (476 matching known genes, 368 matching ESTs, and 48 novel genes). Some of the known genes were previously described cancer genes such as CD24, folate binding protein, c-myc, Her2/neu, mucin, metallothionein or c-jun. Detection of these genes demonstrates the power of our approach. To date, we performed real time-PCR-based expression validation on 34 novel and known genes in 72 tissues (18 normal ovaries,40 ovarian tumors and cell lines), of which 20 genes were confirmed as overexpressed in the tumours. These genes are currently characterized by in-situ hybridisation on tissue sections, and by screening for antibodies and transcripts in patient sera that were collected with the tissues.

Appendix E
Project Two: Figures and Table 2

1. Figures 1-4
2. Table 2

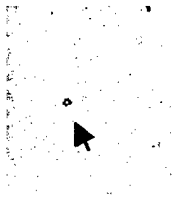


Fig 1. Example of an immunoreactive phage plaque from a primary SEREX screen. The arrowhead indicates a single immunoreactive plaque (dark halo) amongst several hundred non-reactive plaques (clear spots).

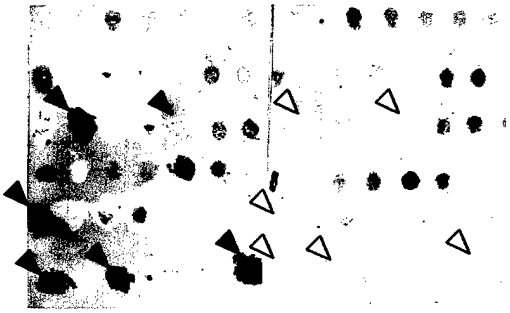


Fig 2. Secondary SEREX screening of ovarian cancer cDNA clones by phage array. The left and right panels show duplicate nitrocellulose membranes containing a 2-D array of recombinant phage clones that were identified in a primary SEREX screen of an ovarian tumor cDNA library. The left panel was immunoblotted with serum from a ovarian cancer patient (stage III , serous) whereas the right panel was immunoblotted with serum from a normal control. Membranes were then probed with a human IgG-specific, AP-conjugated secondary antibody and developed with NBT/BCIP. Immunoreactive phage plaques appear as dark circles, whereas non-reactive phage are clear. The arrows indicate 6 phage that showed a cancer-specific pattern of immunoreactivity with these and other serum samples.

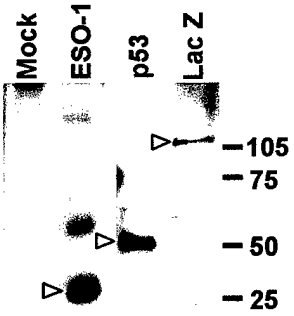


Fig 3. Western blot showing expressing of His-tagged recombinant tumor antigens in mammalian COS7 cells. Cells were transiently transfected with pcDNA3-based expression vectors encoding His-tagged ESO-1, p53 or Lac Z (as a control). Mock transfected cells served as a negative control. Nuclear extracts were prepared, subjected to SDS-PAGE and immunoblotted with a monoclonal antibody to the His tag (Sigma). Antibody detection was by enhanced chemiluminescence. Recombinant proteins are indicated by open arrowheads.

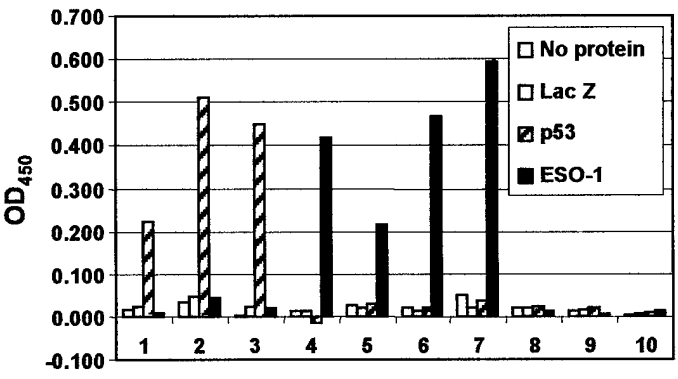


Fig 4. ELISA demonstrating serum antibody responses to p53 and ESO-1 in patients with ovarian cancer. Lysates from COS7 cells (see Fig. 3) expressing His-tagged p53, ESO-1 or, as a negative control, Lac Z were added to nickel-coated ELISA plates. After unbound proteins were washed away, serum from 10 ovarian cancer patients was added at 1:50 dilution, followed by HRP-conjugated goat anti-human IgG secondary antibody. Plates were developed with TMB and read at 450 nm. Patients #1-3 show a serum antibody response to p53, whereas patients #4-7 show a response to ESO-1. Patients #8-10 show no response to either protein.

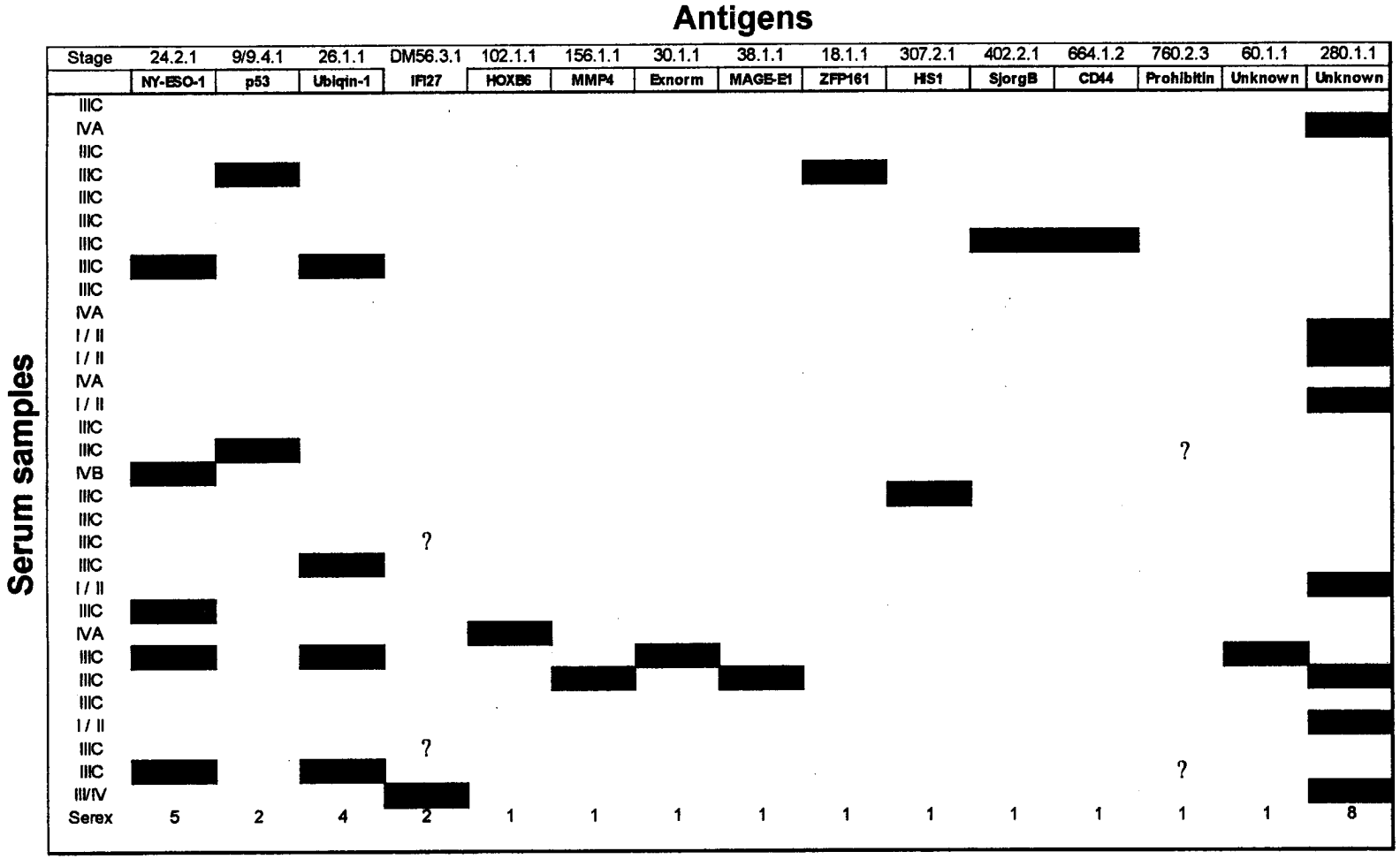


Table 2. Pattern of IgG serum antibody responses to SEREX-defined antigens among 26 late-stage serous, and 5 early-stage LMP serous ovarian cancer patients. Each row shows the results for one ovarian cancer patient. Black cells indicate a positive response, as detected by SEREX or by ELISA. All antigens were negative when tested against a panel of 20 normal control sera from age-matched women.

Appendix F
Study Brochure

1. ORCHID study brochure

Who conducts the study?

This work is a collaboration among experts in oncology, immunology, molecular biology, and statistical methods representing Fred Hutchinson Cancer Research Center, the University of Washington, Virginia Mason Research Center and Swedish Medical Center. Nicole Urban, ScD, of the Fred Hutchinson Cancer Research Center, is the Principal Investigator of the study. Funding for this project is awarded by the Department of Defense Ovarian Cancer Research Program to the Fred Hutchinson Cancer Research Center.

What is the Marsha Rivkin Center for Ovarian Cancer Research™?

In September 1989, Marsha Rivkin was diagnosed with ovarian cancer. Four years later, at the age of 49, she passed away leaving behind five daughters and her husband of 29 years, Dr. Saul Rivkin, an oncologist at Swedish Medical Center in Seattle. In Marsha's honor, and in recognition of the importance of continued work against this deadly disease, the Rivkin family founded the Marsha Rivkin Center for Ovarian Cancer Research.

The mission of the Marsha Rivkin Center is to improve the outcomes for women diagnosed with ovarian cancer, and those at risk for the disease, by encouraging collaborative scientific research, increased education and awareness, and community participation. The Swedish Medical Center Foundation raises funds for the Marsha Rivkin Center, some of which were used to support costs for the pilot phases of the ORCHID study.

What is the Fred Hutchinson Cancer Research Center?

The Fred Hutchinson Cancer Research Center is an independent, non-profit research institution dedicated to developing new knowledge to eliminate cancer. The Hutchinson Center is designated by the National Cancer Institute as a comprehensive cancer center. The Hutchinson Center is a world leader in laboratory, treatment and prevention research.

How can I get more information?

Contact the study office at **(206) 215-6200** or write to the address below:

ORCHID Study
Marsha Rivkin Center
for Ovarian Cancer Research
1221 Madison Street, Suite 1410
Seattle, WA 98104

ORCHID
Ovarian Research Collaboration
Helping to Improve Detection

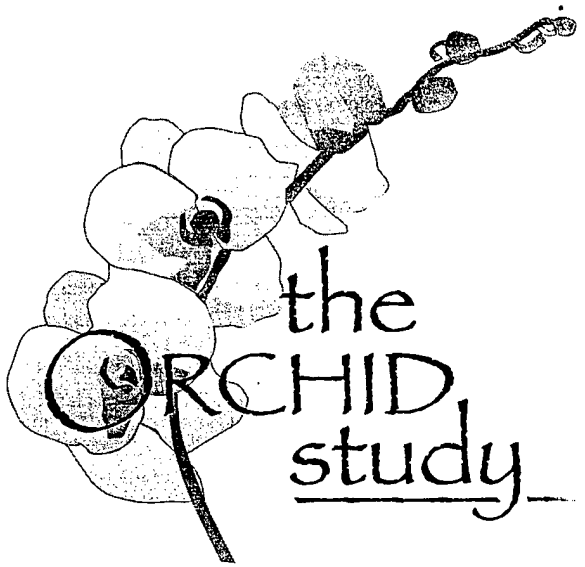
UW

FRED HUTCHINSON CANCER RESEARCH CENTER

Marsha Rivkin Center for Ovarian Cancer Research™

SWEDISH MEDICAL CENTER

VIRGINIA MASON Research Center



the ORCHID study

Improving Ovarian Cancer Screening

What is the ORCHID Study?

ORCHID stands for Ovarian Research Collaboration Helping to Improve Detection. The goal of the ORCHID study is to improve women's health by developing new screening tests to find ovarian cancer early, when it is easiest to treat.

Who may participate in the study?

You are eligible to participate if you:

- Are at least 18 years of age
- Have one or both ovaries
- Are willing to provide blood and/or ovarian tissue specimens

Women with ovarian cancer, women with benign ovarian disease (not cancer), and women with disease-free ovaries are invited to participate in this research.

What does participation in ORCHID involve?

Participation in ORCHID involves filling out a 15-minute questionnaire, giving permission for us to look at your medical records, and donating blood and/or tissue to be used in research analysis. Only women undergoing ovarian-related surgery at a participating hospital are asked to donate tissue. Scientists will look for molecular changes in the tissue and blood of women with and without ovarian cancer to aid in the development of a screening test for ovarian cancer.

What is ovarian cancer screening?

Ovarian cancer screening consists of medical tests that help find ovarian cancer in an early stage when there are no clinical signs of disease. Scientists with the ORCHID study are working to develop accurate methods of screening that can be used in the general population.

If I choose to participate in the study, how will I donate blood?

You can give blood during your regular office visit to your doctor. Women who are undergoing surgery at a participating hospital can donate blood at the time of their surgery.

If I choose to participate in the study, how will I donate tissue?

Some women require surgery to determine the best way to treat their medical condition. If you have surgery as part of your regular medical care and decide to participate in ORCHID, you may choose to donate tissue samples to the study.

During surgery, doctors remove tissue from the woman's ovary and send it to a lab for evaluation. The doctor in the lab, a pathologist, examines a portion of the tissue to help determine the best course of care. It is often not necessary to examine all of the tissue removed, so the remaining unexamined tissue is generally disposed of appropriately. If the patient is a participant in ORCHID and chooses to donate tissue, we will keep a portion of the tissue that is normally discarded and use it in our research.

How will participating in ORCHID affect my medical care?

Participating in this study will not affect your medical care. You will receive the same excellent medical care whether you participate in this study or not.

What will I get out of the study?

Participation in this study will not provide any direct benefit to you, but your contribution to this research effort may help other women at risk for ovarian cancer in the future.

Will I have to pay to participate? Will there be costs to me?

Participation in this study will not cost you any money. We will pay for the costs of blood and tissue collection. Neither you nor your insurance company will be billed for any procedures related to the collection of blood and tissue samples for the ORCHID study.

How much time will I have to give to the study?

The time required to participate in this study will be approximately 15-20 minutes.
We will ask you to read and sign a consent form, complete a questionnaire, and discuss your questions with a member of the research team. Blood and tissue donations generally occur during a regular office visit or medical procedure and do not require extra time.

Appendix G
Specimen Collection and Tracking Forms

1. ORCHID Specimen Collection Form
2. ORCHID Blood Specimen Form
3. ORCHID Specimen Tracking Form

ORCHID - *Specimen Collection Form*

THIS FORM MUST ACCOMPANY ALL PATIENT SPECIMENS SUBMITTED TO THE CORE LABORATORY AND REPOSITORY

UPN: _____ Procedure: _____ Surgeon ID: _____ Procedure date: ____/____/____

Patient name: _____ Institution of Proc.: ☐ SMC ☐ UW ☐ PRV TCS ID: _____
LAST FIRST MI

Informed Consent Verified? ☐ No ☐ Yes Case status: ☐ Incident ☐ Recurrent Chemo prior to surgery? ☐ No ☐ Yes

Specimen Information Place the duplicate barcode in the appropriate section (as applicable) and circle side from which each was removed.

	Primary Ovary	*Contralateral Ovary*	*Metastatic Site*
STM	R L		
Form. OCT	R L		
Formalin	R L	R L	
Frozen	R L	R L	
	R L	R L	
	R L	R L	
	R L	R L	
	R L	R L	
	R L	R L	
	R L	R L	
Total: _____		Total: _____	Total: _____

Comments/Notes: _____

_____

_____ Total frozen

Blood/Ascites

WBC pellets { ☐ ☐ ☐ / ☐ ☐ ☐ } Plasma ☐ Ascites { ☐ / ☐ }

Blood tubes: 10 ml red tops: _____ 5 ml EDTA tubes: _____ Time of collection: ____:_____ AM / PM

ORCHID – *Blood Specimen Form*

UPN: _____ Patient name: _____
LAST FIRST MI

DynaCare requisition no.: _____ Blood drawn in: ❑ surgery ❑ clinic

Date of blood draw: ____/____/____

Processing & Transport

❑ No. tubes submitted for processing: ☐ Red top ☐ EDTA

❑ Transported to freezer holding area after processing _____
INITIALS

In the spaces below, record the 6-digit number on the label of each vial received from the processing labs. Please note if any vial is only partially full, or if the contents are hemolyzed or lipidemic.

Serum

WBC Pellets

Plasma

ORCHID – *Specimen Tracking Form*

Date: October 12, 2000

Specimens sent from: ORCHID Core Repository

- ☐ -20° C freezer
- ☐ **-70° C freezer**
- ☐ liquid nitrogen freezer

Specimen type	**Number**
☐ Tissue	_____
☐ **Serum**	10_____
☐ Other	_____

Time of packaging: ___:_____ AM / PM

Transported on:
- ☐ **dry ice**
- ☐ liquid nitrogen

Specimens packaged for delivery:

FOR NELSON LABORATORY – SPECIMENS FOR ORCHID PROJECT 2.

202774	**100545**
201601	**202605**
201775	**100324**
100724	**100743**
202746	**100578**

Specimens received:

Receiving laboratory:
- ☐ Kiviat laboratory
- ☐ **M. Schummer laboratory (UW)**
- ☐ B. Nelson laboratory (VMMC)
- ☐ N. Disis laboratory (UW)
- ☐ Hellstrom laboratory (PNRI)
- ☐ Univ. of Washington: _____
- ☐ FHCRC: _____

Time of delivery receipt: ___:_____ AM / PM

Date of receipt: ___/___/___

- ☐ Contents above confirmed
- ☐ Contents different as noted: _____
  _____

Investigator/lab personnel signature: _____

TCS initials: _____

Appendix H
Histology and Clinical Data Abstraction Forms

1. ORCHID Specimen Histology Report I
2. ORCHID Specimen Histology Report II
3. ORCHID Clinical Data Form

ORCHID – *Specimen Histology Report I*

UPN: _____

Date of analysis: ____/____/____
Form completed by: _____

Clinical diagnosis: _____ **Site of Primary Ca.:** ☐ Ovarian ☐ Other:

Pathology Diagnosis:

☐ **Normal**

☐$_{03}$ Normal ovarian/tubal tissue
☐$_{05}$ Other normal:

☐ **Malignant**

Serous
☐$_{16}$ Serous carcinoma of LMP
☐$_{17}$ Serous carcinoma
☐$_{19}$ Serous cystadenofibroma
☐$_{20}$ Serous adenofibroma

Mucinous
☐$_{28}$ Mucinous adenocarcinoma of LMP
☐$_{29}$ Mucinous carcinoma
☐$_{30}$ Malignant mucinous adenofibroma
☐$_{31}$ Malignant mucinous cystadenofibroma

☐$_{57}$ Other:

Endometrioid
☐$_{37}$ Endometrioid carcinoma of LMP
☐$_{38}$ Endometrioid adenocarcinoma

Other
☐$_{40}$ Malignant adenosarcoma (mesodermal)
☐$_{41}$ Mesodermal (mullerian) mixed tumor, homo.
☐$_{42}$ Mesodermal (mullerian) mixed tumor, hetero.
☐$_{47}$ Clear cell carcinoma of LMP
☐$_{48}$ Clear cell carcinoma
☐$_{52}$ Brenner tumor of LMP
☐$_{53}$ Malignant Brenner tumor
☐$_{54}$ Undifferentiated carcinoma
☐$_{55}$ Adenocarcinoma, NOS

☐ **Benign**

Serous
☐$_{11}$ Serous cystadenoma
☐$_{12}$ Serous adenofibroma
☐$_{13}$ Serous cystadenofibroma
☐$_{14}$ Proliferating serous adenofibroma
☐$_{15}$ Proliferating serous cystadenofibroma

Mucinous
☐$_{25}$ Mucinous cystadenoma
☐$_{26}$ Mucinous adenofibroma
☐$_{27}$ Mucinous cystadenofibroma

Non-neoplastic
☐$_{06}$ Paraovarian cyst
☐$_{07}$ Functional cyst
☐$_{08}$ Corpus luteum
☐$_{09}$ Inflammatory lesion
☐$_{10}$ Endometriosis

Endometrioid
☐$_{32}$ Endometrioid cystadenoma
☐$_{33}$ Endometrioid adenofibroma
☐$_{34}$ Endometrioid cystadenofibroma
☐$_{35}$ Proliferating endometrioid adenofibroma
☐$_{36}$ Proliferating endometrioid cystadenofibroma

Other
☐$_{39}$ Benign adenofibroma (mesodermal)
☐$_{43}$ Clear cell adenofibroma
☐$_{44}$ Clear cell cystadenofibroma
☐$_{45}$ Proliferating clear cell adenofibroma
☐$_{46}$ Proliferating clear cell cystadenofibroma
☐$_{49}$ Benign Brenner tumor, typical
☐$_{50}$ Metaplastic Brenner tumor
☐$_{51}$ Proliferating Brenner tumor

☐$_{98}$ Other:

Tumor Grade: ☐$_{a}$ well differentiated ☐$_{b}$ moderately differentiated ☐$_{c}$ poorly differentiated

FIGO Stage:

☐$_{01}$ IA	☐$_{04}$ IIA	☐$_{07}$ IIIA	☐$_{10}$ IVA
☐$_{02}$ IB	☐$_{05}$ IIB	☐$_{08}$ IIIB	☐$_{11}$ IVB
☐$_{03}$ IC	☐$_{06}$ IIC	☐$_{09}$ IIIC	☐$_{12}$ IVC

POCRC – *Specimen Histology Report II*

Date of analysis: ____/____/____

Patient ID: _____

Form completed by: _____

A Site: _____
Path. dx.: _____
Necrosis: _____%
Normal cells: _____%
Infilt. by Inflammatory cells: _____%

B Site: _____
Path. dx.: _____
Necrosis: _____%
Normal cells: _____%
Infilt. by Inflammatory cells: _____%

C Site: _____
Path. dx.: _____
Necrosis: _____%
Normal cells: _____%
Infilt. by Inflammatory cells: _____%

D Site: _____
Path. dx.: _____
Necrosis: _____%
Normal cells: _____%
Infilt. by Inflammatory cells: _____%

Site

1 Primary ovarian tumor
2 Contralateral ovary – NL
3 Metastatic tumor
4 Non-ovarian tissue – NL
5 Not known
6 Ovarian tissue – NL
7 Tube – NL
8 Uterus
9 Other (specify)

Differentiation

a well differentiated
b moderately differentiated
c poorly differentiated

Pathology Diagnosis

Non-neoplastic lesions

1 Inadequate
2 Necrosis only
3 Normal ovarian or tubal tissue
4 Normal fibrovascular tissue
5 Normal other (specify)
6 Benign cyst/paraovarian
7 Functional cyst
8 Corpus luteum
9 Inflammatory lesion
10 Endometriosis

Epithelial Tumors

Serous tumors, benign
11 Serous cystadenoma
12 Serous adenofibroma
13 Serous cystadenofibroma
14 Proliferating serous adenofibroma
15 Proliferating serous cystadenofibroma

Serous tumors, malignant
16 Serous carcinoma of LMP
17 Serous carcinoma
19 Serous cystadenofibroma
20 Serous adenofibroma

Mucinous tumors, benign
25 Mucinous cystadenoma
26 Mucinous adenofibroma
27 Mucinous cystadenofibroma

Mucinous tumors, malignant
28 Mucinous adenocarcinoma of LMP
29 Mucinous carcinoma
30 Malignant mucinous adenofibroma
31 Malignant mucinous cystadenofibroma

Endometrioid tumors, benign
32 Endometrioid cystadenoma
33 Endometrioid adenofibroma
34 Endometrioid cystadenofibroma
35 Proliferating endometrioid adenofibroma
36 Proliferating endometrioid cystadenofibroma

Endometrioid tumors, malignant
37 Endometrioid carcinoma of LMP
38 Endometrioid adenocarcinoma

Mesodermal mixed tumors
39 Benign adenofibroma
40 Malignant adenosarcoma
41 Mesodermal (mullerian) mixed tumor, homo.
42 Mesodermal (mullerian) mixed tumor, hetero.

Clear cell tumors, benign
43 Clear cell adenofibroma
44 Clear cell cystadenofibroma
45 Proliferating clear cell adenofibroma
46 Proliferating clear cell cystadenofibroma

Clear cell tumors, malignant
47 Clear cell carcinoma of LMP
48 Clear cell carcinoma

Brenner tumors, benign
49 Benign Brenner tumor, typical
50 Metaplastic Brenner tumor
51 Proliferating Brenner tumor

Brenner tumors, malignant
52 Brenner tumor of LMP
53 Malignant Brenner tumor

Other
54 Undifferentiated carcinoma
55 Adenocarcinoma, NOS
56 Unclassified epithelial tumor
57 Neoplastic other (specify)

98 Non-neoplastic other (specify)
99 Other (specify)

ORCHID – *Clinical Data Form*

This form should be completed 1 to 2 weeks following a participant's surgery; this allows time for all surgical and pathology reports to be submitted to her medical records file.

UPN: _____ Form completed by: _____

Name: _____ Physician ID: _____

Med. records ID: _____ Location of records: ❑1 PGS ❑2 UW Gyn. Onc.

I. Presenting symptoms & duration

❑ H&P not in clinic records

Symptom	Report of symptom during history		Symptom duration (weeks) <4	4-8	>8	Not noted in H & P
1. Pain	❑1 Neg.	❑2 Pos. → Duration?	❑1	❑2	❑3	❑8
2. Distention	❑1 Neg.	❑2 Pos. → Duration?	❑1	❑2	❑3	❑8
3. Bleeding	❑1 Neg.	❑2 Pos. → Duration?	❑1	❑2	❑3	❑8
4. Fatigue	❑1 Neg.	❑2 Pos. → Duration?	❑1	❑2	❑3	❑8
5. Dyspepsia	❑1 Neg.	❑2 Pos. → Duration?	❑1	❑2	❑3	❑8
6. Weight change	❑1 Neg.	❑2 Pos. → Duration?	❑1	❑2	❑3	❑8
7. Bladder changes	❑1 Neg.	❑2 Pos. → Duration?	❑1	❑2	❑3	❑8
8. Bowel changes	❑1 Neg.	❑2 Pos. → Duration?	❑1	❑2	❑3	❑8
9. Other: _____	❑1 Neg.	❑2 Pos. → Duration?	❑1	❑2	❑3	❑8

Comments: _____

II. Pre-operative CA 125 screens

❑ None listed in clinic records

Date of exam	Results		
	U/ml	❑1 Dynacare	❑2 Other laboratory
	U/ml	❑1 Dynacare	❑2 Other laboratory
	U/ml	❑1 Dynacare	❑2 Other laboratory
	U/ml	❑1 Dynacare	❑2 Other laboratory

III. Size of ovarian mass

Abstract from pathology reports.

Date of report	LT	RT	Bidimensional tumor size (note units)
	❑1	❑2	____ x ____
	❑1	❑2	____ x ____

IV. Post-operative diagnosis

Record all that apply.

	Ca	LMP	Ben	Nml	
Rt. ovary	❑1	❑2	❑3	❑4	Dx.: _____
Lt. ovary	❑1	❑2	❑3	❑4	Dx.: _____

Other relevant conditions: ____ Dx.: _____

____ Dx.: _____

____ Dx.: _____

Appendix I
Tissue Protocol

1. ORCHID Tissue Collection, Processing and Transport Protocol

"Use of Novel Technologies to Identify and Investigate Molecular Markers for Ovarian Cancer Screening and Prevention"

TISSUE COLLECTION, PROCESSING AND TRANSPORT PROTOCOL
for
ORCHID CORE

Ovarian Research Collaboration Helping to Improve Detection

Conducted by Investigators at Fred Hutchinson Cancer Research Center (FHCRC), University of Washington (UW) and Virginia Mason Research Center (VM). Funding by the Marsha Rivkin Center for Ovarian Cancer Research and the United States Army Medical Research and Materiel Command (USAMRMC).

Dates of Study: February 1998 – September 2000

Principal Investigator:
Nicole D. Urban, ScD
Fred Hutchinson Cancer Research Center
1100 Fairview Avenue North - MP-804
PO Box 19024
Seattle, WA 98109
(206) 667-4677

Investigators:
Garnet Anderson, PhD (FHCRC)
Nancy Kiviat, MD (UW)
Charles Drescher, MD (FHCRC)
Mary Anne Rossing, PhD (FHCRC)
Jane Kuypers, PhD (UW)
Leona Holmberg, MD (FHCRC)

Location(s) of Study:

Fred Hutchinson Cancer Research Center (FHCRC)
1100 Fairview Avenue North
Seattle, WA 98109
(206) 667-5000

Pacific Gynecology Specialists (PGS)
1101 Madison Street, #1500
Seattle, WA 98104
(206) 215-6595

Marsha Rivkin Center for Ovarian Cancer Research
1221 Madison Street, Suite 1410
Seattle, WA 98104
(206) 386-2419

University of Washington (UW)
School of Medicine
Seattle, WA 98195
(206) 543-2100

Table of Contents

Introduction

Blood and tissue specimens for this research will be obtained from consenting patients identified by physicians practicing at Pacific Gynecology Specialists (PGS) and at the University of Washington (UW), and performing ovarian related surgery on the campuses of Swedish Medical Center and the University of Washington Medical Center.

Tissue collection technicians will be available during patient surgery to collect the appropriate tissue and serum samples for initial transport to the pathology laboratory, and later to the study laboratory/repository for further analysis and storage. Specimens will be placed in freezers in a pre-defined location, pending processing into the inventory control system. The specimen tracking subsystem will uniquely identify each separate item entered into the repository, document its location in the freezer, and its disposition. The transfer of specimens between sites of study will be tracked using a computerized data tracking system. Specimens that are not transported to project laboratories for analyses will remain in storage at the Core repository in a -70°C freezer or a liquid nitrogen freezer.

After characterization of specimens, requests for access to specimens by both Project Investigators and non-Project Investigators will be reviewed. These requests will be granted in the form of defined number of specimens of each type, from a selection criteria defined by final diagnosis and potentially other design criteria. Upon request approval, the specimens will be retrieved and delivered on dry ice to the Project Investigator. Please see section 2.5 about specimen allocation to non-Project Investigators.

1.0 Selection Criteria

Tissue and blood specimens for this research project will be obtained from women undergoing surgery for ovarian-related disorders at Swedish Hospital Medical Center by Pacific Gynecology Specialists surgeons or from women identified through the gynecological oncology service at the University of Washington Medical Center.

Over a period of two years, 500 women with appropriate diagnoses (70 ovarian cancer cases, 430 women with benign disease, or with no ovarian abnormalities) will be recruited for this research study. The age range of the participants will be between 20-80 years old. Only women undergoing ovarian related surgery will be invited to participate in the tissue and serum collection component of the research study. Minority representation will be reflective of the representation seen at PGS and the UW.

1.1 Recruitment

All appropriately identified PGS and UW patients scheduled to undergo surgery for ovarian related disorders, will be considered as potential candidates for this research study. These patients will be invited to participate in the study by the attending physician or nurse at the time of the pre-operative office visit. Interested patients will be provided written information describing the research study to review. Should the patient choose to participate in this protocol, the attending physician, nurse or a research study staff member will review the consent form and other required enrollment documents with the patient.

Patients will be given an opportunity to ask questions and address any concerns they may have about participating. Patients will be informed that the decision to participate will not affect their treatment in any way, and that in agreeing to participate, they reserve a right to terminate their participation at any time without prior notice.

1.2 Data Collection, Flow and Subject Identification

Women who consent to participate will be given a study enrollment packet. This packet includes a study letter or brochure, a participant enrollment form, a medical records release form, and a consent form to be completed at the pre-operative office visit. A self-administered 20-minute questionnaire with a self-addressed, stamped envelope will also be included in the enrollment packet. The patient may complete the questionnaire during the pre-operative office visit, or return it by mail at a later date. All materials will be pre-labeled with a unique packet identification number.

Clinic or research study staff will complete the portion of the patient enrollment form reserved for internal use only. This portion of the patient enrollment form indicates the date of scheduled surgery, institution of enrollment and identification of the enrolling physician, other medical or research study personnel. The enrolling staff member will ensure that the patient has fully completed the informed consent, and will note in the patient's chart that she has been approached and has agreed to participate in the research study. To ensure that this patient will not be approached again about the study, the enrolling staff will be responsible for noting patient participation or refusal in the patient's medical chart.

The completed enrollment forms and two copies of the signed informed consent will be sent to the Clinical, Statistical and Laboratory Coordination Core. The Data Coordinator will enter data from the enrollment forms into the study database on a daily basis. At the time that data from enrollment forms is entered into the study database, each participant will be given a unique participant number (UPN). All self-administered questionnaires, after completion by the participants, will be returned to the Core facility for editing and data entry.

The UPN will be used to label all data collection forms, requisition forms, and transport forms. A unique 6-digit number will be used to label all specimens, with a duplicate of

the label attached to the specimen collection form. The unique label number for each specimen will be linked to each participant's UPN in the specimen inventory database.

1.3 Risks to Subject

All tissue that is obtained for the purposes of this research study is collected only after it has been removed for the purposes of the surgical procedure that the participant is undergoing. All blood (no more than 40 cc) obtained for the purposes of this research study is collected by the anesthesiologist prior to or during the actual surgical procedure. This collection will not pose any additional risks to the participant. The participant may experience potential discomfort by not being given the results of the analyses to be conducted by the research project.

Any precautions possible will be taken to ensure that the participant's risks are minimized or eliminated. These include extensive precautions to maintain the confidentiality of all study records identifying patient information by enforcing and following strict protocols. These procedures include a pledge of confidentiality by all study personnel at Fred Hutchinson Cancer Research Center, Pacific Gynecology Specialists and the University of Washington, data handling procedures, network and password protection, and proper storage and handling of all files and specimens.

Study participants are informed that their personal identity will not be revealed in any publication or release of results. Study participants are also informed that representatives from the U.S. Army Medical Research and Materiel Command will have access to their study records, and may inspect the records of the research in their duty to protect human subjects in research.

1.4 Follow-up Procedures

1.4.a. **Incomplete Enrollment**

With adherence to stringent enrollment procedures, very little participant follow-up for this study is anticipated. However, follow-up may be required if a patient has not fully completed enrollment forms or has not returned the study questionnaire. In such situations, a written request, followed by one telephone call, will be made by the Study Coordinator. A letter will be mailed to the participant if their enrollment materials are incomplete or if their questionnaire has not been received by the Core within thirty days of their enrollment. After fourteen days, a follow up call will be made to the participant if she has not responded to the request. The call script included in the appendix of this protocol will be employed to inquire about the questionnaire or clarify ambiguous or incomplete information on the enrollment forms.

1.5 Medical Monitor

A medical monitor has been assigned to this research study. The medical monitor is Dr. Saul Rivkin of the Swedish Hospital Medical Center Tumor Institute. Dr. Rivkin is an exceptionally qualified physician, who is not associated with the protocol. Dr. Rivkin is fully able to provide medical care to the research subjects for conditions that may arise during the conduct of the study. A short biosketch and Dr. Rivkin's Curriculum Vitae is included with this protocol.

2.0 Specimen Collection

Each day, a report will be generated showing the most up-to-date information on scheduled surgeries for study participants. This report includes the patient name and UPN, surgeon, date, time and location of surgery.

The schedule and amount of specimen to be collected will vary throughout the study. It is anticipated that three to five collections from qualified participants will be conducted per week. The amount of tissue specimen collected for the purposes of the research study will also vary from 1 gram up to 5 grams. Only that which is not needed for the purposes of pathologic diagnosis will be available for ORCHID study collection. This will be determined by the clinical pathologist on a case by case basis.

For each scheduled surgery, a packet containing the UPN specimen collection and processing forms, and a copy of the informed consent, will be assembled and sent to the specimen collection team with the scheduled surgery report. In addition, the collection team will be provided with a pre-assembled specimen kit for tissue and blood collection. The Tissue Collection Specialist will be responsible for ordering, maintaining, and assembling supplies for the specimen kit. The specimen kit will include the following pre-labeled items:

- Three (3) biohazard bags (with foil) for snap frozen primary and metastatic tumor and normal specimens
- One (1) truncated embedding mold for primary tumor/tissue specimens frozen in OCT compound
- Three (3) 15 ml. formalin jars for fixed specimens
- One (1) STM tube for primary tumor tissue
- Two (2) 5 ml. lavender-top EDTA tubes for blood collection
- Three (3) 10 ml. red-top tubes for blood collection

- Ten pre-labeled cryovials for serum, plasma, and WBC pellet collection
- 4 lbs. dry ice
- Biohazard stickers and dry ice labels

2.1 Tissue Collection

The Tissue Collection Specialist, who maintains a log of scheduled surgeries, will work with operating room physicians and personnel to notify them prior to the beginning of a participating patient's surgery. During the entire surgical procedure, the attending surgeon and surgical personnel will be responsible for monitoring the patient's vital signs and condition.

Immediately after the surgeon has removed the necessary tissue and the pathologist has taken what is required for pathologic diagnosis, the Tissue Collection Specialist will be allowed to collect specimens from this removed tissue for the purposes of this study. The tissue samples will be processed according to the guidelines below:

Ovarian Tissue:

- Surgical specimens will be placed in labeled sterile containers containing 0.9% sodium chloride and transported by Tissue Collection Specialist into the processing area located in the frozen section room.
- Under the direction of a clinical pathologist, tissue necessary for clinical evaluation will be removed.
- Tissue used for the proposed studies will be selected from an area representative of the specimen and as free of necrosis as possible.
- In the case of normal ovaries, the surface epithelium will be manually scraped from the ovary and snap frozen, to minimize contaminating stromal tissue.

Frozen Tissue Amounts and Preparation:

- A minimum of 1gm and up to 5 gm of tissue, which will be divided into approximately 1 cm^3 sections Each section will be completely wrapped in aluminum foil and immersed in liquid nitrogen for a minimum of 3 minutes.
- Frozen tissues will then be placed in biohazard bags pre-labeled with the UPN and tissue type.
- Specimens will be stored on dry ice for transport to the core facility.
- For the OCT mold, truncated molds will be pre-labeled with the UPN using a SECURLINE permanent marker.

- Each mold will be partially filled with OCT medium and pre-cooled by holding over (not in) liquid nitrogen until OCT medium loses transparency.
- Approximately 1 gm of tissue will be placed in the mold, covered with OCT medium and immersed into liquid nitrogen until completely solid.
- Specimens will be placed into a UPN labeled biohazard bags and stored on dry ice for transport to the core facility.

Paraformaldehyde-Fixed Tissue Amounts and Preparation:

- A portion of tumor smaller than or equal to 1x1 cm and no thicker than 2 mm will be selected and placed in cold 4% paraformaldehyde and stored for 2 hours at 4°C.
- After a 2 hour fixation, the 4% paraformaldehyde will be discarded and replaced with cold 30% sucrose, and the sample will be stored at 4°C.
- Tissue will initially float in sucrose but when left overnight will sink.
- After the tissue has sunk, but no longer that 24 hours after fixation, the tissue will be imbedded in OCT as described in protocol for tissue preparation. (see above)
- The mold will be placed into a labeled biohazard bag and stored at -70° C until transfer to the liquid nitrogen freezer at the Core laboratory/repository .

2.2 Blood Collection

Prior to each surgery, patient consent will be obtained to collect blood. The research collection team will work with the anesthesiologist and notify him/her prior to the beginning of surgery that a patient is participating in the research study.

***Blood** Collection and Preparation:*

- All requisite serum collection vials and clot tubes will be prepared and labeled with the patient name, UPN, and date of collection prior to surgery.
- At the time of surgery, the anesthesiologist will collect up to 30 cc of whole blood in ~~a~~ non-heparinized, red-top tubes using a vacutainer and 21 ga needle just prior to the surgeon removing tissue samples.
- Blood will be allowed to stand for 30-120 minutes and then stored in the operating room on ice until it is transferred to the Core laboratory/repository.
- At the Core laboratory, the blood will be placed into a refrigerated centrifuge and spun for 10 minutes at 2500 rpm at 4°C.
- After centrifugation, the vials will be placed in styrofoam container holding ice and placed under a hood. The tops of the vials will be swabbed with alcohol and the serum will be removed using a sterile 21 gauge needle and syringe. The serum is then aliquoted into study-labeled 250 ul NUNC tubes.

- NUNC tubes will then be placed into a UPN labeled plastic baggie and stored on ice for transport to the Core laboratory/repository.

WBC Pellet Preparation:

- The anesthesiologist will collect an additional 10 cc of whole blood in lavender-top EDTA tubes using a vacutainer and 21 ga needle just prior to the surgeon removing tissue samples.
- Blood will be stored in the operating room on ice until it is transferred to the Core laboratory within six hours of collection.
- At the laboratory, 0.5 ml of this blood will be added to a 2.0 ml Sarstedt tube. Next, 1 ml of specimen wash solution is added to this tube to lyse the RBCs.
- The Sarstedt tube will then be centrifuged to pellet the WBCs.
- After the WBC pellets have been washed three times with wash solution, the pellets will be transferred to a prelabeled cryovial.
- These WBC pellet cryovials will then be placed in the -70° C for long-term storage.

2.3 Specimen Storage

All tissue and sera obtained by the specimen collection team will be placed on dry ice for transport to the Core facility. The Tissue Collection Specialist will transport all samples to the Core laboratory the same day as collected. Upon receipt at the Core, all frozen tissue specimens will be stored in a liquid nitrogen freezer. Serum, plasma and WBC pellets will be stored in a -70° C freezer.

2.4 GOG/External Specimens

Additional tissue specimens provided by the GOG will be treated as collected specimens. Upon receipt of frozen specimens (shipped overnight on dry-ice by the GOG), each patient will be assigned a PIN (unique PIN will be allocated to GOG samples) and all specimens will be stored in the liquid nitrogen freezer prior to processing and characterization. Any information accompanying the specimens such as date of collection, age of patient at time of surgery, pathology and histology information, and other non-identified demographic data will be entered into the tracking system.

2.5 Specimen Allocation

After characterization in the Laboratory Core, specimens will be made available to Project Investigators. After Project needs have been met, specimens may be made available to non-Project Investigators. In such circumstances, the non-Project

Investigators will be required to complete a review process for use of said specimens. All specimens transferred to non-Project Investigators must receive approval and/or certification from Study Investigators, and the FHCRC IRO. Specimens provided to commercial entities, or Investigators in collaboration with a commercial entities must also receive approval from the FHCRC Human Specimens Committee.

Non-Project Investigators and/or commercial entities will be asked to submit a proposal to this study's Investigators, stating the following: 1) the hypothesis to be tested 2) how the specimens will be used, 3) the amounts and types of specimens requested and 4) preliminary data. In addition, a biostatistical consult will be conducted to ensure that sample sizes are sufficient and that the study is sound in design.

If approved by study Investigators, non-Project Investigator(s) and/or commercial entities will be required to submit an Institutional Review Board application to the FHCRC IRO for research protocol review and approval.

If approved by the FHCRC IRO, and if not a commercial entity, or an investigator(s) involved in a collaboration with a commercial entity, the specimen request will be considered approved. If approved by the FHCRC IRO, and if a commercial entity, or an investigator(s) in a collaboration with a commercial entity, non -Project Investigators will be required to submit application to the FHCRC Human Specimens Committee for research protocol review and approval.

In either situation, upon request approval, the specimens will be retrieved and delivered on dry ice to the non-Project Investigator(s). The Project Coordinator, Suepattra May, will serve as the Repository Gatekeeper and will ensure that specimens and/or corresponding data are provided only to Investigators that are in full compliance with the application protocol. In addition, Ms. May will be responsible for all application materials and other paperwork associated with this process, including completed Confidentiality Pledges.

3.0 Specimen Processing Facility (Laboratory Core)

All tissue specimens will undergo processing at the Core laboratory facility before long-term storage and/or transport to project or non-project laboratories. All red top blood specimens will undergo processing at the Dynacare Laboratory of Pathology Stat Laboratory on the Swedish Medical Center Campus or at the Core laboratory facility before long-term storage and/or transport to project or non-project laboratories. All purple top (EDTA) blood products will undergo processing at the Core laboratory facility before long-term storage and/or transport to project or non-project laboratories

3.1 Equipment

The laboratory facility will be equipped with all necessary specimen processing equipment and supplies. Equipment for this facility includes a refrigerator, two freezers, MicroProbe IHC system, microcentrifuge, PCR hood, thermocycler, hybridization oven, gel electrophoresis tank and gel dryer, vacuum pump and trap.

The Research Technician will conduct all quality assurance of laboratory equipment and arrange for routine maintenance of equipment as recommended by manufacturers. All monitoring and maintenance checks of equipment will be recorded in individual maintenance logs. Freezer temperatures will be monitored daily using a thermometer linked to an alarm system and recorded daily by the Research Technician.

Each of the freezers will be equipped with an eight-hour CO_2 back-up system and a temperature-sensitive alarm system that alerts the building maintenance staff when the interior temperature reached a designated temperature. The CO_2 tank will be checked on a monthly basis to ensure that it has not been emptied. Each month, the alarm system will be tested to ensure that it will sound should the freezer temperature rise above -50°C. The Project Manager will be available for monitoring the freezers during non-office hours. In the case of equipment failure, a back-up freezer space will be available for specimens.

3.2 Laboratory Supplies

The Core will maintain, at minimum, a two-month inventory of specimen collection, processing and transport supplies. The Research Technician and Tissue Collection Specialist will be responsible for ensuring that the minimum requisite levels of supplies are available for the researchers and collection team.

3.3 Labeling

A number of labels specifying a unique specimen identification number will be used for collection and processing of each participant's tissue and blood specimen for the duration of the study. Labels will be used for all collection and processing forms, specimen containers and vials, and logs.

The freezer boxes for each specimen type will be labeled accordingly. Freezer boxes will be pre-labeled with appropriate specimen characterizations, so that as new specimen types are received, the specimens will be added to the appropriate freezer box.

3.4 Computerized Specimen Tracking Program

A computer database will serve as the specimen inventory system that tracks specimens collected by the Tissue Collection Specialist. This system will uniquely identify each separate item entered into the repository, document its location in the freezer, and its disposition as it is transferred to project laboratories for analysis.

3.5 Specimen Transport Preparation

After the specimen requisition process has been completed, the Research Technician will prepare the specimens for transport to the project laboratories. To ensure the integrity of the specimens, the freezer boxes will not be removed from the freezer for processing until all transport supplies are available for performing the transport procedure. The specimens will be packaged in a styrofoam box according to the following procedure:

- A 2" layer of pelleted dry ice nuggets will be placed on the bottom of the styrofoam box;
- The freezer boxes will be placed in a self-sealed, or waterproof sealed plastic bag;
- A biohazard symbol will be affixed to the outside of the plastic bag;
- The sealed plastic bags will be placed in the styrofoam box on top of the dry ice;
- Another 2 lbs. of pelleted dry ice will be layered on top of and around the plastic bags;
- Any empty spaces will be stuffed tightly with newspaper;
- The seams of the styrofoam box will be taped with waterproof tape. No scotch or masking type will be used;
- The styrofoam container will then be placed inside the transport carton;
- The transport carton will be labeled "Diagnostic Specimen" with a grease pen;
- The transport carton will be temporarily stored in the freezer until it is picked up for transport.

If it is noted that one of the vials or containers is cracked, the cracked container will be placed into another larger vial or container and may be transported separately in a sealable plastic bag, at the discretion and evaluation of the laboratory director.

The Research Technician or Data Coordinator will advise the appropriate project laboratory that a shipment of specimen is on its way. The project laboratory personnel will contact the Core Repository staff to inform them that the shipment has been received.

4.0 Specimen Characterization and Analysis

All tissues will be reviewed by the Core facility prior to their analyses in Projects 1 and 2. The review will include the frozen section of the biopsy material that has been submitted for further studies in Projects 1 and 2. For "normal" tissue, multiple levels of tissue submitted as normal will be reviewed. Tumors will be classified according to the WHO classification system as:

1) an epithelial tumors including serous (benign, borderline, malignant), mucinous (benign, borderline, malignant), endometrioid, clear cell, Brenner, mixed epithelial tumors, undifferentiated carcinomas, or unclassified epithelial tumors);
2) a sex cord-stromal tumors;
3) a lipid cell tumors;
4) a germ cell tumors;
5) a gonadoblastomas;
6) a soft tissue tumors (not specific to ovary);
7) an unclassified tumors, a metastatic tumors or
8) a tumor-like conditions.

All formalin-fixed, paraffin-embedded biopsy tissues will be examined by immunohistochemical techniques for the presence of the oncoproteins HER2/neu, Myc and for the intranuclear accumulation of mutant p53 proteins. DNA will also be extracted from fresh tissue and will be tested by PCR-SSCP analysis to screen for mutations in the p53 DNA. These IHC and mutation analyses will be performed utilizing the following protocols.

4.1 Immunohistochemistry (IHC)

- For all IHC assays, the Elite Vectastain ABC kit (Vector Laboratories) will be used with a specific primary antibody for each protein to be assayed.
- Sections four to six microns thick will be cut, mounted on silanated slides, dewaxed, and rehydrated.
- After blocking, the primary antibody will be added.
- The slides will then be washed and the secondary antibody, a biotinylated goat anti-mouse IgG, will be added.
- After a second wash, peroxidase-conjugated avidin will be added, washed and then reacted with the chromogen diaminobenzidine.
- Finally, the slides will be counter-stained with methyl green. Normal tissue will be stained as a negative control. An antibody against keratin AE1/AE3 (Boehringer Mannheim) will be used as a positive control.
- The intensity of staining of the cases will be determined (0 = not greater than the negative control, 1+ = light staining, 2+ = moderate staining, 3+ = heavy staining) and compared to the intensity of staining of the normal ovarian tissue.

4.1 PCR-SSCP for p53 Mutations

- All cases will be screened for mutations in the p53 gene by single-strand conformation polymorphism (PCR-SSCP) analysis.

- DNA will be purified from fresh tissue by Proteinase K digestion, phenol/chloroform extraction and ethanol precipitation.

- Mutations in exons 5-9 will be detected by means of PCR-SSCP analysis (4), amplifying 0.1-1.0 μg of DNA in separate reaction mixes with primer pairs for exons 5, 6, 7, and 8.

- The amplification products will be denatured and run on a polyacrylamide gel. Bands will be visualized by a silver stain.

- To further detect mutations in exons 7-9, a fifth primer pair will be used to generate a PCR fragment containing exons 7-9 which will be digested with the restriction enzyme MspI.

- Any variant band detected by PCR-SSCP analysis which do not conform to the pattern of the common p53 mutations (which serve as positive controls for the assay) will be cut out of the gel, eluted in TE-4, and directly sequenced using dye terminator reactions (utilizing the same primer pairs) and analyzed on an ABI sequencer.

4.3 Preparation of RNA from Peripheral Blood Samples for RT-PCR Assays

- Whole blood will be collected into EDTA anticoagulant tubes and 0.5 ml aliquots will be added to 1 ml of lysis buffer containing 0.4% detergent.

- The unlysed cells will be pelleted and washed two times with the lysis buffer.

- The pellets from 2 ml of whole blood will be combined and resuspended in 1 ml of UltraSpec RNA isolation reagent (Biotecx) and the total cellular RNA will be purified according to the manufacturer's protocol.

- One to two ug of total RNA will be added to a reverse transcriptase PCR reaction using primers specific for the protein of interest and the PCR amplicons will be detected by dot blot hybridization.

4.3 Radioimmunoassay for CA-125

One 250 ul aliquot of serum from the selected women will be obtained for CA-125 detection.

- Immunoradiometric assay of CA-125 levels will be performed using the commercially available RIA kit (Centocor, Malvern, PA.).

5.0 Safety Procedures

All Core personnel handling tissue and blood specimens are required to become familiar with and adhere to the applicable sections of the Protocol for Specimen Collection, Processing and Transport. A copy of any local and state requirements relating to the collection and processing of blood products must be on file with at the Statistical, Clinical and Laboratory Coordination Core.

All blood and tissue specimens will be handled as potentially infectious material. The Core has adopted the Universal Precautions for blood collection and processing. Universal Precautions refers to an approach to infectious disease control which assumes that every direct contact with body fluids is infectious. This approach requires that persons who may be in direct contact with body fluids be protected as though all body fluids contain blood-borne pathogens. All Core clinic and laboratory personnel will be guided by the universal precautions in order to protect all persons from parenteral, mucous membrane and non-intact skin exposures to blood borne pathogens.

It must be noted that any body fluid may contain microorganisms capable of transmitting disease. Therefore, appropriate protective attire must be worn where there is potential for direct contact with any body fluid or tissue. Core personnel will be required to change gloves and wash hands after handling laboratory specimens containing body fluids.

All procedures involving blood or other potentially infectious materials must be performed in a manner which minimizes splashing, spraying and aerosolization of these substances.

Core personnel will adhere to the following guidelines as regards each topic:

- *Hand Washing* - Employees must wash their hands:
  - Immediately after contact with blood or other infectious materials (even if gloves were worn);
  - Before and after using restroom facilities;
  - After removal of gloves and/or other protective clothing;
  - Upon leaving the work area where blood or other infectious materials are present.

- *Personal Protective Equipment* - Personal protective equipment such as fluid resistant gowns, gloves, goggles, and masks must be available and used in areas where blood and or other potentially infectious materials are handled. Supplies such as face shields, head and foot coverings must be available and used when invasive procedures are being carried out.

- *Accessibility of Equipment* - Appropriate protective clothing must be worn when the employee has a potential for exposure to blood and other potentially infectious materials.

- *Removal of Equipment* - Personal protective equipment (disposable clothing) must be removed immediately upon leaving the work area and placed in a labeled infectious waste container for disposal.

- *Gloves* - The use of disposable gloves is mandatory for procedures in which body fluids or other potentially infectious materials are handled. Gloves should be changed when contaminated and prior to entering common areas (such as elevators or restrooms. Latex or vinyl gloves are appropriate. Gloves must be worn when the Core personnel has the potential for direct skin contact with:
    - Blood;
    - Infectious materials;
    - Tissue;
    - Mucous membranes;
    - When handling items or surfaces soiled with blood or other infectious materials.

- Gloves should not be used if they are peeling, cracked or discolored, or if they have punctures, tears, or other evidence of deterioration. If an employee has an open cut or abrasion on the hand(s), the area must be protected with a Band-Aid underneath the glove.

- *Gowns* - Fluid resistant gowns or aprons must be worn if there is a potential for soiling ~~of~~ clothes with blood or other potentially infectious materials.

- *Surgical Caps or Hoods* - Surgical caps or hoods must be worn if there is a potential for splashing or spattering of blood or other potentially infectious materials on the head.

- *Fluid Proof Shoe Covers* - Fluid-proof shoe covers must be worn if there is a potential for shoes to become contaminated with blood or other potentially infectious materials.

- *Masks, Eye Protection and Face Shields* - Masks, eye protection, or chin-length face shields must be worn whenever splash, spray, spatter, droplets or aerosols of blood or other potentially infectious materials may be generated and there is a potential for eye, nose, or mouth contamination.

- *Spill Clean-Up* - All equipment and working surfaces must be properly cleaned and disinfected after contact with blood, tissue or other potentially infectious materials. Broken glassware which may be contaminated must be removed by mechanical means, such as tongs, cotton swabs or forceps. Chemical germicides and disinfectants should be used at recommended dilutions to decontaminate all spills of blood and other potentially infectious materials. All spills must be cleaned immediately while adhering to the following guidelines:
    - Gloves must be worn when wiping up a spill;

- An appropriate disinfectant must be used; and
- Disinfectants (at appropriate dilution) should be poured onto a paper towel for wiping up small spills.
- No trigger type spray bottles or other equipment which would aerosolize the disinfectant should be used.

- *Laundry* - All laundry is assumed to be contaminated. Personnel handling laundry must wear protective gloves. Laundry must be bagged at the location where it was used. If soaking through is likely, double bagging is required.

- *Waste Management* - Employees are required to wear gloves when handling any infectious waste.

- *Labeling* - A label showing the biohazard symbol will be affixed to all containers of infectious waste (i.e. biohazardous and medical waste), refrigerators, and freezers containing blood or other potentially infectious materials. The biohazard symbol must be black on an orange background.

- *Transportation* - All specimens or containers of blood and tissue will be transported within a secondary container (e.g., plastic bag or other container having a liquid tight seal). These materials will be placed in a secondary container and labeled with the biohazard symbol prior to being taken into common areas.

- *Food and Drink* - Eating, drinking, applying cosmetics or lip balm, and handling contact lenses are prohibited in laboratories and other work areas where blood or tissue, or other potentially infectious materials are present.

Please refer to the Appendix 7 – Safety Program Plan for more details.

6.0 Disposition of Data

All documents, data and study records collected for the purposes of this study will be stored indefinitely at the Fred Hutchinson Cancer Research Center. All researchers and staff with access to this information will follow procedures to prevent disclosure of information to anyone who is not an investigator on this study. This includes a pledge of confidentiality by all FHCRC, UW and PGS personnel; data handling procedures, network protection, password protection, proper storage and handling of all files and specimens, and secured facilities. A copy of the pledge of confidentiality is enclosed with this application.

6.1 Biostatistical Reviews

Biostatistical review of all project data will be conducted by the Core Co-Project Director, Garnet Anderson, PhD. Biostatistical reviews will be conducted to understand the behavior of new and known markers jointly. Analyses specific to data generated by the Core Laboratory and additional analyses of the consolidated project data will be performed. These analyses will include the following:

- Describing the joint and unique expression of p53, HER2/neu and Myc in tumor tissue, by disease status and stage.
- Describing the correlation between expression of p53, HER2/neu and Myc in peripheral blood and tumor tissue, by disease status and stage.
- Describing the relationship between serum CA-125 levels and the expression of p53, HER2/neu and Myc in tissue or peripheral blood.
- Describe the relationship between various clinical and epidemiological factors (e.g., disease stage, menopausal status, prior history of cancer, number of ovulatory cycles) and marker levels in blood.

7.0 Protocol Modification

Departure from protocol for individual subjects will not occur in this research study. The research investigators in this project acknowledge and accept their responsibility for protecting the rights and welfare of human research subjects and for complying with all human use and regulatory compliance as determined by the Institutional Review Office of the Fred Hutchinson Cancer Research Center and the Human Subjects Protection Division (HSPD). Research investigators will promptly report proposed changes in previously approved human subject research activities to both the IRO and HSPD. The proposed changes will not be initiated without IRB and HSPD review and approval.

7.1 Reporting of Serious and Unexpected Adverse Events

Serious and unexpected adverse experiences will be immediately reported by telephone to the USAMRMC Deputy Chief of Staff for Regulatory Compliance and Quality (301-619-2165) (non-duty hours call 301-619-2165 and send information by facsimile to 301-619-7803). A written report will follow the initial telephone call within 3 working days. Address the written report to the U.S. Army Medical Research and Materiel Command: ATTN: MCMR-RCQ, 504 Scott Street, Fort Detrick, Maryland 21702-5012.

8.0 Use of Information/Publications Arising From This Study

The personal identity of subject participants will not be revealed in any publication or release of results. All information/publications arising from this

study will be conducted in an ethical manner, as approved by the Institutional Review Office of the Fred Hutchinson Cancer Research Center.

9.0 Personnel to Conduct Project

Principal Investigator:	Nicole D. Urban, ScD	206-667-4677
Project Director	Garnet Anderson, PhD	206-667-4699
Project Director	Nancy Kiviat, MD	206-616-9740
Investigator	Charles Drescher, MD	206-587-0585
Investigator	Leona Holmberg, MD	206-667-6447
Investigator	Mary Anne Rossing, PhD	206-667-5041
Medical Monitor	Saul Rivkin, MD	206-386-2929

10.0 Signature of Principal Investigator

"I have read the foregoing protocol and agree to conduct the study as outlined herein."

Nicole Urban, ScD Date

Appendix J
Characteristics of Participants

1. Table 1: Characteristics of ORCHID Participants who completed the questionnaire
2. Table 2: Clinical Characteristics by Outcome (N=299)
3. Table 3: Clinical Characteristics by Outcome (N=244)
4. Analysis of Project 1RT-PCR data
5. Models 1-4

Table 1: Characteristics of ORCHID Participants who completed the questionnaire (N=299)

Characteristic	Normal	Benign Ovarian Pathology	LMP/ Borderline Ovarian Tumor	Outcome Ovarian Cancer	Other Cancer	Unknown	No Tissue Donation
Demographics							
Age median (range)	52 (28-78)	58 (26-84)	45 (34-70)	63.5 (41-90)	65 (36-86)	57 (35-81)	55.5 (22-85)
Race N(%)							
Native American	2 (3.1)	1 (3.2)	1 (9.1)	0 (0.0)	1 (3.3)	0 (0.0)	1 (1.7)
Asian	3 (4.7)	2 (6.5)	1 (9.1)	4 (6.7)	1 (3.3)	2 (4.6)	2 (3.5)
Black	0 (0.0)	1 (3.2)	0 (0.0	1 (1.7)	1 (3.3)	2 (4.6)	2 (3.5)
Caucasian	57 (89.1)	26 (83.9)	9 (81.8)	54 (90.0)	27 (90.0)	40 (90.9)	50 (86.2)
Other	2 (3.1)	1 (3.2)	0 (0.0)	1 (1.7)	0 (0.0)	0 (0.0)	3 (5.1)
Highest Educational Level N(%)							
<=11th grade	1 (1.6)	1 (3.1)	2 (18.2)	5 (8.5)	2 (6.7)	1 (2.3)	0 (0.0)
High School Grad. or GED	7 (10.9)	7 (21.9)	2 (18.2)	14 (23.7)	8 (26.7)	8 (18.2)	21 (36.2)
Some college/votech	25 (39.1)	10 (31.3)	5 (45.6)	21 (35.6)	11 (36.7)	18 (40.9)	17 (29.3)
College graduate	16 (25.0)	3 (9.4)	2 (18.2)	12 (20.3)	6 (20.0)	11 (25.0)	10 (17.2)
Graduate school/advanced degree	15 (23.4)	11 (34.5)	0 (0.0)	7 (11.9)	3 (10.0)	6 (13.6)	10 (17.2)
Marital Status N(%)							
Currently Married	47 (74.6)	19 (59.4)	6 (54.6)	21 (35.0)	15 (50.0)	23 (53.5)	38 (65.5)
Other	16 (25.4)	13 (40.6)	5 (45.5)	39 (65.0)	15 (50.0)	20 (46.5)	20 (34.5)
Area of Birth N(%)							
USA	55 (88.7)	30 (93.8)	10 (90.9)	51 (86.4)	29 (96.7)	38 (90.5)	55 (94.8)
Europe	2 (3.2)	1 (3.1)	0 (0.0)	4 (6.8)	0 (0.0)	1 (2.4)	1 (1.7)
Asia	3 (4.8)	0 (0.0)	1 (9.1)	2 (3.4)	1 (3.3)	2 (4.8)	1 (1.7)
Other	2 (3.2)	1 (3.1)	0 (0.0)	2 (3.4)	0 (0.0)	1 (2.4)	1 (1.7)
Work Status N (%)							
Work full/part time	45 (71.4)	18 (56.3)	9 (81.8)	29 (48.3)	12 (40.0)	28 (63.6)	35 (60.3)
Retired	10 (15.9)	10 (31.3)	0 (0.0)	24 (40.0)	11 (36.7)	7 (15.9)	14 (24.1)
Other	8 (12.7)	4 (12.5)	2 (18.2)	7 (11.7)	7 (23.3)	9 (20.5)	9 (15.5)
Height Median (range)	64 (59-71)	65.5 (58-72)	67 (61-70)	64 (53-70)	65.5 (56-70)	65 (58-70)	65 (60-72)
Weight Median (range)	145 (98-280)	160 (112-250)	160 (140-200)	149.5 (100-275)	149 (105-260)	145 (95-350)	159.5 (104-270)
BMI Median (range)	24.3 (17.9-51.3)	25.8 (19.6-41.5)	24.9 (22.6-33.9)	25.4 (19.1-44.0)	25.8(18.1-41.4)	25.7(18.5-50-3)	26.3 (18.9-43.0)

Table 1(continued): Characteristics of ORCHID Participants (N=299)

Characteristic	Normal	Benign Ovarian Pathology	LMP/ Borderline Ovarian Tumor	Outcome Ovarian Cancer	Other Cancer	Unknown	No Tissue Donation
Menstrual History							
Age at menarche median (range)	13 (9-17)	13 (9-17)	13 (12-17)	13 (11-16)	13 (11-15)	13 (9-16)	13 (9-17)
Ever Pregnant N(%)							
Yes	56 (87.5)	24 (77.4)	10 (90.9)	51 (86.4)	25 (83.3)	39 (90.7)	44 (75.9)
No	8 (12.5)	7 (22.6)	1 (9.1)	8 (13.6)	5 (16.7)	4 (9.3)	14 (24.1)
Age at First birth median (range)	24 (16-38)	23.5 (17-32)	24 (17-37)	22.5 (16-35)	24 (16-40)	23 (16-30)	23 (17-40
Total # pregnancies > 6 months							
None	4 (7.4)	5 (20.8)	1 (10.0)	3 (6.0)	1 (4.4)	5 (14.3)	2 (4.7)
1	9 (16.7)	3 (12.5)	1 (10.0)	7 (14.0)	5 (21.7)	8 (22.9)	6 (14.0)
2	20 (37.0)	8 (33.3)	2 (20.0)	17 (34.0)	8 (34.8)	12 (34.3)	18 (41.9)
3	17 (31.5)	6 (25.0)	3 (30.0)	13 (26.0)	3 (13.0)	5 (14.3)	8 (18.6)
4+	4 (7.4)	2 (8.3)	3 (30.0)	10 (20.0)	6 (26.1)	5 (14.3)	9 (20.9)
Ever Breastfed N(%)							
Yes	33 (52.4)	9 (29.0)	6 (54.6)	29 (48.3)	12 (40.0)	19 (44.2)	23 (40.4)
No	30 (47.6)	22 (71.0)	5 (45.6)	31 (51.7)	18 (60.0)	24 (55.8)	34 (59.6)
Total months breastfed median (range)	8 (1-34)	7.5 (1-22)	8.5 (1-38)	6 (1-80)	7.5 (1-90)	7 (1-45)	7.5 (2-43)
Ever use birth control pills (BCP)							
Yes	47 (73.4)	20 (62.5)	7 (63.6)	32(55.2)	19 (63.3)	28 (65.1)	36 (64.3)
No	17 (26.6)	12 (37.5)	4 (36.4)	26 (44.8)	11 (36.7)	15 (34.9)	20 (35.7)
Total months BCP median (range)	60 (0-324)	72 (0-276)	36 (12-144)	48 (0-300)	48 (0-216)	60 (0-336)	72 (12-240)
Hysterectomy N (%)							
Yes	18 (34.6)	16 (57.1)	3 (33.3)	29 (56.9)	12 (42.9)	20 (50.0)	20 (37.7)
No	34 (65.4)	12 (42.9)	6 (66.7)	22 (43.1)	16 (57.1)	20 (50.0)	33 (62.3)
Age at hysterectomy median (range)	50 (29-76)	43.5(32-71)	63 (41-70)	52 (28-76)	53.5 (35-78)	48 (30-79)	49 (32-84)
Age at last period median (range)	48 (28-61)	44.5(26-73)	44 (33-52)	48 (28-57)	50 (35-56)	46 (30-56)	47 (22-60)
Number ovulatory cycles median (range)	373 (78-578)	340 (55-727)	292 (124-428)	397 (13-557)	329 (199-543)	375 (91-510)	377 (51-575)
Ever used hormone replacement therapy N(%)							
Yes	31 (48.4)	20 (62.5)	4 (36.4)	41 (68.3)	12 (41.4)	23 (53.5)	29 (50.0)
No	33 (51.6)	12 (37.5)	7 (63.6)	19 (31.7)	17 (58.6)	20 (46.5)	29 (50.0)
Total months used HRT median (range)	48 (0-240)	84 (0-456)	6 (0-12)	96 (0-564)	42 (0-648)	96 (0-324)	36 (0-348)

Table 1(continued): Characteristics of ORCHID Participants (N=299)

Characteristic	Normal	Benign Ovarian Pathology	LMP/ Borderline Ovarian Tumor	Outcome Ovarian Cancer	Other Cancer	Unknown	No Tissue Donation
Personal Medical History							
Ever diagnosed with diabetes N(%)							
Yes	3 (4.8)	5 (15.6)	2 (18.2)	1 (1.7)	4 (13.3)	3 (7.1)	3 (5.2)
No	59 (95.2)	27 (84.4)	9 (81.8)	57 (98.3)	26 (86.7)	39 (92.7)	55 (94.8)
Ever diagnosed with inflamed bowel synd. N(%)							
Yes	7 (11.1)	2 (6.9)	0 (0.0)	7 (12.1)	2 (6.7)	4 (10.3)	4 (7.0)
No	56 (88.9)	27 (93.1)	11 (100.0)	51 (87.9)	28 (93.3)	35 (89.7)	53 (93.0)
Ever diagnosed with fibroids in uterus N(%)							
Yes	34 (54.8)	22 (68.8)	4 (36.4)	19 (33.3)	6 (20.7)	13 (31.7)	23 (40.4)
No	28 (45.2)	10 (21.3)	7 (63.6)	38 (66.7)	23 (79.3)	28 (68.3)	34 (59.7)
Ever diagnosed with endometriosis N(%)							
Yes	6 (10.9)	3 (9.7)	0 (0.0)	5 (9.8)	4 (14.3)	4 (11.1)	8 (14.3)
No	49 (89.1)	28 (90.3)	9 (100.0)	46 (90.2)	24 (85.7)	32 (88.9)	48 (85.7)
Ever diagnosed with benign breast disease N(%)							
Yes	13 (21.7)	14 (46.7)	3 (27.3)	10 (18.5)	7 (23.3)	15 (36.6)	10 (17.2)
No	47 (78.3)	16 (53.3)	8 (72.3)	44 (81.5)	23 (76.7)	26 (63.4)	48 (82.7)
Ever diagnosed with polycystic ovarian dis. N(%)							
Yes	3 (6.3)	0 (0.0)	2 (25.0)	2 (5.0)	1 (3.6)	2 (5.6)	3 (6.1)
No	45 (93.8)	23 (100.0)	6 (75.0)	38 (95.0)	27 (96.4)	34 (94.4)	46 (93.9)
Ever diagnosed with ovarian cyst N(%)							
Yes	27 (51.9)	16 (64.0)	6 (85.7)	10 (23.3)	3 (10.7)	19 (50.0)	20 (38.5)
No	25 (48.1)	9 (36.0)	1 (14.3)	33 (76.7)	25 (89.3)	19 (50.0)	32 (61.5)
Ever diagnosed with breast cancer N(%)							
Yes	8 (12.5)	3 (9.4)	0 (0.0)	5 (8.5)	0 (0.0)	5 (11.4)	2 (3.5)
Age diagnosed with breast cancer median (range)	48.5 (32-59)	50 (35-53)	0	51 (36-62)	0	54 (53-63)	46 (30-62)

Table 1(continued): Characteristics of ORCHID Participants (N=299)

Characteristic	Outcome						
	Normal	Benign Ovarian Pathology	LMP/ Borderline Ovarian Tumor	Ovarian Cancer	Other Cancer	Unknown	No Tissue Donation
Family History: # 1st degree relatives with:							
Breast Cancer N(%)							
0	52 (81.3)	24 (75.0)	10 (90.9)	52 (86.7)	25 (83.3)	39 (88.6)	52 (89.7)
1+	12 (18.8)	8 (25.0)	1 (9.1)	8 (13.3)	5 (16.7)	5 (11.4)	6 (10.3)
Ovarian Cancer N(%)							
0	63 (98.4)	32 (100.0)	11 (100.0)	58 (96.7)	29 (96.7)	42 (95.5)	55 (94.8)
1+	1 (1.6)	0 (0.0)	0 (0.0)	2 (3.3)	1 (3.3)	2 (4.6)	3 (5.2)
Prostate Cancer N(%)							
0	59 (92.2)	28 (87.5)	10 (90.9)	54 (90.0)	27 (90.0)	37 (84.1)	55 (94.8)
1+	5 (7.8)	4 (12.5)	1 (9.1)	6 (10.0)	3 (10.0)	7 (15.9)	3 (5.2)
Lifestyle							
Smoking N(%)							
Current smoker	6 (9.4)	3 (9.7)	1 (9.1)	8 (13.3)	1 (3.5)	8 (18.6)	3 (5.2)
Former smoker	21 (32.8)	11 (35.5)	4 (36.4)	19 (31.7)	14 (48.3)	12 (27.9)	22 (37.9)
Never smoked	37 (57.8)	17 (54.8)	6 (54.6)	33 (55.0)	14 (48.3)	23 (53.5)	33 (56.9)
Ever drink alcohol N (%)							
Yes	47 (73.4)	24 (75.0)	8 (72.3)	38 (63.3)	21 (72.4)	26 (60.5)	35 (61.4)
No	17 (26.6)	8 (25.0)	3 (27.3)	22 (36.7)	8 (27.6)	17 (39.5)	22 (38.6)

Table 2: Clinical Characteristics by Outcome (N=299)

Characteristic	Outcome						
	Normal	Benign Ovarian Pathology	LMP/ Borderline Ovarian Tumor	Ovarian Cancer	Other Cancer	Unknown	No Tissue Donation
Presenting symptoms N (%)							
Pain	26 (40.6)	10 (31.3)	2 (18.2)	43 (71.7)	6 (20.0)	20 (45.6)	2 (3.5)
Distention	12 (18.8)	6 (18.8)	3 (27.3)	39 (65.0)	5 (16.7)	20 (45.6)	2 (3.5)
Bleeding	23 (35.9)	6 (18.8)	3 (27.3)	4 (6.7)	17 (56.7)	5 (11.4)	0 (0.0)
Fatigue	1 (1.6)	2 (6.3)	0 (0.0)	10 (16.7)	2 (6.7)	5 (11.4)	0 (0.0)
Dyspepsia	2 (3.2)	0 (0.0)	2 (18.2)	14 (23.3)	2 (6.7)	9 (20.5)	0 (0.0)
Weight Change	5 (7.1)	2 (6.3)	1 (9.1)	13 (21.7)	2 (6.7)	5 (11.4)	1 (1.7)
Bladder Changes	13 (20.3)	3 (9.4)	0 (0.0)	6 (10.0)	1 (3.3)	7 (15.9)	0 (0.0)
Bowel Changes	7 (10.9)	3 (9.4)	2 (18.2)	13 (21.7)	1 (3.3)	8 (18.2)	1 (1.7)
Other	11 (17.2)	2 (6.3)	1 (9.1)	10 (16.7)	2 (6.7)	6 (13.6)	1 (1.7)
Pre-operative CA-125 median (range)	12 (2-84)	33.6 (9.6-58)	109.9 (4-1587)	427 (8.8-9060)	338(338-338)	44.5 (5.3-4667)	611.2 (578.3-644)

Table 3: Clinical Characteristics by Outcome (N=244)

Characteristic	FIGO Stage				
	0	I	II	III	IV
Grade N(%)					
Well differentiated	1 (25.0)	3 (42.9)	0 (0.0)	0 (0.0)	0 (0.0)
Moderately differentiated	1 (25.0)	4 (57.1)	0 (0.0)	5 (20.8)	2 (28.6)
Poorly differentiated	2 (50.0)	0 (0.0)	1 (100.0)	19 (79.2)	5 (71.4)
Histology					
Ovarian Cancer	2 (1.2)	8 (47.1)	2 (50.0)	41 (87.2)	7 (70.0)
LMP	0 (0.0)	9 (52.9)	2 (50.0)	0 (0.0)	0 (0.0)
Unknown	37 (22.3)	0 (0.0)	0 (0.0)	5 (10.2)	2 (16.7)
HER-2 median (range)	0 (0-8.2)	0 (0-3.3)	1.1 (0.6-2.4)	0 (0-9.2)	0.1 (0-2.7)

Analysis of Project 1RT-PCR data

	NORMAL				CANCER					Modified	Ranking	
Gene	Mean	StdDev	Minimum	Maximum	Mean	StdDev	Minimum	Maximum	T	T	Original	Revised
SLPI	0.00	0.00	0.00	0.00	58.05	79.41	0.00	312.32	2.83	Infinity	15	1
HE4	0.04	0.04	0.01	0.15	32.04	47.13	0.03	190.62	2.63	2282.48	20	2
Mesothelin	1.01	0.89	0.01	3.52	477.99	1125.67	0.00	4453.90	1.64	1519.46	43	3
Mucin1	0.11	0.12	0.02	0.53	31.44	50.97	0.34	206.90	2.38	719.25	26	4
Folate BP	1.06	0.99	0.10	4.19	79.56	86.19	1.71	271.80	3.53	223.45	7	5
CD24	0.75	1.62	0.00	6.80	109.62	172.49	0.90	690.26	2.44	189.22	25	6
Keratin8	0.20	0.14	0.02	0.46	7.35	11.06	0.08	40.34	2.50	144.47	24	7
ESE-1	0.02	0.05	0.00	0.20	1.73	1.91	0.00	5.56	3.47	101.67	9	8
Ku80	0.12	0.48	0.00	1.99	11.79	31.22	0.00	110.14	1.45	68.41	49	9
Lipocalin2	0.50	1.09	0.05	4.59	24.27	36.34	1.38	148.34	2.53	61.41	23	10
BRCA1	0.13	0.10	0.00	0.28	2.29	6.01	0.06	23.76	1.39	58.95	50	11
oviductGP	0.11	0.17	0.01	0.66	3.02	10.35	0.00	40.39	1.09	48.96	57	12
p53	0.41	0.51	0.00	1.00	6.87	4.60	0.00	10.00	5.41	35.92	1	13
Ferritin H	4.20	5.85	1.48	26.30	46.05	158.03	0.32	617.05	1.03	20.20	60	14
Enolase	7.03	4.45	1.31	17.39	37.95	24.07	1.51	92.42	4.90	19.61	3	15
KIAA0762	5.12	2.72	1.47	11.13	22.58	20.08	0.23	56.21	3.34	18.14	10	16
T000M-07-F19	2.33	1.15	0.39	4.51	9.61	10.30	0.00	39.72	2.72	17.92	19	17
Ryudocan (no T069c)	4.21	3.21	0.79	11.17	22.22	21.76	0.00	70.50	3.18	15.84	12	18
RIG-E	97.32	45.01	16.59	204.47	348.61	244.00	16.47	854.06	3.93	15.76	5	19
p27	23.93	31.65	0.92	106.81	198.95	140.58	17.94	448.55	4.72	15.61	4	20
TGF beta 1	0.10	0.08	0.02	0.27	0.55	1.63	0.03	6.43	1.07	15.55	59	21
T000-09-c15	1.87	1.49	0.12	4.62	8.81	12.63	0.21	50.63	2.12	13.15	32	22
MR	0.75	0.72	0.00	2.58	3.54	1.94	0.00	7.06	5.24	10.88	2	23
GAPDH	14.16	8.35	0.68	31.19	43.41	27.92	1.76	94.29	3.91	9.89	6	24
IGF BP2	10.99	7.72	0.25	26.95	36.64	45.76	0.00	144.46	2.14	9.38	31	25
MCAF	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.64	9.32	44	26
BRCA2	1.07	0.68	0.00	2.19	3.31	2.65	0.30	8.79	3.18	9.26	11	27
GPR39	0.01	0.02	0.00	0.06	0.05	0.07	0.00	0.25	2.80	8.82	16	28
PAX2	0.00	0.01	0.00	0.05	0.04	0.07	0.00	0.21	1.89	8.04	34	29
T000M-171-i11	2.03	0.93	0.83	3.74	4.60	3.81	1.04	16.52	2.54	7.76	22	30
MAT-1	0.99	0.54	0.20	2.20	2.41	1.84	0.00	6.49	2.90	7.38	14	31
T000M-97-E17	0.39	0.23	0.04	0.77	0.97	0.96	0.10	3.36	2.29	7.18	27	32

Ross' suggestion

10/27/2000

Analysis of Project 1RT-PCR data

	NORMAL				CANCER				Modified		Ranking	
Gene	Mean	StdDev	Minimum	Maximum	Mean	StdDev	Minimum	Maximum	T	T	Original	Revised
T000M-06-B19	1.71	1.60	0.05	6.67	5.58	6.65	0.18	24.15	2.20	6.81	29	33
T000M-60-F20	0.03	0.05	0.00	0.14	0.14	0.14	0.00	0.48	2.76	6.15	17	34
H2N (2nd ref)	2.10	1.92	0.00	5.81	6.21	8.55	0.00	31.37	1.82	6.02	36	35
T000M-106-H01	0.34	0.43	0.06	1.97	1.23	0.89	0.15	2.82	3.51	5.81	8	36
ProgBP	0.53	0.36	0.05	1.70	1.24	1.21	0.38	4.26	2.21	5.55	28	37
IGF2	0.74	0.68	0.00	3.09	2.03	3.21	0.00	10.55	1.53	5.36	47	38
14.3.3	3.48	3.45	0.31	12.74	9.78	10.82	0.00	42.54	2.16	5.16	30	39
hose-060c2403	8.83	3.20	1.90	14.34	14.63	8.17	1.77	37.39	2.58	5.11	21	40
T000M-76-K21	11.19	10.25	0.14	40.41	29.54	20.57	3.43	66.00	3.13	5.05	13	41
MAGE E1	69.86	38.59	14.65	120.06	129.90	119.34	2.51	451.87	1.86	4.39	35	42
t820-02-o1102	0.69	0.38	0.19	1.55	1.26	1.21	0.00	4.29	1.76	4.22	39	43
ZFP161	0.28	0.21	0.00	0.75	0.58	0.63	0.00	2.15	1.80	4.17	37	44
T000M-139-A09	0.79	0.25	0.34	1.19	1.16	0.46	0.26	1.82	2.72	4.13	18	45
AA447275	0.34	0.17	0.13	0.72	0.57	0.67	0.09	2.48	1.29	3.81	54	46
KIAA0991	4.30	4.52	0.00	19.43	10.33	11.02	0.26	44.63	1.98	3.76	33	47
KIAA0512	2.03	1.20	0.35	4.38	3.52	3.19	0.15	12.02	1.70	3.48	40	48
Calgizzarin	13.25	9.39	2.37	35.07	24.67	25.34	0.00	88.03	1.65	3.43	42	49
KIAA0952	1.56	1.11	0.00	3.59	2.79	2.45	0.31	7.40	1.79	3.13	38	50
T000M-187-K19	19.32	75.38	0.00	311.76	101.88	339.95	0.00	1316.90	0.92	3.09	63	51
1-4	2.24	4.21	0.14	15.55	6.76	18.78	0.10	73.67	0.91	3.03	64	52
T000M-26-I14	0.36	0.23	0.00	0.88	0.58	0.48	0.00	1.83	1.67	2.80	41	53
H53727	0.18	0.28	0.00	1.10	0.46	0.63	0.00	2.07	1.53	2.72	46	54
T000M-14-M15	0.78	0.47	0.03	1.62	1.18	1.25	0.18	4.36	1.18	2.41	56	55
PTEN	5.52	5.24	0.00	17.34	9.96	12.85	0.00	48.92	1.25	2.39	55	56
MDC15	0.51	0.75	0.00	2.54	1.11	1.27	0.01	3.70	1.60	2.25	45	57
E16	14.14	35.87	0.00	149.79	40.92	96.40	0.00	380.96	1.02	2.11	61	58
T000M-34-L01	12.10	0.95	9.96	13.65	12.77	1.46	10.47	15.45	1.51	1.99	48	59
S31iii125-2	18.23	13.12	1.21	49.42	27.04	28.84	3.56	120.62	1.09	1.89	58	60
98118-D01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	1.36	1.85	51	61
KIAA0263	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.03	0.59	1.79	69	62
H2N	0.41	0.51	0.00	1.00	0.73	1.75	0.00	5.00	0.69	1.79	68	63
Actin	35.87	40.18	0.25	126.74	59.56	56.99	0.10	140.84	1.34	1.66	52	64

Ross' suggestion

10/27/2000

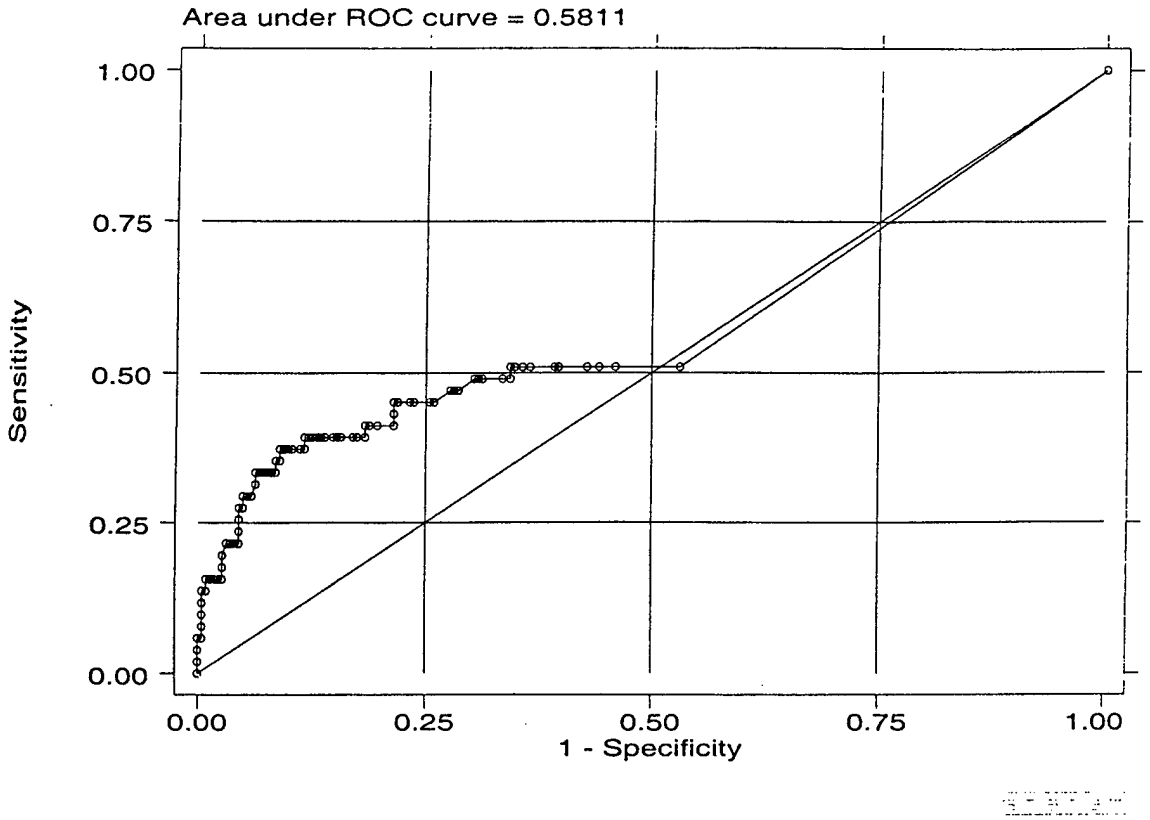
Analysis of Project 1RT-PCR data

	NORMAL				CANCER				Modified		Ranking	
Gene	Mean	StdDev	Minimum	Maximum	Mean	StdDev	Minimum	Maximum	T	T	Original	Revised
S31iii125	8.01	6.75	0.60	26.12	11.07	6.57	1.77	25.80	1.30	1.28	53	65
PLTP	123.14	75.57	12.45	312.87	156.25	147.00	5.01	462.15	0.79	1.24	67	66
T000M-16-A21	1.30	0.99	0.10	4.41	1.73	1.34	0.11	5.57	1.01	1.21	62	67
Bamacan	3.40	2.53	1.22	11.47	4.32	3.76	0.50	14.99	0.80	1.03	66	68
TRC8	4.13	3.82	0.65	16.63	5.22	3.13	0.85	10.25	0.89	0.81	65	69
SP5 (noT069c)	20.42	6.73	6.93	30.78	22.18	12.94	0.00	57.31	0.47	0.74	71	70
c-myc	1.47	2.84	0.00	11.43	2.06	3.05	0.00	10.78	0.57	0.59	70	71
T000M-31-N08	2.16	1.91	0.00	6.26	2.34	2.05	0.00	7.64	0.26	0.27	72	72
N000-11-E24	0.02	0.02	0.00	0.08	0.02	0.04	0.00	0.15	0.10	0.15	73	73
Kadereit	0.73	0.57	0.13	2.26	0.65	0.56	0.06	2.05	-0.41	-0.41	74	74
SAS	5.39	16.34	0.00	65.38	1.90	2.29	0.05	7.70	-0.87	-0.60	76	75
BA46	55.59	31.06	7.09	100.98	48.50	50.31	7.56	187.37	-0.47	-0.64	75	76
TLE4	2.24	1.97	0.19	6.84	0.89	0.97	0.04	3.66	-2.49	-1.93	77	77
flj10561 noStandard	287.76	167.22	0.00	578.04	139.81	155.09	0.00	447.09	-2.60	-2.50	78	78

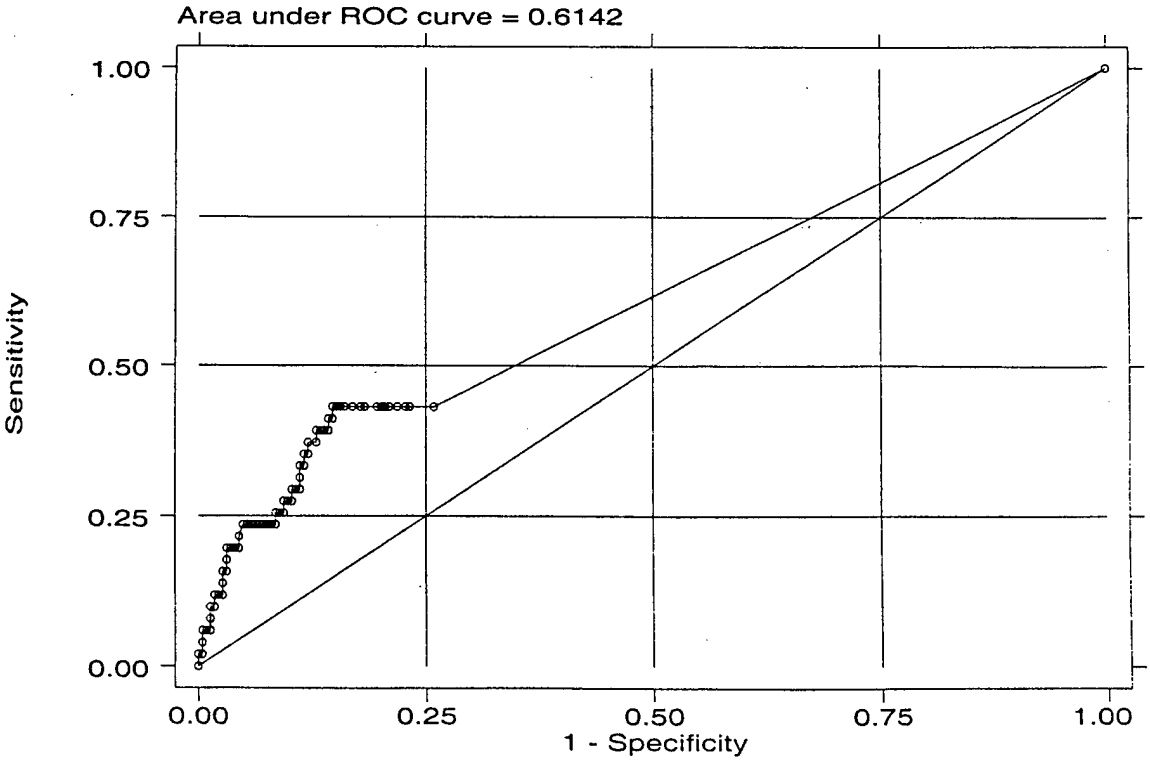
Ross' suggestion

10/27/2000

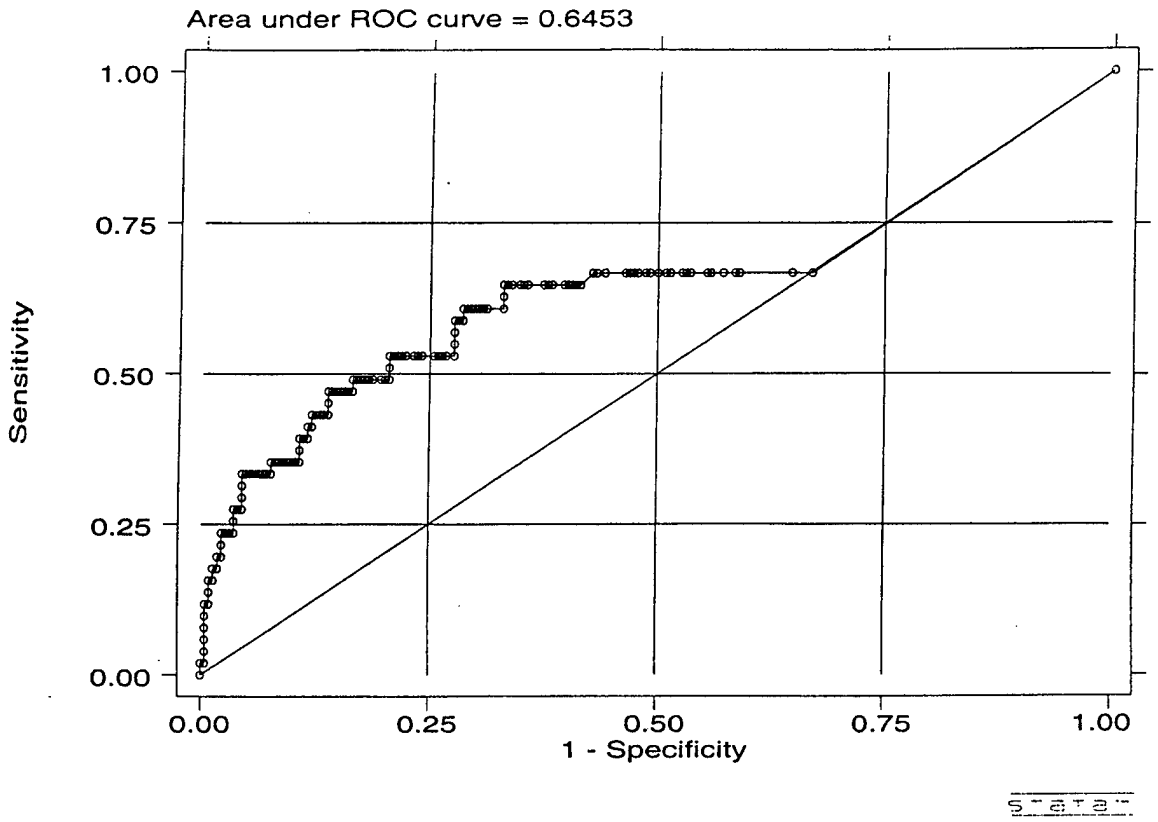
Model 1: logit (p) = $\beta_0 + \beta_1$ ln(p53 + 1)



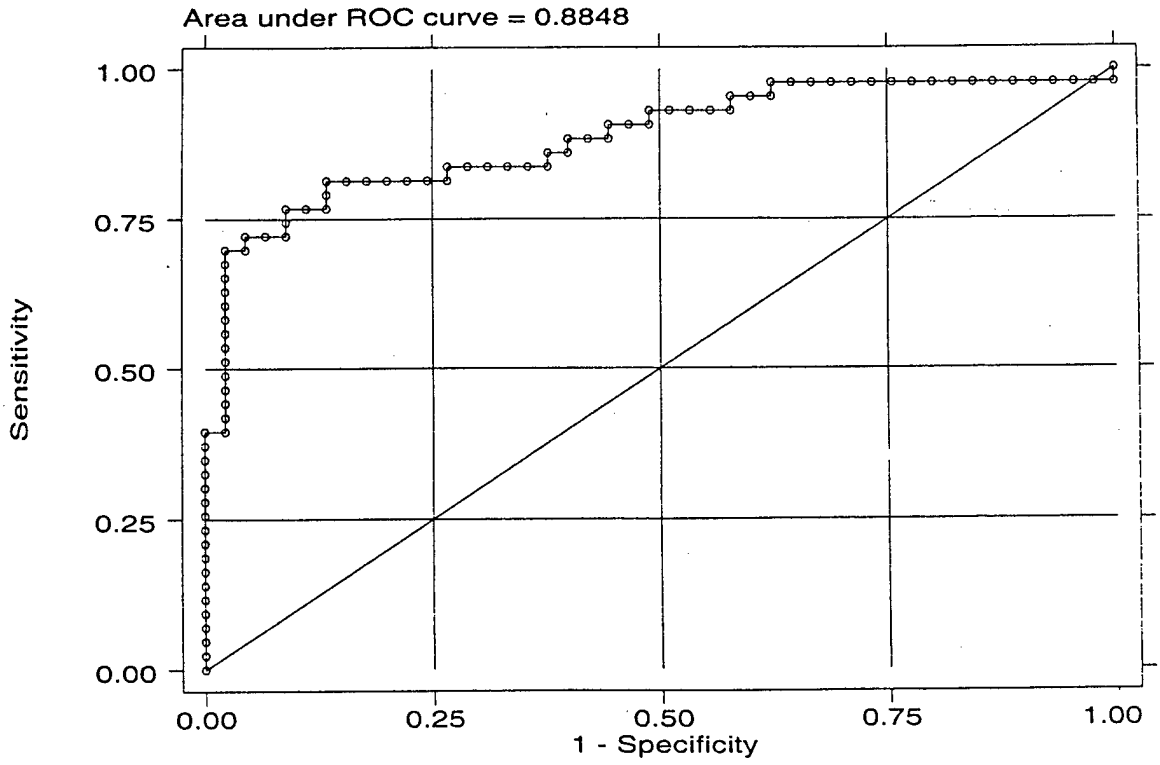
Model 2: logit (p) = $\beta_0 + \beta_1$ ln(h2n + 1)



Model 3: logit (p) = $\beta_0 + \beta_1$ ln(p53 + 1) + β_2 ln(H2N + 1)



logit (p) = $\beta_0 + \beta_1$ Age + β_2 ln(CA-125 + 1



Model 4: logit (p) = $\beta_0 + \beta_1$ Age + β_2 ln(CA-125+1) + β_3 ln(p53 + 1)+ $\beta 4$ ln(H2N + 1)