Commonwealth Technology Research Fund
Sponsored Project

Title:

Cancer Genomics and Development of Diagnostic Tools and Therapies

Virginia Commonwealth University

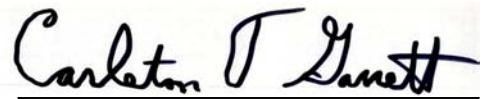
Annual Report

Grant # SE2002-02

7/1/01 -- 12/31/01

Submitted by:

Dr. Carleton Garrett
Director, CTRF Cancer Genomics Project
Dept. of Pathology
(for Dr. Marsha R. Torr)
(Vice President for Research)
Virginia Commonwealth University
1101 East Marshal Street
Richmond, Virginia



(Signature)

The Purpose of this Grant

The purpose of this grant is to recruit extramural funding for research and to develop new business opportunities for the Commonwealth of Virginia. This will be done through development of infrastructure and collaboration between two Virginia universities - Virginia Commonwealth University and George Mason University - and a major health care provider in Virginia – Inova. The focus of this partnership is the development of new diagnostic tools and therapies for use in patients with cancer. This collaboration will enable creation of new intellectual property and enhance competitiveness in recruitment of extramural funding and development of new business opportunities.

Background

Cancer is the second leading cause of death in the nation resulting annually in 553,400 deaths nation wide and 13,000 deaths in Virginia (http://www.cancer.org/eprise/main/docroot/STT/stt_0). These grim statistics indicate that although progress has been made on many fronts in the war against cancer, much remains to be done.

Cancer arises when normal cells within our body begin to behave in an uncontrolled manner. The cancer cells continuously replicate forming masses of cells, which are referred to as *tumors*. The cancer cells also spread throughout the body forming tumors at distant sites, which are called *metastases*. This loss of control occurs because some of the 10,000 or more instructions that controls cell behavior becomes corrupted. These 10,000 or more instructions arises from the genes in the cell's nucleus. The instructions are referred to as *messenger RNA* (mRNA) and the process of creating the instructions is referred to as *gene expression*. It is now clear that the instructions become corrupted because certain important *control genes* in the cancer cell become either inactivated (turned off), or in some cases, inappropriately activated (turned on) by mutations in their DNA. The end result of these mutations of cellular control genes is that the cell now begins to receive instructions that cause its behavior to change from well controlled normal cells to that of an uncontrolled cancer cells.

Over the past 20 years, our understanding has increased dramatically regarding the changes that occur in these sets of instructions between cancer and normal cells. However, the direct application to patient care of this increased understanding has been rather limited. This is due in part to the fact that development of human cancers is a complex process involving mutations in the DNA of multiple control genes with resulting similar complex changes in the instructions that control cell behavior. Unfortunately, the standard experimental techniques that have been available for studying how a specific gene or its instruction controls cell behavior have only been able to evaluate the actions of one or a few genes at a time. Thus, scientists may determine that in a particular model system, that is, a cancer cell line growing in tissue culture or in an experimental animal, that a drug or chemical suppresses the activity of one or a few genes. They also observe that this decreased gene activity causes the cancer cells in the model system to die.

However, until recently the scientists have not been able to view simultaneously the impact of the drug or chemical on the activity of the overall set of instructions that the cell uses to regulate its behavior. Cancer cells in humans have a redundancy of mechanisms for maintaining their most important biological functions such as cell growth and replication. Thus, while the drug treatment may cause the cells growing in a culture dish to die, the cells in a human tumor may express instructions not measured by the scientists which allow the cancer cells in the human to survive.

Recently, new technologies such as microarray analyses have become available which can simultaneously evaluate the thousands of instructions that are present in normal and cancer cells. Theoretically, this presents means of circumventing the problem described above. If a cancer cell is able to alter its instructions and thereby its behavior and in doing so escape the effect of a given therapeutic treatment, we are at least in a position to identify its new instructions and these new instructions can become the targets of new therapeutic agents and treatments. Although the clinical utility of this approach at this juncture is unknown, the limited data available is encouraging (1-4).

Research Plan

The Grant proposes to create new intellectual property through the study of four types of human cancer using microarray technology. That is, it will measure the sets of instructions present in human tumors and then determine which instructions are predictive of good outcome[1] for the patient and which are predictive of a bad outcome. The four types of cancer that will be evaluated are cancers of the brain, breast, ovary and hematopoetic system (leukemias and lymphomas).

While the new technology provides us with an opportunity to gain a more accurate and potentially a more clinically useful understanding of the origins and behavior of human cancer, it raises a host of *technical* and some *ethical* problems. **Efforts to develop an infrastructure to address these technical and ethical issues has been a major focus of this grant over the preceding 6 month period.**

Infrastructure Development

Tissue Acquisition Service

The first issue that the study must deal with is the fact that it will work with human tissue. Human tissue samples are removed from a patient as part of a diagnostic and/or therapeutic procedure (i.e., a procedure whose main purpose is to diagnose and treat the patient's clinical cancer). The standard of medical practice for evaluating such specimens is histopathologic examination by a pathologist. This examination establishes the histopathologic diagnosis and pathologic stage of the patient's disease, which are fundamental for determining the patient's subsequent therapy and prognosis. A second

[1] A good outcome may be defined as increased survival time, an earlier stage at the time of diagnosis, or improved quality of life or function in cancer patients. Poor outcomes may be defined as shorter survival time, later stage at diagnosis, or decreased function among cancer patients.

issue related to the use of human tissue is the need to inform the patient of the planned research use of their residual tissue specimen. Candidate patients need to be informed of the project through the formal process of IRB approved informed consent. Patients who choose not to allow use of their residual specimen for research will not have their residual specimen examined. A third issue relates to the need to obtain a tissue sample that will be useful to the research project. Once tissue has been removed from a patient, the cells present in tissue begin to alter the instructions that control their behavior (5). Thus it becomes important from a research perspective to freeze a piece of the tumor tissue as soon as it is removed from the patient. However, since it is important to not disrupt the specimen prior to review by a pathologist, the sample must be transported as quickly as possible to the pathology laboratory in order to achieve the optimum in patient care and preservation of the specimen for use in the research project.

Clinical and Laboratory Data Analysis

As described above, the use of microarray analysis to study the thousands of cellular instructions controlling cell behavior offers the possibility for improving care for cancer patients. The question, however, is how will we know that a research finding actually adds new understanding to caring for the cancer patient and does not just provide the same information that we currently obtain from standard clinical and laboratory examinations? In order to address this question, we must consider all of the standard clinical and laboratory information that is currently used to describe and classify a particular tumor. For example, we might observe that our new microarray analysis separated a group of patients with breast cancer into those that did well and those that did poorly in response to a particular therapy. However, if we then observed that all patients who did well had tumors of less than 1 cm in diameter and all patients that did poorly had tumors that were greater than 3 cm in diameter, our research finding would not be adding very much new information with regard to diagnosis and prognosis of breast cancer since it is already known that patients with tumors greater than 3 cm in diameter do worse than those with tumors that are one cm in diameter. *However, even in the unlikely event that microarray information did not prove to be of clinical use in the diagnosis or prognosis of breast cancer, it could still have* ***important intellectual property value****.* This is because the microarray analysis is detecting the cellular instructions that actually CAUSE the cancer cells from the larger tumors to behave worse. Thus, if a pharmaceutical company knew WHICH of the cells 10,000 instructions were responsible for the cancer cell's 'bad' behavior, it might be possible for the company to develop a drug that would interact with the molecules making up the cell's instructions and thereby disrupt the ability of the cancer cell to grow.

Database Design

Given the need to analyze thousands of genes and their cellular instructions, as well as the need to link these data to a patient's standard clinical and laboratory information, it is clear that some vehicle is required into which all of these data can be put and organized. This is a critical task and requires the development of new database standards and models. A summary of the work being undertaken in this area by the

scientists involved in this CTRF funded grant is given in the section "Focus Group Update" below.

Data Analysis

In the typical research study, a scientist assembles two populations of patients. One population has the condition for which the scientist is interested (the test group) and the other population does not (the control group). In this CTRF funded grant, the two major populations to be studied are patients who have cancer but who also have a good outcome with their disease and a second population of patients who have the same type of cancer but who have a bad outcome from their disease after a similar course of therapy. The typical approach that the scientist takes in investigating this type of question is to list the different factors that MIGHT be important in determining the different outcomes for the two populations and then perform a type of statistical analysis called *regression analysis* to see which, if any, of the factors that the scientists propose might be important in determining outcome actually ARE important. For example, in cases of breast cancer, a scientist might propose that younger patients and patients with smaller tumors, tumors that have not metastasized to lymph nodes, and tumors that express estrogen receptors might do better than older patients and patients with larger tumors, tumors that have metastasized to lymph nodes and tumors that don't express estrogen receptors.

In the above example, the scientist is seeking to evaluate the importance of 4 factors or variables. These are: a patient's age, a tumor's size, degree of spread (metastasis to lymph nodes) and degree of differentiation (presence or absence of estrogen receptors). *In general, the greater the number of variables that the scientist wants to investigate, the larger must be the size of the two populations of patients.* This fact points to a major challenge in the use of microarray data for evaluating cancer patients. Each microarray study tests for the presence or absence of instructions from 5,000 to 40,000 different genes. The result is that if the scientist desires to know whether the instruction from any individual gene is important in determining whether a cancer behaves badly or not, he might need to test populations of 50,000 to 100,000 patients or greater. To carry out such a study would require a study of truly national proportion. *Fortunately, however, scientists in this CTRF funded grant and elsewhere are developing computational techniques that permit clustering or combining of microarray data which reduces the size of the populations of patients required to draw meaningful conclusions.* The development and evaluation of these computational techniques is one of the major goals of this research grant. The calculation of the necessary sample size that will permit us to answer these questions has not been previously done. The ability to calculate samples sizes would provide much needed information to the newly developing field of using microarray data linked to clinical information to address relevant clinical issues. An update of the efforts by the members of this CTRF grant are presented in the section Focus Group Updates below.

Other examples of the types of important clinical questions that could be answered based on this project include practical issues related to who should be treated

aggressively and identifying populations at highest risk. For example: we know that only about 10% of men with prostate cancer will develop aggressive disease that causes their death. However, we are currently unable to identify who these 10% are, and thus everyone is treated. If we could use the information gained from the combined gene expression, clinical, and risk factor information to identify that population, we could reduce the number of men undergoing such therapies as radical prostatectomy. Another example would be to use these data to try and identify patients with leukemia whom are unlikely to respond to traditional chemotherapy. If we could successfully identify those persons, we could prevent their undergoing chemotherapy and its possible side effects when they are not likely to have a response.

Quality Control and Quality Assurance

The preceding sections have described challenges arising from the logistical complexity of specimen and data acquisition and the analytical complexity of data analysis. At least one more infrastructure issue remains which arises from the fact that tissue will be collected from two geographically separate sites (VCU and Inova) and analyses will be performed on at least 4 separate platforms in 4 different laboratories (MDx, NARF, & C3B at VCU and at GMU). In order to compare test results across the different microarray platforms, standard testing materials that can be analyzed on all of the microarray platforms have to be established. In addition, tissue acquisition procedures must be standardized between VCU and Inova in order to ensure comparability of separately acquired samples. The activities by the members of this CTRF grant are presented in the section Focus Group Updates below.

Progress During the Period 7/1/01 - 12/31/01

A. Project Coordination and Management

This CTRF funded Cancer Genomics project presents a tremendous opportunity to create a win-win situation for two of Virginia's leading universities and for the citizens of the Commonwealth of Virginia in the area of cancer research. It combines the computational bioinformatics expertise at George Mason University with the experience and expertise in medical sciences at Virginia Commonwealth University. The inclusion of the Inova Health Care Organization in Northern Virginia adds additional important clinical expertise to this partnership and expands the clinical base of potential specimens and patients available to the study. As noted in the section *Data Analysis s* above, the size of the population of patients that are studied is very important to the overall success of the project. *This partnership is expected to enable the partners to create new intellectual property and to facilitate competition for extramural funding in this rapidly growing field.*

With this opportunity, however, comes the challenge of creating a collaborative enterprise in which members located at the three geographically separate sites actually work productively together. The major responsibility for leadership and oversight in this area has been delegated by Dr. Torr to Dr. Carleton T. Garrett, MD, PhD, Professor of Pathology at VCU who currently serves as Interim Program Director for the CTRF Cancer Genomics project. Dr. Garrett is also the Project PI for the Tissue Acquisition Service at VCU and medical director of the Molecular Diagnostics laboratory at VCU. The Molecular Diagnostics laboratory is one of the three laboratories at VCU which will be responsible for analysis of patient residual specimens.

The initial goals of the Office of the Program Director are to: 1) establish monthly and annual meetings 2) initiate funding; 3) establish working groups to address task focused issues; and, 4) assemble required reports to CTRF.

Establish monthly and annual meetings:

The kickoff meeting of the leadership team for the grant was held August 23, 2001 at VCU. The need to establish focus groups that addressed each of the major areas of emphasis was discussed. The leadership team agreed to hold meetings approximately once a month alternating between VCU and either Inova or GMU. Subsequently, the leadership team has held meetings on October 1, 2000, at GMU and on November 19, 2001, simultaneously at both VCU and Inova via teleconference. **This latter format was well received and clearly permits the widest opportunity for attendance at leadership team meetings, communication between the geographically dispersed partners, and efficient use of professional time.** The next leadership team meeting will be held in January 2001.

In addition to the leadership team meetings, the members of the Database Design, Data Analysis and QAQC focus groups met at GMU on September 28, 2001 to address common issues.

Initiate funding:

Due to some final negotiations between VCU and the Virginia DPB over rebudgeting issues, the appropriation funding the first year of the grant was not received at VCU until the end of July. In August 2000 the VCU office of Grants and Contracts identified the funds and began the process of establishment of ledger 5 accounts to hold CTRF funds and linked ledger 4 accounts to monitor institutional matching funds. In addition, two subcontracts were created to fund the projects at GMU and Inova.

Establish working groups to address task focused issues:

During the months of September and October 2001, the leadership team established focus groups to address each of the critical tasks facing the grant in each of the areas of: 1) tissue banking; 2) clinical and laboratory data; 3) database design; 4) data

analysis; and, 5) quality assurance and quality control (QAQC). Table 1 outlines the groups and their leadership.

Table 1: Focus Group Leaders			
Focus Group	GMU	VCU	Inova
TissueBank	Geraldine Grant	Suhail Nasim	Barrie Cook
Clinical and Laboratory Data	----	Lynne Penberthy Suhail Nasim	James Cooper Barrie Cook James Burgess
Database Design	Curtis Jamison	Lynne Penberthy Greg Miller	Mike Sheriden
Data Analysis	Vikas Chandhoke	Greg Buck	---
QA/QC	Alan Christensen	Andrea Ferreira-Gonzalez	---

Assemble required reports to CTRF:

This document represents the efforts of the combined leadership of the CTRF funded grant PIs and the focus group leadership.

B. Listservs and Websites

In order to facilitate communication among members and focus groups, seven listservs have been created which are shown in Table 2. An effort has been made where there is overlapping interest of the members of two focus groups (ex. database design and data analysis) to at least include the leadership members of each focus group on the other corresponding focus group's subscription list.

Table 2. CTRF Cancer Genomics Project Listservs	
Listserv Name *(listname)*	Function
CG-TISBK	Tissue Bank
CG-CLNDT	Clinical and Pathology Data
CG-DBDSN	Database Design
CG-ANLDT	Analyze Data (Data Analysis)
CG-QAQC	QAQC
CG-LDRPI	Focus Group Leaders and PIs
CG-MEMBS	All Members

Members send messages to the appropriate listserv by addressing the message to *listname*@VENUS.VCU.EDU. Members may unsubscribe by submitting a message with the words SIGNOFF *listname* to:LISTSERV@VENUS.VCU.EDU. Addition to a

listserv requires the recommendation by a project PI who submits the individuals name and E-mail address to the Program Director (C.Garrett).

The CTRF grant expects to have its own website in operation by May 2002. The latter will provide cancer Genomics updated information about the project and provide links for members to obtain data generated by the project as well as links to member's home websites.

C. Milestones

In Table 3 are listed the important milestones that the CTRF grant leadership intend to accomplish during the first year of the grant. The initial version of the list was adopted by the leadership team at its October 1, 2001, meeting. Since then it has been revised to reflect the growing experience of project personnel. As is reported in section D. below, *excellent progress has been made in respect to all of the objectives*.

Table 3: CTRF Ca Genomics Project Research Plan Outline and Proposed Milestones - Year 1		
	Milestone	**Date**
Tissue Bank		
1	Submit to IRB	**10/1/2001 to 2/1/2002**
2	Hire Staff	**10/1/2001 to 3/1/2002**
3	Construct a protocol for tissue acquisition based on external experience and internal expertise	**10/1/2001 to 2/1/2002**
4	Determine tissue bank information that is required for resource tracking and management	**10/1/2001 to 2/1/2002**
5	Tissue Collection	**3/1/2002 to end of project**
Clinical and Pathology Laboratory Data		
1	Determine the most appropriate control tissue for brain, breast, ovarian, and hematopoetic cancers. Can they can be expected to be obtained from the specimen from which the tumor is obtained?	**10/1/2001 to 2/1/2002**
2	Determine most efficient way to effectively accomplish consenting process.	**10/1/2001 to 2/1/2002**
3	Identify clinical data that is needed for analysis. Determine most appropriate (user friendly) mechanism(s) for obtaining clinical data (?nurse coordinator, ?other).	**10/1/2001 to 2/1/2002**
4	Identify pathology laboratory data that is needed for analysis. Determine most appropriate (user friendly) mechanism(s) for obtaining pathology data (?use residents/fellows to manually encode data, ?automate encoding through use of 'indexing of text terms' in SQL database, ?other).	**10/1/2001 to 2/1/2002**

Database Design		
Phase I		
1	Create or identify existing databases into which expression microarray data can be stored in electronic format in real time at this juncture.	**10/1/2001 to 12/31/2001**
2	Determine appropriate field types for storing clinical, laboratory and tissue bank data needed in Phase I databases	**11/15/2001 to 2/1/2002**
3	Create or identify existing databases into which clinical, laboratory, tissue bank information, and expression microarray can be stored in electronic format in real time at this juncture.	**11/15/2001 to 2/1/2002**
4	Create links between separate databases containing clinical, laboratory, and tissue bank data.	**11/15/2001 to 6/30/2002**
Phase II		
1a	Develop specifications for design of a statistical and scientific database (CTRF SSDB) which will permit analysis of micro array data with clinical (including clinical laboratory) data.	**10/1/2001 to 4/1/2002**
2a	Create CTRF SSDB with links to Phase I databases	**1/1/2002 to 5/1/2002**
3a	Populate CTRF SSDB	**5/1/2002 to 6/15/2002**
Data Analysis		
1	Survey and enumerate data mining approaches and existing software tools to perform analysis of micro array data with clinical (including clinical laboratory) data.	**10/1/2001 to 12/1/2001**
2	Select software covering range of data mining approaches judged to be most statistically reliable and that will be compatible with anticipated CTRF SSDB structure.	**12/1/2001 to 5/31/2002**
3	Develop questions to begin to test the relevance of expression data to stratify patients according to: 1) observed outcome; and, 2) predicted outcome based on clinical and pathologic stage and tumor grade (Note: there may not be sufficient time in the project to complete #1. This will require creation of an ad hoc subcommittee comprised of physicians and pathologists as well as data analysis personnel.)	**1/1/2002 to 5/31/2002**
4	Develop questions to begin to test the relevance of expression data to explain disease mechanism.	**1/1/2002 to 5/31/2002**
5	Initiate analyses using most applicable software tools to answer above questions.	**6/1/2002 to end of project**

Quality Assurance and Gene Expression		
1	Initiate GeneChip and spotted array development VCU; develop human cDNA spotted array GMU.	**10/1/2001 to 2/1/2002**
2	Develop criteria for intralaboratory and interlaboratory assessment of quality using standard materials and shared specimens.	**10/1/2001 to 3/31/2002**
3	Develop at least 1 standard material for intra- and inter-laboratory testing	**10/1/2001 to 3/31/2002**
4	Initiate interlaboratory testing comparisons program using one or more standard materials	**4/1/2002 to 6/30/2002**
5	Develop additional testing materials as required	**7/1/2002 to 12/31/2002**
6	Initiate interlaboratory testing comparisons program using one or more shared specimens	**7/1/2002 to 10/1/2002**
7	Perform initial evaluation of quality measurements on available data	**7/1/2002 to 10/1/2002**

D. Focus Group Updates

Tissue Bank (Suhail Nasim, MD, PhD and Kristen Atkins, MD)

Members of the Tissue Bank Focus Group met several times to review the procedures necessary to obtain the residual tissue samples required for support of the CTRF Cancer Genomics Project. In collaboration with Dr. Lynne Penberthy of the Massey Cancer Center, a protocol has been developed for a tissue acquisition system (TAS). Patients will be consented by either a coordinator at the Massey Cancer Center or by one of the TAS staff. After informed consent has been obtained, a member of the TAS staff will arrange to be at the location where the tissue sample will be removed (operating room or location where a bone marrow aspiration will be performed). At the time that the tissue sample has been removed from the patient, the TAS staff member will immediately take the tissue specimen to the surgical pathologist where it will be immediately grossed and a sample of tumor and normal tissue removed and frozen in liquid nitrogen.

The TAS staff will then assign the frozen sample a tissue bank number and place the sample in a predetermined location in a -80^{o}C freezer with liquid nitrogen back-up. The TAS staff will enter the number and location into the appropriate fields of a database that is used to monitor specimen availability. Specimens will be reviewed for adequacy by a surgical pathologist aided by the TAS staff and the specimen dissected as required to ensure that the main part of the tumor specimen consists of tumor cells.

Clinical and Pathology Laboratory Data (Lynne Penberthy MD)

Members of the Tissue Bank Focus Group were queried regarding what parameters they considered to be essential in establishing the diagnosis of cancer of the breast, ovary, brain, and hematopoetic systems. A compilation of these findings is

presented in Table 4 below. These parameters will be specifically coded into the CTRF Phase I database being developed (see next section *Database Design* for further details). The CTRF Phase I database will also have at its disposal, hundreds of clinical and laboratory data items for each case which are currently part of the MCC database and Pathology Research Database. In addition, in the process of consenting patients for obtaining their residual samples, primary clinical data will be obtained through questionnaires dealing with information on health history, behavioral risk factor status, family history for cancer and certain other diseases, baseline functional status and quality of life measures.

Table 4 - Clinical and Histopathological Parameters of Diagnostic and Prognostic Importance

	Breast Cancer	Ovarian Cancer	Brain Cancer	Hematopoetic Cancer	
				Leukemia	Lymphoma
Clinical:	age post-chemo or radiotherapy	age uni- vs bilateral	Age Karnovsky functional score	Age sex WBC Hct plt. Count diff hx chemo-/radiation therapy hx myelodysplasia HIV status	age sex clinical stage +/- 'B' symptoms HIV status post-transplant pt. hx autoimmune disease hx lymphomogenic drugs (methotrexate, dilantin, etc.)
Gross:	size	Solid vs. cystic Size			site of biopsy size of node/tumor marrow status (pos. or neg.)
Histologic:	histologic type histologic grade inflammatory ca in-situ ca presence/extent hormone receptor Her-2/neu lymphovascular invasion margin status lymph node status	Category: epithelial germ cell sex cord stromal Subclassification serous papillary mucinous endometrioid pure or mixed grade invasion lymph node status	Necrosis Increased Cellularity Polymorphic Nuclei Endothelial Proliferation increased mitotic figures vascular proliferation satellitosis WHO Classification MIB 1 Mitotic Index GFAP staining	histologic dx (W.H.O. classification) immunophenotype cytogenetics molecular	histologic dx (W.H.O. classification) immunophenotype cytogenetics molecular

Database Design (Lynne Penberthy, MD, Greg Miller PhD, Brian Cassel, PhD)

The issue of database design has been approached from two perspectives that are described in Table 3 as 'Phase I' and 'Phase II' databases. *This is necessitated by the fact that at this juncture NO database exists which has been designed to hold both a large amounts of clinical data, which contains significant amounts of categorical information, AND the large amounts of data associated with microarray studies which are principally quantitative in nature.*

Focus group members primarily from VCU have been concentrating on constructing a database based on conventional software platforms, primarily SQL 2000, which can bring together the multiplicity of clinical data, including cancer center, clinical laboratory and histopathology data, together with microarray data produced by the Affymetrix GeneChip instrumentation. The basic components for this database are listed in Table 5 below. This approach will rely on commercially available software tools some of which are referred to by Dr. Chandhoke in his summary of data analysis efforts below, to carry out data analysis.

The second effort involves Dr. Buck's laboratory here at VCU and Dr. Chandhoke's faculty at GMU. They are focusing on adapting an open database source called GeneX http://genex.ncgr.org/ to meet the database needs of this project. This effort is described further in Dr. Chandhoke's discussion of data analysis and in Dr. Buck's update of his laboratory's activities for this grant.

Presented here below is a summary of efforts to create the Phase I database, which will combine clinical and microarray data.

Table 5 Data Sources

Data Source	Source	Key Components	Utility
Massey Cancer Center Information System (MCCIS)	Decision Support *Inpatient Claims* *Outpatient Claims*	All Diagnoses & Procedures (ICD-9 and CPT codes) Utilization, Costs, Charges & Reimbursement	Information on Comorbidity, treatment to permit adjustment for differences in severity Outcomes related to cost
	Cancer Registry	Clinical Variables for cancer (site, stag, histology, morphology, treatment, dates of diagnosis, treatment)	Permits controls for confounding on severity measures for the cancer
	Pharmacy	Chemotherapy	Specific characterization of chemotherapy
	Research Database	Information on clinical trial enrollment status	
Laboratory Information System (LIS)	Surgical Pathology	Details on pathology of cancer, Snomed codes	Permits adjustment for differences in specifics of cancer (more detailed than cancer registry)
	Clinical Pathology	Tumor Markers	Independent variables assessed for predicting outcomes
		Laboratory tests (I.e. BUN, glucose)	Provide independent measures of comorbidity and severity of illness
Gene Expression Data	Microarray results from Affymetrix gene chips	Patterns of protein expression for approximately ?11,000 genes on the chip	Predicting variation in outcomes associated with variation in genetic expression across cancers and between cancer patients and and control tissue
Health Interview Data	Interview data collected from cases and controls by the study coordinator	MOS SF 36 Measures of QOL Functional Status Health History *Family History* *Current and prior health conditions* *Health Risk Behavior Patterns*	Serve as both outcomes and predictor variables in assessing differences in outcomes. Independent predictor variables to adjust for differences among cases and between cases and controls.

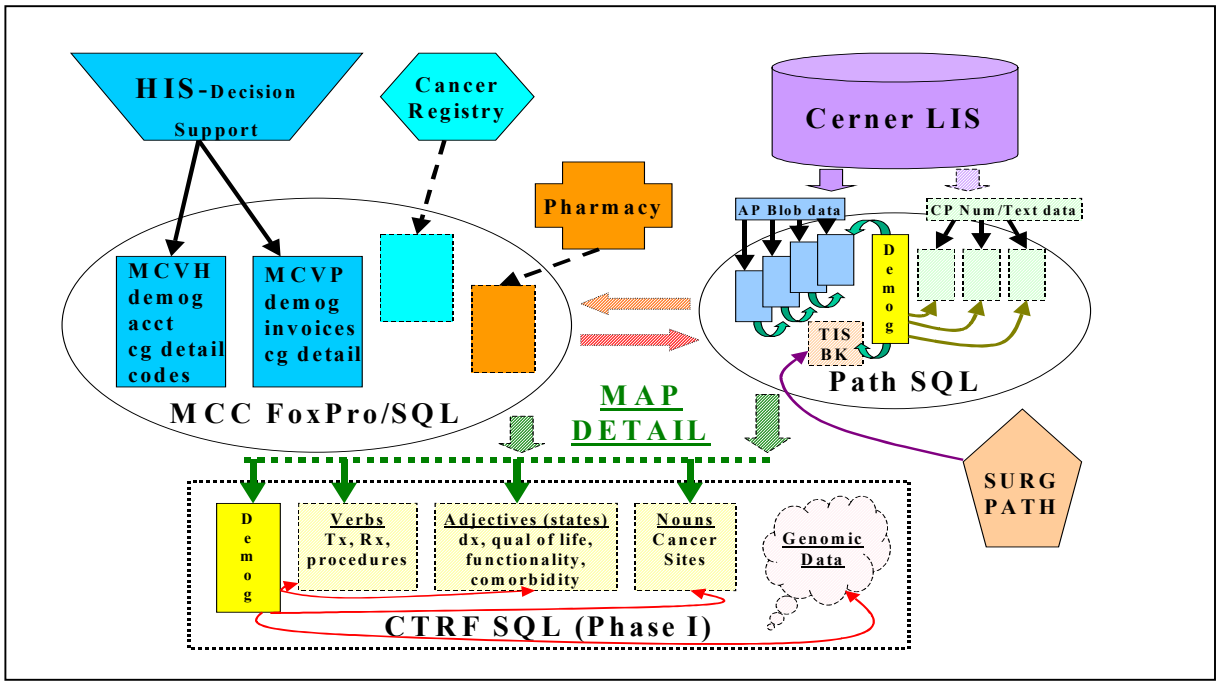


Figure 1

Phase I - Database Resources - Current Status

Massey Cancer Center (MCC)

The major current sources of data are the MCVH and MCVP information systems, the Cerner LIS and the Pharmacy database. Further, there are in the MCC additional data sources including a Clinical Trials Research Database, Bone Marrow Transplant Database, and Cancer Registry. MCC acquires charge detail that includes varying amounts of diagnostic and therapeutic information from MCVH and MCVP though efforts of HIS Decision Support. This information is parsed and imported into one or more of seven tables in the Fox Pro database (see Figure 1 above).

VCU Pathology

The major existing resource for Pathology is the Cerner LIS which went live in Oct of 1999. Recently, VCU Pathology Information Systems (PATHIS) has begun to build a research database (PATHRDB) using the MS SQL2000 platform (see Figure 1 above). Current efforts have focused on developing a mechanism for extracting AP Surgical Pathology reports and associated SNOMED codes from Cerner. A major limitation of data at present is that they are mainly in free text format which leads to major problems in using the existing data for data analysis. PATHIS has established processes to import selective clinical laboratory data. Also, current plans include creation of tables within the PATHRDB for accumulation and Tissue Bank data on stored specimens (Figure 1 above).

Phase I - Developing Specifications and Plan

Review of the schema in Figure 1 helps to identify specific tasks that must be accomplished in order to build a comprehensive Phase I database that can be linked with genomic data. These include:

Creation of a NAME TABLE in each database that will enable linkage of event and account information to one personal identifier. This table will also be essential for anonymizing data which must take place to allow its use in research studies. The specification must include:

a. Definition of mapping algorithms for populating new GRAMMATICAL table structures in CTRF Phase I database (Verbs, Adjectives, Nouns).

b. Defining a process for encoding AP 'text' data.

c. Identifying which additional clinical laboratory data should be stored in the PATHRDB.

d. Determine how to link the Affymetrix GeneChip microarray data, which is stored in an MS SQL format, with the clinical data in a manner permitting regression analysis.

Data Analysis (Vikas Chandhoke, PhD and Greg Buck, PhD)

Research in molecular biology has produced an avalanche of genetic information, including the DNA sequences of many complete genomes and whole genome-level analysis of gene expression in normal cells and cancer cells. In many cases this avalanche of information has overwhelmed the initial attempts at analysis. However, retrospective attempts to reanalyze multiple data sets with respect to a single question ("data mining") are proving to be productive. Data mining is particularly likely to be helpful in clinical studies using DNA microarrays ("gene chips") to analyze gene expression in human cancer cells. Such studies involve massive data sets with large numbers of genes (~40,000) and patients (hundreds to thousands). We expect information such as patient outcomes, genetic susceptibilities, gene functions, disease subtypes, and new drugs to become available in the coming months and years. With the addition of this new data, retrospective data mining of our gene expression databanks is likely to yield critical new insights into human health and disease.

High-density microarray technology is allowing the rapid expansion of gene expression databases. With the influx in samples examined and in the expanding number of conditions monitored, this data accumulation is growing exponentially. It is clear that computational means are absolutely required in order to separate the genome noise from the relevant gene expression signals. This gene expression data mining endeavor calls for the merging of human genomic “common sense” with massive unbiased computational power. Several bioinformatics tools are available for an initial screen of the data. More

will be developed in order to customize the data mining to the needs of the Cancer Genomics program.

It is cost-effective to import public domain software, especially from sites and organizations with solid reputations in the area of providing such toolsets. One such organization is the National Institutes of Health, in particular, NCBI and NHGRI. The NCBI site not only has GenBank, the repository for DNA sequence, gene, genome and protein databases, but also several sequence analyses tools. An example is the "Protein Information Resource" (http://pir.georgetown.edu/pirwww/search/pirnref.shtml),

Specific examples of software for microarray gene expression data mining are "GeneX," NCGR's collaborative internet database and toolset for gene expression data http://genex.ncgr.org/; "Gene Expression Omnibus (GEO)," NCBI's public gene expression repository in development; and "ArrayDB," NHGRI's gene expression software. Other public domain tools are Albert Einstein College of Medicine's "Microarray Analysis Tool (MAT)," B. Dyysvik's "J-Express: Java program for analyzing microarray data," and EMBL's "ArrayExpress." In addition, the Whitehead Institute's "Molecular Pattern Recognition" software, which focuses on the analyses and interpretation of large-scale microarray gene expression data sets, may be very useful.

There are many commercial vendors offering similar gene expression data mining software. GeneDataAG's (Switzerland) "GeneData Workbench and GeneData Expressionist," Lion Bioscience AG's (Germany) arraySCOUT," Rosetta Inpharmaceutical's (USA) "Resolver Expression Data Analysis System," Scanalytics's (USA) "Microarray Suite," and GCG's (USA) "SeqArray."

These tools offer methodology for analyzing microarray data signals, such as "pattern recognition," extracting relevant signal from noise and archiving data. Once archived, the data is re-analyzed with respect to gene expression as well as to gene identity and function. Additionally, the commercial packages allow the context of the genes to be displayed: gene cluster analysis, gene order (genome co-linearity) and metabolic pathway mapping. Once characterized, the genes are linked to currently available public and private (commercial) databases.

These toolsets are available and being evaluated for strengths and weaknesses, and their applications to the Cancer Genomics data sets are being explored. With the bioinformatics researchers in residence (Fryxell, Jamison, Seto and Kinser) at GMU, a custom relevant package incorporating some or all of the above software may be developed for this unique data set. One data-clustering algorithm developed at GMU provides a general and optimal solution to the number of higher-order clusters in a complex data set (such as a cluster of patients that are particularly susceptible to lung cancer). This algorithm was specifically designed to be scalable to massive data sets. Additional algorithms are under development to automatically extract the functions of these genes, provide a 3-dimensional visualization of the clusters, and determine the statistical uncertainty in cluster boundaries. These efforts are computationally expensive, and require multiple parallel processors and very large databases. One of the

researchers at GMU has already developed two computational tools, GeneOrder2.0 and OrthologLocator, to do genome data mining analyses. Similar tools have been developed and are being developed independently by several commercial vendors, such as those listed above.

Quality Assurance and Gene Expression (Alan Christensen, PhD and Andrea Ferreira-Gonzales, PhD)

The QAQC Focus Group has as its two main objectives:

1) Develop criteria for intralaboratory and interlaboratory assessment of quality using standard materials and shared specimens.

   With regard to this objective, progress will most likely be achieved once some data is generated from objective #2.

2) Develop at least one standard material for intra- and interlaboratory testing.

   a) Standard materials are used in the *Affymetrix GeneChip arrays* that allow to quantitate gene expression and can also be used as spike controls inter-samples:

      i) 20X GeneChip Eukaryotic Hybridization Control Kit: Each commercially available Affymetrix GeneChip probe array contains probe sets for several prokaryotic genes, which serve as hybridization controls when mixed with eukaryotic cRNA samples. The genes are: *BioB, BioC, BioD* (from the biotin synthesis pathway from *E. coli*), *Cre* (recombinase gene from P1 bacteriophage), *dap, thr, trp, phe*, and *lys* (from *B. subtilis*). These controls are provided as antisense labeled cRNA in known concentration, therefore they are used as hybridization standards for estimating the abundance of RNA transcripts in the sample. In addition these controls can be used to normalize expression levels between experiments during software analyses.

   b) Three possible standard materials useful for *spotted arrays* are:

      i) Lambda 1 A polyA RNA (standard RNA). Lambda l A polyA RNA will be included in probe labeling. The l A cDNA will be printed on the cDNA arrays (dilution series). This will permit comparison between slides and allow independent normalization.

      ii) Stratagene Universal Human Reference RNA (a commercially available human reference RNA). Stratagene Universal Human Reference RNA provides a consistent internal control as well as allows data comparison between different microarray experiments. Prepared from 10 different cell lines to optimize gene coverage.

iii) End-labeled primer. Primer will be labeled with a third color fluorescent dye. Primer will hybridize to sequences present in all of the amplified cDNA fragments. This will allow determination of concentration of DNA in arrayed PCR products and for comparison between slides.

E. PI Lab Updates

VCU Molecular Diagnostics Laboratory [MDxL] (Carleton T. Garrett, MD, PhD - Project PI)

In February of 2001, the VCU Molecular Diagnostics Laboratory (MDxL) into 6,400 sq ft of new laboratory space that had been built at cost of 1.5 million dollars from VCU Health System funds. The space includes 4,400 sq ft of laboratory bench space and 2,000 sq ft of office and storage space. The laboratory space was carefully designed with differential pressure in specimen handling and analysis areas to eliminate any contamination that can occur during the testing process. The Laboratory is CLIA certified and performs clinical as well as research molecular testing with state of the art equipment including automated DNA sequencing and real time PCR instrumentation.

The major objectives for MDxL for the first year of this grant are: (1) establish a tissue acquisition service (TAS); (2) acquire and validate testing using the Affymetrix GeneChip instrument for detection of changes in gene expression and somatic genetic changes in human cancer; (3) work in collaboration with members of the Massey Cancer Center to develop a database and protocols for undertaking data analysis that will allow determination of the use of gene expression and somatic mutational changes in human tumors as stratifiers for determining diagnosis and prognosis of these cancers.

Progress with regard to the objectives #1 and #3 have been described above in Section D. Focus Group Updates *Tissue Bank* and *Database Design*. In this section the progress made with objective #2 in validating and testing the Affymetrix GeneChip instrumentation for detection of changes in gene expression and somatic genetic changes in human cancer will be summarized.

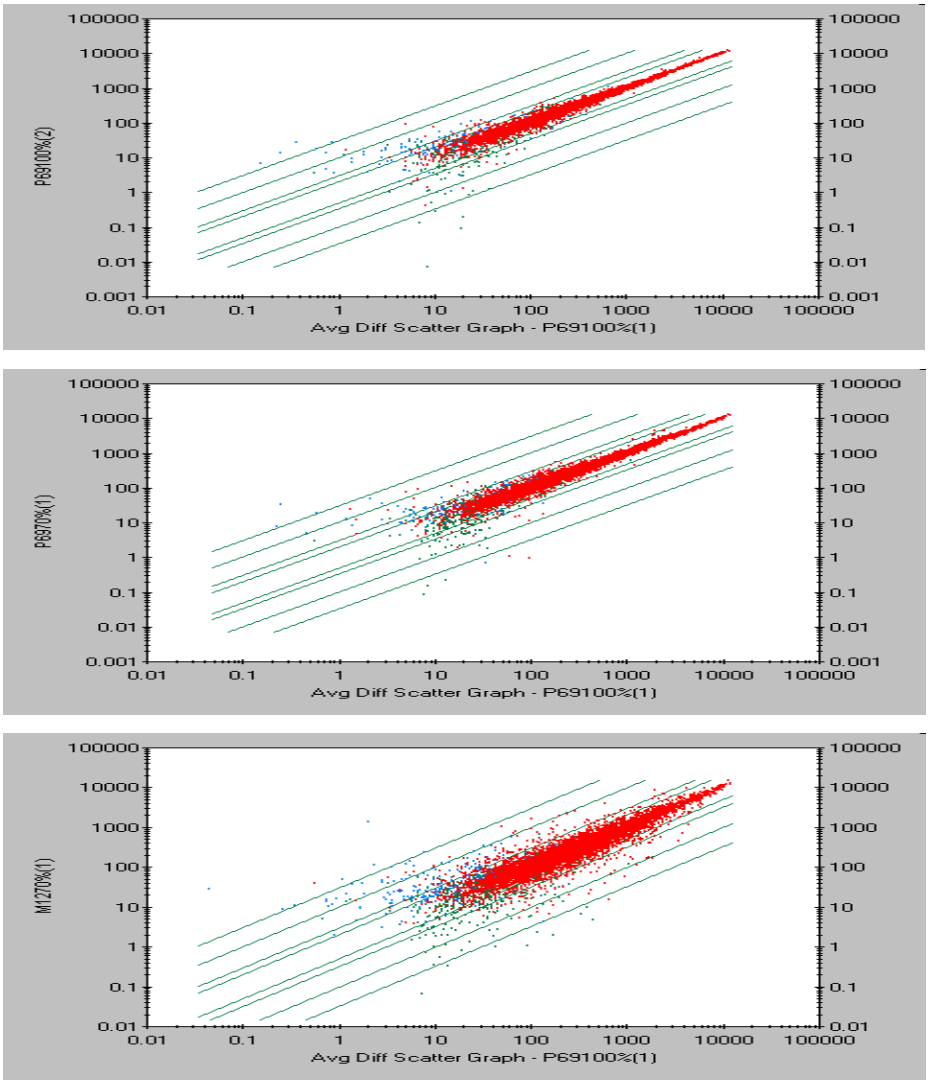
Work toward achieving objective #2 is underway. To begin the process of validating the Affymetrix GeneChip instrument for the studies proposed in this grant, it was decided to initiate testing using an established model for the study of human prostate cancer that has been developed in the VCU Department of Pathology. In this model human prostate epithelial cells were immortalized through expression of the SV40 large T-antigen. Next, increasingly aggressive sublines of the original (parental) cell line were generated. The original cell line (P69) is immortal but does not form tumors when transplanted into nude mice. A derivative cell line is consistently tumorigenic and metastatic (M12) when injected with nude mice (6). In results described below, P69 cells are referred to as 'Benign' cells since they are non-tumorigenic and M12 cells as 'Cancer' cells since they are both tumorigenic and metastatic.

Two types of studies are underway. The first study looks for differences in *gene expression* (that is, differences in the cellular instructions that regulate cell behavior) thag are present in the benign (P69) cells and in the cancer (M12) cells. The second study looks for changes in the DNA of the cancer (M12) cells, which may be responsible for the alterations in its gene expression and thus its tumorigenic and metastatic behavior.

Changes in Gene Expression in Benign (P69) and Cancerous (M12) Human Prostate Epithelial Cell

The human prostate epithelial cell model investigated here is in many ways an ideal system with which to begin validation studies. The cancer (M12) cells are derived directly from the benign (P69) cells and the samples that are examined in the studies described below are either 100% pure benign (P69) cells or 100% pure cancer (M12) cells. Thus, any differences observed in the genes being expressed by the cells presumably related to differences in biological activity and not to the fact that the benign and cancerous cells arose from different organs or are from different individuals.

In the study described below, three sets of cells were examined. The 'control' cells consisted of benign (P69) cells, which had ceased their growth (100% confluency). The second sample consisted of benign (P69) cells which were still growing (70% confluent) and the third sample consisted of cancer (M12) cells which were also still growing (70% confluent). Each of the samples was run in duplicate. In the three panels of Figure 2 below, the level of genes expressed in each of the samples is compared with the sample of benign (P69) cells that had ceased growth. The top panel shows a comparison of one sample of non-growing benign (P69) cells was compared with the duplicate sample of non-growing benign (P69) cells. The distribution of expressed genes (cellular instructions) clusters very tightly along a 45° diagonal line indicating that there is very little difference detected in the two samples as would be would be expected. In the middle panel, the expressed genes of the non-growing benign (P69) cells are compared with those of the growing benign (P69_ cells and again the results indicate that there is little between the genes expressed in these two types of benign cells. The results of comparison of non-growing benign (P69) cells and cancerous (M12) cells are shown in the lower panel. Clearly, there are significant changes in the genes that are expressed in the cancer (M12) cells when compared with the non-growing benign (P69) cells. Moreover, as seen with reference to the middle panel, these changes can not be explained simply by the fact that the cancer cells are growing.



Benign (not growing) vs. Benign (not growing)

vs. Benign (growing)

vs. Cancer (growing)

Figure 2

In order to more closely determine the changes in gene expression present in the benign and cancer cells, the results were subjected to further analysis. The findings are summarized in Table 6 below. The analysis examined 10,099 genes. The Affymetrix GeneChip instrument scored each gene as either expressed (transcripts [instructions] from that gene were present in the sample) or not expressed (transcripts [instructions] from that gene were absent from the sample). If transcripts from a gene were present, the amount of transcript was also measured. As summarized in Table 6, of the 10,099 genes evaluated, 13 were found to be expressed two fold or greater when compared to the non-growing benign (P69) cells. Six of the 13 genes over expressed in the sample of growing benign (P69) cells matched genes found in the group of 198 over expressed genes of the cancer (M12) cells. A similar analysis was conducted for genes that showed 2 fold or greater *reduced expression* in growing benign (P69; number 5) and cancer (M12; number 318) cells as well as for genes *induced* in the growing benign (P69; number 14) and

cancer (M12; number 116) cells and for genes repressed in the growing benign (P69; number 46) and cancer (M12; number 125) cells.

In summary, the findings of this initial study are very encouraging. In this model system the Affymetrix system was able to identify consistent changes in duplicate samples of benign (P69) and cancer (M12) cells which made biological sense. That is, one would expect that cancer cells would demonstrate a greater degree of change in gene expression than growing benign cells. This was clearly seen in our study. Moreover, in each case, one or more of the genes that changed its level of expression in the sample of growing benign (P69) cells was also found amongst the group of genes that changed its level of expression in the cancer (M12) cells. The likelihood that this was due to some random event is very unlikely. First there is the enormous size of the population from which the samples were drawn. Second is the fact that in no case did one see matches between genes from growing benign and cancer cells where the directions of change were not the same. Further experiments and more extensive data analysis are planned

Table 6: Differences Observed in 10,099 Genes Between Benign Non-Growing Prostate Cells and Benign Growing Prostate Cells and Prostate Cancer Cells

Change Over Control	Increase ≥ 2 Fold		Decrease ≥ 2 Fold		Induced		Repressed	
Sample	Benign	Cancer	Benign	Cancer	Benign	Cancer	Benign	Cancer
Total	13	198	5	318	14	116	46	125
#Match(1)	6	--	1	--	7	--	19	--
Matching Genes	tyrosine kinase receptor (axl) Human Bax gamma mRNA alpha adducin mRNA cDNA IMAGE-195893 cDNA IMAGE-1009907 cDNA IMAGE-47383		ras interactor (RIN1) mRNA		FGF receptor (K-sam) desmocollin type 4 KIAA1038 protein Xsome 1 transcript KIAA0495 transmembrane protein potassium channel Kir3.2 ORF1 & ORF2 from Xsome X		osteoblast specific factor 2 (OSF-2os) Mel-18 protein KIAA0421 mRNA KIAA1106 mRNA transmembrane receptor UNC5C clone 24694 mRNA sequence zinc finger mRNA ZNF132 transcription factor SL1 KIAA0764 mRNA cDNA, IMAGE-1706635 retinoic acid-inducible endogenous retroviral DNA mannan-binding lectin-associated serine protease-2 cytokeratin 15 sushi-repeat mRNA (SRPUL) cartilage specific proteoglycan Na+-independent neutral and basic amino acid transporter ras interactor (RIN1) mRNA bone morphogenetic mRNA 2A (BMP-2A) mammaglobin B precursor	

Note: (1) The number of genes in the benign growing prostate cells showing the indicated change which matched genes showing this change in the prostate cancer cells.

Loss of Heterozygosity Detected by Single Nucleotide Polymorphism (SNP) Analysis in Human Prostate Epithelial Cell Lines of Differing Malignant Potential.

Human cancers arise by a combination of genetic changes such as discrete mutations and chromosomal alterations that lead to activation of genes responsible for cell growth and mobility (oncogenes) and inactivation of genes that suppress cell growth and mobility (tumor suppressor genes). Detection of inactivated tumor suppressor genes can be accomplished through detection of portions of a tumor cell's DNA in which one of the two pairs of chromosomes has been lost. Such regions in a tumor cell's DNA are said to show loss of heterozygosity (LOH) (7). Searching for LOH used to be an extremely arduous and labor intensive task but with the advent of microarrays and the availability of DNA chips that contain SNPs spread across the entire human genome, screening for LOH has become easier. To test this approach for detecting LOH in human cancer, we employed the prostate epithelial cell lines described above (6). To test the potential of SNP array hybridization for the detection of genome-wide LOH in human tumors, we measured allelic loss and retention patterns for the human prostate cell line M12. 1,494 SNP loci were interrogated by performing 24 multiplex PCR amplifications followed by hybridization to a high-density oligonucleotide array (HuSNP™).

In the benign (P69) cells, 306 informative SNPs were identified. When compared to the cancer (M12) cells, at least 5 regions with loss of heterozygosity (LOH) were identified. These were mainly located in chromosomes 3p12-p14, 3p21-p22, 11q23-q24, 11q25 and 19p13. These regions harbor several tumor suppressor genes that might be implicated in prostate cancer development and progression.

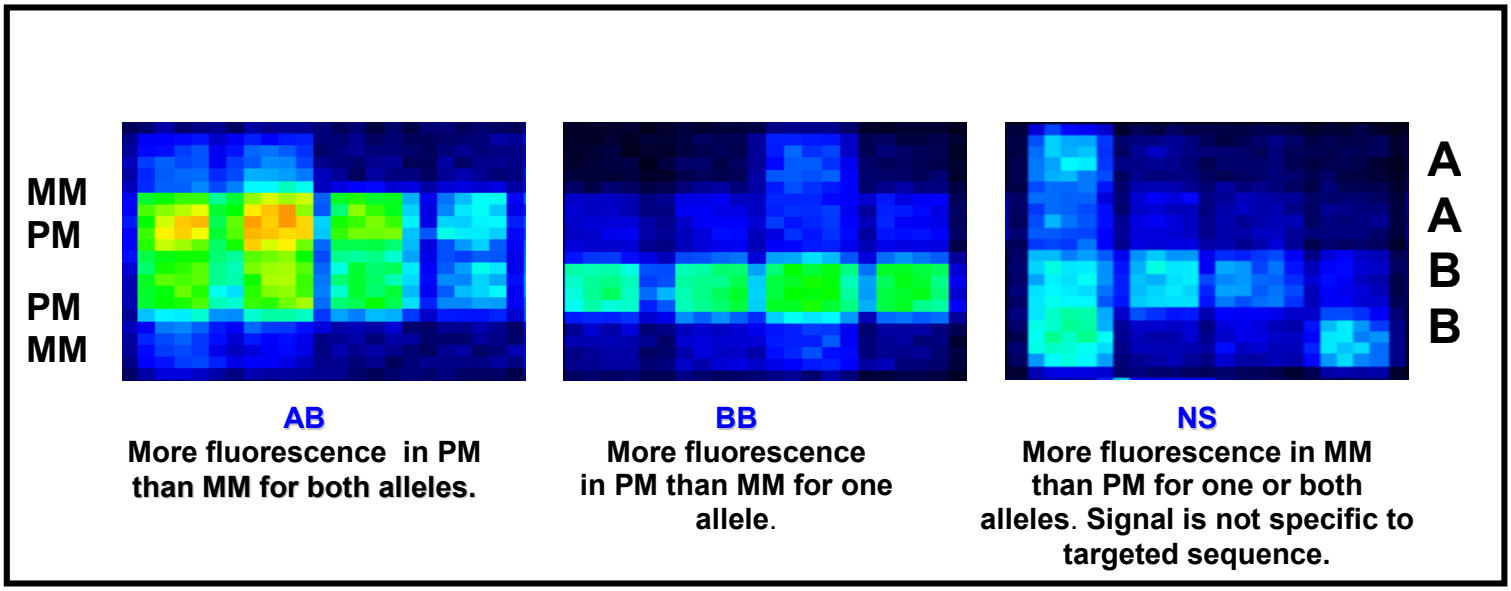


Figure 3: Enlargement of cells from an Affymetrix HuSNP™ array showing signals from heterozygous, homozygous and absent signals.

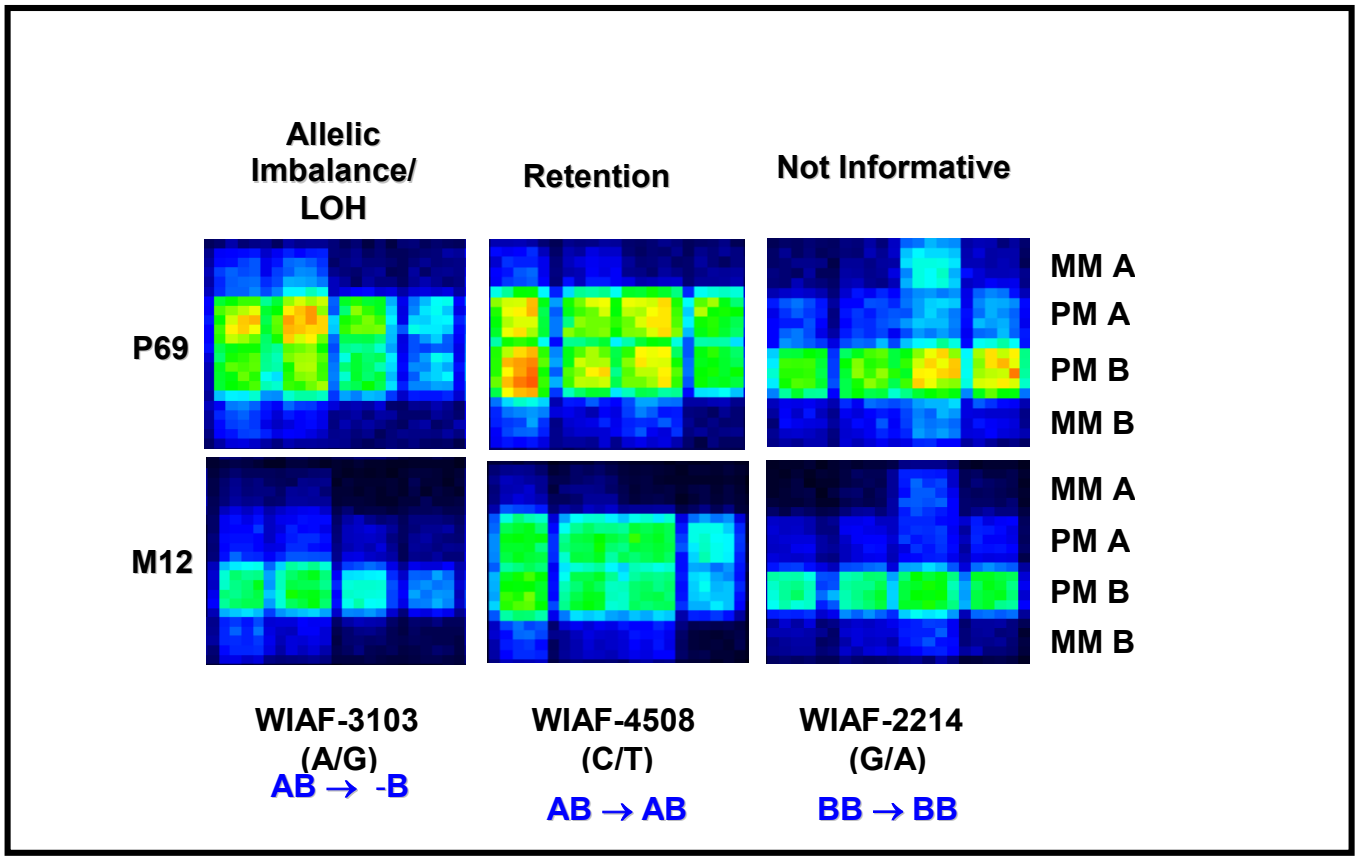


Figure 4: Enlargement of cells from an Affymetrix HuSNP™ array showing signals for heterozygous locus showing LOH, heterozygous locus showing preservation of heterozygosity and uninformative homozygous locus.

VCU Nucleic Acids Research Facility [NARF] (Greg Buck, PhD - Project PI)

In the first year of the project, we had planned to achieve the following milestones:

1. Complete the renovations of the MicroArraying Core (MAC) of the Nucleic Acids Research Facilities (NARF).
2. Acquire the Spotted Array and Affymetrix GeneChip instrumentation for the MAC.
3. Hire an Assistant Professor to operate the MAC.
4. Test and validate the instrumentation in the MAC.
5. Participate with Massey Cancer Center and Biostatistics to hire a bioinformatics faculty member to assist with the analysis and database generation in this project. Begin to develop software and databases to coordinate the development of the expression data to be obtained in this project.

Our progress on each of these points is described below.

1. The renovations of the MAC are now completed. Begun in August, 2001, the lab was repainted, refurnished, rewired and new Ethernet ports added. The renovation was completed in September, 2001. Funds for this renovation were provided from institutional resources but were not included as institutional matching for this project.
2. We acquired the instrumentation for the MAC as follows: 1. PCR instrumentation. Four DNA Thermocyclers, two 96-well and two 384-well ($25,000), were purchased from BioRad in July, 2001. 2. BioRobotics MicroGrid Total Array System II ($106,000) DNA spotter was purchased and installed. 3. Packard ScanArray 5000XL ($92,000) was purchased and installed. The Affymetrix GeneChip Array System ($212,000) was installed. $91,875 of the cost of this equipment was used as institutional matching funds for this project.
3. A search was initiated in Spring, 2000, for an assistant professor to operate the MAC. Dr. Darrell Mallonee, who has ten years of experience as a senior scientist was hired. One half of Dr. Mallonee's salary ($26,000 + 6500 fringe) is listed as institutional matching for this project.
4. Test and validate the instrumentation in the MAC. We have tested and validated the Affymetrix GeneChip system and are now operating this instrument regularly (see example of results of a study investigating effect of infection of mouse liver cells by T.cruzi below). Dr. Mallonee attended a course at BioRobotics, Inc., in Cambridge, England, to learn the operation of the BioRobotics Spotter, and has been intensively instructed on site in the operation of the Packard Scanner. We are awaiting an upgrade on the Spotter for full operational activity. Funds from this award are being expended to ensure the validation of these instruments.
5. A search has been begun for a bioinformatics faculty member who will be paid in part from this project and in part from Biostatistics and the Massey Cancer Center. Candidates are currently being reviewed, and we anticipate hiring a new faculty member in upcoming months. Funds from this project will be used to support a fraction of this new faculty member's salary.
6. We have begun to address the question of database development and data mining. First, we are planning to implement a version of the GeneX Database that was developed at NCGR for the *Arabidopsis* genome project. This is an open source, freeware product that must be modified for our use. We are installing this database on LINUX based MAC computers, for eventual movement to the Bioinformatics Computational Core Laboratory Suite (BCCL). Our operations systems analysis is installing the program, and in collaboration with colleagues at UVA, GMU and VaTech, we are modifying the software to accept Affymetrix uploads. We anticipate that this work will be completed by January or February, 2002. This work is being supported by other institutional funds, not from this project.

Effect of T.cruzi infection on gene expression in mouse heart cells using the Affymetrix GeneChip instrument and MicroArray Suite software version 4.0

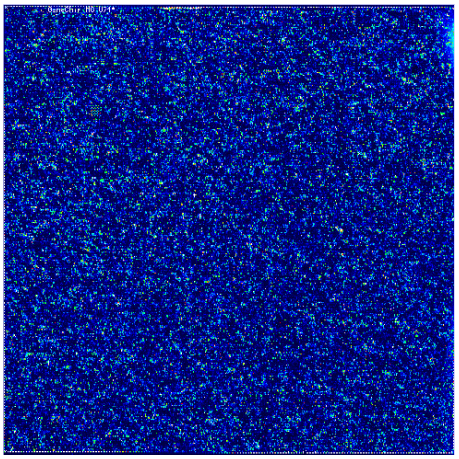


Figure 5: Example GeneChip Result (DNA Chip)

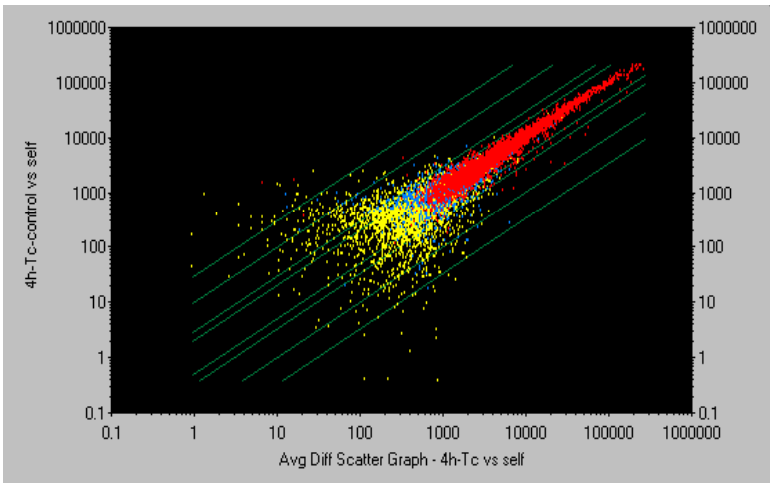


Figure 6:
Analysis: Total genes probed: ~12,500
Expressed: ~4,500
Not expressed: ~8,000

Table 7 - Results of Infection of Mouse Heart Cells with T. cruzi

Induced Genes	*Repressed Genes*
Interleukin 6	Thrombin receptor
Mac inflamm protein 2	Raf-related oncogene
GM-CSF	RNAse P protein p30
Interleukin 1 alpha	Pre B-cell transcription factor 1
B cell leukemia/lymphoma 3	N-acetylglucosamine kinase
TK	Microtubule associated protein 6
Cell signaling repressor 3 -SOCS-3	Transcriptional regulator mEnx-1
Selectin	Ras-like GTP binding protein (Rem)
LIM domain, protein kinase	Neuronal tyrosine/threonine
Fas-binding protein	phosphatase 1
	S-AKAP84
TGN38A	Groucho related gene 1 (grg1)
Apoptosis inhibitor 2	
Cytokine A3	
Interleukin 1 beta	
Platelet derived Growth factor	

VCU Center for Bioelectronics, Biosensors and Biochips [C3B] (Anthony Guiseppi-Elie, Sc.D. - Project PI)

The C3B is concerned with the development and application of second generation genomic devices and systems and with the application of genomic data in molecular diagnostic and pharmacogenomic biochips. The use of spotted oligo arrays represent a recent significant development in microarraying and is being pursued at the C3B. Current efforts are directed at three related areas: i) surface chemistries for reproducible cDNA and oligo array manufacture, ii) implementation of this gene set on a second generation bioelectronic biochip.

In this first year, we have addressed the following project goals:

1. Complete the build-out and installation of the Advanced Microarraying Facility (AMF) of the C3B.
2. Acquire where possible, and design and build where necessary, the biochip microarraying instrumentation for the AMF.
3. Hire a Research Technician, Graduate Student, and Research Assistant Professor to support and conduct research using the C3B's AMF.
4. Test and validate the instrumentation in the C3B's AMF.
5. Begin to address technical issues related to standards and quality control in the use of cDNA and oligo arrays.
6. Seek to play a role in guiding the policies and procedures for the procurement and banking of tissues for expression microarraying.
7. Begin to define the data analysis needs for the development of the expression database of this project.

The C3B is approaching completion of its Advanced Microarraying Facility (AMF). The major hardware items [Packard MultiProbe II Liquid Handler ($ 68,000), MJ Research DNA ThermoCycler ($ 14,000), Cartesian Technologies PixSys 5000SQ Microarrayer ($ 82,531), Packard BioChip Technologies ScanArray 4000 ($ 54,000)] have each been procured. The systems are currently being tested and validated through preliminary experiments (see below).

Engineering modifications are currently being made to adapt available instrumentation to the needs of the project. This includes the installation of novel tip cleaning procedures suitable for oligo printing and the use of fiber-optic inspection cameras on the print heads.

A Research Technician, Ms. Stephanie Smith, has been hired to support the facility. Ms. Smith holds the B.S. degree (2001) in Biological Science (cum laude) from Texas A&M University. A doctoral-level graduate student has been appointed to the project. Mr. Scott Taylor holds the B.S degree (1997) in Biology from Radford University and the M.S. degree (pending) in Biology from VCU. Current efforts are

aimed at recruiting a Research Assistant Professor to support the facility and pursue active research in the area of advanced genomic biochips.

The current experimental program aims to test and validate the instrumentation of the facility while addressing the role of surface and blocking chemistries in array signal quality. Printing is in progress of a spotted 30-mer oligo array at six concentrations that vary from 100ng/mL to 1ng/mL onto glass chips possessing varying chemistries. Chip surface chemistries being evaluated include the well-established poly-L-lysine, 3-aminopropyltrimethoxysilane, a commercially available polyaminodendrimer and un-modified glass. Post-spotting blocking chemistries being evaluated include the use of bovine serum albumen and acetic anhydride.

The C3B has also initiated spotted array testing using spotted arrays manufactured from a 1,920 gene-set specifically targeted at human cancers. The C3B facility is developing its 1,920 oligo DNA chip from 50-mers purchased for $ 8,000 from MGW. This supply of oligos is capable of producing ca. 1000 chips. The 20 96-well plates obtained will be array printed in early February and will then be available for test hybridizations aimed at the development and implementation of quality assurance and control protocols and standards.

The C3B's AMF is not a high throughput facility and does not possess the throughput capacities of the cDNA laboratory or the core laboratories of the other partners. The value of the C3B's approach is found in its work with the cancer biologists, biostatisticians and bioinformaticians to identify a limited suite of genes that may be specifically employed for a targeted diagnostic purpose. Such a purpose is the molecular genetic staging of primary brain tumors. The C3B will therefore receive only brain tumor samples, along with the other laboratories, and this then will serve as the basis for comparison among the various techniques. Brain tumor samples will therefore be analyzed by GeneChip®, cDNA and oligo array technologies.

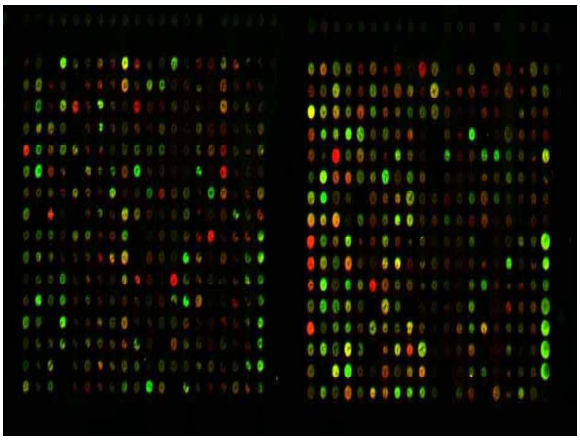


Figure 7: An example of a spotted DNA microarray with two-color detection of hybridization of probe (surface) with target (sample). The two sectors of 16 x 20 elements is shown following hybridization with two differentially labeled cDNA preparations, Cy3 (pseudocolored green) and Cy5 (pseudocolored red). The overlaying of the green and red images produces the image shown. The hue of each spot, ranging from green to red, indicates the relative expression level for the gene specific for each spot. (Image courtesy Packard Biochip Technologies).

GMU (Vikas Chandhoke, PhD - Project PI and Alan Christensen, PhD - Coinvestigator)

The GMU spotted array facility has initiated spotted array testing using rodent cDNA spotted arrays. The facility is developing a 5,000 cDNA chip, as well as a chip containing the full human cDNA set (40,000+ cDNAs). The amplifications for the 5k cDNA chip are about 25% completed and should be finished by December 1. These chips will be printed in the first half of December and will be used initially for test hybridizations. For the chip containing the full human cDNA set, the laboratory can complete ~20 - 96 well plate PCRs per day maximum. This will allow the laboratory to complete all amplifications by December 21, 2001, and begin construction of full human cDNA spotted arrays in January of 2002.

Shown below are results from a study of gene expression analysis by micro array. These slides are rat chips hybridized with H4IIE rat liver cancer cells exposed to a toxin (Jet propulsion 8) JP8 jet fuel. The panel on the right shows two of the genes whose expression has been altered by the toxin.

Figure 8: H4IIE cells exposed to 10µg/ml JP8 Jet fuel over 2 hours.

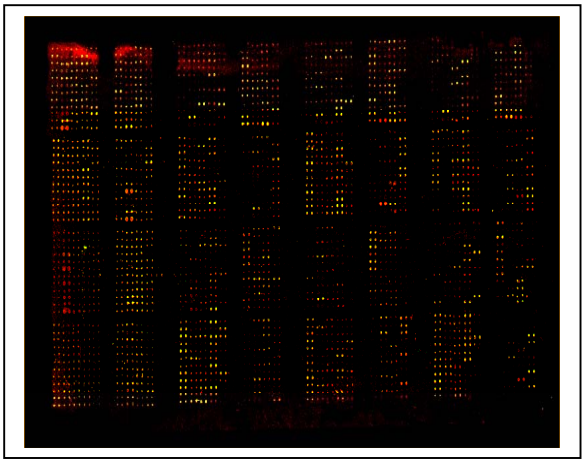
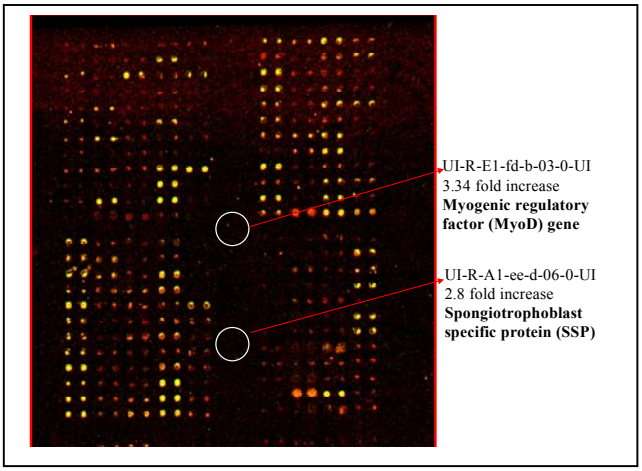


Figure 9: Over two fold increase in gene expression



Specific Reportables:

- **Intellectual property reporting - licenses, patents, etc.**

At this juncture barely 5 months into the grant, we have not yet created new intellectual property. However, given the combined competencies of the partner institutions, the evolving synergy as reflected in this report and the substantial opportunities afforded by this rapidly expanding field, we are confident that the investigators will provide appropriate return on investment in this category

- **Publications**

1) Dumur CI, Ferreira-Gonzalez A, Wilkinson DS, Ware JL and Garrett CT. Use of Affymetrix single nucleotide polymorphism arrays (HuSNP™) for assessing loss-of-heterozygosity (LOH) in prostate cell lines. J Mol Diag 3:209, 2001.
2) Brad Windle and Anthony Guiseppi-Elie "Microarrays and Gene Expression Profiling Applied to Drug Research" In Burger's Medicinal Chemistry, 6th Edition, Donald J. Abraham, Ph.D., Editor**,** John Wiley & Sons, Inc. (2002) **(submitted for publication)**

- **New applications, identified**

1) Dr. Lynne Penberthy has identified a program announcement entitled "COHORT STUDIES IN CANCER EPIDEMIOLOGY" (Release Date: October 11, 2001; PA NUMBER: PAS-02-009; National Cancer Institute). The purpose of this Program Announcement (PA) is to coordinate the submission, review, and funding of epidemiological cohort studies, and covers applications characterized by their cohort design and direct costs of $500,000 or more in any study year. Dr. Penberthy expects to apply for the June deadline after the CTRF Cancer Genomics Project has activated accrual of samples and will utilize the infrastructure of the CTRF grant to support her application.

- **Any other discoveries**

It is anticipated that the procedures established for handling of human tissues for the studies described in this grant will provide the basis for protocols to handle human tissue for clinical evaluation. While this may not result in patentable intellectual property the findings will be critical for the application of microarray technology to real-time patient care.

- **Federal money leveraged**

Once the infrastructure is established, it is likely to be used to help support research studies already supported by federal and private funding.

- **Private research money leverage**

Once the infrastructure is established, it is likely to be used to help support research studies already supported by federal and private funding.

- **Advancement of technological and economic development in Virginia as a result of this project.**

Successful establishment of the infrastructure described in this grant is certain to contribute to the advancement of technological and economic development in Virginia.

Budget and Matching Funds Summary

On the page 31 of this Annual Report is a document summarizing funding and expenditures for this CTRF Cancer Genomics Project. The document was prepared by the VCU Grants and Contracts Accounting and Effort Reporting Office (Margie R. Booker, Director).

As observed in the document, the project partners have already made significant matching fund expenditures in support of this grant. Due to the brief time that the project has been active only a limited amount of the funds allocated by the Commonwealth have been expended. However, expenditure of the current funds is anticipated by the end of the first full year of operation. Achievement of the goals and objectives of this grant are dependent on receipt of the remaining funds committed by the Commonwealth.

VIRGINIA COMMONWEALTH UNIVERSITY
GRANTS AND CONTRACTS ACCOUNTING

Project Title: Cancer Genomics and Development of Diagnostic Tools and Therapies
For Period Ending November 30, 2001
Home Account 5-35281

	Budget	Expenditures	Balance
VCU	$1,000,000.00	$23,487.78	**$976,512.22**
Match	$2,026,691.00	$385,949.96	**$1,640,741.04**
Project Total	**$3,026,691.00**	**$409,437.74**	**$2,617,253.26**

Certification

I certify that, to the best of my knowledge, the information above is correct and that all expenditures are in accordance with the grant.

Margie R Booker

Signature of Authorized Official

Margie R. Booker, Director Grants and Contracts Accounting
Name and Title

December 18, 2001

References

1. Perou CM, Sρrlie T, Elsen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumors. Nature 406: 747-752, 2000.

2. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403: 503-511, 2000.

3. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Sci 286: 531-537, 1999.

4. Sρrlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Hohnsen H et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. PNAS 98:10869-10874, 2001.

5. Huang J, Qi R, Quackenbush J, Dauway E, Lazaridis E, Yeatman T. Effects of ischemia on gene expression. J Surg Res 99:2220-227, 2001

6. Bae VL, Jackson-Cook CK, Maygarden SJ, Plymate SR, Chen J, Ware JL: Metastatic sublines of an SV40 large T antigen immortalized human prostate epithelial cell line. Prostate 34:275-82,1998.

7. Brown MA: Tumor suppressor genes and human cancer. Adv Genet 36:45-135, 1997.