Title:

Cancer Genomics and Development of Diagnostic Tools and Therapies

Virginia Commonwealth University

Annual Report

Grant # SE2002-02

1/1/03 -- 12/31/03

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)
# Table of Contents

- Title Page: 1
- Table of Contents: 2
- Summary of Progress 1/1/03 - 12/31/03: 3-5
- Specific Reportables
  - Publications and Presentations: 5-8
  - New applications, identified: 8-9
  - Federal money leveraged: 9
- Summary Statement: 10
- Summary of Project Financials (Expenditures and Match): 11

## Detailed Information
- Overview: 12
- Research Plan: 12
- Project Coordination and Management: 12-13
- Focus Group Organization: 13-14
- Tissue Acquisation Service: 14-15
  - Tissue Quality Assurance: 15
  - RNA Exchange: 15-16
- Tissue and Gene Expression Information Utilization: 16
- Tumor (Tissue) Cellular Heterogeneity: 16-18
- Clinical Data: 18-19
- Database Design
  - VCU Clinical Database: 19
    - Pathology/Massey Cancer Center Database Project: 19
  - GeneX Database at VCU: 20
  - GMU Informatics, 2003: 20
- Quality Assurance/Quality Control: 21
- Interlaboratory Comparison Studies
  - Preliminary Comparison of Gene Expression Measurements of Human Reference RNA from Affymetrix and Custom Spotted Array Platforms: 21-22
  - Quality control and quality assurance in the era of microarray: Results from an interlaboratory sample exchange: 22-23
- VCU Molecular Diagnsotics Laboratory
  - Reproducibility in Microarray data analysis from small sample preparations: 23-24
- VCU Nucleic Acid Research Facility (NARF): 25
- VCU C3B Laboratory: 25-26
- GMU: 27-28
- Data Analysis
  - Gene Expression Data Analysis (NCCLS Draft Document): 29-31
  - Preliminary Analysis of Different Algorithms for Breast Cancer Classification using Some of the Preceding Methods and Approaches: 31-32
- Supplemental Information
  - Tissue and Gene Expression Information Utilization Documents: 33-34
  - Clinical Data (VCU): 34-35
  - Custom Spotted Array(Microarray) QA/QC Procedures (GMU): 35-36
The Purpose of this Grant

Develop infrastructure and collaboration between two Virginia universities - Virginia Commonwealth University and George Mason University - and a major health care provider in Virginia – Inova to recruit extramural funding for research and develop new business opportunities for the Commonwealth of Virginia through creation of new diagnostic tools and therapies for use in patients with cancer.

Milestones and Progress To Date

A. Project Coordination and Management

1) Conduct regular meetings of project leadership – **Done** – the project PIs meet monthly on the first non-holiday Monday of each month. Meetings are held simultaneously at VCU and Inova and are facilitated through televideo conferencing. Dr. Garrett alternates between the two sites on a monthly basis. Combined meetings between the three institutions have been held by the database design, data analysis and tissue bank focus groups over the past year. In addition, Dr. Garrett holds a regular meeting each Tuesday and Thursday morning with Focus Group leaders dealing with issues of tissue acquisition, clinical data sets, database design and QC/QA at VCU.

2) **Initiate 3rd Year funding** – **NOT Received from Commonwealth of Va as of 12/15/03**

3) **Working groups** to address task focused issues – **Done**

4) Assemble required reports to CTRF
   a. 2nd year Progress Report **Done (Submitted 10/15/03)**
   b. Annual Report – ([this report](http://www.ctrf-cagenomics.vcu.edu/))

5) Listservs and Website to facilitate communication among members and focus groups – **Done** – ([http://www.ctrf-cagenomics.vcu.edu/](http://www.ctrf-cagenomics.vcu.edu/))

B. Project Milestones

1) **Tissue Banking**
   a. Tissue Acquisition System
      i. Collect residual human tumor samples @ VCU **Ongoing - 448 specimens collected to date** (IRB protocol approved at VCU April 2002).
      ii. Collect residual human tumor samples @ Inova **Ongoing 71 specimens collected to date** (IRB protocol final approval at Inova March 2003).

   b. Broaden types of human tumors collected **In progress** – since initiating the original grant, additional tumor types now collected at VCU and Inova now include Head and Neck, Colo-rectal & Lung; in addition, at Inova Liver tissue is also collected. It is anticipated that all GI cancers and all GU cancers will be added during year III of the grant.
2) Clinical Data
   a. Determine an appropriate data set of clinical and histopathological information to
      associate with each tissue sample – **Done**
      i. clinical data – primary data (exposure/family history from questionnaires)
      ii. clinical data - encoded claims data
         1. clinical data – cancer registry data subset forwarded to the Virginia State Tumor
            Registry.
         2. clinical data - histopathology and clinical pathology from clinical laboratory
            computer system

3) Database Design - Create a database(s) into which clinical, laboratory, tissue bank
   information, and expression microarray can be stored.
   a. Database to hold Tissue Bank and Primary Clinical Data – **Done** – this database is
      currently housed at VCU and is accessed via internet and VPN card from Inova to
      enter their Tissue Bank and Primary Clinical Data.
   b. Data Warehouse to aggregate Clinical and microarray Data and tools for Data
      Analysis – **In progress** –
      i. A SQL2000 database has been developed by the VCU pathology and Massey
         Cancer Center informatics personnel for storage of multidimensional clinical data
         for the CTRF Cancer Genomics Project.
      ii. A version of the open source GeneX database (http://genex.csbc.vcu.edu/) has
          been instatiated in the Center for Biocomplexity in the Life Sciences Center at
          VCU which permits entry of both Affymetrix and custom spotted array data. At
          present it does not have capability to include complex multidimensional clinical
          data.

4) QC/QA - Develop criteria for intralaboratory and interlaboratory assessment of quality using
   standard materials and shared specimens.
   a. Develop at least 1 standard material for intra- and inter-laboratory testing – **Done** –
      Both VCU and GMU investigators have performed QC analysis on their respective
      microarray platforms using a commercial preparation of human RNA provided by
      Stratagene.
   b. Initiate intra- and interlaboratory testing comparisons – **In progress** – Each of the
      microarray laboratories is evaluating shared RNA from human breast, ovarian and
      brain cancers. Data is being combined and analyzed jointly.
   c. Study of microarray surface chemistries and blocking strategies -- VCU investigators
      have undertaken to evaluate different approaches with the view to evolving standards.

5) Data Analysis -
   a. Data mining and statistical tools for analysis of microarray data with clinical
      (including clinical laboratory) data. – **In progress**
      i. Drs. Archer, Dumur and Garrett are major contributors to analysis section of the
         NCCLS guideline for clinical use of microarrays “Diagnostic Nucleic Acid
         Arrays—Proposed Guideline”. This document provides extensive review of
         methods for data analysis that are relevant to the Cancer Genomics Project.
ii. Drs Buck and Jamison have focused on the use of the open source database and tools associated with GeneX (http://genex.csbc.vcu.edu/)

6) Chip Fabrication –
   a. GMU has fabricated an H1 & H2 human 40,000 gene chip set cDNA microarray. Each chip in the set contains 20,000 genes. Done VCU-C3B has begun fabrication of a 10k gene set oligo microarray. VCU-Microarray Core will begin fabrication of a 30k gene set oligo-based microarray.
   b. Appropriate controls, standards and protocols are being evaluated and developed for implementation by all three chip-fabrication laboratories. – In progress

C. Specific Reportables

Publications and Presentations
5) Ginder, GD, Massey Cancer Center. NIH CaBIG Site Visit Review Panel, Massey Cancer Center, Richmond, VA, September 8, 2003.
6) Penberthy, L, Overview of Bioinformatics at the Massey Cancer Center. NIH CaBIG Site Visit Review Panel, Massey Cancer Center, Richmond, VA, September 8, 2003.
9) Garrett, CT, CTRF Cancer Genomics Project. NIH CaBIG Site Visit Review Panel, Massey Cancer Center, Richmond, VA, September 8, 2003.
11) Miller, WG. Athena Society Conference, Samos, Greece, September 2003, on Trends and Issues Facing Laboratory Medicine; session title “Informatics – How to Keep Track Of, and Do Something With All The Data”
13) Van Everdink WJ, Baranova A, Lummen C, Tyazhelova T, Looman MWG, Ivanov D, Verlind E, Pestova A, Faber H, van der Veen AY, Yankovsky N, Vellenga E, Buys CHCM. “RFP2, c13ORF1 and FAM10A4 are the most likely tumor suppressor gene candidates for B-cell chronic lymphocytic leukemia” Cancer Genetics and Cytogenetics, in press. CGC 6699


15) Gorreta F., Del Giacco L., Schlauch K., VanMeter A., Grant G., Christensen A., Chandhoke V., Goodman Z., Younoszai A, Elariny, H., Ong J., Younossi ZM. “Gender Differences in Hepatic Gene Expression of Patients with Non-Alcoholic Steatohepatitis (NASH)” American Association for Study of Liver Disease, October 2003 (Submitted) Boston, Massachusetts


23) Yankovsky NK, Ivanov DV, Tyazhelova TV, Pestova AA, Skoblov M, Baranova A. “Human genes RFP2, RFP2OS, KCNRG and C13ORF1 as a new candidates for TSG deleted in 13q14 area.” Human Genome Meeting, April 27-30, 2003, Cancun, Mexico


New applications, identified

1) Karen Kurdziel, submitted 6/13/02 a response to RFP-NIH-27018 “Early Clinical Trials of Imaging Agents”. (This contract will permit the VCU Molecular Imaging Center to respond to subsequent specific RFPs for development of new imaging agents. 30 million dollars has been set aside to support applications in this programmatic area. Tissue acquisition and molecular diagnostic testing form critical components of the application.), total amount "approved" for disbursement is $2,245,168 Total award (Direct + Indirect) over the 5-years (Sept 2002-2007) Status Funded

2) Gordon Ginder, Lung Cancer Gene Expression Molecular Genetics Project, 7/1/02 - 6/30/04, $409,700 Status Funded

3) Carleton Garrett, MD, PhD, “Using Molecular Genetics to Target 'High Risk' Youth Smokers”, 7/1/03 – 12/31/03, $50,000 Status Funded.

4) Anthony Guiseppi-Elie, Genomic Classification of Primary Brain Tumors and Development of Diagnostic Biochips 01/01/04-12/31/07 - $200,000 Status Unfunded.

5) Peter O’Connell, Clinical and Molecular Measurements Predicting Optimal Breast Cancer Therapies $2,000,000 Status Pending

6) Andrew Yeudall, Site-dependent gene expression and oral cancer outcome, 7/1/04-6/30/09, $250,000/yr ($1,250,000 dc) Status Pending

7) Lynne Penberthy/Greg Miller, caBIG Initiative – 3 subprojects submitted for total direct cost $500,000 for one year. Status Pending
8) Curtis Jamison, A Virtual Data Warehouse for Microarray Data, NIH, $350,000, Status Unfunded
9) Curtis Jamison, Clustering with Knowledge for Gene Expression Data, NIH, $450,000, Status Pending
12) Vikas Chandhoke, HIV research, Acambis 2003, $325,000, Status Funded.
14) Ken Alibek, Multiple Chronic Viral Infections Induce Prostate Cancer, CDMRP, Status Pending.
15) Ancha Baranova, Human KCNRG Gene Suppressing K+ Channel Activity as a Putative Tumor Suppressor Gene for Prostate Carcinoma, CDMRP, Status Pending
16) Dos Santos, Wagner G F32 Fellowship for Dos Santos, Wagner G. Medicine, Surgery, Title: Inhibition Of Ribonucleotide Reductase In Gliomas. Description: Dr. dos Santos is working as a senior fellow on brain cancers with Dr. William Broaddus at VCU. Funds from the CTRF have nucleated a new project that will look at the effects of radiation resistance in gliomas. This data is being used as a pilot for an RO1 submission, NIH 1F32CA099384 $51,904, Status - Funded
17) Buck G, Purchase of Robotic Liquid Handler, Description: This project will provide resources to purchase a robotic liquid handler for the MicroArraying Core in the Nucleic Acids Research Facilities at VCU. This liquid handler will support spotted arraying technologies that can be applied in future years to related projects by all VCU investigators. 04/01/2003-03/01/2004, NIH Divison of Biomedical Technology, $210,541 Status - Funded.
18) Buck G (PI of VCU Sub, UMD primary), Cryptosporidium parvum genomics, pathogenesis and vaccinology, Description: The major goals of this project are to develop a panel of vaccinogen candidates for the bioterrorism agent C. parvum. Though not directly dealing with cancer, the technology developed in the Cancer Genomics CTRF was applied directly in this project, certainly enhancing our funding competitiveness. NIH-NIAID, 09/04/2003-2/29/2008, NIH U34 AI57168, $1,034,885 , Status – Funded.
19) Buck G, Functional Genomics of Metacyclogenesis in T. cruzi, Description: The major goals of this project are to identify gene expression patterns during metacyclogenesis in T. cruzi. Though not directly dealing with cancer, the technology developed in the Cancer Genomics CTRF was applied directly in this project, certainly enhancing our funding competitiveness. NIH-NIAID, 07/01/2002-03/01/2007, NIH RO1 RO1AI50196, $491,000 (For Year 2003-2004) (DC only), Status – Funded.

**Federal money leveraged** – Expertise developed on this project has been applied to other areas of research including in VCU’s NARF an analysis of host response to parasite infection. Though not directly funded by this project, the analysis tools and strategies developed were utilized. This has led to the successful competition for two large NIH R01’s and a third R01 has been submitted. Thus, the project is leading directly to new federal funding in the Commonwealth. See #11 & 17-19 above. Also, the CTRF Director has been approached by several NIH funded
investigators with requests to assist them in procuring human residual samples for their research. These requests will be considered as the Tissue Banking efforts at VCU and Inova develop.

D. Summary Statement

The CTRF supported Cancer Genomics Project has completed or made significant progress in each of its objectives. Collaboration and cooperation amongst the project PIs and other members continues to increase. Due to delay in transfer of FY03 funds to the program, there has been some delay in initiating new activities. However, this temporary delay should be alleviated by early January 2004 after receipt of FY03 funds. Note that in spite of this delay, THE AWARDEES HAVE ALREADY EXCEEDED THE REQUIRED DOLLAR-FOR-DOLLAR MATCH OF $3,000,000 FOR THE ENTIRE PROJECT REFLECTING BOTH THE SUBSTANTIAL IMPORTANCE THAT THEY ASSOCIATE WITH THE GOALS AND OBJECTIVES OF THE PROJECT AND THEIR COMMITMENT TO ITS SUCCESS. Based on the ongoing level of commitment and cooperation among the partners and the degree to which infrastructure has been developed to date, continued excellent progress is expected during the remainder of the grant.
Summary of Project Financials (Expenditures and Match)

<table>
<thead>
<tr>
<th></th>
<th>Cumulative Budget</th>
<th>Cumulative Expenditures</th>
<th>Cumulative Balance</th>
</tr>
</thead>
<tbody>
<tr>
<td>VCU</td>
<td>$2,000,000.00</td>
<td>$1,780,444.00</td>
<td>$219,556.00</td>
</tr>
<tr>
<td>Match</td>
<td>$3,836,919.00</td>
<td>$3,398,507.97</td>
<td>$438,411.03</td>
</tr>
<tr>
<td><strong>Project Total</strong></td>
<td><strong>$5,836,919.00</strong></td>
<td><strong>$5,178,951.97</strong></td>
<td><strong>$657,967.03</strong></td>
</tr>
</tbody>
</table>

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.

[Signature]

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

December 16, 2003
Overview

Cancer is the second leading cause of death in the nation resulting annually in 553,400 deaths nationwide and 13,000 deaths in Virginia (http://www.cancer.org/eprise/main/docroot/STT/stt_0). Balanced against these grim statistics are the significant advances that have been made in recent years in understanding the basic mechanisms by which cancer develops. Cancer arises when normal cells begin to replicate and migrate throughout the body in an uncontrolled manner. This loss of control occurs because some of the 10,000 or more instructions that control cell behavior becomes corrupted. Moreover, this large number of instructions provides for a multiplicity of mechanisms for maintaining their grow and spread through out the body. Most scientist believe now days that in order to fully understand human cancer cells medical science needs to: 1) measure all of the thousands of instructions from a single cancer so that the patterns of expression of these instructions can be understood; and, 2) measure these instructions in primary human cancers not just in model systems such as tissue culture or animal model tumors. This grant provided to the Awardees (VCU, GMU and Inova) directly addresses these two objectives.

Research Plan

The Grant proposes to 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. Initially, four types of cancer were chosen for evaluation (brain, breast, ovary and hematopoetic system (leukemias and lymphomas). However, this number of tumors has now been extended to include lung, head and neck, colon and hepatobiliary and it is anticipated in the next year it will be extended to include essentially all types of human cancer.

Infrastructure

While 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 logistical, technical and ethical issues. The project has been quite successful in addressing all of these issues which are summarized below and which except for the topic of information and specimen utilization, are described in detail in the Annual report of last year (http://www.ctrf-cagenomics.vcu.edu/SE2002_02AnnualReport20021231.pdf).

Project Coordination and Management

This grant represents a collaborative effort among members located at the three geographically separate sites. 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 serves as Program Director for the CTRF Cancer Genomics project.

\(^1\) 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.
**Leadership meetings:** A monthly leadership meeting is held on the first Monday of each month except where there is a conflict due to a holiday or other extenuating circumstance. Meetings are held simultaneously at Inova in Fairfax VA and at VCU in Richmond VA and connected via televideo conference. Dr. Garrett rotates his attendance each month between the Fairfax and Richmond locations.

**Website and Listservs:** In order to facilitate communication among members and focus groups, eight listservs have been created which are shown in Table 2.

<table>
<thead>
<tr>
<th>Listserv Name (listname)</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>CG-TISBK</td>
<td>Tissue Bank</td>
</tr>
<tr>
<td>CG-CLNLDT</td>
<td>Clinical and Pathology Data</td>
</tr>
<tr>
<td>CG-DBDSN</td>
<td>Database Design</td>
</tr>
<tr>
<td>CG-ANLDT</td>
<td>Analyze Data (Data Analysis)</td>
</tr>
<tr>
<td>CG-QAQC</td>
<td>QAQC</td>
</tr>
<tr>
<td>CG-FBCHP</td>
<td>Chip Fabrication</td>
</tr>
<tr>
<td>CG-LDRPI</td>
<td>Focus Group Leaders and PIs</td>
</tr>
<tr>
<td>CG-MEMBS</td>
<td>All Members</td>
</tr>
</tbody>
</table>

The Project's website went live in May of 2002 ([http://www.ctrf-cagenomics.vcu.edu/](http://www.ctrf-cagenomics.vcu.edu/)). It contains information regarding the organization and goals of the project, status of progress, reportables, new funding opportunities, links to other internet resources and publicly available data produced by the project. It is maintained by Jo Ann Breaux, Assistant Administrator for the CTRF Cancer Genomics Project.

**Reporting Requirements:** As required by Virginia DPB, the Project submitted a progress report in October 2003.

**Review of Project PI Progress and Cost Share:** Progress of projects is reviewed monthly at the Leadership Team Meetings as are project expenses and cost share contributions.

**Focus Group Organization**

In order to address the complex issues of the CTRF Cancer Genomics Project, key areas of emphasis (Focus Groups) have been identified and individuals given responsibility for leadership in these areas. The figure below outlines the groups and their leadership.

![Focus Group Leaders](image-url)
Tissue Acquisition Service

At the heart of this project to study human cancer is the acquisition of the human cancer specimens. The process by which these specimens are acquired must deal with the ethical considerations related to the fundamental reason for there to be a the specimen in the first place – which is to provide treatment for the patient. The process must also to protect patient privacy. In April of 2002, the VCU IRB approved a tissue acquisition protocol to acquire residual samples of ovarian, breast, brain, and hematopoetic cancers. A similar plan was approved at Inova in November 2002. The acquisition process ensures that the primary purpose of the sample to diagnose and treat the patient's clinical cancer is preserved through review of all specimens by a pathologist prior to removal of a sample of tumor for research. Once tissue has been removed from a patient, it becomes important from a research perspective to freeze a piece of the tumor tissue as soon as possible. Acquisition to date is summarized in the figures below:

Figure 1. Tissue acquisition as of 11/30/03 at:

<table>
<thead>
<tr>
<th>VCU</th>
<th>Specimen Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast</td>
<td>64</td>
</tr>
<tr>
<td>Bone Marrow &amp; Blood</td>
<td>299</td>
</tr>
<tr>
<td>Ovary</td>
<td>28</td>
</tr>
<tr>
<td>Head &amp; Neck</td>
<td>35</td>
</tr>
<tr>
<td>Lymphoma</td>
<td>13</td>
</tr>
<tr>
<td>Brain</td>
<td>21</td>
</tr>
<tr>
<td>Lung</td>
<td>14</td>
</tr>
<tr>
<td>Colo-Rectal</td>
<td>17</td>
</tr>
<tr>
<td>Uterus</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>448</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Inova</th>
<th>Specimen Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brain</td>
<td>14</td>
</tr>
<tr>
<td>Breast</td>
<td>32</td>
</tr>
<tr>
<td>Colo-Rectal</td>
<td>5</td>
</tr>
<tr>
<td>Lung</td>
<td>13</td>
</tr>
<tr>
<td>Ovary</td>
<td>10</td>
</tr>
<tr>
<td>Liver</td>
<td>1</td>
</tr>
<tr>
<td>Head &amp; Neck</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>71</td>
</tr>
</tbody>
</table>

Simply knowing that a piece of cancerous tissue is from breast or colon or brain is wholly inadequate. It is essential to know both the tumor’s histopathological classification and basic clinical information about the tumor. Figure 2 demonstrates the type of information contained in the Project Database that is used to monitor tissue acquisition and processing.

Figure 2. Examples of histopathological classification of Project specimens

Hematopoietic Neoplasia – Diagnosis Distribution
Tissue Quality Assurance

On August 1, 2002, the CTRF Leadership Team consisting of all of the PIs convened meeting via televideo conference at Inova and VCU to consider and adopt a tissue utilization plan. The plan defined preanalytical handling of tissue including tissue acquisition, tissue storage and supplying of tissue to Project PI Labs; preparation of RNA from tissue samples; quality assurance measures that each laboratory analyzing samples must agree to; and, parameters to monitor quality of each of the critical steps in the testing process.

RNA Exchange

The following samples have been exchanged between VCU partners and GMU. Each of these samples has been processed at VCU by the Molecular Diagnostics and NARF laboratories using the Affymetrix GeneChip platform and GMU using a custom made cDNA arrays. To date, the data has been forwarded to VCU in a raw data format.
The ovarian devitalization specimens consist of separate samples of RNA taken from tissue allowed to sit at room temperature for up to 2 hr to study the effect of devitalization on gene expression.

**Tissue and Gene Expression Information Utilization**

The creation of a human specimen repository and a database of associated clinical information leads at some point in time to the analysis of the human samples and creation of test results. In order to achieve this, VCU is in the process of forming a protocol and oversight committee (*Human Tissue and Specimen Information Utilization Committee*) and it is expected that Inova and GMU will do likewise or will elect to participate together with their colleagues at VCU. The committee members will be comprised of the Focus Group leadership plus other key faculty and personnel associated with the project. Dr. Garrett will be the committee’s Chair. The functions of the committee will be:

- to provide access and monitor utilization of tissue and associated clinical data for inter- and intra-institutional investigators
- to provide oversight of data utilization to assure that:
  - tissue and data are utilized most efficiently and with the inclusion of the appropriate experts for each project
  - data are analyzed using appropriate biostatistical methods
  - participating investigators maintain appropriate QA/QC methods
  - the research interests of the participating principal investigators are protected while insuring that the data comprise a research community resource
  - patient confidentiality and privacy are maintained

Two forms currently planned for use by the committee are included under “Tissue and Gene Expression Information Utilization Documents” in the **Supplemental Information** section at the end of this report.

**Tumor (Tissue) Cellular Heterogeneity**

In a study initially designed to evaluate the effect of devitalization on ovarian tumor tissue, it was observed that two samples taken at zero time displayed significant differential gene expression (Figure 3). One hundred and five genes were found to be more than 3 fold higher expressed in the sample that had 50% tumor than was observed in the sample judged to be 90%
tumor. The portion of the 50% tumor sample that was not cancer was histologically observed to be ovarian stroma. The gene expression from another zero time sample, which demonstrated histologically 95% normal stroma, was then examined. Based on a gene list comprising the 105 differentially expressed genes between 50% Tumor and 90% Tumor, these genes were all found to be more highly expressed in normal stroma as compared to 90% Tumor, and they were mainly grouped in the Extracellular Matrix category according to the Gene Ontology classification plotted at the right side of the gene cluster (Figure 4).

Figure 3. Scatter Plots Demonstrating Differential Gene Expression Related to Content of Normal Stroma

Figure 4. Cluster Analysis of Differentially Expressed Genes Observed in Cancer vs Normal Stroma

The significance of this finding is that genes are frequently considered to be differentially expressed if there is greater than 2 fold change.

Thus, if the investigators did not know that the two samples of tumor were from the same patient and processed at the same time, the 105 genes observed to be differentially expressed might have been incorrectly related to a difference in biological behavior such as tumor diagnosis or prognosis or other clinical features associated with the patients from whom the tumors were derived.
The results clearly indicate that careful histopathological characterization of the tumor samples studied for gene expression is critical for accurate evaluation of gene expression changes.

Clinical Data

Microarray analysis simultaneously evaluates thousands of cellular instructions controlling cell behavior. While this offers the possibility for identifying new approaches to diagnosing and stratifying patients with regard to risk of poor outcome, it is really only useful for this purpose if it adds new diagnostic or prognostic information not provided by standard clinical and laboratory examinations? Thus, it is important in any analyses of human tumor samples to evaluate the findings from microarray tests with standard clinical and laboratory findings. In this regard the CTRF Cancer Genomics Project researchers have agreed to utilize a set of primary data elements obtained from questionnaires given to each participating patient and a set of data elements that correspond to the data elements forwarded to the Virginia State Tumor Registry on each cancer patient diagnosed and treated within the Commonwealth.

Correlation of microarray analysis with standard clinical findings and clinical laboratory results can also lead to creation of potentially valuable new intellectual property in addition to whatever value it may provide from the standpoint of identifying new biomarkers for diagnosis or prognosis. Microarray analysis detects the cellular instructions that actually CAUSE cancer cell grow and metastasize. Thus, the tests identify potential targets for pharmaceutical companies to develop a drugs against which could disrupt the ability of the cancer cell to grow.
A summary of data elements being collected in the above schema are listed in “Clinical Data” under the “Supplemental Information” section at the end of this report.

The efforts underway at VCU to bring together clinical data and microarray data are shown in the figure above and described in more detail in the section Database Design below.

**Database Design**

**VCU Clinical Database**

The effort at VCU to create a clinical data repository that can be integrated with microarray data is shown in the above figure. At this juncture, a specific ‘Gene Expression’ (Gene Exprsn) database has not been selected. The most likely candidate, at this juncture is a version of the open source database GeneX (see below).

**Pathology/Massey Cancer Center Database Project**

The effort to create a data warehouse for data mining in the CTRF Cancer Genomics Project is leveraging an already ongoing project to create a common Pathology Information Systems/Massey Cancer Center Informatics data repository. The level 1 databases shown in the above Figure already exist. Using Microsoft SQL2000, Pathology IS has created a shadow of the Cerner laboratory information system database into which it has successfully imported anatomical pathology and clinical pathology laboratory data. Recently, this shadow database has been expanded to include tables for cancer registry data, claims data, and the data from the Tissue Bank database. A summary of the structure and concepts embodied in this effort are indicated in the figures below:

Figure 5. VCU Pathology/MCC Data Warehouse Project
GeneX Database at VCU

At VCU the main effort to create a database for managing micrarray data has been conducted by the Nucleic Acid Research Facility (NARF). Brad Freeman, DBA and supported in part on this grant, has been working in a team with Jae Lee, Tom Laudeman and Jodi Kanter at UVA, and Daniela Puiu, Li min Wen, Mike Davis and Greg Buck at VCU to create a database for archiving and basic analysis of Affymetrix microarray data. The system, now installed at http://genex.csbc.vcu.edu, permits automated uploading of data into a data archiving system, download at selected sites, and a limited ‘analysis tree’. This work has been submitted for publication in Biotechniques (see Lee et al., above). We are currently attempting to adapt the code of the database to permit uploading and archiving of custom spotted array data. All code is ‘open source’ and therefore the utility and availability of the program will be an important asset. Adam Putnam, Ford Sleeman, and Chris Wash, funded as student hourlies in this grant, have also contributed significantly to the development and maintenance of these databases.

GMU Informatics, 2003

GMU has completed the computational infrastructure to collect microarray data. A LIMS system (BASE) has been deployed to collect the raw data, and analysis solutions have been identified and installed (GeneSpring, R/BioConductor). A hardware firewall has been installed to protect the data, and network issues have been resolved. Efforts to create the clinical database and to work out methods for data exchange are currently underway.

Create or identify existing databases into which expression microarray data can be stored in electronic format in real time at this juncture.

- Identified GeneX and BASE as candidate microarray database.
- Worked with GeneX developers and UVa to modify GeneX to accept both cDNA and Affymetrix gene expression data.
- Instantiated new version of GeneX.
- Defined new LIMS schema for data management.

Identified existing databases into which clinical, laboratory, tissue bank information, and expression microarray can be stored in electronic format in real time at this juncture.

- Examined several available clinical databases and found none to be sufficient in terms of performance and flexibility.
- Used CGO as starting basis to generate new clinical schema.
- Currently implementing clinical database.

Establishing ODBC links between separate databases containing clinical, laboratory, and tissue bank data.

- In progress.
- cDNA microarray data exchanged with VCU, awaiting Affymetrix data on same samples.
Quality Assurance/Quality Control

Prior to initiating studies with RNA from human residual specimens, the testing laboratories agreed to analyze a sample of RNA obtained from Stratagene (Product Number 740000-41; Human Reference RNA) using a protocol devised within each testing laboratory that would evaluate possible sources of variance within that laboratory's specific platform. A manuscript summarizing the experience of the VCU Molecular Diagnostics Laboratory experience with this analyte has been submitted (Dumur et al, JMD). An effort to compare and evaluate this data between each of the participating laboratories (see below) is ongoing. In addition, a second interlaboratory comparison examining human breast cancer RNA has begun.

Interlaboratory Comparison Studies

Preliminary Comparison of Gene Expression Measurements of Human Reference RNA from Affymetrix and Custom Spotted Array Platforms (Dumur, CI)

Assumptions:
The intensity values for every channel can be considered as single experiments similarly as the Affymetrix conditions.

Data Analysis:
The intensity values for the 2 channels from the two Human cDNA spotted microarrays (Human-1 and Human-2) run at GMU were compared with the Universal Human Reference RNA from Stratagene (yellow data). The intensity values were calculated by subtracting the background for every spot in every channel.

For every channel, a relative expression value was calculated for every spot. The relative expression value was calculated as the ratio of the intensity value for every spot and the intensity value of an arbitrary chosen gene.

From the 39,835 different cDNA clones located in the two Human chips, 8,191 share the same Unigene name with probe sets present in the HG-U133A Affymetrix chip (querying the same genes).

In order to compare the Affymetrix data from one of the 16 chips run in the QAQC study, with the yellow data from GMU, we calculated relative expression values the 8,191 probe sets using the same arbitrary gene.

The relative expression values, or fold changes relative to the arbitrary gene, were log2-transformed, to better visualize the results.

Results and Conclusion:
From the 8,191 data points, 2,921 and 2,724 genes, for channel 1 and channel 2 respectively, changed in the same direction as the same genes in the Affymetrix data set.
A subset of 1,072 and 846 genes, for channel 1 and channel 2 respectively, changed in very similar magnitude as the same genes in Affymetrix data set as depicted in the scatterplots in Figure 6.

Figure 6.

Quality control and quality assurance in the era of microarray: Results from an interlaboratory sample exchange (Dumur CI, Mallonee D, Buck G, Archer K, Garrett CT, Nasim S, Wilkinson DS and Ferreira-Gonzalez A)

The authors established an interlaboratory sample exchange to assess interlaboratory variability for the Affymetrix HG-U133A high-density oligonucleotide microarrays. They used a the Stratagene Universal Human Reference RNA, as well as RNA from 9 ovarian tumors. The RNA quality was assessed with a bioanalyzer and 28S/18S ratios were calculated for every sample. The Reference RNA systematically showed lower 28S/18S ratios than the ovarian tumor RNA samples. Expression values for every chip were calculated and significantly altered probe sets were identified. From the experiment using the Reference RNA, they found 6624 significantly altered probe sets with 217 of them showing a 3-fold or greater difference between the two laboratories. When the samples exchanged between the 2 laboratories corresponded to identical aliquots of carefully quality controlled total RNA isolated from 9 different ovarian tumors, they found 3002 significantly altered probe sets although only 51 of them showed a 3-fold or greater difference between the two laboratories. In both experiments, a limited number of probe sets corresponding to rRNAs were systematically higher in one laboratory than the other. This may be due to technical differences during the cRNA labeling process. Overall the results showed that a better reproducibility in microarray analysis was obtained when a good quality RNA was exchanged.
The statistical significance of altered probe sets between the two laboratories was assessed with paired t-tests and multivariate permutation tests where parametric p values are used and the permutation distribution of these p values is determined. A probe set was declared to be significant when the p-value was less than the $\alpha$-level=0.0001. Thus, for every different algorithm, scatterplots were created with only the significantly altered probe sets. Those probe sets showing less than 3-fold change were plotted as black dots and those altered probe sets with greater than 3-fold change were plotted as red dots. Among the red dots we clearly identified two groups of highly altered probe sets. One group contained probe sets that were systematically **DOWN** in Lab1 compared to Lab2 (green circle), corresponding to probes that query for ribosomal RNA (rRNA) species. A second group of probe sets was identify that contained probe sets that were **UP** in Lab1 compared to Lab2 (blue circle), corresponding to probes querying for endogenous labeling spike controls used only by Lab1 in the ovarian and breast tumor samples analyzed.

High-density oligonucleotide microarray technology has proven to be an excellent approach for gene expression analysis in different human cancers. This technique assesses the expression of thousands of genes simultaneously, but requires a minimum of 5 µg of the total RNA sample to run an experiment. This total RNA requirement poses a challenge when studying small unique clinical samples, like biopsies. Recently, a new protocol for small samples has been released by Affymetrix\textregistered, that uses as little as 1 ng of total RNA in a linear amplification method to run an experiment. The amplification process involves two successive rounds of cDNA synthesis and in vitro transcription. This method may cause a bias in the gene expression data if underrepresented transcripts end up being undetected. Here we evaluated the impact of the amplification process in the gene expression data obtained from 5 µg and 100 ng of total RNA isolated from the same human ovarian tumors. The expression data was analyzed with different algorithms and statistical tools to determine which genes were affected by the amplification process. Our results showed that a relatively small number of probe sets were significantly
affected by the amplification protocol, but the overall gene expression profile of the samples analyzed was not significantly altered.

Figure 8.

A. Hierarchical clustering of the genes and samples analysis using a Pearson (centered) correlation between the eight chips based on the 447 labeling protocol-specific probe sets.

B. Scatterplots of average log-expression of the labeling 447 protocol-specific probe sets within the two classes (n= 4 arrays per class): Small Samples vs. Standard protocol. The dashed lines show the cutoff and the red dots show those significant probe sets with 3-fold or greater difference between the two classes.
During 2003 NARF started processing human tumor samples for comparison of results with the other labs working on this project. They have completed the processing of 9 ovarian cancer samples, 7 breast cancer samples, 3 brain cancer samples and 5 ovarian samples from a time course study. Labeled RNA from these samples was hybridized to Affymetrix HG-U133A chips. The data obtained will be analyzed for gene expression at present/absent as well as quantitative levels. Comparison of their results with the results obtained by the other labs will help to provide useful data on the reproducibility of Affymetrix microarray data from different labs. They will also be able to assess cross-platform reproducibility issues associated with comparison to spotted array results. This project is being performed across several of the subprojects at VCU and GMU and the results will be compared.

Dr. Kellie Archer, whose salary is funded 33% through this subproject of the CTRF, has been working to develop new technologies for the statistically significant analysis of Affymetrix arrays. Dr. Archer has also participated in analysis of data generated in the Cancer Genomics project and will continue to play a significant role.

VCU C3B Laboratory

The VCU C3B Laboratory has made substantial progress over the previous 12 months. The accomplishments include:

- **Designed** and established a microarray fabrication facility housed in a class 1000 clean room.
- **Evaluated and Procured** a 10K oligonucleotide gene library
- **Identified** a brain tumor specific gene set comprising 341 genes that are to be included in our current C3B 10K Oligo library
- **Designed** oligonucleotide microarray to be useful in cross-platform comparison studies between Affymetrix and GMU microarrays – C3B 10K Oligo Microarray
- **Concluded** a microarray design parameter study: “Surface Chemistries and Blocking Strategies for DNA Microarrays” Manuscript to be submitted to *Journal of Nucleic Acids Research*
- **Produced 50** C3B 10K Oligo Microarrays.

Since the primary focus of the C3B effort is to investigate changes in gene expression in brain cancer, the first challenge was identification of an appropriate set of genes to study. In addition, the set of genes needed to include sufficient number of controls that could be compared between GMU and VCU microarray platforms.
Importantly, from the standpoint of quality control were the studies to define surface chemistries and blocking strategies for DNA custom spotted microarrays and the subsequent production of the C3B 10K Oligo Microarray and evaluation. The Laboratory is currently proceeding with testing of the reference RNA already evaluated by each of the other CTRF laboratories.

### C3B 10K Array process Variability Study using the Stratagene Universal Human Reference mRNA

<table>
<thead>
<tr>
<th>Parameter to be assessed</th>
<th># of Slides</th>
<th>Protocol comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slide variability</td>
<td>5</td>
<td>Slides taken from arrayer deck as in figure 1. mRNA labeled with Alexa 555</td>
</tr>
<tr>
<td>Polymerase variability</td>
<td>4</td>
<td>Two RT reactions hybridized in duplicate</td>
</tr>
<tr>
<td>Labeling variability</td>
<td>4</td>
<td>One RT reaction aliquoted into 4 labeling reactions and hybridized onto one slide per labeling reaction</td>
</tr>
<tr>
<td>Temporal cRNA stability</td>
<td>4</td>
<td>One RT reaction, one labeling reaction aliquoted into 4 samples to be applied at different times: e.g.: Fresh, 4hr, 8hr, and Freeze-Thaw</td>
</tr>
</tbody>
</table>

*Dynamic Range* | *3'/ 5' and Affy* | *Spiking controls* |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamic Range</td>
<td>3'/ 5' and Affy</td>
<td>Spiking controls</td>
</tr>
<tr>
<td>All other colors are housekeeping genes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The study consisted of printing 20 slides with ~5,200 genes (cDNAs). The reference RNA was then labeled in 5 separate reactions with either Cy3 or Cy5 dye. The 5 Cy3 reactions were pooled and the Cy3 and Cy5 labeled RNA hybridized to each of 5 slides on 4 separate days (20 slides total). Results indicated little variability from separate labeling reactions. Moreover, the normalized log reg/green ratios for greater than 96% of genes were tightly clustered around the expected value of 1.0 (0.5<ratio<2.0). Thus, the overall reliability of the custom spotted array system was judged to be high.

<table>
<thead>
<tr>
<th>Variance</th>
<th>Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labeling Reaction</td>
<td>Pooled Reactions vs. Individual Reactions</td>
</tr>
<tr>
<td>Slide Variation</td>
<td>Changes between individual slides for pooled reactions (factoring in effect of different hybridization chambers over separate runs)</td>
</tr>
<tr>
<td>Hybridization Chamber Differences</td>
<td>Changes of mean between chamber hybridizations for multiple runs (factoring in effect of between slide variation).</td>
</tr>
<tr>
<td>Run to Run Variability factors: Wash solutions hybrid oven temp handling – other</td>
<td>Between run comparisons of gene expression intensities controlled for hybridization chamber over multiple runs</td>
</tr>
</tbody>
</table>

Variability of separately labeled cDNA vs. pooled cDNA, including the variation between slides and hybridizations

Frequency distribution of Cy5/Cy3 ratio

Day 1 96.3 %
Day 2 97.8 %
Day 3 99.2 %
Day 4 99.6 %
**Reference development and study:** Cross comparison of different experimental data in the custom spotted array platform requires the introduction of a common reference into the analysis of a large number of samples. GMU has investigated the effect of the introduction of a reference into a study, and how it interferes with the final results. They performed their studies using a pool of 10 different cell line RNAs as a common reference. The reference was used to compare samples and also to evaluate chip-to-chip variability. In parallel they developed a custom reference specifically designed using the entire collection of clones printed on the microarray. In comparison to the “pool of cell lines” reference that is able to represent statistically 50% of the genes printed, their custom ‘complete’ preparation is capable of represent up to 97%. In addition use of this complete reference gives more reproducibility within the results in “self vs self” experiments. An extensive study also demonstrated that the custom reference preparation results in less deviations of the analysis of different RNA samples compared to the pool of 10 cell lines.

**Detailed Quality Control Measures Applied to Custom Spotted Arrays:** See *Custom Spotted Array(Microarray) QA/QC Procedures (GMU)* in the Supplemental Information section at the end of this report.

**Data Analysis**

The general approach to evaluating microarray and clinical data is shown in the diagram below. In essence, one develops a model consisting of one or more outcome measurements and parameters (predictor variables) that one believes may be important in causing the outcome.

\[
Y = A + B_1X_1 + B_2X_2 + B_3X_1X_2 + \ldots + B_kX_k (+ E)
\]

- **Y** = Outcome
- **X_1** = Genomic Signature
- **X_2** = Environment – Behavior – Exposure
- **X_3** = Comorbidity
- **X_4** = Treatment
- **X_5** = Demographic
- **E** = Unmeasured, Mismeasured

Members of the VCU Massey cancer center have had extensive experience in this type of analysis with regard to studying the relationship between outcomes and predictor variables obtained from joining Cancer Registry, claims, and clinical trials data.
Another resource to the Project in the area of data analysis are efforts by Drs. Garrett and Archer. Dr. Garrett is an advisor and Dr. Archer a member on the NCCLS subcommittee that is developing a guideline for the use of microarrays in clinical testing. This guideline is currently under still development. A summary of the this effort in the area of microarray expression analysis prepared by Drs. Garrett and Archer is given below.

_Gene Expression Data Analysis (NCCLS Draft Document - Garrett and Archer)_

1 **Data Elements**
   The data elements that are presented for evaluation in expression microarrays, at this juncture, arise from two general schemas. These are high density oligonucleotide arrays which have been commercialized by Affymetrix, Inc and so-called custom spotted arrays. Affymetrix arrays are created through synthesis of oligonucleotides directly onto a silicon wafer using a process similar to that used in creating chips used in microprocessors (Affymetrix: [http://www.affymetrix.com/technology/manufacturing/index.affx](http://www.affymetrix.com/technology/manufacturing/index.affx)). Custom spotted arrays are created by transferring the specific probe, generally a cloned fragment of cDNA to the surface of a glass slide using a robotic spotting apparatus (Pat Brown: [http://cmgm.stanford.edu/pbrown/mguide/index.html](http://cmgm.stanford.edu/pbrown/mguide/index.html)).

2.1 **Confounding Covariate Parameters**

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Sources of Obscuring Variation in Microarray Measurements</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sample handling (degree of physical manipulation, time from extripation to freezing)</td>
</tr>
<tr>
<td>2</td>
<td>Microarray manufacture[custom spotted arrays] (missed spots, mix-ups, degraded probes)</td>
</tr>
<tr>
<td>3</td>
<td>Sample processing (extraction procedure, RNA integrity &amp; purity, RNA labeling)</td>
</tr>
<tr>
<td>4</td>
<td>Processing differences (hybridization chambers, washing modules, scanners)</td>
</tr>
<tr>
<td>5</td>
<td>Personnel differences</td>
</tr>
<tr>
<td>6</td>
<td>Random differences in signal intensity in a data set which co vary with the biological process</td>
</tr>
</tbody>
</table>

A standard that has been put forth as a minimal collection of the above parameters is MIAMI (Minimum information about a microarray Experiment) which is promoted by the Microarray Gene Expression Data Society (MGED: [http://www.mged.org/](http://www.mged.org/)).

3 **Low Level Analysis**

The process of extracting spot or probe intensity signals and correcting them for sources of variation present in items #2-4 in Table 1 are part of the process of _Low Level Analysis_.

3.1 **Extraction of spot or probe intensity signals**

Image analysis is required to translate the raw data from a microarray experiment (i.e., the set of images) into a numerical form. General issues of image analysis include addressing, the process of identifying the location or coordinates of the probes; segmentation, the process of classifying the pixels as either signal or background; and intensity extraction, the process of calculating the intensity measures for each probe.

3.1.1 **Custom Spotted Arrays**

Tool:
- QuantArray
- ScanAnalyze (available at [http://rana.lbl.gov/EisenSoftware.htm](http://rana.lbl.gov/EisenSoftware.htm))
R/Spot package – Bioconductor: http://www.bioconductor.org/

3.1.2 Affymetrix Arrays

3.2 Calculation of Expression Levels, Background, Scaling and Normalization

3.2.1 Custom Spotted Arrays

3.2.2 Affymetrix Arrays

Microarray Suite 5.0 (MAS 5.0)
Tools:
A description of the algorithms is found at:

Model Based Expression Index (MBEI)
Tools:
dChip: http://www.biostat.harvard.edu/complab/dchip/
R/Affy package - Bioconductor: http://www.bioconductor.org/

Robust Multiarray Assay (RMA)
Tools:
R/Affy package – Bioconductor: http://www.bioconductor.org/
Irazarry Publications:
http://www.biostat.jhsph.edu/~ririzarr/papers/index.html

4 Gene Filtering and Identification of Differentially Expressed Genes

Tools
R/Bioconductor: http://www.bioconductor.org/

4.1 High Level Data Analysis – Unsupervised Learning Algorithms

Unsupervised learning involves the aggregation of samples into groups
based on similarity of their respective expression patterns without making any a priori
assumptions.

Methods of Clustering
Hierarchical clustering
Self Organizing Maps

Tools
R/Bioconductor:

4.2 High Level Data Analysis – Supervised Learning and Classification Procedures

Supervised learning incorporates knowledge of the phenotypic class information of the
samples being examined. Supervised learning procedures are applied to the training set to
identify the changes in gene expression, which best make the phenotypic class
distinction. These features are then applied to an independent test set to validate the
ability of the features to make that distinction.

Methods of Supervised Learning
Linear or Quadratic Discriminants
k-Nearest Neighbors
Weighted Voting
Naïve Bayes
5. Clinical Utility

In deciding what classification algorithm to utilize, theoretically, this would be determined by which algorithm provided the highest level of clinical sensitivity and specificity when applied to a given data set. In fact, however, as pointed out by Salzberg, determining whether an apparent difference between the performance of classification algorithms is statistically significantly different is also a matter that requires further investigation and the development of standards.

Methods:

- Binomial test of comparisons
- Analysis of variance

**Preliminary Analysis of Different Algorithms for Breast Cancer Classification using Some of the Preceding Methods and Approaches** (Dumur CI)

Several statistical tools based on different algorithms are currently utilized by the scientific community to analyze DNA microarray data. These tools identify genes that are differentially expressed among defined biological systems. The goal of this study was to evaluate the classification proficiency of the currently available algorithms used to find relevant genes in different biological systems. In this study 35 breast tumors were analyzed comprising mainly infiltrating ductal carcinomas and infiltrating lobular carcinomas. These tumors were classified as follows: 20 tumors were estrogen receptor (ER) positive and 15 ER negative. In addition, 11 tumors were classified as T1, 19 as T2 and 5 as T3 or higher; and 17 tumors were classified as N0, 10 as N1 and 8 as N2 or higher. Expression values were calculated with MAS 5.0, RMA and MBEI algorithms. Lists of statistical significant genes were created with two different sets of tools: BRBArrayTools and SAM. Finally, lists of common genes for all these algorithms were created for every one of the 3 histopathological clinical classifications. The biological model to test these different algorithms was the subclassification of the 35 breast tumor samples, according to their ER status, previously determined by immunohistochemistry, as well as Tumor and Lymph Node classifications. Our results showed that 3 different gene lists were successful to cluster all the breast tumor samples into the accurate tumor classification according to their gene expression profiles. From these results we concluded that it would be advantageous to evaluate several different algorithms simultaneously and generate agreement gene lists for the discovery of biologically relevant genes.
Designed to detect the gene encoding for SOD2, the probe sets in cluster #2, were found to agree with numerous publications. Among probe sets designed to detect the GATA-3 gene, the ER(+) gene, as expected. Also, the cluster #1, were found to detect the ESR1 gene, which is generally downregulated in ER(+) breast tumors.

372 probe sets that were significantly different between the two classes. These probe sets were clustered into two separate branches: cluster #1, with probes that were overexpressed in ER(+) tumors compared to ER(-) tumors; and cluster #2, with probes that were downregulated in ER(+) compared with ER(-) tumors. Among cluster #1, were found the probe set designed to detect the ESR1 gene, as expected. Also, the probe set designed to detect the GATA-3 gene was overexpressed in the ER(+) tumors, in agreement with numerous publications. Among the probe sets in cluster #2, were found the one designed to detect the gene encoding for SOD2, which is generally downregulated in ER(+) breast tumors.

**Figure 9 - Hierarchical clustering of the genes and samples analysis using a Pearson (centered) correlation between the 35 breast tumors using 372 probe sets that were significantly different between the two classes.**

372 probe sets that were significantly different between the ER classes were clustered into two separate branches: cluster #1, with probes sets that were overexpressed in ER(+) tumors compared with ER(-) tumors; and cluster #2, with probes sets that were down regulated in ER(+) compared with ER(-) tumors. Among cluster #1, were found the probe set designed to detect the ESR1 gene, as expected. Also, the probe set designed to detect the GATA-3 gene was overexpressed in the ER(+) tumors, in agreement with numerous publications. Among the probe sets in cluster #2, were found the one designed to detect the gene encoding for SOD2, which is generally downregulated in ER(+) breast tumors.

**Analysis Example - Assessment of ER Classification**

**Principal Components Analysis (PCA)**

![Multidimensional scaling, using centered correlation](image)

**SAM / BRB-ArrayTools**

<table>
<thead>
<tr>
<th>No. of ER Class Predictor Probe Sets</th>
<th>838</th>
<th>839</th>
<th>837</th>
</tr>
</thead>
<tbody>
<tr>
<td>640</td>
<td>652</td>
<td>709</td>
<td></td>
</tr>
</tbody>
</table>

**372 in common**

**Prediction for new samples using compound covariate predictor (CCP)**

\[
CCP = \sum (1-372) \log(t-value) \times \log(MeanExpValue) - log(MeanExpValue)
\]

- **CCP > 0**
  - Sample belongs to Class 1: ER(-)
- **CCP < 0**
  - Sample belongs to Class 2: ER(+)

---

**Analysis Example - Assessment of ER Classification**

**Prediction for new samples using compound covariate predictor (CCP)**

\[
CCP = \sum (1-372) \log(t-value) \times \log(MeanExpValue) - log(MeanExpValue)
\]

- **CCP > 0**
  - Sample belongs to Class 1: ER(-)
- **CCP < 0**
  - Sample belongs to Class 2: ER(+)

---

**Figure 9 - Hierarchical clustering of the genes and samples analysis using a Pearson (centered) correlation between the 35 breast tumors using 372 probe sets that were significantly different between the two classes.**

372 probe sets that were significantly different between the ER classes were clustered into two separate branches: cluster #1, with probes sets that were overexpressed in ER(+) tumors compared with ER(-) tumors; and cluster #2, with probes sets that were down regulated in ER(+) compared with ER(-) tumors. Among cluster #1, were found the probe set designed to detect the ESR1 gene, as expected. Also, the probe set designed to detect the GATA-3 gene was overexpressed in the ER(+) tumors, in agreement with numerous publications. Among the probe sets in cluster #2, were found the one designed to detect the gene encoding for SOD2, which is generally downregulated in ER(+) breast tumors.

**Analysis Example - Assessment of ER Classification**

**Prediction for new samples using compound covariate predictor (CCP)**

\[
CCP = \sum (1-372) \log(t-value) \times \log(MeanExpValue) - log(MeanExpValue)
\]

- **CCP > 0**
  - Sample belongs to Class 1: ER(-)
- **CCP < 0**
  - Sample belongs to Class 2: ER(+)

---

**Figure 9 - Hierarchical clustering of the genes and samples analysis using a Pearson (centered) correlation between the 35 breast tumors using 372 probe sets that were significantly different between the two classes.**

372 probe sets that were significantly different between the ER classes were clustered into two separate branches: cluster #1, with probes sets that were overexpressed in ER(+) tumors compared with ER(-) tumors; and cluster #2, with probes sets that were down regulated in ER(+) compared with ER(-) tumors. Among cluster #1, were found the probe set designed to detect the ESR1 gene, as expected. Also, the probe set designed to detect the GATA-3 gene was overexpressed in the ER(+) tumors, in agreement with numerous publications. Among the probe sets in cluster #2, were found the one designed to detect the gene encoding for SOD2, which is generally downregulated in ER(+) breast tumors.
Supplemental Information:

*Tissue and Gene Expression Information Utilization Documents*

---

**Request Form for Use of CTRF Cancer Genomics Data**

Request Date: ________________________________

Investigators:  
PI_________________________________________
Co-Investigator __________________________________
Co-Investigator __________________________________

**Study Objective:** (Please list specific aims or objectives)

**Project Summary:** (Please provide a summary of methods and study design as well as proposed analyses)

**Data Requested:** (Please specify study population, data fields or tables requested.)

**Study Outcome:** (List the ultimate objective whether publication, pilot work, quality control purposes)

□  IRB Approval Applicable  YES  □  NO  □

If IRB N/A please list reason: (eg. Used for hospital QC) _______________________

If Yes IRB Number __________________________
**Data Use Agreement**

CTRF Cancer Genomics Project

*Please complete one form for each person who will be using or accessing the data for your project.*

- Project Title: ________________________________________
- Project Approval Date: ________________________________
- PI Name: ___________________________________________
- Investigator Name: ___________________________________
- Analyst: ____________________________________________

I agree to conditions of use the CTRF Cancer Genomics data released to me for the following project_____________________________________________

As part of this project I will act as ___PI, _____ Investigator, ________ Analyst.

As a condition of use for these data I agree that I will not share these data with anyone not listed on this document.

I will use the data for purposes as stated in the attached request form which has been approved by the Tissue and Data Utilization Review Board (Approval #__________).

Any use of the data for purposes other than those agreed to in this document may result in a reversal of permission to use these data and restriction of access to and use of additional existing or future Project data.

IRB Status: Exempt [ ] Approval Number # ___________________

<table>
<thead>
<tr>
<th>Signature of PI</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Printed Name of PI</td>
<td>Date</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Signature of Data User</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Printed Name of Data User</td>
<td>Date</td>
</tr>
</tbody>
</table>

---

**Clinical Data (VCU)**

‘Basic’ clinical data.

1. Demographics, risk factors, quality of life, functioning: Interviews.
3. Demographics, more tumor info, courses of treatment, status at follow-up; also previous and later cancers: Cancer Registry.
4. Other lab data (tumor markers, side-effects, pre-existing conditions, etc.): Pathology Shadow database.
5. Other diagnoses (history, co-morbidities, outcomes, side-effects) and treatments: Claims database.
6. Other data sources: clinical trials (e.g., tx detail, adverse events), bone marrow transplant (e.g., degree of match with donor).

All of these data sources will be joined and mined to create variables that are suitable for higher level analyses.

Higher level analysis -- these are the kinds of variables to be analyzed with gene expression data.

1. Demographics: age, sex, race / ethnicity, etc.
2. Risk factors, family history.
3. QALYs: Quality-adjusted life years: QoL multiplied by survival time from diagnosis.
4. Disease-free survival: Time from remission (no cancer detected) to recurrence.
5. Response to treatment: remission (or lack of progression) relative to side effects; comparison of txs.
6. All together: Outcomes analyses with gene expression [including experimental meta data], demographics, risk-factors, co-morbidities, treatments, side-effects.

Custom Spotted Array(Microarray) QA/QC Procedures (GMU)

- **Agarose gel analysis**: PCR aliquots from a selection of amplifications were analyzed by agarose gel electrophoresis in order to monitor the yield and specificity of the reactions of PCR products.

- **POPO-3 staining**: POPO-3 is an intercalating dye that binds to double strand DNA. It was used to monitor the number of spots containing PCR products and how many failed either the amplification or printing procedure.

- **Printing Carry-Over**: A number of pin washes protocols were assessed in order to avoid sample carry-over during the printing and the possible consequent spot and library contamination.

- **Hybridization and washes stringencies**: Laboratories standard procedures were optimized in order to reduce background noise and to increase specificity during hybridization.

- **“Self vs self” test**: Before comparing two different RNA populations, an error model was created. This entailed labeling identical RNA samples with Cy3 and Cy5 dye respectively, followed by hybridization. A series of hybridizations (26 slides) was use to evaluate results variability and reproducibility.

- **Negative Controls analysis**: The previously described error models enabled us to determine signal from blank spots (where nothing was printed), SSC spots (where only the resuspension buffer was printed), and from negative controls spots (where PCR negative controls where printed).
The analysis revealed that no contamination was present in the library PCR products. Extensive study of the results allowed us to filter out, from the data set, genes that presented low intensities values and high variability (no gene expression or low gene expression). Additionally, background signals as well as blank spots, SSC spots and negative controls were used to evaluate signal and background uniformity.

- **Positives Controls**: A series of lambda DNA was printed on each chip as a positive control. Lambda RNA was then added to the each RNA sample as an internal standard to monitor the efficiency of labeling and as positive control of the hybridization.

- **Real Time PCR validation**: Post data analysis, confirmatory Real Time PCRs were performed to validate the results. From our data we obtained consistent and strong agreement between microarray data and Real Time data.

- **Accuracy of the cDNA library**:
  - *Hybridization with specific probes*: single gene probes were generated labeling PCR products. The hybridizations confirmed the correct positions of the clones selected.
  - *Random Sequencing*: several of the library clones were selected at random and sequenced. Moreover, sequencing was used to verify the identities of the genes highlighted in a microarray experiment.