ANTICANCER RESEARCH *24*: 441-448 (2004)

Microarrays in Cancer Research

GERALDINE M. GRANT[1], AMANDA FORTNEY[1], FRANCESCO GORRETA[1], MICHAEL ESTEP[1], LUCA DEL GIACCO[1], AMY VAN METER[1], ALAN CHRISTENSEN[1], LAKSHMI APPALLA[1], CHAHLA NAOUAR[1], CURTIS JAMISON[2], ALI AL-TIMIMI[2], JEAN DONOVAN[3], JAMES COOPER[3], CARLETON GARRETT[4] and VIKAS CHANDHOKE[1]

[1]*Department of Molecular Biology and Microbiology, George Mason University, 10900 University Blvd, Manassas VA 20110;*
[2]*School of Computational Science, George Mason University, 10900 University Blvd, Manassas VA 20110;*
[3]*INOVA Fairfax Hospital, 3300 Gallows Road, Falls Church, VA 22042-3300;*
[4]*Department of Pathology, Virginia Commonwealth University, Richmond, VA 23298-0662, U.S.A.*

Abstract. *Microarray technology has presented the scientific community with a compelling approach that allows for simultaneous evaluation of all cellular processes at once. Cancer, being one of the most challenging diseases due to its polygenic nature, presents itself as a perfect candidate for evaluation by this approach. Several recent articles have provided significant insight into the strengths and limitations of microarrays. Nevertheless, there are strong indications that this approach will provide new molecular markers that could be used in diagnosis and prognosis of cancers (1, 2). To achieve these goals it is essential that there is a seamless integration of clinical and molecular biological data that allows us to elucidate genes and pathways involved in various cancers. To this effect we are currently evaluating gene expression profiles in human brain, ovarian, breast and hematopoetic, lung, colo-rectal, head and neck and biliary tract cancers. To address the issues we have a joint team of scientists, doctors and computer scientists from two Virginia Universities and a major healthcare provider. The study has been divided into several focus groups that include; Tissue Bank Clinical & Pathology Laboratory Data, Chip Fabrication, QA/QC, Tissue Devitalization, Database Design and Data Analysis, using multiple microarray platforms. Currently over 300 consenting patients have been enrolled in the study with the largest number being that of breast cancer patients. Clinical data on each patient is being compiled into a secure and interactive relational database and integration of these data elements will be accomplished by a common programming interface. This clinical database contains several key parameters on each patient including demographic (risk factors, nutrition, co-morbidity, familial history), histopathology (non genetic predictors), tumor, treatment and follow-up information. Gene expression data derived from the tissue samples will be linked to this database, which allows us to query the data at multiple levels. The challenge of tissue acquisition and processing is of paramount importance to the success of this venture. A tissue devitalization timeline protocol was devised to ensure sample and RNA integrity. Stringent protocols are being employed to ascertain accurate tumor homogeneity, by serial dissection of each tumor sample at 10μM frozen sections followed by histopathological evaluation. The multiple platforms being utilized in this study include Affimetrix, Oligo-Chips and custom-designed cDNA arrays. Selected RNA samples will be evaluated on each platform between the groups. Analysis steps will involve normalization and standardization of gene expression data followed by hierarchical clustering to determine co-regulation profiles. The aim of this conjoint effort is to elucidate pathways and genes involved in various cancers, resistance mechanisms, molecular markers for diagnosis and prognosis.*

The complete mechanisms of normal cell growth and survival are as yet a mystery. Cancer is the unregulated, uncontrolled, abnormal growth of cells, due to genetic mutations and alterations, which lead to abnormal expression of genes that control these processes (3-5). To unravel the putative mechanism involved in this abnormal growth is a complex and vast task requiring powerful tools, on multiple levels. In the last 2 decades, and especially the last 8 years, there have been enormous strides in the availability, versatility and integration of new technologies and resources in the field of molecular biology (6, 7). With the introduction of microarray technology, the completion of the human genome project and the emergence of the field of bioinformatics, many previous boundaries have been eliminated and the sky has become the limit, or has it? (8, 9). Early reports in this field have indicated that microarray technology, and the analysis of gene

Correspondence to: Geraldine M. Grant, Department of Molecular Biology and Microbiology, 10900 University Blvd, PWII, MSN 4D7, Manassas VA 20110, U.S.A. Tel: 703-993-4292, Fax: 703-993-4393, e-mail: ggrant1@gmu.edu

Key Words: Microarray, cancer, genomics, across-platform.

Table I. *Tissue bank.*

Tissue bank count	
Tissue type	Specimen count
Brain	31
Bone Marrow and Blood	254
Breast	90
Colo-rectal	17
Lung	19
Ovary	27
Liver	1
Lymphoma	10
Head and Neck	25

Table II. *Quality control/quality assurance of microarrays.*

Sample QA/QC	**Purpose**
Tissue devitalization	To ensure that the correct protocol for sample collection is in place.
Sample sharing	For sample comparison and platform comparison with the same starting material.
Chip QA/QC	**Purpose**
Plate orientation	To determine correct orientation of the 384 well plates on the chip.
Random sequencing	To ensure the integrity of the library.
Known gene probing	To ensure the integrity of the chip addressing of the clones.
Negative controls	To investigate the signal specificity and threshold levels for filtering on non-expressed genes.
POPO Staining	To Investigate the integrity of the printing, and the concentration of the cDNA printed.
Reference RNA	Chip to chip validation.
Yellow test	Chip and labeling validation. Error model for analysis protocols.
Housekeeping genes, Lambda concentration curve	Chip integrity, spot saturation.

expression this technology delivers, is capable of providing powerful and previously unattainable prognostic information for several types of cancer (10-13).

To this end, 2 academic institutions and 2 hospital systems, George Mason University (GMU), Virginia Commonwealth University (VCU), INOVA Health System and the Massy Cancer Center (all USA), have together embarked on an ambitious collaboration, funded by The Commonwealth Technology Research Fund (CTRF http://www.cit.org/ctrf-main.asp) (Strategic institutional enhancement program) to join their considerable and complementary research strengths in a collaborative, strategic, basic and translational research initiative in the field of cancer genomics. (http://www.ctrf-cagenomics.vcu.edu/)

The main aim of this project is to find the hidden correlations between gene expression, patient demographics, treatment regimens and outcomes, by enrolling a significant patient base at both the Massy Cancer Center and INOVA. One of the most exciting and promising outcomes of this project is the assembly of a working infrastructure between these institutions, through which an invaluable tissue bank has been created and a live exchange of data and technologies has been achieved.

There are a number of complex aspects and levels to this project, which must be finely orchestrated to ensure the success of this endeavor and rely heavily on the ability of each of the institutions involved to work together as a team. This complexity begins with: 1) the enrollment of each patient and patient consent; 2) the coordination of the tissue bank assembly, requiring the timely collection and storage of the samples at the time of resection. This task involves the cooperation and acceptance of the surgical staff at each site: 3) the assembly and storage of the encoded patient and tumor information, in compliance with privacy regulations-Health Insurance Portability and Accountability Act (HIPAA); 4) storage/banking of the hospital released samples at -80˚C.

This paper will focus mainly on the roles played by GMU and INOVA Fairfax in this enterprise, including the collection of patient samples, the fabrication of the cDNA chips and data analysis. In addition the essential quality control/quality assurances (QA/QC) of the entire processes, in conjunction with VCU are discussed.

Materials and Methods

Tissue bank/collection of samples

INOVA/Massy Cancer Center. Only residual tissue samples from diagnostic resected specimens, which are judged by the patient's primary and consulting physicians, including the pathologist handling the patient's case, to not be essential for diagnosis or pathologic staging of the patient's disease, are collected. Prior to surgery: 1) patients are identified for inclusion in the project from information supplied by nurse coordinators of participating surgeons; 2) informed consent is obtained from the patients; 3) samples from: a) brain cancer; b) ovarian cancer; c) breast cancer; d) leukemia; e) colo-rectal cancers and liver cancers, are collected and frozen immediately in liquid nitrogen and assigned a unique ID number; 4) the specimens are then sent to the pathology department and stored at -80˚C until released by the pathologist for expression analysis; and finally 5) a final copy of the surgical pathology report for each specimen in obtained and entered into the database (Figures 1 and 3).

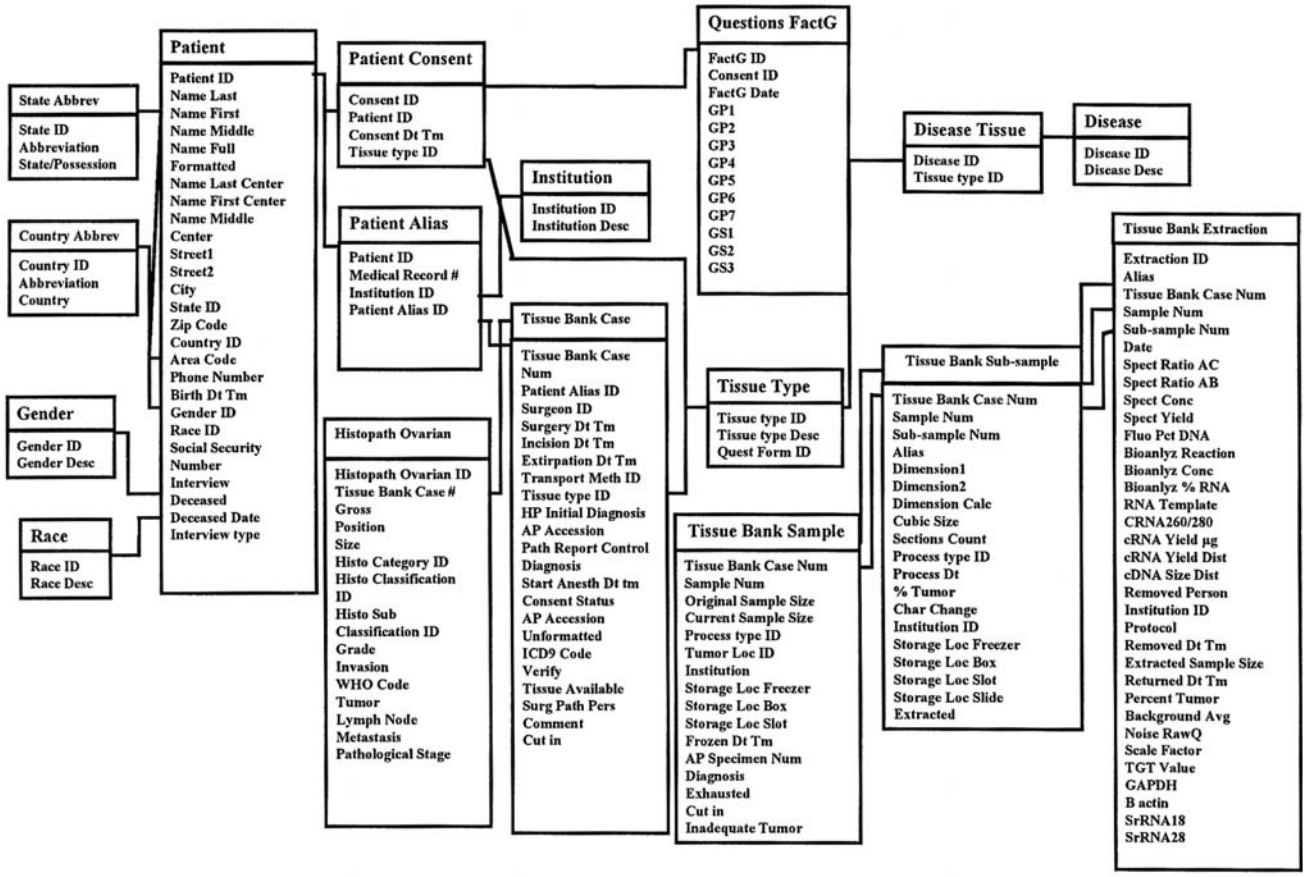

Figure 1. *Patient information flowchart.*

GMU/VCU. The project surgical pathologist carries out the final histological evaluation of residual tissue at the time of sectioning of the specimens, for RNA extraction for gene expression analysis. This procedure ensures, to the extent possible, selection of slices that are homogeneous for either normal or neoplastic tissue.

Tissue preparation and RNA extraction. Tissue specimens ranging in size from approximately 1 cm^3 - 3.5 cm^3 are sectioned frozen using a Leica 1850 Cryostat with high profile blades (Leica Microsystems). Initially, a 5-µm slice was taken, fixed and stained with Hematoxylin and Eosin (H&E). The team pathologist then determined the percentage of tumor present in the sample by viewing this initial slide. The team pathologist identified the location of any non-neoplastic tissue that is then excised from the frozen block. Once prepared in this manner, each tumor sample is then measured and, using an algorithm (based on size of the sample and OCT content, VCU), a determined number (n) of 10-µm sections are taken. Each set of (n) slices are placed in 10 mL of Trizol® (Invitrogen) and stored at -80˚C for extraction. This process is repeated again until the tumor is either (a) exhausted or (b) enough sample tubes have been collected to provide an adequate amount of RNA for hybridization.

RNA extraction. Total RNA was extracted using the Trizol® procedure (Invitrogen) with the following modifications. The sample (in 10 mL Trizol®) was thawed from -80˚C for 5 minutes at room temperature followed by the addition of 2 mL chloroform. The sample was then centrifuged for 15 minutes and the upper phase transferred to a new RNase/DNase-free tube. To this fresh tube an amount of 100% isopropanol equal to the initial volume of Trizol® was added and the mix placed at -20˚C for 2 hours to precipitate. After 2 hours the sample was centrifuged at 4000 g for 30 minutes and the resulting pellet washed once with ice-cold 70% (v/v) ethanol. The final pellet was then allowed to air dry followed by resuspension in 40 µL of RNase/DNase-free water between 55˚C and 65˚C for 10 minutes. The solubilized RNA was then passed through an RNAezy® (Qiagen) cleanup column and stored at -80˚C.

RNA amplification. Total RNA (1.5 µg) was amplified using the MessageAmp™ aRNA Kit (Ambion) according to the manufacturer's recommendations. The aRNA was quantified in a spectrophotometer and its purity was monitored by the ratio of absorbance (A260/A280). The aRNA quality was monitored by electrophoresis on agarose gels. The average size of the aRNA was evaluated with Agilent 2100 Bioanalyzer (Agilent).

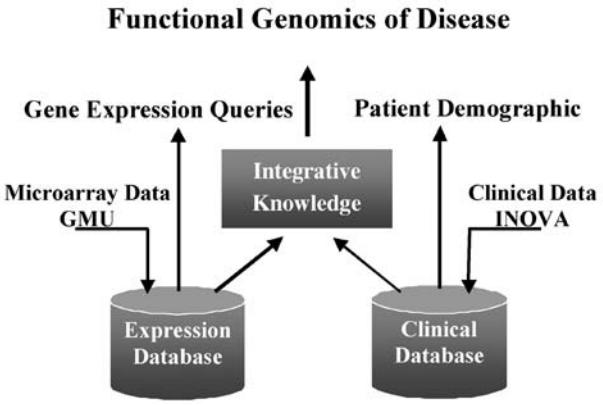

Figure 2. *Database, integrative knowledge base.*

Labeling of cDNA. RNA (10μg) or aRNA (4μg) was reverse transcribed and labeled according to The Institute for Genomic Research (TIGR) protocol (http://www.tigr.org/tdb/microarray/protocolsTIGR.shtml). Briefly, RNA was heated at 70˚C for 10 minutes in a total volume of 18.5 μl in the presence of 6 μg of random hexamer primers (3mg/ml) (Invitrogen). The reaction was cooled on ice for 1 minute. After brief centrifugation, 6 μl of 5X first-strand buffer (Invitrogen), 3 μl of 0.1 M DTT, 0.6 μl of 50X dNTPs mix containing 25mM dATP, dCTP, dGTP, 15mM dTTP(Fisher Scientific) and 2mM aminoallyl-dUTP (Sigma) were added. To this, 400 units of SuperScript II reverse transcriptase (Invitrogen) were added to the reaction. The mixture was then incubated at 42˚C overnight. RNA was hydrolyzed and the cDNA was purified using Microcon YM-30 (Millipore) according to the manufacturer's instructions. The aminoallyl-labeled cDNA was coupled with the Cyanine3 (Cy3) and Cyanine5 (Cy5) dye esters (Amersham Biosciences). After purification with QIAquick PCR purification kit columns (Qiagen), the samples were dehydrated to dryness.

Due to the lack of a "normal" control, we employ a reference RNA for each tumor made up of a mixture of 3 tumor-specific established cell lines. This reference RNA is then used across the entire experiment, allowing us to make comparisons across multiple groups. Total RNA was extracted from each set of 3 cell lines and pooled to create a reference RNA, in addition to the Stratagene's Quantitative PCR Human Reference Total RNA Reference RNA, which was employed

Microarray chip production. The GMU Microarray facility (http://www.gmu.edu/centers/genomics/resources/microarra) has designed and produced a human cDNA 40,000 Unigene (Research Genetics/Invitrogen) gene microarray (2 chips 22,000 genes/chip).

Array fabrication. The Human I and Human II chips were constructed from 417 plates of the Human cDNA library (Research Genetics). The complete list of genes with accession numbers is available at http://www.gmu.edu/centers/genomics/keys. cDNA inserts were amplified directly from clones in culture using universal library primers, GF200F (5'-CTGCAAGGCGATTAAGTTGGGTAAC) GF200R (5'-GTGAGCGGATAACAATTTCACACAGGAAACAGC).

Amplification products were purified using Multiscreen™ PCR plates (Millipore), then dehydrated to dryness and resuspended in 30μl of 3X SSC. Selected aliquots were monitored by agarose gel electrophoresis after purification to ensure (a) presence of and purity of PCR product (b) correct size of product. The collection of amplified cDNAs was printed on poly-L-lysine-coated slides in a single replicate using Gene Machines OGR-03 OmniGrid Microarrayer with SMP3 pins (Telechem International).

Microarray controls. Negative controls consisting of no-template PCR amplifications were also printed on the microarrays. Blank controls are areas that were not printed. Blanks and negative controls were distributed in several sub-arrays to monitor the background in different areas on the slide surface and the background uniformly.

A set of housekeeping genes was printed to monitor their behavior and the accuracy of the normalization protocols. Lambda DNA was also printed as positive control. Lambda RNA is spiked into each RNA sample as an internal standard to monitor the efficiency of labeling and as positive control of the hybridization (Table III).

Prehybridization. Arrays of spotted cDNAs were first rehydrated in a humidity chamber, inverted over 1X SSC, for 1 minute and 30 seconds, denatured (95˚C for 4 seconds), and finally UV cross-linked at 650 Mj. The microarrays were then incubated at 45˚C for 45 minutes in prehybridization buffer containing 5X SSC, 0.1% SDS, 1% bovine serum albumin (BSA, Sigma). After a single wash in RNase/DNase-free water, the slides were dipped once in isopropanol and air-dried.

Hybridization. Cy3-labeled cDNA was combined with labeled Cy5 cDNA in a total volume of 45μl hybridization buffer (25% formamide, 5X SSC, 0.1% SDS) for each slide, denatured at 95˚C for 3 minutes and applied to a prehybridized microarray slide under a lifterslip™ (Erie Scientific). The microarray slide was incubated at 45˚C overnight in a sealed humidified hybridization chamber (CMT-hybridization chamber, Corning Costar). After overnight hybridization slides were washed twice in 1X SSC, 0.2% SDS (10 minutes, 45˚C), twice in 0.1X SSC, 0.1% SDS (10 minutes, 45˚C), twice in 0.1X SSC (10 minutes, 45˚C), rinsed once in RNase/DNase-free water, and dried by brief centrifugation (400g 1 min).

Scanning and normalization. All image acquisition was carried out using the ScanArray Express HT confocal laser scanner with setting at 75% of photomultiplier tube, 75% of laser power and 10 μm of pixel resolution. Images were acquired by ScanArray Express 2.0 software (GSI Lumonics) and processed with Quantarray 3.0 software (Packard Bioscience).

Quality assurance/quality control (QA/QC). To ensure accuracy in each step of this process and to ensure each team is operating within the same guideline, a series of QA/QC experiments were performed. These tests are divided into two categories: Sample QA/QC and Process QA/QC (Table II).

Tissue devitalization. To ensure that the freezing process preserved the integrity of the samples RNA, large samples, that did not contain necrotic tissue, were divided immediately at resection and frozen at –80˚C at time 0, 15, 30, 60 and 130 minutes. The RNA from each sample was then extracted and analyzed by 260/280 OD, gel electrophoresis and Bioanalyzer™. Each group, as they receive

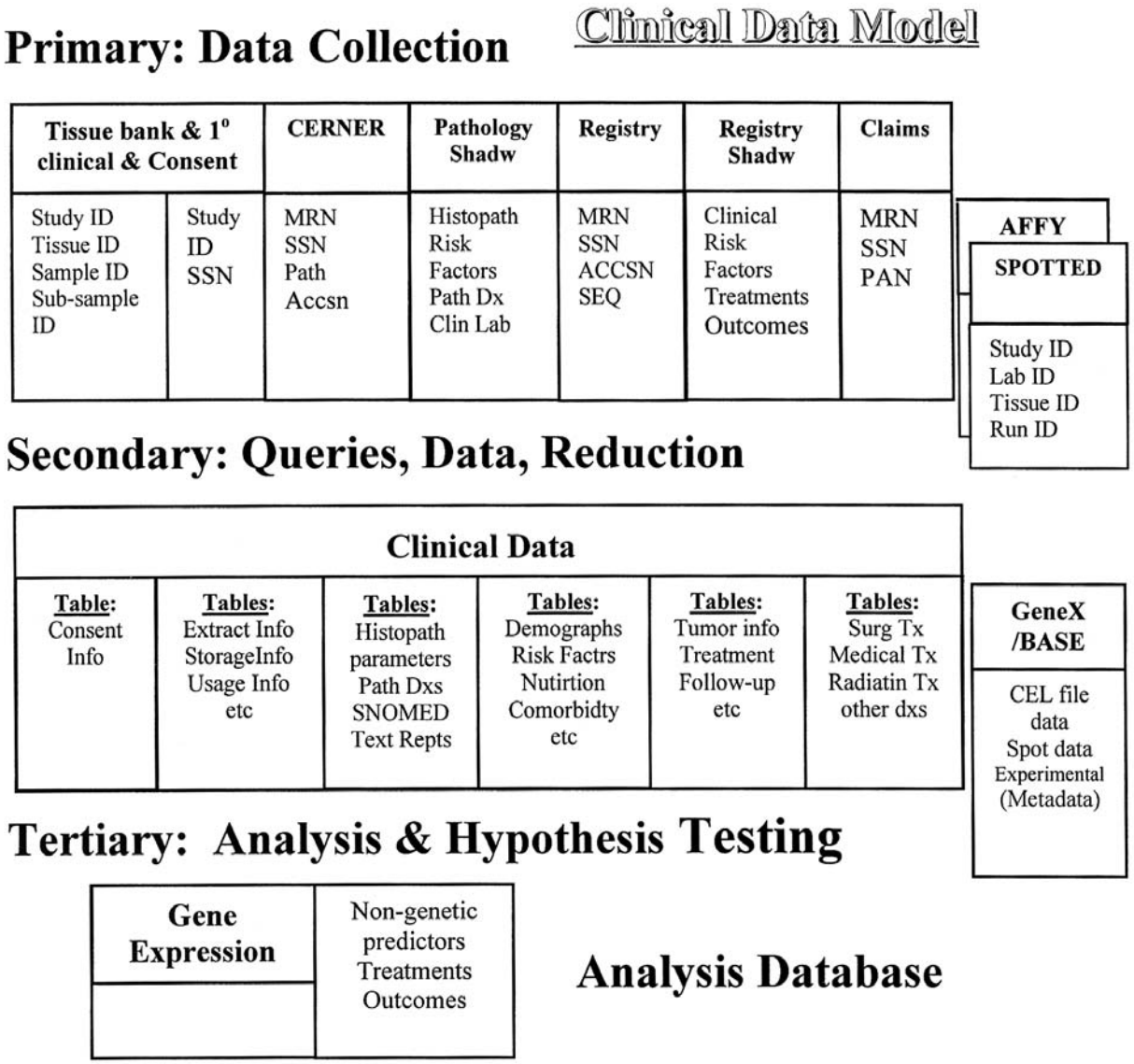

Figure 3. *Clinical data model.*

samples large enough, carry out this procedure and the resulting RNA is then shared between both GMU and VCU. This allows both groups to compare each step of the process.

Plate orientation and known gene probing

Plate orientation. To ensure the correct orientation of each of the 384 well plates loaded on the Omnigrid Micorarrayer, random clones and their destination addresses were selected such that no two selected clones would lie next to each other. Plasmid DNA was extracted from each of the selected clones and the insert DNA was amplified by PCR. Each insert was labeled using the Megaprime Kit (Amersham/Pharmacia) according to the manufacturer's protocol and using fluorescent label. The resulting labeled genes were combined and hybridized to the chip, as previously mentioned. Based on the known location of each of the clones on the microarray, a specific pattern of fluorescent spot would result. If the pattern was incorrect, it was possible to determine if the plates were place on the array deck in the inverted orientation.

Known gene probing. Ten genes were randomly selected from the library, the only essential criteria being that each of the clones were contained within the same vector. Each clone was then *in vitro* transcribed to RNA using MAXIscript kit (Ambion). The resulting RNA was then labeled as previously mentioned and hybridized to H1 and H2. This technique ensures proof of accurate gene position on the microarray and the specificity of hybridization. In addition this step reinforces our sequencing and plate orientation experiments.

Table III. *Housekeeping genes.*

Housekeeping genes	Abbreviation	3' Accession number
Ribosomal Protein L19	RPL19	AA083485
Glucose-6-Phosphate Dehydrogenase	G6PD	AA424938
Vimentin	VIM	AA486321
Tubulin, Alpha 1	TUBA1	AA180742
Phosphofructokinase,	PFKP	AA608558
Mitochondrial Ribosomal Protein L19	MRPL19	AA521243
Lactate Dehydrogenase A	LDHA	AA497029
Angio-Associated, Migratory Cell Protein	AAMP	AA452848
Actin, Beta	ACTB	AA031770
Ribosomal Protein S27a	RPS27A	AA625632
Phosphoglycerate Mutase 1	PGAM1	AA676970
Ribosomal Protein L11	RPL11	AA680244
Phosphoglycerate Kinase 1	PGK1	AA599187
Non-POU-Domain-Containing, Octamer-Binding	NONO	AA056465
Rho GDP Dissociation Inhibitor (GDI) Alpha	ARHGDIA	AA453756
Asparagine Synthetase	ASNS	AA894927
Hypoxanthine Phosphoribosyltransferase 1	HPRT1	N47312
Rho GDP Dissociation Inhibitor (GDI) Alpha	ARHGDIA	AA459400
Aldolase A, Fructose-Bisphosphate	ALDOA	AA775241
Beta-2-Microglobulin	B2M	AA670408
Heat Shock 90kd Protein 1, Alpha	HSPCA	N62400
Heat Shock 90kd Protein 1, Alpha	HSPCA	AA199881
Ribosomal Protein L29	RPL29	AI018613
Ribosomal Protein S3	RPS3	AA046713
Ribosomal Protein L19	RPL19	AA707531
Phosphoglycerate Kinase 1	PGK1	AA426516
Ribosomal Protein L19	RPL19	AA983933
Rho GDP Dissociation Inhibitor (GDI) Alpha	ARHGDIA	AA099160

Random sequencing. Random clones were selected and sequenced. The resulting sequences were compared with the known sequences for the clones accession number and clone ID. Since it is essentially beyond our financial resources to sequence each of our 40,000 clones, this method allows us to ensures accurate knowledge of the genes arrayed

Negative controls. Negative controls were included within the chip printing process. These controls are areas on the chip where the spots were a) PCR-negative controls; b) resuspension buffer only (3X SSC); c) areas included in the array format that were not printed (blanks). These negative controls test for background level and interference of PCR product/contamination.

POPO -3 staining. POPO'-3 (Molecular Probes) DNA staining was performed in triplicate to monitor the number of spots that contained PCR products and to identify genes that failed either the PCR amplification or the printing procedure. First, the 1mM stain was diluted 10,000 fold in 1X TE buffer and each slide was incubated at room temperature for 4 minutes. The microarray was then incubated for 1 minute in a primary 1X TE buffer solution and then for 3 additional minutes in a secondary 1X TE buffer solution. Using a swinging bucket centrifuge, the slides were spun-dry at 650g at room temperature for 3 minutes. The microarrays were next scanned in the Cy3 channel using ScanArray Express HT.

The intensity of a spot was considered significant if it was recorded to be higher than the median local background plus two standard deviations calculated in each sub-array.

Reference RNA. For each of our experiments we used Universal Human Reference RNA, Stratagene (Cat #740000), which is composed of RNA from 10 cell lines. This RNA was labeled as previously mentioned for either the Cy3 channel or the Cy5 channel. This use of human reference RNA allows us to have a non-variable internal standard in each of our experiments. Reference RNA allows us to look at labeling efficiency over our experiments, hybridization efficiency and slide-to-slide variability.

Yellow test/dye swap/labeling & hybridization efficiency

Yellow test. The yellow test or "self to self" test uses the same cDNA independently labeled with Cy3 and Cy5. The cDNAs were then combined in equal quantities and hybridized with the arrays. This test demonstrates labeling efficiency and hybridization efficiency. The resulting hybridization scan presents all yellow spots and the normalized Cy5/Cy3 ratio are supposed to be equal to one. This experiment quantifies the variability and determines the number of false-positives in a comparison between 2 preparations of the same cDNA population. This experiment also provides a control to set all the parameters for the analysis (filtering threshold, normalization protocols...) in order to reduce the number of false-positives and to identify a "*cut-off threshold*" for the identification of up- or down-regulated genes in a comparison of two different RNA populations.

Dye swap. To investigate labeling bias an additional dye swap experiment was carried out where first, two populations of RNA were labeled one Cy3, one Cy5 and hybridized. Second, the dyes (Cy3 and Cy5) were reversed and hybridized, to achieve a complete reverse of the order. Furthermore, combining individually Cy5-labeled cDNA with pooled Cy3-labeled cDNA, followed by dye swap, tested variability in the labeling step due to dye batch variation and technique.

Housekeeping genes, Lambda concentration curve. A series of 10 Lambda DNA spots were included on each array at concentrations between 20μg-0.002μg. Lambda RNA (Panvera) was labeled as previously mentioned and spiked into the hybridization reactions. A set of housekeeping genes (Table III) was also included in the printing of each chip for monitoring.

Data management and analysis. The CTRF Cancer Genomics project generates data from a variety of diverse sources. Broadly, the data can be characterized into three categories: clinical, experimental and metadata. Clinical data is that derived from the

patient, including clinical tests, pathology reports and demographic information. The experimental data is the actual gene expression data, derived from either two-color cDNA or one-color Affymetrix microarray gene expression experiments. Finally, the metadata is the data about the provenance of the experimental data, including the sample handling and experimental protocols used. Each category of data is handled differently, but with the ultimate goal of combining with the other data in the analysis phase.

Clinical data is often difficult to deal with. The canonical sources of data are typically dispersed across the hospital, and are often in different formats (both electronic and non-electronic). Additionally, the HIPAA regulations demand that identifying information be scrubbed from patient data before it is used for research purposes. Thus a major challenge of this project has been the collation of clinical data into a single, comprehensive data ware house structure. This has involved the creation of subsidiary databases which hold scrubbed extracts from the original data sources prior to the federation step, which creates a collated data warehouse which can be used in analysis (Figure 3).

In contrast, the handling of the experimental data and the metadata is a much more straightforward proposition, as much prior work on how to handle this data has been done previously. Specifically, the Minimum Information About Microarray Experiments (MIAME) (14) standard defines the data required to fully analyze a microarray experiment, while the Microarray and GeneExpression Markup Language (MAGEML) (15) specification allows for the exchange of microarray data using the eXtensible Markup Language (XML). The MIAME standard has been implemented in several different databases, including GeneX (16) and BASE (17). The latter database also supports MAGEML and has limited LIMS capability.

Our data management system thus consists of a redacted clinical database (built using the Sybase relational database management system) and the BASE software, which serves as the experimental and metadata database. To help protect the integrity of these data, the databases are treated as write-once, read-many repositories, meaning that after the initial deposition, all data elements become read only. When required for analysis, copies of data sets are extracted from the repositories and distributed using the GeNet software system. Data security is provided by firewalls and enhanced by housing the database servers in a limited access computer facility.

Data analysis is accomplished using a suite of software tools. Basic gene expression data analysis is done using the GeneSpring software (Silicon Genetics). More complex analyses and analysis of clinical data is performed using the S+ or R statistical software packages, utilizing BioConductor (18) or custom-made scripts. Standard data mining techniques such as clustering and classification are also applied to the data, both using an unsupervised methodology as well as partially supervised by the inclusion of prior biological knowledge incorporated *via* agent technology (19).

Data analysis in this collaborative study is complicated by the use of both two-color cDNA (GMU) and Affymetrix (VCU) gene expression technology. Although there are currently no algorithms proven to enable coordinate analysis with the two different platforms (20), representative subsets of samples (see QA/QC) are being processed in parallel by both institutions, using both protocols. Thus when methods for cross-platform comparisons become available, either through our current endeavors or other's investigations; the two data sets will be comparable *via* the common subset.

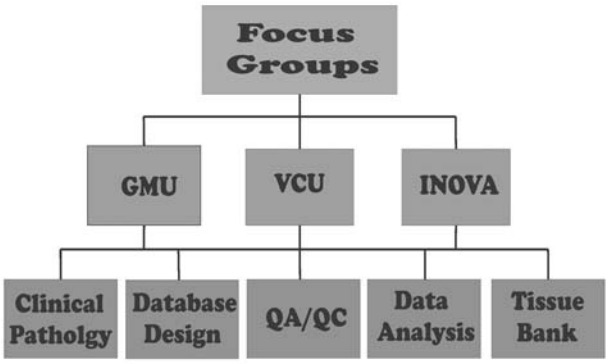

Figure 4. *Focus groups.*

Discussion

With any venture of this size, between two universities, two hospital systems and an interdisciplinary team, a strong commitment on all sides and dedication to the collaboration is required. As a result of the CTRF Strategic institutional enhancement program, we believe that, at this point, we have attained the required infrastructure that will allow us to continue down the road to eventually achieving our ultimate aim of finding the hidden correlations between gene expression, patient demographics, treatment regimens and outcomes.

We (GMU, VCU, INOVA) have amassed a diverse and expanding tissue bank (Table I) and a detailed HIPAA compliant patient record (Figure 1). As there are so many important tasks to be addressed, each task has been broken down to manageable pieces. At each site an appointed individual was identified who was responsible for leading each focus group (Figure 4), and inter-institute lists allowed each of the group members to communicate globally *via* the web. In addition, to keep each focus group on track, monthly video-conferencing meetings have been essential, giving us the opportunity to address technical and administrative issues as they occur.

One of the most intricate tasks has been coordinating the collection of the samples. This involved, enlisting surgeons, enrolling patients, assembling and completing each of the questionnaires and gaining internal review board (IRB) approval. These tasks have were intensified as each of these procedures required the involvement of a large number of people not used to working together on a regular basis *i.e.* medical staff and researchers. Furthermore, collection of the samples must take place at the time of surgery to ensure the integrity of the samples. This is a high-tension environment (the operating room) where the attending staff is constantly rotating. To address this issue a full-time individual at each site was employed, who is on call for each surgery, works in

close concert with surgeons and their staff to identify individuals who may meet the studies criteria. The long-term storage and tracking of each sample, the patient history and follow up *etc*, is of paramount importance to the success of this study. To ensure this integrity of each sample, the use of LIMS, back-up freezers and sample sharing are essential.

As already mentioned in the data analysis section, at this point direct comparison between both institutions, the cDNA microarray data and the affymetrix data, is not currently directly possible. However, research is on going to resolve this issue, which will allow us to directly compare our gene expression results online and move closer to achieving our goal.

With regard to the financial operation of this CTRF award, each institution contributed 100% in matching funds, and budget compliance on this project required annual renewal and documentation of the matching funds. This mechanism ensured institutional commitment toward the endeavor and provided a mechanism to establish infrastructure that could be used across other areas of research. The overall impact of this approach will be assessed by overall research productivity as determined by number of publications and grants generated in related areas.

Acknowledgements

We would like to thank our colleagues and collaborators, at INOVA Fairfax and VCU. INOVA Fairfax Hospital: Dr. Barry Cook, Dr. James Cooper, Jean Donavan, Rene Brenner and Mike Sheriden.

VCU, Richmond: Dr.Carleton Garrett, Cynthia Losco, Dr. Suhail Nasim, Dr. Lynne Penberthy, Dr. Gregory Miller, Dr. Gregory Buck, Andrea Ferreira-Gonzales, Dr Catherine Dumur and Dr. Anthony Guiseppi-Elie.

References

1 van't Veer L *et al*: Gene expression profiling predicts clinical outcome of breast cancer. Nature *415(6871)*: 530-6, 2002.
2 Bertucci F *et al*: DNA arrays in clinical oncology: promises and challenges. Laboratory Investigations *83(3)*: 305, 2003.
3 Knudson A: Hereditary cancer, oncogenes, and antioncogenes. Cancer Res *45*: 1437-1443, 1985.
4 Knudson A: Cancer genetics. Am J Med Genet *111(1)*: 96-102, 2002.
5 Schwechheimer K and Cavenee W: Genetics of cancer predisposition and progression. Clin Investig *71(6)*: 488-502, 1993.
6 Brown P and Botstein D: Exploring the new world of the genome with DNA microarrays. Nat Genet *21(1 Supp)*: 33-7, 1999.
7 Duggan D *et al*: Expression profiling using cDNA microarrays. Nat Genet *21(1 Suppl)*: 10-4, 1999.
8 Geschwind D: DNA microarrays: translation of the genome from laboratory to clinic. Lancet Neurol *2(5)*: 275-82, 2003.
9 Macgregor P: Gene expression in cancer: the application of microarrays. Expert Rev Mol Diagn *3(2)*: 185-200, 2003.
10 Golub T *et al*: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science *286(5439)*: 531-7, 1999.
11 Alizadeh A *et al*: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature *403*: 503-11, 2000.
12 Perou C *et al*: Molecular portraits of human breast tumours. Nature *406(6797)*: 747-52, 2000.
13 Emmert Buck M *et al*: Molecular profiling of clinical tissue specimens: feasibility and applications. Am J Pathol *156(4)*: 1109, 2000.
14 Brazma A *et al*: Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet *29*: 365-71, 2001.
15 Spellman P *et al*: Design and implementation of microarray gene expression markup language. Genome Bio *3*(RESEARCH0046.), 2002.
16 Magalam H *et al*: GeneX: An open source gene expression database and integrated tool set. IBM Systems J *40*: 552-69, 2001.
17 Saal L *et al*: Bioarray software environment: a platform for comprehensive management and analysis of microarray data. Genome Biol *3*: software0003.1-0003.6, 2002.
18 Dudoit S: Gentleman RC *et al*: Open source software for the analysis of microarray data. Biotechniques *Suppl*: 45-51, 2003.
19 Al-Timimi A *et al*: Knowledge discovery in a microarray data warehouse. Submitted, 2003.
20 Kuo W *et al*: Analysis of matched mRNA measurements from two different microarray technologies. Bioinformatics *18*: 405-12, 2002.

Received November 17, 2003
Accepted January 5, 2004