

Pediatric Genome and Expression Database

Introduction

Significant advances have been made in the treatment of pediatric cancers with overall reductions of mortality rate and increased overall survival. However the challenge remains for high-risk, disease where the survival rate has remained <30% for patients for for over a decade. One apparent reason for lack of improvement in the survival rates for these high-risk cancers has been that we have reached the maximal tolerated dosage for many of the currently utilized chemotherapeutic agents. Thus it is not possible to further increase the dose for these drugs without causing significant morbidity and mortality to the patients. There is therefore is an urgent need for the identification and validation of new targets for treating these high-risk tumors. The field of genomics and in particular the DNA microarray technology (1, 2) has accelerated the search and identification of new targets for these incurable cancers.

The Human Genome Project (HGP) has served as a platform for the launching and the development of genomic tools that allow rapid and global surveys of whole genomes. These resources, combined with rapid increases in computing processing speeds and world wide internet access, have stimulated the explosion of techniques utilizing genomic-based tools in cancer research and the microarray technology is one of these techniques. Because of the successes of the HGP it is now possible to perform quantitative gene expression levels of >30,000 genes in a single experiment with small quantities of RNA.

The Oncogenomics section (POB, CCR, NCI) has sought to apply genomic approaches to: 1) identify differentially expressed genes in pediatric malignancies, focusing in particular on high-risk disease, 2) identifying which of these genes are the “best” therapeutic targets, and 3) validating and translating the “best” molecular targets to the clinic, and 4) disseminating the data to the public. To achieve these goals we have been applying DNA microarray techniques for investigating the gene expression levels of: 1) of pediatric solid tumors, 2) xenografts, and 3) normal organ tissue that are currently utilized for pre-clinical testing of drugs. The overall goal for our section is to translate genomic approaches to the clinic thereby improving survival and the quality of life for children with high-risk cancer.

The experiments outlined above are generating large amounts of data that no single investigator can analyze in its entirety and constitutes a valuable resource for the community. The dissemination of this data to the general research community through an easy to use web site will ensure its full utilization. We plan therefore to develop a web accessible resource for those researching in pediatric cancer so that investigators throughout the world will have access to the data. The data will be useful for a wide variety of investigators including basic scientists, immunologists, translational scientists, pathologists and clinicians.

Specific Aims

The long-term goal of the proposed work is to establish a web accessible database of gene expression profiles from a wide range on normal and malignant tissues from pediatric samples.

The specific aims of the project are:

1. Develop a web accessible database of gene expression that will initially be populated by:

- i. Pre treatment tumor tissue and cell lines
- ii. Pediatric cancer Xenografts
- iii. Pediatric Normal Organ Tissue

Over the past 3 years we have performed gene expression profiling on over 500 samples of tissue. Over the next 3 years will accumulate an additional 300 more tissue expression profiles including 200 from data obtained from the Affymetrix platform. After initial data normalization and filtering of poor quality data, the experiments will be placed in a database, grouped according to samples and experiments. Investigators will then query the data for the entire data set or within a set of pre-chosen experiment or sample set. Investigators will also be able to download the selected data set for their own individual analysis. The database will be expandable to include high quality data from other investigators performing DNA microarray experimentation on the Affymetrix platform.

2. Develop and implement tools to unify cDNA and Affymetrix data.

We are performing a set of hybridizations for a widely variety of samples of different histological diagnosis on both the Affymetrix and cDNA platforms. For each gene we will determine a “standard curve”. By this and other means we will be able to import data generated by investigators using the Affymetrix platform as well as allow these investigators to export our data, in a format compatible to their platform.

3. Implement web-based algorithms for visualization and analysis and of gene expression data

Investigators will be able to utilize web-based statistical and machine learning algorithms that have already been developed in the lab, for their set of chosen experiments. They will be able to visualize data for the samples and selected genes. They will be able to export the entire data sets and use other statistical tools of their choosing.

Background and Significance:

The overall aim of cancer therapy is to develop specific therapies to improve the quality of life as well as the survival rates with the minimization of side effects of treatment. This has been partially achieved with the overall mortality rate for childhood cancer has declining nearly 40 percent in the last 25 years, a decrease of 2.6 percent per year (3) (see Figure 1 A, B below), with a current overall survival rate of 60-65%. However, despite this improving trend, several fundamental problems remain in the management of pediatric cancers. First, the choice of chemotherapeutic agents for treatment of cancer is largely empirical in nature, based on their efficacy in clinical trials and not on targeting specific genes, proteins, or pathways known to be active in that cancer. Second, the majority of these drugs target all dividing cells, including those in normal bone marrow and mucosa, which often leads to severe dose limiting toxicity. Third, other apparently idiopathic, sometimes fatal toxicities such as cardiomyopathy, may occur as maximal tolerated doses are reached. Fourth, there is currently no cure for the 35% of patients with the most aggressive disease (e.g. high-risk neuroblastoma see Figure 1 C taken from (4)).

A: Incidence & Mortality **B: Percent Surviving** **C: Survival in high-risk neuroblastoma**

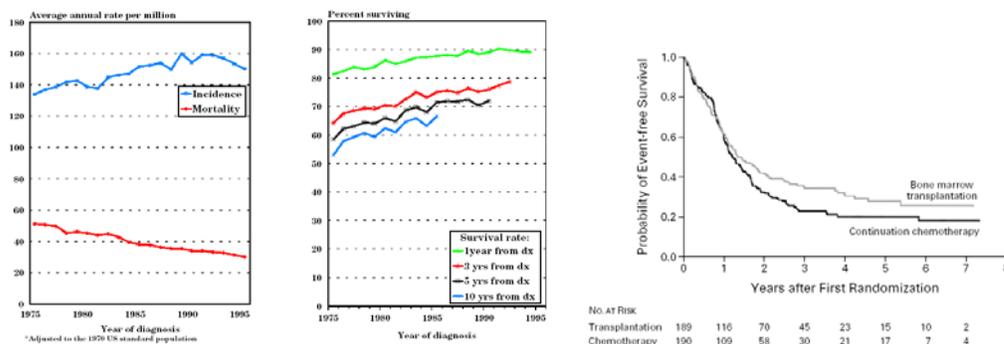


Figure 1

Additionally, despite very careful clinical and pathological prognostic stratifications, 5-30% of patients with apparently “low-risk” cancers will die from their cancer, while a similar percentage in the “high-risk” groups will survive. Therefore, there is a need for the identification for more accurate prognostic markers as well as the identification of new and specific targets for therapy in these high risk cancers. This has been the impetus for an increasing emphasis on using global genomic approaches to explore the molecular features of high-risk cancers, correlating these with diagnosis and prognosis. For the first time in scientific history, the tools and resources generated by the Human Genome Project (HGP), which include the microarray technology, allow us to tackle these fundamental problems that are evident in the management of high-risk diseases in a global and high throughput manner.

DNA microarrays were developed as a high throughput technology to monitor gene expression at the transcription level using cDNA clones whose inserts were PCR amplified, purified, and printed onto glass slides, to interrogate fluorescently labeled RNA samples (5). Monitoring global gene expression levels by DNA microarrays provides an additional tool for elucidating tumor biology as well as the potential for

molecular diagnostic classification of cancer (6). Several studies have demonstrated that gene expression profiling using DNA microarrays is able to classify tumors with a high accuracy, and discover new cancer classes (6-10).

DNA microarray technology therefore is one of the most important recent breakthroughs in experimental molecular biology, allowing monitoring of gene expression on genomic scale, and is already creating considerable amounts of valuable data. With more and more laboratories acquiring this technology, the amount of gene expression data has grown exponentially. Currently these data are scattered across various Internet sites.

Two publicly funded centers accept and make available gene expression data. In the United States, the National Center for Biotechnology Information (NCBI), a part of the National Library of Medicine within the National Institutes of Health provides the Gene Expression Omnibus (GEO) database, at <http://www.ncbi.nlm.nih.gov/geo/>. In Europe, the European Bioinformatics Institute (EBI), a part of the European Molecular Biology Laboratory, provides the ArrayExpress database, at <http://www.ebi.ac.uk/arrayexpress/>.

However although most of these public databases may be of use to highly skilled bioinformaticians who can download the data, parse it and perform local data analysis, they are not readily useful or searchable for non-microarray aficionados, who nevertheless would find the data very useful. Two recent interactions with outside investigators highlight the importance of developing publicly accessible gene expression databases. Dickens et al. (11) recognized that the overexpression of the inducible enzyme, cyclooxygenase-2 (COX-2), has been discovered in a variety of adult solid tumors and numerous studies have shown COX-2 inhibitors to have significant antiproliferative effects. Therefore, they wanted to determine the expression of COX-2 in pediatric sarcomas, and there was no public database to undertake this simple search. The investigators thus approached my laboratory for this answer. We manually searched our databases for COX-2 expression in rhabdomyosarcoma (RMS), osteosarcoma (OS), and Ewing sarcoma (EWS) samples. COX-2 expression was detected in 52/59 (88.1%) tumors on our database. They subsequently confirmed this by immunohistochemistry in independent sample sets and found an increased COX-2 expression in metastatic rhabdomyosarcoma and osteosarcoma, though it did not reach significance. The degree of COX-2 immunoreactivity did not vary significantly with other clinicopathologic features such as age, gender, or histologic classification. They concluded from this study that the majority of these pediatric sarcoma samples express COX-2 to varying degrees and testing the efficacy of COX-2 inhibitors in the treatment of pediatric sarcomas were warranted.

A second question was recently posed by outside investigators. Since attenuated viruses derived from Adenovirus (Ad) has been shown to kill tumor cells (oncolysis) and are currently in clinical trials for selected cancers, Rice et al. Who were working on Ewing's sarcomas posed to us the question what are the levels of Adenoviral (Ad) receptors in Ewing's sarcomas (12)? We searched our Ewing's database and found that both EWS cell lines and tumors expressed coxsackie-adenovirus receptor (CAR) and alpha(v)-integrins, showed high levels of gene transduction. Subsequent experiments showed EWS

to be highly sensitive to viral oncolysis. Furthermore expression was confirmed by immunohistochemistry. Thus the authors concluded that treatment of EWS may be possible using conditionally replicative adenoviruses.

A third area of usefulness for public pediatric gene expression databases is to be able to determine the level of expression of potential target genes, e.g. CAR as described above in normal, vital organs. For instance it would be prudent to avoid using Ad to treat a cancer if any of these organs also express this target gene at high levels since it may induce oncolysis in vital tissues. This is also of importance for those developing immunotherapy, e.g. vaccines, to avoid potentially serious autoimmune disease. No such public repository of gene expression levels in normal tissues currently exists.

A fourth area that will be addressed in our proposal is the ability to combine data sets obtained from divergent microarray technologies. There are several large-scale DNA microarray platforms that are currently utilized, including double-stranded cDNA as used in our laboratory as well as single stranded short 25mers (Affymetrix), mid-sized 30mer (Amersham) or long 50-70mers (Compugen or Operon) oligonucleotides. Due to the availability utilization of these multiple platforms, it has been difficult for one investigator to compare his results with another. The two most popular methods utilize the cDNA and Affymetrix platforms. We are proposing to develop explore simple algorithms for transforming cDNA data to Affymetrix and vice versa for individual genes, which will allow scientific investigators to compare their results with our data and allow us to import their data.

Establishing a growing repository of gene expression for pediatric cancers and normal organ would therefore be advantageous for the following reasons:

1. Allow translational physician scientists to determine the level of expression of a target protein that there is laboratory or clinical evidence of effectiveness in certain cancers for a variety of other pediatric cancers.
2. Allow the searching of the expression levels of target genes in normal organ tissue.
3. By combining and up loading data obtained by different laboratories that have used the Affymetrix platform, the repository will create the ability to build up progressively detailed gene expression profiles and will give access to this information to third parties.
4. It will facilitate the cross-validation of data obtained by different laboratories.
5. It will enable other bioinformatics groups, to participate in the data analysis and to explore new methods and develop new tools for such analysis.
6. It will promote a public sharing of crucial data.
7. It will create a public resource that can be referenced by the scientific literature, allowing articles to discuss data that have been deposited in the database.
8. By releasing our data into the public domain without restrictions on intellectual property, we expect to make more efficient the process of new therapy development for children with cancer.

Recently the National Cancer Institute Center for Bioinformatics (NCICB) has launched an important project (<http://cabig.nci.nih.gov/>): the cancer biomedical informatics grid (caBIG), to expedite the cancer research communities' access to data and key bioinformatics platforms. They will work in partnership with the cancer research community, in creating a common, extensible informatics platform that integrates diverse data types and supports interoperable analytic tools. Their platform will allow research groups to tap into the rich collection of emerging cancer research data while supporting their individual investigations. We will work closely with this group to ensure that our data is compatible with the infrastructure being developed by caBIG and also make our data publicly available through caBIG database. The leadership of the caBIG has agreed to provide guidance to our group, so that the data is fully compliant with the developing standards, and therefore be of maximum utility to the worldwide research community.

Preliminary Studies and Data Accumulated

Our first study has been the identification of expression profiles in the small round blue tumors (SRBCTs) (6) where we identified 93 genes whose expression levels were able to reliably diagnose these cancers. Of these genes, 41 have not been previously reported to be associated with the respective cancers. Several of these 93 genes are potential targets for therapy in these cancers, but since we were limited by a 6567-element array, the largest array available at that time, it was not certain that these represent the best possible targets. We hypothesized that: 1) using larger arrays with more extensive coverage of the human genome and analyzing more cancer samples, we will confirm our previous findings and discover newer, perhaps better molecular targets. 2) cDNA microarray analysis may identify genes whose expression levels will predict prognosis of patients, and 3) By performing expression analysis on normal tissues, we will be able determine which of these genes are expressed at low levels or not expressed at all in important organs and therefore targeting these genes thus will reduce the risk of potentially serious side effects of therapy.

Human cDNA Microarrays developed at the Oncogenomics Section

In order to broaden the coverage of our cDNA arrays we obtained a 42 481 cDNA plasmid set from Research Genetics (<http://mp.invitrogen.com/>) and PCR amplified, purified all these inserts. From this we have produced high quality cDNA microarrays and each entire set has been printed in two chips each with approximately 21 000 probes/spots. A summary of the number of Unigene clusters and unique genes represented in the entire set is shown below.

Clones	Unigene clusters	Known Genes	EST
42 578	23 389	10 270	13 119

RNA Probe Production: We have modified the RNA amplification method of Wang *et al.* (13) and combined it with indirect labeling using amino-allyl dUTP. This has produced consistent hybridization and lowered the total RNA requirement from 100-200µg to as

little as 300-500ng. This has reduced the tissue requirement from 200mg to 5mg, thus allowing us to process diagnostic needle core biopsies for DNA microarrays. This RNA amplification method, originally described by Eberwine (14), is established in the microarray field, and does not lose representation of the original RNA sample. The control RNA for all of our experiments will be a panel of 7 cell lines, shown in the table left, (NB, neuroblastoma; ERMS, embryonal rhabdomyosarcoma; CML, chronic myeloid leukemia; CCa, cervical carcinoma; Sarc, undifferentiated sarcoma; EWS, Ewing's sarcoma and BL is Burkitt's lymphoma). These cell lines were chosen to represent a broad category of cancers, primarily pediatric.

Cell Line	Type
CHP212	NB
RD	ERMS
K562	CML
Hela	CCa
A204	Sarc
RDES	EWS
CA46	BL

Static Data: Expression Profiles of Tumor and Normal Samples

The initial focus in the lab has been in the identification of expression profiles that correlate with prognosis. We have worked primarily on neuroblastoma (NB), Wilms tumors (WT), Ewing's sarcoma (EWS) and rhabdomyosarcoma (alveolar; ARMS, embryonal; ERMS and botryoid; BRMS). A summary of the samples for which there is high quality expression data available in tabular form is shown below.

Neuroblastoma

Diagnosis	Number
Stage 1	38
Stage 2	18
Stage 3	21
Stage 4 MYCN not amplified	63
Stage 2 MYCN amplified	1
Stage 3 MYCN amplified	4
Stage 4 MYCN amplified	26
Stage 4S	18
Stage 4S MYCN amplified	2
Unknown	23
Total/good quality	214/170

Others

Diagnosis	Number
Tumors	
ARMS	13
ERMS	13
EWS	23
Favorable Histology WT Non-relapsing	14
Favorable Histology WT Relapsing	13
Others	60
Total	136
Cell Lines	
Alveolar RMS	10
ERMS	3
EWS	10
NB	13
Total	172

Prediction of clinical outcome using gene expression profiling and artificial neural networks for patients with neuroblastoma

Our primary focus has been in neuroblastoma biology, and we initially analyzed our data to identify gene expression signatures in poor risk neuroblastoma. Currently, patients with neuroblastoma (NB) are classified into risk-groups (e.g. according to the Children’s Oncology Group (COG) risk-stratification) to guide physicians in the choice of the most appropriate therapy. Despite this careful stratification the survival rate for patients with high-risk NB remains less than 30%, and it is not possible to predict which of these high-risk patients will survive or succumb to the disease. We have therefore performed gene expression profiling using our cDNA microarrays containing 42578-clones and utilized artificial neural networks (ANNs) to develop an accurate predictor of survival for each individual patient with NB. Using principal component analysis we found that NB tumors exhibited inherent prognostic specific gene expression profiles. Subsequent ANN-based prognosis prediction using expression levels of all 37920 good-quality clones achieved 88% accuracy (see Figure 2 below).

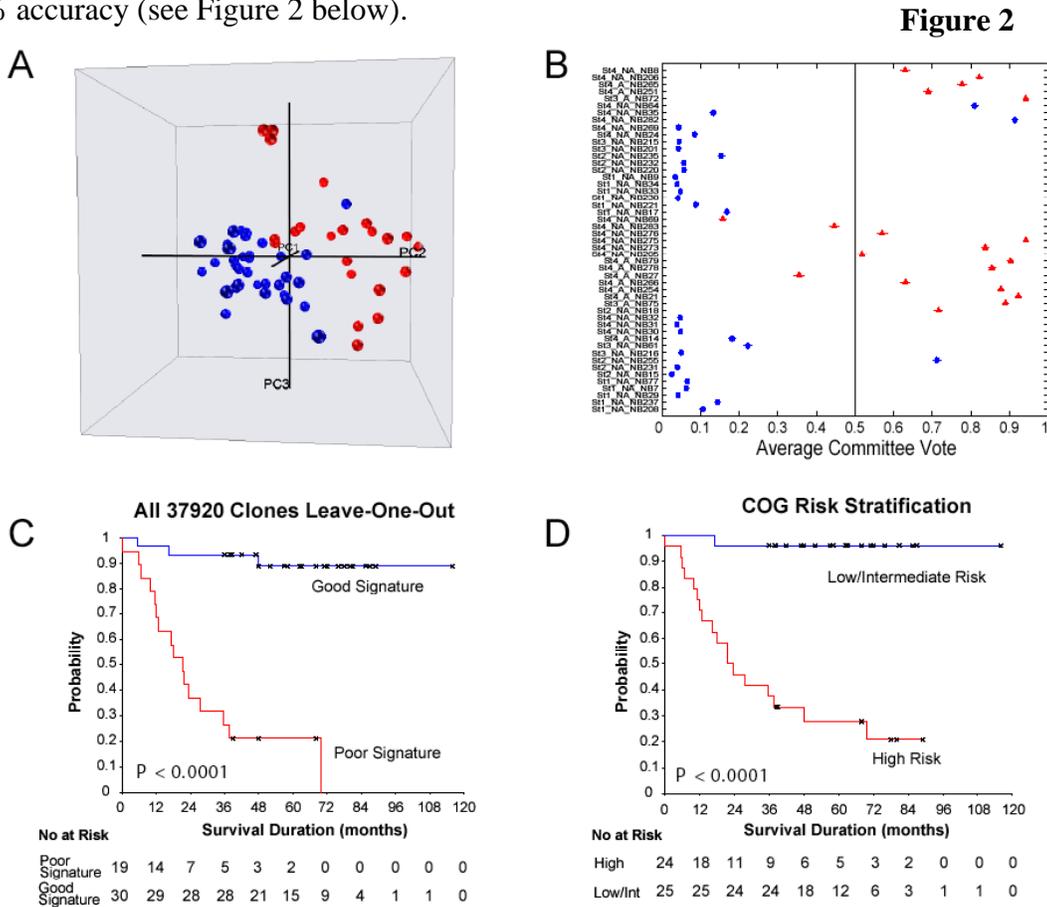


Figure 2: Predicting the outcomes of neuroblastoma without gene selection. A, Plot of the top 3 principal components (PCs) of the 56 NB samples using all quality-filtered 37920 clones demonstrates some separation according to the clinical outcome. Red spheres represent poor-outcome patients, while blue spheres represent good-outcome patients. **B,** ANN voting results for outcome prediction of the 49 unique NB patients using 37920 clones without any further clone selection in a leave-one-out prediction

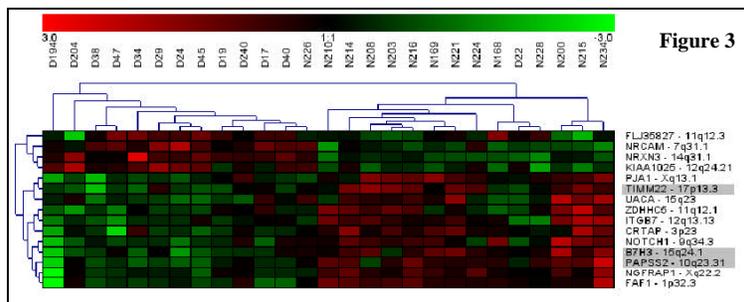
scheme. (Samples labels; St=stage, NA=MYCN non-amplified, A= MYCN amplified, followed by sample name). Symbols represent ANN average committee votes for each sample, while the length of the horizontal lines represents the standard error. Red triangles represent poor-outcome, and blue circles represent good-outcome NBs. Vertical line at 0.5 is the decision boundary for outcome prediction (i.e., good signature<0.5, poor signature>0.5). **C**, Kaplan-Meier curves of survival probability for the 49 NB patients derived from the results in Fig. 2B. **D**, Kaplan-Meier curves of survival probability for the 49 NB patients using the current COG risk stratification

Furthermore, using an ANN-based gene minimization strategy we identified only 19 genes, including two previously reported prognostic markers, MYCN and CD44 that correctly predicted outcome for 98% of these patients. In addition, these 19 predictor genes were able to further partition COG stratified high-risk patients into two subgroups according to their survival status (P=0.0005). Our findings provide evidence of a gene expression signature that can predict prognosis independent of currently known risk factors and could assist physicians in the individual management of patients with high-risk neuroblastoma.

Once the manuscript is accepted for publication we will release the entire database to the scientific community including all clinical annotations for further exploration of our data. A recent exciting development has been a collaboration with the United Kingdom Children’s Cancer Study Group (UKCCSG) in which they have agreed to provide ~200 “anonymized” samples from patients with neuroblastoma entered into their national studies who have had a > 3 year follow up. For these samples will perform DNA microarrays on the Affymetrix platform and compare and validate our findings detailed above. With this data we hope to identify confirm a set of robust markers for predicting prognosis as well as explore biological processes associated with poor risk neuroblastoma. We believe this will be a rich source of information for neuroblastoma biologists, physicians and translational scientists.

Prognostic Classification of Relapsing Favourable Histology Wilms Tumour using cDNA Microarray Expression Profiling and Support Vector Machines

We have performed a similar study in favorable histology Wilms tumor to investigate the



differences in gene expression between the 85% of patients achieving long-term survival with the 15% that relapse early in therapy using t-test and the machine learning algorithm support vector machines (SVM). We identified 15 differentially

expressed genes by these methods that were generalisable (Figure 3 above), i.e. able to predict the relapse in an independent set. This manuscript has been accepted for publication (15). Likewise this data will be good resource to Wilms tumor biologists.

Characterization of Xenograft Models of Childhood Cancers:

This array project is being conducted in collaboration with the Children’s Oncology Group (COG) Phase 1 Consortium and the Cancer Therapy Evaluation Program (CTEP) and was partially externally funded by a grant from the NCI. It is in accord with some of the requirement embodied in the “Best Pharmaceuticals for Children Act” which stated that “the Director of the National Cancer Institute shall expand, intensify, and coordinate the activities of the Institute with respect to research on the development of preclinical models to evaluate which therapies are likely to be effective for treating pediatric cancer.”

There are three main objectives of this project.

- (1) To determine which of the currently utilized pediatric cancer models (xenograft) most closely resembles the cancer(s) of origin by cDNA microarrays.
- (2) To develop tissue and cell microarrays and protein lysate arrays of this preclinical panel, for the purpose of identifying cancer-related molecular targets and signaling pathways that may be therapeutically exploited to improve the outcome for children that suffer from cancer.
- (3) To provide these tissue and protein arrays as reagents to qualified investigators in the extra- and intra-mural research community.

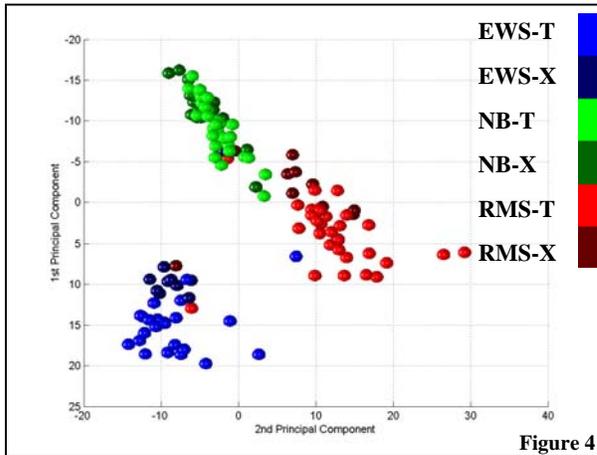
This project will facilitate pediatric cancer drug development by identifying the best xenograft models that are the most similar to the cancer of origin, and will be a resource to use in identifying molecular targets and cell signal profiles, and for subsequent testing new anti-cancer agents for their potential activity against childhood cancers. This resource has the potential to increase the pace of new therapeutic target discovery for childhood cancers and to facilitate the clinical use of new molecularly targeted agents active against childhood cancers.

Preliminary Data Analysis of Xenograft Models of Childhood Cancers:

Diagnosis	Number
Ewing’s	9
Rhabdomyosarcoma	10
Neuroblastoma	17
Osteosarcoma	10
Rhabdoid tumors	3
Wilms	7
Ependymoma	5
Medulloblastoma	7
ALL	10
Other	8
Total	86

We received 86 xenografts from several cancer treatment centers in the USA and 10 acute lymphoblastic leukemias from Australia. Initial microarray experiments showed that 3 of the xenografts were composed entirely of mouse tissue. For the data analysis we initially focused on the xenografts for which we had the human tumor counterpart from patient biopsies. This included rhabdomyosarcomas (RMS), Ewing’s (EWS) and neuroblastoma (NB). We performed principal component analysis and plotting of the first two components is shown in Figure 4. This demonstrated that the majority of the xenografts clustered closely to the respective tumor

biopsy samples validating that the global pattern of gene expression of these xenografts were indeed very similar to their cancers of origin.



This was confirmed by hierarchical clustering using all quality-filtered genes (>37 000 clones). We found that 7 samples did not cluster with their respective cancers. We are currently undertaking more sophisticated tests including, weighted gene analysis (9), signal to noise statistics or “Golub score” (8), F-test (16), Support Vector Machines (SVM) (15, 17) principal component analysis (PCA) and artificial neural networks (ANN) (6). The ultimate goal for our lab is to identify a short list of genes that

optimally classifies these cancers and may be potential targets for therapy. This panel of xenografts together with their gene expression profiles will be a valuable resource to researchers involved in childhood cancer drug development that would like to see if specific target genes are expressed in their cancer of interest. They can then use the readily available xenografts to test their hypothesis that therapy against that target will be effective.

Normal Samples

As discussed earlier, the non-specific nature of current cancer therapeutics, the lack of response in high-risk disease, and the severe, sometimes fatal, organ toxicities have driven cancer therapeutics towards rationally designed and specific molecularly targeted therapy. It is hoped that targeting genes expressed only in cancer, and not in normal tissue will minimize the side effects of therapy. These unique genes can also be used as tumor antigens for the development of vaccine-based immune therapies for childhood cancer (a primary interest of our branch of Pediatric Oncology Branch, NCI), which will reduce the risk of serious autoimmune manifestations. To achieve this aim, we have performed gene expression analysis on 185 normal organ samples. The tissues were obtained from the Maryland Brain and Tissue Bank (<http://medschool.umaryland.edu/btbank/>). The median age of the subjects from which the samples were obtained was 19yrs (range 2-40yrs), with a median post mortem interval (PMI) of 11hrs (range 4-19hrs). Approximately 85% (158) yielded intact RNA and had high quality gene expression

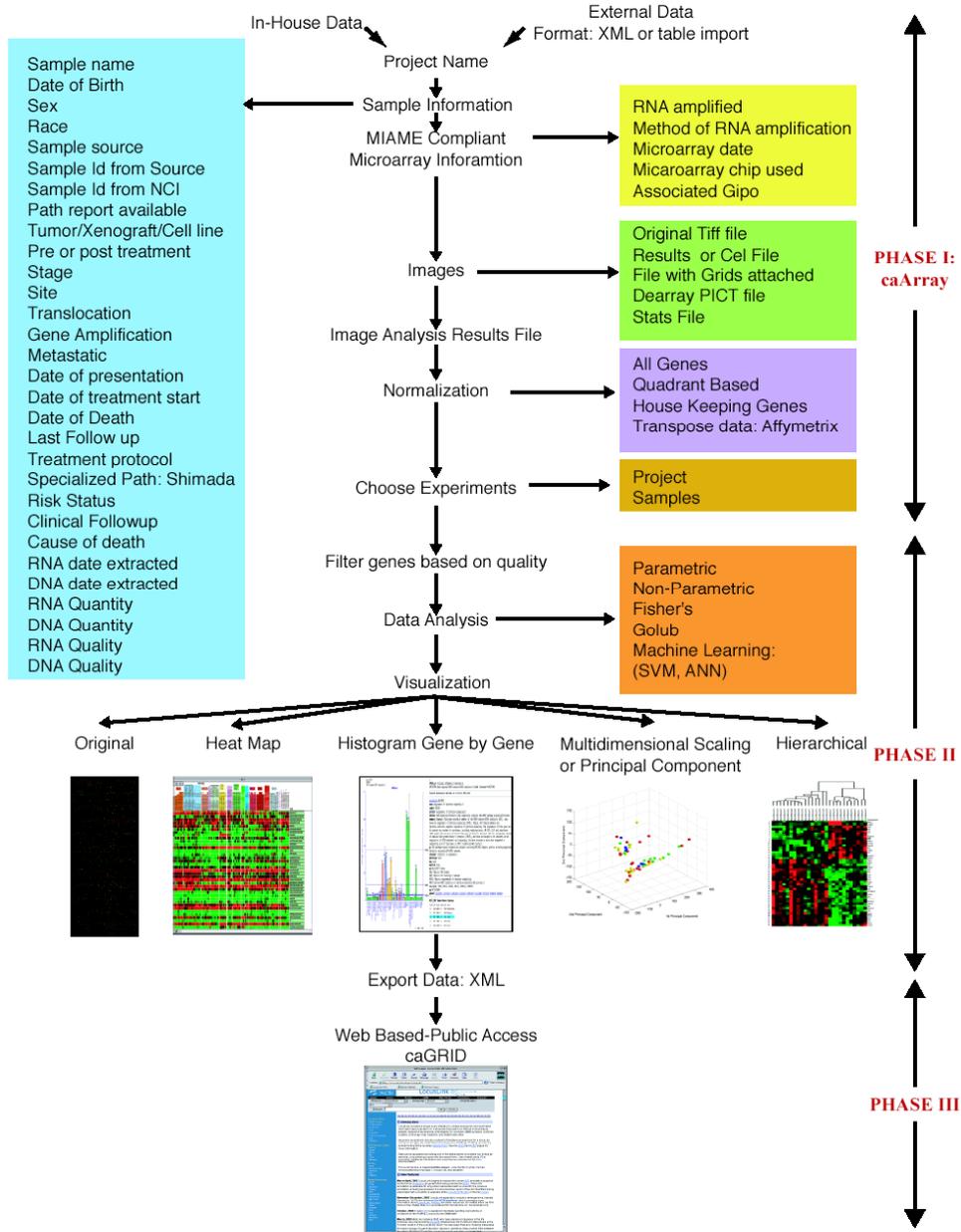
Organ	Number
Adrenal	9
Bladder	9
Cerebellum	6
Cerebrum	7
Colon	8
Heart	7
Ileum	10
Kidney	10
Liver	10
Ovary	5
Pancreas	6
Prostate	8
Skeletal Muscle	9
Spleen	10
Stomach	10
Testicle	7
Ureter	8
Uterus	10
Lung	9
Total	158

data (see above).

Experimental Design /Methods

The work will be carried out in three overlapping phases and is summarized by the figure 5 below:

Figure 5



- **PHASE I** Collate the data. Work with caBIG/NCICB to map our required data elements to the current structure of the caArray and secondary databases. Unification of Affymetrix and cDNA data.
- **PHASE II** Develop a simple, user-friendly website to import, export, and query our data from the database. Integrate bioinformatic tools into the website.
- **PHASE III** Public access to the database and all of our tools.

Phase I

Collaboration with caBIG/NCICB

Only high quality microarray data will be entered into the database. The top-level structure of the database will be a table where rows represent the genes, and columns the experiments. Each clone will have a unique number (plate position or Affymetrix identifier) which will be linked/mapped at a lower level with the NCBI (<http://www.ncbi.nlm.nih.gov/>) database. The gene/experiment cell contains numbers describing the expression level (relative or absolute) of the particular gene in the particular experiment, and a quality measurement. The original fluorescent images of the arrays, as well as the image analysis files which includes mean and total red and green intensities, background intensities and union area, will also be stored in the archive, and can be accessed if necessary. The annotation of each of the arrays will be done in collaboration with NCICB and will include a set of fields that are likely to be unique to pediatric cancers. For examples in neuroblastoma it is critical to include prognostic features such as age at presentation, presence of amplification of genes and Shimada histology.

We will use the caArray database (<http://caarray.nci.nih.gov/>) platform for the organization and storage of the data. caArray is a microarray database with open interfaces, and a user interface that is designed to be MIAME (minimum information about a published microarray based gene expression experiment) compliant. The MIAME is a set of required information for each of the microarray experiment that has been agreed by an international panel of experts and is the format for data submission for the publication involving microarray data : see http://www.mged.org/Workgroups/MIAME/miame_mage-om.html.

The new N-tier architecture of the caArray will allow seamless integration with other NCICB data sources: clinical data, animal models, genomic data, ontologies and controlled vocabularies, and the NCBI data databases. In phase 1 we will deploy the caArray database at our local site. We will use the Oracle database as the backend data storage system. We will have the technical help of the NCICB for this task. CaArray allows the exporting of data in XML format and facilitates data exchange between research centers throughout the world. We will also be able to import into our database other investigators data, which have been deposited in this format or in tabular format.

Unification of cDNA and Affymetrix Data.

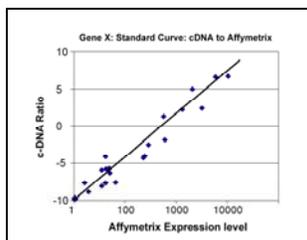


Figure 6

Of the 23 598 unique Unigenes on our cDNA microarrays, 16 384 (71%) overlap with the Affymetrix U133 set. We will perform microarray experiments on a set of 20 histologically distinct and varied samples (including our control RNA) on the Affymetrix platform. All samples will be hybridized in duplicate. The samples will be chosen which have the highest variance along 1st principal component, when genes are used as

variables. We will plot standard curves (see Figure 6 for an example for a theoretical gene x) for each of the unique Unigenes which will be subsequently used to transform the data across each platform. We will test the robustness of the standard curves using an independent set of hybridizations. Once successful we will be able to import other public data sets as well as have our data set be compatible with the Affymetrix data. We will work closely with the bioinformaticians at the NCBI on this project.

Phase II: Develop a simple, user-friendly website to import, export, and query our data from the database. Integrate bioinformatic tools into the website.

During this phase we make the data web accessible, initially internally to ensure easy operability. We will incorporate into our web site our in house MATLAB based analytic tools that can be implemented on a set of data or experiments chosen by the investigator. A non exhaustive set of tools include: 1) Fisher's weighted criterion, 2) t-test or F-test (depending on the number of classes), 3) Max Pairwise t-test, 4) TNoM (Threshold Number of Misclassifications), 5) Mutual Information, 6) Golub Weight, 7) Center Only, 8) Wilcoxon/Kruskall-Wallis (depending on the number of classes).

We will also include two machine learning algorithms; Artificial Neural Networks (ANN) (6) and Support Vector Machines (SVM) (15). All these tools are fully functional, but not web accessible and not readily available to the public. All the individual software will be made into executable files such that investigators may download and analyze the data these on their own personal computer, or use their own personal tools. All the experiments and genes will be fully annotated (by NCICB) and linked to the public databases including NCBI and EBI.

PHASE III Public access to the database and all of our tools.

Once working the database will be opened up to outside investigators, and publicly available. All the data will be exportable to the caBIG for further public distribution in an XML format. This format will allow other investigators to analyze the data using all available tools distributed by the caBIG.

Layman's Summary

All the genes are coded for in the DNA, like a computer code or sequence and the human genome project is the identifying of all of the human genes and their sequences. With the DNA microarray technology we are now able to measure the level (how much) a gene is expressed for 10 of thousands of genes human in a single experiment. We have measured the level of expression of more than 23 000 genes for several hundred pediatric tumors of different types with a primary focus on neuroblastoma. We have also determined the level of expression of the same genes in normal organs. We are attempting to identify which genes are expressed at high or low levels in particular cancers. These can be used to diagnose these cancers. We are also finding out which genes are differently expressed in cancers from patients who die from the disease compared to those that do not. Once these genes are identified we next want to find out if these genes contribute to the behavior of the cancer (biological effects of the genes) and whether if we disrupt the

Collaborators and Letters of Support

1. Dr. Ken Buetow, Director of NCI Center for Bioinformatics and Dr. Mervi Heiskanen, caArray project director, microarray informatics.
2. Dr. Kathy Pritchard-Jones, Chair Biological Studies Committee for the United Kingdom Childhood Cancer Study Group, UK.
3. Dr. Malcolm Smith, Assoc Branch Chief, Pediatrics, Cancer Therapy Evaluation Program, NCI.
4. Dr. Ching C Lau, Associate Professor Pediatrics, Baylor College of Medicine, Texas.

References

1. Khan, J., Bittner, M. L., Chen, Y., Meltzer, P. S., and Trent, J. M. DNA microarray technology: the anticipated impact on the study of human disease. *Biochim Biophys Acta*, *1423*: M17-28, 1999.
2. Khan, J., Saal, L. H., Bittner, M. L., Chen, Y., Trent, J. M., and Meltzer, P. S. Expression profiling in cancer using cDNA microarrays. *Electrophoresis*, *20*: 223-229, 1999.
3. Linet, M. S., Ries, L. A., Smith, M. A., Tarone, R. E., and Devesa, S. S. Cancer surveillance series: recent trends in childhood cancer incidence and mortality in the United States. *J Natl Cancer Inst*, *91*: 1051-1058., 1999.
4. Matthay, K. K., Villablanca, J. G., Seeger, R. C., Stram, D. O., Harris, R. E., Ramsay, N. K., Swift, P., Shimada, H., Black, C. T., Brodeur, G. M., Gerbing, R. B., and Reynolds, C. P. Treatment of high-risk neuroblastoma with intensive chemotherapy, radiotherapy, autologous bone marrow transplantation, and 13-cis-retinoic acid. Children's Cancer Group. *N Engl J Med*, *341*: 1165-1173., 1999.
5. Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, *270*: 467-470, 1995.
6. Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., and Meltzer, P. S. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med*, *7*: 673-679., 2001.
7. Khan, J., Simon, R., Bittner, M., Chen, Y., Leighton, S. B., Pohida, T., Smith, P. D., Jiang, Y., Gooden, G. C., Trent, J. M., and Meltzer, P. S. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res*, *58*: 5009-5013, 1998.
8. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, *286*: 531-537, 1999.
9. Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., Sampas, N., Dougherty, E., Wang, E., Marincola, F., Gooden, C., Lueders, J., Glatfelter, A., Pollock, P., Carpten, J.,

- Gillanders, E., Leja, D., Dietrich, K., Beaudry, C., Berens, M., Alberts, D., and Sondak, V. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, *406*: 536-540, 2000.
10. Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Jr., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Staudt, L. M., and et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, *403*: 503-511, 2000.
 11. Dickens, D. S., Kozielski, R., Khan, J., Forus, A., and Cripe, T. P. Cyclooxygenase-2 expression in pediatric sarcomas. *Pediatr Dev Pathol*, *5*: 356-364., 2002.
 12. Rice, A. M., Currier, M. A., Adams, L. C., Bharatan, N. S., Collins, M. H., Snyder, J. D., Khan, J., and Cripe, T. P. Ewing sarcoma family of tumors express adenovirus receptors and are susceptible to adenovirus-mediated oncolysis. *J Pediatr Hematol Oncol*, *24*: 527-533, 2002.
 13. Wang, E., Miller, L. D., Ohnmacht, G. A., Liu, E. T., and Marincola, F. M. High-fidelity mRNA amplification for gene profiling. *Nat Biotechnol*, *18*: 457-459. http://www.nature.com/nbt/journal/v418/n454/full/nbt0400_0457.html http://www.nature.com/nbt/journal/v0418/n0404/abs/nbt0400_0457.html, 2000.
 14. Eberwine, J., Yeh, H., Miyashiro, K., Cao, Y., Nair, S., Finnell, R., Zettel, M., and Coleman, P. Analysis of gene expression in single live neurons. *Proc Natl Acad Sci U S A*, *89*: 3010-3014., 1992.
 15. Williams, R. D., Hing, S., Greer, B. T., Whiteford, C. C., Wei, J. S., Natrajan, R., Kelsey, A., Rogers, S., Campbell, C., Pritchard- Jones, K., and Khan, J. Prognostic Classification of Relapsing Favourable Histology Wilms Tumour using cDNA Microarray Expression Profiling and Support Vector Machines. *Genes Chromosomes Cancer*, *In Press*, 2004.
 16. Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O. P., Wilfond, B., Borg, A., and Trent, J. Gene-expression profiles in hereditary breast cancer. *N Engl J Med*, *344*: 539-548., 2001.
 17. Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, *16*: 906-914., 2000.