

Project Creates Repository for Microarray Datasets

As the number of microarray studies increases and more cancer researchers use gene expression technology to search for cancer biomarkers or possible therapeutic targets, those who want to compare datasets face a daunting task: obtaining and compiling all this information from researchers scattered around the world. A new online tool called OncoPrint assembles these datasets in one easily searchable site to help researchers take advantage of the vast amount of data available.

Arul Chinnaiyan, M.D., who spearheaded the effort to create the new database, had realized there was a need for a central repository of microarray information when he was asked whether the genes that he found to be differentially expressed in prostate cancer, his field of study, were expressed in other cancer types.

"We wanted to compare data from different platforms and do meta-analyses," but there were no tools to do this, said Chinnaiyan, associate professor of pathology at the University of Michigan in Ann Arbor. He developed OncoPrint "so the average cancer biologist could take advantage of the wealth of information that's out there," he said.

With some pilot funds from his university, Chinnaiyan worked with colleagues at the Institute of Bioinformatics in India and Johns Hopkins University in Baltimore for a year to develop the site (<http://www.oncoPrint.org>). The site went online in late November 2003 with 65 datasets that included 4,702 microarrays and more than 47 million data points and covered 18 cancer types, and it has already gathered more than 1,000 users in 18 countries. Registration is free for academic and nonprofit users.

The information is amassed from publicly available datasets. Because there are no requirements

that microarray datasets be stored in a central location, Chinnaiyan's staff searches the literature for new studies and contacts these researchers for their data.

Users can search the results of a single microarray dataset, look for a gene's activity across multiple datasets, or search for multiple genes across datasets. (An animated online tutorial provides instructions.) Chinnaiyan's group has integrated data analysis with other resources, such as gene ontology annotations and a database of therapeutic targets. "One of the challenges of this has been to develop new data mining tools so people can take advantage of the data," Chinnaiyan said.

OncoPrint 2.0, currently in beta version but scheduled to be released in its final version within weeks, has more datasets (90, including more than 6,000 microarrays and nearly 71 million data points) and new features, such as pathway analysis.

Mark A. Rubin, M.D., associate professor of pathology at Harvard Medical School in Boston, an OncoPrint user, called the site "a convenient resource" that "offers a way of [study comparison] that couldn't be done in the past."

There are numerous programs available to comb through data in any one laboratory, but to analyze data from across laboratories, a researcher must contact each laboratory separately or

download the data from each laboratory's Web site. It is not impossible, Rubin said, but it is inconvenient.

In addition, even after researchers receive all of the datasets, they often have problems comparing them because of the different platforms that the various groups use to collect their data. OncoPrint normalizes the data. "We try to keep everything consistent within each dataset and then make qualitative comparisons across datasets," said Chinnaiyan.

For example, if a researcher finds that gene X is upregulated in prostate cancer, he can then go into OncoPrint and look at the status of gene X in studies of other cancers.

One of the values that Rubin has found in OncoPrint is the ability to do virtual validation studies. In one of the first studies to use the site, published in June in *Cancer Research*, Rubin and his colleagues looked at the status of the gene TPD52 in prostate cancer. With OncoPrint, they were able to confirm previous work that found that TPD52 was overexpressed in breast cancer as well as find that the gene was overexpressed in a number of other tumor types.

OncoPrint can also act as a "discovery engine," by generating discoveries or hypotheses that then need to be followed up with other research, Chinnaiyan said. In the first of these discoveries, published in the June 22 issue of the *Proceedings of the National Academy of Sciences*, Chinnaiyan and his colleagues reported finding 67 genes that were universally activated in most types of cancers. Another set of 69 genes was commonly activated only in aggressive undifferentiated cancers, the type that often results in poorer patient outcomes.

Chinnaiyan's group plans to continue adding datasets to OncoPrint in addition to developing new data mining tools to do gene correlation and anti-correlation studies. They also hope to be able to take a broader approach to the data in the future and look at the activation and repression of entire pathways within cancers.

—Sarah L. Zielinski

The screenshot shows the OncoPrint website interface. At the top, it says "ONCOPRINT cancer microarray database". Below that, there's a navigation menu with "Home", "About us", "Genes", "Study", "Meta", "Submit", and "Contact Us". A search bar contains "ERBB2 : Hs.323910". Below the search bar, there's a table of search results. The table has columns for "Study Name", "Class 1 (C1)", "Class 2 (C2)", "Report", "R1", "R2", "R3", "R4", "R5", "R6", "R7", "R8", "R9", "R10", "R11", "R12", "R13", "R14", "R15", "R16", "R17", "R18", "R19", "R20". The table lists various studies and their corresponding gene expression data for ERBB2.

Study Name	Class 1 (C1)	Class 2 (C2)	Report	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16	R17	R18	R19	R20
Alizadeh_Lymphoma	Blood B-cells	Diffuse Large B-cell Lymphoma	Details	0.521	0.139	1.136	0.269	>1	View														
Alizadeh_Lymphoma	Blood B-cells	Follicular Lymphoma	Details	-0.204	-0.326	0.802	0.3951	>1	View														
Beer_Lung	Normal Lung	Lung Adenocarcinoma	Details	0.467	0.677	-4.691	1.76-5	0.801	View														
Bhattacharjee_Lung	Normal Lung	Lung Adenocarcinoma	Details	-0.08	0.587	-4.816	0.0001	0.0075	View														
Bhattacharjee_Lung	Normal Lung	Lung Carcinoid	Details	-0.08	-0.526	2.814	0.009	0.5303	View														
Bhattacharjee_Lung	Normal Lung	Small Cell Lung Carcinoma	Details	-0.08	-0.562	1.548	0.1499	>1	View														
Bhattacharjee_Lung	Normal Lung	Squamous Cell Lung Carcinoma	Details	-0.08	-0.37	1.628	0.1131	>1	View														
Chen_Liver	Normal Liver	Hepatocellular Carcinoma	Details	0.295	0.119	2.224	0.0275	>1	View														
Dhanasekaran_Prostate	BPH and Normal Prostate	Prostate Cancer	Details	-1.138	-0.921	-0.923	0.3627	>1	View														
Dyrskot_Bladder	Normal Bladder	Bladder Cancer	Details	0.574	0.876	-1.062	0.3568	>1	View														
Frisson_Salivary	Normal Salivary Gland	Salivary Adenoid Cystic Carcinoma	Details	-0.148	0.008	-3.302	0.0144	0.8474	View														
Garber_Lung	Normal Lung	Lung Adenocarcinoma	Details	0.383	1.251	-3.733	0.0030	0.2292	View														
Garber_Lung	Normal Lung	Squamous Cell Lung Carcinoma	Details	0.389	-0.078	1.517	0.1487	>1	View														
Luoh_Prostate	Normal Prostate	Prostate Cancer	Details	0.391	0.267	1.17	0.2523	>1	View														

OncoPrint users can search for one or more genes across multiple datasets. Results can be viewed in a variety of formats, such as tables, graphs, or gene expression heatmaps.