D578–D581 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh003

The PEPR GeneChip data warehouse, and implementation of a dynamic time series query tool (SGQT) with graphical interface

Josephine Chen, Po Zhao, Donald Massaro[1], Linda B. Clerch[2], Richard R. Almon[3,4], Debra C. DuBois[3,4], William J. Jusko[4] and Eric P. Hoffman*

Center for Genetic Medicine, Children's National Medical Center, 111 Michigan Avenue NW, Washington, DC 20010, USA, [1]Department of Medicine, [2]Department of Pediatrics, Georgetown University School of Medicine, Washington, DC, USA, [3]Department of Biological Sciences and [4]Department of Pharmaceutical Sciences, SUNY Buffalo, Buffalo, NY, USA

Received June 29, 2003; Revised and Accepted August 5, 2003

ABSTRACT

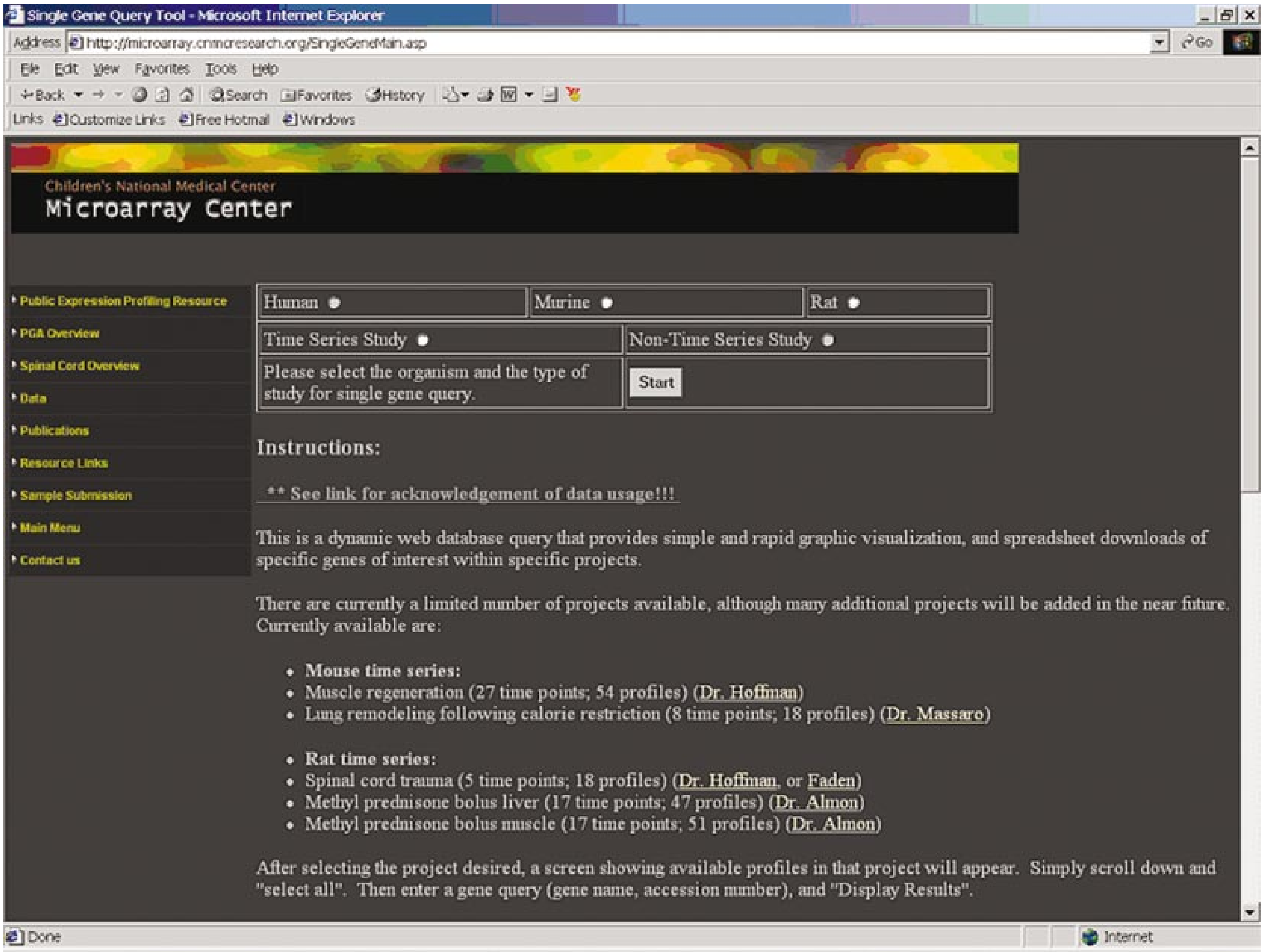
Publicly accessible DNA databases (genome browsers) are rapidly accelerating post-genomic research (see http://www.genome.ucsc.edu/), with integrated genomic DNA, gene structure, EST/splicing and cross-species ortholog data. DNA databases have relatively low dimensionality; the genome is a linear code that anchors all associated data. In contrast, RNA expression and protein databases need to be able to handle very high dimensional data, with time, tissue, cell type and genes, as interrelated variables. The high dimensionality of microarray expression profile data, and the lack of a standard experimental platform have complicated the development of web-accessible databases and analytical tools. We have designed and implemented a public resource of expression profile data containing 1024 human, mouse and rat Affymetrix GeneChip expression profiles, generated in the same laboratory, and subject to the same quality and procedural controls (Public Expression Profiling Resource; PEPR). Our Oracle-based PEPR data warehouse includes a novel time series query analysis tool (SGQT), enabling dynamic generation of graphs and spreadsheets showing the action of any transcript of interest over time. In this report, we demonstrate the utility of this tool using a 27 time point, *in vivo* muscle regeneration series. This data warehouse and associated analysis tools provides access to multidimensional microarray data through web-based interfaces, both for download of all types of raw data for independent analysis, and also for straightforward gene-based queries. Planned implementations of PEPR will include web-based remote entry of projects adhering to quality control and standard operating procedure (QC/SOP) criteria, and automated output of alternative probe set algorithms for each project (see http://microarray.cnmcresearch.org/pgadatatable.asp).

INTRODUCTION AND DATABASE DESCRIPTION

PEPR provides centralized Affymetrix expression profiling data to the public research community, typically before publication in primary research papers. Data released through PEPR are generated within a single centralized research group (Children's National Medical Center, Microarray Center), with projects originating internally and referred from external institutions. Currently, 1024 Affymetrix arrays representing 38 projects (13 human; 25 mouse/rat) are released to the public. PEPR is an Oracle-based web solution, which permits researchers seamless access to an Affymetrix-only expression profiling database through our web browser without requiring their own Affymetrix software. The web interface also enables users to export many forms of data associated with any particular profile, including raw image files (.dat), processed image files (.cel) and interpretation files (.txt). It allows researchers to perform on-line queries of expression profiles by any number of experimental variables (tissue, species, chip type, etc.). Other built-in functions include searching by GenBank Accession ID and gene name (gene-based cross-profile search). These search functions return signal (Avg Diff) values and Present/Absent Calls (MAS5) for all profiles in PEPR. We also designed and implemented an automated back-end process that disseminates all available PEPR profile data into NCBI Gene Expression Omnibus (GEO) database (http://www.ncbi.nih.gov/geo/) (1). Public users can easily access deposited data in GEO as well as original data files in the PEPR database through a corresponding link created during the direct deposit process.

To our knowledge, the PEPR data warehouse is the largest such public resource adhering to quality control and standard operating procedures (QC/SOP). However, we recognized that the utility of PEPR is dependent on some familiarity with bioinformatics aspects of microarray experiments, where files could be downloaded and analyzed with any method desired.

*To whom correspondence should be addressed. Tel: +1 202 884 6011; Fax: +1 202 884 6014; Email: ehoffman@cnmcresearch.org


Single Gene Query Tool - Microsoft Internet Explorer
Address http://microarray.cnmcresearch.org/SingleGeneMain.asp
File Edit View Favorites Tools Help
Back Search Favorites History
Links Customize Links Free Hotmail Windows
Children's National Medical Center
Microarray Center
Public Expression Profiling Resource
PGA Overview
Spinal Cord Overview
Data
Publications
Resource Links
Sample Submission
Main Menu
Contact us
Human ○
Murine ○
Rat ○
Time Series Study ○
Non-Time Series Study ○
Start
Done
Internet

Please select the organism and the type of study for single gene query.

Instructions:

** See link for acknowledgement of data usage!!!

This is a dynamic web database query that provides simple and rapid graphic visualization, and spreadsheet downloads of specific genes of interest within specific projects.

There are currently a limited number of projects available, although many additional projects will be added in the near future. Currently available are:

- **Mouse time series:**
- Muscle regeneration (27 time points; 54 profiles) (Dr. Hoffman)
- Lung remodeling following calorie restriction (8 time points; 18 profiles) (Dr. Massaro)

- **Rat time series:**
- Spinal cord trauma (5 time points; 18 profiles) (Dr. Hoffman, or Faden)
- Methyl prednisone bolus liver (17 time points; 47 profiles) (Dr. Almon)
- Methyl prednisone bolus muscle (17 time points; 51 profiles) (Dr. Almon)

After selecting the project desired, a screen showing available profiles in that project will appear. Simply scroll down and "select all". Then enter a gene query (gene name, accession number), and "Display Results".

Figure 1. Initial database query for the time series query tool.

To begin to build true user-friendly web-based data analysis tools that do not require experience in formatting and interpretation of microarray data, we designed and implemented a Single Gene Query Tool (SGQT) (see http://microarray.cnmcresearch.org/singlegenemain.asp).

SINGLE GENE QUERY TOOL (SGQT)

Our initial implementation of SGQT is for time series data, which we present here. We provide an entry screen that defines the data subset selections that are available for the user to search (Fig. 1). The specific projects available fitting the search criteria are then presented, and selection of one project leads to a list of all profiles associated with the project. In the example we describe here, a 54 profile, 27 time point muscle regeneration series was selected, with two different muscles profiled at each time point on U74A microarrays containing ~12 000 probe sets (2,3). The user is asked to select the profiles to be studied ('select all' is the option used here to query all 54 profiles). A web browser-style search query is then evoked, and entry of any text or probe set then queries genome databases for all genes and probe sets matching the query. For example, entry of 'myosin' will identify myosin heavy chains, light chains, binding proteins, etc. The user then selects the desired gene from the pull down result menu. Query of 'myogenin' returns only a single probe set, which, when selected ('submit') then triggers the database query tool. The tool then dynamically extracts data from the .cel files for the myogenin probe set from the 54 profile (12 000 probe sets/profile) data set, including signal (normalized hybridization intensity), and absent/present calls (Affymetrix MAS 5.0 determinations). The tool then aligns all data into a time series, and graphs replicates for each time point (Fig. 2), as well as calculating the average of the replicates, graphing the average, and drawing a graph line through the averages for all time points (Fig. 2). The tool also calculates the average signal for each time point, and the fold-change relative to time 0 (based upon array-normalized intensities) (Fig. 2).

The resulting on-line graph has mouse-overs containing data associated with each data point (time point, signal, present/absent call), and for the arithmetic average (time point, average signal, fold-change relative to time 0) (Fig. 2). The mouse-over shown in Figure 2 is for the arithmetic average of replicates, with the pop-up window indicating the fold-change from time 0. Clicking over any data point links to a series of databases (Unigene, GenBank, LocusLink,

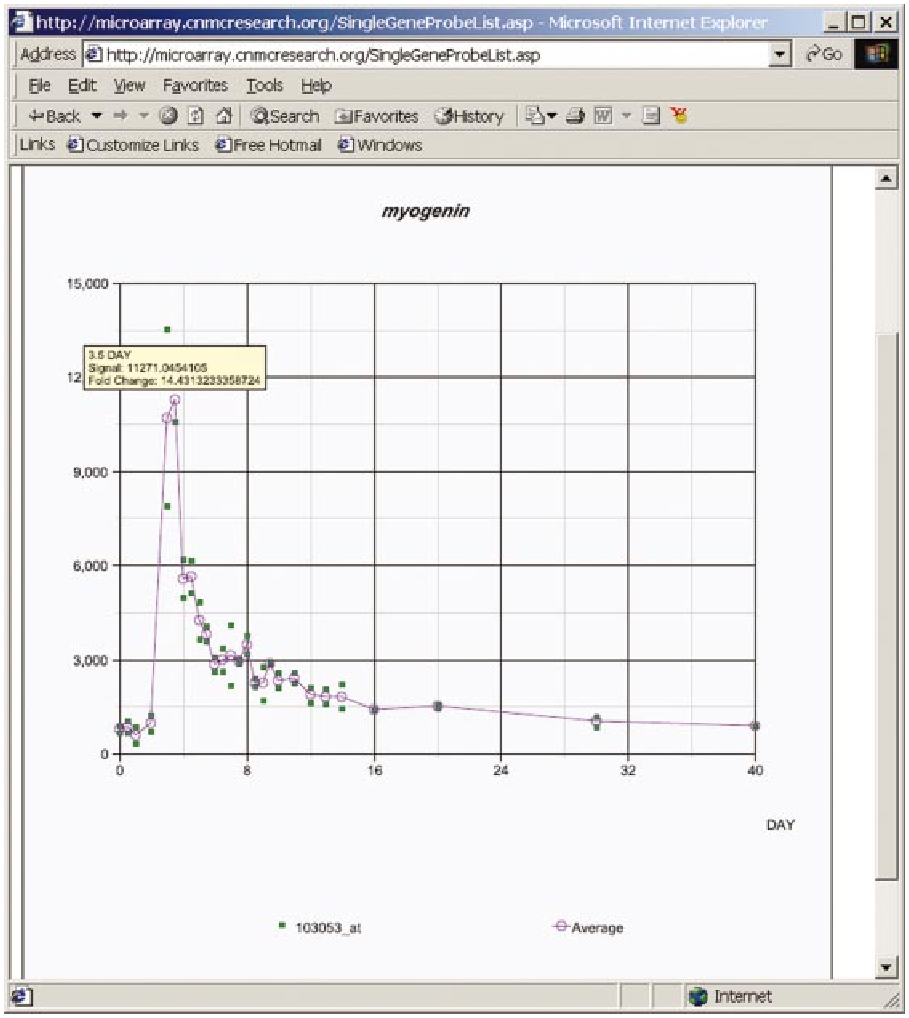

http://microarray.cnmcresearch.org/SingleGeneProbeList.asp - Microsoft Internet Explorer
Address http://microarray.cnmcresearch.org/SingleGeneProbeList.asp
Go
File Edit View Favorites Tools Help
Back
Search Favorites History
Links Customize Links Free Hotmail Windows
myogenin
15,000
12
9,000
6,000
3,000
0
0
8
16
24
32
40
DAY
3.5 DAY
Signal: 11271.0454105
Fold Change: 14.4313233358724
103053_at
Average
Internet

Figure 2. Graphic output of the time series query for myogenin in muscle regeneration. Muscle degeneration/regeneration was induced with intramuscular injection of cardiotoxin, and two different muscles profiled at the indicated time points following injection [see (2) for detailed methods]. Shown is the dynamic database query output of myogenin probe set data contained in 54 U74A Affymetrix microarrays for 27 time points (0–40 days). The green data points are individual expression profiles, with two different muscles profiled and graphed per time point. The purple circle is the average of the replicates, with the graph drawn between the average at each time point. The *y* axis is the relative expression level ('signal') using Affymetrix MAS 5.0. The mouse-over shown corresponds to the average at 3.5 days, and provides both average signal and fold-change relative to time 0 (14.4-fold increase in expression).

Affymetrix) containing information on the gene of interest, as well as access to the download for the original data set (.cel, .dat, or .txt files). The tool also writes a dynamically generated spreadsheet containing all the information in the graph and this appears as a link above the graph. This spreadsheet can be downloaded, and analyzed using any desired graphics or statistical package. It should be emphasized that all visualizations and spreadsheets are dynamic queries of the web Oracle database. The dynamic search and output of the 54 profile murine regeneration series shown here is typically completed in approximately 15 s.

The five time series currently implemented for the tool are a murine *in vivo* 27 time point muscle regeneration series (54 U74A profiles) (2,3), an 8 time point murine lung calorie restriction time series (18 U74A profiles) (4) (D.Massaro and L.B.Clerch, unpublished data), a 5 time point rat spinal cord damage series (18 U34A profiles) (5), and two 17 time point methylprednisone bolus time series in rat (47 profiles in liver and 51 profiles in muscle) (6,7). It is important to note that many experimental variables, such as diurnal variation in gene expression, should be considered when interpreting time series data; for example, in the Massaro and Clerch calorie restriction studies, non-restricted and calorie-restricted mice were killed at the same time. We will continue to add additional time series to the tool, and plan to implement a collection of time series and non-time series data comparisons and visualizations to the PEPR resource.

To our knowledge, the time series query tool described here is the first expression profile data analysis tool that requires no prior knowledge of microarray data format or data interpretation. This tool is useful due to the quality control and replicates available for each time point, and simple visualization, interpretation and download of these. Future implementations of our data warehouse will allow input of externally generated data that conform to minimum experimental design criteria, and our QC/SOP benchmarks (see http://microarray.cnmcresearch.org/pgaoutline-qcofsamples.asp) via a web interface with automated QC/SOP checks. As PEPR is built upon a standardized platform of Affymetrix-only data adhering to QC/SOP, all internally- and externally-generated data within PEPR should be intrinsically comparable. A new implementation of PEPR including many projects able to be queried by the SGQT tool described here is expected in late 2003. The updated PEPR will also include a choice of probe set algorithm for data display (MAS 5.0, dCHIP, RMA and ProbeProfiler).

MATERIALS AND METHODS

Expression profiling

All expression profiles were generated using total RNA, with *in vitro* transcription yielding biotinylated cRNA for hybridization to Affymetrix GeneChips (see http://microarray.cnmcresearch.org/pgaoutline-qcofsamples.asp). Only one of the 38 projects utilized two-round amplifications from limiting sample (8), and this is clearly indicated in the mouse-over for that project (see http://microarray.cnmcresearch.org/pgadatatable.asp).

Data analysis

We provide .dat, .cel, and .txt interpretation files using Affymetrix MAS 5.0 for all microarrays and projects. Other methods of normalization and probe set interpretation can be used by downloading any desired file types. The single gene query tool uses raw .cel file data, normalized via a common target intensity between all profiles in the project, and provides information on 'present/absent' call determinations, but does not use these for data analysis purposes. We have recently shown that the Affymetrix MAS 5.0 probe set interpretation method provides good signal/noise ratios for expression profiling projects using tissue samples (9).

ACKNOWLEDGEMENTS

We thank John Quakenbush for valuable comments on the manuscript. Supported by grants from the NIH (NHLBI U01 HL66614-01 'Programs in Genomic Applications' HOPGENE; NINDS N01-NS-1-2339 (spinal cord); NINDS 3R01 NS29525-09 (muscular dystrophy), the Muscular Dystrophy Association USA (muscle regeneration); NHLBI GM24211 (methylprednisone series); NHLBI HL20366 and HL27413 (alveolus formation).

REFERENCES

1. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
2. Zhao,P., Iezzi,S., Sartorelli,V., Dressman,D. and Hoffman,E.P. (2002) Slug is downstream of myoD: Identification of novel pathway members via temporal expression profiling. *J. Biol. Chem.*, **277**, 20091–20101.
3. Zhao,P., Seo,J., Wang,Z., Wang,Y., Shneiderman,B. and Hoffman,E.P. *In vivo* filtering of *in vitro* MyoD target data: An approach for identification of biologically relevant novel downstream targets of transcription factors. *C. R. Biol.*, in press.
4. Hoffman,E.P., Massaro,D., Massaro,G. and Clerch,L. Expression profiling as a tool for diagnosis and pathway discovery: Experimental design and technical considerations. In Lenfant,C. and Massaro,D. (eds), *Lung Remodeling*. Marcel Dekker, New York, NY, in press.
5. Di Giovanni,S., Knoblach,S.M., Brandoli,C., Aden,S.A., Hoffman,E.P. and Faden,A.I. (2003) Gene profiling in spinal cord injury shows role of cell cycle in neuronal death. *Ann. Neurol.*, **53**, 454–468.
6. Almon,R.R., DuBois,D.C., Pearson,K.E., Stephan,D.A. and Jusko,W.J. (2003) Gene arrays and temporal patterns of drug response: Corticosteroid effects on liver functional and integrative genomics. *Funct. Integr. Genomics*, August 20, [Epub ahead of print].
7. Jin,J.Y., Almon,R.R., DuBois,D.C. and Jusko,W.J. (2003) Modeling of corticosteroid pharmacogenomics in rat liver using gene microarrays. *J. Pharmacol. Exp. Ther.*, **307**, 93–109.
8. Hittel,D.S., Kraus,W.E. and Hoffman,E.P. (2003) Skeletal muscle dictates the fibrinolytic state after exercise training in overweight men with characteristics of metabolic syndrome. *J. Physiol. (Lond.)*, **548**, 401–410.
9. Seo,J., Bakay,M., Zhao,P., Chen,Y.W., Clarkson,P., Shneiderman,B. and Hoffman,E.P. (2003) Interactive color mosaic and dendrogram displays for signal/noise optimization in microarray data analysis. *IEEE ICME*, **III**, 461–465.