CGI: Java Software for Mapping and Visualizing Data from Array-based Comparative Genomic Hybridization and Expression Profiling

Joyce Xiuweu-Xu Gu[1], Michael Yang Wei[1], Pulivarthi H. Rao[1], Ching C. Lau[1], Sanjiv Behl[2] and Tsz-Kwong Man[1]

[1]Department of Pediatrics, Baylor College of Medicine and Texas Children's Cancer Center, Houston, Texas. [2]Department of Computer Science, University of Houston, Victoria, Texas.

Abstract: With the increasing application of various genomic technologies in biomedical research, there is a need to integrate these data to correlate candidate genes/regions that are identified by different genomic platforms. Although there are tools that can analyze data from individual platforms, essential software for integration of genomic data is still lacking. Here, we present a novel Java-based program called CGI (Cytogenetics-Genomics Integrator) that matches the BAC clones from array-based comparative genomic hybridization (aCGH) to genes from RNA expression profiling datasets. The matching is computed via a fast, backend MySQL database containing UCSC Genome Browser annotations. This program also provides an easy-to-use graphical user interface for visualizing and summarizing the correlation of DNA copy number changes and RNA expression patterns from a set of experiments. In addition, CGI uses a Java applet to display the copy number values of a specific BAC clone in aCGH experiments side by side with the expression levels of genes that are mapped back to that BAC clone from the microarray experiments. The CGI program is built on top of extensible, reusable graphic components specifically designed for biologists. It is cross-platform compatible and the source code is freely available under the General Public License.

Keywords: aCGH, expression profiling, visualization, correlation, and data integration

Introduction

With the advent of genomic technologies, DNA and RNA-based microarrays are becoming more accessible to biomedical researchers. One of the common DNA platforms is array-based Comparative Genomic Hybridization (aCGH), which can identify DNA copy number aberrations in the genome (Pinkel, 1998; Man, 2004). There are many software tools that have been developed to analyze aCGH data (Jong 2004; Margolin, 2005; Chen, 2005; Cheung, 2005; Price, 2005; Kim, 2005) and expression microarray data (Sykacek, 2005; Shamir, 2005; Saraiya, 2005; Li, 2001; Vaquerizas, 2005; Bumm, 2002; Saeed, 2003); however, no tool is currently available for the biologist to integrate these two types of data. One of the main challenges is that once the significant BAC clones or genes are identified, it is very difficult to correlate the DNA copy number and RNA expression results. This is because the significant genes may not lie within the corresponding BAC clones even though they are located in the same chromosomal region. Therefore, a more precise method of matching is needed in order to properly correlate these two types of data.

A typical way to perform the matching is to manually search the UCSC Genome Browser (http://genome.ucsc.edu/) to make sure the significant genes lie within the significant BAC clones. However, this type of manual search is very laborious and error prone if the numbers of BAC clones and genes are large. Thus, it is important to develop a user friendly and flexible tool that can match, correlate and display the aCGH and expression profiling data. Since it is common to identify hundreds to thousands of significant genes by either expression profiling or aCGH experiments, our program can further assist researchers to select genes that are found to be significant by both types of experiments, or genes that may not be identified by using either type of technique alone.

To address this issue, we developed a Java-based, stand-alone program that uses MySQL database (http://www.mysql.com) as a backend to store the BAC clones and gene information downloaded from UCSC database. This information is used to match the user-provided BAC clones in aCGH experiments

Correspondence: Tsz-Kwong Man, Texas Children's Cancer Center, 6621 Fannin St. MC 3-3320, Houston, TX 77030-2399,Tel: (832) 824-4682; Email: ctman@txccc.org

and genes in expression profiling experiments. After that, the correlation coefficients and p-values of the matched BAC clone-gene pairs will then be computed and displayed in various formats for data visualization and comparison.

Software Designs

The CGI software is based on an object-oriented framework designed to conduct searches for features/genes in RNA expression-profiling experiments that mapped back to corresponding BAC clones in aCGH experiments. The program combines bioinformatic data matching from databases with simple correlation analysis. The software is organized into three functional modules (Data, Annotation, and Correlation). The Data module contains DNA copy number and RNA expression data and links them with the Annotation module by interacting with the MySQL database that holds a variety of different types of genomic information including chromosomal localization, Unigene ID, and gene annotation data. Information in the database is used to match the BAC clones and the genes provided by the users. The Correlation module calculates the Pearson correlation coefficients and p-values between the DNA copy numbers and expression values of matched BAC clone-gene pairs in different experiments. It also displays DNA copy numbers of a specific BAC clone in different aCGH experiments and the associated gene expression values in microarray experiments for easy data visualization and comparison.

Data Importing

A simple graphical user interface (GUI) prompts users to enter user name, password, database name, and the locations of the aCGH and RNA expression-profiling files (Fig. 1). The aCGH file contains FISH-mapped BAC Clone IDs, cytobands, and normalized log ratios representing DNA copy numbers from aCGH experiments. The RNA expression-profiling file contains Unigene IDs, gene symbols, and log-ratios (dual channel arrays) or log intensities (Affymetrix or oligo-based arrays) of gene expressions in a set of experiments involving identical cases as in the aCGH experiments.

Data Querying and Mining

The program offers two ways to query the data. First, BAC Clone IDs in an aCGH input file are used to query the MySQL database, which stores data downloaded from the UCSC database at

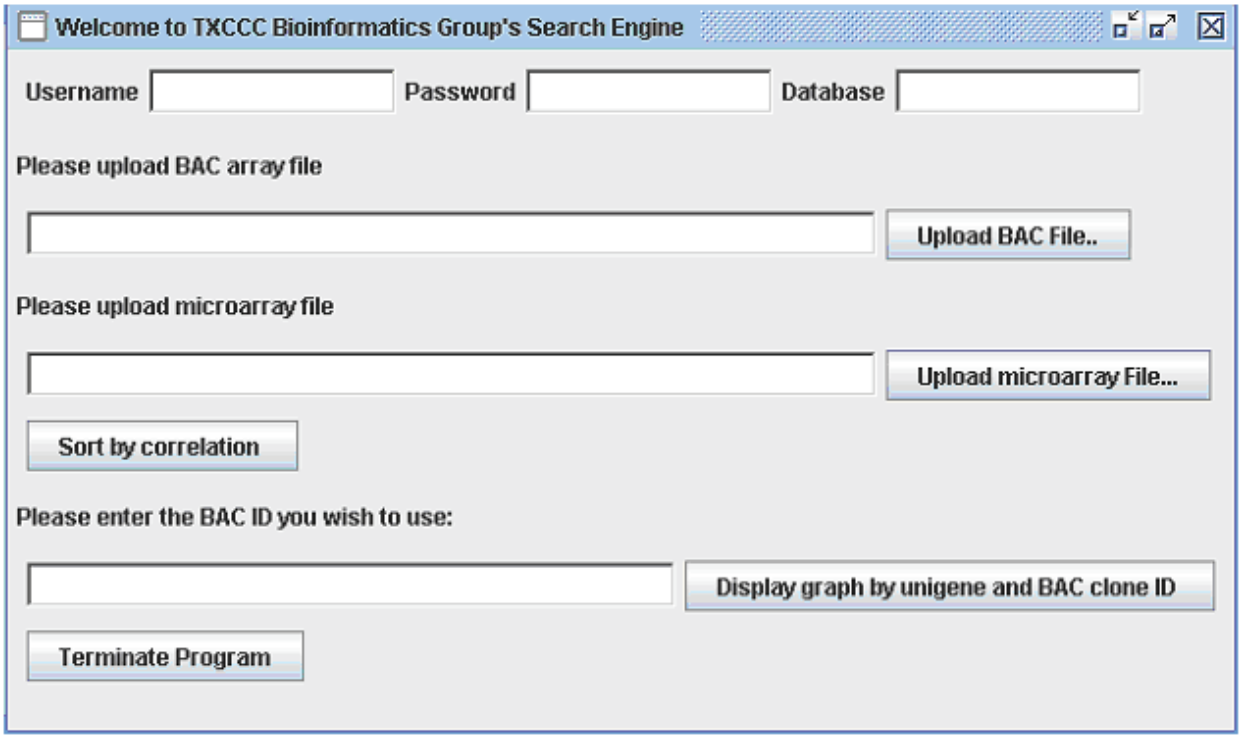
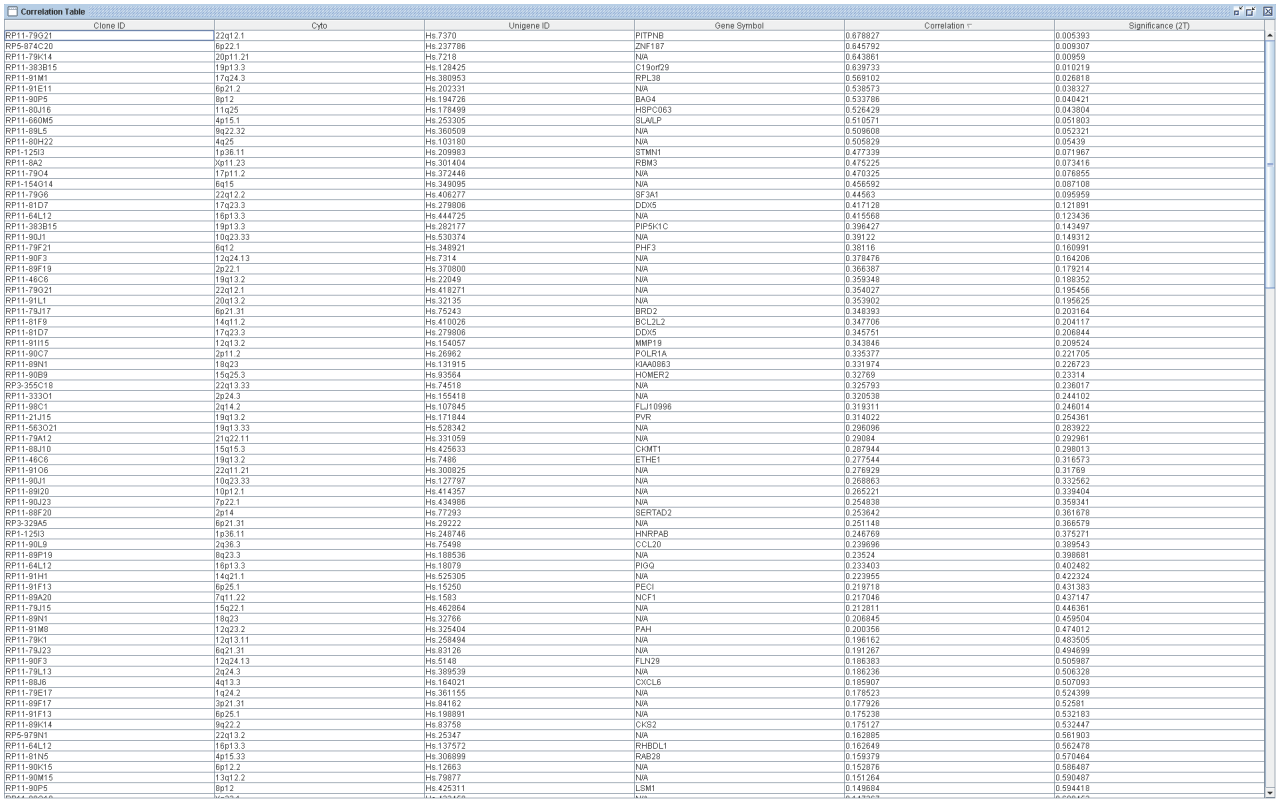


Figure 1. Graphic user interface of CGI for data importing and analysis.

URL:http://genome.ucsc.edu/cgi-bin/hgTables—fishClone and uniGene_2 tables. The two tables are first downloaded by the user and imported to the MySQL databases as described in the installation manual (see supplemental information). The Unigene IDs of the genes that reside in each BAC clone in the aCGH input file are retrieved based on chromosome number and their physical locations by SQL commands. Secondly, these Unigene IDs are used to match with the features/genes provided in a RNA expression-profiling input file, so that the matched BAC clone-gene pairs will be identified. The DNA copy numbers and gene expression values of the matched BAC clone-gene pairs will then be extracted from the input files and their Pearson correlation coefficients and p-values are computed by an internal correlation functions. Finally, the correlation coefficients and p-values of the BAC clone-gene pairs will be tabulated together with their BAC Clone IDs, cytobands, Unigene IDs, and gene symbols provided by the input files. If there are multiple genes within a BAC clone, the program will replicate the DNA copy number data of that BAC clone and correlate with the expression data of each of the other genes that are mapped to the BAC clone.

Data Visualization

CGI uses a correlation table to display a global overview of BAC Clone ID, Cytoband, their corresponding Unigene IDs, gene symbols and Pearson correlation coefficients and p-values of the matched BAC clones and features/genes. The table view is very flexible and the data in the table can be sorted dynamically in an ascending or descending order based on the correlation coefficients, BAC Clone ID, cytoband location, Unigene ID, etc (Fig. 2). It can also change the order of the columns to display different views according to user's preference. Besides the table view, users can also visualize in detail the DNA copy number of a specific BAC Clone and the expression of its associated genes by entering the BAC Clone ID into the text box provided in the GUI (Fig. 1). The CGI program will display two graphic windows if the input BAC Clone ID matches one or more Clone IDs in the correlation table. One window displays three line graphs representing the DNA copy number changes of the queried BAC Clone in aCGH experiments and the expression values of its associated genes in RNA expression-profiling experiments (Fig. 3). The second window displays the DNA copy number data and RNA expression data



Correlation Table

Clone ID	Cyto	Unigene ID	Gene Symbol	Correlation	Significance (2T)
RP11-79Q21	22q12.1	Hs.7370	PITPNB	0.678827	0.005393
RP5-874C20	6p22.1	Hs.237786	ZNF187	0.645792	0.009307
RP11-79K14	20p11.21	Hs.7218	N/A	0.643861	0.00959
RP11-383B15	19p13.3	Hs.128425	C19orf29	0.639733	0.010219
RP11-91M1	17q24.3	Hs.380953	RPL38	0.569102	0.026818
RP11-91E11	6p21.2	Hs.202331	N/A	0.538573	0.038327
RP11-90P5	8p12	Hs.194726	BAG4	0.533786	0.040421
RP11-80J16	11q25	Hs.178499	HSPC063	0.526429	0.043804
RP11-660M5	4p15.1	Hs.253305	SLAILP	0.510571	0.051803
RP11-89L5	9q22.32	Hs.360509	N/A	0.509608	0.052321
RP11-80H22	4q25	Hs.103180	N/A	0.505829	0.05439
RP1-125I3	1p36.11	Hs.209983	STMN1	0.477339	0.071967
RP11-8A2	Xp11.23	Hs.301404	RBM3	0.476225	0.073416
RP11-79O4	17p11.2	Hs.372446	N/A	0.470325	0.076955
RP11-154O14	8q15	Hs.348095	N/A	0.458592	0.087108
RP11-79G6	22q12.2	Hs.406277	SF3A1	0.44563	0.095959
RP11-81D7	17q23.3	Hs.279806	DDX5	0.417128	0.121891
RP11-64L12	16p13.3	Hs.444725	N/A	0.415568	0.123436
RP11-383B15	19p13.3	Hs.292177	PIP5K1C	0.398427	0.143497
RP11-90J1	10q23.33	Hs.530374	N/A	0.39122	0.149312
RP11-79F21	6q12	Hs.348921	PHF3	0.38116	0.160991
RP11-90F3	12q24.13	Hs.7314	N/A	0.378476	0.164206
RP11-89F19	2p22.1	Hs.370800	N/A	0.368387	0.179214
RP11-48C6	19q13.2	Hs.22049	N/A	0.359348	0.188352
RP11-79O21	22q12.1	Hs.418271	N/A	0.354027	0.195456
RP11-91L1	20q13.2	Hs.32135	N/A	0.353902	0.195625
RP11-79J17	6p21.31	Hs.75243	BRD2	0.348393	0.203164
RP11-81F9	14q11.2	Hs.410026	BCL2L2	0.347706	0.204117
RP11-81D7	17q23.3	Hs.279806	DDX5	0.345751	0.206844
RP11-91I15	12q13.2	Hs.154057	MMP19	0.343846	0.209524
RP11-90C7	2p11.2	Hs.26962	POLR1A	0.335377	0.221705
RP11-89N1	18q23	Hs.131915	KIAA0863	0.331974	0.226723
RP11-90B9	15q25.3	Hs.93564	HOMER2	0.32769	0.23314
RP3-355C18	22q13.33	Hs.74518	N/A	0.325793	0.236017
RP11-333O1	2p24.3	Hs.155418	N/A	0.320538	0.244102
RP11-98C1	2q14.2	Hs.107845	FLJ10996	0.319311	0.246014
RP11-21J15	19q13.2	Hs.171844	PVR	0.314022	0.254361
RP11-563O21	19q13.33	Hs.528342	N/A	0.296096	0.283922
RP11-79A12	21q22.11	Hs.331059	N/A	0.29084	0.292961
RP11-88J10	15q15.3	Hs.425633	CHMT1	0.287944	0.298013
RP11-46C6	19q13.2	Hs.7486	ETHE1	0.277544	0.316573
RP11-91O6	22q11.21	Hs.300825	N/A	0.276929	0.31769
RP11-90J1	10q23.33	Hs.127797	N/A	0.268863	0.332562
RP11-89I20	10p12.1	Hs.414357	N/A	0.268221	0.339404
RP11-90J23	7p22.1	Hs.434986	N/A	0.254838	0.359341
RP11-88F20	2p14	Hs.77293	SERTAD2	0.253642	0.361678
RP3-329A5	6p21.31	Hs.29222	N/A	0.251148	0.366578
RP1-125I3	1p36.11	Hs.248746	HNRPAB	0.248769	0.375271
RP11-90L9	2q36.3	Hs.75498	CCL20	0.239696	0.389543
RP11-89P19	8q23.3	Hs.188536	N/A	0.23524	0.398681
RP11-64L12	16p13.3	Hs.18079	PIGQ	0.233403	0.402482
RP11-91H1	14q21.1	Hs.525305	N/A	0.223955	0.422324
RP11-91F13	6p25.1	Hs.15250	PECI	0.219718	0.431383
RP11-89A20	7q11.22	Hs.1583	NCF1	0.217046	0.437147
RP11-79J15	15q22.1	Hs.462864	N/A	0.212911	0.446361
RP11-89N1	18q23	Hs.32766	N/A	0.206845	0.459504
RP11-91M8	12q23.2	Hs.325404	PAH	0.200356	0.474012
RP11-79K1	12q13.11	Hs.258494	N/A	0.196162	0.483505
RP11-79J23	6q21.31	Hs.83126	N/A	0.191267	0.494699
RP11-90F3	12q24.13	Hs.5148	FLN29	0.186383	0.505987
RP11-79L13	2q24.3	Hs.389539	N/A	0.186236	0.506328
RP11-88J6	4q13.3	Hs.164021	CXCL6	0.185907	0.507093
RP11-79E17	1q24.2	Hs.361155	N/A	0.178523	0.524399
RP11-89F17	3p21.31	Hs.84182	N/A	0.177926	0.52581
RP11-91F13	6p25.1	Hs.198891	N/A	0.175238	0.532183
RP11-89K14	9q22.2	Hs.83758	CKS2	0.175127	0.532447
RP5-979N1	22q13.2	Hs.25347	N/A	0.162805	0.561903
RP11-64L12	16p13.3	Hs.137572	RHBDL1	0.162649	0.562478
RP11-81N5	4p15.33	Hs.306999	RAB28	0.159379	0.570464
RP11-90K15	6p12.2	Hs.12663	N/A	0.152878	0.586487
RP11-90M15	13q12.2	Hs.79877	N/A	0.151264	0.590407
RP11-90P5	8p12	Hs.425311	LSM1	0.149684	0.594418

Figure 2. Table view to display the correlations of the matched BAC clones in aCGH and genes in expression microarray experiments.

as separate bar graphs for better visualization of the individual experiments if the number of matched genes is high (Fig. 4). This function provides a graphical visualization of the correlation between a BAC clone and its matching genes/features.

Application

To test this program, we have analyzed a previously published dataset that contains data from both aCGH arrays (Man, 2004) and cDNA microarrays (Man, 2005) of a set of pediatric osteosarcoma patients. We found several genes with RNA expressions correlating with the DNA copy numbers in the corresponding BAC clones ($r > 0.5$, $n = 15$, $p < 0.05$, Fig. 2). One of the highly correlated genes (ZNF187) is mapped back to the BAC clone RP5-874C20, which is one of the most frequently amplified regions (6p21.1) in osteosarcoma (Man, 2004). ZNF187 or SRE-ZBP is induced by serum response and may regulate oncogene c-fos by binding to its serum response element (Attar, 1992). We have validated matching and correlation results of CGI by manually searching the UCSC genome browser to confirm the match between BAC Clone ID and Unigene ID, and recalculated their correlation coefficients using an independent method.

Discussion

We have developed the CGI program, which provides a simple yet powerful tool for matching, correlating, and visualizing aCGH and gene expression-profiling results simultaneously in multiple experiments. This tool is useful because it correlates the results from DNA profiling with those from RNA expression-profiling experiments in order to identify genes that are important at both DNA and RNA levels. The genes that are significantly altered in both sets of experiments add more confidence to the biological significance of these genes and therefore warrant further investigation. It also alleviates the need for manual matching between BAC clones on the aCGH arrays and the features in gene expression arrays using public databases. For data analysis, it provides a visualization tool and correlation calculations with an interactive and flexible interface. We have also implemented error detection routines to handle the database connection, e.g. user needs to enter

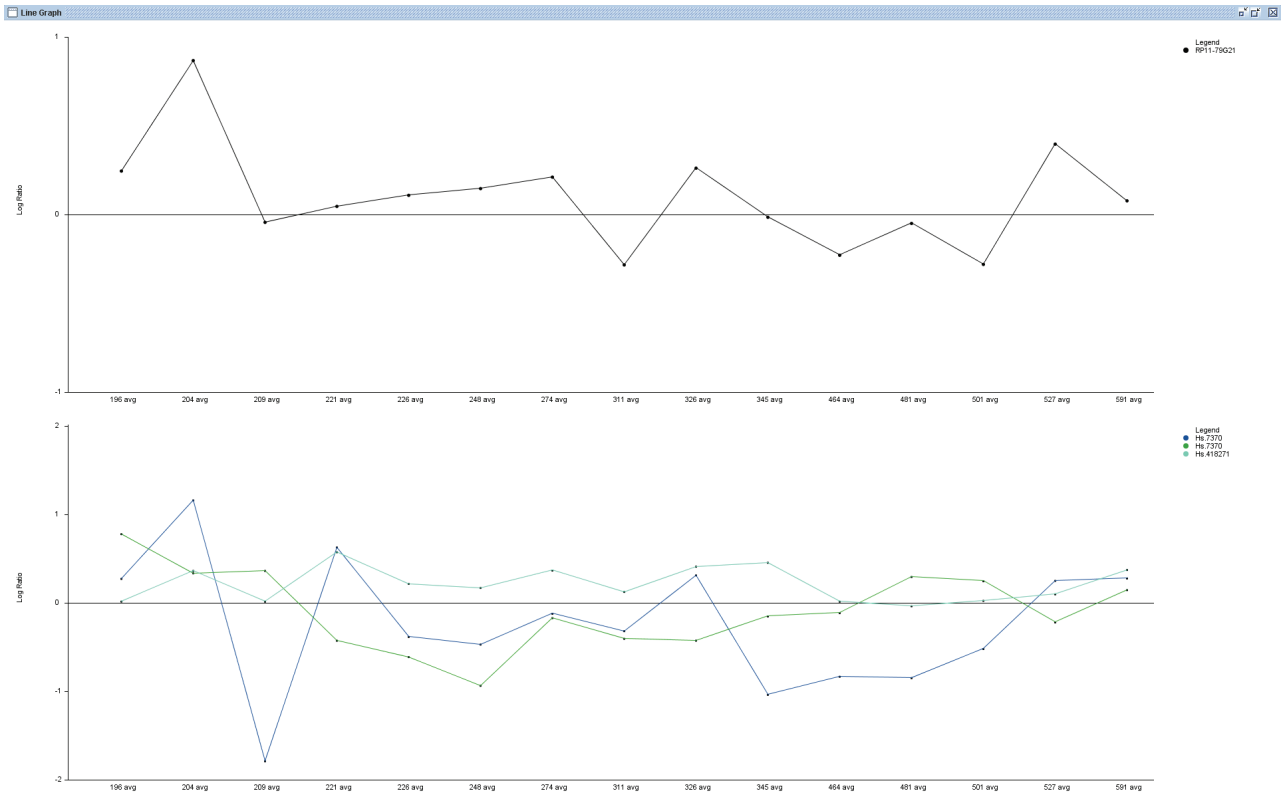


Figure 3. Line graph viewer. The top panel is the copy number changes of a BAC clone in a set of aCGH experiments. The bottom panel is the gene expression values of three corresponding genes that matched the BAC clone in expression microarray experiments using the same experimental cases.

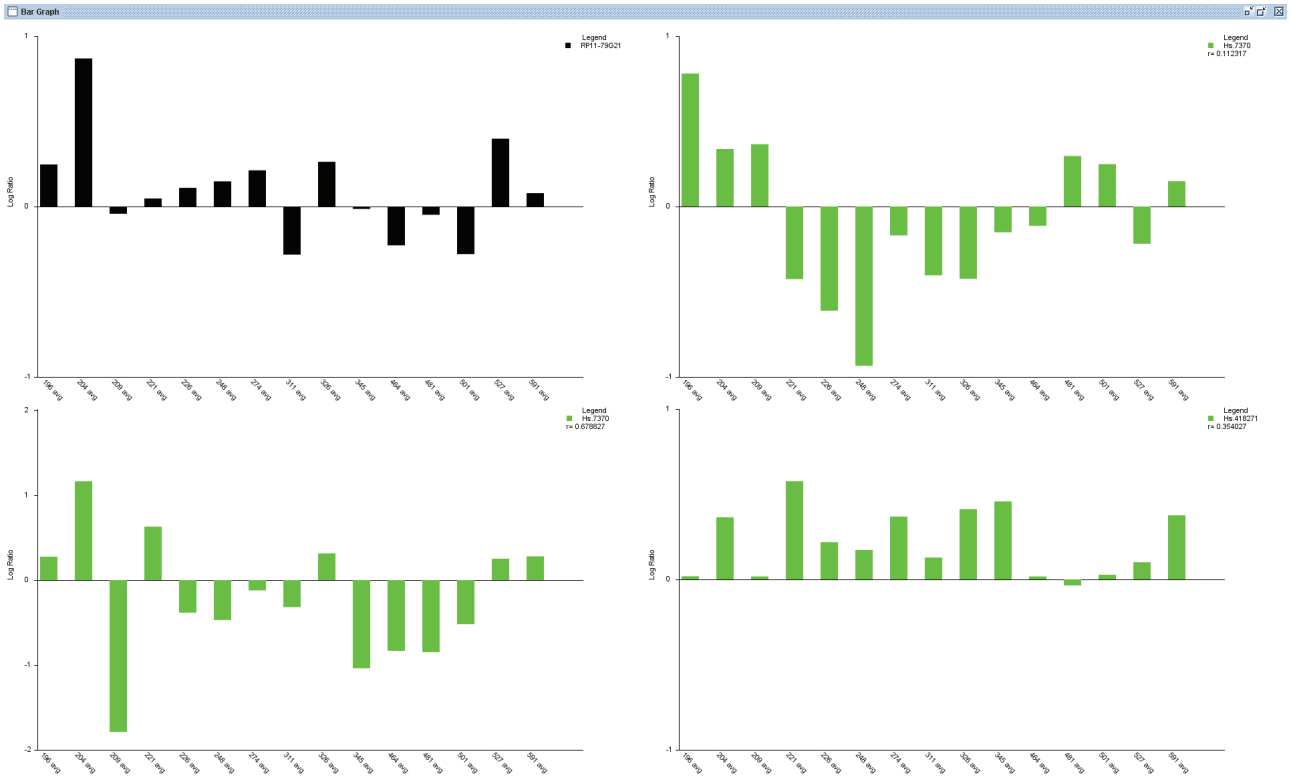


Figure 4. Bar graphic viewer. The black bar graph shows DNA copy numbers (on the Y-axis) of the BAC clone from 15 different experiments (numbered on X-axis). The relative gene expression values (on the Y-axis) of three corresponding genes mapped to the BAC in expression array experiments were displayed in three separate green bar graphs for clearer visualization.

username, password, and database name for secure connection. The number of experiments in the input files are also checked to ensure comparability of the data. This software was implemented in an object-oriented language, Java, to ensure portability across different operating systems. It is a stand-alone program, which is designed for users to install and run on their own local machine. Therefore, unlike other web-based analytical tools, the users do not need any server support, and are not affected by Internet traffic, server-side problems and downtime. Instead of using flat file data storage, the CGI program also provides fast data access and transfer from MySQL database, which is freely available via the web site (http://www.MySQL.com).

Different from some analytical tools, such as BioConductor (www.bioconductor.org), which uses command line interface, the CGI program uses an easy-to-use and intuitive graphical interface for bench biologists to perform the analysis without any prior computational background. A detailed description on how to install the databases and program is also provided in the supplementary information. The software framework that we employ supports the development of more sophisticated visualization and analytical functions in the future through its open API for Java-based plug-ins. The program is coded in Java reusable object classes, thus promoting a rapid development of future program extensions.

Two similar efforts of comparing aCGH and expression-profiling have been published recently. Kingsley et al have recently developed a web-based system, Magellan, which explores the quantitative relationship between aCGH and mRNA expression data (Kingsley, 2006). Magellan computes the relationship of aCGH and expression based on common annotation values between the two sets of experiments. Shankar et al has also developed a program mainly to visualize aCGH and expression data (Shankar, 2006). In contrast to these other two programs, the CGI program is standalone program, which does not require Internet connection and is not affected by the server-side problem. In addition to visualizing the data, the main strength of CGI is to provide an easy-to-use interface for fast matching and correlation of these two types of genomic data using a relational database. Once these candidate genes are identified, they can be

subjected to additional analyses using other existing analytical tools. Since the software is developed in the object-oriented language Java, it can interact with other programs currently available for aCGH and microarray analysis, such as the BioConductor packages. It is straightforward to include other computational algorithms to extend the analytical capability of the program. The modular design of this program also adds flexibility and extensibility for the development of more functions and plug-ins in the future. In summary, we have developed an easy-to-use program CGI to map, correlate, and visualize aCGH and expression profiling data.

Acknowledgements

We would like to thank Jaya Visvanthan, and Jianhe Shen for the preparation of the aCGH and microarray data used in this study. We also thank Alexander Yu, Wei-chun Hsu, and Richard Lowry for their help in programming, and Carolyn Pena for her assistance in manuscript preparation. The study is supported by grants from NIH CA81465, the Robert J. Kleberg, Jr. and Helen C. Kleberg Foundation, the Gillson Longenbaugh Foundation, and the Cancer Fighters in Houston (CCL), as well as the Sarcoma Foundation of America and Fleming and Davenport Award (TKM).

References

Attar, R.M. and Gilman, M.Z. 1992. Expression cloning of a novel zinc finger protein that binds to the c-fos serum response element. *Mol. Cell. Biol.*, 12:2432–43.

Bumm, K and Cheng, M. 2002. CGO: utilizing and integrating gene expression microarray data in clinical research and data management. *Bioinformatics*, 18:327–8.

Chen, W., Erdogan, F., Ropers, H.H. et al. 2005. CGHPRO—a comprehensive data analysis tool for array CGH. *BMC Bioinformatics*, 6:85.

Cheung, S.W., Shaw, C.A., Yu, W. et al. 2005. Development and validation of a CGH microarray for clinical cytogenetic diagnosis. *Genet. Med.*, 7:422–32.

Jong, K., Marchiori, E., Meijer, G. et al. 2004. Breakpoint identification and smoothing of array comparative genomic hybridization data. *Bioinformatics*, 20:3636–7.

Kim, S.Y., Nam, S.W., Lee, S.H. et al. 2005. ArrayCGH: a web application for analysis and visualization of array-CGH data. *Bioinformatics*, 21:2554–5.

Kingsley, C.B., Kuo, W.L., Polikoffm, D. et al. 2006. Magellan: A Web Based System for the Integrated Analysis of Heterogeneous Biological Data and Annotations; Application to DNA Copy Number and Expression Data in Ovarian Cancer. *Cancer Informatics*, 1:10–21.

Li, C. and Wong, W.H. 2001. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci.*, 98:31–36.

Man, T.K., Lu, X.Y., Kim, J. et al. 2004. Genome-wide array comparative genomic hybridization analysis reveals distinct amplifications in osteosarcoma. *BMC Cancer*, 4:45.

Man, T.K., Chintagumpala, M., Visvanathan, J. et al. 2005. Expression profiles of osteosarcoma that can predict response to chemotherapy. *Cancer Research*, 65:8142–50.

Margolin, A.A., Greshock, J., Naylor, T.L. et al. 2005. CGHAnalyzer: a stand-alone software package for cancer genome analysis using array-based DNA copy number data. *Bioinformatics*, 21:3308–11.

Pinkel, D., Segraves, R., Sudar, D. et al. 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genet.*, 20:207–11.

Price, T.S., Regan, R., Mott, R. et al. 2005. SW-ARRAY: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data. *Nucleic Acids Res.*, 33:3455–64.

Saeed, A.I., Sharov, V., White, J. et al. 2003. TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*, 34:374–8.

Saraiya, P., North, C. and Duca, K. 2005. An insight-based methodology for evaluating bioinformatics visualizations. *IEEE Trans. Vis. Compute.Graph*, 11:443–56.

Shamir, R., Maron-Katz, A., Tanay, A. et al. 2005. EXPANDER–an integrative program suite for microarray data analysis. *BMC Bioinformatics*, 6:232.

Shankar, G., Rossi, M.R., McQuaid, D.E. et al. 2006. aCGHViewer: A Generic Visualization Tool For aCGH data. *Cancer Informatics*, 2:36–43.

Sykacek, P., Furlong, R.A. and Micklem, G. 2005. A friendly statistics package for microarray analysis. *Bioinformatics*, 21:4069–70.

Vaquerizas, J.M., Conde, L., Yankilevich, P. et al. 2005. GEPAS, an experiment-oriented pipeline for the analysis of microarray gene expression data. *Nucleic Acids Res.*, 33:W616–W620.