GeneX Va: VBC open source microarray database and analysis software

Jae K. Lee[1], Tom Laudeman[1], Jodi Kanter[1], Teela James[1], Mir S. Siadaty[1], William A. Knaus[1], Alyson Prorok[1], Yongde Bao[1], Brad Freeman[2], Daniela Puiu[2], Li min Wen[2], Gregory A. Buck[2], Karen Schlauch[3], Jennifer Weller[3], and Jay W. Fox[1]

[1]University of Virginia School of Medicine, Charlottesville, [2]Virginia Commonwealth University, Richmond, and [3]George Mason University, Manassas, VA, USA

BioTechniques 36:__-__ (April 2004)

Developed by the Virginia Bioinformatics Consortium (VBC), GeneX Va is an open source, freeware database and bioinformatics analysis software for archiving and analyzing Affymetrix GeneChip® data. It provides an integrated framework for management, documentation, and analysis of microarray experiments and data to support a range of users, from individual research laboratories to institutional microarray facilities. GeneX Va also provides web-based access to a PostgreSQL relational database system with a comprehensive security system. Data can be extracted from the database and delivered to interactive or scriptable statistical analysis protocols. The security system allows each investigator to manage their own array data and analysis output files and also provides custom access privileges for other users, groups, and internal/external collaborators. The analysis interface uses "Analysis Trees," an innovative user interface that allows researchers to interactively create a tree-structured flow chart of analysis routines. The latest GeneX Va software is available from and can be freely downloaded at the Sourceforge web site http://va-genex.sourceforge.net. To allow researchers to access the database and analysis capabilities of the GeneX Va system, microarray data from many VBC GeneChip experiments have been deposited into a public section of the GeneX Va system at the University of Virginia. The VBC GeneX Va sites, which include documentation, are at http://genes.med.virginia.edu/ of the University of Virginia and at http://genex.csbc.vcu.edu/ of the Virginia Commonwealth University.

INTRODUCTION

Microarray technologies are rapidly emerging for genome-wide gene expression studies in biology and medical sciences (1,2). Researchers who have performed microarray experiments seek reliable tools for identifying significant changes in gene expression and extracting the biological pathways that may be responsible for such changes. These investigations require user-friendly databases that link gene sets to their biological functions and appropriate statistical tools for microarray data analysis (3). However, processing and analyzing microarray data with large genome information databases is often overwhelming for individual researchers due to the size and complexity of these data sets. In order to support individual researchers, many research institutions have established shared microarray facilities that conduct microarray experiments and manage expression data using commercial or custom-built database systems for reporting, reviewing, and analyzing the data (4,5).

Unlike sequence data, gene expression data are context-dependent, and interpretations of experimental results may differ depending on the combinations of experimental conditions, RNA samples, array instrumentation, and software processing techniques (6). The problem of sample documentation for gene expression experiments is well recognized. For example, the Microarray Gene Expression Data (MGED) Society has recommended the minimum information about a microarray experiment (MIAME) guidelines to specify the types of descriptive information that should be reported with each microarray experiment (7). Such metadata are vital to provide meaning and guidance to a set of experimental gene expression data and enable one to replicate its experimental results. MIAME standards have been adopted by many scientific journals as publication standards and by public and private microarray databases (8).

Under the collaboration of the Virginia Bioinformatics Consortium (VBC), the University of Virginia (UVA) and the Virginia Commonwealth University (VCU) have sought to address their expression array data management needs by adopting and extending GeneX, an open source gene expression data management and analysis system (9). In the original GeneX, the database schema, specific tools, and a web-based interface were developed in support of the uploading and annotation of cDNA microarray data. This GeneX system did not have tools for handling oligonucleotide microarray data (GeneChip®; Affymetrix, Santa Clara, CA, USA), and it also lacked data security features required by a shared core research support facility comprising a diverse group of users. In order to satisfy these requirements, GeneX Va has been significantly refined and further developed. These changes focus primarily on the microarray experiment workflow encountered in biomedical research (Figure 1A). Important features supported by GeneX Va are summarized in Supplementary Table S1, GeneX Va Gene Expression Database Overview. Supplementary Tables S1–S4 and Supplementary Figure S1 can be found on the BioTechniques web site at http://www.BioTechniques.com/April04/LeeSupplementary.html.

Currently, there exist a number of public microarray database systems. ArrayExpress at the European Bioinformatics Institute (EBI), which has been developed in compliance with the MIAME guidelines, is a public database of microarray data with online submission, database query, and certain clustering and functional filtering tools (8). Gene Expression Omnibus (GEO) at the U.S. National Center for Biotechnology Information (NCBI) serves as a hub to archive public microarray data (3), and the Stanford Microarray database (SMD) also provides public access to their extensive gene expression data, mainly in cDNA microarray

format on human and many other species (4). While these and many other well-known systems have been highly recognized for their public services and rich gene expression databases, these cannot be used as a database system to support individual institutions or groups for their local data archiving and analysis, especially with secured access that is required in biomedical research (5). Furthermore, stable versions of their release codes, especially as open source software, are not available, so customization is either expensive or impossible based on these systems. The GeneX Va system is currently used to perform and manage individual investigators' daily gene expression research at the UVA (http://genes.med.virginia.edu/) and at the VCU (http://genex.csbc.vcu.edu/) and is readily installed by independent laboratories or institutions. The VBC GeneX Va site also provides many public microarray data sets with full MIAME documentation from the biomedical research experiments performed at the VBC institutions.

SYSTEM USE AND USER INTERFACE

GeneX Va comprises a relational database, a web interface, and a statistical analysis suite. There is also a repository for each user's files and a comprehensive security system. All interaction with the system is via secure web pages; no special software is required for end users. Sample documentation can be the most burdensome for individual microarray researchers, because it is often difficult for them to determine and maintain all relevant information with a standardized format for an independent use and analysis of each microarray data set. GeneX Va seeks to reduce this burden by integrating the collection of sample information with the process of submitting samples to the microarray core facility, which greatly enhances the value of the expression data. For example, Figure 1B depicts a schematic diagram that is followed by investigators as they create and document an array study and order(s) for specific chip hybridizations. Using the system, an investigator initiating an array experiment first creates an "array study" and submits an order to the microarray core facility. GeneX Va's interface then requires users to enter the essential sample descriptive data and collects information required by MIAME guidelines at the start of an array study. During this initial data entry step users also create consistent names for each RNA sample, which are carried over and used to identify the actual sample tubes and chips of the array experiment, along with date and array chip information to form a unique identifying ID. Once each microarray experiment is completed, the array core center automatically uploads microarray intensity data from the array scanner to the GeneX Va database. The expression data can then be exported locally in text format and/or analyzed using the GeneX Va analysis interface. At each step, a series of relevant procedures are supported by the user interface. A step-by-step guide for these procedures (Supplementary Document S1, Step-by-Step Guide to GeneX Va), GeneX Va version 1-6.9, and an additional user manual (Supplementary Document S2) are available at the BioTechniques web site at http://www.BioTechniques.com/April04/LeeSoftware.html and can also be obtained at the Sourceforge site (http://sourceforge.net/projects/va-genex/).

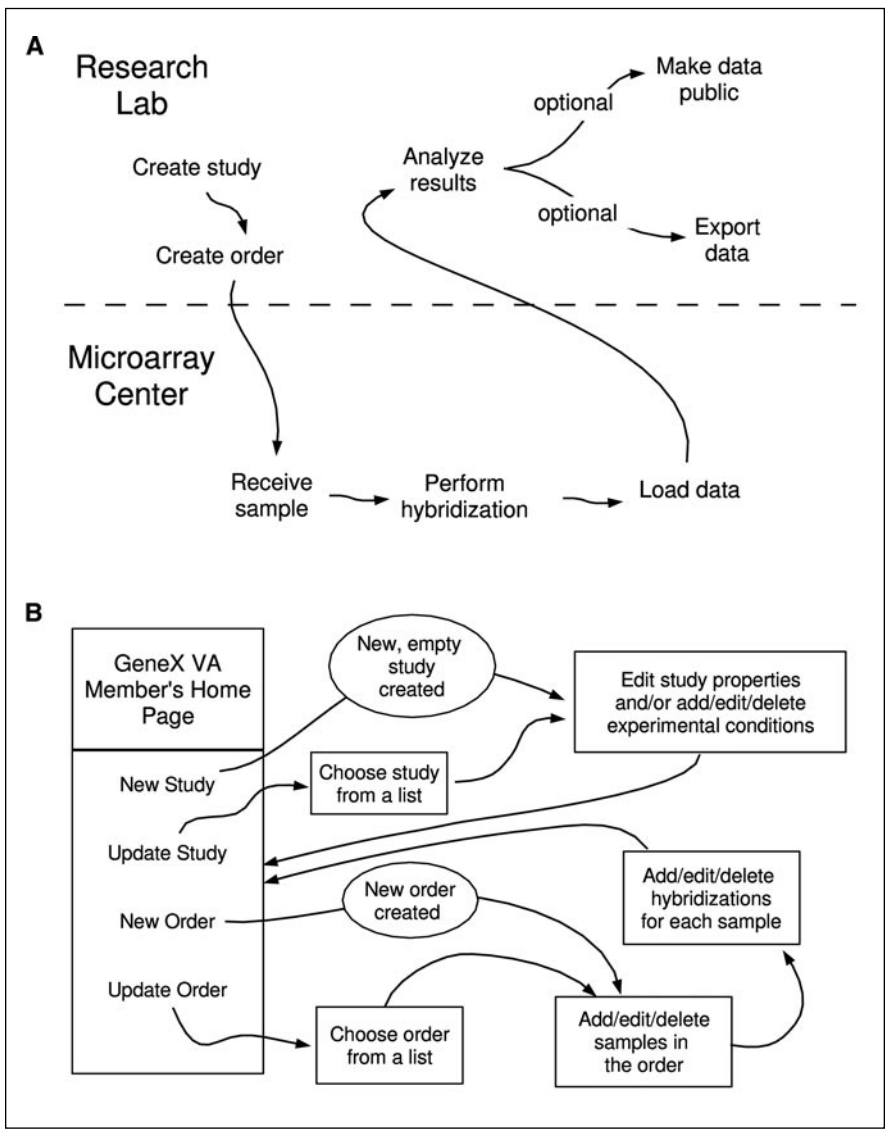

Figure 1. Workflow for microarray experiments. (A) Information and sample transfer between researchers and the microarray core facility. (B) Steps required to create a new study and describe its experimental conditions, supported by GeneX Va.

SYSTEM DESIGN AND SECURITY

The GeneX Va database is built upon PostgreSQL, a mature, open source relational database system, with Perl DBI and application programming interfaces (APIs). The database schema was originally based on open source GeneX 1.05 software, but has been significantly modified and expanded to reflect the workflow and security requirements mentioned above. GeneX Va now contains seven conceptual categories of tables to define experimental information (including RNA samples), array layout, array measurements, spotted sequence/gene information, analysis, ID and security, and administration (see Supplementary Figure S1, GeneX Va database schema). The last two categories of tables implement the comprehensive security system that can manage access to microarray data from clinical samples.

Although there is substantial sentiment within the scientific community that gene expression data should be publicly available, most individual researchers feel that data need only be made accessible after the results are accepted for publication. To gain acceptance by the diverse group of researchers at the UVA, our first modification of the initial GeneX (version 1.0.4) software package was the addition of a versatile security module. This module implements security at the database level, using row-level security on the data tables, and allows users and groups to be defined such that data and metadata can be shared in many ways. Investigators may conduct multiple array studies with different collaborators, and a single microarray data set may be used for several studies, with different subsets shared among researchers at different institutions. Unpublished array data and related patient information must be stored securely under each user's access and control, with legitimate access monitored.

GeneX Va security is implemented at both the web interface for server login and at the database level for accessing database records. The database security model has users, groups, and primary investigators (PIs), the latter of which are the users registered in the PI table with primary owner privileges of each array data set. The UNIX®-like permission scheme provides different levels of data and database security, allowing different accessibility and management privileges for curators, array center personnel, data owners, users, and groups. In order to assure security of the system and data as it transits the network, GeneX Va uses Secure Sockets Layer (SSL) via https and web security implemented via sessioning.

ANALYSIS INTERFACE

While data management, documentation, and security are essential for large-scale microarray core facilities, individual investigators are most interested in flexible, reliable, and informative analysis protocols that allow them to focus on the biologically important changes revealed by the array experiment. Interpretation of microarray results requires a series of analysis procedures from quality control to functional analysis, which can involve many different combinations of analysis options and parameters. Most current microarray database systems [e.g., ArrayExpress, National Cancer Institute's (NCI's) mAdb, and Stanford Microarray database (SMD)] provide various analysis routines such as clustering and functional analysis tools, but many of these analysis routines cannot be linked together to produce a complete, consistent reusable analysis method (4,5,8). Flexible analysis strategies become critical when combining analysis of data from diverse sources. GeneX Va has a powerful and flexible analysis interface, AnalysisTrees, which allows users to interactively create, save, and freely modify a tree-structured flow chart of analysis routines with selectable analysis options for these analysis demands.

Analysis on the GeneX Va system is initiated by using the web-based query tools to retrieve all or a subset of a relevant array data set. Data retrieval can be based either on existing array studies and experimental conditions or on a virtual array study and/or virtual experimental conditions created by researchers. The latter utility is important because researchers may want to analyze gene expression data (either their own or publicly accessible data) by forming a new array data set from a combination of different array experiments (e.g., an array experiment with activated T cell subpopulations and the other with natural killer and dendritic cells for a combined immune-response microarray study). Having established an experimental data set, users then construct analysis methods (linked procedures), which can be visualized by using the GeneX Va analysis interface. Analysis Trees represent hierarchically linked analysis procedures with branches for several analysis routines. Each node of the analysis tree is defined with its input, output, and connections to other compatible analysis nodes (Figure 2). Any number of compatible analysis routines can be added at each node with different selectable options and parameter values for each analysis routine. Each of these analysis routines has various options, as summarized in Supplementary Table S2, GeneX Va Analysis Routines. All analysis (intermediate and output) files are saved, together with their history (log) files, to the researcher's account. A detailed guide for the use of the GeneX Va analysis interface can be found at the GeneX Va web sites (Supplementary Document S2, Analyzing your Data with Analysis Trees). New analysis routines can be written in R, Perl, or C (and many other scripting and programming languages) and can be added to the system based on a plug-in architecture (using in-house Perl utilities such as Rwrapper). Currently, the functionality of each newly added analysis routine is carefully tested, and its compatibility with input and output routines is validated by the GeneX Va developers at the UVA.

OPEN SOURCE SOFTWARE AND INSTALLATION

The GeneX Va system is completely based on open source software, including a Linux server, Apache web server, and PostgreSQL relational database. The web interface and internal control scripts are written in standard Perl. Most statistical analysis routines are written in R, an open source

statistical programming environment (http://www.r-project.org), and some functions written in C and Perl. These analysis routines are developed by the members of the UVA GeneChip/Microarray Bioinformatics (GMB) core and the current GeneX development group based at George Mason University (GMU), and adapted from open source analysis routines, such as the Bioconductor packages (http://www.bioconductor.org/). For the download of the up-to-date software package with its detailed installation instructions, we maintain a Sourceforge site at http://va-genex.sourceforge.net/ or http://sourceforge.net/projects/va-genex/. The complete GeneX Va software package is compact, and its installation can be performed within a few hours, although additional pre-installations of open source packages are required. Documentation can be found in the installation support package. GeneX Va software releases (including analysis routines) are validated based on a standard testing procedure implemented by engineers independent of the development of each component.

PUBLIC MICROARRAY DATA AND ACCESS

We have developed a standard procedure, the majority of which is automated, to consistently and efficiently post public array data in the GeneX Va system with its relevant MIAME information. For guest users, the GeneX Va site at the UVA (http://genes.med.virginia.edu) currently provides 17 public array data sets from various experiments in biomedical research, including human transcriptional profiles for melanoma, Alzheimer and Parkinson diseases, antigen immune response, and various cancer and human diseases as summarized in Supplementary Table S3, VBC Public Microarray Data Sets. Additional array data will be continuously added.

FUTURE DEVELOPMENT

We are currently adding various analysis routines to perform more comprehensive investigations on microarray data and to integrate various functional and annotation analysis tools that can effectively triage genes based on each investigation goal, including many open source analysis procedures that are well recognized and validated for their usefulness, such as *affy, genefilter,* and *multtest* in the Bioconductor packages and others. In parallel with these refinements of GeneX Va analysis capability, we will make our analysis server available via a public Web interface in the near future.

Customized cDNA array technologies are also currently used in experiments at VBC institutions, and specific user interfaces for cDNA array data are under development. One of the most challenging tasks associated with the use of custom arrays is in large part due to the vast number of array layouts that would need to be supported. Archiving the manufacturing and technical aspects of array fabrication becomes problematic. Future enhancements include plans to support documentation of a large number of layouts as well as plans to support counting-type of gene expression technology such as serial analysis of gene expression (SAGE) and amplified fragment length polymorphism (AFLP) soon. We note that under the GeneX 2.0 development, some of these features have already been instated.

Integration of the GeneX Va data-

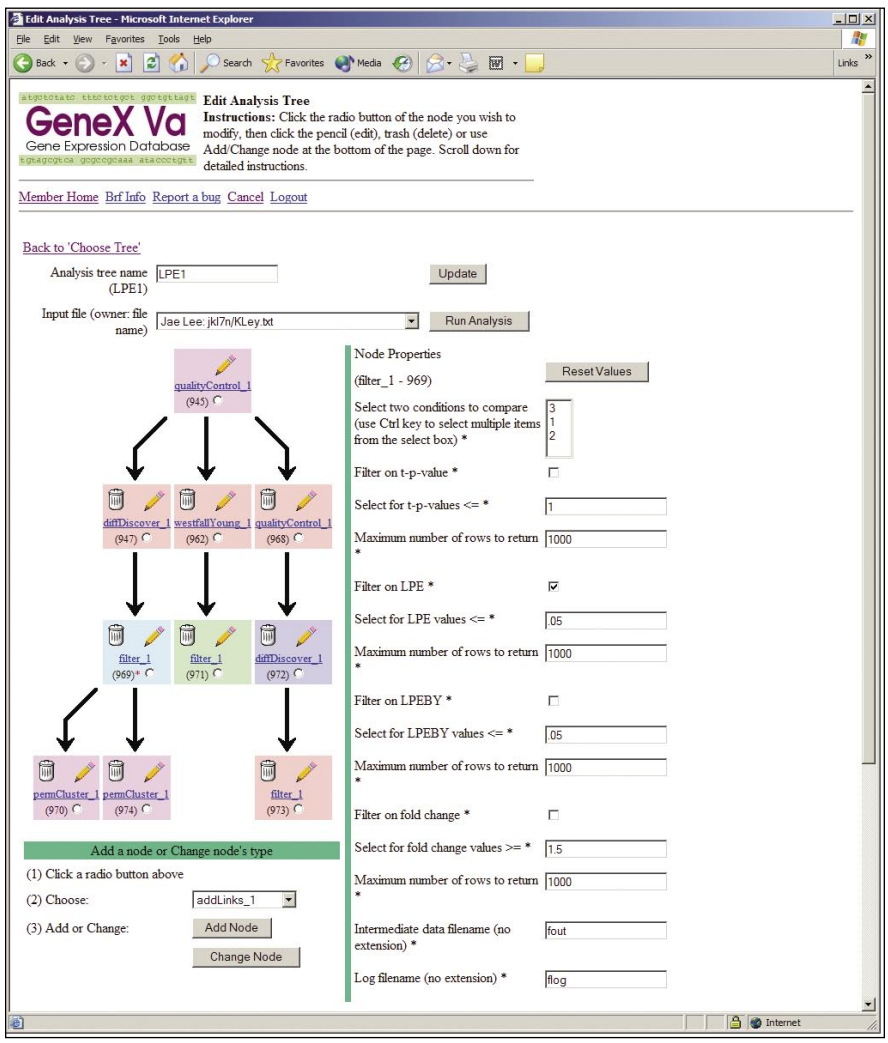

Figure 2. An analysis tree: the GeneX Va graphical analysis interface. An extended analysis tree is drawn with three branches: (*i*) a standard procedure of differential discovery, filtering, and clustering analyses; (*ii*) an analysis with the Westfall-Young statistical test; and (*iii*) an analysis with different quality control (QC) options.

base with complementary genomic and clinical data across multiple institutions is forthcoming. In doing so we will ensure that any patient-related data is stored and exchanged in compliance with the Health Insurance Portability and Accountability Act (HIPAA). The proposed interface between the GeneX Va system and clinical databases will consequently provide anonymity for all patient-identifiable information. Array and personal clinical data will be combined by a properly authorized user and then analyzed as dictated by each predefined investigation goal. As our GeneX Va system is integrated with various medical and clinical data sources, security issues will become increasingly critical. We will monitor the effectiveness of our security implementation and update as necessary to remain current with new regulations. We are also in the process of developing tools to monitor and control client use. Through our collaboration with the GeneX 2.0 developers at GMU, we will consider integrating the microarray gene expression-markup language (MAGE-ML) and MAGE-object model (MAGE-OM) concepts and the XML data exchange format.

We also plan to provide more effective tools for the storage and analysis of microarray data, to enhance usability by linking with existing tools/applications and incorporating new features, and to continuously improve the software by responding to security issues, user needs, and integrating better technologies as they become available. All of these future developments must be carefully considered to tie with web services of database and software interoperability of analysis tools, which will be continuously evaluated by system developers, array center, and individual investigators for efficient workflow creation and management.

ACKNOWLEDGMENTS

GeneX Va development is supported by Virginia's Commonwealth Technology Research Fund (CTRF) grant, Center for Innovative Technology (CIT) grant BIO-02-004, and the University of Virginia Pratt Fund.

REFERENCES

1. **Sander, C.** 2000. Genomic medicine and the future of health care. Science *287*:1977-1978.
2. **Tusher, V., R. Tibshirani, and C. Chu.** 2001. Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. Proc. Natl. Acad. Sci. USA *98*:5116-5121.
3. **Knudtson, K.L., C. Griffin, D.A. Iacobas, K. Johnson, G. Khitrov, S. Levy, A. Massimi, N. Nowak, et al.** 2003. A Current Profile of Microarray Laboratories: the 2002-2003 ABRF Microarray Research Group Survey of Laboratories Using Microarray Technologies (http://www.abrf.org/ResearchGroups/Microarray/EPosters/MARG_Survey_Poster2003.pdf).
4. **Gollub, J., C.A. Ball, G. Binkley, J. Demeter, D.B. Finkelstein, J.M. Hebert, T. Hernandez-Boussard, H. Jin, et al.** 2003. The Stanford Microarray Database: data access and quality assessment tools. Nucleic Acids Res. *31*:94-96.
5. **Gardiner-Garden, M. and T.G. Littlejohn.** 2001. A comparison of microarray databases. Brief Bioinform. *2*:143-158.
6. **Bassett, D.E., Jr., M.B. Eisen, and M.S. Boguski.** 1999. Gene expression informatics—it's all in your mine. Nat. Genet. *21(1 Suppl)*:51-55.
7. **Stoeckert, C.J., H.C. Causton, and C.A. Ball.** 2002. Microarray databases: standards and ontologies. Nat. Genet. *32*:469-473.
8. **Brazma, A., H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena, E. Holloway, M. Kapushesky, et al.** 2003. ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic Acids Res. *31*:68-71.
9. **Mangalam, H., J. Stewart, J. Zhou, K. Schlauch, M. Waugh, G. Chen, A.D. Farmer, G. Colello, and J.W. Weller.** 2001. GeneX: an open source gene expression database and integrated tool sets. IBM Systems J. *40*:552-569.
10. **Lee, J.K.** 2002. Discovery and validation of microarray gene expression patterns. LabMedica International *2*:8-10.

Received 22 December 2003; accepted 11 February 2004.

Address correspondence to Jae K. Lee, Department of Health Evaluation Sciences, Hospital West Complex, Room 3181, University of Virginia School of Medicine, P.O. Box 800717, Charlottesville, VA 22908-0717, USA. e-mail: jaeklee@virginia.edu