D756–D759 *Nucleic Acids Research, 2007, Vol. 35, Database issue* *Published online 7 November 2006*
doi:10.1093/nar/gkl798

Gene Aging Nexus: a web database and data mining platform for microarray data on aging

Fei Pan[1], Chi-Hsien Chiu[1], Sudip Pulapura[1], Michael R. Mehan[1], Juan Nunez-Iglesias[1], Kangyu Zhang[1], Kiran Kamath[1], Michael S. Waterman[1], Caleb E. Finch[1,2,*] and Xianghong Jasmine Zhou[1,*]

[1]Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089, USA and [2]Andrus Gerontology Center, University of Southern California, Los Angeles, CA 90089, USA

Received August 15, 2006; Revised September 20, 2006; Accepted October 2, 2006

ABSTRACT

The recent development of microarray technology provided unprecedented opportunities to understand the genetic basis of aging. So far, many microarray studies have addressed aging-related expression patterns in multiple organisms and under different conditions. The number of relevant studies continues to increase rapidly. However, efficient exploitation of these vast data is frustrated by the lack of an integrated data mining platform or other unifying bioinformatic resource to enable convenient cross-laboratory searches of array signals. To facilitate the integrative analysis of microarray data on aging, we developed a web database and analysis platform 'Gene Aging Nexus' (GAN) that is freely accessible to the research community to query/analyze/visualize cross-platform and cross-species microarray data on aging. By providing the possibility of integrative microarray analysis, GAN should be useful in building the systems-biology understanding of aging. GAN is accessible at http://gan.usc.edu.

INTRODUCTION OF GAN

The recent development of high-throughput technologies resulted in an enormous volume of genomic data to understand the genetic basis of aging. Among those, the microarray technology allows us to measure the expression level of all genes in a genome simultaneously. Thus far, more than 80 microarray studies have directly addressed aging-related expression patterns in diverse model organisms and under different conditions. We define 'aging-related' data to be those datasets which include adult age as a variable and which include diseases with strong adult age-group dependency, e.g. Alzheimer's disease. Given the large number of aging-related microarray datasets comprising tens of millions of measurements, we see many advantages to collecting them on a single platform for integrative analysis: (i) Aging-related signals are generally more subtle than those disease-related signals (e.g. cancer), therefore very difficult to detect. Identifying recurrent signals across multiple datasets enhances signal/noise separation, and can elucidate essential transcriptional features in aging. (ii) As the available aging microarray datasets generally measure different aspects of aging (e.g under different endogenous conditions and exogenous perturbations), combining those datasets can complement each other in revealing the transcriptional mechanisms of the aging. (iii) Comparing genomic expression profiles across species may reveal evolutionary conserved mechanisms in aging processes, as exemplified in McCarrol *et al*. (1).

However, technical obstacles complicate the integration of multiple microarray datasets. A key problem is the existence of the diverse microarray platforms. For instance, human aging expression profiles were conducted using various Affymetrix chips (HuGeneFL, HG_U95A, HG_U95Av2, HG-U133A, etc.) and cDNA arrays (Incyte Genomics and customized). Gene expression values generated by different platform technologies are not necessarily comparable. Even within the same technology, the alternate choice of experimental parameters by different laboratories can cause systematic variation among datasets that often exceeds the capability of statistical normalization. Several recent studies (2–6) proposed meta-analysis approaches to integrate multiple microarray studies. By first extracting expression patterns from individual microarray studies and then identifying recurrent signals, those approaches can enhance signal-to-noise separation. Following this principle, the public database OncoMINE (7) facilitates the identification of genes differentially expressed between cancer and normal tissues or among different cancer subtypes across a large collection of microarray data. Given the large accumulation of aging-related microarray

*To whom correspondence should be addressed. Tel: +1 213 740 1758; Fax: +1 213 740 0853; Email: cefinch@usc.edu
*Correspondence may also be addressed to Xianghong Jasmine Zhou. Tel: +1 213 740 7055; Fax: +1 213 740 2409; Email: xjzhou@usc.edu

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors

data, there is a great need for analytical tools and software platforms to extract aging-relevant information across datasets.

Here we report the development of a web database Gene Aging Nexus (GAN) freely accessible to the biogerontological-geriatric research community to query/analyze/visualize aging-related microarray data. GAN consists of two parts: (i) a database of microarray datasets measuring aging-related expression patterns; (ii) a data mining platform to facilitate the identification of recurrent expression patterns across multiple datasets and species. The database is intended as a shared repository for microarray data on aging as well as to provide easy access to microarray data in a user-friendly way. The web-based data mining platform allows users to perform integrative analysis to derive customized differentially expressed gene lists or co-expressed gene pairs, and apply functional annotation tools. Although GAN's focus is on aging, the architecture is general and could be adapted to other subject-specific knowledge mining platforms for efficient and accessible usage of the public microarray data.

GAN'S DATABASE SERVER FOR AGING-RELATED MICROARRAY DATA

GAN database includes microarray datasets generated from both Affymetrix and cDNA platforms. We collected 42 aging-related microarray datasets for the model organisms human, mouse, rat, fruit fly, worm and yeast from the NCBI GEO database, Stanford Microarray Database (SMD), individual publication websites and personal communications. This comprises more than 14 176 000 gene expression measurements from over 800 microarray experiments. A breakdown of the datasets is shown in Table 1. The datasets can be categorized into four classes: (i) Gene expression profiling of different age groups in various tissues and organisms, e.g. human frontal cortex, human kidney, human muscle, mouse retinal pigmented epithelium, mouse coronary artery, mouse cerebellum, mouse hematopoietic stem cells, rat nervous system, rat muscle, rat glia, fly head, worm, etc. (ii) Effect of different perturbations on aging, e.g. oxidative stress, caloric restriction or GH/IGF-1 signaling disruption. (iii) Aging effect on metabolism, e.g. glucose metabolism and neuroinflammation. (iv) Studies on Alzheimer's disease, e.g. Alzheimer's disease at different severity, comparison to control and animal models.

Normalized data from NCBI and SMD are directly imported. For other datasets, if the raw image CEL files are available, we have re-done the image processing, background subtraction and normalization procedure by using the Bioconductor software. Although it is not realistic for us to undertake a full critical assessment of the quality of each dataset, in our framework, pooling multiple datasets to discover recurrent patterns is in fact a way to filter out poor quality data. Conversely our framework will draw user's attention to those datasets which produce consistent results, thus focus on the datasets with high quality.

Table 1. Categorization of aging-related microarray databases in GAN

Species	Datasets	Experiments	Platforms	
			Affymetric	cDNA
Homo sapiens	12	428	12	0
Rattus norvegicus	5	81	5	0
Mus musculus	17	222	12	5
Drosophila melanogaster	2	15	2	0
Caenorhabditis elegans	5	66	1	4
Saccharomyces cerevisiae	1	8	1	0
Total	42	820	33	9

ANALYSIS TOOLS

GAN's analysis tools include a data visualization module, a co-expression analysis module, a differential expression analysis module and a functional analysis module.

Data visualization module

GAN provides a user-friendly web interface to browse the collected datasets along with detailed dataset annotations (Figure 1). Datasets are categorized by organisms and laboratories to expedite the selection process. After datasets are selected, a brief summary of those datasets will be shown, in order for the user to decide whether to further process the datasets. Once the dataset 'view' option is chosen, the expression matrix of the selected dataset will be displayed, and users may use gene ID, gene symbol, Unigene or GenBank accession no. to search the dataset. Expression levels of a gene can be visualized using bar chart.

Differential expression analysis module

Differential expression analysis will be performed for individual datasets, and genes with frequent differential patterns will be identified across multiple datasets. User may load multiple datasets (from different platforms or different species) from the GAN database. Genes on the different array platforms will be linked via their UnigeneIDs, and homologs of different species will be linked based on the matches in the NCBI HomoloGene Database. For each dataset, the user may select experiments to construct the 'case' and the 'control' groups. Age groups may be defined as young and old, and some examples of treatment groups include, for example, caloric restriction and control, LPS stimuli and control. We have implemented the *t*-test and the Mann–Whitney test to assess the statistical significance of differential expression in individual datasets. The significance is then adjusted for multiple testing with the *Q*-value, a counterpart of the *P*-value in the context of false discovery rate (8). The statistical significance estimate can be combined with user-defined fold change threshold to select the differentially expressed genes. Based on differential expression analysis results derived from individual datasets, users may select those genes that are differentially expressed in at least *m* out of the total *n* selected datasets. In this way, we may identify genes that consistently demonstrate differential expression pattern across comparable conditions or across different species.

Co-expression analysis module

This module is used to derive the gene–gene co-expression relationship across one or multiple datasets. Users may submit the ID of a gene list to be analyzed, and upload one or more datasets from the GAN database for analysis. The

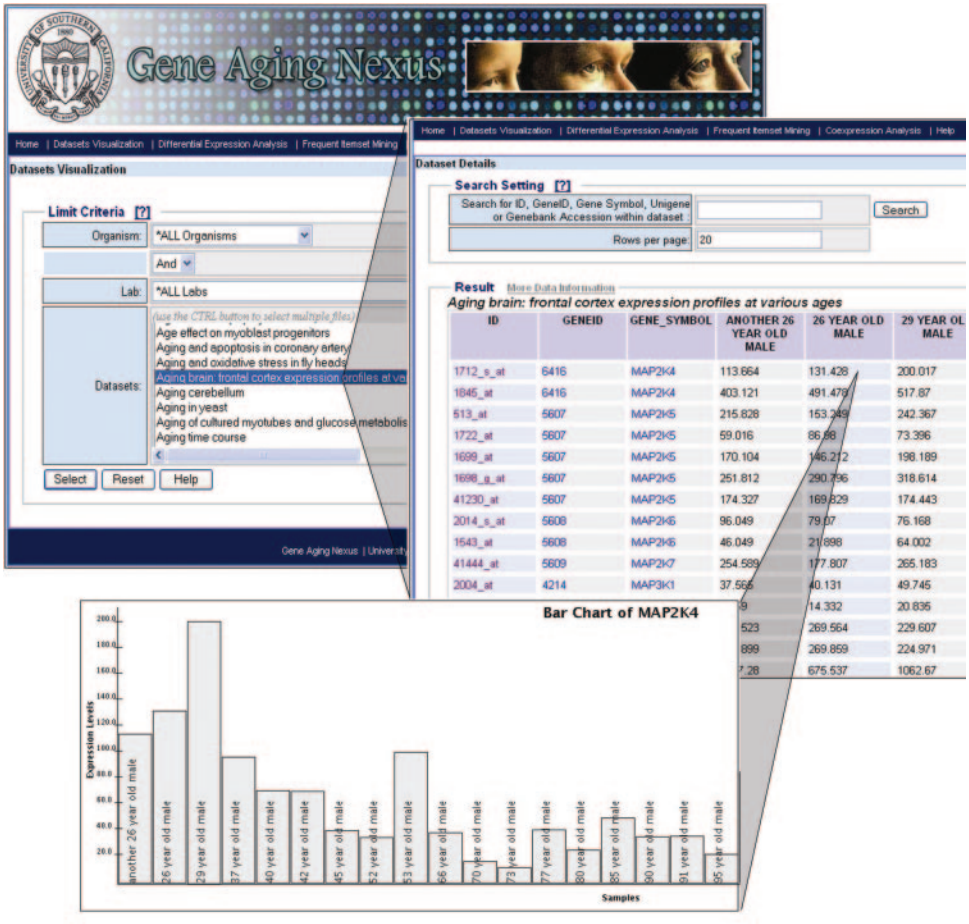

Figure 1. The data visualization interface of GAN.

pairwise expression correlation among selected genes in selected datasets will be displayed in a table format, which allows the users to capture recurrent co-expression relationships. We provide two correlation estimates: Pearson's correlation and 'Jackknife correlation'. Here, the Jackknife correlation is defined as the minimum of the absolute value of leave-one-out Pearson correlation coefficient estimates. This estimate is robust against single experiment outliers yet still sensitive to overall similarities in expression patterns.

Functional analysis module

To facilitate the discovery of novel pathways or novel candidate genes involved in the aging process, we annotate genes with relevant gene ontology descriptions. Given a differentially expressed gene set, we first map each gene onto the Gene Ontology (GO) functional categories, and then evaluate the statistical significance of functional enrichment of each of these categories in the given gene set. We used the hypergeometric distribution to model the probability of observing at least m genes from a gene set of size n by chance in a functional category containing M genes from a total genome size of N genes. Owing to testing a large family of hypothesis simultaneously, we employ stringent Bonferroni correction for multiple testing adjustment. If a GO functional category is statistically significantly enriched in a differentially expressed gene set, the related biological pathway may be activated in the corresponding aging-related conditions. Furthermore, if this gene set contains genes of unknown functions, those genes may be assigned to that particular GO functional category.

DETAILED SYSTEM AND DATABASE ARCHITECTURE OF GAN

Internally, GAN is designed as a layered architecture system, composed of 'presentation layer', 'integration and data analysis layer', 'data access layer' and 'information system layer'. The first layer, the presentation layer, is implemented by JavaServer Pages (JSP) and JavaServer Faces (JSF). It is responsible for providing user interfaces, forming field validation logic and passing users' requests to the web server. The second layer, the integration and data analysis layer, runs on a Tomcat web application server and is responsible for handling requests from the presentation layer. It is the core layer of GAN implemented by java objects and servlets that contain logic to perform the bioinformatics computation and security control for the system. The third layer, the data access layer, contains Data Access Objects (DAO) that are responsible for performing queries on the underlying databases to gather information that are used for bioinformatics

analyses in the 'integration and data analysis layer'. DAO is also responsible for storing the analysis results into the databases. Also, this layer uses the Spring framework to provide connection pooling and transaction management mechanism so that database connections created by one user can be reused by another user, thus saving time and memory. Finally, the information system layer contains databases that store the integrated datasets and the analysis results for each user.

The major responsibilities of this system, which include presentation, analysis and database accessing, are separated into components in each layer. Owing to the layered architecture, it is easy to add functionalities and to change user interfaces in the future without altering the codes too much. This makes the system easy to maintain. Most of the user interfaces are developed using JSF, which is a technology designed to separate the presentation logic and business logic clearly and provide many build-in rich UI components which allow developers to build web application quickly, clearly and easily. Also, with suitable 'Integration Development Environment' tools, it allows developers to drag and drop those build-in UI components and to generate corresponding code automatically. The layered architecture is conducive to future maintenance and scaling up, thus providing the flexibility required to adapt to the future demand from the research community.

The GAN database is composed of three PostgreSQL databases: the Dataset database, the GeneInfo database and the User database. The Dataset database is used to store microarray data related to aging. The GeneInfo Database contains gene annotations as well as information for linking genes in different platforms or across different species. The User database stores user identity, links to files used by the user and metadata files generated by the user.

CONCLUSION AND FUTURE DIRECTIONS

Microarray data are noisy. Identifying recurrent signals from independent microarray studies provide an effective means to separate signal from noise. To facilitate the integrative analysis of microarray data on aging, we developed the web database and analysis platform 'Gene Aging Nexus' which is freely accessible to the research community to query/analyze/visualize cross-platform and cross-species microarray data on aging. The database is also expected to link the genomic information from different species to facilitate the discovery of candidate genes that are involved in aging through the genome-wide comparative analysis. In future, we will implement more sophisticated meta-analysis approaches to extract signals from multiple microarray datasets, e.g. network modules from multiple co-expression networks derived from independent microarray datasets (9) and statistical significance of recurrent differential expression patterns (4,5). We will also systematically curate biological archival information needed to interpret the expression patterns, e.g. known genes related to aging, more functional annotation information (e.g. BIOCARTA and KEGG pathways), gene sequences, transcription regulation information and protein–protein interactions. By integrating such data, users will be able to better interpret the results derived from integrative microarray analysis results, to assign unknown genes to aging-related pathways and to predict transcriptional regulation. As GAN is continuously expanding, this system is designed in such a way that maintenance and scaling up are easy to implement. We hope that GAN can significantly facilitate the re-use of the vast amount of existing aging-related microarray data and reduce the necessity to re-generate data.

ACKNOWLEDGEMENT

The work was supported by a pilot grant from the Seaver foundation, the NIH Grants R01GM074163, P50HG002790, P01AG14751, R01AG13499, the NSF grant 0515936 and R01AG23173 (Carl Cotman PI, subcontract to CE Finch). Funding to pay the Open Access publication charges for this article was provided by the Seaver pilot grant.

Conflict of interest statement. None declared.

REFERENCES

1. McCarroll,S.A., Murphy,C.T., Zou,S., Pletcher,S.D., Chin,C.S., Jan,Y.N., Kenyon,C., Bargmann,C.I. and Li,H. (2004) Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nature Genet*., **36**, 197–204.
2. Choi,J.K., Yu,U., Kim,S. and Yoo,O.J. (2003) Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, **19**(Suppl. 1), I84–I90.
3. Lee,H.K., Hsu,A.K., Sajdak,J., Qin,J. and Pavlidis,P. (2004) Coexpression analysis of human genes across many microarray datasets. *Genome Res*., **14**, 1085–1094.
4. Rhodes,D.R., Barrette,T.R., Rubin,M.A., Ghosh,D. and Chinnaiyan,A.M. (2002) Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res*., **62**, 4427–4433.
5. Rhodes,D.R., Yu,J., Shanker,K., Deshpande,N., Varambally,R., Ghosh,D., Barrette,T., Pandey,A. and Chinnaiyan,A.M. (2004) Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc. Natl Acad. Sci. USA*, **101**, 9309–9314.
6. Zhou,X.J., Kao,M.C., Huang,H., Wong,A., Nunez-Iglesias,J., Primig,M., Aparicio,O.M., Finch,C.E., Morgan,T.E. and Wong,W.H. (2005) Functional annotation and network reconstruction through cross-platform integration of microarray data. *Nat. Biotechnol*., **23**, 238–243.
7. Rhodes,D.R., Yu,J., Shanker,K., Deshpande,N., Varambally,R., Ghosh,D., Barrette,T., Pandey,A. and Chinnaiyan,A.M. (2004) ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia*, **6**, 1–6.
8. Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.
9. Hu,H., Yan,X, Huang,Y., Han,J. and Zhou,X.J. (2005) Mining coherent dense subgraphs across massive biological networks for functional discovery. *Bioinformatics*, **21**(Suppl. 1), i213–i221.