Nucleic Acids Research, 2005, Vol. 33, Database issue ***D533–D536***
doi:10.1093/nar/gki117

Cancer gene expression database (CGED): a database for gene expression profiling with accompanying clinical information of human cancer tissues

Kikuya Kato[1,2,*], Riu Yamashita[2,3], Ryo Matoba[2], Morito Monden[4], Shinzaburo Noguchi[5], Toshihisa Takagi[6] and Kenta Nakai[3]

[1]Osaka Medical Center for Cancer and Cardiovascular Diseases, 1-3-2 Nakamichi, Higashinari-ku, Osaka 537-8511, Japan, [2]Taisho Laboratory of Functional Genomics, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0101, Japan, [3]Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan, [4]Department of Surgery and Clinical Oncology and [5]Department of Surgical Oncology, Graduate School of Medicine, Osaka University, 2-2 Yamadaoka, Suita, Osaka 565-0871, Japan and [6]Department of Computational Biology, Graduate School of Frontier Sciences, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8561, Japan

Received August 13, 2004; Revised and Accepted October 20, 2004

ABSTRACT

Gene expression profiling of cancer tissues is expected to contribute to our understanding of cancer biology as well as developments of new methods of diagnosis and therapy. Our collaborative efforts in Japan have been mainly focused on solid tumors such as breast, colorectal and hepatocellular cancers. The expression data are obtained by a high-throughput RT–PCR technique, and patients are recruited mainly from a single hospital. In the cancer gene expression database (CGED), the expression and clinical data are presented in a way useful for scientists interested in specific genes or biological functions. The data can be retrieved either by gene identifiers or by functional categories defined by Gene Ontology terms or the Swiss-Prot annotation. Expression patterns of multiple genes, selected by names or similarity search of the patterns, can be compared. Visual presentation of the data with sorting function enables users to easily recognize of relationships between gene expression and clinical parameters. Data for other cancers such as lung and thyroid cancers will be added in the near future. The URL of CGED is http://cged.hgc.jp.

INTRODUCTION

Gene expression profiling of human cancer tissues is an emerging approach based on genomics, whose main focus is the characterization of differences among cancer tissues from individual patients (1,2). The results are intended to be applied to therapeutics, especially for diagnosis, based on heterogeneity in gene expression among tumor tissues of individual patients.

We developed the cancer gene expression database (CGED) to present gene expression data and clinical information on human cancer tissues. The data in CGED were obtained through collaborative efforts of Nara Institute of Science and Technology and Osaka University School of Medicine to identify genes of clinical importance. Although the main objective of the project is to identify genes for diagnosis or potential therapeutic targets, the database provides a rich source of information for basic research.

The data in this database are quite unique both in analytical and clinical aspects. First, all the data were obtained using adaptor-tagged competitive PCR (ATAC-PCR) (3,4), an advanced version of quantitative RT–PCR. Generally, RT–PCR produces data of better quality than those based on hybridization-based techniques (5). Second, tissue samples were obtained mainly from a single hospital. This eliminates deterioration of clinical data by a difference in medical practice, commonly found in data collected from multiple hospitals. In addition, the racial background is the same, because all the patients are oriental. This feature also eliminates potential effects of a racial background on gene expression. The high quality of the expression and clinical data is demonstrated by the successful identification of prognostic genes in breast (6), colorectal (7) and hepatocellular cancers (8).

In CGED, considering the characteristics of database systems, we present expression and clinical data in a way useful for individual scientists interested in specific subjects of

*To whom correspondence should be addressed. Tel: +81 6 6972 1181 (ext. 4105, 4304); Fax: +81 6 6973 5691; Email: katou-ki@mc.pref.osaka.jp cancer molecular biology. CGED enables experimental scientists to access the information of interest, without handling expression data matrices.

GENERAL FEATURES OF GENE EXPRESSION DATA AND CLINICAL INFORMATION

Most tissue samples were obtained from patients in the Osaka University Medical School Hospital or the Osaka Medical Center for Cancer and Cardiovascular Diseases. The studies were approved by the Institutional Review Board, Osaka University Medical School or the Osaka Medical Center, and written informed consent was obtained from each patient.

The flow chart of the data production is summarized in Figure 1. We at first made 3′ directed cDNA libraries using mRNA purified from target cancer tissues, and performed EST sequencing. Inserts of the libraries were 3′ end cDNA fragments generated by MboI digestion (9). The number of reads varied from 4000 to 20 000. Then, we selected genes for assay by descending order of abundance. In addition, we selected genes related to each cancer by a literature survey, and subjected them for ATAC-PCR assay. These processes guaranteed that all assayed genes were expressed in the target cancer tissue. Gene expression data by ATAC-PCR were obtained as relative expression levels against control samples. We usually used mixtures of tumor tissues from several patients as controls to detect any slight differences among tumor tissues from different patients. All the data were logarithmically converted after normalization by the median of samples, and then by that of genes. So far the following projects have been finished, and data deposited in CGED.

Breast cancer. Analyzed samples: 98 primary breast tumor tissues and 10 normal breast tissues; number of genes analyzed, 2412. Breast cancer has several molecular markers correlating with malignancy, including estrogen and progesterone receptors, erbB2 and p53. These markers are also included in clinical parameters.

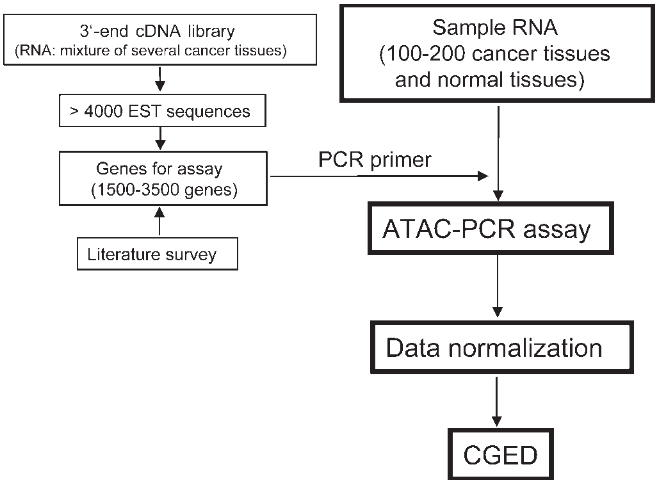


Figure 1. Outline of the gene expression data production. Genes for assays were selected from a pool of genes appearing in the EST collection (the left part of the figure). The data production pipeline from tissues to the database is shown in the right part.

Colorectal cancer. Analyzed samples: 100 colorectal cancer tissues and 11 normal colon tissues; number of genes analyzed, 1536.

Hepatocellular cancer. Analyzed samples: 120 hepatocellular cancer tissues, 86 non-tumor (adjacent to tumor) tissues and 32 normal liver tissues; number of genes analyzed, 3072. The main etiological factor of hepatocellular cancer is virus infection, i.e. hepatitis B virus (HBV) and hepatitis C virus (HCV). Records of infection with these viruses are included in clinical parameters. Because of the limitation of samples displayed by mosaic plots, we randomly selected 100 samples for CGED.

Esophageal cancer. Analyzed samples: 160 esophageal cancer tissues; number of genes analyzed, 1904. Clinical records include stage classification, depth of invasion, number of metastatic lymph nodes.

DESCRIPTION OF THE DATABASE

The objective of this database is not to provide full access to our data matrices (the entire data matrix can be downloaded from http://genome.mc.pref.osaka.jp/CGIP.html), but to present information useful for scientists interested in specific genes and biological functions. Those interested in identifying diagnostic or target genes should instead use entire data matrices. The structure of web pages in CGED is shown in Figure 2. Pages are marked by Roman numbers, which are referred to in the following description of CGED.

QUERY

The top page (I in Figure 2) of CGED has two types of data retrieval systems. One is a conventional data retrieval system based on gene names and gene identifiers such as GenBank accession number. In addition, data of a group of genes can be retrieved using Gene Ontology (GO) terms (10) or words in functional annotation of Swiss-Prot (11). The latter function enables users to select a group of genes belonging to a specific functional category. For example, when ‘oncogene’ is entered as a keyword, a list of genes in the CGED, categorized as oncogene-related, either by GO or the Swiss-Prot functional annotation, is displayed on the search result page (II in Figure 2). Each gene has a link to its information page (V in Figure 2) including various identification numbers such as RefSeq, GenBank, UniGene and LocusLink with links to these websites. It also includes GO terms and functional annotation of the Swiss-Prot. Expression patterns of genes can be displayed in two ways: comparison of expression patterns of selected genes and expression pattern similarity search.

COMPARISON OF EXPRESSION PATTERNS

Genes and a cancer type should be selected from check boxes and from the menu on the right-hand side of the search result page (II), respectively. After clicking the display button in ‘selected genes’ in the menu on the right, gene expression patterns are displayed as a mosaic plot, which is a popular presentation method of microarray data (Gene Expression and Clinical Data Display, III in Figure 2) (12). This page is accompanied by a list of displayed genes in another window

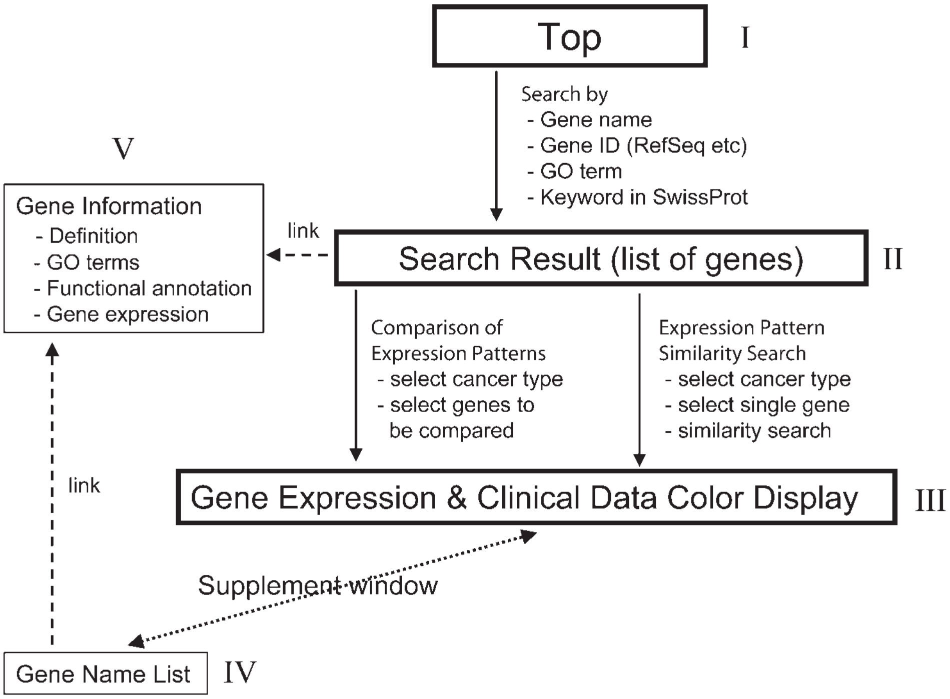
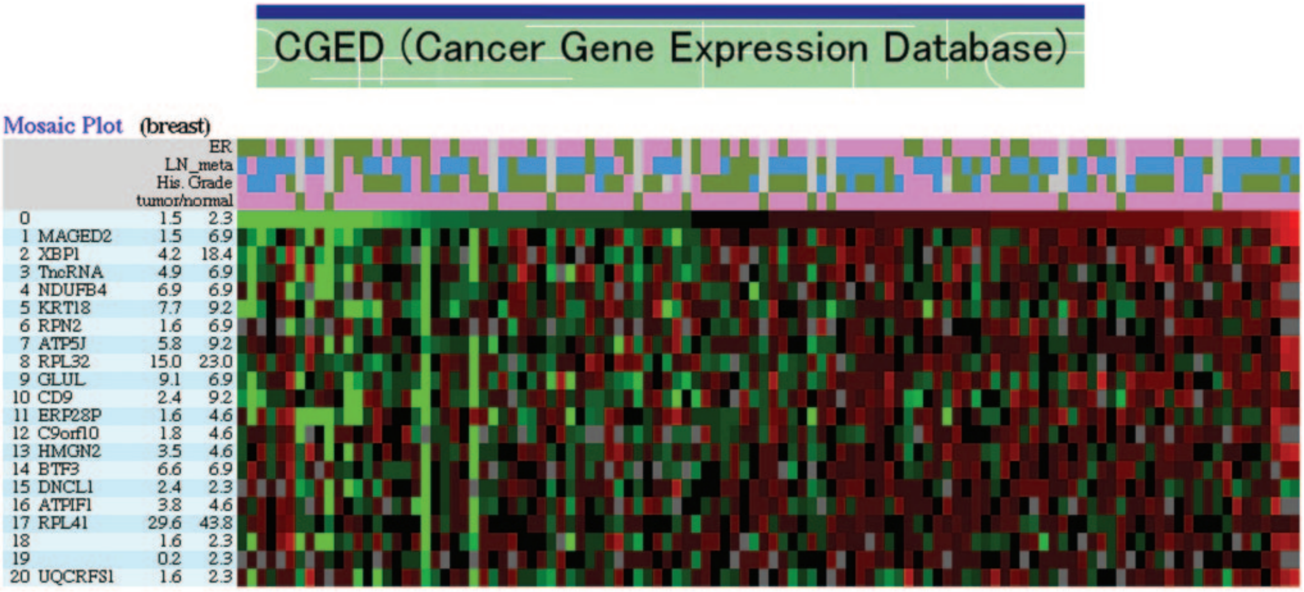


Figure 2. Outline of CGED. Web pages constituting CGED are marked by Roman numbers. In the top page (I), the database is either searched by gene identifiers or by keywords. From the search result page (II), expression patterns of multiple genes are displayed and compared. From the same page, by selecting a single gene, genes with a similar expression pattern are searched and displayed (III). Page III is accompanied by a gene name list (IV). Each gene listed in pages II and IV has a link to its information page (V).


CGED (Cancer Gene Expression Database)

Mosaic Plot (breast)

ER
LN_meta
His. Grade
tumor/normal

0		1.5	2.3
1	MAGED2	1.5	6.9
2	XBP1	4.2	18.4
3	TncRNA	4.9	6.9
4	NDUFB4	6.9	6.9
5	KRT18	7.7	9.2
6	RPN2	1.6	6.9
7	ATP5J	5.8	9.2
8	RPL32	15.0	23.0
9	GLUL	9.1	6.9
10	CD9	2.4	9.2
11	ERP28P	1.6	4.6
12	C9orf10	1.8	4.6
13	HMGN2	3.5	4.6
14	BTF3	6.6	6.9
15	DNCL1	2.4	2.3
16	ATPIF1	3.8	4.6
17	RPL41	29.6	43.8
18		1.6	2.3
19		0.2	2.3
20	UQCRFS1	1.6	2.3

Figure 3. Color display of the gene expression patterns and clinical information. Clinical information (top part of the graph) and gene expression data (bottom part of the graph) of 98 breast cancer tissues and 10 normal tissues are represented by small color boxes. A legend of color boxes for clinical information is attached to each graph (not shown in this figure). Gene expression levels are schematically represented by a color gradient from light red to light green: light red, high level of gene expression; black, middle level; light green, low level. The details of gene names are shown in another window (not shown in this figure). Figures next to symbol names represent transcript abundance by EST frequencies in all of our cDNA libraries (left) and in the library of the selected cancer (right). The graph is a result of searching for genes whose expression patterns are similar to that of zinc-alpha2-glycoprotein (the top gene numbered 0) and the following sort by its expression levels. Zinc-alpha2-glycoprotein is one of the prognostic genes of breast cancer (6), and its expression level is correlated with estrogen receptor (ER).

(IV in Figure 2). Each gene has a link to its information page (V). This list also provides information on transcript abundance by EST frequencies in all of our cDNA libraries and in the library of the selected cancer.

GENE EXPRESSION PATTERN SIMILARITY SEARCH

At first, a gene of interest and a cancer type are selected by clicking a radio button and from the menu on the right of the search result page (II). Then, by clicking the display button in 'similarity search', CGED displays a predefined number of genes by the order of similarity of expression pattern (III). Currently, Euclidean distance is used as the measure of similarity. The page displaying gene expression patterns (III) is also accompanied by a list of displayed genes (IV).

SORTING BY GENE EXPRESSION OR CLINICAL PARAMETERS

The results of multiple gene display and the similarity search (III) can be sorted. The samples can be sorted either by the order of gene expression of a selected gene or by clinical parameters such as the presence and absence of metastasis. Sorting is useful for easy recognition of relationships between gene expression and clinical parameters. An example is shown in Figure 3.

COMMENTS AND FUTURE PLANS

Gene expression data matrices of human cancers are usually distributed as flat files. However, it is not an easy task for experimental biologists to analyze the large data matrix by themselves. The aim of the CGED is to bypass the need of analyzing the complex data matrix, and to enable direct access to the information the individual scientist is interested in.

Data analysis of glioblastoma and thyroid cancers is in progress. In addition, expression data acquisition of lung cancer will be finished within the fiscal year 2004. These data will be deposited in CGED after approval by the local committee consisting of participants of the project. Some of the patients recruited for these studies are being clinically followed up, and the clinical information will be regularly updated.

ACKNOWLEDGEMENTS

We thank Dr Kyoko Koizumi, Ms Shizuko Muro, Drs Yukinori Kurokawa and Kazuyuki Oda for managing the projects. We also thank Ms Noriko Ueno, Satoko Maki-Kinjo, Keiko Miyaoka-Ikegami and Mihoko Yoshino for their technical assistance. This work was supported by a Grant-in-Aid for Scientific Research on the Priority Area 'Genome Information Science' from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

REFERENCES

1. Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A. *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
2. Alizadeh,A.A., Eisen,M.B., Davis,R.E., Ma,C., Lossos,I.S., Rosenwald,A., Boldrick,J.C., Sabet,H., Tran,T., Yu,X. *et al*. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.
3. Kato,K. (1997) Adaptor-tagged competitive PCR: a novel method for measuring relative gene expression. *Nucleic Acids Res*., **25**, 4694–4696.
4. Matoba,R., Kato,K., Kurooka,C., Maruyama,C., Sakakibara,Y. and Matsubara,K. (2000) Correlation between gene functions and developmental expression patterns in the mouse cerebellum. *Eur. J. Neurosci*., **12**, 1357–1371.
5. Holland,M.J. (2002) Transcript abundance in yeast varies over six orders of magnitude. *J. Biol. Chem*., **277**, 14363–14366.
6. Iwao,K., Matoba,R., Ueno,N., Ando,A., Miyoshi,Y., Matsubara,K., Noguchi,S. and Kato,K. (2002) Molecular classification of primary breast tumors possessing distinct prognostic properties. *Hum. Mol. Genet*., **11**, 199–206.
7. Muro,S., Takemasa,I., Oba,S., Matoba,R., Ueno,N., Maruyama,M., Yamashita,R., Sekimoto,M., Yamamoto,H., Nakamori,S. *et al*. (2003) Identification of expressed genes linked to the malignancy of human colorectal carcinoma by parametric clustering of quantitative expression data. *Genome Biol*., **4**, R21.
8. Kurokawa,Y., Matoba,R., Takemasa,I., Nagano,H., Dono,K., Nakamori,S., Umeshita,K., Sakon,M., Ueno,N., Oba,S. *et al*. (2004) Molecular-based prediction of early recurrence in hepatocellular carcinoma. *J. Hepatol*., **41**, 284–291.
9. Matoba,R., Kato,K., Saito,S., Kurooka,C., Maruyama,C., Sakakibara,Y. and Matsubara,K. (2000) Gene expression in mouse cerebellum during its development. *Gene*, **241**, 125–131.
10. The Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Res*., **11**, 1425–1433.
11. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nuceic Acids Res*., **31**, 365–370.
12. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.