Significant underexpression of some genes, such as those marked as Group α in Fig. 2A, is observed in both poorly and moderately differentiated samples. Such genes include SLC22A1 (solute carrier family 22, member 1), CYP2A6 (cytochrome P450, subfamily IIA, polypeptide 6), CYP2A7 (cytochrome P450, subfamily IIA, polypeptide 7), ALB (albumin), and FETUB (fetuin B), etc. Genes in Group β show a tendency of increased expression in the order of poorly, moderately, and well-differentiated tumors. This group includes ADH1A (alcohol dehydrogenase 1A), HFL3 (H factor (complement)-like 3), and AFM (afamin), etc. For some genes, significant variations in expression in moderately differentiated samples are observed.

Many of these genes are related to the function of hepatocyte cells. Using the Onto-Express software [24] based on the Gene Ontology database, we confirmed statistically significant overrepresentation of functional categories like immune response ($N = 9$, $P < 0.0002$), oxidoreductase activity ($N = 7$, $P < 0.0002$), lipid transporter activity ($N = 4$, $P < 0.00001$), and so on (Supplementary Fig. 7A).

A small number of genes marked as Group γ are highly expressed in poorly differentiated samples, but not in well-differentiated samples or normal livers. These genes are tissue-specific genes for fetal liver. Among these genes are AFP (alpha-fetoprotein), FACL4 (fatty acid–coenzyme A ligase, long-chain 4), MKI67 (antigen identified by monoclonal antibody Ki-67), and MCM7 (MCM7 minichromosome maintenance-deficient 7). Among these genes, AFP and Ki-67 are known markers whose high expression is related to poor prognosis [48,49].

Note that the 64 transcripts shown in Fig. 2A are selected through two criteria: they are specifically expressed in normal liver, and their expression varies among HCC samples. These 64 transcripts represent only a small part of 175 liver-specific genes. There are other liver-specific genes that are still highly expressed even in poorly differentiated samples. As shown in Supplementary Fig. 8, even poorly differentiated HCCs do not lose completely their liver-specific expression of many genes. This observation gives us some justification for using tissue-specific expression signatures in the interpretation of expression data to address some other questions such as the identification of the origin of tumors. This will be discussed in the following sections.

Neuronal and glial-specific expression signatures in brain tumors

Next, we study the expression of brain-specific genes in embryonal tumors of the central nervous system (CNS). We use dataset A of Pomeroy et al. [25], which consisted of medulloblastoma (MD, $N = 10$), supratentorial primitive neuroectodermal tumor (PNET, $N = 6$), CNS atypical teratoid/rhabdoid tumor (CNS AT/RT, $N = 5$), renal and extrarenal AT/RT ($N = 5$), nonembryonal malignant glioma (MG, $N = 10$), and normal cerebella ($N = 4$). From the dataset, the original study reports that medulloblastomas are molecularly distinct from other brain tumors.

From our list of brain-specific genes, we retrieved data from this dataset and performed unsupervised clustering. As shown in Fig. 3A, the samples are divided into two major groups. The glioma and medulloblastoma group shows high expression of many brain-specific genes, which is not observed in the PNET and AR/AT groups. With our gene subset, no difference is observed between CNS and non-CNS AR/AT tumors. Malignant gliomas and medulloblastomas are further distinguished by their high expression of two clusters of genes marked as Cluster α and Cluster β, respectively. Included in Cluster α are genes such as GFAP (glial fibrillary acidic protein) and OLIG2 (oligodendrocyte lineage transcription factor 2), which are known to be markers of glia cells. On the other hand, genes in Cluster β are mainly neuron related. For Cluster β genes, functional analysis with Onto-Express software [24] also revealed statistically significant enrichment of genes with functions related to transmission of nerve impulses ($N = 6$, $P < 0.00005$), neurophysiological processes ($N = 6$, $P < 0.003$), and neurontransmitter transport ($N = 2$, $P < 0.002$) as shown in Supplementary Fig. 7B. Therefore, our clustering results suggest that glioma and medulloblastoma carry expression signatures of glia and neuron cells, respectively.

For further confirmation, we plotted the expression pattern of these genes in different parts of the normal nervous system (Fig. 3B). Clearly, genes in Cluster α are highly expressed in corpus callosum and spinal cord while genes in Cluster β are specifically expressed in thalamus, cerebellum, hippocampus, and amygdala. As spinal cord and corpus callosum are enriched in glias and contain less nerons, this result clearly indicates that gliomas carried a glia-specific expression signature and medulloblastoma show neuronal origin, which is in agreement with the current understanding of the origins of these tumors. Therefore, comparative analyses of normal and cancer expression profiles are useful for studying the cell lineage of tumors.

Breast tumors with two distinct types of differentiation

To study the expression of breast-specific genes in breast cancer, we started with a list of 57 genes that are breast specific or breast selective (highly expressed in several tissues including the breast). From this list, we selected 26 genes that show significant variation in expression among the 21 breast cancer samples in the dataset of Su et al. [21]. The expression of these genes in our normal tissue database is shown in Fig. 4A. Among these genes are several keratins (KRT14, KRT15, and KRT17) that are highly expressed in the skin and breast. Another important gene in the list is estrogen receptor 1 (ESR1) that is defined as a tissue-selective gene for the breast and uterus. In addition to these 26 genes, we intentionally included three more

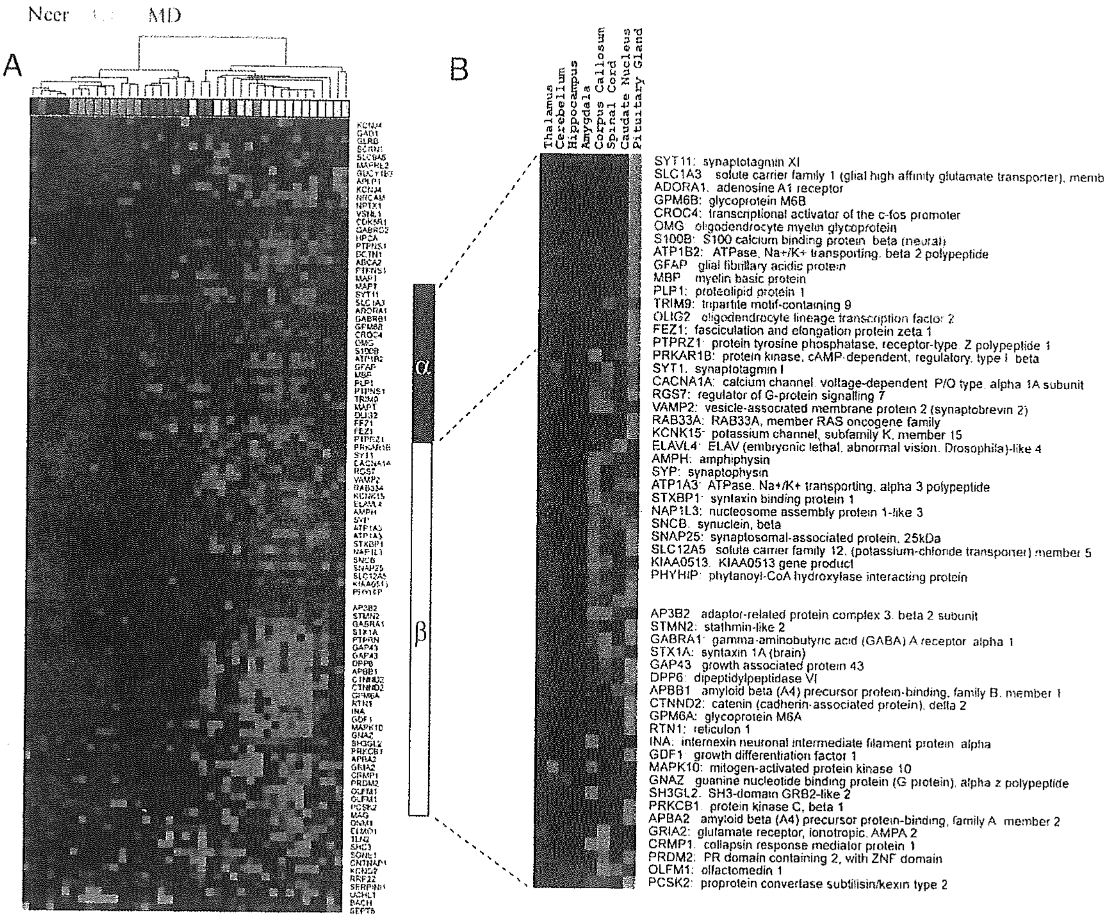


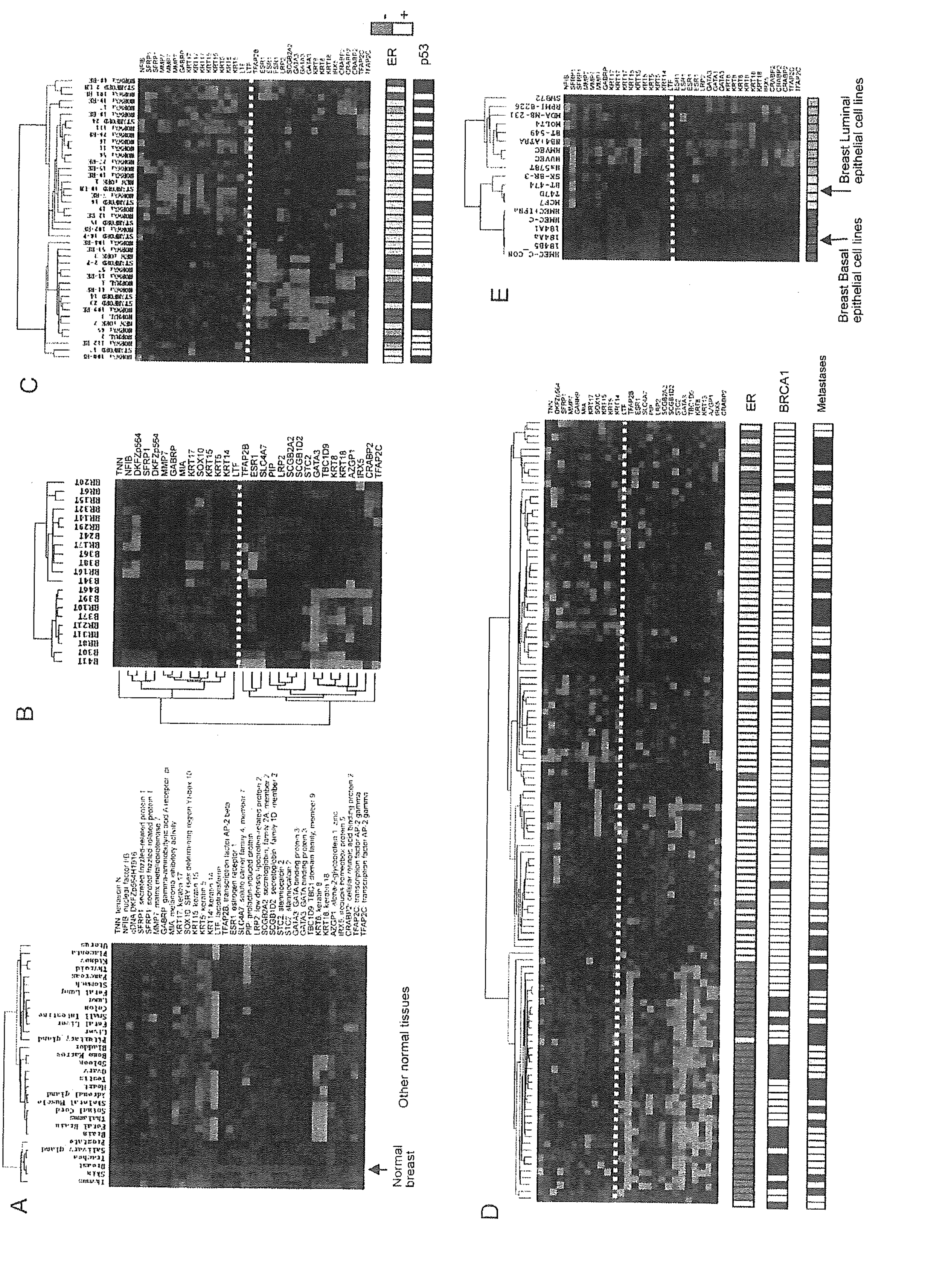
Fig. 3. (A) Expression of brain-specific genes in various types of brain tumors. The samples are divided into two major groups, a glioma and medulloblastoma group, and a PNET and AR/AT group. The first group shows higher expression of many brain-specific genes while the second does not. Within the first group, malignant gliomas and medulloblastomas are characterized by their high expression of two clusters of genes marked as Cluster α and Cluster β, respectively. (B) The expression pattern of Group α and β genes in different parts of the normal nervous system. Genes in Cluster α are highly expressed in corpus callosum and spinal cord while genes in Cluster β are specifically expressed in thalamus, cerebellum, hippocampus, and amygdala.

keratin genes, KRT5, KRT8, and KRT18 as markers for different epithelial cells [53]. Although they are not in our list of breast-specific or breast-selective genes, their expression levels are also higher in the breast than in most other tissues (Fig. 4A) and are added for the discussion on tumor origin.

We then performed clustering analysis of these 29 genes in 21 breast cancer samples from the dataset of Su et al. [21]. The result is shown in Fig. 4B. Surprisingly, these genes form two groups. Overexpression of these two groups in cancer samples seems to be mutually exclusive. This is quite different from the univariant behavior of liver-specific genes in liver cancers, in which the expression levels of liver-specific genes are increased as one group from poorly differentiated to well-differentiated tumors. Breast tumors seem to have two distinct types of differentiation.

This interesting expression pattern is confirmed by two larger breast cancer datasets shown in Figs. 4C and D. In these two figures, hierarchical clustering of the samples is performed while the genes are arranged in the same order as in Fig. 4B (same for Figs. 4A and 4E). Note that the dataset of Perou et al. [26] shown in Fig. 4C is obtained with cDNA microarrays while the dataset of van't Veer et al. [2] in Fig. 4D is based on a kind of oligonucleotide microarray that is different from the Affymetrix GeneChip used by Su et al. in Fig. 4B. Moreover, patient samples are

Fig. 4. Expression of breast-specific genes in breast cancer. (A) Expression pattern of these genes in normal tissues. (B) Hierarchical clustering analysis of the expression data of these genes in 21 breast cancer samples from the dataset of Su et al [21]. Note that the resultant order of genes is used throughout this figure. (C) Expression of these genes in a breast cancer dataset of Perou et al. [26]. In the color bar for clinical ER status, blue indicates ER positive and red indicates ER negative. In the color bar for p53 mutation, black and white indicate the presence and absence of p53 mutations, respectively. In both color bars, gray indicates that the information is not available. (D) Expression of these genes in the dataset of Van'T Veer et al. [2]. ER status and mutations of BRCA1 and status of distant metastases are indicated at the bottom using the same coloring scheme as in C. (E) Expression of these genes in breast basal epithelial cell lines (red), breast luminal epithelial cell lines (blue), and other types of cell lines (grey). Data are from Ref. [26].


A
Normal breast
Other normal tissues
B
TNN
NFIB
DKFZp564
SFRP1
DKFZp564
MMP7
GABRP
MIA
KRT17
SOX10
KRT15
KRT5
KRT14
LTF
TFAP2B
ESR1
SLC4A7
PIP
LRP2
SCGB2A2
SCGB1D2
STC2
GATA3
TBC1D9
KRT8
KRT18
AZGP1
IRX5
CRABP2
TFAP2C
C
ER
p53
-
+
D
ER
BRCA1
Metastases
E
Breast Basal epithelial cell lines
Breast Luminal epithelial cell lines

collected by three different laboratories from different populations. Despite these differences, the same pattern is observed in three independent datasets. All these data suggest that breast cancers could exhibit two types of differentiation.

To gain insight into the two types of cancers, the expression of these 29 genes in various cell lines is shown in Fig. 4E (data from Ref. [26]). It is found that the two types of gene expression pattern correspond well to breast basal epithelial cell lines (HMEC and 184Aa) and luminal epithelial cell lines (MCF7, T47D, BT-474, and SK-ER-3), respectively. Such expression patterns are not observed in other types of cell lines such as those derived from breast carcinosarcoma (Hs578T), shown on the right side of Fig. 4E. This is also consistent with the expression pattern of several markers for different cell types in the breast. Keratins 5/6 and 17 are conventional markers for breast basal epithelial cells while keratins 8 and 18 are markers for breast luminal epithelial cells. Therefore, breast cancer samples can exhibit basal-like differentiation or luminal-like differentiation.

Combined with clinical information given at the bottom of these figures, we observed that a basal-like expression pattern is usually seen in ER− breast cancers while a luminal-like expression pattern is mostly observed in ER+ breast cancers. In addition, there are some ER− samples that show neither basal nor luminal differentiation, which are shown on the right sides of Figs. 4C and D. Some of them are characterized to overexpress erbB2 [26,27]. While the basal-like group is homogeneous, luminal like samples are heterogeneous and might be further divided into several subtypes [27]. Our result agrees with previous report that gene expression patterns of breast cancer are divided into two big clusters in association with ER status [2,26]. ER+/luminal subtypes of breast cancers usually have a good prognosis while those with an ER−/basal-like expression pattern are more invasive. This has been observed repeatedly in several studies (see s.1b in Ref. [26], Fig. 1a in Ref. [2], and Ref. [32]).

For many of the genes shown in Fig. 4, differential expression in ER+ and ER− tumors has been reported previously [2,26]. Our results linked such observations with their expression pattern in normal tissues: many of the differentially expressed genes between subtypes of breast tumors are highly expressed in normal breast. It is surprising that a small set of breast-specific genes seems to contain genes highly expressed in both ER+ and ER− tumors, in a seemingly unbiased manner.

The normal breast epithelium consists of a luminal epithelial layer and a basal myoepithelial layer. RNA samples for the normal breast are extracted from this heterogeneous tissue as a mixture of these microscopic organizations. Hence both basal and luminal cells contribute to the tissue specificity observed in the gene expression pattern. Breast-specific genes actually contain basal-specific and luminal-specific genes as shown by the cell line data in Fig. 4E. Since breast tumors could display basal- or luminal-like differentiation, we could separate these two types of tumors with a small set of breast-specific genes. This is a phenomenological explanation for the expression pattern in Fig. 4.

Our observation seems to suggest that these two types of breast tumors might originate from different cell types within the normal breast epithelium. But it might also be possible that they all come from the same myoepithelial cells and some later undergo a drastic change in global gene expression during progression of dedifferentiation. Further discussion is available in the Supplementary Information. Whatever the molecular mechanism, our analysis revealed that breast tumors exhibit two types of differentiation that could be related to two types of epithelial cells within the normal breast.

Heterogeneity of lung cancers

The following two sections deal with lung cancer, which is more heterogeneous than liver and breast cancers discussed above. We reanalyzed a dataset of lung cancers (N = 186) and normal lung (N = 17) [28]. The cancer samples are histologically divided into lung adenocarcinomas (AD, N = 127), squamous cell lung carcinomas (SQ, N = 21), pulmonary carcinoids (COID, N = 20), small-cell lung carcinomas (SCLC, N = 6), and other adenocarcinomas (N = 12). For each sample, gene expression data of 12,600 transcripts are also obtained with U95A oligonucleotide arrays. This array covers 22 of the 32 lung-specific genes identified in the present study. About 77% (17/22) of these transcripts are called present in all 17 normal lung samples. In contrast, most of them (68%) are called absent in at least 7 of the 8 normal liver samples noted in theprevious section [28]. A similar percentage (64%) of these genes are absent in at least 40 of the 50 normal prostate samples in another microarray dataset [29].

Because there are so few lung-specific transcripts and lung tumors are known to have greater heterogeneity, expression data are retrieved from this dataset for our list of 2503 tissue-specific and tissue-selective genes associated with all tissue types. Hierarchical clustering is performed after variation filtering. From the result shown in Fig. 5, we noted several features. First of all, high expression of lung-specific genes is observed in normal lung and some adenocarcinomas. Expression of those genes varies among adenocarcinomas, indicating degree of differentiation, as discussed in the case of liver cancer.

Another feature is high-level expression of skin-specific genes in SQ samples. Such genes include galectin 7 (LGALS7), desmoglein 3 (DSG3), plakophilin 1 (PKP1), and keratin 16 (KRT16). KRT16 is a member of keratin family known as markers for squamous tumors. When analyzed with Onto-Express software, this gene list shows strong correlation with ectoderm development (N = 5, P = 0.0), and contains many cytoskeleton genes (N = 6, P <

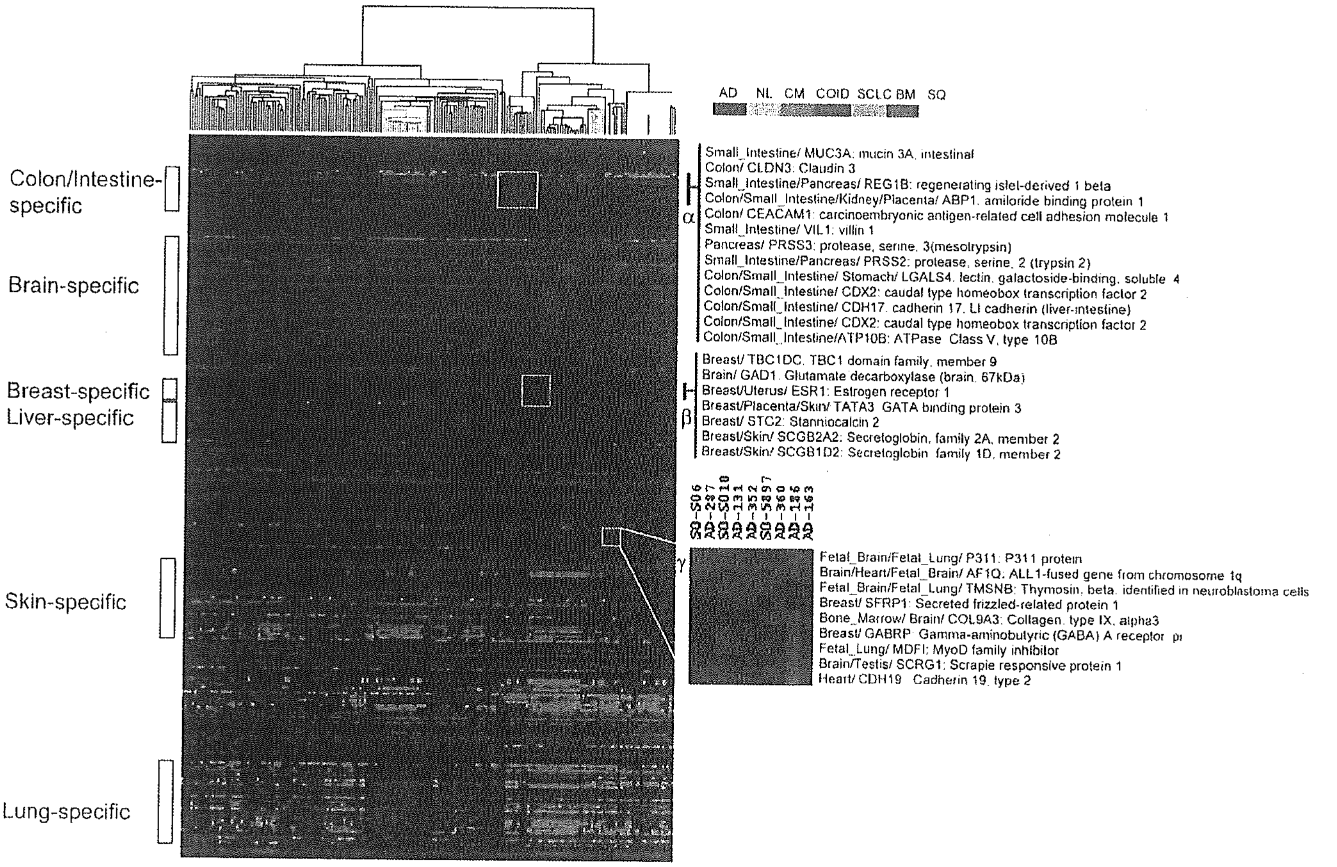


Fig. 5. Clustering analysis of a dataset of lung cancer using all 2503 of the tissue-specific/selective genes. Branches are marked according to clinical diagnosis: normal lung (NL), gray; lung adenocarcinoma (AD), black; squamous cell lung carcinomas (SQ), yellow; pulmonary carcinoids (COID), blue; and small-cell lung carcinomas (SCLC), green. Some lung adenocarcinoma samples are diagnosed as colon metastasis (CM, pink), or breast metastasis (BM, red). Gene groups: α, a set of colon/intestine-specific genes highly expressed in CM samples; β, six breast-specific genes highly expressed in a BM sample; γ, two breast-specific genes and some genes highly expressed in fetal tissues. As marked at the left side, this figure also shows higher expression of brain-specific genes in COID and SCLC samples and skin-specific genes in SQ samples. In addition, one AD sample shows very high expression of dozens of liver-specific genes.

0.00012). The high expression of skin-related genes in SQ samples is reasonable as this type of lung tumor is believed to originate from bronchial epithelium.

Similarly, high-level expression of brain-specific genes is observed in COID samples. Typical genes include GRIA2 (glutamate receptor, ionotropic, AMPA 2), SLC4A3 (solute carrier family 4, anion exchanger, member 3), SYT1 (synaptotagmin I), SNAP25 (synaptosomal-associated protein, 25 kDa), and APLP1 (amyloid beta (A4) precursor-like protein 1), etc. Part of these genes, such as SYT1, SNAP25 and APLP1, are also highly expressed in SCLC. This gene cluster overlaps with the Cluster α in Fig. 3; functional analysis with Onto-Express also confirmed strong link to neurogenesis ($N = 3$, $P < 0.00034$). Such observations agree with general understanding that SCLC and COID are neuroendocrine tumors.

In summary, we observed higher expression of lung-specific genes in AD cancers, skin-specific genes in SQ cancers, and brain-specific genes in SCLC and COID. These expression signatures reveal the origin and cell lineage of these tumors, which illustrated the usefulness of studying tissue-specific gene expression in cancers.

Primary sites of metastatic cancer

We identified a set of colon/intestine-specific genes that are highly expressed in a cluster of 12 samples (Group α in Fig. 5). Clinical and histological information shows that 7 of these samples are metastases of colon cancer. Therefore, this cluster may represent metastatic cancer from the colon.

In the original study, it is found that these samples form a cluster with quite different expression signatures from other lung cancer samples and that these tumors express some genes (such as galectin-4, cadherin 17, and *c-myc*) that are known to be overexpressed in colon carcinoma. These authors concluded that this cluster of 12 samples may be colon metastasis. In our study, the high-level expression of dozens of colon/intestine-specific genes lead us to a similar conclusion. While their conclusion is based on reported markers from the literature, ours solely makes use of a gene expression database of normal tissues. So our approach might be helpful for the diagnosis of metastatic cancer from organs that are not as well-studied as colon cancer.

We also observed overexpression of several liver or fetal liver-specific genes in one lung tumor (AD368). This is also noted in the original study, as some of these genes such as albumin are associated with liver. Although this sample is not clinically identified as metastasis, it carries a liver-specific expression signature, which can be clearly seen in the middle of Fig. 5.

Metastatic cancers from some other organs could be difficult to identify. For example, the dataset contains one sample (AD352) that is diagnosed as breast metastasis and another three samples (AD163, AD186, and AD172) as probably breast metastasis. Of these four samples, only one (AD163) showed high expression of six breast-specific genes including ESR1. These genes are marked as Group β in Fig. 5. The other three samples do not have such an expression signature. However, two of them (AD352 and AD186) are found in a cluster of eight samples, characterized by high expression of a group of nine genes (Group γ), including two breast-specific genes, SFRP1 and GABRP; this indicates a weak breast-specific expression signature. This group also includes several genes that are highly expressed in fetal tissues: p311, AF1Q (ALL1-fused gene from chromosome 1q), TMSNB (Thymosin, beta), and MDFI (MyoD family inhibitor). This seems to suggest that these tumors are more aggressive and that they might be metastasis from distant organs.

A closer look at the genes in Groups β and γ revealed something interesting. In the previous section we show that breast tumors could have two distinct differentiations. In fact, all of the six breast-specific genes in Group β belong to those given in the lower part of Fig. 4B, characteristic of a luminal/ER+ tumor type. Thus AD163 is probably metastasis of a luminal-like/ER+ breast cancer. On the other hand, Group γ genes include two breast-specific genes, SFRP1 and GABRP, which are characteristic of basal-like/ER− tumors. Therefore the samples AD352 and AD186 might be from this tumor subtype. Because the expression pattern of the two subtypes of breast tumors are quite different, AD163 are found in a different branch of the clustering tree in Fig. 5. This might explain why it is difficult for original authors [28] to identify such breast metastasis.

For confirmation, we constructed a set of marker genes based on results shown in Figs. 4 and 5. Markers for two types of breast cancers are the same as in Fig. 4, while those for colon and liver cancers are selected from the highlighted regions of Fig. 5. In addition, 19 markers for lung adenocarcinoma are taken from Ref. [21]. As shown in Fig. 6A, these genes are specifically expressed in primary colon, breast, liver, and lung cancers in the dataset of Su et al. [21].

Then we examined the expression of these genes in the lung cancer dataset of Bhattacharjee et al. [28]. For simplicity, only those diagnosed as lung adnocarcinoma are examined. As shown in Fig. 6B, we observed higher expression of a colon-specific gene cluster in 12 lung tumors, most of which are diagnosed as colon metastasis. We also observed overexpression of dozens of liver-specific genes in one sample (AD368). In agreement with clinical diagnosis, one sample (AD163) clearly shows an expression pattern similar to that of luminal-like breast cancer. Meanwhile, three samples (AD352, AD186, and AD131) exhibit expression signatures of basal-like breast cancer. Two of them (AD352 and AD186) are diagnosed as breast metastasis. Totally, we identified 17 tumors that might have originated from distant organs. Nine of them are confirmed by clinical diagnosis. All of these 17 samples show underexpression of genes specific for lung adenocarcinoma (Fig. 6B). Therefore, the expression pattern of these marker genes provides useful information about tumor origin.

Sample AD172 was diagnosed as probably breast metastasis, but did not show either of the two breast-specific expression patterns. On the contrary, AD131 was diagnosed as primary lung adenocarcinoma, but shows an expression profile similar to that of basal-like breast cancer. These are some discrepancies between our prediction and diagnosis.

With these marker genes, it is possible to train some machine-learning algorithms to predict tumor origins. The data shown in Fig. 6A were used to train a prototype matching algorithm described in Ref. [44] (available at http://www.jsbi.org/journal/GI14.html), which is similar to the one proposed in Ref. [45] but emphasizes the minimization of false positive errors. When tested with the lung dataset of Fig. 6B, the algorithm makes confident predictions for 16 of the 17 secondary tumors in agreement with clinical diagnosis. Only three false positive predictions are made for the remaining 112 primary lung adenocarcinomas. Therefore, with a set of carefully selected tissue-specific genes, it is possible to predict the origin of tumors with high accuracy.

Discussion

Through expression profiling of a spectrum of normal human tissues, we identified sets of tissue-specific genes, and then studied their expression in cancers by analyzing a wealth of previously published DNA microarray datasets. Through unsupervised clustering of tissue-specific genes differentially expressed in tumors from the same anatomical site, we identified groups of coexpressed genes characteristic of different cell types within the organ, thus revealing cell lineage of tumor subtypes. Similar observations are made in liver, brain, and breast, as well as lung tumors.

The expression pattern of tissue-specific genes in tumors could be univariant (liver cancer), bivariant (breast cancer), or multivariant (brain and lung cancers). We identified a set of liver-specific genes whose expression in HCC changes according to the degree of tumor differentiation (Fig. 2). This set of genes can be used to classify tumors into differentiation categories more accurately than

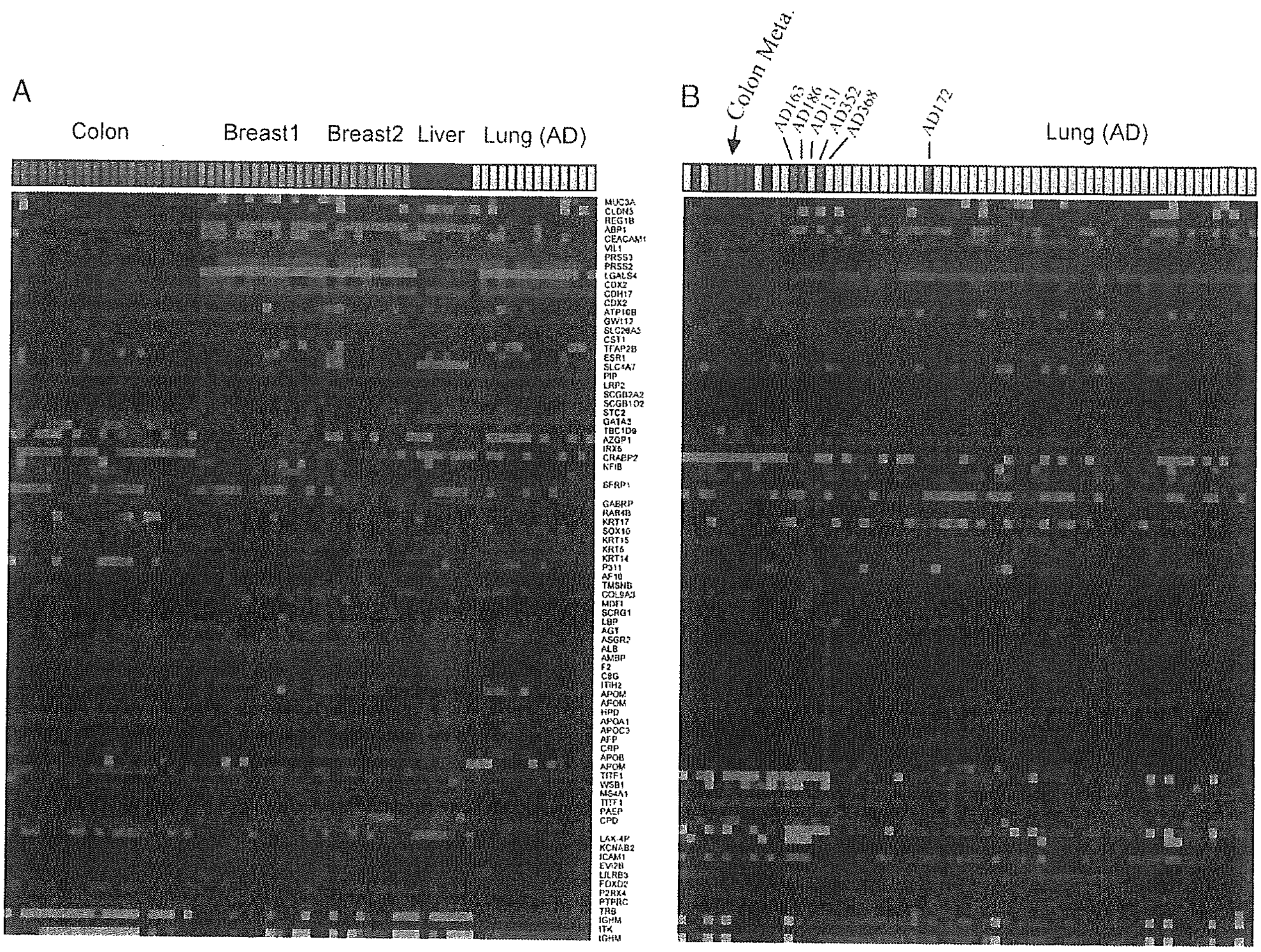


Fig. 6. Prediction of tumor origin with selected markers genes. Predictor genes for colon and liver cancers are selected from the result shown in Fig. 5, while those for two major types of breast cancers are from Fig. 4. Predictors of lung adenocarcinoma are from Ref. [21]. (A) Expression of these genes in the dataset of primary colon, breast, liver, and lung cancers (data from Ref. [21]). (B) The expression of these genes in the dataset of lung tumor dataset [28]. Some samples are diagnosed as colon or breast metastases, indicated by red and green, respectively.

using the global expression profile. For brain tumors, we identified neuron-specific expression signatures in medullobrastoma and glia-specific signatures in glioma. No such feature is observed for rhabdoid and PNET. We also found a small set of 26 genes that are highly expressed in the normal breast but are divided into two groups, whose expression in breast tumors is mutually exclusive and defines two types of differentiation (Fig. 4). We observed that different subtypes of lung cancers show different patterns of tissue specificity, e.g., high expression of skin-specific genes in SQ and high expression of brain-specific genes in SCLC and COID (Fig. 5). In addition, expression signatures of primary sites is detectable in lung tumors originating from colon, liver, and breast. Notably, we were able to detect lung tumors with expression profiles resembling two subtypes of breast cancers. Summarizing these results, we selected molecular markers that can be used to predict tumor origins (Fig. 6).

DNA microarrays are powerful tools for studying cancer. But biological interpretation of the obtained levels of gene expression is often challenging. Our work shows that categorization of genes according to their tissue specificity is useful for the interpretation of the data of cancer. Starting from a small set of a normal tissue gene expression dataset, we reanalyzed multiple cancer datasets of liver, breast, brain, and lung cancers, and obtained valuable information on tumor differentiation, molecular heterogeneity, and tumor origin. Such information is often difficult to extract when each dataset is analyzed independently in a stand-alone manner. This illustrated the far-reaching benefits of systematic studies on normal tissues. The creation of a collective normal control panel that includes gene expression datasets of a spectrum of normal tissues is beneficial for research on tumors in all organs.

As a proof-of-concept study, the present work used pooled RNA to reduce the cost of biological replicates, a strategy supported by some recent comparative studies [46,47]. Although we showed that our list of tissue-specific genes are already useful for analyzing gene expression data of various cancers, further work is needed to refine these lists by including biological replicates in a systematic study on normal tissue gene expression. During the preparation of the current manuscript a larger database of normal tissue gene expression was published (see Ref. [52]).

As tumors are the result of uncontrolled proliferation of certain cells within an organ, they are more homogeneous than normal organs and could serve as natural subject for studying expression signatures of individual cell types. The expression patterns shown in Fig. 2 to Fig. 5 contain many well-known markers for different cell types, such as ALB for heptocyte, GFAP and OLIG2 for glia cells, and the keratin genes for basal and luminal epithelial cells. Other genes in the list might serve as potential candidates for new markers. It might be possible to take advantage of the homogeneity of cell population in tumors and gain insights on expression signatures of different cell types from the expression profiles of tumors.

As these expression patterns are cell-type specific, we should be able to detect common transcription factor binding motifs on the promoter regions of these genes in the human genome. For the gene lists shown in Figs. 2, 3, and 4, we extracted upstream sequences and compared the occurrence of known transcription factor motifs with a group of control genes (see Supplementary Information for more details). We found statistically significant enrichment of motifs for hepatic nuclear factors (HNF1, HNF3, HNF4, and HNF6) in the hepatocyte-specific genes (Fig. 2), and neuron-restrictive silencer factor (NRSF) for neuron-specific genes marked as Group β in Fig. 3. Without the combination of normal and cancer expression profiles, such regulatory motifs would be more difficult to detect.

In summary, we demonstrated the importance of integrating tissue specificity into the interpretation of the expression profiles of tumors, especially for the study of tumor differentiation, cell lineage, and metastasis. Systematic, large-scale studies on normal tissue gene expression profiles could both give rise to baseline controls in basic data analysis and be used to define each gene's breadth of expression in normal tissues. Knowing how genes are expressed under normal physiological conditions is important for dissecting complicated cancer transcriptomes.

Materials and methods

Sample preparation

Twenty-five total RNA specimens were purchased from Clontech (Palo Alto, CA), Ambion (Austin, TX) and Strategene (La Jolla, CA). In order to define breadth of expression accurately at a reasonable cost, we tried to cover as many tissue types as possible by using pooled RNA samples. Each specimen represents a human organ. We used RNA samples pooled from 2 to 84 donors to avoid differences at the individual level. But still many specimens from single donors are included because of the difficulty in obtaining healthy tissues. We also purchased seven poly(A) RNA specimens of spinal cord and several brain regions such as corpus callosum, hippocampus, thalamus, pituitary gland, caudate, and amygdala. In addition to these purchased RNAs, we obtained tissue specimens of liver, stomach, lung, and fetal lung from individuals with informed consent. The specimens were immediately preserved in liquid nitrogen for further analysis. Total RNAs were extracted from these specimens by using ISOGEN (Isogen Life Science, Industrieweg 66-68, 3606 AS Maarssen, Netherlands). For further demographic information, please refer to the Supplementary Information.

Microarray experiments

Total RNA or Poly(A) RNA was used to synthesize cRNA which was then hybridized to HG-U133A oligonucleotide array (Affymetrix, Santa Clara, CA) according to standard protocols as described previously [18].

Data acquisition

After hybridization, all scanned images were visually inspected for artifacts and overall quality. Affymetrix's MicroArray Suite 5.0 software was used to analyze image files. The software calculates a "signal" to characterize each gene's expression level based on the difference between the densities of mutiple pairs of perfect match (PM) and mismatch (MM) probes. In addition, it also produces a "detection *P* value" to indicate how confidently a gene's expression is detected. If the densities on most PM probes are significantly larger than their corresponding MM probes, the algorithm will return a smaller *P* value. Usually, a gene is considered present if $P < 0.05$, and absent if $P > 0.06$. Absent calls indicate that the corresponding expression data are not reliable. Raw DNA microarray data have been deposited with NCBI Gene Expression Omnibus (GEO) under accession: GSE2361. The data is also available at the authors' web site: http://www.genome.rcast.u-tokyo.ac.jp/normal/.

Data normalization

Normalization is done among the probe sets with present calls in each array. After the top and bottom 5% are removed, the average of the logarithm of signals produced by these probe sets is centered to the logarithm of a positive number, here 160, to be comparable with a target density of 100 in global scaling for most tissues. Scores are then transformed by an inverse logarithm. This kind of procedure is preferred when comparing multiple tissue types because the total number of present calls varies significantly with tissues, which leads to biases to the default global scaling method. Finally signals smaller than 10 are set to 10.

Selection of tissue-specific genes

We consider a gene specific to a tissue type if it is exclusively highly expressed in this tissue. An example of the expression pattern of tissue-specific genes is shown in Supplementary Fig. 2. To select such genes, we used t test and several empirical criteria. Suppose a gene's expression level (g) is the highest in a certain tissue, for example, liver. We first require that this score is associated with a present call. Then the expression level g is compared with the mean (m) and standard deviation (SD) observed in the rest of the tissues. This gene is considered liver-specific if (a) $g > m + 3SD$, (b) $g/g2 > 2$, and (c) $g > 160$ (d) $g2 < 150$, where $g2$ is the second highest expression score in all the tissues. To avoid missing lowly expressed tissue-specific genes, we also included genes that meet an alternative criterion: a gene must be confidently present in this tissue (detection P value < 0.02) and absent (detection P value > 0.08) in all others. Also the absolute expression level must meet condition (b).

In addition to tissue-specific genes, which are exclusively expressed in one particular organ, there are some genes whose expression is restricted to two or more organs or anatomical sites. As an example, the expression pattern of cytokeratin 20 (KRT20), which is highly expressed in stomach, colon, and small intestine, is shown in Supplementary Fig. 3. To define such tissue-selective genes, we used Sprent's nonparametric method [19]. For each gene, the log-transformed signal values of all tissues are used to calculate a median and median absolute deviation (MAD). Then those tissues with a signal larger than median by more than 5 MAD (equivalent to 3.375 SD in normal distribution) are considered significant. The number of tissues with significantly higher expression must be smaller than 8. The usage of median and MAD are preferred over the mean and SD because they are more robust and less sensitive to outliners, e.g., extremely large signal values in a few tissues.

Clustering analysis

A filtering process is applied to eliminating genes whose expression does not show much variance among the samples in question. A gene should show more than a 2-fold change between the maximum and the median. Also the absolute difference should be larger than 100. Then the data are log-transformed, and the gene vector is median-centered and divided by SD. Average linkage hierarchical clustering is done using the Cluster and Treeview program [20] with Pearson's correlation coefficient as a distance metrics.

Public gene expression datasets and metaanalysis

In addition to our own data, we also use two normal tissue gene expression database, namely HuGe Index database (Ref. [13], available at http://www.hugeindex.org) and Gene Expression Atlas database (Ref. [15], http://www.expression.gnf.org/). To study the expression of tissue-specific genes in cancers, we analyzed a dataset of multiple cancer types (Ref. [21], http://www.carrier.gnf.org/welsh/epican/), a liver cancer dataset (Ref. [22], http://www.lsbm.org/db/), two datasets of breast cancer (Refs. [26,27], http://www.genome-www.stanford.edu/breast_cancer/molecularportraits/, and Ref. [2], http://www.rii.com/publications/vantveer.htm), and a lung cancer dataset (Ref. [28], http://www-genome.wi.mit.edu/cancer/). Several datasets of other cancer types are also used in our study of maintenance genes. A full list of data sources is available in Supplementary Table 1.

Most of these datasets are based on Affymetrix GeneChip systems (HuGeneFL, HG-U95A, or HG-U133A), for which annotation information about probe sets are available at http://www.affymetrix.com. We also used one dataset of cDNA microarrays. Mapping between these different datasets is performed according to the latest version of UniGene (as for May 2003) by using the SOURCE database (Ref. [31], http://www.source.stanford.edu).

Classification of HCC samples

We tested two sets of predictor genes for classifying HCC samples into well, moderate, and poorly differentiated tumors. This first set consists of 64 liver-specific transcripts shown in Fig. 2A; the other set includes 3536 genes passed through a variation filter (max/min > 2, max−min > 100). A standard k-nearest neighbor (kNN) algorithm with (k = 4) was employed to classify each of 25 tumors withheld from training. To make a positive prediction, a winning type must receive a percentage of votes larger than a certain margin over all other types. This threshold is adjusted from 0, 10, 30%, 50, 70, and 90% to produce the ROC curve in Fig. 2B.

Gene ontology analysis

Statistical association of gene lists with GO categories are performed with the Onto-Expression software [24]), available at http://www.vortex.cs.wayne.edu. Binominal distribution is used to calculate the P value at which the list is enriched by genes belonging to a certain function category.

Promoter analysis

To search for cell-specific promoter binding motifs, we extract promoter sequences from 2500 bp upstream to 500 bp downstream transcription starting site (TSS) using the Promoser web service ([50], http://www.biowulf.bu.edu/zlab/PromoSer/). As a control group, we also extract similar sequences of 1144 maintenance genes. We developed a set of Perl scripts to scan these sequences for binding sites of known transcription factors included in the TRANSFAC database [51]. Then we calculated the P value of over-representation for each motif by comparing the frequency between each cluster and the control group according to hypergeometric distribution.

Acknowledgments

The authors are indebted to Hirokazu Taniguchi for help with tissue acquisition, Hiroko Meguro for technical assistance, Yoshitaka Hippo, Naoko Nishikawa, Chen Yongxin, and Guo Yongqiu for stimulating discussions and Jiang Fu for proofreading. This work was partially supported by Grants-in-Aid for Scientific Research (S) 16101006 from The Ministry of Education, Culture, Sports, Science and Technology, Japan (to H.A.), and Health and Labour Sciences Research Grants (to H.A.). This work has been supported in part by NIH and Daniel F. and Ada L. Rice Foundation (to S.M.W.).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2005.04.008. Supplementary data is also available at the authors' web site: http://www.genome.rcast.u-tokyo.ac.jp/normal/.

References

[1] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (1999) 531–537.

[2] L.J. Van 'T Veer, H. Dai, M.J. Van de Vijver, Y.D. He, A.A.M. Hart, M. Mao, H.L. Peterse, K. Van der Kooy, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bernards, S.H. Friend, Gene expression profiling predicts clinical outcome of breast cancer, Nature 415 (2002) 530–536.

[3] J.E. Staunton, D.K. Slonim, H.A. Coller, P. Tamayo, M.J. Angelo, J. Park, U. Scherf, J.K. Lee, W.O. Reinhold, J.N. Weinstein, et al., Chemosensitivity prediction by transcriptional profiling, Proc. Natl. Acad. Sci. USA 98 (2001) 10787–10792.

[4] M.A. Shipp, K.N. Ross, P. Tamayo, A.P. Weng, J.L. Kutok, R.C.T. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G.S. Pinkus, T.S. Ray, M.A. Koval, K.W. Last, A. Norton, T.A. Lister, J. Mesirov, D.S. Neuberg, E.S. Lander, J.C. Aster, T.R. Golub, Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning, Nat. Med. 8 (2002) 68–74.

[5] E.J. Yeoh, M.E. Ross, S.A. Shurtleff, W.K. Williams, D. Patel, R. Mahfouz, F.G. Behm, S.C. Raimondi, M.V. Relling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X.D. Zhou, J.Y. Li, H.G. Liu, C.H. Pui, W.E. Evans, C. Naeve, L. Wong, J.R. Downing, Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling, Cancer Cell 1 (2002) 133–143.

[6] Y. Moreau, S. Aerts, B. De Moor, B. De Strooper, M. Dabrowski, Comparison and meta-analysis of microarray data: from the bench to the computer desk, Trends Genet. 19 (2003) 570–577.

[7] The Gene Ontology Consortium, Gene Ontology: tool for the unification of biology, Nat. Genet. 25 (2000) 25–29.

[8] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, M. Hattori, The KEGG resources for deciphering the genome, Nucleic Acids Res. 32 (2004) D277–D280.

[9] International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome, Nature 409 (2001) 860–921.

[10] J.C. Venter, et al., The sequence of the human genome, Science 291 (2001) 1304–1351.

[11] J.A. Warrington, A. Nair, M. Mahadevappa, M. Tsyganskaya, Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes, Physiol. Genomics 2 (2000) 143–147.

[12] V.E. Velculescu, et al., Analysis of human transcriptomes, Nat. Genet. 23 (1999) 387–388.

[13] L.L. Hsiao, et al., A compendium of gene expression in normal human tissues, Physiol. Genomics 7 (2001) 97–104.

[14] A. Saito-Hisaminato, T. Katagiri, S. Kakiuchi, T. Nakamura, T. Tsunoda, Y. Nakamura, Genome-wide profiling of gene expression in 29 normal human tissues with a cDNA microarray, DNA Res. 9 (2002) 35–45.

[15] A.I. Su, M.P. Cooke, K.A. Ching, Y. Hakak, J.R. Walker, T. Wiltshire, A.P. Orth, R.G. Vega, L.M. Sapinoso, A. Moqrich, A. Patapoutian, G.M. Hampton, P.G. Schultz, J.B. Hogenesch, Large-scale analysis of the human and mouse transcriptomes, Proc. Natl. Acad. Sci. USA 99 (2002) 4465–4470.

[16] H. Caron, B. van Schaik, M. van der Mee, F. Baas, G. Riggins, P. van Sluis, M.C. Hermus, R. van Asperen, K. Boon, P.A. Voute, S. Heisterkamp, A. van Kampen, R. Versteeg, The human transcriptome map: clustering of highly expressed genes in chromosomal domains, Science 291 (2001) 1289–1292.

[17] A.E. Lash, C.M. Tolstoshev, L. Wagner, G.D. Schuler, R.L. Strausberg, G.J. Riggins, S.F. Altschul, SAGEmap: a public gene expression resource, Genome Res. 10 (2000) 1051–1060.

[18] Y. Hippo, H. Taniguchi, S. Tsutsumi, N. Machida, J.M. Chong, M. Fukayama, T. Kodama, H. Aburatani, Global gene expression analysis of gastric cancer by oligonucleotide microarrays, Cancer Res. 62 (2002) 233–240.

[19] P. Sprent, N.C. Smeeton, Applied Nonparametric Statistical Methods (Texts in Statistical Science), Chapman and Hall, London, 2000.

[20] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, Proc. Natl. Acad. Sci. USA 95 (1998) 14863–14868.

[21] A.I. Su, J.B. Welsh, L.M. Sapinoso, S.G. Kern, P. Dimitrov, H. Lapp, P.G. Schultz, S.M. Powell, C.A. Moskaluk, H.F. Frierson Jr., G.M. Hampton, Molecular classification of human carcinomas by use of gene expression signatures, Cancer Res. 61 (2001) 7388–7393.

[22] Y. Midorikawa, S. Tsutsumi, H. Taniguchi, M. Ishii, Y. Kobune, T. Kodama, M. Makuuchi, H. Aburatani, Identification of genes associated with dedifferentiation of hepatocellular carcinoma with expression profiling analysis, Jpn. J. Cancer Res. 93 (2002) 636–643.

[23] Y. Midorikawa, S. Tsutsumi, K. Nishimura, N. Kamimura, M. Kano, H. Sakamoto, M. Makuuchi, H. Aburatani, Distinct chromosomal bias of gene expression signatures in the progression of hepatocellular carcinoma, Cancer Res. 64 (2004) 7263–7270.

[24] P. Khatri, S. Draghici, G.C. Ostermeier, S.A. Krawetz, Profiling gene expression using Onto-Express, Genomics 79 (2002) 266–270.

[25] S.L. Pomeroy, et al., Prediction of central nervous system embryonal tumour outcome based on gene expression, Nature 415 (2002) 436–442.

[26] C.M. Perou, P.O. Brown, D. Botstein, et al., Molecular portraits of human breast tumours, Nature 406 (2000) 747–752.

[27] T. Sorlie, et al., Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications, Proc. Natl. Acad. Sci. USA 98 (2001) 10869–10874.

[28] A. Bhattacharjee, W.G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E.J.

Mark, E.S. Lander, W. Wong, B.E. Johnson, T.R. Golub, D.J. Sugarbaker, M. Meyerson, Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses, Proc. Natl. Acad. Sci. USA 98 (2001) 13790–13795.

[29] D. Singh, et al., Gene expression correlates of clinical prostate cancer behavior, Cancer Cell 1 (2002) 203–209.

[30] B. Dasarthy, Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques, IEEE Computer Society Press, Los Alamitos, CA, 1991.

[31] M. Diehn, G. Sherlock, G. Binkley, H. Jin, J.C. Matese, T. Hernandez-Boussard, C.A. Rees, J.M. Cherry, D. Botstein, P.O. Brown, A.A. Alizadeh, SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data, Nucleic Acids Res. 31 (2003) 219–223.

[32] S. Gruvberger, M. Ringner, Y. Chen, S. Panavally, L.H. Saal, A. Borg, M. Ferno, C. Peterson, P.S. Meltzer, Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns, Cancer Res. 61 (2001) 5979–5984.

[33] X.J. Ge, S. Tsutsumi, H. Aburatani, S. Iwata, Reducing false positives in molecular pattern recognition, Genome Informatics 14 (2003) 34–43.

[34] R. Tibshirani, T. Hastie, B. Narasimhan, G. Chu, Diagnosis of multiple cancer types by shrunken centroids of gene expression, Proc. Natl. Acad. Sci. USA 99 (2002) 6567–6572.

[35] J.H. Shih, et al., Effects of pooling mRNA in microarray class comparisons, Bioinformatics 20 (2004) 3318–3325.

[36] C.M. Kendziorski, Y. Zhang, H. Lan, A.D. Attie, The efficiency of pooling mRNA in microarray experiments, Biostatistics 4 (2003) 465–477.

[37] M. Gotoh, T. Nakatani, T. Masuda, Y. Mizuguchi, M. Sakamoto, R. Tsuchiya, H. Kato, K. Furuta, Prediction of invasive activities in hepatocellular carcinomas with special reference to alpha-fetoprotein and des-gamma-carboxyprothrombin, Jpn. J. Clin. Oncol. 33 (2003) 522–526.

[38] C. Couture, H. Raybaud-Diogene, B. Tetu, I. Bairati, D. Murry, J. Allard, A. Fortin, p53 and Ki-67 as markers of radioresistance in head and neck carcinoma, Cancer 94 (2002) 713–722.

[39] A.S. Halees, D. Leyfer, Z. Weng, Promoser: a larger-scale mammalian promoter and transcription start site identification service, Nucl. Acids. Res. 31 (2003) 3554–3559.

[40] E. Wingender, X. Chen, E. Fricke, R. Geffers, R. Hehl, I. Liebich, M. Krull, V. Matys, H. Michael, R. Ohnhäuser, M. Prüß, F. Schacherer, S. Thiele, S. Urbach, The TRANSFAC system on gene expression regulation, Nucleic Acids Res. 29 (2001) 281–283.

[41] A.I. Su, T. Wiltshire, S. Batalov, H. Lapp, K.A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M.P. Cooke, J.R. Walker, J.B. Hogenesch, A gene atlas of the mouse and human protein-encoding transcriptomes, Proc. Natl. Acad. Sci. USA 101 (2004) 6062–6067.

[42] L. Ronnov-Jessen, O.W. Petersen, M.J. Bissell, Cellular changes involved in conversion of normal to malignant breast: importance of the stromal reaction, Physiol. Rev. 76 (1996) 69–125.

A meta-clustering analysis indicates distinct pattern alteration between two series of gene expression profiles for induced ischemic tolerance in rats

Makoto Kano, Shuichi Tsutsumi, Nobutaka Kawahara, Yan Wang, Akitake Mukasa, Takaaki Kirino and Hiroyuki Aburatani

Physiol. Genomics 21:274-283, 2005. First published Feb 15, 2005;
doi:10.1152/physiolgenomics.00107.2004

Physiol Genomics 21: 274–283, 2005.
First published February 15, 2005; doi:10.1152/physiolgenomics.00107.2004.

A meta-clustering analysis indicates distinct pattern alteration between two series of gene expression profiles for induced ischemic tolerance in rats

Makoto Kano,[1] Shuichi Tsutsumi,[2] Nobutaka Kawahara,[3,4] Yan Wang,[3] Akitake Mukasa,[2,3] Takaaki Kirino,[3,4] and Hiroyuki Aburatani[2]

[1]*Intelligent Cooperative System, Department of Information Systems, Research Center for Advanced Science and Technology, University of Tokyo, Tokyo;* [2]*Genome Science Division, Research Center for Advanced Science and Technology and* [3]*Department of Neurosurgery, Faculty of Medicine, University of Tokyo, Tokyo; and* [4]*Solution-Oriented Research for Science and Technology/Japan Science and Technology, Kawaguchi, Saitama, Japan*

Submitted 5 May 2004; accepted in final form 11 February 2005

Kano, Makoto, Shuichi Tsutsumi, Nobutaka Kawahara, Yan Wang, Akitake Mukasa, Takaaki Kirino, and Hiroyuki Aburatani. A meta-clustering analysis indicates distinct pattern alteration between two series of gene expression profiles for induced ischemic tolerance in rats. *Physiol Genomics* 21: 274–283, 2005. First published February 15, 2005; doi:10.1152/physiolgenomics.00107.2004.—We have developed a visualization methodology, called a "cluster overlap distribution map" (CODM), for comparing the clustering results of time series gene expression profiles generated under two different conditions. Although various clustering algorithms for gene expression data have been proposed, there are few effective methods to compare clustering results for different conditions. With CODM, the utilization of three-dimensional space and color allows intuitive visualization of changes in cluster set composition, changes in the expression patterns of genes between the two conditions, and relationship with other known gene information, such as transcription factors. We applied CODM to time series gene expression profiles obtained from rat four-vessel occlusion models combined with systemic hypotension and time-matched sham control animals (with sham operation), identifying distinct pattern alteration between the two. Comparisons of dynamic changes of time series gene expression levels under different conditions are important in various fields of gene expression profiling analysis, including toxicogenomics and pharmacogenomics. CODM will be valuable for various types of analyses within these fields, because it integrates and simultaneously visualizes various types of information across clustering results.

time series; transcription factor; visualization

ADVANCES IN MICROARRAY TECHNOLOGIES have made it possible to comprehensively measure gene expression profiles. Observation of dynamic changes of gene expression levels provides important markers to clarify cellular responses, differentiation, and genetic regulatory networks. In particular, a comparison of dynamic changes of time series gene expression levels under various conditions (e.g., administration of different drugs) is expected to make a major contribution to the understanding of complex biological processes. In general, we observe the influence of each condition through the results of clustering analysis, which is the most popular analysis for gene expression profiles. Therefore, a comparison between the results of clustering analyses in different conditions will allow interpretation of different macroscopic phenomenon that occurred under those conditions. However, although many clustering algorithms, including hierarchical clustering (1, 2, 4, 15), k-nearest neighbor (17), and self-organizing maps (10, 13, 16) have been proposed, there are few effective methods to effectively compare clustering results under different conditions. We have defined four issues to be addressed for a comparison of clustering results, especially for a comparison of time series gene expression data under two different conditions: changes in the composition of the cluster sets, changes in the expression patterns, integration with known other gene information, and threshold problems.

Changes in the Composition of the Cluster Sets

In this report, we focused on hierarchical clustering, since it is the most popular method for gene expression analysis. Here we define the composition of a cluster set as the hierarchical structure of clustering results and "cluster set" as the set of all clusters in the structure. A comparison of clusters' compositions shows which clusters are conserved in different conditions and how the genes in a cluster for one condition are distributed into a cluster set under another condition. Genes that cluster under a single condition may possibly be regulated by the same factors for that condition. However, under different conditions, some of those genes would be regulated by other factors and generate different clusters. Thus changes in the cluster compositions could provide key information for interpreting the effects of the different conditions. To get a full picture of the relationships of two cluster sets, the overlap between each pair of clusters under the two different conditions should be evaluated. However, since clustering analysis, especially hierarchical clustering, almost always generates a great number of clusters, there are a very large number of combinations of clusters. Simple line connections of the genes between the dendrograms of two hierarchical clustering results (14) provide insufficient information about the relationships between the clusters. Therefore, an effective presentation method that provides a full picture of the relationships of the cluster sets would be desirable.

Recently, a statistical model for performing meta-analysis of independent microarray data sets was proposed (12). This model revealed, for example, that four prostate cancer gene expression data sets shared significantly similar results, independent of the method and technology used. However, in a comparison of the cluster sets based on different conditions, the objective is not to confirm that several data sets share significantly similar results, but to detect the differences be-

Article published online before print. See web site for date of publication (http://physiolgenomics.physiology.org).

Address for reprint requests and other correspondence: M. Kano, Intelligent Cooperative System, Dept. of Information Systems, Research Center for Advanced Science and Technology, Univ. of Tokyo, Tokyo 153-8904, Japan (E-mail: mkano@cyber.rcast.u-tokyo.ac.jp).

tween them. Several statistical algorithms have been proposed for evaluating how clusters based on expression profiles include genes with well-known functions (3, 17). However, the number of clusters that were compared was limited, and an effective presentation method was not required in those situations.

Changes in the Expression Pattern

Where two clusters under different conditions have a statistically meaningful number of genes in common, it is also important to examine the differences in their expression patterns. The differences of macroscopic phenomena that the conditions exhibit result from the differences of expression of multiple, rather than single, genes. Therefore, the genes whose expression patterns changed in a similar fashion between different conditions provide markers for the different phenomena. In other words, if the genes in a certain cluster based on one condition also constitute a cluster for another condition, but the expression patterns are greatly different between the two conditions, then these genes are causally implicated in the phenotypic difference.

In general, there will be many false candidate genes whose expression patterns coincidentally match between the two different conditions. Therefore, it is important to simultaneously evaluate the statistical significance of the overlaps between clusters and the differences in their expression patterns.

Integration with Other Known Gene Information

In gene expression analysis, it is important to biologically interpret the results after integrating them with other known gene information. Therefore, changes in the composition of the cluster sets and changes in the expression patterns between different conditions should be associated with other known gene information such as transcription factors.

Threshold Problems

In a comparison of cluster sets on gene expression profiles, we have to handle four types of thresholds: *1*) a threshold for generating clusters for each condition; *2*) a threshold for evaluating the number of common genes that two clusters have; *3*) a threshold for evaluating the differences in the expression patterns between two clusters; and *4*) a threshold for evaluating the relationship with other known gene information. Among these, determining the threshold for generating clusters is most challenging, because the clustering result strongly depends on this threshold, and a change of this threshold greatly affects the number and composition of clusters. It is generally difficult to determine optimal values for these four types of thresholds, and the results of analysis are greatly affected by the threshold values specified. Arbitrary selection of thresholds involves a risk of overlooking important genes, so the number of thresholds should be reduced, and, if used, it is necessary to allow users to interactively change the thresholds.

We focused on visualization technology to address these four issues. Interactive visualization is effective for handling ambiguous threshold problems and for providing a wide variety of information at one time. In previous work, we developed a "cluster overlap distribution map" (CODM), which is a visualization method for comparing cluster sets based on different sets of gene expression profiles (7). In this report, we extended it for time series gene expression analysis. In the CODM, the relationships of all possible pairing of clusters can be examined, and both the changes in the composition of the cluster sets and the changes in the expression patterns of the clusters can be effectively visualized as three-dimensional (3D) histograms, without any arbitrary thresholds. In addition, relationships with other known gene information such as transcription factors can also be elucidated. We applied the CODM to a comparison between the gene expression data sets of double ischemia rats and sham control rats (with sham operation) and confirmed that CODM identified distinct patterns between the two.

CODM, available on our web site (**http://www.genome.rcast.u-tokyo.ac.jp/CODM**), runs on a PC with Windows 2000 or Windows XP. Memory requirement is in proportion to the square of the number of genes to be analyzed. The analysis for ~4,000 genes, represented in this paper, required ~250 megabytes. In addition, since the analysis results of the CODM are visualized by use of the OpenGL, a machine with a graphics board with a hardware accelerator for the OpenGL is recommended.

MATERIALS AND METHODS

Experiment Design

In this report, CODM is illustrated using time series gene expression data sets obtained from rat four-vessel occlusion models combined with systemic hypotension and time-matched control animals with sham operation. In the experiment, we used 2-min ischemia rats with induced ischemic tolerance (tolerant rats, TOL) and rats with sham operation (sham rats, SHAM), after confirming the histological outcomes. Note that the sham rats did not acquire ischemic tolerance. Three days after the operation, we conducted a 6-min ischemia operation on the two groups. Because of their ischemic tolerance, very little neuronal death of CA1 hippocampal neurons was observed in the tolerant rats (9). With duplicate assessments of 6 time points ({0 h, 1 h, 3 h, 12 h, 24 h, 48 h} × 2) after the second ischemia, microdissected CA1 regions from each of the two groups were subjected to oligonucleotide-based microarray analysis.

All animal-related procedures were conducted in accordance with guidelines for the care and use of laboratory animals set out by the National Institutes of Health and were approved by the committee for the use of laboratory animals in the University of Tokyo. More detailed experimental design is described in our previous report (8).

GeneChip Experiment

Five micrograms of total RNA from each sample was used to synthesize biotin-labeled cRNA, which was then hybridized to a high-density oligonucleotide array (GeneChip Rat RG-U34A array, Affymetrix) essentially following a previously published protocol (6). The arrays contain probe sets for 8,737 rat genes and expressed sequence tags (ESTs), which were selected from Build 34 of the UniGene Database (derived from GenBank 107, dbEST/11-18-98). Sequences and GenBank accession numbers of all probe sets are available from the Affymetrix home page (**http://www.affymetrix.com/index.affx**). Washing and staining was performed in a Fluidics Station 400 (Affymetrix) using the protocol EukGE-WS2. Scanning was performed on an Affymetrix GeneChip scanner to collect primary data. The Affymetrix Microarray Suite v4.0 was used to calculate the average difference for each gene probe on the array, which was shown as an intensity value of gene expression defined by Affymetrix using their algorithm. The average difference has been shown to quantitatively reflect the abundance of a particular mRNA molecule in a

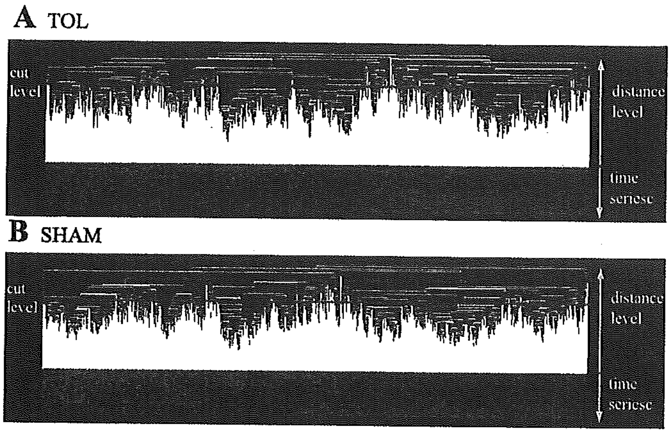


Fig. 1. Hierarchical clustering of TOL (*A*) and SHAM (*B*). We obtained time series ({0 h, 1 h, 3 h, 12 h, 24 h, 48 h} × 2) microarray data from rats with induced ischemic tolerance (tolerant rats, TOL) and rats with sham operation (sham rats, SHAM). In the analysis, we used these data sets as 12 time point ({0a, 0b, 1a, 1b, 3a, 3b, ..., 48a, 48b} = $\{T_i\}$ ($i = 1,2,\ldots,12$)) data sets on TOL and SHAM, respectively. After preprocessing and normalization, hierarchical clustering analysis based on Euclidian distances was then performed for each data set independently.

population (6). To allow comparison among multiple arrays, the average differences were normalized for each array by assigning the mean of overall average difference values to be 100. This data set has been submitted as GSE1357 to the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (**http://www.ncbi.nlm.nih.gov/geo/info/linking.html**)

Preprocessing and Clustering

In the following analysis, we used data sets as 12 time point ({0a, 0b, 1a, 1b, 3a, 3b, ..., 48a, 48b} = $\{T_i\}$ ($i = 1,2,\ldots,12$)) data sets on TOL and SHAM, since the CODM does not depend on the intervals of the time points.

Standard clustering analysis for gene expression profiles is based on the correlation coefficients between genes. Therefore, this approach cannot handle genes with expression profiles that have almost no changes for a condition. However, if the expression profiles of those genes have meaningful changes in expression levels for other conditions, then these provide markers to interpret the influence that the conditions exerted, because these are possibly regulated by different factors. To handle those genes and to align the baselines of the expression patterns between the different data sets, preprocessing (i.e., filtering and normalization) was conducted for all of the data sets where TOL and SHAM were merged. More specifically, 3,363 probes with mean expressions above 50 and coefficient of variance (CV = standard deviation/mean) above 0.1 were selected. After logarithmic transformation of the gene expression data, the expression levels were normalized to satisfy the following equations:

$$\sum_{i}^{12}(x_i + y_i) = 0 \qquad (1)$$

$$\sum_{i}^{12}(x_i^2 + y_i^2) = 1 \qquad (2)$$

where x_i and y_i are normalized expression levels of a gene at time point T_i ($i = 1,2,\ldots12$) on conditions TOL and SHAM, respectively. Using these normalized data sets, we performed hierarchical clustering analysis based on Euclidian distances, for each data set independently. Clustering analysis using Euclidian distances instead of correlation coefficients allows us to handle genes whose expression levels are downregulated or upregulated. In addition, due to the common normalization, gene expression patterns can be compared within a data set and between data sets.

In general, Euclidian-distance-based clustering after normalization, in terms of mean and standard deviation, is equivalent with correlation-coefficient-based clustering. That is, a Euclidian-distance-based clustering analysis for the merged data of TOL and SHAM with the above preprocessing is equivalent with a correlation-coefficient-based clustering analysis for the original merged data. In the analysis of the CODM, the preprocessing is conducted for the merged data, and Euclidian-based clustering is individually conducted for each data. Roughly speaking, this analysis provides us with results similar to those of normal correlation-coefficient-based clustering, while it allows us to handle genes with expression profiles that have changes for only one condition but not for the other.

As Fig. 1, *A* and *B*, shows, there are a large number of clusters generated at various levels. Although the composition and number of cluster sets depend on the threshold value of the distance, it is generally difficult to identify an optimum value. These aspects make it difficult to compare cluster sets derived from different sources.

The Cluster Overlap Distribution Map

The CODM is a visualization methodology for pair-wise comparison between cluster sets generated from different gene expression data sets. In this methodology, two types of cluster sets (i.e., dendrograms of hierarchical clustering results) are mapped, respectively, to the *x*-axis and to the *y*-axis, and the relationship between them is displayed as a 3D histogram (Fig. 2). In this report, the dendrogram of TOL is mapped to the *x*-axis, and that of SHAM is mapped to the *y*-axis. The statistical evaluation values of the overlaps between two clusters selected from the respective cluster sets are displayed as the height of the blocks (Fig. 2). More specifically, we evaluated the number of common genes between the two different clusters by using hypergeometric probability distributions (17). Assuming that the generation of gene clusters is a random selection from among the total set of genes, the probability of observing at least *k* overlapping genes between randomly selected n_1 genes and n_2 genes from among all of the *g* genes is given by:

$$P(g,n_1,n_2,k) = 1 - \sum_{i=k}^{k-1} \frac{{}_{n_2}C_i \cdot {}_{g-n_2}C_{n_1-i}}{{}_{g}C_{n_1}} [=P(g,n_2,n_1,k)] \qquad (3)$$

When the *P* value is small, the overlap is regarded as statistically meaningful. Thus we defined the evaluation value of the overlap as:

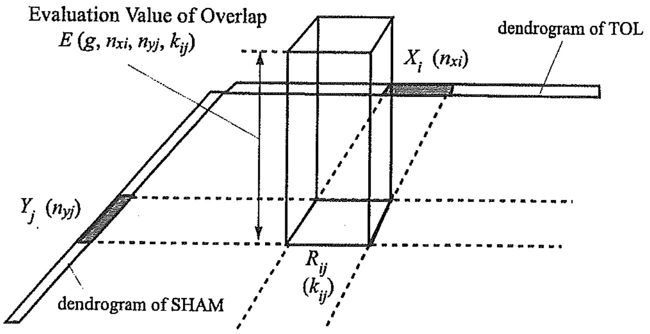


Fig. 2. Overlap block of two clusters. The dendrogram of TOL is mapped to the *x*-axis, and that of SHAM is mapped to the *y*-axis. Then, for the area (R_{ij}) determined by a cluster on the *x*-axis (X_i) and a cluster on the *y*-axis (Y_j), a block whose height represents $E(g,n_{xi},n_{yj},k_{ij})$ (statistical evaluation values of the overlaps between X_i and Y_j) is displayed, where *g* is the total number of genes, n_{xi} is the number of genes in X_i, n_{yj} is the number of genes in Y_j, and k_{ij} is the number of overlap genes between X_i and Y_j.

$$E(g,n_1,n_2,k) = -\log_{10}P(g,n_1,n_2,k) \qquad (4)$$

Then in the area (R_{ij}) determined by a cluster on the x-axis (X_i) and a cluster on the y-axis (Y_j), a block whose height represents $E(g,n_{xi},n_{yj},k_{ij})$ is displayed, where n_{xi} is the number of genes in X_i, n_{yj} is the number of genes in Y_j, and k_{ij} is the number of overlapping genes between X_i and Y_j (Fig. 2). We term this block an "overlap block." Note that the number of UniGenes, to which probes in a cluster correspond through their original GenBank accession number, was used as the number of genes. In this report, all 8,737 probes on RG-U34A were corresponding to 5,249 UniGenes (g = 5,249).

For hierarchical clustering, there are a large number of clusters generated at various distance levels. Our algorithm examines the overlaps of the genes between all combinations of two clusters with smaller "distance level" values than the "cut level," which is a threshold value specified by users (Fig. 1). In other words, we evaluated and visualized any clusters with a smaller distance level than the cut level, even if they were included in other clusters. Note that conventional hierarchical clustering does not focus on subclusters that are included in other clusters. Since all of the statistically significant combinations between cluster sets can be visualized simultaneously, users can grasp the overall picture of the relationships between the two different cluster sets.

In the CODM, all of the clusters are dealt with equally without regard to their difference level (i.e., their homogeneity). Even if they are included in other clusters, all of the statistical significance of the number of common genes between clusters is simultaneously visualized. Therefore, there is a risk that a small overlap block may be hidden by a large block. For example, assume that the clusters X_j and Y_n are included in X_i and Y_m respectively. Then, if the evaluation value E_{jn} is less than E_{im}, then the small block B_{jn} will be hidden in the large block B_{im} (Fig. 3*A*). To avoid this problem, the CODM allows the user to change the cut level interactively. That is, if the user decreases the cut level, some small blocks that are hidden in larger blocks will emerge. Therefore, in consideration of the homogeneity of clusters and the relationships with other gene information, the user can find important genes displayed as blocks in the CODM.

Color of Each Overlap Block

Since the statistical significance of the number of common genes between two different clusters is represented as the height of a block, the color of a block can be used to represent other information. In the current prototype, the CODM provides three color modes.

1) Redundant visualization. The first mode is a representation of the evaluation values of overlaps using a gray scale. This redundant representation helps users comprehend the distribution of the relative evaluation values of overlaps.

2) Similarity of expression patterns. The second mode is a representation of the similarity of expression patterns between two clusters, from red to blue. The similarity $f(T,S)$ of expression patterns between *cluster T* on TOL and *cluster S* on SHAM was defined using the average of the square of the Euclidean distance between them. Assuming that N_{TS} is the number of common genes in T and S, x_{ki} and y_{ki} are normalized expression levels of a common gene k at time T_i on TOL and SHAM, respectively. The similarity $f(T,S)$ was defined as follows:

$$f(T,S) = 1 - \frac{1}{N_{TS}}\sum_{k=1}^{N_{TS}}\sum_{i=1}^{12}(x_{ki} - y_{ki})^2 \qquad (5)$$

Since $\{x_{ti}\}$ and $\{y_{si}\}$ (i = 1,2,. . .12) satisfy *Eqs. 1* and *2*, the range of $f(T,S)$ is −1 to 1, and $f(T,S)$ can be rewritten as follows (See APPENDIX):

$$f(T,S) = \frac{1}{N_{TS}}\sum_{k=1}^{N_{TS}}\sum_{i=1}^{12}2x_{ki}y_{ki} \qquad (6)$$

In the CODM, the similarity $f(T,S)$ was represented as the color of the block from red ($f(T,S)$ = 1) to blue ($f(T,S)$ = −1). Roughly speaking, red indicates that expression patterns between the two clusters are similar, and blue indicates they have a negative correlation. In addition, purple ($f(T,S)$ = 0) indicates they have no correlation, or genes of one cluster have no changes in expression levels, i.e.,

$$\forall\, x_{ki} \approx 0 \text{ or } \forall\, y_{ki} \approx 0$$

As mentioned above, if genes in a certain cluster based on SHAM also constitute a cluster in TOL, but the expression level in SHAM is significantly different from that in TOL, then these genes provide potential markers for the cause of ischemic tolerance. Strong candidates will appear as tall blue or purple blocks. CODM allows users to easily look for such blocks, with interactively controlling the thresholds.

3) Relationship with a known gene classification. The third type of information is a representation of the relationship between overlapping genes and a known gene classification. If statistically significant representation of genes within a particular class is observed among the overlapping genes, then the block is color coded according to the class. The level of statistical significance of the representation of genes within a particular class is evaluated using *Eq. 3*, where g is the total number of genes that are classified by the known classification, n_1 is the number of genes that are classified by the known classification among overlapping genes, n_2 is the total number of genes within a class based on the known gene classification, and k is the observed number of genes found in both the given overlapping genes and the given class according to the known gene classification.

In this report, we associated overlapping genes with eight types of transcription factors (HIF, ARNT, and EGR families) that were reported to have a relationship with ischemia (5, 8, 18, 19). We extracted complete sequences of 1.0 kb upstream and 0.1 kb downstream for 2,816 UniGenes among the 5,249 UniGenes corresponding to 8,737 probes on the RG-U34A microarray. The 1.1-kb sequences of the 2,816 UniGenes were searched to determine whether they correspond to the TRANSFAC matrices v7.2 (11) with the threshold set to "minimum false negative." Table 1 shows the names of the transcription factors, the number of UniGenes that correspond to each transcription factor, and the thresholds for matching. In CODM, we color

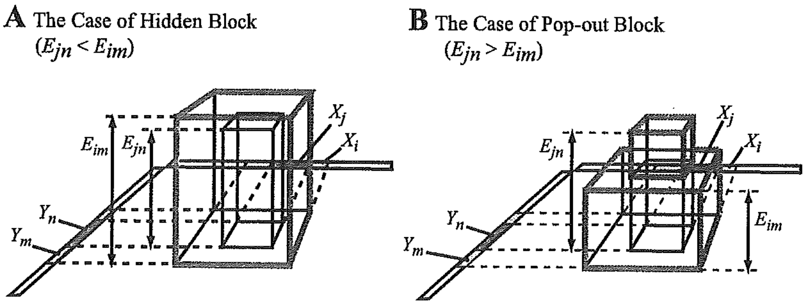


Fig. 3. Relationships of two blocks. In CODM, all of the clusters are dealt with equally, regardless of their difference levels (i.e., their homogeneity). Even if they are included in other clusters, all of the statistical significance of the number of common genes between clusters is simultaneously visualized. There is a risk that a small overlap block may be hidden in a large block. Assume that the clusters X_j and Y_n are included in X_i and Y_m, respectively. Then, if the evaluation value E_{jn} is less than E_{im}, the small block B_{jn} will be hidden within the large block B_{im} (*A*).

Table 1. *Transcription factors linked to ischemia*

Transcription Factor	No. of UniGenes	Thresholds
V$AHRARNT_01	540	0.92
V$AHRARNT_02	4	0.91
V$HIF1_Q3	955	0.55
V$HIF1_Q5	507	0.87
V$EGR1_01	143	0.87
V$EGR2_01	92	0.89
V$EGR3_01	26	0.93
V$ENGFIC_01	143	0.88

In the cluster overlap distribution map (CODM), changes in the composition of the cluster sets and changes in the expression patterns between different conditions were associated with 8 types of transcription factors (HIF, ARNT, and EGR families), which are all known to mediate response to ischemia. We extracted UniGenes that contain putative binding sites for the transcription factors and correspond to probes on RG-U34A GeneChips (Affymetrix, Santa Clara, CA). Shown are the names of the transcription factors, the number of UniGenes, and the thresholds for matching.

coded overlap blocks that contain statistically meaningful numbers of genes with putative transcription factor binding sites. If an overlap block represents statistical significance for multiple transcription factors' putative binding sites, then only a single transcription factor with the highest evaluation value was visualized. However, the CODM allows users to click overlap blocks and browse description messages (in a console window) for the relationships with all of the transcription factors.

RESULTS AND DISCUSSION

Figure 4 shows the visualization results of the comparison between TOL and SHAM in the mode of redundant visualization, the similarity of the expression patterns, and the relationships with known gene classifications (transcription factors). In Fig. 4, the cut level for the distance for hierarchical clustering was 0.74, and all overlap blocks with 2.0 or higher evaluation values are displayed as a 3D histogram. As Fig. 4 shows, the CODM provides not only a 3D mode but also a two-dimensional (2D) mode where users can see a projected overhead view of the 3D mode. In the 3D mode, the statistical significance of the overlaps between clusters and the differences in expression levels between the clusters can be simultaneously represented, since we can use the height and color of blocks. However, it is somewhat difficult to recognize the expression patterns of clusters that generate an overlapping block. For this purpose, the 2D mode is better, although the 2D mode of CODM can visualize only a single species of information at a time, i.e., the statistical significance of the overlaps or the differences in expression levels between clusters, or relationships with known gene classification. Therefore, it is useful to interactively change the mode as required. Exploration by changing the color mode and the 2D and 3D modes allowed us to pick up three potentially important overlap blocks (Fig. 4). The information for these three overlap blocks is shown in Table 2, their gene lists are shown in the Supplemental Material, and their expression patterns are shown in Fig. 5. (The Supplemental Material is available at the *Physiological Genomics* web site.)[1]

As stated above, we assumed that there are four issues for a comparison of clustering results: changes in the composition of the cluster sets, changes in the expression patterns, relationships with other known gene information, and threshold problems. The CODM enables us to address these issues as follows.

Changes in the Composition of the Cluster Sets

As shown in Fig. 4, *A* and *B*, the CODM can intuitively visualize changes in the composition of the cluster sets as 3D histograms. That is, the dissimilarity of the expression level under SHAM divides each cluster on TOL into specific subclusters, and these subclusters are displayed along the *y*-axis. In the same manner, the relationships between each cluster of SHAM and all of the clusters of TOL are displayed on the *x*-axis. If a clustering analysis is conducted for the merged data of TOL and SHAM, then these subclusters would be scattered and it would be difficult to intuitively observe the relationships of the compositions of the cluster sets.

Changes in the Expression Pattern

A comparison of the dynamic changes of gene expression level across time under various conditions provides a useful tool for interpreting complex biological processes. However, there are generally many false candidate genes whose expression patterns between two different conditions are different purely by chance. For the comparison between TOL and SHAM, only 357 probes (of the 3,363 selected probes) had 0.8 or higher correlation coefficient values of expression pattern between the two conditions. On the other hand, 756 probes had negative correlation coefficient values. As stated above, the difference of macroscopic phenomena that the conditions exhibit results from the difference of expression of not a single gene but of multiple genes. Therefore, it is quite important to search for genes whose expression patterns changed in a similar fashion between different conditions. Figure 4, *C* and *D*, shows that the CODM can simultaneously depict the statistical significance of the overlaps between clusters and the differences in their expression patterns. In this mode, tall blocks colored blue or purple, such as *blocks B* and *C*, would be good candidates, since their similarities of expression patterns were negative (−0.28 and −0.23), while the two clusters under different conditions share a statistically meaningful number of common genes (E = 53.3 and E = 34.8). Note that the objective of the CODM is to identify such potentially important pairs of clusters from massive combinations. To further understand the significance of the expression patterns, it would be a desirable approach to combine CODM with other visualization tools for line graphical view of expression patterns, as shown in Fig. 5. The expression of genes in TOL in *block B* was upregulated, compared with SHAM, at early stage, i.e., 1 h, 3 h, and 12 h. On the other hand, the expression of genes in TOL in *block C* was downregulated, compared with SHAM, at early stage, i.e., 1 h, and 3 h. Once again, CODM enabled us to easily detect candidate genes of this type.

Integration with Other Known Gene Information

In gene expression analysis, interpretation and validation of the results should be performed in the context of what is already known about the genes being analyzed. CODM allows us to associate the results with other such gene information and

[1] The Supplemental Material (Supplemental Tables S1–S3) for this article is available online at **http://physiolgenomics.physiology.org/cgi/content/full/00107.2004/DC1.**

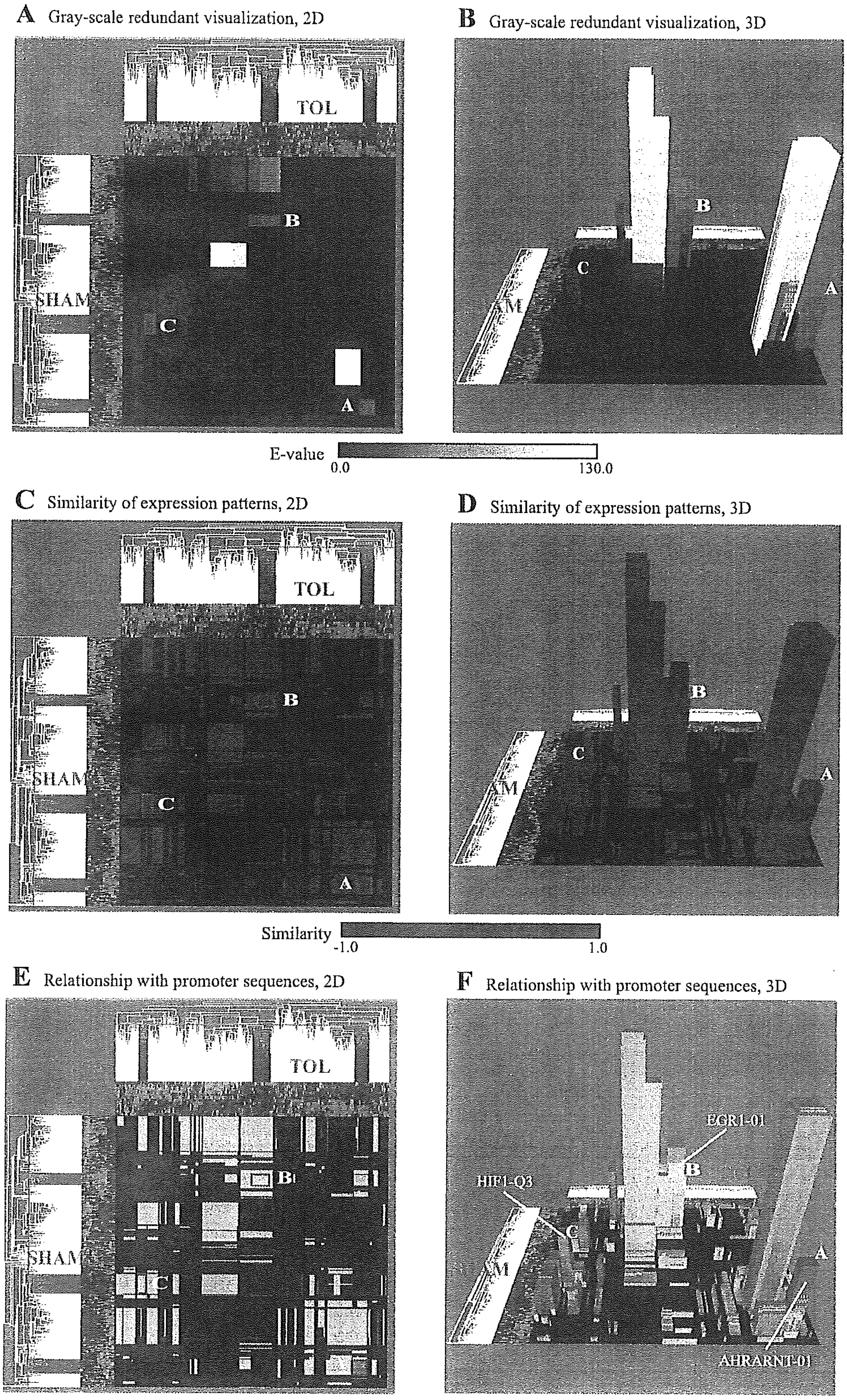


Fig. 4. Visualizations for comparison of clustering results of TOL and SHAM. These are visualization results of the comparisons between TOL and SHAM in the mode of redundant visualization (*A* and *B*), similarity of the expression patterns (*C* and *D*), and the relationships with transcription factors (*E* and *F*). Here, the cut level of the distance for hierarchical clustering was 0.74, and all of the overlap blocks with 2.0 or higher evaluation values are displayed as three-dimensional (3D) histograms. As shown, the CODM provides not only a 3D mode (*B*, *D*, and *F*) but also a two-dimensional (2D) mode (*A*, *C*, and *E*) where users can see a projected overhead view of the 3D mode. In the mode showing the relationships with the transcription factors (*E* and *F*), we considered the relationships with 8 types of transcription factors (HIF, ARNT, and EGR families) that are known to mediate response to ischemia. Here, only overlap blocks with 2.0 or higher evaluation values of the number of genes with putative transcription factor binding sites were color coded. Where an overlap block represents statistical significance for multiple transcription factors' putative binding sites, only the transcription factor with the highest evaluation value was visualized. Exploration through changing the color mode and the 2D and 3D mode allowed us to pick up three potentially important overlap blocks that represented high evaluation values of the number of genes with the binding sites ($E > 2.0$).

narrow down candidates. Figure 4, *E* and *F*, shows the relationships between eight types of transcription factors (HIF, ARNT, and EGR families; see Table 1) that were reported to have a relationship with ischemia (5, 8, 18, 19). In Fig. 4, overlap blocks with 2.0 or higher evaluation values for the representation of genes with putative transcription factor binding sites were color coded. Table 2 shows that overlap *blocks A*, *B*, and *C* implied a relationship with the transcription factors ($E > 2.0$). This example illustrates the utility of representing relationships with other known gene-associated information by

Table 2. *Information about 3 overlap blocks*

Overlap Block	No. of UniGenes in Cluster of TOL	No. of UniGenes in Cluster of SHAM	No. of Common UniGenes (Evaluation Value)	Similarity $f(T,S)$	Binding Sites of Transcription Factors: No. of Genes (Evaluation Value)
A	156	147	54 (E = 46.9)	0.42	V$AHRARNT_01:14 ($E$ = 2.10)
B	190	132	60 (E = 53.3)	−0.28	V$EGR1_01:6 ($E$ = 2.01)
C	99	207	43 (E = 34.8)	−0.23	V$HIF1_Q3:11 ($E$ = 2.33)

Exploration with CODM allowed us to pick up 3 potentially important "overlap blocks.' The "No. of UniGenes in Cluster of TOL(/SHAM)" is the number of UniGenes which correspond to probes included in a cluster of TOL(/SHAM). The "No. of Common UniGenes" is the number of common genes shared between the clusters of TOL and SHAM, and its statistical evaluation value, (E,) is shown in parentheses. The "Similarity f (T,S)" is the similarity of the expression patterns between the clusters of TOL and SHAM. The range of similarity f (T,S) is −1 (dissimilar) to 1 (similar). The "Binding Sites of Transcription Factors" shows the name of putative binding sites of transcription factors, the number of common genes that share the same binding sites, and the E value of the number of common genes with the same binding sites, if the evaluation value is 2.0 or higher. TOL, induced ischemic tolerance; SHAM, shamoperation.

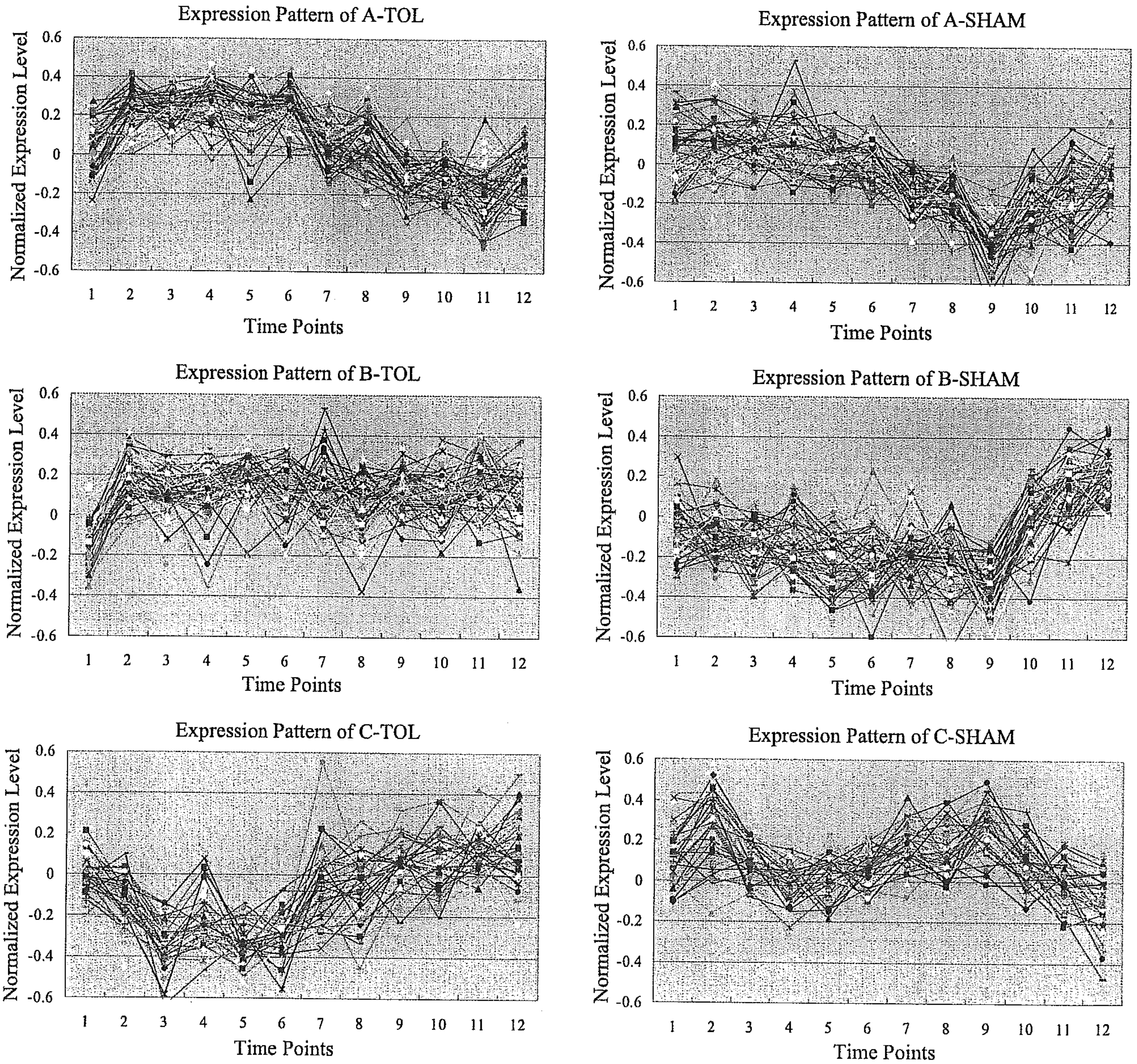


Fig. 5. Expression patterns of genes in the three overlap blocks. These are the expression patterns of common genes for the three overlap blocks that were picked up through exploration with CODM (Fig. 4). The "Expression Patterns of Cluster $T_i(/S_i)$" (i = a,b,c) are the expression patterns of the common genes of the *overlap block i* in TOL(/SHAM).

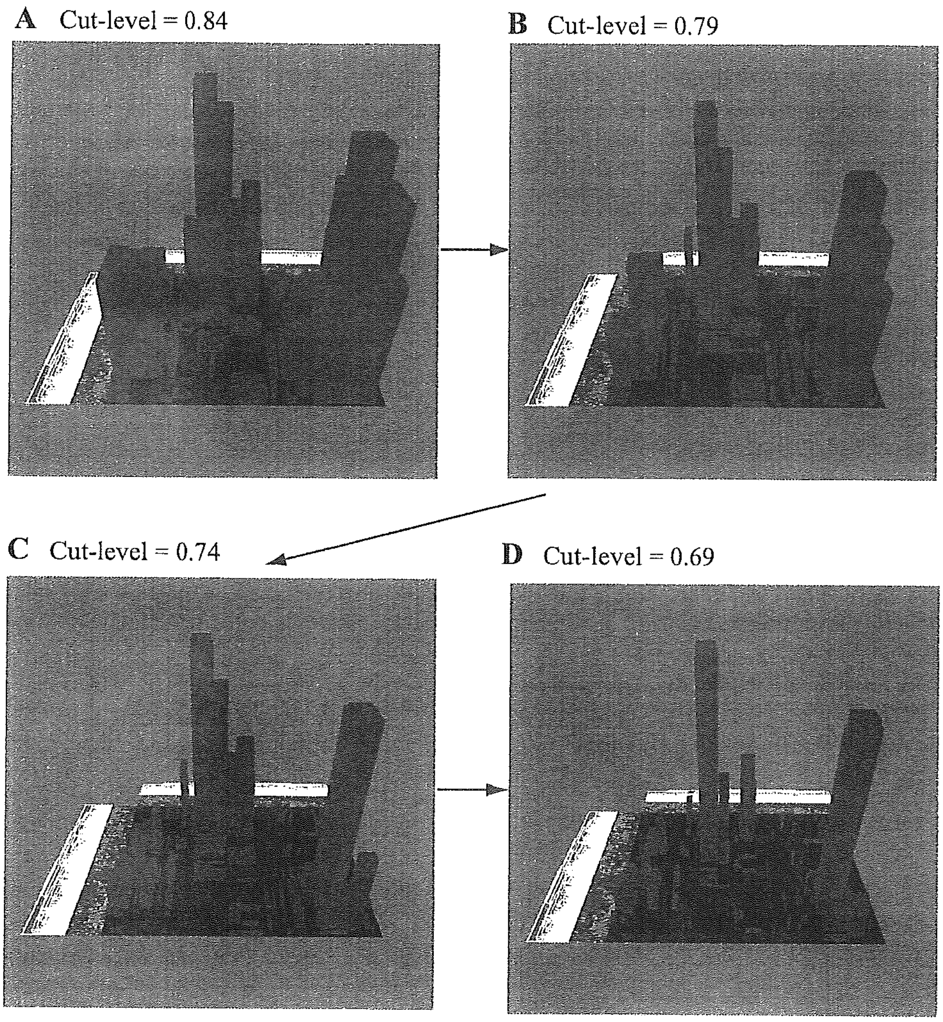


Fig. 6. Interactive changes of cut levels. In CODM, there is a risk that a small overlap block may be hidden in a large block. To avoid this problem, CODM allows the user to change the cut level interactively. If the user decreases the cut level, then some small blocks that are hidden in larger blocks will emerge. By considering the homogeneity of clusters and the relationships with other gene information, the user can find important genes displayed as blocks in the CODM.

use of the color of overlap blocks, although it may be difficult to extract biological conclusions because of the limited number of genes with the putative binding sites in the overlap blocks. If binding site information from more genes becomes available, then more detailed analysis of results will be possible. Furthermore, representation of relationships with other known gene classifications should provide us with deeper insights.

Threshold Problems

Arbitrary selection of thresholds involves a risk of overlooking important genes. In a comparison of cluster sets on gene expression profiles, there are four types of thresholds: *1*) a threshold for generating clusters for each condition; *2*) a threshold for evaluating the number of common genes that two clusters share; *3*) a threshold for evaluating the differences in the expression patterns between two clusters; and *4*) a threshold for evaluating the relationship with other known gene information. The CODM reduces the number of thresholds and allows users to interactively change the thresholds as follows.

1) Threshold for generating clusters for each condition. Since conventional hierarchical clustering does not focus on subclusters that are included in other clusters, there is a risk that the important subclusters could be overlooked. In the CODM, overlaps of genes between any two clusters of TOL and SHAM are statistically evaluated, even if these are included in other clusters. In addition, the CODM allows users to interactively change the cut level, to reduce the risk that a small overlap block may be hidden in a large block (Fig. 6). Therefore, by considering the homogeneity of clusters and the relationships with other known gene information, the user should be able to find the important genes displayed as blocks.

2) Threshold for evaluating the number of common genes shared by two clusters. In CODM, the statistical significance of the number of common genes between two different clusters is represented as the height of a block, and statistical significances of the overlap of all combinations of clusters are displayed as a 3D histogram at the same time. Therefore, without the selection of an arbitrary threshold, the distribution of the statistical significance of the overlap is effectively displayed. Although (to reduce the rendering load) Fig. 4 shows only overlap blocks with 2.0 or higher evaluation values of the overlap, users can interactively change this value.

3) Threshold for evaluating the differences in the expression patterns between two clusters. CODM represents the differences in the expression patterns between two clusters by the color of the blocks ranging from red to blue. Therefore, the distribution of differences in the expression patterns of all