ELSEVIER

International Journal of Cardiology 103 (2005) 105–106

International Journal of Cardiology

www.elsevier.com/locate/ijcard

Letter to the Editor

Large collateral conus branch to the left anterior descending branch of the coronary artery in a subject with angina pectoris demonstrated by multislice computed tomography

Nobusada Funabashi*, Miki Asano, Issei Komuro

Department of Cardiovascular Science and Medicine, Chiba University Graduate School of Medicine, 1-8-1 Inohana, Chuo-ku, Chiba City, Chiba 260-8670, Japan

Received 20 March 2004; accepted 17 June 2004
Available online 5 February 2005

A 70-year-old male presented to our hospital for chest pain. To evaluate the coronary arteries, electrocardiogram-gated enhanced multislice computed tomography (CT) (Light Speed Ultra, General Electric, Milwaukee, WI, USA) was performed with a 1.25-mm slice thickness, helical pitch 3.25. Thirty seconds after intravenous injection of 100 ml of iodinated contrast material (350 mgI/ml), CT scanning was performed with retrospective ECG-gated reconstruction and volume data were transferred to a workstation (Virtual Place Office Azemoto Tokyo Japan).

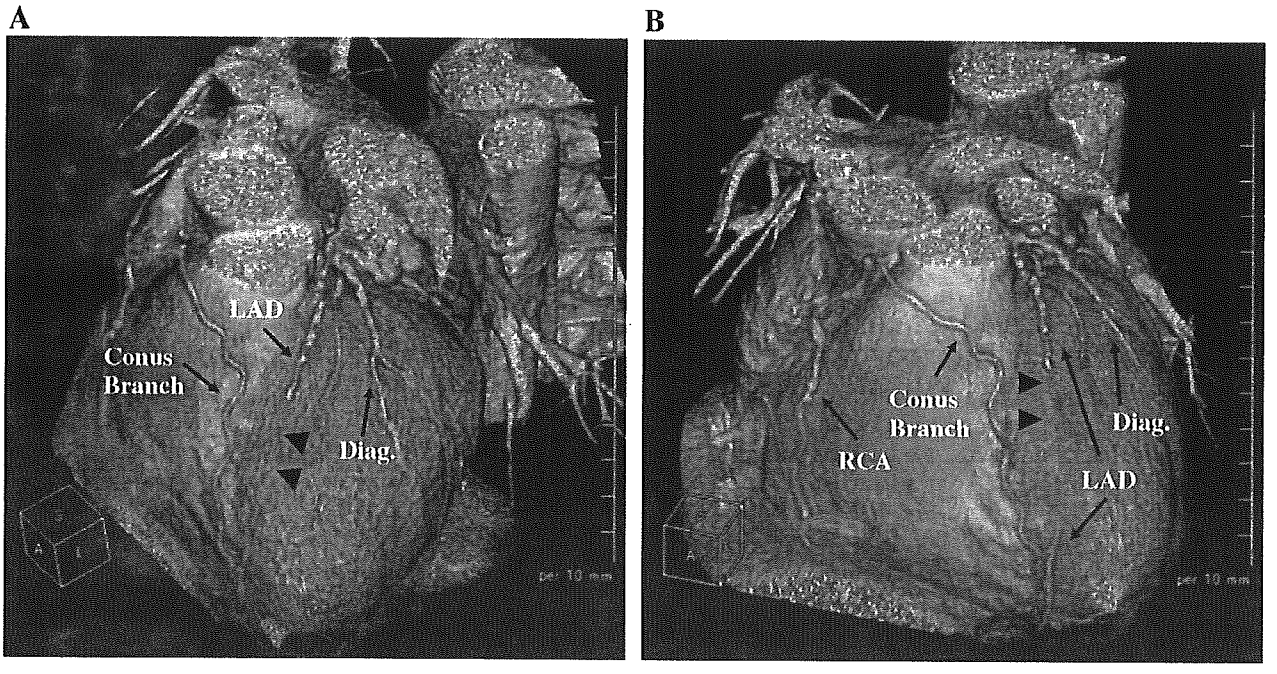


Fig. 1. Three dimensional volume-rendering images of enhanced electrocardiogram-gated multislice computed tomography revealed total occlusion of the left anterior descending branch (LAD) (arrowheads) and a large conus branch originating from the right coronary artery (RCA) that fed into the distal part of the LAD. Diag. indicates diagonal branch.

* Corresponding author.
E-mail address: nobusada@ma.kcom.ne.jp (N. Funabashi).

0167-5273/$ - see front matter
doi:10.1016/j.ijcard.2004.06.017

Three dimensional volume-rendering images revealed total occlusion of the left anterior descending branch (LAD) (arrowheads, Fig. 1A,B) and a large conus branch originating from the right coronary artery (RCA) that fed into the distal part of the LAD. Conventional coronary angiograms revealed the same findings of total occlusion of the LAD (arrowheads, Fig. 2A) and a collateral conus branch originating from the RCA that fed into the distal part of the LAD (Fig. 2B). As blood flow of the collateral artery was good, normal motion of the left ventricle was revealed by a selective left ventriculogram, and his chest pain was improved by the oral administration of nitroglycerin and a β-blocker, no interventions were performed.

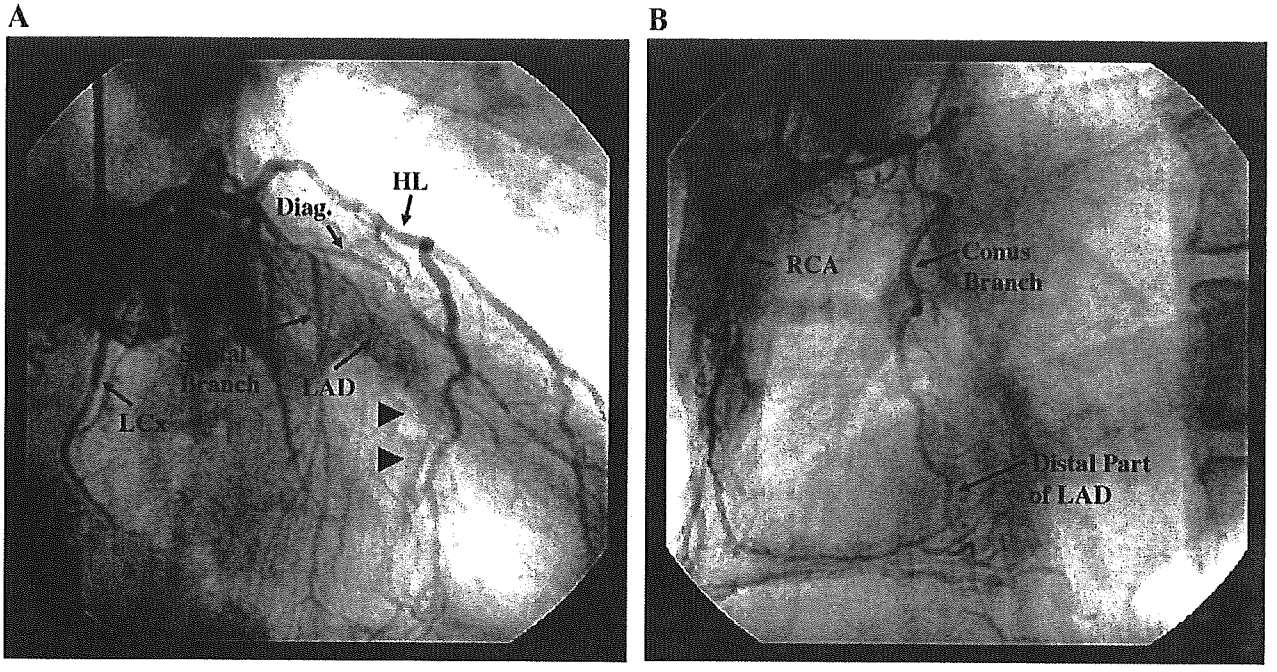


Fig. 2. Conventional coronary angiograms revealed the same findings of total occlusion of the LAD (arrowheads, A) and a collateral conus branch originating from the RCA that fed into the distal part of the LAD (B). Diag., HL, and LCx indicate diagonal branch, high lateral branch, and left circumflex branch, respectively.

ELSEVIER

International Journal of Cardiology 106 (2006) 270–272

International Journal of Cardiology

www.elsevier.com/locate/ijcard

Letter to the Editor

Atherosclerotic right internal thoracic arterial aneurysm demonstrated by multislice computed tomography

Yuriko Niitsuma[b], Nobusada Funabashi[a,*], Mizuho Imamaki[b], Issei Komuro[a], Masaru Miyazaki[b]

[a]*Department of Cardiovascular Science and Medicine, Chiba University Graduate School of Medicine, 1-8-1 Inohana, Chuo-ku, Chiba City, Chiba 260-8670, Japan*

[b]*Department of General Surgery, Chiba University Graduate School of Medicine, Chiba, Japan*

Received 30 October 2004; accepted 31 December 2004
Available online 1 April 2005

Keywords: Internal thoracic arterial aneurysm; Multislice computed tomography; Coronary artery bypass graft

A 74-year-old male presented with weight loss; stomach cancer was diagnosed that required surgery. As he had chest pains on effort, a conventional coronary angiogram was performed, which revealed severe stenosis of the left main branch, and a coronary artery bypass graft (CABG) was indicated.

Evaluation of the aorta and internal thoracic artery (ITA) was done using ECG-gated enhanced multislice computed tomography (CT) (Light Speed Ultra 16, General Electric, Milwaukee, Wisconsin) with a 1.25 mm slice thickness, helical pitch 6.00. CT scanning was performed 30 s after intravenous injection of 100 ml of iodinated contrast material (350 mgI/ml). Axial source (Fig. 1A) and sagittal view (Fig. 1B) of multiplanar reconstruction image and volume-rendered images (Fig. 1C and D), revealed normal findings except for a right ITA (RITA) aneurysm (arrowheads) and aortic arch calcification. The surgeons had planned to connect the left ITA to the left circumflex branch and, as stomach cancer precluded the use of the gastro-epiploic artery, they chose the RITA to connect to the left anterior descending (LAD) artery. During the CABG procedure the RITA aneurysm (10 mm in diameter) was resected (arrow Fig. 2A). The radial artery was connected to the proximal portion of the RITA and the radial arterial graft to LAD. Pathological examination of the resected aneurysm revealed plaque (★) with thickened intima, fragmentation of the membrane elastic interna (arrow) (Fig. 2B) and excessive cholesterin deposition in intima. Therefore the RITA aneurysm was considered due to atherosclerosis rather than specific arteritis or systemic connective tissue disease.

* Corresponding author. Tel.: +81 43 222 7171.
E-mail address: nobusada@ma.kcom.ne.jp (N. Funabashi).

0167-5273/$ - see front matter
doi:10.1016/j.ijcard.2004.12.081

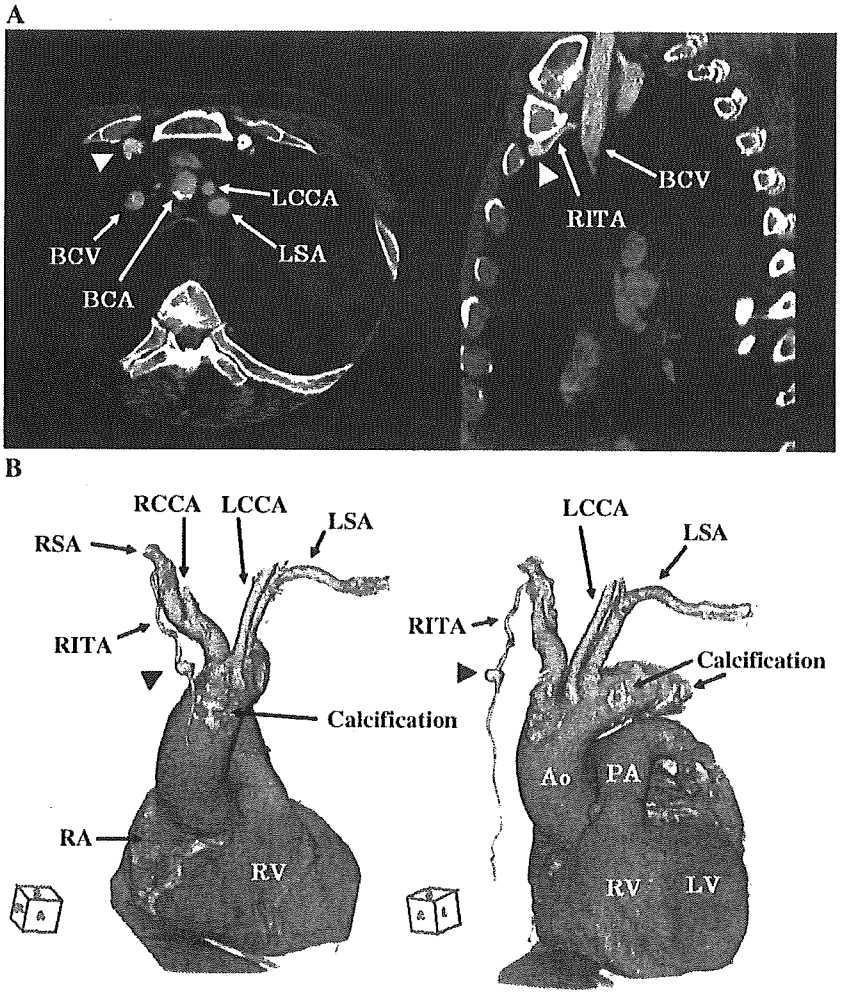


Fig. 1. A and B: Axial source (A) and sagittal view (B) of multiplanar images of enhanced electrocardiograph-gated multislice computed tomography show the aneurysm of the right internal thoracic artery (RITA) (arrowheads). (C) and (D) Volume-rendered images of enhanced electrocardiogram-gated multislice computed tomography show the aneurysm of the right internal thoracic artery (RITA) (arrowheads). Calcification of aortic arch can also be observed. BCV (brachiocephalic vein), BCA (brachiocephalic artery), LCCA (left common carotid artery), LSA (left subclavian artery), RSA (right subclavian artery), RCCA (right common carotid artery), RA (right atria), RV (right ventricle), Ao (aorta), PA (pulmonary artery), and LV (left ventricle).

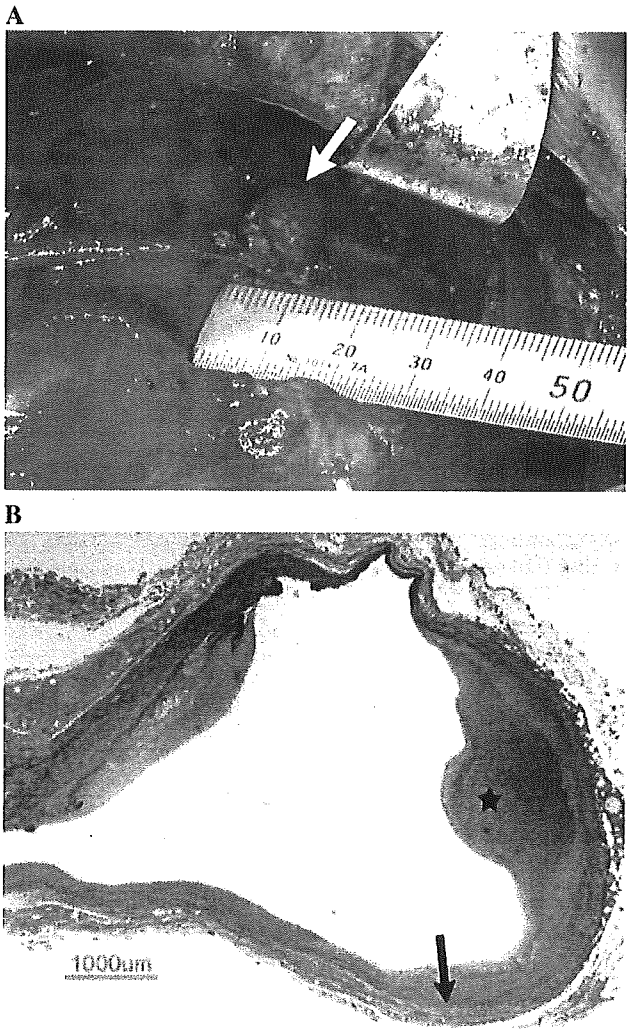


Fig. 2. (A) Intraoperative photograph of the aneurysm of the right internal thoracic artery (RITA) (arrow). (B) Histological section of the wall of the aneurysm of the RITA in Elastica van Gieson stein represented plaque (★) with thickened intima and fragmentation of the membrane elastica interna (arrow). Original magnification ×20.

ELSEVIER

International Journal of Cardiology 107 (2006) 130–131

International Journal of Cardiology

www.elsevier.com/locate/ijcard

Letter to the Editor

Patency of gastroepiploic arterial graft to left circumflex branch with distal portion of the anastomotic site demonstrated by multislice computed tomography

Nobusada Funabashi*, Issei Komuro

Department of Cardiovascular Science and Medicine, Chiba University Graduate School of Medicine, Chiba University Hospital, 1-8-1 Inohana, Chuo-ku, Chiba City, Chiba 260-8670, Japan

Received 2 January 2005; accepted 6 January 2005
Available online 5 April 2005

A 77-year-old male presented to our hospital with chest pain on effort 5 years previously. Conventional coronary angiogram (CAG) revealed occlusion in the proximal left circumflex branch (LCx) and right coronary artery (RCA), with distal collateral arteries and occlusion of the ostium of the left subclavian artery (LSA). He underwent a coronary artery bypass connecting the aortic root to the mid portions of the RCA using a saphenous vein graft (SVG), and a gastroepiploic arterial (GEA) graft to the distal LCx. An artificial graft was also implanted from the ascending aorta to the mid portion of the LSA. Postoperatively, CAG revealed a patent GEA graft but completely occluded ostium of the SVG. Five years later, although asymptomatic, electrocardiogram (ECG)-gated enhanced multislice computed tomography (CT) (Light Speed Ultra 16, General Electric, Milwaukee, WI) was performed with 1.25-mm slice thickness, helical pitch 6.00. Thirty seconds after intravenous injection of 100 ml of iodinated contrast material (350 mg/ml), CT scanning was performed with retrospective ECG-gated reconstruction and volume data were transferred to a workstation (Virtual Place Office Azemoto, Tokyo).

Volume-rendered images revealed occlusion of the LSA and patent mid and distal portions fed by the artificial graft (Fig. 1A,B). The SVG was completely occluded at the ostium. The proximal LCx was occluded and the distal portion of the anastomotic site of the GEA graft was visualized (arrowheads Fig. 2A,B), findings identical to the previous CAG.

* Corresponding author.
E-mail address: nobusada@ma.kcom.ne.jp (N. Funabashi).

0167-5273/$ - see front matter
doi:10.1016/j.ijcard.2005.01.049

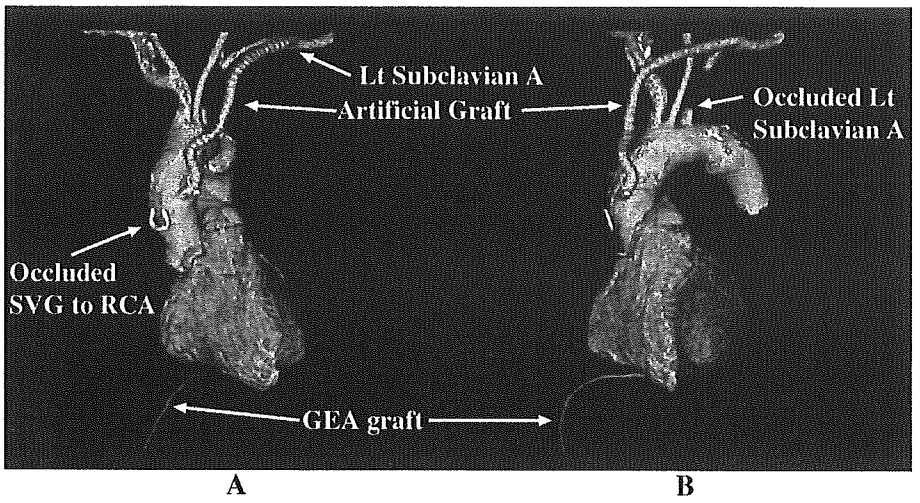


Fig. 1. Volume-rendered images of enhanced ECG-gated multislice computed tomography from the anterior view (A) and left anterior view (B) revealed the occluded proximal portion of the left subclavian artery (Lt Subclavian A), and the mid and distal portions of the Lt Subclavian A fed by the artificial graft from the ascending aorta with good patency. The saphenous vein graft (SVG), which should have connected to the right coronary artery (RCA), was completely occluded at the ostium of the graft. The gastroepiploic arterial (GEA) graft could also be visualized.

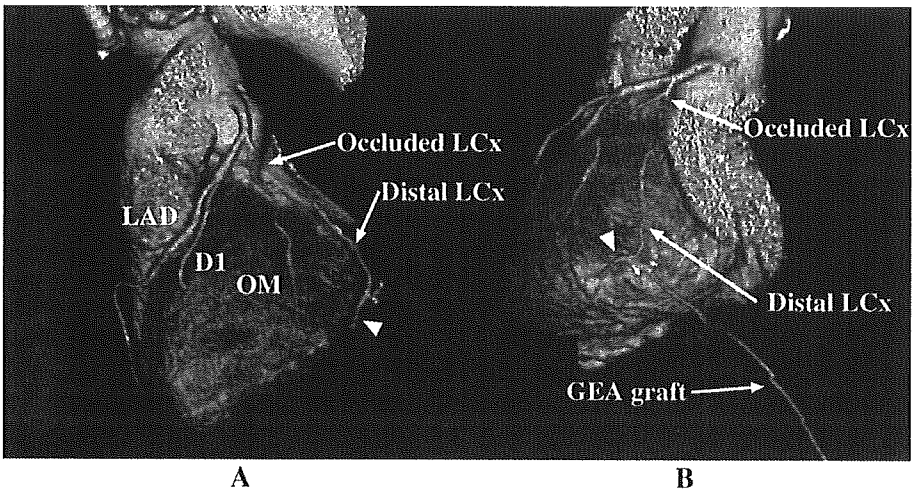


Fig. 2. Volume-rendered images of enhanced ECG-gated multislice computed tomography from the left superior posterior view (A) and posterior view (B) revealed that the proximal portion of the left circumflex branch (LCx) was occluded. The GEA graft was connected to the distal portion of the LCx and the distal portion of the anastomotic site of the GEA graft was visualized (Arrowheads). LAD, D1, and OM indicate left anterior descending branch, 1st diagonal branch, and obtuse marginal branch, respectively.

Available online at www.sciencedirect.com

SCIENCE DIRECT

ELSEVIER

Genomics 86 (2005) 127 – 141

GENOMICS

www.elsevier.com/locate/ygeno

Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues☆

Xijin Ge[a,b,*], Shogo Yamamoto[a], Shuichi Tsutsumi[a], Yutaka Midorikawa[a], Sigeo Ihara[a], San Ming Wang[b], Hiroyuki Aburatani[a,c,*]

[a]*Genome Science Division, Research Center for Advanced Science and Technology, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8904, Japan*

[b]*Center for Functional Genomics, ENH Research Institute, Northwestern University Feinberg School of Medicine, 1001 University Place, Evanston, IL 60201, USA*

[c]*CREST, Japan Science and Technology Corporation (JST), Japan*

Received 9 November 2004; accepted 12 April 2005
Available online 9 June 2005

Abstract

A critical and difficult part of studying cancer with DNA microarrays is data interpretation. Besides the need for data analysis algorithms, integration of additional information about genes might be useful. We performed genome-wide expression profiling of 36 types of normal human tissues and identified 2503 tissue-specific genes. We then systematically studied the expression of these genes in cancers by reanalyzing a large collection of published DNA microarray datasets. We observed that the expression level of liver-specific genes in hepatocellular carcinoma (HCC) correlates with the clinically defined degree of tumor differentiation. Through unsupervised clustering of tissue-specific genes differentially expressed in tumors, we extracted expression patterns that are characteristic of individual cell types, uncovering differences in cell lineage among tumor subtypes. We were able to detect the expression signature of hepatoctyes in HCC, neuron cells in medulloblastoma, glia cells in glioma, basal and luminal epithelial cells in breast tumors, and various cell types in lung cancer samples. We also demonstrated that tissue-specific expression signatures are useful in locating the origin of metastatic tumors. Our study shows that integration of each gene's breadth of expression (BOE) in normal tissues is important for biological interpretation of the expression profiles of cancers in terms of tumor differentiation, cell lineage, and metastasis.

Keywords: Tissue-specific gene; Tumor differentiation; DNA microarray data interpretation; Breadth of expression; BRCA1; ESR1

Introduction

Genome-wide expression profiling with DNA microarrays has been widely used to identify new cancer subtypes and expression signatures associated with prognosis [1–5]. The expression data of thousands of tumor samples, each characterized by the expression levels of up to ~40,000 transcripts, are being quickly accumulated in the public repositories (reviewed in [6]). Due to technological limitations and the inherent complexity of the gene regulatory mechanism, such data are often noisy and extremely multivariate, leading to difficulties in data interpretation.

Besides the need for robust computational tools, integration of additional biological information about genes is essential for uncovering molecular mechanisms underling expression profiles. For example, functional categories from the Gene Ontology (GO) consortium [7], KEGG databases of molecular interaction pathways [8], and genome sequences of promoters are playing important roles in understanding a cluster of genes defined by expression profiling. In this paper, we introduce another kind of information that

☆ DNA microarray data from this article have been deposited with NCBI Gene Expression Omnibus (GEO) under accession: GSE2361.

* Corresponding authors. Hiroyuki Aburatani is to be contacted at Genome Science Division, RCAST, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8904, Japan. Fax: +81 3 5452 5355. Xijin Ge is to be contacted at ENHRI, 1001 University Place, Evanston, IL 60201, USA. Fax: +1 224 364 5003.

E-mail addresses: haburata-tky@umin.ac.jp (H. Aburatani), xge@northwestern.edu (X.J. Ge).

0888-7543/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2005.04.008

concerns each gene's expression pattern in a panel of normal tissues.

Only a small portion of the 30,000–40,000 protein-coding genes [9,10] in the human genome are essential to the survival of individual cells, hence are constitutively expressed in different types of tissues [11]. Transcription of most genes is regulated by a cell differentiation process, and thus is often highly variable among tissue/cell types and developmental stages. While ubiquitously expressed genes (so-called maintenance genes [11]) play key roles in basic cellular processes, tissue-specific genes are related to the functioning of particular organs. Although it is still difficult to obtain expression profiles for individual cell types that constitute normal organs, genome-wide expression profiles of bulk tissues has been carried out by serial analysis of gene expression (SAGE) and DNA microarrays [11–17]. For each gene, such studies define its breadth of expression (BOE) in normal tissues, which tell where a certain gene is expressed under normal physiological conditions. Categorization of genes based on BOE might serve as additional sources of information to help us decipher the complex expression profiles observed in cancers.

In this paper, we performed additional microarray experiments of normal tissues to search extensively for tissue-specific genes and then systematically reanalyzed previously published DNA microarray data of various cancers. We employed oligonucleotide microarrays to measure the expression of ~20,000 transcripts in 3 fetal and 33 adult normal human tissues (full list is given in Fig. 1A). Pooled RNA samples are used to maximize tissue coverage, which is important for defining tissue specificity. We retrieved data from a collection of previously published datasets of liver, brain, breast, and lung cancers. Then we focused on the genes that are specifically expressed in certain normal tissues but are differentially expressed in tumors arising from the same anatomical sites. Our strategy is to create a small but carefully selected dataset of normal tissue gene expression profiles, and use it as a seed to reanalyze large datasets in the public domain.

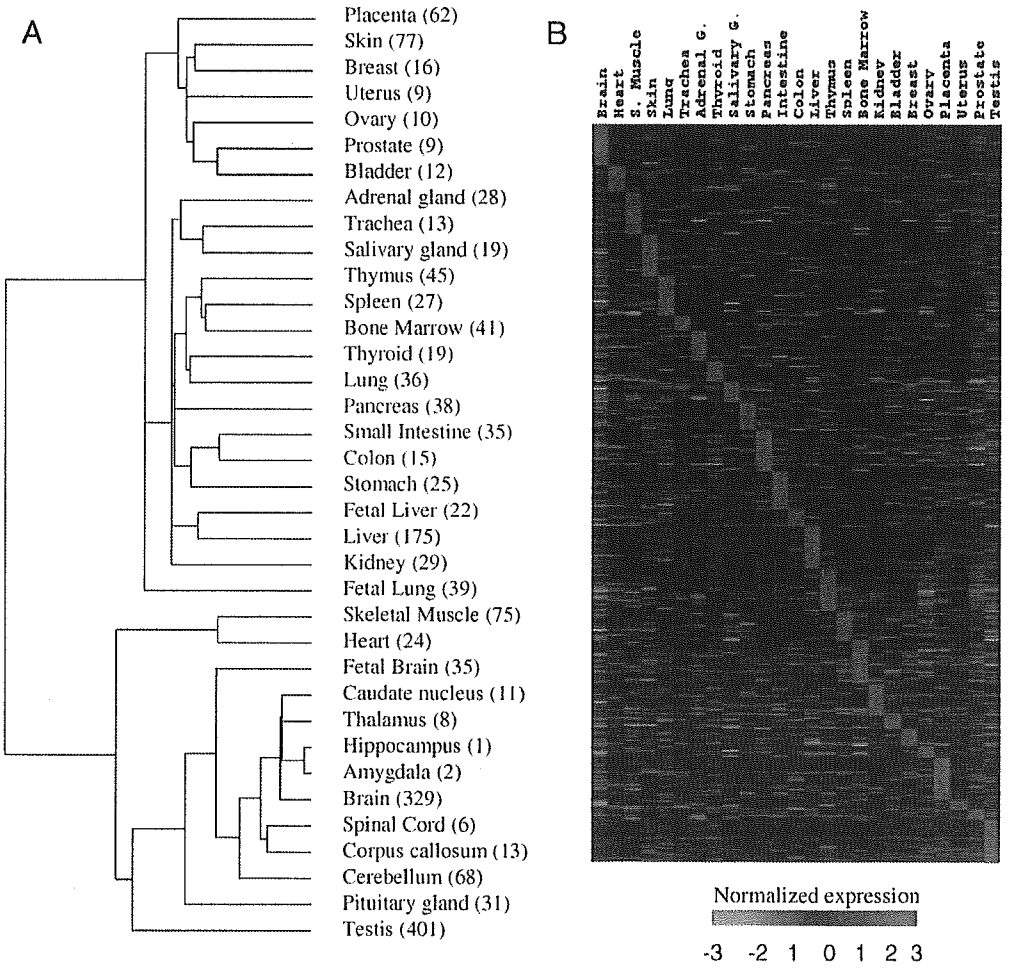


Fig. 1. (A) Hierarchical clustering of gene expression data of 36 types of normal human tissues. Data for 7396 genes are used to generate the cluster tree. Numbers in brackets indicate the number of tissue-specific genes. (B) Expression pattern of tissue-specific genes. Red denotes high expression and green low expression. Only the top 40 with highest specificity are shown for each tissue.

Results

Expression profiling of normal tissues

Using the Affymetrix U133A array, we performed expression profiles of 36 common normal human tissues, each represented by a pooled RNA sample (see Supplementary Information for details). The raw data are available at our web site: http://www.genome.rcast.u-tokyo.ac.jp/normal/, and can also be queried through a graphical web interface at http://www.lsbm.org. After eliminating genes with little variation (max/min > 2, max−min > 100, see Method for details) in their expression among tissues, we performed hierarchical clustering of the remaining 7396 probe sets. The result is shown in Fig. 1A. As expected, whole brain, brain regions, fetal brain, and spinal cord form a group related to the nervous system. Other closely related organs also aggregate, such as {colon, small intestine, stomach}, {heart, skeletal muscle}, {skin, breast} etc. The hierarchical tree of different tissues defined by expression patterns might reflect the intrinsic similarities between these tissues as a result of development.

Tissue-specific genes

We identified 1956 probe sets showing exclusively high expression in one of the 36 tissues (see Method for details). These probes map to 1687 UniGene clusters. The number of tissue-specific genes associated with each tissue is also given in Fig. 1A. The expression patterns of these genes are shown in Fig. 1B. Only the top 40 genes with highest Z scores are given. A full list of these genes can be found in the Supplementary Information and our web site http://www.genome.rcast.u-tokyo.ac.jp/normal/. We identified 401 testis-specific genes, 329 brain-specific genes, and 175 liver-specific genes. The remaining tissues have much fewer specific genes. For example, less than 20 are found for trachea, breast, colon, bladder, prostate, ovary, and uterus. This can be understood from the fact these tissues are less specialized and more similar to each other in their physiological organization.

We also identified 920 "tissue-selective" transcripts that are highly expressed in several related tissues (see Method for details). Unlike tissue-specific genes, tissue-selective genes are highly expressed in multiple tissues. For example, we identified 25 genes whose expression is restricted to colon and small intestine, 10 for heart and skeletal muscle, 9 for kidney and liver, 10 for brain and testis, etc. These 920 tissue-selective genes represent 816 UniGene clusters. A full list is available in the Supplementary Information and at our web site http://www.genome.rcast.u-tokyo.ac.jp/normal/. Together with the 1687 tissue-specific genes, we identified 2503 genes whose expression is strongly associated with specific tissues.

Ideally, multiple independent biological replicates representing each tissue type are needed to obtain a robust set of tissue-specific genes. Also more tissue types need to be covered to better define expression specificity. But due to the difficulty in obtaining normal samples we used commercially available pooled RNA. Because of such limitations in the resultant data and our empirical selection criteria, our lists of tissue-specific genes might be subject to false positive and false negative errors. This should be taken into account when using these lists for data interpretation. We reasoned that although tissue specificity of individual genes might be unreliable, it should still be possible to use these lists in a statistical sense by observing the coexpression of a group of such genes.

To validate our list of tissue-specific genes, we used the HuGe Index database [13], which contains biological replicates for some tissues. As shown in Supplementary Fig. 4, tissue-specific expressions of most of these genes can be seen in the HuGe Index database. A similar agreement with the Gene Expression Atlas database [15] is also observed (Supplementary Fig. 5). This agreement between independent datasets supports the effectiveness of sample pooling in our study and the reliability of our lists of tissue-specific genes.

In the following sections we study the expression of these tissue-specific genes in various cancers, starting from simple univariant liver cancer to multivariant lung tumors.

Hepatocyte-specific expression signature and differentiation of liver cancer

It is known that tumor cells sometimes could be transformed into a less differentiated state via a dedifferentiation process [22]. At the molecular level, one would expect the expression of tissue-specific genes to be decreased or lost. To confirm this, we reanalyzed a dataset of hepatocellular carcinoma (HCC) [23], which contains the data of 8 normal liver and 25 HCC samples. The cancer samples are further classified into three categories, namely well-differentiated ($N = 8$), moderately differentiated ($N = 12$), and poorly differentiated ($N = 5$). The expression levels of 12,600 transcripts are obtained with U95A oligonucleotide arrays (Affymetrix, Santa Clara, CA). Of the 175 liver-specific genes identified in the U133A array, 141 are covered by U95A arrays. So we retrieved expression data of these 141 transcripts.

We noted that 9 (6.4%) of these transcripts are often called "absent" in at least 4 normal liver samples, and hence do not show consistent tissue specificity. But the majority (129, or 75%) of these transcripts are called "present" in all of the 8 normal liver samples. When the expressions of these transcripts in other tissues are examined, only 5% (7/141) are all present in 17 normal lung samples [28]. Instead, 72% (102/141) are called absent in at least 80% of the normal lung samples. Even for those that are called present in both tissues, their expression levels in liver are on average 13.5 times higher. Similarly 77% (109/141) of these genes are called absent in at least 80% of normal prostate samples in another microarray datasets [29]. This assured us again the consistency of tissue specificity of the majority of these genes.

We also included 19 transcripts that are specifically expressed in fetal liver. After eliminating some genes with a

variation filter (max−min > 100, max/min > 2), we performed unsupervised clustering analysis with the remaining 64 transcripts. As shown in Fig. 2A, except Cluster γ, we observed a general tendency of increased levels of expression of these transcripts in the order of poorly, moderately, and well-differentiated tumor. Well-differentiated HCC samples are found to form a subcluster characterized by high expression of liver-specific genes. The remaining samples are further divided into two smaller groups, one dominated by poorly differentiated samples and the other moderately differentiated samples. Samples are arranged in the cluster tree according to their degree of differentiation.

Such classification is difficult if we analyze the expression of all genes in the microarray. For comparison, hierarchical clustering with the 3536 of 12,000 transcripts passed a variation filter is shown in Supplementary Fig. 6. Poorly differentiated HCC samples could also be distinguished, but unsupervised global analysis failed to distinguish moderately differentiated from well-differentiated HCC. This is further confirmed by a receiver operating characteristic (ROC) curve given in Fig. 2B. We applied a *k*-nearest neighbor (*k*NN) algorithm [30] to classify samples into well, moderate, or poorly differentiated HCC. We first used the 64 liver-specific transcripts and then all 3536 genes. Using a series of thresholds of percentage vote for making positive predictions, we observed the specificity and sensitivity of prediction in leave-one-out cross-validation. The ROC curve suggests that without further supervised gene selection the list of 64 liver-specific transcripts outperforms the global expression profiles for the classification of samples by degrees of differentiation. This is because these samples are greatly variable in other clinicopathological parameters, such as sex, age, viral infection, invasiveness, etc. By focusing on liver-specific genes, we are able to filter other factors and gain information about tumor differentiation with a higher signal-to-noise ratio.

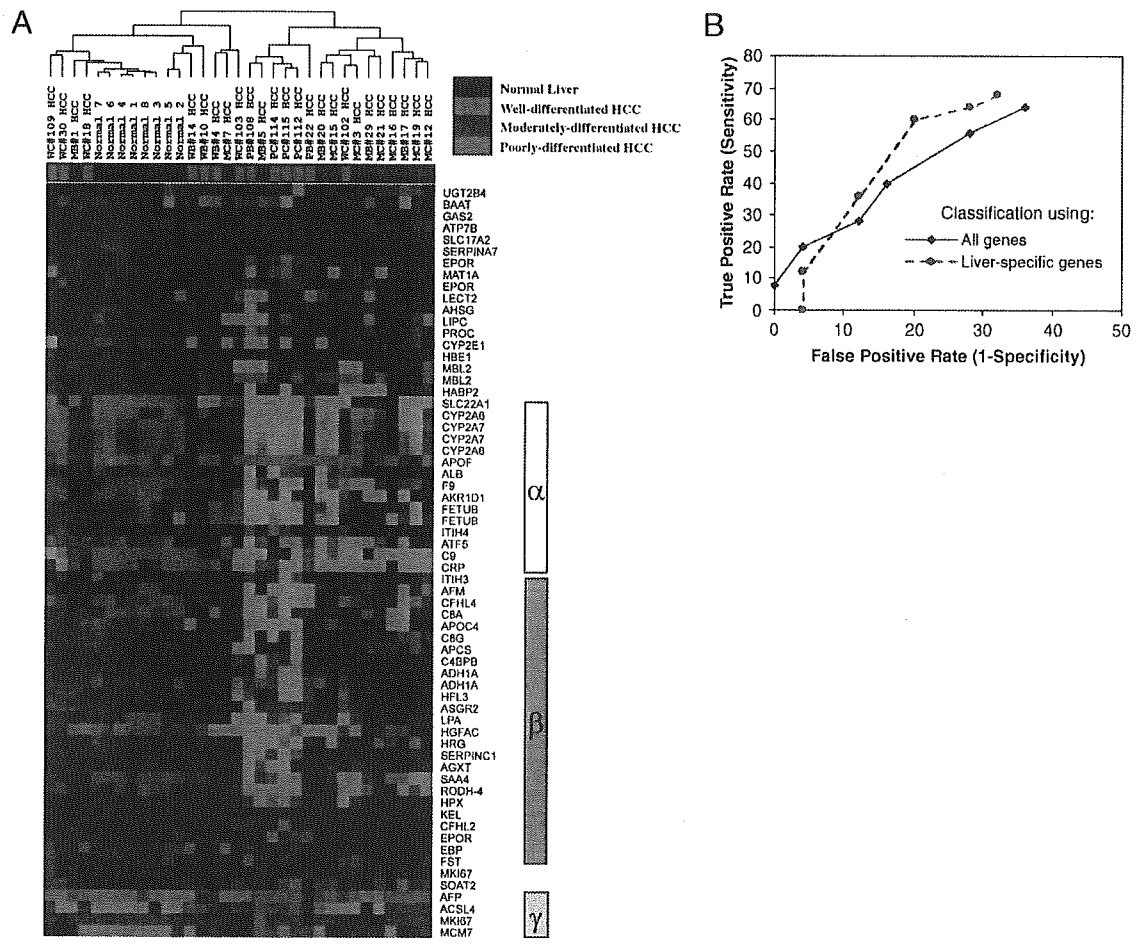


Fig. 2. (A) Unsupervised clustering analysis of a hepatocellular carcinoma (HCC) dataset using liver-specific genes. Well-differentiated HCC (green) and normal liver samples (black) form a subcluster characterized by high expression of liver-specific genes. The remaining samples are further divided into two smaller groups, one dominated by poorly (red) and the other moderately differentiated HCC (blue). While genes in Group α are underexpressed in both moderately and poorly differentiated HCCs, those in Group β show a tendency of increased expression according to degrees of tumor differentiation. Genes in Group γ are overexpressed in poorly differentiated HCC samples. This group includes some fetal-liver specific genes. (B) ROC curve showing that liver-specific transcripts outperform the whole gene set in classifying tumor into well, moderate, or poorly differentiated HCC.

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

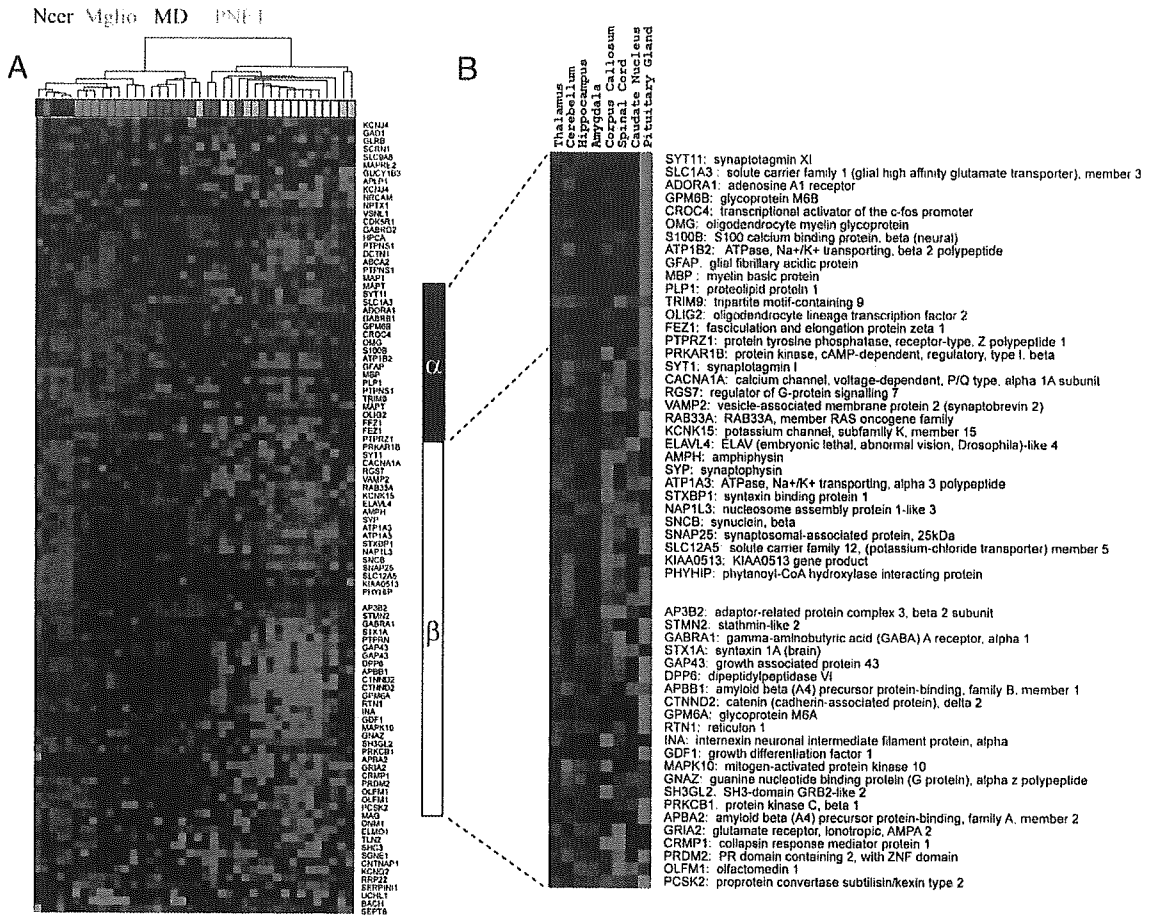


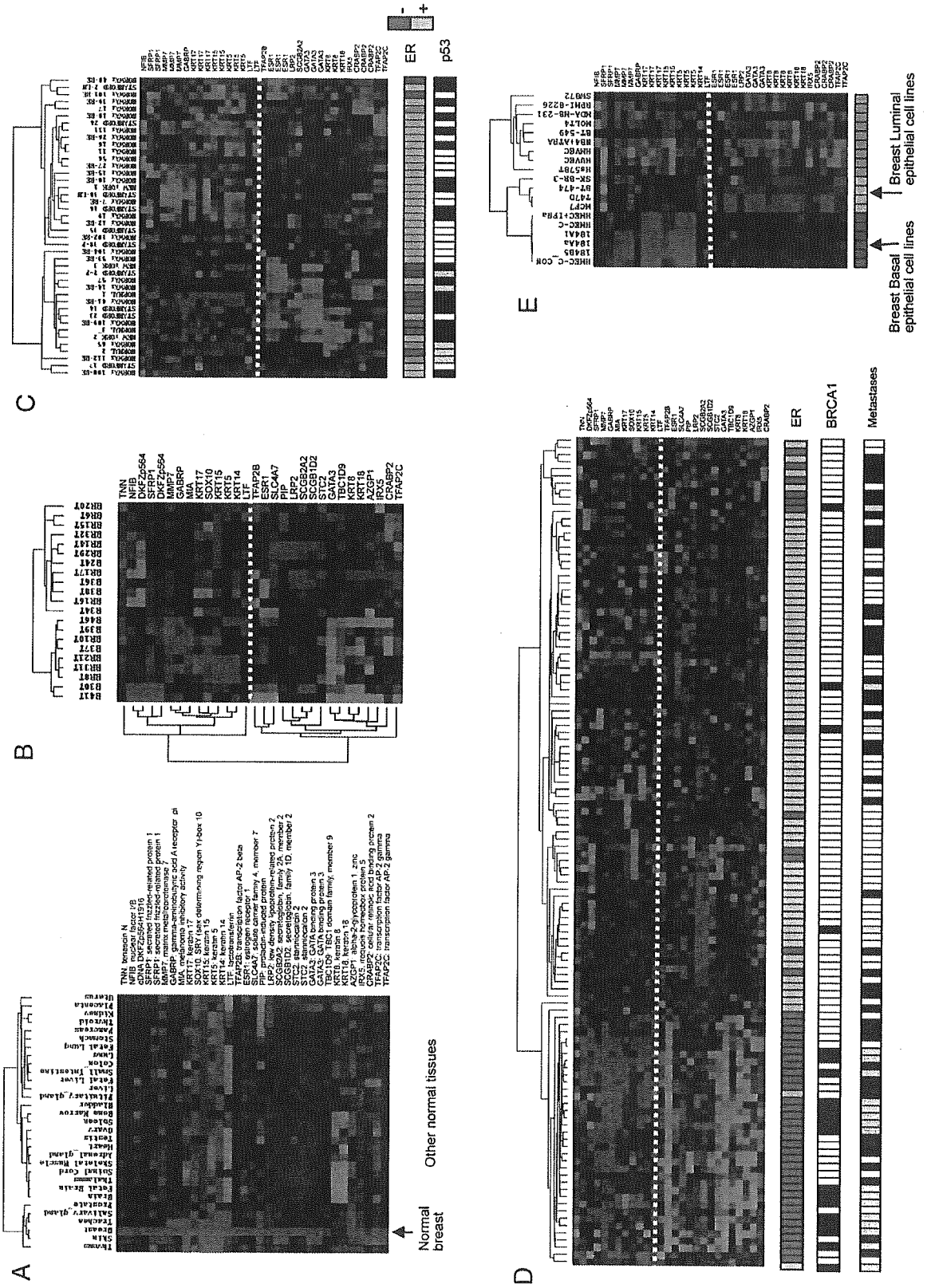
Fig. 3. (A) Expression of brain-specific genes in various types of brain tumors. The samples are divided into two major groups, a glioma and medulloblastoma group, and a PNET and AR/AT group. The first group shows higher expression of many brain-specific genes while the second does not. Within the first 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
B
C
D
E
Normal breast
Other normal tissues
ER
p53
ER
BRCA1
Metastases
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 <

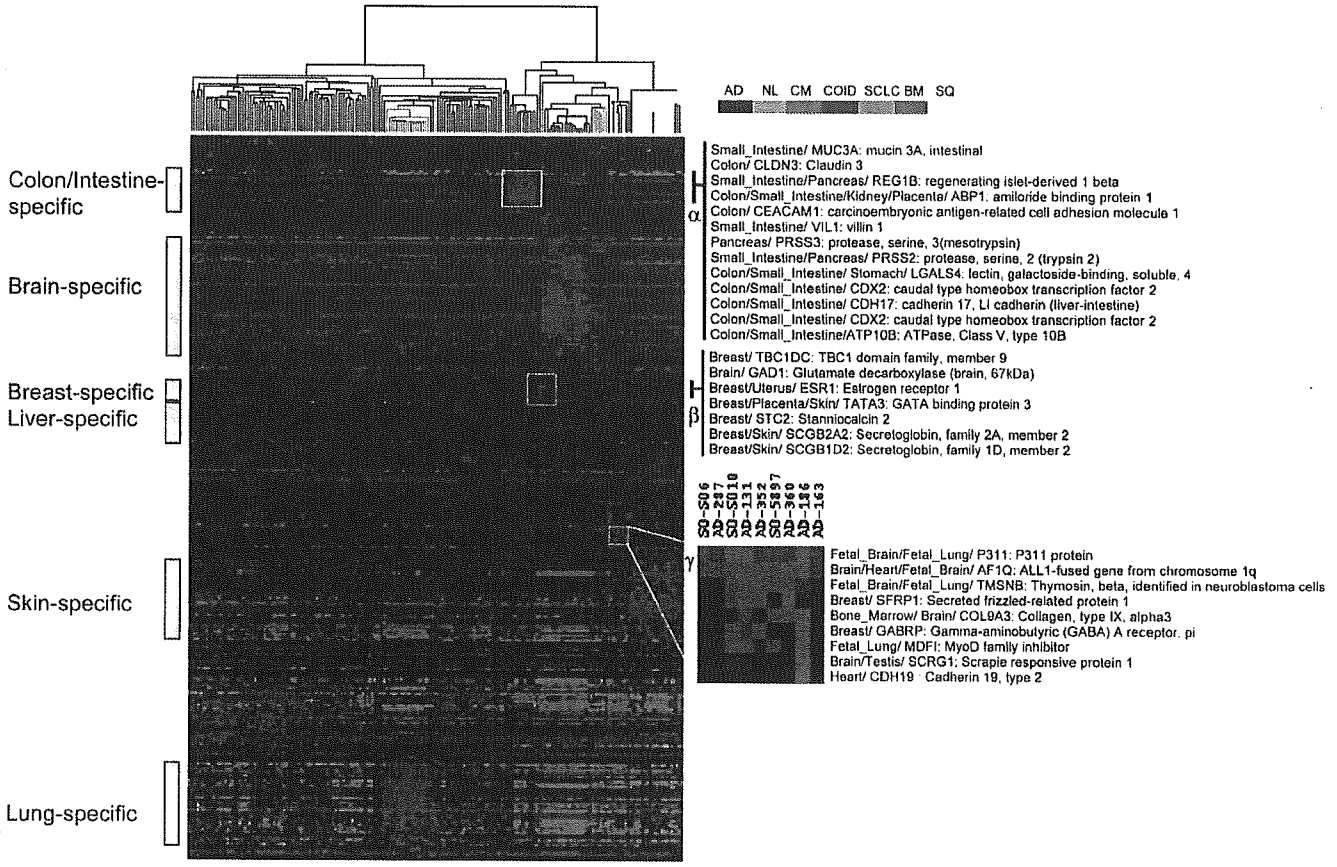


Fig. 5. Clustering analysis of a dataset of lung cancer using all 2503 of the tissue-specific/selective genes. Branches are marked according to clinical diagnosis: normal lung (NL), gray; lung adenocarcinoma (AD), black; squamous cell lung carcinomas (SQ), yellow; pulmonary carcinoids (COID), blue; and 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

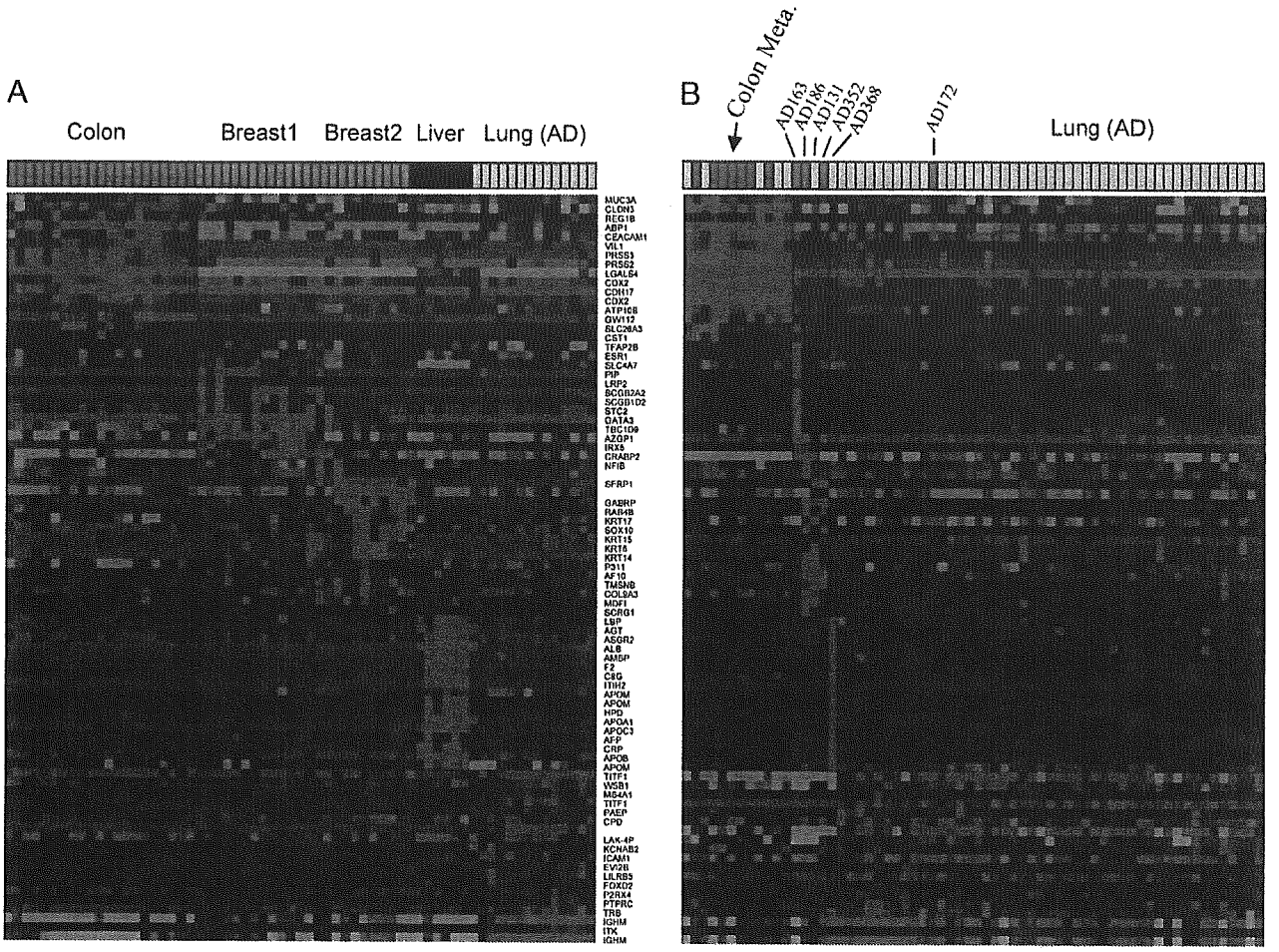


Fig. 6. Prediction of tumor origin with selected markers genes. Predictor genes for colon and liver cancers are selected from the result shown in Fig. 5, while those for two major types of breast cancers are from Fig. 4. Predictors of lung adenocarcinoma are from Ref. [21]. (A) Expression of these genes 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