Physiol Genomics 13: 249–262, 2003.
First published March 18, 2003; 10.1152/physiolgenomics.00186.2002.

Identification of endothelial cell genes by combined database mining and microarray analysis

Michael Ho,[1] Eugene Yang,[1] George Matcuk,[1] David Deng,[2] Nick Sampas,[2] Anya Tsalenko,[2] Raymond Tabibiazar,[1] Ying Zhang,[1] Mary Chen,[1] Said Talbi,[1] Yen Dong Ho,[1] James Wang,[1] Philip S. Tsao,[1] Amir Ben-Dor,[2] Zohar Yakhini,[2] Laurakay Bruhn,[2] and Thomas Quertermous[1]

[1]*Donald W. Reynolds Cardiovascular Clinical Research Center, Division of Cardiovascular Medicine, Stanford University School of Medicine, Stanford, California 94305; and*
[2]*Agilent Technologies, Inc., Palo Alto, California 94304*

Submitted 30 December 2002; accepted in final form 14 March 2003

Ho, Michael, Eugene Yang, George Matcuk, David Deng, Nick Sampas, Anya Tsalenko, Raymond Tabibiazar, Ying Zhang, Mary Chen, Said Talbi, Yen Dong Ho, James Wang, Philip S. Tsao, Amir Ben-Dor, Zohar Yakhini, Laurakay Bruhn, and Thomas Quertermous. Identification of endothelial cell genes by combined database mining and microarray analysis. *Physiol Genomics* 13: 249–262, 2003. First published March 18, 2003; 10.1152/physiolgenomics.00186.2002.—Vascular endothelial cells maintain the interface between the systemic circulation and soft tissues and mediate critical processes such as inflammation in a vascular bed-selective fashion. To expand our understanding of the genetic pathways that underlie these specific functions, we have focused on the identification of novel genes that are differentially expressed in all endothelial cells, as well as restricted groups of this cell type. Virtual subtraction was conducted employing gene expression data deposited in public databases and 384 genes identified.[1] These genes were spotted on custom microarrays, along with 288 genes identified through subtraction cloning from TGF-β-stimulated endothelial cells. Arrays were evaluated with RNA samples representing endothelial cells cultured from four vascular sources and five non-endothelial cell types. These studies identified 64 pan-endothelial markers that were differentially expressed with at least a threefold difference (range 3- to 55-fold). In addition, differences in gene expression profiles among endothelial cells from different vascular beds were identified. Validation of these findings was performed by RNA blot expression studies, and a number of the novel genes were shown to be expressed under angiogenic conditions in the developing mouse embryo. The combined tools of database mining and transcriptional profiling thus provide expanded knowledge of endothelial cell gene expression and endothelial cell biology.

transcriptional profiling; vascular; expression database; cell specificity; angiogenesis

IN ADDITION TO SERVING A BROAD range of physiological functions, the endothelial cell is widely appreciated to have a central role in the development of vascular disease. There has been a longstanding interest in endothelial cell-specific gene expression, to identify specific pathways for therapeutic targeting in this cell type. Primarily, genes specifically expressed in endothelial cells have been identified through molecular cloning of molecules that mediate vascular wall-specific processes. Endothelial genes identified in this fashion include platelet endothelial cell adhesion molecule, vascular endothelial cell adhesion molecule, and endothelin-1, among others. In addition, a small number of directed efforts have sought to expand the repertoire of endothelial cell genes through directed cloning efforts, employing differential display and other molecular methods (18, 37). At least one effort has employed the publicly available gene expression databases to identify endothelial genes (19). Unfortunately, such databases are biased, depending on the cell types chosen for study and the depth to which the libraries were sequenced.

The development of high-throughput transcriptional profiling employing microarray technology has dramatically expanded the ability to characterize patterns of gene expression in isolated cells and normal and diseased tissues. Originally applied to basic questions regarding metabolic pathways in model organisms such as yeast, there is now widespread use of the technology to answer a broad range of biological questions. Arrays have been employed to develop molecular classifications of lymphoid tumors and melanoma, to link metastasis of breast cancer to gene expression, and to provide new clues to the molecular basis of glioblastoma and melanoma (6, 24, 30, 32, 34, 39). Additional imaginative uses of microarrays include the study of downstream targets of developmental transcription factors and the genetic basis of circadian rhythms (25, 31, 36). Transcriptional profiling with microarrays provides a novel approach to characteriz-

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

Address for reprint requests and other correspondence: T. Quertermous, Division of Cardiovascular Medicine, Stanford Univ. School of Medicine, 300 Pasteur Drive, Falk CVRC, Stanford, CA 94305 (E-mail: tomq1@stanford.edu).

[1]The microarray data derived through these experiments have been deposited in the GEO expression database at the NCBI and has been given the accession number GPL217, with others pending. Primary data and supplementary material associated with this manuscript are being deposited at the following website: **http://quertermous.stanford.edu**.

ing cell-specific gene expression and could be employed to identify endothelial cell-specific genes. Indeed, microarray methodology has already been employed to identify genes in endothelial cells that are responsive to cytokines and physical forces or activated in the context of in vitro morphogenesis models (1, 3, 8, 9, 12, 13, 21, 26–28, 40–43).

As a method to identify minimally characterized genes that are selectively expressed by endothelial cells, we have coupled cloning-based gene expression databases that are available to the public with the high-throughput capability of microarrays. Using bioinformatics tools and queries designed to screen for those genes with the highest level of differential expression, we have identified 384 genes. These genes, along with 288 TGF-β-responsive genes, were placed on a custom-spotted cDNA microarray and probed with RNA samples derived from endothelial and non-endothelial cells. It was possible to identify over 64 clones that were differentially expressed in endothelial cells, with a number of these being novel expressed sequence tags (ESTs) or putative expressed sequences deduced from the human genome. Interestingly, the cultured cells from different vascular beds were found to exhibit specific patterns of gene expression, likely reflecting differences in their gene expression in vivo. Thus these studies define a panel of unique endothelial genes and identify putative new vascular bed restricted markers.

MATERIALS AND METHODS

Cell culture. Human umbilical vein endothelial cells (HUVEC), human aortic endothelial cells (HAEC), human coronary artery endothelial cells (HCAEC), human lung microvascular endothelial cells (HMVEC), human aortic smooth muscle cells (HASMC), human mammary epithelial cells (HMEC), normal human astrocytes (NHA), and normal human epidermal keratinocytes (NHK) were primary cultured cells obtained from Clonetics (San Diego, CA). All primary human cells were from single donors ages 33 to 54 yr, except NHA (donor 18 wk), HASMC and pulmonary HMVEC (donor 3 yr), and HUVEC (donated at birth). HCAEC and HAEC were from male donors, other cells were from females, or the gender of the source was not known. Cryopreserved primary cells were received at the following passages: HCAEC passage 3, HAEC passage 3, HUVEC passage 1, HMVEC passage 4, HASMC passage 3, HMEC passage 7, NHK passage 1, and NHA passage 2. For consistency, each cell type was cultured for two more passages after they were received prior to RNA isolation. Therefore, RNA was obtained from endothelial cells that were passaged no more than six times from the original culture. HepG2 (human hepatocellular carcinoma cell line) cells were obtained from the American Type Culture Collection (Manassas, VA). All four endothelial cell types were plated on 100-mm culture dishes precoated with 2% gelatin (Sigma, St. Louis, MO) and cultured in M199 containing 15% FBS (HyClone, Logan, UT), endothelial cell growth supplement (20 μg/ml, Sigma), heparin (20 U/ml, Sigma), glutamine, and penicillin. HASMC were cultured in smooth muscle cell basal medium (modified MCDB 131, Clonetics) and the following growth supplements: 0.5 ng/ml hEGF, 5 μg/ml insulin, 2 ng/ml hFGF-B, 5% FBS, and Clonetics GA-1000 (gentamycin, amphotericin B) at 1:100. NHA were cultured in Astrocyte basal medium (CCMD 190, Clonetics) and the following growth supplements: 20 ng/ml hEGF, 25 μg/ml insulin, 25 ng/ml progesterone, 50 ng/ml transferrin, Clonetics GA-1000 at 1:100, and 5% FBS. NHK were cultured in keratinocyte basal medium-2 (CCMD 151, Clonetics) and the following growth supplements: 0.03 mg/ml bovine pituitary extract, 0.1 ng/ml hEGF, 5 μg/ml insulin, 0.5 μg/ml hydrocortisone, 0.05 mg/ml transferrin, Clonetics GA-1000 at 1:100, and 5% FBS. HMEC were cultured in mammary epithelial cell basal medium (modified MCDB 170, Clonetics) and the following growth supplements: 52 ng/ml bovine pituitary extract (BPE), 10 ng/ml hEGF, 0.5 ng/ml hydrocortisone, 5 ng/ml insulin, GA-1000 at 1:100, and 5% FBS. HepG2 were cultured in minimum essential medium (Eagle's) with 2 mM L-glutamine and Earle's BSS adjusted to contain 1.5 g/l sodium bicarbonate, 0.1 mM non-essential amino acids, 1.0 mM sodium pyruvate, and 10% FBS.

RNA isolation. RNA was isolated from cells employing a combination of Trizol (Life Technologies, Rockville, MD) and RNeasy columns (Qiagen, Valencia, CA) techniques. Briefly, media was removed, and 2 ml Trizol was used per 3×10^6 cells. Cells were sheared through a 21-gauge needle, this solution was extracted with chloroform, and the supernatant was mixed with 500 μl of 70% ethanol for every milliliter Trizol used. This mixture was then loaded and eluted from an RNeasy column for further purification. RNA quality and concentration were evaluated by gel visualization and spectrophotometric analysis.

Database screening methodology. Top "endothelial-enriched" candidate genes were compiled from the UniGene, Serial Analysis of Gene Expression (SAGE), and BodyMap libraries as follows.

Using the "library differential display" feature of UniGene (**http://www.ncbi.nlm.nih.gov/UniGene/ddd.cgi?ORG=Hs**), we generated a list of the top 100 genes representing those differentially expressed between endothelial cell lines (*pool A*) and all nonvascular cell lines (*pool B*). A metric was developed termed the score; score = *pool A*/(*pool B* + 0.00001), with the "0.00001" factor added to prevent division by zero and to compensate for undue inflation of the score that division by a small fraction would create. The genes were then sorted by the highest score, with all genes already identified with a UniGene ID. Only those genes with a score greater than 10.4 were used in the final compilation of genes (a total of 29 genes).

The SAGE database at the National Center for Biotechnology Information (NCBI) (**http://www.ncbi.nlm.nih.gov/SAGE/sagexpsetup.cgi**) was queried as follows. The xProfiler tool of SAGE was used to generate two pools of cell lines to perform a virtual subtraction between endothelial and all non-endothelial cell types. In *pool A*, only HMVEC and HMVEC+VEGF cell lines were selected. *Pool B* was derived from all libraries that represented non-endothelial cell lines. No tissue libraries were included in the analysis, as they were likely to contain significant numbers of microvascular endothelial cells. The subtraction was performed with a factor difference of 2.0 and coefficients of variance cutoffs of 0%. From this, a list of 65,505 SAGE tags was generated and downloaded, along with the number of "hits" in *pool A* and *pool B*. This list was then resorted based upon the highest ratio of *pool A* to *pool B* (with zero hits in *pool B* changed to 0.9 to prevent division by zero). A score metric was devised whereby the score = *group A*/*group B*, and all genes were then sorted by this value. In some instances, a gene with a high score was only identified by the SAGE tag, without information on the UniGene ID or gene name. In such cases, a search using the SAGE tag to gene mapper (**http://www.ncbi.nlm.nih.gov/SAGE/index.cgi?cmd=tagsearch**) was

performed to identify the likely gene or EST represented by the tag. Genes with a score of 2.2222 or greater constituted an initial compilation of genes (a total of 368 genes). However, of these 368, no reliable matches for a UniGene cluster were found based on the SAGE tag information for 112, leaving 256 to be included in the final compilation.

Putative endothelial genes were also identified from the BodyMap gene expression database (**http://bodymap.ims.u-tokyo.ac.jp/**) (16, 22). The Gene Ranking System was utilized to select for genes preferentially expressed in the intima layer of the human aorta. This generated a list of 200 genes along with the number of times each gene was found to be present in aortic intima vs. other tissues [score = no. intima aorta/(total − intima aorta + 0.0001)], and the genes were ranked based upon the highest ratio. After the genes were sorted by this score, a UniGene number was found for the top 188 genes, when possible, based on either the gene name, gene sequence, or some combination thereof. Of these 188 genes, no significant matches were found for only 10 genes, leaving 178 BodyMap genes used in the final compilation.

The top genes from each virtual subtraction were compiled (178 from BodyMap, 29 from UniGene, and 256 from SAGE for a total of 463 genes). Two of these 463 genes were later found to best match genes from the mouse and rat and were therefore excluded. Of these 461 genes, 384 genes (UniGene numbers) were ordered from Research Genetics.

Cloning by suppression subtractive hybridization. HAEC (passage 5) were serum starved for 24 h without ECGS or heparin, then treated with 3 ng/ml human recombinant TGF-β1 (R&D Systems, Minneapolis, MN), and harvested at 30-min, 5-h, and 24-h intervals. RNAs isolated from these harvested cells were pooled. A PCR-based cDNA subtraction and normalization methodology was employed using reagents supplied in the PCR Select cDNA Subtraction kit (Clontech, Palo Alto, CA) to identify genes preferentially expressed in TGF-β-stimulated endothelial cells (10). Tester DNA was derived from 2 μg of TGF-β-stimulated HUVEC poly(A)$^+$ RNA, and driver DNA was derived from 2 μg of poly(A)$^+$ RNA from growth-arrested HUVEC. The reverse subtraction was performed by reversing the tester and driver. Subtraction hybridization was performed according to the manufacturer's instructions, and the 288 products of secondary PCR were cloned into plasmid vectors and evaluated by nucleotide sequence analysis. Clones were identified by employing the DNA sequence in automated BLAST searches of the NCBI nucleotide databases.

Microarray construction and hybridization. For cDNA probe preparation, the 384 virtual subtraction and 288 subtraction hybridization clones were amplified by PCR employing flanking sequences of cloning vectors, according to standard methodology. Five microliters of PCR reaction were visualized on 1% agarose gels for quality determination, and PCR was repeated and optimized until all clones gave a single band. PCR reactions were purified on a Qiagen BioRobot 3000. In addition, the 288 subtraction hybridization clones were PCR amplified and purified by both Microcon-96 filtrate (Millipore, Bedford, MA) and ArrayIt kits (TeleChem International, Sunnyvale, CA). There were not appreciable differences in microarray results from PCR products purified by the different methods; however, results are reported separately for the three different preps such that there are 1,248 probes representing 672 clones. DNA microarrays were printed on glass slides employing Agilent's SurePrint inkjet technology (Agilent Technologies, Palo Alto, CA). The microarrays contain 6 repeat features for each of the 1,248 probes. For a description of the performance features of Agilent's deposition cDNA microarrays with respect to uniformity, sensitivity, precision, and accuracy in gene expression profiling assays, see (**http://www.chem.agilent.com/scripts/LiteraturePDF.asp?iWHID=27667**).

Sample labeling and hybridization to the arrays was performed as follows. Ten micrograms of total RNA from cultured cells was reverse-transcribed in the presence of 400 U Superscript II RNase H$^-$ reverse transcriptase (Invitrogen, Carlsbad, CA), 25 μM dCTP and 100 μM each dATP, dTTP, and dGTP, 25 μM Cy3- or Cy5-dCTP (NEN Life Science, Boston, MA), 4 μM 5′-T16N-2′ DNA primer, and 27 U of RNase inhibitor (Amersham, Piscataway, NJ). The labeling was carried out at 42°C for 1 h. After degradation of unlabeled RNA by RNase I, labeled cDNAs were purified with a Qiagen PCR cleanup kit. Microarray hybridization was performed at 65°C overnight in 25 μl of hybridization solution containing Agilent's deposition hybridization buffer, 5 U of PolydA$_{40–60}$ (Amersham), 5 μg of yeast tRNA (Sigma), 10 μg of human *Cot*I DNA (Invitrogen), and Cy3- and Cy5-prelabeled HCV deposition control targets (Qiagen Operon, Valencia, CA). At the end of hybridization, microarrays were first washed in 0.5× SSC/0.01% SDS for 5 min at room temperature, then washed in 0.06× SSC wash buffer for 10 min. Finally, microarrays were dried by centrifugation. At least two separate cultures of each cell type were employed for RNA preparation and hybridization. HUVEC RNA served as a common reference for all of these experiments. At least four hybridizations were performed for each cell type, and dye reversals were conducted for all cell type measurements.

Scanning, background subtraction, and normalization of array data. The microarrays were scanned on an Agilent model G2565AA microarray scanner system, and the images were quantified using Agilent G2567AA Feature Extraction software version A.5.0.92. Background subtraction and normalization methods not included in the commercial software were applied for this data set. For background subtraction, the array was divided into 30 localized regions, and for each region the average of the weakest 5% of features was subtracted from the raw signal for each feature. Given that the microarrays used in this study were designed to be highly enriched for genes known to be expressed in endothelial cells, standard two-color DNA microarray normalization methods that rely on the assumption that the distributions of signals for both the samples in the red and green channels were not considered appropriate. As the endothelial cell types tended to have higher expression for most of the probes compared with the non-endothelial cell types, we expected the samples to have overall biases of up- and downregulation of genes. For this reason, a set of normalization probes was selected that appeared least regulated across the set of all arrays in the study. We employed an approach similar to one used by Schadt et al. (35) for normalization between pairs of single-color microarrays. Those investigators used features that were members of two-dimensional longest order-preserving sequences (LOPS) to normalize all features on an array. In our approach, each probe was scored by the length of the 2N-dimensional LOPS of which it was a member. Probes within these longest, or almost longest, order-preserving sequences tend to be the most "housekeeping-like," since their rankings within each sample are highly conserved across all of the samples. Once the probes were scored, the 174 least regulated probes, corresponding to 1,044 features (6 replicate features per probe) on each array, were used for normalization. The normalization probe signals were fitted to a straight line in log(red signal) vs. log(green signal) space, and these fit parameters were applied to calculate normalized signals for all features. Saturated, non-uniform, and

population-outlier features were omitted from the normalization sets on an array-by-array basis. Signals from the six replicate features for each probe were combined by averaging normalized feature signals, while eliminating bad or outlier features.

Data analysis. Several statistical methods were employed to identify those genes that were likely to be differentially expressed in the cell types under consideration. In particular, we used significance analysis of microarrays (SAM; X-Mine, Brisbane, CA) as well as parametric and nonparametric methods developed at Agilent Laboratories.

SAM is a widely used variant of permutation analysis developed specifically for microarrays. The algorithm employs replicate experiments to develop a measure of variance that is used to test whether observed differences in gene expression, in two cell-type partitions, are likely to be real (**http://www-stat.stanford.edu/~tibs/SAM/**) (38). The SAM algorithm employed was executed by the SAM Isolator program (X-Mine). SAM also computes the false detection rate (FDR) for any set of genes defined by how sharply they individually separate two cell type classes. The FDR calculation uses a pseudo-random permutation process, and the actual numbers assume sample independence. While this assumption is valid when two homogenous classes are considered, it is not easily justified when different numbers of replicates represent different subclasses. The ranking of genes using SAM is, however, valid and this analysis yields strong results in two-way classifications such as in Fig. 1.

Parametric and nonparametric (distribution free) scoring methods developed at Agilent Laboratories (**http://www.labs.agilent.com/resources/techreports.html**) were also used to identify differentially expressed genes and to assess the significance of the observed differences. Some of these methods are applicable for data with more than two classes, as described below. Parametric methods assume a certain distribution for expression values of every gene within each given class (e.g., cell type) and then score genes according to how separate the class-specific distributions are. The parametric method employed here was the Gaussian error score (discussed below). Distribution free scores, in contrast, are not based on parametric assumptions. The Kolmogorov-Smirnov score and the Wilcoxon rank-sum test are classic examples (7, 17). The use of nonparametric scores for microarray data analysis has been described (4, 5). The threshold-number-of-misclassifications (TNoM) score was applied to the current data. The exact *P* values for the TNoM score can be computed using combinatorial approaches with the underlying assumption of independence of samples within one class (5). This assumption is not valid for the set of experiments we consider here since we include replicate experiments for the same cell type.

The Gaussian error score (GER) is applicable in the general multi-class case (such as the comparison of HAEC vs. HCAEC vs. HMVEC), where we seek genes that separate all classes from each other. In Fig. 2*B* the expression patterns of single genes are shown for the replicate experiments with HAEC (red), HCAEC (green), and HMVEC (blue). For each cell type a Gaussian curve (shown in Figs. 1–5 with the respective class color) is fit to the data. Intuitively, the more separate the three distributions are, the more relevant the gene to the distinction between the different cell types. To formalize this intuition we compute, for each sample, t, of class (cell type) $\Gamma(t)$ its error score,

$$\mathrm{Err}_g(t) = 1 - \frac{p(e_g(t)|\mu_{\Gamma(t)},\sigma_{\Gamma(t)})p(\Gamma(t))}{\sum_{\Gamma} p(e_g(t)|\mu_{\Gamma},\sigma_{\Gamma})p(\Gamma)}$$

where $e_g(t)$ is the expression level of the gene g in sample t, $p(\cdot|\mu,\sigma)$ is the Gaussian density with parameters μ and σ, and $p(\Gamma)$ is the prior probability of the class Γ . The Gaussian error score (GER) of g is then defined by

$$\mathrm{GER}(g) = \sum_{j=1}^{m} \mathrm{Err}_g(t_j)$$

One disadvantage of parametric scores is their sensitivity to the effects of outliers. Very large or small ratio values can push the distributions apart and yield a score much better than reality supports. Another disadvantage is that parametric scores are usually based on homogeneous distributions within classes. This is not the case in some of these analyses, e.g., comparison of endothelial to non-endothelial cells as in Fig. 1.

Northern blot and in situ hybridization. For each of the nine different cell types, 10 μg of total RNA was electrophoresed on a 1% formaldehyde agarose gels and transferred onto nylon membranes. The RNA was immobilized on the membranes by baking at 80°C for 1 h. Full-length inserts were cut from EST plasmids and radiolabeled with [^{32}P]dCTP by random priming and used as probes for Northern blots. Insert sizes were as follows: AW770514, 1.1 kb; AA256482, 0.69 kb; CD31, 1.5 kb; AI261621, 1.0 kb; AI422298, 0.92 kb; AA428201, 1.58 kb; and multimerin, 0.79 kb. Blots were hybridized at 42°C for 16 h in the presence of 48% formamide and 10% dextran sulfate. After hybridization, the membranes were washed at high-stringency conditions, 65°C with 0.2× SSC buffer and 0.5% SDS. Visualization was achieved by exposure to Kodak Biomax MS film (Eastman Kodak, Rochester, NY).

For mouse embryo in situ hybridization, mouse orthologs of the putative endothelial restricted ESTs were identified through HomoloGene, or through identification of mouse genes by BLAST searches (NCBI) that gave reciprocal best hits, and corresponding IMAGE clones were obtained from Research Genetics. Orthologs were identified as follows: AA4 (mouse AA145088), EST AW772163 (mouse BG172926), and EST W81545 (mouse AA797701). The whole mount in situ preparations of mouse embryo at embryonic *day 8.5* (*E8.5*) and *E9.5* were done according to Henrique et al. (14). Briefly, digoxigenin-labeled probes were transcribed from linearized cDNA template and applied to embryos digested with proteinase K. Then embryos were incubated with BM purple substrate and images were taken with a CCD digital camera.

RESULTS

Database mining, array construction, and hybridization. Three publicly available gene expression databases were identified and used for this analysis: SAGE (**http://www.ncbi.nlm.nih.gov/SAGE/index.cgi?cmd=expsetup**), UniGene (**http://www.ncbi.nlm.nih.gov/UniGene/ddd.cgi?ORG=Hs**), and BodyMap (**http://bodymap.ims.u-tokyo.ac.jp/human/gene_ranking.php**). A metric was employed with each database to identify those SAGE tags that were sequenced in endothelial cell libraries and not in other cell type libraries. UniGene IDs were recorded for each gene identified, or obtained by searching the appropriate database for assignment of a UniGene number. At UniGene, it was possible to employ the "library differential display" feature, and to identify 29 genes that were differentially expressed in cultured endothelial cells. The SAGE site had information from HMVEC

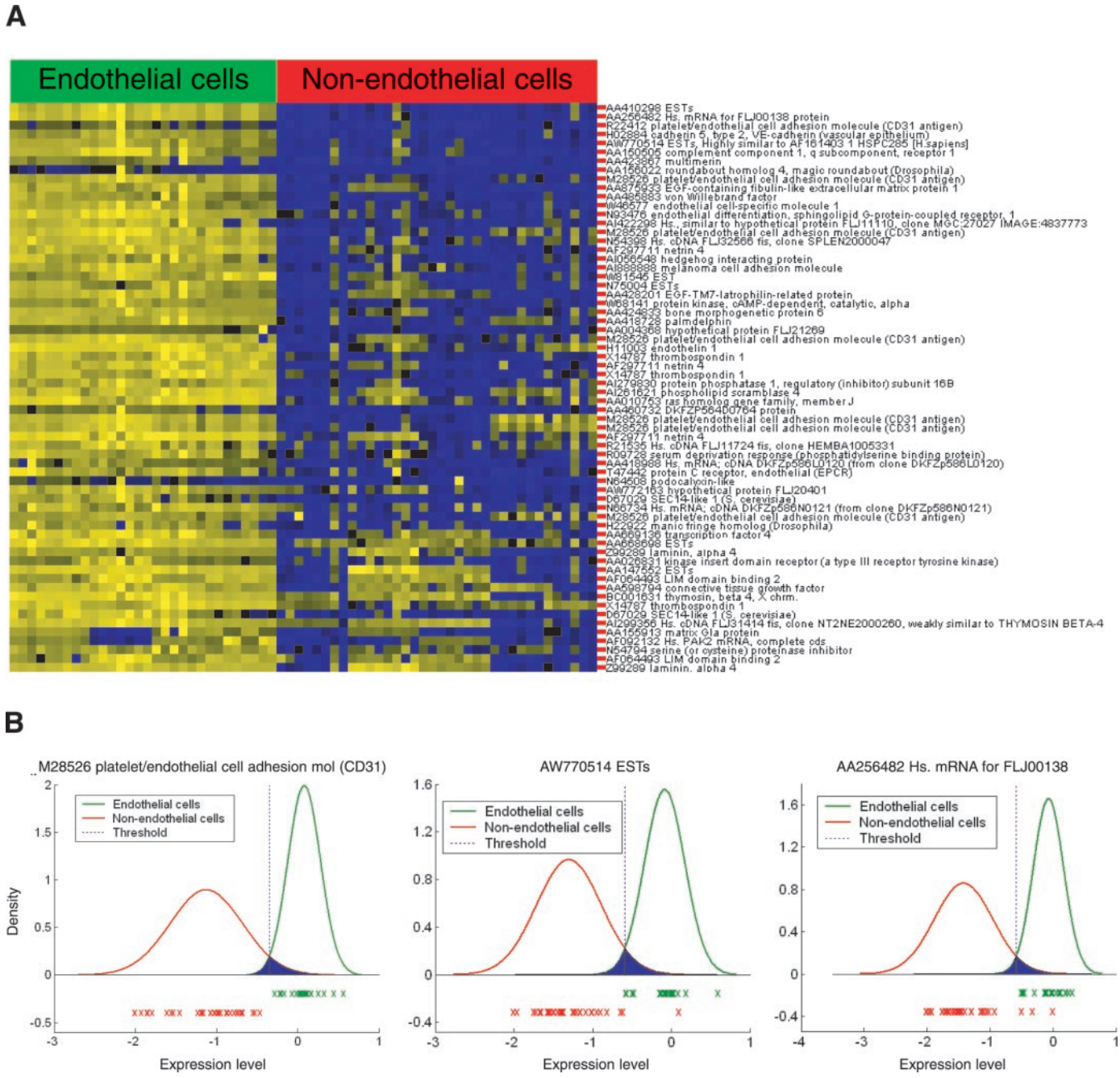

Fig. 1. Identification of genes preferentially expressed in various cultured human endothelial cell types. Primary endothelial cells, including those from umbilical vein (HUVEC), coronary artery (HCAEC), aorta (HAEC), and the pulmonary microcirculation (HMVEC), were compared with non-endothelial cell types including primary cultured human epithelial cells (HMEC), astrocytes (NHA), aortic smooth muscle cells (HASMC), keratinocytes (NHK), and the HepG2 liver tumor cell line. Thirty hybridizations were performed with endothelial cell RNAs, and 36 hybridizations with non-endothelial cell RNAs, to a microarray was constructed with endothelial cell genes identified in expression databases and cloned from TGF-β-stimulated endothelial cells. Significance analysis of microarrays (SAM) was employed to identify those genes expressed at a significantly higher level across the different endothelial cell types, compared with the non-endothelial cells (38). The analysis shown was associated with a false detection rate (FDR) of 0.0, making it unlikely that any of the 64 genes were falsely identified. *A*: heat map is organized with individual hybridizations arranged along the *x*-axis, with relative ratios of expression (cellular RNA compared with HUVEC reference) indicated by color. The order of the genes reflects decreasing SAM score, or *d* statistic value. Color intensity is scaled within each row, so that the highest expression corresponds to bright yellow, and the lowest expression corresponds to bright blue. *B*: a Gaussian error model was also used for classification, and the separation of signals between different cell types is shown for three representative genes. In each graph the *x*-axis represents log (red/green ratios). The colored crosses indicate actual data of the respective classes, and the curves represent best Gaussian fits to the actual data. The threshold indicated represents the intersection of the two curves, and the highlighted region is the Bayesean probability of error when using the curves for classification into the two depicted classes.

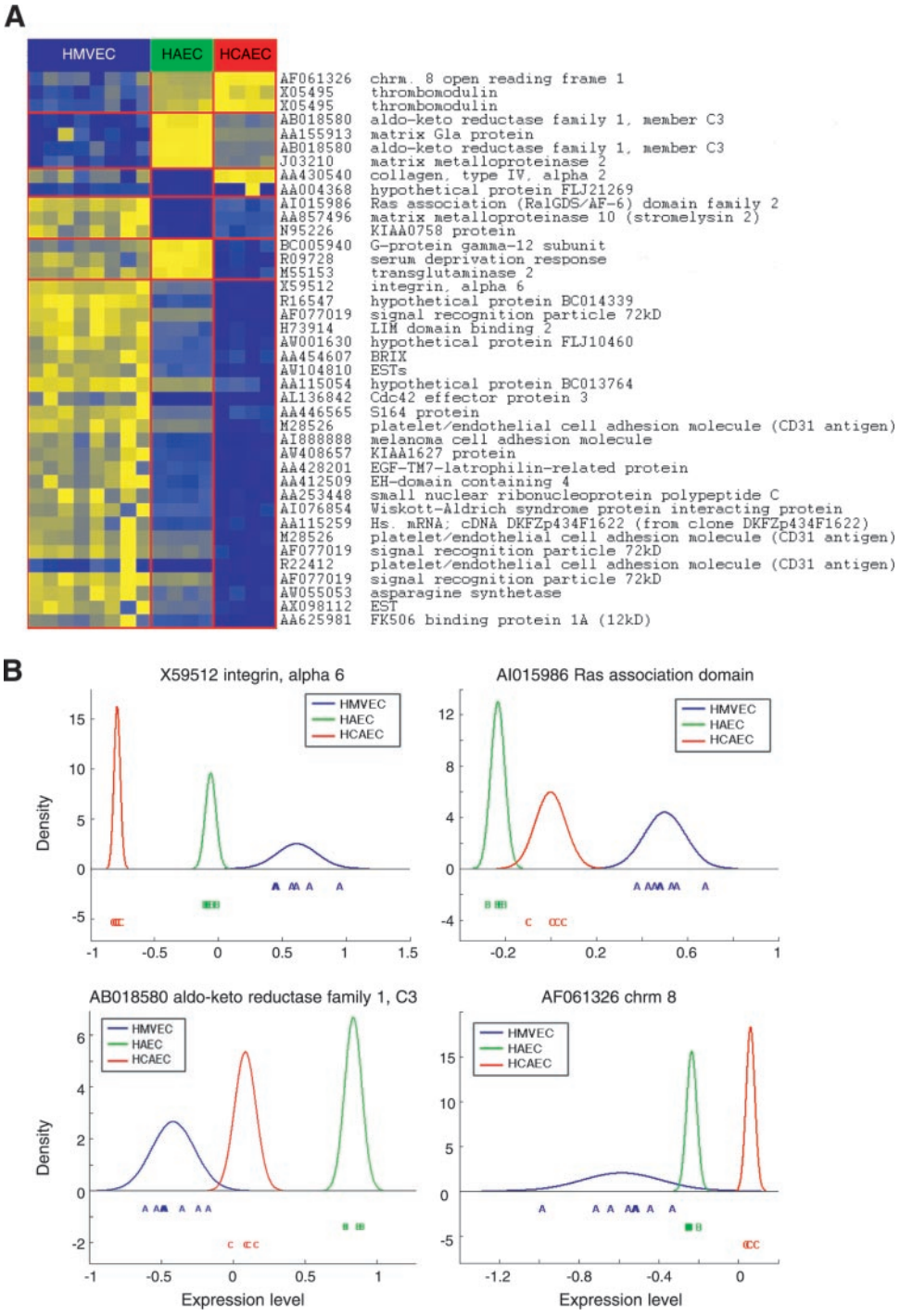

Fig. 2. Identification of genes that are differentially expressed in HAEC, HCAEC, and HMVEC. *A*: for each gene, we computed a Gaussian error score applied to three classes corresponding to the three different endothelial cell types (5). Shown are the 40 top-scoring genes grouped by pattern of expression observed in these three classes. Genes in the *top* group have the highest expression in HCAEC and the lowest expression in HMVEC. In the next group, genes have the highest expression in HAEC and the lowest in HMVEC. The last group shows genes with the highest expression in HMVEC and the lowest expression in HCAEC. *B*: the separation of signals between the different cell types is shown with Gaussian distribution plots for six representative genes. The letter designations in the Gaussian curves represent the hybridization signals for HMVEC (blue "A"), HAEC (green "B"), and HCAEC (red "C").

cells in the presence or absence of vascular endothelial cell growth factor, and 256 genes were identified as enriched in these cells. Expression data at BodyMap at the University of Tokyo allowed comparison of human aortic intimal tissue to other nonvascular tissue types (16, 22). The intima in normal blood vessel wall is primarily endothelial cells and extracellular matrix, so this comparison was appropriate. Using BodyMap data, we initially identified 200 differentially expressed genes and obtained UniGene numbers. IMAGE clones representing the identified UniGene clusters were ordered from Research Genetics.

We complemented our in silico approach to identify potential endothelial enriched clones by performing subtraction suppression hybridization on endothelial cells stimulated with TGF-β, a known modulator of angiogenic vascular formation (29). From this library, 288 TGF-β-regulated endothelial cell genes were isolated and sequenced, identified through sequence database searches, and included in the construction of the

custom spotted cDNA array. Accession numbers for these clones, as well as the virtual subtraction clones are available at **http://quertermous.stanford.edu**.

Genes identified through database mining and subtraction cloning were spotted on treated glass slides with inkjet printer methodology. Experiments described here involved hybridizations to microarrays derived from two separate printings. Hybridization to these arrays with 10 μg total RNA gave consistent and reproducible results between arrays. At least four hybridizations were performed with at least two RNA samples, and dye reversal experiments were performed for every RNA, providing reproducibility data. Primary cultured human endothelial cells derived from the umbilical vein (HUVEC), coronary artery (HCAEC), aorta (HAEC), and pulmonary microcirculation (HMVEC) were employed as models of endothelial cell lineage. Primary cultured human astrocytes (NHA), keratinocytes (NHK), mammary epithelial cells (HMEC), and aortic smooth muscle cells (HASMC), as well as the HepG2 hepatic cell tumor line, were employed as models of non-endothelial cell lineage.

Identification of endothelial cell enriched genes. To identify genes that might serve as markers consistently expressed by endothelial cells in different size vessels in various tissues, RNA samples isolated from all of the cultured cell types were hybridized to the arrays, and the resulting data were evaluated with a number of different analytical tools. One method employed for data analysis was the statistical analysis of microarrays (SAM), which rigorously scores differences in gene expression through consideration of the variance exhibited by each gene on the array (38). To provide rigid criteria for endothelial cell enriched expression, the FDR was set to 0.0, minimizing the detection of insignificant differences in gene expression (38). The results of this analysis are presented in graphical format in the form of a heat map (Fig. 1) as well as numerical format (Table 1). A heat map is a graphical representation of relative gene expression using two colors to depict up- or downregulation of individual genes. In the heat maps presented here, yellow represents high log ratios of relative expression, and blue represents low relative expression levels. Color intensity is scaled within each row, so that the highest expression corresponds to bright yellow, and the lowest expression corresponds to bright blue.

The 64 genes that were thus identified included well-characterized endothelial markers such as CD31, VE cadherin, multimerin, the KDR vascular endothelial cell receptor, and von Willebrand factor (Fig. 1, Table 1). These genes showed a relative expression between endothelial cells and non-endothelial cells that varied between 55-fold and 5-fold. Additional genes known to be preferentially expressed in endothelial cells were also identified, including endothelial differentiation sphingolipid G-protein coupled receptor (EDG-1), melanoma adhesion receptor (MCAM), endothelin-1, ICAM2, protein C receptor, thymosin-β4, and plasminogen activator inhibitor type I. The detection of these transcripts validated the methodology employed. Interestingly, a number of genes that have been characterized in the context of other tissues and cell types were identified, including netrin-4 precursor, cAMP-dependent protein kinase, bone morphogenetic protein 6, manic fringe, matrix gla protein, and hedgehog interacting protein. While clearly not specific for endothelial cells in vivo, these genes did show higher levels of expression in the endothelial lineage (10- to 5-fold), suggesting some requirement for their biological function in this cell type.

Previously uncharacterized genes that shared some sequence similarity with known genes were identified, including ESTs similar to fibulin and Ras. Among the most differentially expressed genes were a large group that had been characterized only as ESTs with no specific annotation, and included AA410298, AA256482, AI261621, and AI422298. During the course of this work, AA410298 has been noted to contain a hypothetical zinc finger domain, AI261621 was assigned to the phospholipid scramblase 4 UniGene cluster, and AA256482 was noted to have significant homology to SHB (Src homology 2 domain containing adaptor protein B). Interestingly, a greater percentage of clones identified through SAM analysis as showing endothelial cell preferential expression came from the database mining gene set as opposed to the genes identified through subtraction cloning of TGF-β-responsive transcripts. Also of note, this statistical analysis did not find genes that were consistently expressed at lower levels by endothelial cells compared with non-endothelial cells. In some cases, individual genes were represented more than once on the microarray due to redundancy between the gene sets, and the corresponding features were handled independently in the data analysis. For this reason, they appear multiple times on the heat map.

These data were also evaluated with a Gaussian error model (GER), which employs the overlap between Gaussian curves as a ranking function, and an analogous gene list was generated through this approach (Table 1). The separation of signals between the different endothelial and non-endothelial cells is shown for the well-known endothelial enriched marker CD31, as well as two novel ESTs identified through this work (Fig. 1*B*). A comparison of results obtained with SAM, the GER, and the TNoM score is presented in the Table 1 (see METHODS).

Identification of vascular bed-selective genes. Considerable attention has been paid to differences in gene expression exhibited by endothelial cells depending on their association with different sizes and types of vessels as well as their tissues of origin (2). To investigate this issue with this dataset, analyses were conducted to identify differences in gene expression between the HAEC, HCAEC, and the HMVEC (Fig. 2). Analysis methods included the GER model for two- and multiway comparisons, and SAM for two-way comparisons. Differentially expressed genes among the different endothelial cell types included thrombomodulin, aldo-keto reductase family member 3, Ras association domain family 2, and matrix metalloproteinase 10, as

Table 1. *Endothelial cell-specific genes identified by various data analysis algorithms*

Acc. No.	Gene Name	Fold Difference	SAM Score	SAM Expect	GER Score	GER Rank	TNoM Score
AA410298	ESTs	29.7	7.33	1.43	0.027	1	0
AA256482	Hs. mRNA for FLJ00138 protein	28.6	7.21	1.31	0.052	6	3
R22412	platelet/endothelial cell adhesion molecule (CD31 antigen)	55.5	6.91	1.25	0.102	20	3
H02884	cadherin 5, type 2, VE-cadherin (vascular epithelium)	23.5	6.64	1.20	0.046	3	1
AW770514	ESTs, Highly similar to AF161403 1 HSPC285 [*H. sapiens*]	20.4	6.51	1.16	0.067	11	1
AA150505	complement component 1, q subcomponent, receptor 1	29.1	6.48	1.13	0.130	28	4
AA423867	multimerin	48.3	6.16	1.10	0.191	45	7
AA156022	roundabout homolog 4, magic roundabout (*Drosophila*)	14.4	5.99	1.08	0.113	22	0
M28526	platelet/endothelial cell adhesion molecule (CD31 antigen)	14.4	5.96	1.06	0.115	23	3
AA875933	EGF-containing fibulin-like extracellular matrix protein 1	25.0	5.93	1.04	0.098	18	3
AA485883	von Willebrand factor	10.0	5.88	1.03	0.065	10	1
W46577	endothelial cell-specific molecule 1	27.6	5.85	1.02	0.130	27	3
N93476	endothelial differentiation, sphingolipid G-protein-coupled recept, 1	10.4	5.67	1.00	0.080	13	0
AI422298	Hs., similar to hypothetical protein FLJ11110	11.9	5.43	0.99	0.145	35	4
M28526	platelet/endothelial cell adhesion molecule (CD31 antigen)	12.9	5.41	0.98	0.057	7	3
N54398	Hs. cDNA FLJ32566 fis, clone SPLEN2000047	10.9	5.24	0.97	0.101	19	4
AF297711	netrin 4	5.1	5.18	0.96	0.062	8	2
AI056548	hedgehog interacting protein	10.5	5.16	0.95	0.181	42	3
AI888888	melanoma cell adhesion molecule	11.6	5.10	0.94	0.111	21	2
W81545	EST	10.6	5.04	0.93	0.148	36	1
N75004	ESTs	11.2	4.91	0.92	0.170	40	5
AA428201	EGF-TM7-latrophilin-related protein	12.8	4.91	0.91	0.193	46	5
W68141	protein kinase, cAMP-dependent, catalytic, alpha	5.2	4.86	0.90	0.044	2	0
AA424833	bone morphogenetic protein 6	6.5	4.80	0.90	0.121	25	2
AA418728	palmdelphin	6.3	4.79	0.89	0.091	16	1
AA004368	hypothetical protein FLJ21269	7.0	4.68	0.88	0.217	53	2
M28526	platelet/endothelial cell adhesion molecule (CD31 antigen)	11.2	4.63	0.88	0.190	44	2
H11003	endothelin 1	6.5	4.62	0.87	0.129	26	4
X14787	thrombospondin 1	3.2	4.61	0.86	0.177	41	4
AF297711	netrin 4	5.7	4.60	0.86	0.083	14	2
X14787	thrombospondin 1	3.2	4.58	0.85	0.177	41	4
AI279830	protein phosphatase 1, regulatory (inhibitor) subunit 16B	6.9	4.54	0.84	0.063	9	2
AI261621	phospholipid scramblase 4	9.8	4.50	0.84	0.091	17	3
AA010753	ras homolog gene family, member J	6.2	4.44	0.83	0.168	39	6
AA460732	DKFZP564D0764 protein	5.5	4.38	0.83	0.140	34	4
M28526	platelet/endothelial cell adhesion molecule (CD31 antigen)	14.4	4.34	0.82	0.115	23	3
M28526	platelet/endothelial cell adhesion molecule (CD31 antigen)	14.4	4.33	0.82	0.115	23	5
AF297711	netrin 4	5.1	4.09	0.81	0.062	8	2
R21535	Hs. cDNA FLJ11724 fis, clone HEMBA1005331	9.0	4.05	0.81	0.290	79	5
R09728	serum deprivation response (phosphatidylserine binding protein)	4.9	3.91	0.80	0.247	67	6
AA418988	Hs. mRNA; cDNA DKFZp586L0120 (from clone DKFZp586L0120)	5.0	3.89	0.80	0.136	33	3
T47442	protein C receptor, endothelial (EPCR)	5.0	3.83	0.79	0.301	84	6
N64508	podocalyxin-like	8.2	3.80	0.79	0.398	145	7
AW772163	hypothetical protein FLJ20401	4.4	3.76	0.78	0.258	72	6
D67029	SEC14-like 1 (S. cerevisiae)	2.9	3.69	0.78	0.354	110	9
N66734	Hs. mRNA; cDNA DKFZp586N0121 (from clone DKFZp586N0121)	4.1	3.63	0.77	0.226	58	7
M28526	platelet/endothelial cell adhesion molecule (CD31 antigen)	8.8	3.58	0.77	0.223	56	9
H22922	manic fringe homolog (*Drosophila*)	6.2	3.57	0.77	0.316	88	6
AA669136	transcription factor 4	3.7	3.56	0.76	0.245	64	6
AA668698	ESTs	3.4	3.54	0.76	0.212	52	3
Z99289	laminin, alpha 4	4.9	3.47	0.75	0.222	55	9
AA026831	kinase insert domain receptor (a type III receptor tyrosine kinase)	5.5	3.44	0.75	0.331	96	9
AA147552	ESTs	5.8	3.42	0.75	0.133	31	8
AF064493	LIM domain binding 2	5.9	3.41	0.74	0.367	120	12
AA598794	connective tissue growth factor	4.2	3.31	0.74	0.320	91	11
BC001631	thymosin, beta 4, X chrm.	2.6	3.31	0.73	0.121	24	4
XI4787	thrombospondin 1	3.2	3.29	0.73	0.177	41	4
D67029	SEC14-like 1 (*S. cerevisiae*)	3.7	3.26	0.73	0.136	32	6
AI299356	Hs. cDNA FLJ31414 fis, clone NT2NE2000260	3.6	3.26	0.72	0.250	68	8
AA155913	matrix Gla protein	5.6	3.23	0.72	0.555	205	12
AF092132	Hs. PAK2 mRNA, complete cds	4.5	3.19	0.72	0.539	401	15
N54794	serine (or cysteine) proteinase inhibitor, clade E (nexin, PAI1)	3.8	3.18	0.71	0.220	54	4
AF064493	LIM domain binding 2	5.9	3.16	0.71	0.367	120	12
Z99289	laminin, alpha 4	5.7	3.15	0.71	0.130	29	8

Significance analysis of microarrays (SAM) identified 64 genes with significantly higher expression in endothelial cells (four types) versus non-endothelial cells (5 types) with a false detection rate of 0.0. The genes are ranked by SAM score, and the expected SAM score (SAM Expected) is shown for comparison. The SAM score is a *t* statistic with the denominator modified by the addition of a small positive constant to ensure independence of variance and gene expression level (variance is high at low expression levels). Expected SAM scores represent the "null distribution" in this data set. After random permutation of values between groups within each gene, SAM scores are ranked across the data set, and the process is repeated several hundred times. The "expected" score for each gene is then the average SAM score across the permutations for its rank position in the actual analysis (38). Fold difference was calculated as the difference between average log ratio expression in one class and log ratio expression in the other class. For this computation, average expression was the mean value calculated after excluding the lower 25% and the top 25% of values. The Gaussian error (GER) rank is the relative position of the gene when all genes are ordered on the basis of the GER score, which is calculated on the basis of overlap between Gaussian curves (see MATERIALS AND METHODS). The threshold number of misclassifications (TnoM) is a nonparametric score representing how well a gene separates two sample classes (5). TNoM counts the minimal number of errors committed by using a threshold on the expression values for this separation.

well as a number of ESTs and hypothetical proteins deduced from the human genome sequence. When the analyses were focused still further, a large number of markers were identified that distinguished between the two types of large vessel endothelial cells, the HAEC and the HCAEC (Fig. 3*A*). Representative genes expressed at higher levels in HAEC included those encoding factors involved in coagulation, including multimerin, thrombospondin, tissue factor pathway inhibitor, and von Willebrand factor. Representative genes expressed at higher levels in HCAEC included the chemokine mononuclear cell attractant cytokine (MCP-1) and the endothelial cell-selective protein kinase A scaffold protein gravin. There were differences in gene expression for extracellular matrix factors, including the basement membrane protein collagen type IV α2, and the matrix gla protein that regulates calcification of matrix.

Since microvascular endothelial cells have markedly different functions from those lining conduit vessels, the specific expression profile of this cell type was investigated. When microvascular cells were compared with all other endothelial cell types, employing the GER model and SAM, several hundred genes were found to be preferentially expressed in HMVEC, and ~30 such genes are shown in the heat map in Fig. 3*B*. Examination of this set of genes provided insight into the unique functions of the microvasculature. For example, selective expression of the vascular endothelial cell growth factor receptor KDR was detected by SAM in HMVEC (data not shown). This result was consistent with known physiology, since the angiogenic pro-

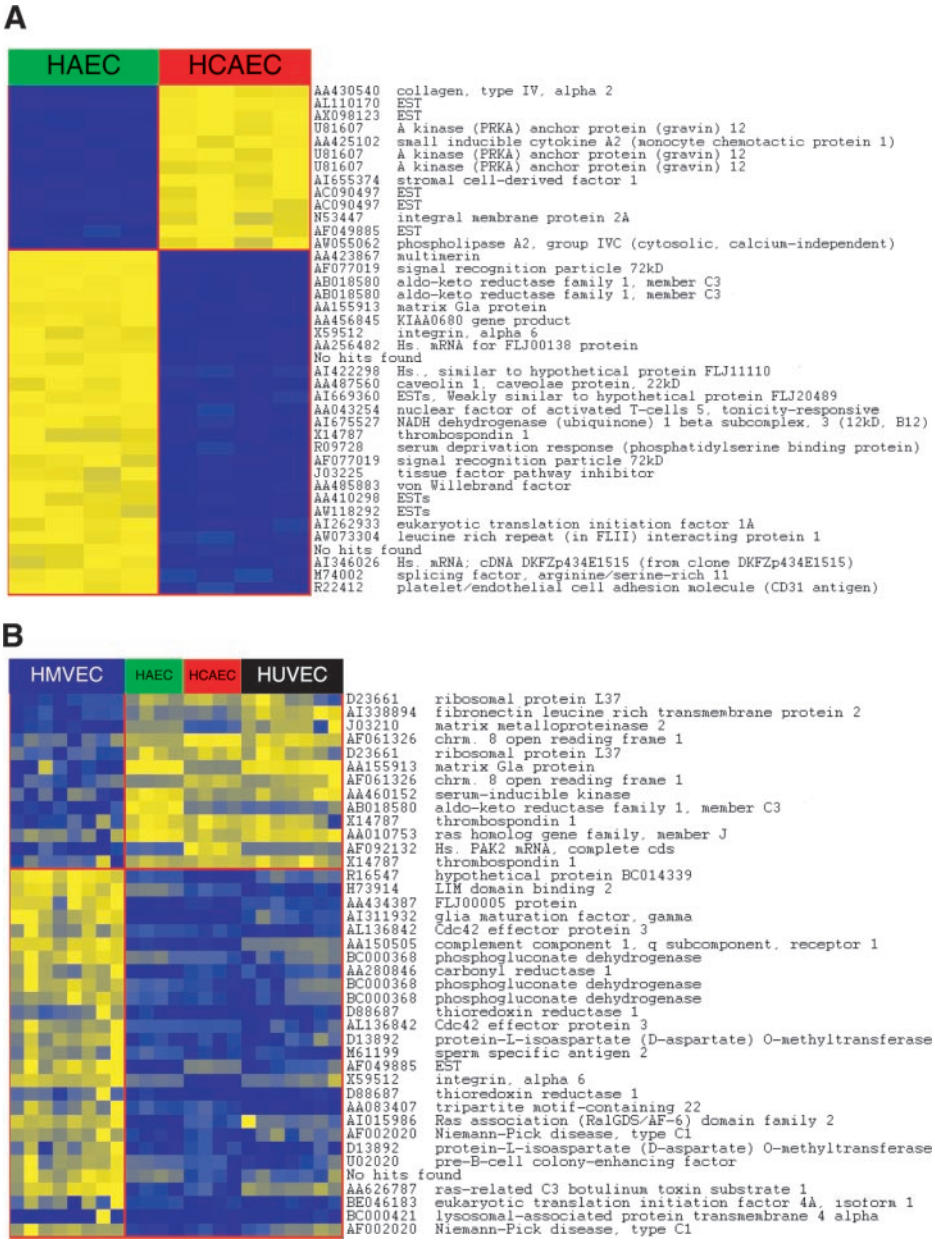

Fig. 3. Identification of genes that can distinguish between HCAEC and HAEC and genes that can distinguish HMVEC from other endothelial cell types. *A*: a portion of the genes found to be preferentially expressed in HAEC (*bottom*) and HCAEC (*top*), employing the Gaussian error model. The gene name "No hits found" was given to genes that did not match any known sequences in the publicly available databases. *B*: genes were identified that were preferentially expressed at higher levels (*bottom*) and lower levels (*top*) in HMVEC compared with other endothelial cell types. The Gaussian error model was employed for this analysis.

gram activated by this receptor is mediated at the level of the microcirculation. Differences in expression of factors that mediate cytoplasmic signaling pathways were also evident, and included PAK2 and cdc42 effector protein 3. Many of the known genes identified through this analysis encode proteins that have not been studied in the context of endothelial cell biology and those genes found to be more highly expressed in HMVEC included LIM binding domain 2, glia maturation factor-γ, thioredoxin reductase, integrin α6-subunit, and pre-B cell colony-enhancing factor. A small number of genes were expressed at lower levels in HMVEC compared with all other endothelial cell types, and these included matrix metalloproteinase 2, matrix gla protein, and thrombospondin (Fig. 3*B*).

RNA blot and in situ hybridization studies. To validate the microarray methodology and confirm the data analysis, a number of RNA blot experiments were conducted. ESTs AW770514, AA256482, AI261621, AI422298, and AA428201 were chosen because of their relative endothelial cell enriched expression, as well as positive controls CD31 and multimerin (Fig. 4*A*). Comparison was made between the Northern blot data, and the quantitative information provided by the array experiments as presented by heat map (Fig. 1) and visualized with Gaussian curves (Fig. 4*B*). Each of the ESTs was found to be primarily endothelial cell enriched, with no detectable expression in the non-endothelial cell lanes. Notable exceptions were AI261621 and AA428201, which showed significant hybridization to the HASMC RNA sample. Even after correcting for differences in RNA loading between the lanes, differences in hybridization signal were observed between the different endothelial cell types for some genes. Multimerin to a great extent and CD31 to a lesser extent showed more hybridization to the HAEC RNA sample. The EST AI422298 showed a single band with greater hybridization to the HAEC, and AW770514 had two specific bands with both showing somewhat greater hybridization to the HAEC RNA. EST AI261621 showed less hybridization to HUVEC RNA than the other genes. These differences correlated with the data obtained by microarray (see, for example, Fig. 3*A*). The separation of signals between the HAEC and HCAEC cell types is shown with Gaussian distribution curves for four of the genes evaluated by Northern blot (Fig. 4*B*).

To evaluate in vivo expression of ESTs identified as endothelial cell enriched, in situ hybridization was performed with staged mouse embryos to determine whether these genes might be expressed in early stages of angiogenic blood vessel development (Fig. 5). Of the 10 genes studied, five exhibited unique embryonic endothelial cell expression patterns. The patterns varied from expression in all endothelial cells, as observed for one of the ESTs that was determined to represent the AA4 gene (Fig. 5*A*), to expression in a restricted group of endothelial cells (20). For instance, the mouse ortholog of human EST AW772163 revealed expression restricted to the vitelline veins and aorta but also showed a temporal expression pattern in the somites (Fig. 5*B*). The mouse ortholog of human EST W81545 revealed highly restricted expression in the developing blood vessels in the brain and the yolk sac (Fig. 5C).

DISCUSSION

Although considerable attention has been paid to the concept of "endothelial cell-specific" gene expression, in fact virtually all of the genes that have been characterized in this regard exhibit expression in at least several additional cell types. The CD31 gene is considered to be one of the most specific for the endothelial cell; however, it is well known to be expressed by monocytes and a number of other leukocyte cell populations. Although the concept is flawed, the recognition that certain genes which are preferentially expressed by the endothelial cell lineage may have unique and important functional roles in this cell has led to significant findings. This has been most striking for genes that were cloned and shown to have important roles in embryonic vascular development. For instance, receptor tyrosine kinases such as tek/tie-1 and tie-2 and their ligands the angiopoietins have provided important new insights into the basic biology of blood vessel assembly (11). The recently described endothelial cell lipase gene, LIPG, serves as an example of a gene identified through efforts aimed at endothelial cell gene cloning (15, 33). Characterization of the lipid metabolism enzyme encoded by this gene has significantly expanded our understanding of the role of the vascular wall in lipid metabolism and suggests novel pathways for atherosclerotic disease development.

Through experiments reported here we have sought to expand the repertoire of genes that are specifically or preferentially expressed in the endothelial cell. Several known endothelial marker genes were shown to be exclusively expressed in the cultured cells, including CD31, von Willebrand factor, VE cadherin, multimerin, and ESM-1, confirming that endothelial markers in vitro are likely to have in vivo relevance. A total of 64 clones were found to be differentially expressed in endothelial cells with at least a threefold difference (range 3–55). Forty-four of these came from the "virtual subtraction" clone set, validating the utility of the expression database searching algorithms. A number of these genes were ESTs, representing an uncharacterized group of encoded proteins that likely serve functions that are highly restricted to the vessel wall and are thus of greatest interest. These genes are predicted to have diverse cellular functions, because of their conserved protein motifs. For instance, the Src homology 2 (SH2) domain (EST AA256482) is predicted to mediate cytoplasmic signaling functions, and EGF and lectin C-type domains are predicted to mediate protein-protein interactions such as ligand receptor binding (EST AW770514). Genes that are not specific for endothelial cells, but preferentially expressed in the endothelial cell with a greater than threefold higher level of expression, are also interesting since they provide insights into a number of biological processes that are potentially important for this vascular cell type.

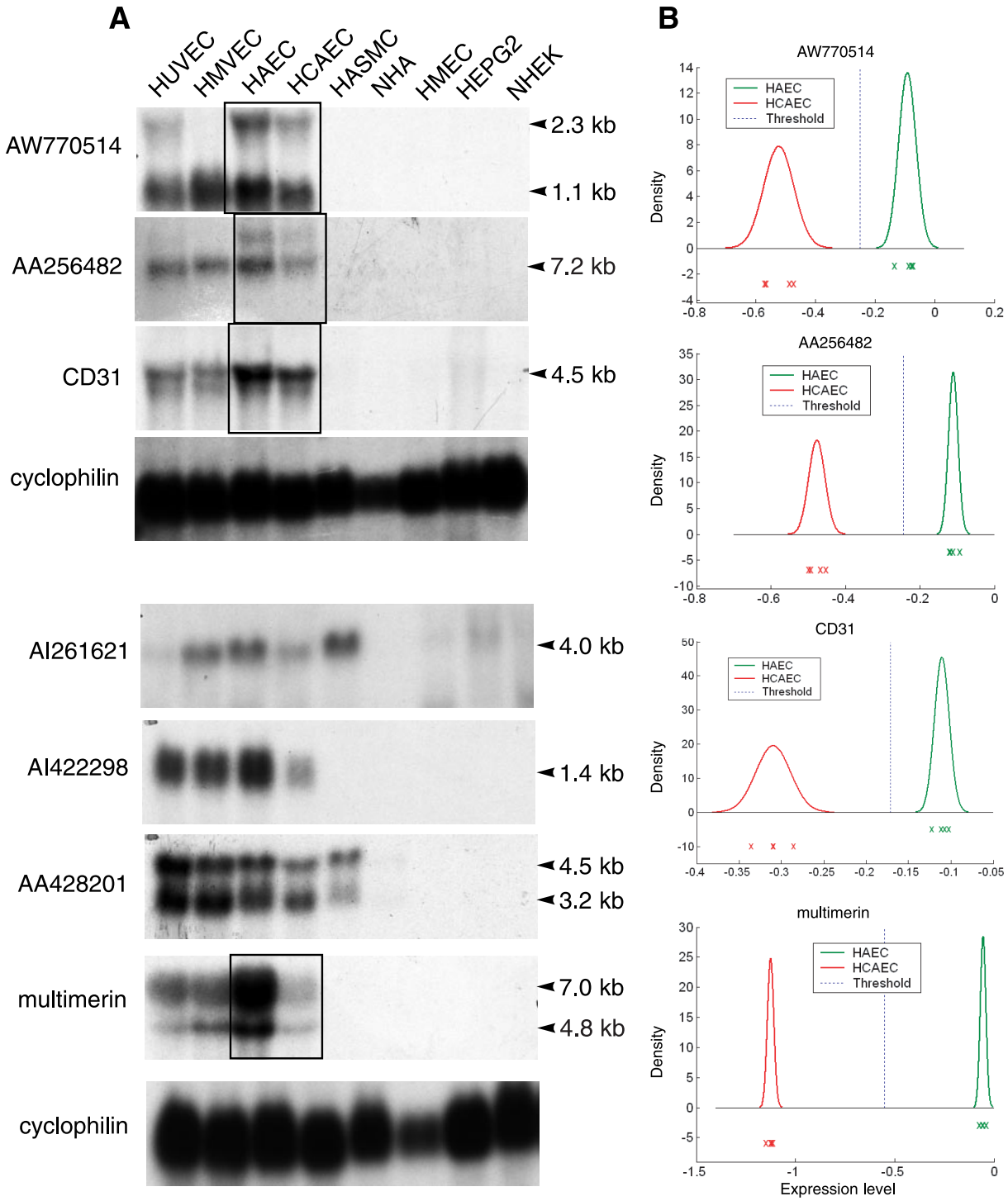

Fig. 4. Northern blot validation of endothelial cell-specific gene expression. *A*: endothelial and non-endothelial RNAs were evaluated by RNA blot studies employing radiolabeled probes for five putative endothelial cell-specific ESTs as well as two positive control genes, CD31 and multimerin. ESTs AW770514, AA256482, and AI422298 are highly endothelial restricted, whereas AI261621 and AA428201 show expression in smooth muscle cells as well as endothelial cells. The relative preponderance of the multimerin, AW770514, AA256482, CD31, and AI422298 messages in HAEC confirms the microarray data. *B*: Gaussian distributions of normalized red/green ratios derived from microarray studies investigating cell-specific expression of genes shown in *A*.

Interestingly, a number of genes were identified that have been characterized in different settings to have fundamental roles in the development or function of other tissues but have not been previously linked to the endothelial cell. In some cases, important reagents have been developed, and there have been extensive studies elucidating the basic cellular functions of the encoded factors. If these genes are shown to be ex-

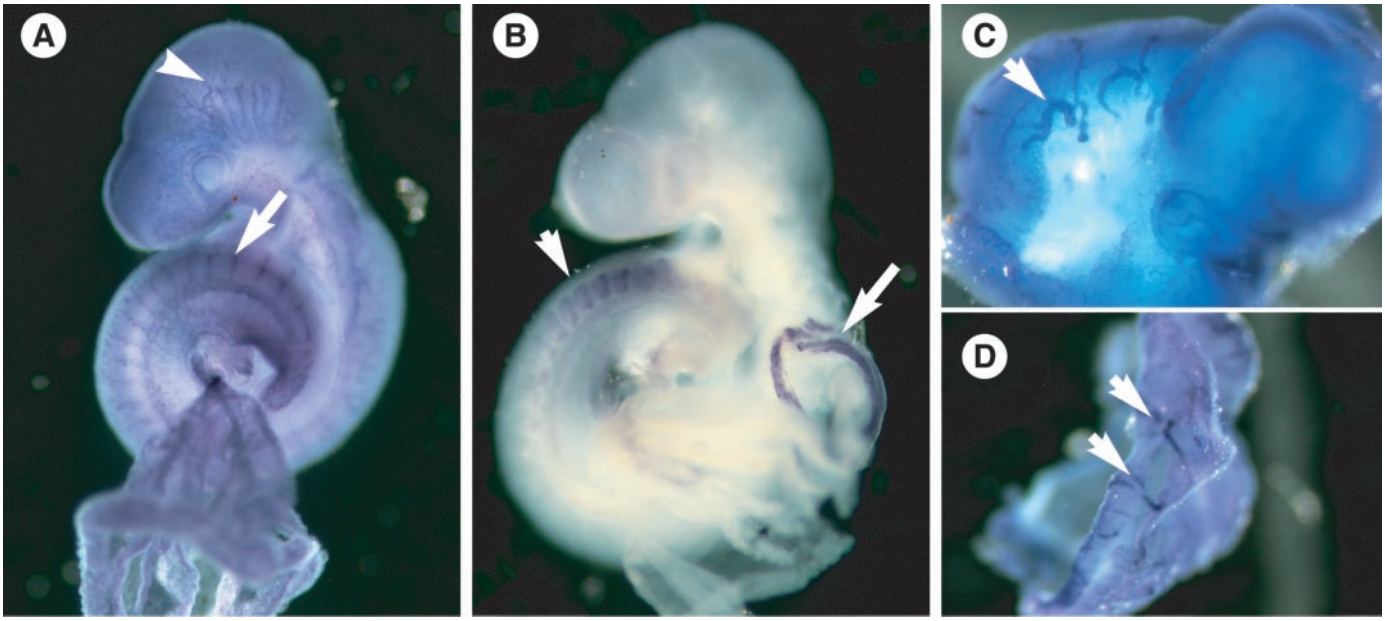

Fig. 5. Whole mount in situ hybridization of putative endothelial cell-specific ESTs identified through data mining and expression analysis. Digoxigenin-labeled cRNA probes were synthesized from cDNAs representing mouse orthologs of human sequences and hybridized to embryonic *day 9.5* mouse embryos. *A*: one gene identified through these studies was a gene initially characterized in hematopoietic cells, AA4 (mouse AA145088) (20). The in situ pattern suggests expression in all endothelial cells at this stage of development, including those forming the intersomitic vessels (arrow) and those in the developing brain (arrowhead). *B*: in situ hybridization with the mouse ortholog of human EST AW772163 (mouse BG172926) revealed expression in endothelial cells of the aorta and vitelline veins (arrows), as well as cells of the developing somites (arrowhead). The mouse ortholog of human EST W81545 (mouse AA797701) showed hybridization to endothelial cells in vascular structures in the developing brain (*C*, arrow) and yolk sac (*D*, arrows).

pressed by endothelial cells in vivo, then available reagents and insights may be rapidly applied to investigate the functional role of these factors in the vessel wall. For instance, the netrin-4 precursor gene was identified through these studies as being preferentially expressed in most cultured endothelial cells. Although this gene has been identified in nervous system cells and epithelial cells, it has not been previously noted to be expressed by endothelial cells. Netrins are secreted proteins that direct axon extension and cell migration during neural development, providing a framework for nervous system wiring. It is an intriguing possibility that endothelial cells may express this netrin family member during vascular formation, linking these two developmental processes characterized by cell migration and cell-cell recognition and communication.

The specific functions of endothelial cells vary, depending on the size and nature of the vascular structure and the tissue where they reside. For instance, endothelial cells in the spleen and other lymphoid tissues express specific addressins that direct lymphoid cell homing. Increased definition of such differentially expressed endothelial cell genes will provide for more complete understanding of the signals that regulate selective endothelial cell expression and the spectrum of phenotypes that this differential gene expression confers. When expression profiles were compared between the endothelial cells isolated from the aorta, coronary artery, and microvascular circulation, there was evidence of significant differences in gene expression. Genes were identified that clearly distinguished between endothelial cells from the three vascular beds, extending our notion of endothelial cell variability. Comparison of gene expression profiles was particularly informative for the comparison of HAEC vs. HCAEC, with a large number of genes showing preferentially higher expression in one or the other cell type. For instance, genes such as von Willebrand factor and multimerin have been considered to be universal endothelial cell markers, but these data suggest that they are preferentially expressed, along with the tissue factor pathway inhibitor and thrombospondin, in aortic endothelial cells. Also, the gene encoding a bone development factor, matrix gla protein, has significantly higher specific expression in aortic endothelial cells, where it may regulate calcification of the vessel wall and thus contribute to vascular disease in individuals with risk factors. Such insights will provide important new experimental directions for investigation of vascular disease. Also, these data further draw into question the concept of "endothelial cell-specific" gene expression, as these data clearly indicate even the most endothelial enriched genes are differentially expressed between different vascular beds.

A large number of genes were expressed at higher levels in the microvascular cells, some of which are likely linked to the functional role of the microvasculature in angiogenesis. For instance, PAK2 and Cdc42 interacting protein are differentially expressed between microvascular cells and other endothelial cell

types. These factors, which are part of the Rho/Rac/Cdc42 pathway, mediate critical links between activation pathways and cytoskeleton and gene expression. Such pathways are critical for cell movement and shape changes that are characteristic features of angiogenesis.

We cannot be certain whether the information derived from this approach is predictive of gene expression patterns in vivo. However, there is reason to believe that those genes identified as pan-endothelial cell markers will show endothelial enriched patterns of expression in vivo. First, a number of well-known genes with highly restricted in vivo endothelial-specific expression were identified with this methodology. Second, a number of novel ESTs identified through this work were shown to be expressed in vivo in a vascular endothelial cell pattern in the developing embryo. Interestingly, these markers were not expressed in all endothelial cells in the embryo but showed different patterns of gene expression, providing surprising new insights regarding the heterogeneity of developing vascular beds. Of course, at later stages of development or in the adult, these genes may show a more generalized endothelial pattern of expression. Also, the novel genes that did not show evidence of endothelial expression in the embryo might show endothelial expression in normal or pathophysiological conditions in the adult. Perhaps the most interesting aspect of these microarray experiments is the identification of putative vascular bed-specific markers. This area of investigation remains one of great interest, as it has been appreciated for some time that endothelial cells from different organs have different functional profiles. Evidence provided here that specific genes are preferentially expressed in endothelial cells from one vascular bed vs. another will require further study, with in vivo techniques such as in situ hybridization.

Previous efforts have employed data mining strategies to identify endothelial cell-specific genes. One group employed xProfiler as we did, along with a direct dbEST recursive BLAST algorithm, and combined these two methods to identify four new ESTs, including one novel gene that was felt to be a member of the roundabout family of factors that regulate neuronal development (19). Our work validated three of the ESTs that were identified by these investigators, "ESTs-ECSM2" (Hs.30089/AA410298), "ESTs-ECSM3" (Hs.8135/AA150505), and "ESTs (Hs.233955)". Another of their ESTs was found to be expressed by non-endothelial cell types and was not included in our list of endothelial-specific genes. Use of additional expression databases, and the addition of transcriptional profiling as a mechanism to confirm endothelial cell restricted expression, has allowed the identification of a larger and more well-defined group of novel genes. Our approach combining data mining and transcriptional profiling is similar to one employed in the study of neural crest-derived melanocyte gene expression (23). These investigators used database analysis to identify a set of expressed sequences that were derived primarily from neural crest-melanocyte tissues, with these genes then being employed for microarray construction and expression profiling. Until such time that all genes, or preferably all exons, are available for study on microarrays, it will be advantageous to have comprehensive tissue-appropriate gene sets on microarrays employed for transcriptional profiling.

We thank to Jennifer King and Euan Ashley for helping with the final data analyses and countless loose ends. We also thank Peter Tsang for contributions to the computational tools used for the data analysis.

This work was supported by the Donald W. Reynolds Cardiovascular Clinical Research Center at Stanford University.

REFERENCES

1. **Abe M and Sato Y.** cDNA microarray analysis of the gene expression profile of VEGF-activated human umbilical vein endothelial cells. *Angiogenesis* 4: 289–298, 2001.
2. **Auerbach R, Alby L, Morrissey LW, Tu M, and Joseph J.** Expression of organ-specific antigens on capillary endothelial cells. *Microvasc Res* 29: 401–411, 1985.
3. **Bell SE, Mavila A, Salazar R, Bayless KJ, Kanagala S, Maxwell SA, and Davis GE.** Differential gene expression during capillary morphogenesis in 3D collagen matrices: regulated expression of genes involved in basement membrane matrix assembly, cell cycle progression, cellular differentiation and G-protein signaling. *J Cell Sci* 114: 2755–2773, 2001.
4. **Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, and Yakhini Z.** Tissue classification with gene expression profiles. *J Comput Biol* 7: 559–583, 2000.
5. **Ben-Dor A, Friedman N, and Yakhini Z.** Class discovery in gene expression data. *Proceedings of the Fifth International Conference on Computational Biology 2001*, p. 31–38.
6. **Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, Sampas N, Dougherty E, Wang E, Marincola F, Gooden C, Lueders J, Glatfelter A, Pollock P, Carpten J, Gillanders E, Leja D, Dietrich K, Beaudry C, Berens M, Alberts D, and Sondak V.** Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 406: 536–540, 2000.
7. **Chakravarti L and Roy J.** *Handbook of Methods of Applied Statistics*. New York: Wiley, 1967, vol. 1.
8. **Chen BPC, Li YS, Zhao Y, Chen KD, Li S, Lao J, Yuan S, Shyy JYJ, and Chien S.** DNA microarray analysis of gene expression in endothelial cells in response to 24-h shear stress. *Physiol Genomics* 7: 55–63, 2001.
9. **Dekker RJ, van Soest S, Fontijn RD, Salamanca S, de Groot PG, VanBavel E, Pannekoek H, and Horrevoets AJ.** Prolonged fluid shear stress induces a distinct set of endothelial cell genes, most specifically lung Kruppel-like factor (KLF2). *Blood* 100: 1689–1698, 2002.
10. **Diatchenko L, Lau YF, Campbell AP, Chenchik A, Moqadam F, Huang B, Lukyanov S, Lukyanov K, Gurskaya N, Sverdlov ED, and Siebert PD.** Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proc Natl Acad Sci USA* 93: 6025–6030, 1996.
11. **Gale NW and Yancopoulos GD.** Growth factors acting via endothelial cell-specific receptor tyrosine kinases: VEGFs, angiopoietins, and ephrins in vascular development. *Genes Dev* 13: 1055–1066, 1999.
12. **Garcia-Cardena G, Comander J, Anderson KR, Blackman BR, and Gimbrone MA Jr.** Biomechanical activation of vascular endothelium as a determinant of its functional phenotype. *Proc Natl Acad Sci USA* 98: 4478–4485, 2001.
13. **Gerritsen ME, Soriano R, Yang S, Ingle G, Zlot C, Toy K, Winer J, Draksharapu A, Peale F, Wu TD, and Williams PM.** In silico data filtering to identify new angiogenesis targets from a large in vitro gene profiling data set. *Physiol Genomics* 10: 13–20, 2002. First published May 15, 2002; 10.1152/physiolgenomics.00035.2002

14. **Henrique D, Adam J, Myat A, Chitnis A, Lewis J, and Ish-Horowicz D.** Expression of a Delta homologue in prospective neurons in the chick. *Nature* 375: 787–790, 1995.
15. **Hirata K, Dichek HL, Cioffi JA, Choi SY, Leeper NJ, Quintana L, Kronmal GS, Cooper AD, and Quertermous T.** Cloning of a unique lipase from endothelial cells extends the lipase gene family. *J Biol Chem* 274: 14170–14175, 1999.
16. **Hishiki T, Kawamoto S, Morishita S, and Okubo K.** BodyMap: a human and mouse gene expression database. *Nucleic Acids Res* 28: 136–138, 2000.
17. **Hollander M and Wolfe DA.** *Nonparametric Statistical Methods.* New York: Wiley, 1973.
18. **Horrevoets AJ, Fontijn RD, van Zonneveld AJ, de Vries CJ, ten Cate JW, and Pannekoek H.** Vascular endothelial genes that are responsive to tumor necrosis factor-α in vitro are expressed in atherosclerotic lesions, including inhibitor of apoptosis protein-1, stannin, and two novel genes. *Blood* 93: 3418–3431, 1999.
19. **Huminiecki L and Bicknell R.** In silico cloning of novel endothelial-specific genes. *Genome Res* 10: 1796–1806, 2000.
20. **Jordan CT, McKearn JP, and Lemischka IR.** Cellular and developmental properties of fetal hematopoietic stem cells. *Cell* 61: 953–963, 1990.
21. **Kallmann BA, Wagner S, Hummel V, Buttmann M, Bayas A, Tonn JC, and Rieckmann P.** Characteristic gene expression profile of primary human cerebral endothelial cells. *FASEB J* 16: 589–591, 2002.
22. **Kawamoto S, Yoshii J, Mizuno K, Ito K, Miyamoto Y, Ohnishi T, Matoba R, Hori N, Matsumoto Y, Okumura T, Nakao Y, Yoshii H, Arimoto J, Ohashi H, Nakanishi H, Ohno I, Hashimoto J, Shimizu K, Maeda K, Kuriyama H, Nishida K, Shimizu-Matsumoto A, Adachi W, Ito R, Kawasaki S, and Chae KS.** BodyMap: a collection of 3′ ESTs for analysis of human gene expression information. *Genome Res* 10: 1817–1827, 2000.
23. **Loftus SK, Chen Y, Gooden G, Ryan JF, Birznieks G, Hilliard M, Baxevanis AD, Bittner M, Meltzer P, Trent J, and Pavan W.** Informatic selection of a neural crest-melanocyte cDNA set for microarray analysis. *Proc Natl Acad Sci USA* 96: 9277–9280, 1999.
24. **MacDonald TJ, Brown KM, LaFleur B, Peterson K, Lawlor C, Chen Y, Packer RJ, Cogen P, and Stephan DA.** Expression profiling of medulloblastoma: PDGFRA and the RAS/MAPK pathway as therapeutic targets for metastatic disease. *Nat Genet* 29: 143–152, 2001.
25. **Mayanil CS, George D, Freilich L, Miljan EJ, Mania-Farnell B, McLone DG, and Bremer EG.** Microarray analysis detects novel Pax3 downstream target genes. *J Biol Chem* 276: 49299–49309, 2001.
26. **McCormick SM, Eskin SG, McIntire LV, Teng CL, Lu CM, Russell CG, and Chittur KK.** DNA microarray reveals changes in gene expression of shear stressed human umbilical vein endothelial cells. *Proc Natl Acad Sci USA* 98: 8955–8960, 2001.
27. **Murakami T, Mataki C, Nagao C, Umetani M, Wada Y, Ishii M, Tsutsumi S, Kohro T, Saiura A, Aburatani H, Hamakubo T, and Kodama T.** The gene expression profile of human umbilical vein endothelial cells stimulated by tumor necrosis factor alpha using DNA microarray analysis. *J Atheroscler Thromb* 7: 39–44, 2000.
28. **Peale FV Jr and Gerritsen ME.** Gene profiling techniques and their application in angiogenesis and vascular development. *J Pathol* 195: 7–19, 2001.
29. **Pepper MS.** Transforming growth factor-β: vasculogenesis, angiogenesis, and vessel wall integrity. *Cytokine Growth Factor Rev* 8: 21–43, 1997.
30. **Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, and Botstein D.** Molecular portraits of human breast tumours. *Nature* 406: 747–752, 2000.
31. **Pilpel Y, Sudarsanam P, and Church GM.** Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat Genet* 29: 153–159, 2001.
32. **Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JY, Goumnerova LC, Black PM, Lau C, Allen JC, Zagzag D, Olson JM, Curran T, Wetmore C, Biegel JA, Poggio T, Mukherjee S, Rifkin R, Califano A, Stolovitzky G, Louis DN, Mesirov JP, Lander ES, and Golub TR.** Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* 415: 436–442, 2002.
33. **Rader DJ and Jaye M.** Endothelial lipase: a new member of the triglyceride lipase gene family. *Curr Opin Lipidol* 11: 141–147, 2000.
34. **Rickman DS, Bobek MP, Misek DE, Kuick R, Blaivas M, Kurnit DM, Taylor J, and Hanash SM.** Distinctive molecular profiles of high-grade and low-grade gliomas based on oligonucleotide microarray analysis. *Cancer Res* 61: 6885–6891, 2001.
35. **Schadt EE, Li C, Ellis B, and Wong WH.** Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J Cell Biochem Suppl* 37: 120–125, 2001.
36. **Storch KF, Lipan O, Leykin I, Viswanathan N, Davis FC, Wong WH, and Weitz CJ.** Extensive and divergent circadian gene expression in liver and heart. *Nature* 417: 78–83, 2002.
37. **Topper JN, Cai J, Qiu Y, Anderson KR, Xu YY, Deeds JD, Feeley R, Gimeno CJ, Woolf EA, Tayber O, Mays GG, Sampson BA, Schoen FJ, Gimbrone MA Jr, and Falb D.** Vascular MADs: two novel MAD-related genes selectively inducible by flow in human vascular endothelium. *Proc Natl Acad Sci USA* 94: 9314–9319, 1997.
38. **Tusher VG, Tibshirani R, and Chu G.** Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 98: 5116–5121, 2001.
39. **van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, and Friend SH.** Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415: 530–536, 2002.
40. **Weston GC, Haviv I, and Rogers PA.** Microarray analysis of VEGF-responsive genes in myometrial endothelial cells. *Mol Hum Reprod* 8: 855–863, 2002.
41. **Zhang S, Day INM, and Ye S.** Microarray analysis of nicotine-induced changes in gene expression in endothelial cells. *Physiol Genomics* 5: 187–192, 2001.
42. **Zhao B, Bowden RA, Stavchansky SA, and Bowman PD.** Human endothelial cell response to gram-negative lipopolysaccharide assessed with cDNA microarrays. *Am J Physiol Cell Physiol* 281: C1587–C1595, 2001.
43. **Zhou J, Jin Y, Gao Y, Wang H, Hu G, Huang Y, Chen Q, Feng M, and Wu C.** Genomic-scale analysis of gene expression profiles in TNF-α treated human umbilical vein endothelial cells. *Inflamm Res* 51: 332–341, 2002.