In Silico Cloning of Novel Endothelial-Specific Genes

Lukasz Huminiecki and Roy Bicknell[1]

Molecular Angiogenesis Laboratory, Imperial Cancer Research Fund, Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DS, UK

The endothelium plays a pivotal role in many physiological and pathological processes and is known to be an exceptionally active transcriptional site. To advance our understanding of endothelial cell biology and to elucidate potential pharmaceutical targets, we developed a new database screening approach to permit identification of novel endothelial-specific genes. The UniGene gene index was screened using high stringency BLAST against a pool of endothelial expressed sequence tags (ESTs) and a pool of nonendothelial ESTs constructed from cell-type–specific dbEST libraries. UniGene clusters with matches in the endothelial pool and no matches in the nonendothelial pool were selected. The UniGene/EST approach was then combined with serial analysis of gene expression (SAGE) library subtraction and reverse transcription polymerase chain reaction to further examine interesting clusters. Four novel genes were identified and labeled: endothelial cell–specific molecules (*ECSM*) *1–3* and *magic roundabout* (similar to the axon guidance protein roundabout). In summary, we present a powerful novel approach for comparative expression analysis combining two datamining strategies followed by experimental verification.

In the postgenomic era, data analysis rather than data collection will present the biggest challenge to biologists. Efforts to ascribe biological meaning to genomic data, whether by identification of function, structure, or expression pattern, are lagging behind sequencing efforts (Boguski 1999). Here, we describe the use of two independent strategies for differential expression analysis combined with experimental verification to identify genes specifically or preferentially expressed in vascular endothelium.

The first strategy was based on an EST cluster expression analysis in the human UniGene gene index (Schuler et al. 1997). Recurrent gapped `BLAST` searches (Altschul et al. 1997) were performed at very high stringency against expressed sequence tags (ESTs) grouped into two pools. The two pools comprised endothelial cell and nonendothelial cell libraries derived from dbEST (Boguski et al. 1995). The second strategy used another datamining tool: SAGEmap `xProfiler`. `xProfiler` is a freely available online tool, which is a part of the NCBI's Cancer Genome Anatomy Project (CGAP) (Cole et al. 1995; Strausberg et al. 1997).

These two approaches alone produced a discouragingly high number of false positives. However, when both strategies were combined, predictions proved exceptionally reliable and four novel candidate endothelial-specific genes have been identified. For two of these genes, full-length cDNAs have been identified in sequence databases. Another gene (EST cluster) corresponds to a partial cDNA sequence from a large-scale cDNA sequencing project and contains a region of similarity to the intracellular domain of human roundabout homolog 1 (ROBO1).

[1]Corresponding author.
E-MAIL bicknelr@icrf.icnet.uk; FAX 44 (0)-1865-222431.
Article and publication are at www.genome.org/cgi/doi/10.1101/gr.150700.

RESULTS

UniGene/EST Gene Index Screen

A pool of endothelial ESTs and a pool of nonendothelial ESTs were extracted using the Sequence Retrieval System (SRS) from dbEST. The endothelial pool consisted of 11,117 ESTs from nine human endothelial libraries (Table 1). The nonendothelial pool included 173,137 ESTs from 108 human cell lines and microdissected tumor libraries (Table 2). ESTs were extracted from dbEST, release April 2000. Multiple-FASTA files were transformed into BLAST searchable databases using the pressdb program. Table 3 shows the expression status of five known endothelial cell–specific genes in these two pools: von Willebrand factor (*vWF*; Ginsburg et al. 1985); two vascular endothelial growth factor receptors, fms-like tyrosine kinase 1 (*FLT1*; Shibuya et al. 1990) and kinase insert domain receptor (*KDR*; Matthews et al. 1991); tyrosine kinase receptor type tie

Table 1. Nine Human Endothelial Libraries from dbEST

Human aortic endothelium, 20 sequences, in vitro culture
Human endothelial cells, 346 sequences, primary isolate
Human endothelial cell (Y.Mitsui), 3 sequences, in vitro culture
Stratagene endothelial cell 937223, 7171 sequences, primary isolate
Aorta endothelial cells, 1245 sequences, primary isolate
Aorta endothelial cells, TNF treated, 1908 sequences, primary isolate
Umbilical vein endothelial cells I, 9 sequences
HDMEC cDNA library, 11 sequences, in vitro culture
Umbilical vein endothelial cells II, 404 sequences

10:1796–1806 ©2000 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/00 $5.00; www.genome.org

Table 2. Nonendothelial dbEST Libraries

1. Activated T-cells I
2. Activated T-cells II
3. Activated T-cells III
4. Activated T-cells IV
5. Activated T-cells IX
6. Activated T-cells V
7. Activated T-cells VI
8. Activated T-cells VII
9. Activated T-cells VIII
10. Activated T-cells X
11. Activated T-cells XI
12. Activated T-cells XII
13. Activated T-cells XX
14. CAMA1Ee cell line I
15. CAMA1Ee cell line II
16. CCRF-CEM cells, cyclohexamide treated I
17. cDNA library of activated B cell line 3D5
18. Chromosome 7 HeLa cDNA Library
19. Colon carcinoma (Caco-2) cell line I
20. Colon carcinoma (Caco-2) cell line II
21. Colon carcinoma (HCC) cell line
22. Colon carcinoma (HCC) cell line II
23. HCC cell line (matastasis to liver in mouse)
24. HCC cell line (matastasis to liver in mouse) II
25. HeLa cDNA (T. Noma)
26. HeLa SRIG (Synthetic retinoids induced genes)
27. Homo sapiens monocyte-derived macrophages
28. HSC172 cells I
29. HSC172 cells II
30. Human 23132 gastric carcinoma cell line
31. Human breast cancer cell line Bcap 37
32. Human cell line A431 subclone
33. Human cell line AGZY-83a
34. Human cell line PCI-O6A
35. Human cell line PCI-O6B
36. Human cell line SK-N-MC
37. Human cell line TF-1 (D.L.Ma)
38. Human exocervical cells (CGLee)
39. Human fibrosarcoma cell line HT1080
40. Human fibrosarcoma cell line HT1080-6TGc5
41. Human gastric cancer SGC-7901 cell line
42. Human GM-CSF-deprived TF-1 cell line (Liu, Hongtao)
43. Human HeLa (Y. Wang)
44. Human HeLa cells (M. Lovett)
45. Human Jurkat cell line mRNA (K. Thiele)
46. Human K562 erythroleukemic cells
47. Human lung cancer cell line A549.A549
48. Human nasopharyngeal carcinoma cell line HNE1
49. Human neuroblastoma SK-ER3 cells (M. Garnier)
50. Human newborn melanocytes (T. Vogt)
51. Human pancreatic cancer cell line Patu 8988t
52. Human primary melanocytes mRNA (I.M. Eisenbarth)
53. Human promyelocytic HL60 cell line (S. Herblot)
54. Human retina cell line ARPE-19
55. Human salivary gland cell line HSG
56. Human White blood cells
57. Jurkat T-cells I
58. Jurkat T-cells II
59. Jurkat T-cells III
60. Jurkat T-cells V
61. Jurkat T-cells VI
62. Liver HepG2 cell line
63. LNCAP cells I
64. Macrophage I
65. Macrophage II
66. Macrophage, subtracted (total cDNA)

Table 2. (*Continued*)

67. MCF7 cell line
68. Namalwa B cells I
69. Namalwa B cells II
70. NCI_CGAP_Br4
71. NCI_CGAP_Br5
72. NCI_CGAP_CLL1
73. NCI_CGAP_GCB0
74. NCI_CGAP_GCB1
75. NCI_CGAP_HN1
76. NCI_CGAP_HN3
77. NCI_CGAP_HN4
78. NCI_CGAP_HSC1
79. NCI_CGAP_Li1
80. NCI_CGAP_Li2
81. NCI_CGAP_Ov5
82. NCI_CGAP_Ov6
83. NCI_CGAP_Pr1
84. NCI_CGAP_Pr10
85. NCI_CGAP_Pr11
86. NCI_CGAP_Pr16
87. NCI_CGAP_Pr18
88. NCI_CGAP_Pr2
89. NCI_CGAP_Pr20
90. NCI_CGAP_Pr24
91. NCI_CGAP_Pr25
92. NCI_CGAP_Pr3
93. NCI_CGAP_Pr4
94. NCI_CGAP_Pr4.1
95. NCI_CGAP_Pr5
96. NCI_CGAP_Pr6
97. NCI_CGAP_Pr7
98. NCI_CGAP_Pr8
99. NCI_CGAP_Pr9
100. Normal human trabecular bone cells
101. Raji cells, cyclohexamide treated I
102. Retinal pigment epithelium 0041 cell line
103. Retinoid treated HeLa cells
104. Soares melanocyte 2NbHM
105. Soares_senescent_fibroblasts_NbHSF
106. Stratagene HeLa cell s3 937216
107. Supt cells
108. T, human adult rhabdomyosarcoma cell line

Table 3. Five Known Endothelial-Specific Genes in the dbEST Pools

Known endothelial-specific gene	Hits in the nonendothelial pool	Hits in the endothelial pool
von Willebrand factor (*vWF*)	1	27
FLT1 VEGF receptor	—	—
KDR VEGF receptor	1	—
TIE1 tyrosine kinase	—	5
TIE2/TEK tyrosine kinase	—	2

The number of expressed sequence tags (ESTs) in the endothelial pool is relatively small (~11, 117), and not all known endothelial genes are represented.

(*TIE1*; Partanen et al. 1992); and tyrosine kinase receptor type tek (*TIE2/TEK*; Vikkula et al. 1996).

Optimizing the BLAST E-value was crucial for the success of BLAST identity-level searches. Too high an E-value would result in gene paralogs being reported. In contrast, too low (stringent) an E-parameter would result in many false negatives, i.e., true positives would not be reported because of sequencing errors in EST data; ESTs are large-scale, low-cost single pass sequences and have a high error rate (Aaronson et al. 1996). In this work an E-value of 10e–20 was used in searches against the nonendothelial EST pool and a more stringent 10e–30 value was used in searches against the smaller endothelial pool. These values were deemed optimal after a series of test BLAST searches.

SAGE Data and SAGEmap `xProfiler` Differential Analysis

Internet-based SAGE library subtraction (SAGEmap `xProfiler`) was used as the second datamining strategy for the identification of novel endothelial-specific or preferentially endothelial genes. Two endothelial SAGE libraries (SAGE_Duke_HMVEC and SAGE_Duke_HMVEC + VEGF with a total of 110,790 sequences) were compared with 24 nonendothelial cell line libraries (full list in Table 4, total of 733,461 sequences). Table 5 shows the status of expression of the five reference endothelial-specific genes in these two SAGE pools.

Combined Data Gives Highly Accurate Predictions

Twenty known genes were selected in the UniGene/EST screen (Table 6). These genes had no matches in the nonendothelial pool and at least one match in the endothelial pool. The list contained four endothelial-specific genes: *TIE1* (Partanen et al. 1992), *TIE2/TEK* (Vikkula et al. 1996), *LYVE1* (Banerji et al. 1999), and *multimerin* (Hayward et al. 1998), indicating ~20% accuracy of prediction. Other genes on the list, although certainly preferentially expressed in the endothelial cells, may not be endothelial specific. To improve on the prediction accuracy, we decided to combine UniGene/EST screen with the `xProfiler` SAGE analysis. Table 7 shows how data from the two approaches were combined. Identity-level BLAST searches were performed on mRNAs (known genes) or phrap-computed contigs (EST clusters representing novel genes) to investigate how these genes were represented in the endothelial and nonendothelial pool. Subsequent experimental verification by reverse transcription polymerase chain reaction (RT-PCR; Fig. 1) proved that the combined approach was 100% accurate, i.e., genes on the `xProfiler` list that had no matches the nonendothelial EST pool and at least one match in the endothelial pool were indeed endothelial specific.

DISCUSSION

There have been several reports of computer analysis of tissue transcriptosomes. Usually an expression profile

Table 4. Twenty-four Nonendothelial Cell Serial Analysis Gene Expression (SAGE)–Cancer Genome Anatomy Project

Symbol	Description
SAGE_HCT116	Colon, cell line derived from colorectal carcinoma
SAGE_Caco_2	Colon, colorectal carcinoma cell line
SAGE_Duke_H392	Brain, Duke glioblastoma multiforme cell line
SAGE_SW837	Colon, cancer cell line
SAGE_RKO	Colon, cancer cell line
SAGE_NHA(5th)	Brain, normal human astrocyte cells harvested at passage 5
SAGE_ES2-1	Ovarian clear cell carcinoma cell line ES-2, poorly differentiated
SAGE_OVCA432-2	Ovary, carcinoma cell line OVCA432
SAGE_OV1063-3	Ovary, carcinoma cell line OV1063
SAGE_Duke_mhh-1	Brain, c-myc negative medulloblastoma cell line mhh-1
SAGE_Duke-H341	Brain, c-myc positive medulloblastoma cell line H341
SAGE_HOSE_4	Ovary, normal surface epithelium
SAGE_OVP-5	Ovary, pooled cancer cell lines
SAGE_LNCaP	Prostate, cell line, androgen dependent
SAGE_HMEC-B41	Cell culture HMEC-B41 of normal human mammary epithelial cells
SAGE_MDA453	Cell line MDA-MB-453 of human breast carcinoma
SAGE_SKBR3	ATCC cell line SK-BR-3, human breast adenocarcinoma
SAGE_A2780-9	Ovary, ovarian cancer cell line A2780
SAGE_Duke_H247_normal	Brain, glioblastoma multiforme cell line H247
SAGE_Duke_H247_Hypoxia	Brain, Duke glioblastoma multiforme cell line H247, grown under 1.5% oxygen
SAGE_Duke_post_crisis_fibroblasts	Skin, postcrisis survival fibroblast cell line
SAGE_Duke_precrisis_fibroblasts	Skin, large T antigen transformed human fibroblasts clones
SAGE_A	Prostate, cancer cell line, induced with synthetic androgen
SAGE_IOSE29-11	Ovary, surface epithelium line

Table 5. Five Known Endothelial-Specific Genes in the CGAP SAGE Pools

Known endothelial-specific gene	Tags in the nonendothelial SAGE libraries	Tags in the endothelial SAGE libraries
von Willebrand factor (vWF)	1 (colon carcinoma cell line)	80
FLT1 VEGF receptor	—	—
KDR VEGF receptor	1 (IOSE29 ovarian surface epithelium cell line)	6
TIE1 tyrosine kinase	17 (ovarian tumour and normal ovarian epithelium cell lines)	27
TIE2/TEK tyrosine kinase	4 (ovarian carcinoma and glioblastoma multiforme cell lines)	2

TIE1 and *TIE2/TEK* have multiple hits in the nonendothelial pool (most in normal or carcinoma cell lines of ovarian origin). vWF is most endothelial specific, having 80 hits in the endothelial pool and only one hit in the nonendothelial pool.

is constructed, based on the number of tags assigned to a given gene or a class of genes (Bernstein et al. 1996; Welle et al. 1999; Bortoluzzi et al. 2000). An attempt can be made to identify tissue-specific transcripts: for example, Vasmatzis et al. (1997) described three novel genes expressed exclusively in the prostate by in silico subtraction of libraries from the dbEST collection. Purpose-made cDNA libraries may also be used. Ten candidate granulocyte-specific genes have been identified by extensive sequence analysis of cDNA libraries derived from granulocytes and eleven other tissue samples, namely a hepatocyte cell line, fetal liver, infant liver, adult liver, subcutaneous fat, visceral fat, lung, colonic mucosa, keratinocytes, cornea, and retina (Itoh et al. 1998).

An analysis similar to the dbEST-based approach taken by Vasmatzis et al. (1997) is complicated by the fact that endothelial cells are present in all tissues of the body and endothelial ESTs are contaminating all bulk tissue libraries. To validate this, we used three well-known endothelial-specific genes—*KDR*, *FLT1*, and *TIE-2*—as queries for BLAST searches against dbEST. Transcripts were present in a wide range of tissues, with multiple hits in well-vascularized tissues (e.g., placenta, retina), embryonic tissues (liver, spleen), or infant tissues (brain). In addition, we found

Table 6. Results of the UniGene/EST Screen

Description	UniGene ID	Endothelial hits
TIE1 receptor endothelial tyrosine kinase	Hs.78824	5
Cytosolic phospholipase A2; involved in the metabolism of eicosanoids	Hs.211587	3
Branched chain alpha-ketoacid dehydrogenase	Hs.1265	2
CGMP-dependent protein kinase; cloned from aorta cDNA, strongly expressed in well vascularised tissues like aorta, heart, and uterus (Tamura et al. 1996)	Hs.2689	2
Lymphatic vessel endothelial hyaluronan receptor *1-LYVE1* (Banerji et al. 1999)	Hs.17917	2
TRAF interacting protein: TNF signalling pathway	Hs.21254	2
Multimerin: a very big endothelial-specific protein; binds platelet factor V, can also be found in platelets (Hayward et al. 1996)	Hs.32934	2
Diubiquitin (a member of the ubiquitin family); reported in dendrytic and B lymphocyte cells; involved in antigen processing; this is first evidence that it is also present in endothelial cells (Bates et al. 1997)	Hs.44532	2
Beta-transducin family protein; also a homolog of D. melanogaster gene notchless: a novel WD40 repeat containing protein that modulates Notch signalling activity	Hs.85570	2
TIE2/TEK receptor endothelial tyrosine kinase	Hs.89640	2
BCL2 associated X protein (BAX)	Hs.159428	2
Sepiapterin reductase mRNA	Hs.160100	2
Retinoic acid receptor beta (RARB)	Hs.171495	2
ST2 receptor: a homolog of the interleukin 1 receptor	Hs.66	1
Mitogen activated protein kinase 8 (MAPK8)	Hs.859	1
ERG gene related to the ETS oncogene	Hs.45514	1
PP35 similar to *Escherichia coli* yhdg and *Rhodobacter capsulatus* nifR3	Hs.97627	1
Interphotoreceptor matrix proteoglycan; strongly expressed in retina and umbilical cord vein (Felbor et al. 1998)	Hs.129882	1
Methylmalonate semialdehyde dehydrogenase gene	Hs.170008	1
HTLV-I–related endogenous retroviral sequence	Hs.247963	1

Twenty known genes were selected in the UniGene/EST screen (no hits in the nonendothelial pool and at least one hit in the endothelial pool). At least four of these genes are known endothelial-specific genes: *TIE1, TIE2/TEK, LVYE1,* and multimerin, indicating ~20% prediction accuracy. Other genes, although certainly preferentially expressed in the endothelial cells, may not be endothelial specific.

Table 7. `xProfiler` Differential Analysis was Combined with Data from the UniGene/EST Screen Achieving 100% Certainty of Prediction

Unigene ID	Gene description	XPROFILER prediction certainty	Hits in endothelial EST pool	Hits in nonendothelial EST pool
Hs.13957	**ESTs–ECSM1**	**97**	**4**	**0**
Hs.30089	**ESTs–ECSM2**	**96**	**8**	**0**
Hs.8135	**ESTs–ECSM3**	**91**	**4**	**0**
Hs.111518	**magic roundabout, distant homology to human roundabout 1**	**100**	**4**	**0**
Hs.268107	**multimerin**	**92**	**5**	**0**
Hs.155106	calcitonin receptor-like receptor activity modifying protein 2	97	0	0
Hs.233955	ESTs	96	0	0
Hs.26530	serum deprivation response (phosphatidylserine-binding protein)	94	3	1
Hs.83213	fatty acid binding protein 4	100	3	1
Hs.110802	**von Willebrand factor**	**100**	**25**	**1**
Hs.76206	**cadherin 5, VE-cadherin (vascular endothelium)**	**100**	**4**	**1**
Hs.2271	**endothelin 1**	**98**	**9**	**2**
Hs.119129	collagen, type IV, alpha 1	100	4	6
Hs.78146	**platelet/endothelial cell adhesion molecule (CD31 antigen)**	**99**	**18**	**5**
Hs.76224	EGF-containing fibulin-like extracellular matrix protein 1	100	37	9
Hs.75511	connective tissue growth factor	100	34	48

`xProfiler's` output lists genes with 10× number of tags in the endothelial pool than in the nonendothelial pool of SAGE-CGAP libraries. Hits corresponding to these genes in the endothelial and nonendothelial EST pools were identified by identity-level BLAST searches for mRNA (known genes) or phrap-computed contig sequences (EST clusters representing novel genes). Genes are sorted according to the number of hits in the nonendothelial EST pool. Known and predicted novel endothelial specific genes are in bold.

that simple subtraction of endothelial EST libraries against all other dbEST libraries failed to identify any specific genes (data not shown).

Two very different types of expression data resources were used in our datamining efforts. The UniGene/EST screen was based on EST libraries from dbEST. There are nine human endothelial libraries in the current release of dbEST, with a relatively small total number of ESTs (11,117). Some well-known endothelial-specific genes are not represented in this data set (Table 3). This limitation raised our concerns that genes with low levels of expression would be overlooked in our analysis. Therefore, we used another type of computable expression data: CGAP SAGE libraries. SAGE tags are sometimes called small ESTs (usually 10–11 bp in length). Their major advantage is that they can be unambiguously located within the cDNA: they are immediately adjacent to the most 3′ NlaIII restriction site. Although there are only two endothelial CGAP SAGE libraries available at the moment, they contain an impressive total of ~111,000 tags—a data set approximately ten times bigger than the 11,117 sequences in the endothelial EST pool. The combined approach proved very accurate (Fig. 1; Table 8) when verified by RT-PCR. We report here identification of four novel highly endothelial-specific genes: endothelial cell–specific molecule 1 (*ECSM1*; UniGene entry Hs.13957), endothelial cell–specific molecule 2 (*ECSM2*; UniGene entry Hs.30089), endothelial cell–specific molecule 3 (*ECSM3*; UniGene entry Hs.8135), and *magic roundabout* (UniGene entry Hs.111518). For a comprehensive summary of data available on these genes, see Table 8.

ECSM1 has no protein or nucleotide homologs. It codes for a small protein of ~103 aa, the longest and most upstream open reading frame (ORF) identified in the contig sequence.

BLAST searches against the EMBL patent database revealed that *ECSM2* corresponds to the cDNA from the patent "cDNA encoding novel polypeptide from human umbilical vein endothelial cell" (Shibayama et al. 1997), EMBL acc. E10591. A 205-aa polypeptide coded by this cDNA is a transmembrane protein with a suggested role in cell adhesion in that it is serine and proline rich, although no exact function has yet been identified.

ECSM3 was found to be identical with the matrix remodeling–associated gene 4 (*MXRA4*, cDNA sequence acc. AW888224) recently identified in a screen of 40,000 genes from 552 human cDNA libraries (M. Walker, pers. comm.). The strategy was based on the assumption that coexpression implies similar function (guilt by association). In total, eight novel genes coexpressed with 21 known matrix remodelling–associated genes were identified. In the human genome, *ECSM3* is very closely associated with another endothelial-specific gene, *AA4* (Clq/MBL/SPA receptor, C1qRp, Ly68). AA4 is a transmembrane protein expressed in vascular endothelial cells, aortic hematopoietic clusters, and fetal liver hematopoietic progenitors during fetal development, with a proposed role in the development of vascular and hematopoietic systems, espe-

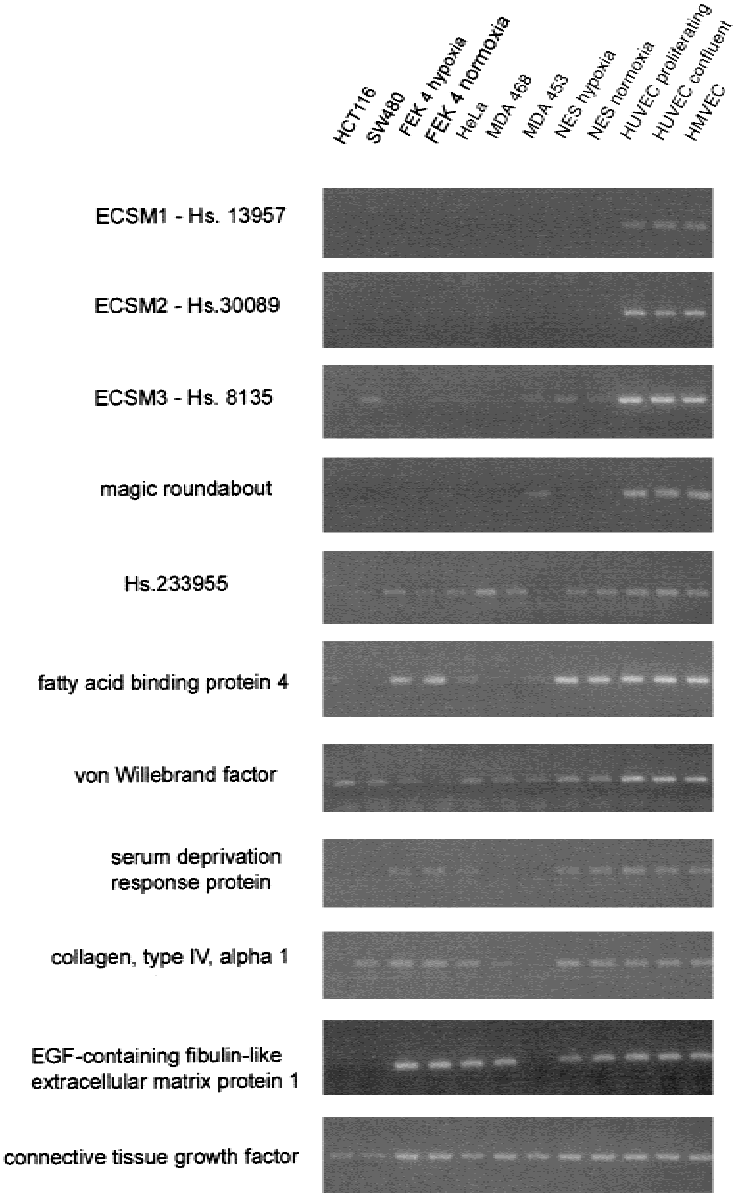

Figure 1 Experimental verification by reverse transcription polymerase chain reaction (RT-PCR). Candidate endothelial-specific genes predicted by the combination of the UniGene/EST screen and `xProfiler` serial analysis of gene expression (SAGE) differential analysis (Table 8) were checked for expression in three endothelial and nine nonendothelial cell cultures. Endothelial cultures were as follows: HMVEC (human microvascular endothelial cells), HUVEC (human umbilical vein endothelial cells) confluent culture, and HUVEC proliferating culture. Nonendothelial cultures were as follows: normal endometrial stromal (NES) cells grown in normoxia and NES grown in hypoxia, MDA 453 and MDA 468 breast carcinoma cell lines, HeLa, FEK4 fibroblasts cultured in normoxia and FEK4 fibroblasts cultured in hypoxia, SW480, and HCT116, the last two being colorectal epithelium cell lines. *ECSM1* and *ECSM2* showed complete endothelial specificity, whereas ECSM3 and *magic roundabout* were very strongly preferentially expressed in the endothelium. Interestingly, all these novel genes appear more specific than the benchmark endothelial-specific gene, *von Willebrand factor*.

cially cell-to-cell adhesion and/or signaling (Petrenko et al. 1999). By analyzing the 123,832-bp genomic clone AL118508, we found that *ECSM3* is an immediate genomic neighbor of *AA4*. Both genes were contained within the 8000-bp sequence: *MXRA4* has one exon, and *AA4* has two exons and a small intron. *MXRA4* contains a 402-bp region of strong homology (64.4% identity, E = 1.3e–24) to the 3′ untranslated region (UTR) of the mouse (and not human) *AA4* mRNA (acc. AF081789). Such an endothelial-specific gene cluster suggests existence of a functional gene expression domain (for a review on expression domains, see Dillon et al. 2000). It is also possible that *MXRA4* is a recent evolutionary insertion into the *AA4* locus, and it now exploits a part of the *AA4* regulatory sequence located in the former 3′UTR of the *AA4* gene. Because mouse *AA4* genomic structure is not available, it's impossible to say whether a gene similar to *MXRA4* is located in the vicinity. BLAST search of the full-length *MXRA4* cDNA against the mouse EST database reveals only two similar ESTs that both belong to the mouse *AA4* transcript, suggesting that *MXRA4* is not present at all in the mouse genome.

BLAST searches for the Hs.111518 contig identified a cDNA clone (GenBank acc. AK000805) with a long ORF of 417 (accession no. BAA 91382). This sequence is rich in prolines and has several regions of low amino-acid complexity. BLAST `PRODOM` search (protein families database at Human Genome Project Resource Centre) identified a 120-bp region of homology to the cytoplasmic domain conserved family of transmembrane receptors involved in repulsive axon guidance (ROBO1 DUTT1 protein family; E = 4e–07). Homology was extended to 468 aa (E = 1.3e–09) when a more rigorous analysis was performed using `ssearch` (Smith and Waterman 1981), but the region of similarity was still restricted to the cytoplasmic domain. The ROBO1 DUTT1 family comprises the human roundabout homolog 1 (ROBO1), the mouse gene DUTT1, and the rat ROBO1 (Kidd et al. 1998, Brose et al. 1999). Because of this region of homology, we called the gene represented by Hs. 111518 *magic roundabout*. In addition, `BLAST SBASE` (protein domain database at Human Genome Project Resource Centre) suggested a region of similarity to the domain of the intracellular neural cell adhesion molecule long domain form precursor (E = 2e–11). It should be noted that the true protein product

Table 8. Summary of Available Information on ECSM1-3 and Magic Roundabout

	UniGene cluster ID and size	Full-length cDNA	Longest ORF	Transmembrane segments, signal peptide	Mapping information Genomic context Genomic clones	Description
ECSMl	Hs.13957 1100 bp		103 aa confirmed with 5′RACE		Genomic neighbour: Tropomyosin dbSTS G26129 and G28043 Chr. 19 Gene Map 98: Marker SGC33470, Marker stSG3414, IntervalD19S425-D19S418 AC005945, AC005795 (partial identity)	
ECSM2	Hs.30089 1023 bp	Identical to the full-length sequence in the "cDNA encoding novel polypeptide from human umbilical vein endothelial cell" patent (Shibayama et al. 1997) 983 bp	205 aa	Clear signal peptide (SignalP), two predicted transmembrane domains (TopPred2 and DAS)		Transmembrane protein, possibly an adhesion molecule
ECSM3	Hs.8135 1694 bp	AW888224-MXRA4 Matrix remodelling–associated gene 4 (Walker et al. 2000) 3047 bp			Genomic neighbour: endothelial specific Clq/MBL/SPA receptor (C1qRp, AA4, Ly68)—both genes contained within only 8 kb-region dbSTS G06859 Chr. 20, Gene Map 98: Marker sts-W72082, Interval D20S182-D20S106 AL118508, AC011137, HSJ737E23, and AL118508.27 on the NCBI Map Viewer	Endothelial specific gene involved in matrix remodelling, possibly a novel metalloproteinase or ECM protein Possible 26 bp regulatory sequence shared with the endothelial-specific gene endothelin-converting enzyme-1 (E = 3e–04): CTTCCTGAAGCCTTGCTCCCTCCACC Possible regulatory sequence shared with the endothelial specific mouse AA4 gene (ClqRp, Ly68) 3′UTR (E = 2e–24)
Magic roundabout	Hs.111518 2076 bp	Partial cDNA FLJ20798 fis, clone ADSU02031 (acc. AK000805) 1496 bp	417 aa	One transmembrane domain predicted by TopPred2 and DAS No signal peptide detected in the available 417 aa ORF (SignalP); however, the true protein product is very likely to be larger	Genomic neighbour: integral transmembrane protein 1 (ITM1) dbSTS G14646 and G14937 Chr. 11, Gene Map 98: Marker SHGC-11739, Interval D11S1353-D11S93	468 aa region of homology to the cytoplasmic portion of the roundabout axon guidance protein family: human ROBO1, rat ROBO1, and mouse dutt1 (E = 1.3e–0.9) ORF has no apparent upstream limit. This and size comparison to ROBO1 (1651 aa) suggest that true protein is very likely to be much larger Possible alternative polyA sites: the cDNA clone from adipocyte tissue seems to be polyadenylated in a different position to the sequence from the UniGene contig

for *magic roundabout* is likely to be larger than the 417 aa coded in the AK000805 clone because the ORF has no apparent upstream limit, and size comparison to human roundabout 1 (1651 aa) suggests a much bigger protein.

Recently, intriguing associations between neuronal differentiation genes and endothelial cells have been discovered. For example, a neuronal receptor for vascular endothelial growth factor (VEGF) neuropilin 1 (Soker et al. 1998) was identified. VEGF was traditionally regarded as an exclusively endothelial growth factor. Processes similar to neuronal axon guidance are now being implicated in guiding migration of endothelial cells during angiogenic capillary sprouting. Thus, ephrinB ligands and EphB receptors are involved in demarcation of arterial and venous domains (Adams et al. 1999). It is possible that magic roundabout may be an endothelial-specific homolog of the human *roundabout 1* involved in endothelial-cell repulsive guidance, presumably with a different ligand because similarity is contained within the cytoplasmic (i.e., effector) region and guidance receptors are known to have highly modular architecture (Bashaw and Goodman 1999).

It should be noted that expression of endothelial-specific genes is not usually 100% restricted to the endothelial cell. *KDR* and *FLT1* are both expressed in the male and female reproductive tract: on spermatogenic cells (Obermair et al. 1999), on trophoblasts, and in decidua (Clark et al. 1996). *KDR* has been shown to define hematopoietic stem cells (Ziegler et al. 1999). *FLT1* is also present on monocytes. In addition to endothelial cells, *vWF* is strongly expressed in megakaryocytes (Nichols et al. 1985; Sporn et al. 1985) and, in consequence, is present on platelets. Similarly, *multimerin* is present both in endothelial cells (Hayward et al. 1993) and platelets (Hayward et al. 1998). Generally speaking, endothelial and hematopoietic cells descend from same embryonic precursors: hemangioblasts and many cellular markers are shared between these two cell lineages (for review, see Suda et al. 2000). A surprising result of our RT-PCR analysis was that the genes identified here (*ECSM1–3* and *magic roundabout*) appear to show greater endothelial specificity (Fig. 1) than does the classic endothelial marker *von Willebrand factor*.

As stated before, vascular endothelium plays a central role in many physiological and pathological processes and it is known to be an exceptionally active transcriptional site. Approximately 1000 distinct genes are expressed in an endothelial cell. In contrast, red blood cells were found to express 8 separate genes, platelets to express 22, and smooth muscle to express 127 (Adams et al. 1995). Known endothelial-specific genes attract much attention from both basic research and the clinical community. For example, endothelial-specific tyrosine kinases—*TIE1*, *TIE2/TEK*, *KDR*, and *FLT1*—are crucial players in the regulation of vascular integrity and angiogenesis (Sato et al. 1993,1995; Alello et al. 1995; Fong et al. 1995; Shalaby et al. 1995). Angiogenesis is now widely recognized as a rate-limiting process for the growth of solid tumors. It is also implicated in the formation of atherosclerotic plaques and restenosis. Finally endothelium plays a central role in the complex and dynamic system regulating coagulation and hemostasis.

Our combined datamining approach, together with experimental verification, is a powerful functional genomics tool. This type of analysis can be applied to many cell types, not just endothelial cells. The challenge of identifying the function of discovered genes remains, but bioinformatics tools such as structural genomics or homology and motif searches can offer insights that can then be verified experimentally.

METHODS

Database Sequence Retrieval

Locally stored UniGene files (Build #111, release date May 2000) were used in the preparation of the final version of this paper. The UniGene Web site can be accessed at the http://www.ncbi.nlm.nih.gov/UniGene/. UniGene files can be downloaded from the ftp repository at ftp://ncbi.nlm.nih.gov/repository/unigene/. Representative sequences for the human subset of UniGene (the longest EST within the cluster) are stored in the file Hs.seq.uniq, whereas all ESTs belonging to the cluster are stored in a separate file called Hs.seq.

Sequences were extracted from the dbEST database accessed locally at the Human Genome Project Resource Centre using the SRS (SRS version 5) getz command. This was performed repeatedly using a PERL script for all the libraries in the endothelial and nonendothelial subsets, and sequences were merged into two multiple-FASTA files.

Selection Criteria for Nonendothelial EST Libraries

Selection of 108 nonendothelial dbEST libraries was largely manual. Initially, the list of all available dbEST libraries (http://www.ncbi.nlm.nih.gov/dbEST/libs_byorg.html) was searched using the keyword "cells" and the phrase "cell line". Although this search identified most of the libraries, additional keywords had to be added for the list to be full: "melanocyte," "macrophage," "HeLa," and "fibroblast." In some cases, the detailed library description was consulted to confirm that the library is derived from a cell line/primary culture. We also added a number of CGAP microdissected-tumor libraries. For that, Library Browser (available at http://www.ncbi.nlm.nih.gov/CGAP/hTGI/lbrow/cgaplb.cgi) was used to search for the keyword "microdissected."

UniGene Gene Index Screen

The UniGene gene transcript index was screened against the EST division of GenBank, dbEST. Both UniGene and dbEST were developed at the National Centre for Biotechnology Information (NCBI). UniGene is a collection of EST clusters corresponding to putative unique genes. It currently consists of four data sets: human, mouse, rat, and zebrafish. The human

data set is comprised of approximately 90,000 clusters (UniGene Build #111, May 2000). By means of very high stringency BLAST identity searches, we aimed to identify those UniGene genes that have transcripts in the endothelial and not in the nonendothelial cell-type dbEST libraries. University of Washington BLAST2, which is a gapped version, was used as BLAST implementation. The E-value was set to 10e–20 in searches against the nonendothelial EST pool and to 10e–30 in searches against the smaller endothelial pool.

Although UniGene does not provide consensus sequences for its clusters, the longest sequence within the cluster is identified. Thus, this longest representative sequence (multiple-FASTA file Hs.seq.uniq) was searched using very high stringency BLAST against the endothelial and nonendothelial EST pool. If such representative sequence reported no matches, the rest of the sequences belonging to the cluster (UniGene multiple-FASTA file Hs.seq) followed as BLAST queries. Finally, clusters with no matches in the nonendothelial pool and at least one match in the endothelial pool were selected using PERL scripts analyzing BLAST textual output.

xProfiler SAGE Subtraction

xProfiler enables an online user to perform a differential comparison of any combination of 47 SAGE libraries with a total of ~2,300,000 SAGE tags using a dedicated statistical algorithm (Chen et al. 1998). xProfiler can be accessed at http://www.ncbi.nlm.nih.gov/SAGE/sagexpsetup.cgi. SAGE itself is a quantitative expression technology in which genes are identified by typically a 10- or 11-bp sequence tag adjacent to the cDNA's most 3′ NlaIII restriction site (Velculescu et al. 1995).

The two available endothelial cell libraries (SAGE_Duke_HMVEC and SAGE_Duke_HMVEC + VEGF) defined pool A, and 24 (see Table 4 for list) nonendothelial libraries together built pool B. The approach was verified by establishing the status of expression of the five reference endothelial-specific genes in the two SAGE pools (Table 5) using Gene to Tag Mapping (http://www.ncbi.nlm.nih.gov/SAGE/SAGEcid.cgi). Subsequently, xProfiler was used to select genes differentially expressed between the pools A and B. The xProfiler output consisted of a list of genes with a 10-fold difference in the number of tags in the endothelial compared with the nonendothelial pool sorted according to the certainty of prediction. A 90% certainty threshold was applied to this list.

The other CGAP online differential expression analysis tool, Digital Differential Display (DDD), relies on EST expression data (source library information) instead of using SAGE tags. We attempted to use this tool similarly to SAGEmap xProfiler but have been unable to obtain useful results. Five out of nine endothelial and 64 out of 108 nonendothelial cell libraries used in our BLAST-oriented approach were available for online analysis using DDD (http://www.ncbi.nlm.nih.gov/CGAP/info/ddd.cgi). When such analysis was performed, the following were the 15 top scoring genes: annexin A2, actin γ 1, ribosomal protein large P0, plasminogen activator inhibitor type I, thymosin β 4, peptidyloprolyl isomerase A, ribosomal protein L13a, laminin receptor 1 (ribosomal protein SA), eukaryotic translation elongation factor 1 α 1, vimentin, ferritin heavy polypeptide, ribosomal protein L3, ribosomal protein S18, ribosomal protein L19, and tumor protein translationally controlled 1. This list was rather surprising as it did not include any well-known endothelial-specific genes, did not have any overlap with SAGE results (Table 8), and contained many genes that in the literature are reported to be ubiquitously expressed (i.e., ribosomal proteins, actin, vimentin, ferritin). A major advantage of our UniGene/EST screen is that instead of relying on source library data and fallible EST clustering algorithms, it actually performs identity-level BLAST comparisons in search of transcripts corresponding to a gene.

Mining Data on UniGene Clusters

To quickly access information about UniGene entries (e.g., literature references, sequence tagged sites, homologs, references to function), online resources were routinely used: NCBI's UniGene and LocusLink interfaces and Online Mendelian Inheritance in Man.

ESTs in UniGene clusters are not assembled into contigs, so before any sequence analysis, contigs were created using phrap assembler (for documentation on phrap, see http://bozeman.mbt.washington.edu/phrap.docs/phrap.html).

To analyze genomic contigs AC005795 (44,399 bp) and AL118508 (123,832 bp) containing *ECSM1* and *ECSM3*, respectively, NIX Internet interface for multiapplication analysis of large unknown nucleotide sequences was used. For further information on NIX, see http://www.hgmp.mrc.ac.uk/NIX/. Alignments of *ECSM1* and *ECSM3* against AC005795 and AL118508 were obtained using the NCBI interface to the Human Genome: The NCBI Map Viewer. For further information on the NCBI Map Viewer, see http://www.ncbi.nlm.nih.gov/genome/guide/.

To search for possible transmembrane domains and signal sequences in translated nucleotide sequences, three Internet-based applications were used: DAS, http://www.biokemi.su.se/~server/DAS/ (Cserzo et al. 1997); TopPred2, http://www.biokemi.su.se/~server/toppred2/ (Heijne 1992); and SignalP, http://www.cbs.dtu.dk/services/SignalP/ (Nielsen et al. 1997).

Computing Resources

Computing resources of the Oxford University Bioinformatics Centre (http://www.molbiol.ox.ac.uk) and the Human Genome Project Resource Centre (http://www.hgmp.mrc.ac.uk) were used.

Detailed information on PERL scripts used in this work, may be obtained from L.H. (lucash@icrf.icnet.uk).

Experimental Verification

To experimentally verify specificity of expression, we used RT-PCR. RNA was extracted from three endothelial and seven nonendothelial cell types cultured in vitro. Endothelial cultures were as follows: HMVEC (human microvascular endothelial cells), HUVEC (human umbilical vein endothelial cells) confluent culture, and HUVEC proliferating culture. Nonendothelial cultures were as follows: normal endometrial stromal (NES) cells grown in normoxia and NES grown in hypoxia, MDA 453 and MDA 468 breast carcinoma cell lines, HeLa, FEK4 fibroblasts cultured in normoxia and FEK4 fibroBLASTs cultured in hypoxia, SW480, and HCT116, the last two listed being colorectal epithelium cell lines.

If a sequence tagged site was available, dbSTS PCR primers were used and cycle conditions suggested in the dbSTS entry followed. Otherwise, primers were designed using the Primer3 program. Primers are listed in Table 9.

Table 9. List of Primers Used in Reverse Transcription Polymerase Chain Reactions

Gene	Primers (sequence or GenBank Accession for the STS)
ECSM1–Hs.13957	G26129
ECSM2–Hs.30089	5′-TGG GAG AAG CAG GCA GTA TT-3′
	5′-CAG CTG CCC TGT GAC TAC AA-3′
ECSM3–Hs.8135	G06859
Magic roundabout–Hs.111518	G14937
calcitonin receptor–like receptor activity modifying 2	G26129
Hs.233955	G21261
fatty acid binding protein 4	5′-TGC AGC TTC CTT CTC ACC TT-3′
	5′-TCA CAT CCC CAT TCA CAC TG-3′
vWF	5′-TGT ACC ATG AGG TTC TCA ATG C-3′
	5′-TTA TTG TGG GCT CAG AAG GG-3′
serum deprivation response protein	G21528
collagen type IV, alpha 1	G07125
EGF-containing fibulin-like extracellular matrix protein 1	G06992
connective tissue growth factor	5′-CAA ATG CTT CCA GGT GAA AAA-3′
	5′-CGT TCA AAG CAT GAA ATG GA-3′

dbSTS primers were used if a UniGene entry contained a sequence tagged site (STS). Otherwise, primers were designed using the Primer3 programme.

Tissue Culture Media, RNA Extraction, and cDNA Synthesis

Cell lines were cultured in vitro according to standard tissue culture protocols. In particular, endothelial media were supplemented with endothelial-cell growth supplement (ECGS; Sigma) and heparin (Sigma) to promote growth. Total RNA was extracted using the RNeasy Minikit (Qiagen) and cDNA synthesized using the Reverse-IT 1st Strand Synthesis Kit (ABgene).

ACKNOWLEDGMENTS

We received extensive and patient help from many people in the British bioinformatics community, especially Drs. Sarah Butcher and John Peden from the Oxford University Bioinformatics Centre. We also thank Drs. Michael Göern and Ken Smith from the Imperial Cancer Research Fund laboratories for generous help with tissue culture techniques and preparation of RNA's for RT-PCR and Prof. Adrian Harris and Dr. Chris Norbury for stimulating discussions.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Aaronson, J.S., Eckman, B., Blevins, R.A., Borkowski, J.A., Myerson, J., Imran, S., and Elliston, K.O. 1996. Toward the development of a gene index to the human genome: An assessment of the nature of high-throughput EST sequence data. *Genome Res.* **6:** 829–845.

Adams, M.D., Kerlavage, A.R., Fleischmann, R.D., Fuldner, R.A., Bult, C.J., Lee, N.H., Kirkness, E.F., Weinstock, K.G., Gocayne, J.D., White, O., et al. 1995. Initial assessment of human gene diversity and expression patterns based on 83 million nucleotides of cDNA sequence. *Nature* **377:** 3–174.

Adams, R.H., Wilkinson, G.A., Weiss, C., Diella, F., Gale, N.W., Deutsch, U., Risau, W., and Klein, R. 1999. Roles of ephrinB ligands and EphB receptors in cardiovascular development: Demarcation of arterial/venous domains, vascular morphogenesis, and sprouting angiogenesis. *Genes & Dev.* **13:** 295–306.

Aiello, L.P., Pierce, E.A., Foley, E.D., Takagi, H., Chen, H., Riddle, L., Ferrara, N., King, G.L., and Smith, L.E.H. 1995. Suppression of retinal neovascularization in vivo by inhibition of vascular endothelial growth factor (VEGF) using soluble VEGF-receptor chimeric proteins. *Proc. Natl. Acad. Sci.* **92:** 10457–10461.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Banerji S., Ni, J., Wang, S.X., Clasper, S., Su, J., Tammi, R., Jones, M., and Jackson, D.G. 1999. LYVE-1, a new homologue of the CD44 glycoprotein, is a lymph-specific receptor for hyaluronan. *J. Cell. Biol.* **144:** 789–801.

Bashaw, G.J. and Goodman, C.S. 1999. Chimeric axon guidance receptors: The cytoplasmic domains of slit and netrin receptors specify attraction versus repulsion. *Cell* **97:** 917–926.

Bates, E.E., Ravel, O., Dieu, M.C., Ho, S., Guret, C., Bridon, J.M., Ait-Yahia, S., Briere, F., Caux, C., Banchereau, J., et al. 1997. Identification and analysis of a novel member of the ubiquitin family expressed in dendritic cells and mature B cells. *Eur. J. Immunol.* **27:** 2471–2477.

Bernstein, S.L., Borst, D.E., Neuder, M.E., and Wong, P. 1996. Characterization of the human fovea cDNA library and regional differential gene expression in the human retina. *Genomics* **32:** 301–308.

Boguski, M.S. 1999. Biosequence exegesis. *Science* **286:** 453–455.

Boguski, M.S. and Schuler, G.D. 1995. ESTablishing a human transcript map. *Nat. Genet.* **10:** 369–371.

Bortoluzzi, S., d'Alessi, F., Romualdi, C., and Danieli, G.A. 2000. The human adult skeletal muscle transcriptional profile reconstructed by a novel computational approach. *Genome Res.* **10:** 344–349.

Brose, K., Bland, K.S., Wang, K.H., Arnott, D., Henzel, W., Goodman, C.S., Tessier-Lavigne, M., and Kidd, T. 1999. Slit proteins bind Robo receptors and have an evolutionarily conserved role in repulsive axon guidance. *Cell* **96:** 795–806.

Chen, H., Centola, M., Altschul, S.F., and Metzger, H. 1998. Characterization of gene expression in resting and activated mast cells. *J. Exp. Med.* **188:** 1657–1668.

Clark, D.E., Smith, S.K., Sharkey, A.M., and Charnock-Jones, D.S. 1996. Localization of VEGF and expression of its receptors flt

and KDR in human placenta throughout pregnancy. *Hum. Reprod.* **11:** 1090–1098.
Cole, K.A., Krizman, D.B., and Emmert-Buck, M.R. 1999. The genetics of cancer—a 3D model. *Nat. Genet.* **21:** 38–41.
Cserzo, M., Wallin, E., Simon, I., von Heijne, G., and Elofsson, A. 1997. Prediction of transmembrane α-helices in prokaryotic membrane proteins: The dense alignment surface method. *Protein Eng.* **6:** 673–676.
Dillon, N. and Sabbattini, P. 2000. Functional gene expression domains: Defining the functional unit of eukaryotic gene regulation. *Bioessays* **7:** 657–665.
Felbor, U., Gehrig, A., Sauer, C.G., Marquardt, A., Kohler, M., Schmid, M., and Weber, B.H.F. 1998. Genomic organization and chromosomal localization of the interphotoreceptor matrix proteoglycan-1 (IMPG1) gene: A candidate for 6q-linked retinopathies. *Cytogenet. Cell. Genet.* **81:** 12–17.
Fong, G.H., Rossant, J., and Breitman, M.L. 1995. Role of the Flt-1 receptor tyrosine kinase in regulating the assembly of vascular endothelium. *Nature* **376:** 65–69.
Gerhold, D. and Caskey, C.T. 1996. It's the genes! EST access to human genome content. *Bioessays* **18:** 973–981.
Ginsburg, D., Handin, R.I., Bonthron, D.T., Donlon, T.A., Bruns, G.A., Latt, S.A., and Orkin, S.H. 1985. Human von Willebrand factor (vWF): Isolation of complementary DNA (cDNA) clones and chromosomal localization. *Science* **228:** 1401–6.
Hayward, C.P., Bainton, D.F., Smith, J.W., Horsewood, P., Stead, R.H., Podor, T.J., Warkentin, T.E., and Kelton, J.G. 1993. Multimerin is found in the α-granules of resting platelets and is synthesized by a megakaryocytic cell line. *J. Clin. Invest.* **91:** 2630–2639.
Hayward, C.P., Cramer, E.M., Song, Z., Zheng, S., Fung, R., Masse, J.M., Stead, R.H., and Podor, T.J. 1998. Studies of multimerin in human endothelial cells. *Blood* **91:** 1304–1317.
Hayward, C.P.M., Rivard, G.E., Kane, W.H., Drouin, J., Zheng, S., Moore, J.C., and Kelton, J.G. 1996. An autosomal dominant, qualitative platelet disorder associated with multimerin deficiency, abnormalities in platelet factor V, thrombospondin, von Willebrand factor, and fibrinogen and an epinephrine aggregation defect. *Blood* **87:** 4967–4978.
Heijne, G. 1992. Membrane Protein Structure Prediction, Hydrophobicity Analysis and the Positive-inside Rule. *J. Mol. Biol.* **225:** 487–494.
Itoh, K., Okubo, K., Utiyama, H., Hirano, T., Yoshii, J., and Matsubara, K. 1998. Expression profile of active genes in granulocytes. *Blood* **15:** 1432–1441.
Kidd, T., Brose, K., Mitchell, K.J., Fetter, R.D., Tessier-Lavigne, M., Goodman, C.S., and Tear, G. 1998. Roundabout controls axon crossing of the CNS midline and defines a novel subfamily of evolutionarily conserved guidance receptors. *Cell* **92(2):** 205–15.
Matthews, W., Jordan, C.T., Gavin, M., Jenkins, N.A., Copeland, N.G., and Lemischka, I.R. 1991. A receptor tyrosine kinase cDNA isolated from a population of enriched primitive hematopoietic cells and exhibiting close genetic linkage to c-kit. *Proc. Natl. Acad. Sci.* **88:** 9026–9030.
Nichols, W.L., Gastineau, D.A., Solberg, L.A., and Mann, K.G. 1985. Identification of human megakaryocyte coagulation factor V. *Blood* **65:** 1396–1406.
Nielsen, H., Engelbrecht, J., Brunak, S., and Heijne, G. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10:** 1–6.
Obermair, A., Obruca, A., Pohl, M., Kaider, A., Vales, A., Leodolter, S., Wojta, J., and Feichtinger, W. 1999. Vascular endothelial growth factor and its receptors in male fertility. *Fertil. Steril.* **72:** 269–275.
Partanen, J., Armstrong, E., Makela, T.P., Korhonen, J., Sandberg, M., Renkonen, R., Knuutila, S., Huebner, K., and Alitalo, K. 1992. A novel endothelial cell surface receptor tyrosine kinase with extracellular epidermal growth factor homology domains. *Mol. Cell Biol.* **12:** 1698–1707.
Petrenko, O., Beavis, A., Klaine, M., Kittappa, R., Godin, I., and Lemischka, I.R. 1999. The molecular characterization of the fetal stem cell marker AA4. *Immunity* **10:** 691–700.
Sato, T.N, Qin, Y., Kozak, C.A., and Audus, K.L. 1993. Tie-1 and tie-2 define another class of putative receptor tyrosine kinase genes expressed in early embryonic vascular system. *Proc. Nat. Acad. Sci.* **90:** 9355–9358.
Sato, T.N., Tozawa, Y., Deutsch, U., Wolburg-Buchholz, K., Fujiwara, Y., Gendron-Maguire, M., Gridley, T., Wolburg, H., Risau, W., and Qin, Y. 1995. Distinct roles of the receptor tyrosine kinases Tie-1 and Tie-2 in blood vessel formation. *Nature* **376:** 70–74.
Schuler, G.D. 1997. Pieces of the puzzle: Expressed sequence tags and the catalog of human genes. *J. Mol. Med.* **75:** 694–698.
Shalaby, F., Rossant, J., Yamaguchi, T.P., Gertsenstein, M., Wu, X.F., Breitman, M.L., and Schuh, A.C. 1995. Failure of blood-island formation and vasculogenesis in Flk-1-deficient mice. *Nature* **376:** 62–65.
Shibayama, S., Hirano, J., and Ono, H. 1997. cDNA encoding novel polypeptide from human umbilical vein endothelial cell. European Patent Office. Publication number: 0 682 113 A2.
Shibuya, M., Yamaguchi, S., Yamane, A., Ikeda, T., Tojo, A., Matsushime, H., and Sato, M. 1990. Nucleotide sequence and expression of a novel human receptor-type tyrosine kinase gene (flt) closely related to the fms family. *Oncogene* **5:** 519–524.
Smith, T.F. and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* **147:** 195–197.
Soker, S., Takashima, S., Miao, H.Q., Neufeld, G., and Klagsbrun, M. 1998. Neuropilin-1 is expressed by endothelial and tumor cells as an isoform-specific receptor for vascular endothelial growth factor. *Cell* **92:** 735–745.
Sporn, L.A., Chavin, S.I., Marder, V.J., and Wagner, D.D. 1985. Biosynthesis of von Willebrand protein by human megakaryocytes. *J. Clin. Invest.* **76:** 1102–1106.
Strausberg, R.L., Dahl, C.A., and Klausner, R.D. 1997. New opportunities for uncovering the molecular basis of cancer. *Nat. Genet.* **15:** 415–416.
Suda, T., Takakura, N., and Oike, Y. 2000. Hematopoiesis and angiogenesis. *Int. J. Hematol.* **71:** 99–107
Tamura, N., Itoh, H., Ogawa, Y., Nakagawa, O., Harada, M., Chun, T.H., Suga, T., Yoshimasa, T., and Nakao, K. 1996. cDNA cloning and gene expression of human type I-α cGMP-dependent protein kinase. *Hypertension* **27:** 552–557.
Vasmatzis, G., Essand, M., Brinkmann, U., Byungkook, L., and Pastan, I. 1997. Discovery of three genes specifically expressed in human prostate by expressed sequence tag database analysis. *Proc. Natl. Acad. Sci.* **95:** 300–304.
Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. 1995. Serial analysis of gene expression. *Science* **270:** 484–487.
Vikkula, M., Boon, L.M., Carraway 3rd, K.L., Calvert, J.T., Diamonti, A.J., Goumnerov, B., Pasyk, K.A., Marchuk, D.A. ,Warman, M.L., Cantley, L.C., et al. 1996. Vascular dysmorphogenesis caused by an activating mutation in the receptor tyrosine kinase TIE2. *Cell* **87:** 1181–1190.
Welle, S., Bhatt, K., and Thornton, C.A. 1999. Inventory of high-abundance mRNAs in skeletal muscle of normal men. *Genome Res.* **9:** 506–513.
Ziegler, B.L., Valtieri, M., Porada, G.A., De Maria, R., Muller, R., Masella, B., Gabbianelli, M., Casella, I., Pelosi, E., Bock, T., et al. 1999. KDR receptor: A key marker defining hematopoietic stem cells. *Science* **285:** 1553–1558.

Received June 1, 2000; accepted in revised form August 29, 2000.