DNA RESEARCH 2, 37–43 (1995)

Prediction of the Coding Sequences of Unidentified Human Genes. III. The Coding Sequences of 40 New Genes (KIAA0081-KIAA0120) Deduced by Analysis of cDNA Clones from Human Cell Line KG-1

Takahiro NAGASE, Nobuyuki MIYAJIMA, Ayako TANAKA,
Takashi SAZUKA, Naohiko SEKI, Shusei SATO, Satoshi TABATA,
Ken-ichi ISHIKAWA, Yutaka KAWARABAYASI, Hirokazu KOTANI, and Nobuo NOMURA*
Kazusa DNA Research Institute, 1532-3 Yana Uchino, Kisarazu, Chiba 292, Japan

(Received 14 February 1995)

Abstract

We isolated full-length cDNA clones from size-fractionated cDNA libraries of human immature myeloid cell line KG-1, and the coding sequences of 40 genes were newly predicted. A computer search of the GenBank/EMBL databases indicated that the sequences of 14 genes were unrelated to any reported genes, while the remaining 26 genes carried some sequences with similarities to known genes. Significant transmembrane domains were identified in 17 genes, and protein motifs that matched those in the PROSITE motif database were identified in 11 genes. Northern hybridization analysis with 18 different cells and tissues demonstrated that 10 genes were apparently expressed in a cell-specific or tissue-specific manner. Among the genes predicted, half were isolated from the medium-sized cDNA library and the other half from the small-sized cDNA library, and their average sizes were 4 kb and 1.4 kb, respectively. As judged by Northern hybridization profiles, small-sized cDNAs appeared to be expressed more ubiquitously and abundantly in various tissues, compared with that of medium-sized cDNAs.

Key words: full-length cDNA sequence; unidentified human gene; protein motif; mRNA expression; chromosomal location; myeloid cell line KG-1

1. Introduction

To accumulate information on the structure of unidentified human genes, we have begun a sequencing project of full-length cDNA clones of human cells.[1] The cell source used was human immature myeloid cell line KG-1. As to the sequencing strategy, size-fractionated cDNA libraries were constructed, from which cDNA clones that carry unreported sequences were first isolated. Then, the sizes of the mRNA corresponding to these clones were analyzed by Northern hybridization, and the entire nucleotide sequences of clones that comprised nearly full-length transcripts were deduced.[1] By using this strategy, we have already predicted the coding sequences of 80 new genes.[1,2] In this paper, we report the coding sequences of an additional 40 genes newly determined and their sequence features as well as expression profiles.

Communicated by Mituru Takanami

* To whom correspondence should be addressed. Tel. +81-438-52-3930, Fax. +81-438-52-3931

2. Materials and Methods

The source of cDNA libraries and the methods used for sequence analysis, chromosomal mapping of cDNA clones and computer analysis of sequences are identical to those used in the previous study.[1] Expression profiles in human tissues were examined using human multiple tissue Northern blots (Human MTN blots) from Clontech (California, USA).

3. Results and Discussion

3.1. Sequence features of analyzed cDNA clones

The clones that carried inserts longer than 2 kb and those that retained 1–2 kb inserts were randomly selected from the libraries constructed from the medium- and small-sized cDNA classes, respectively. After the 5′-terminal sequences of the inserts were determined by single-run sequencing, similarities of the sequences were searched using the FASTA program, and the clones which showed no significant similarities to reported sequences were selected. The sizes of their inserts were then compared with those of corresponding transcripts by Northern hybridization, and the clones that carried inserts

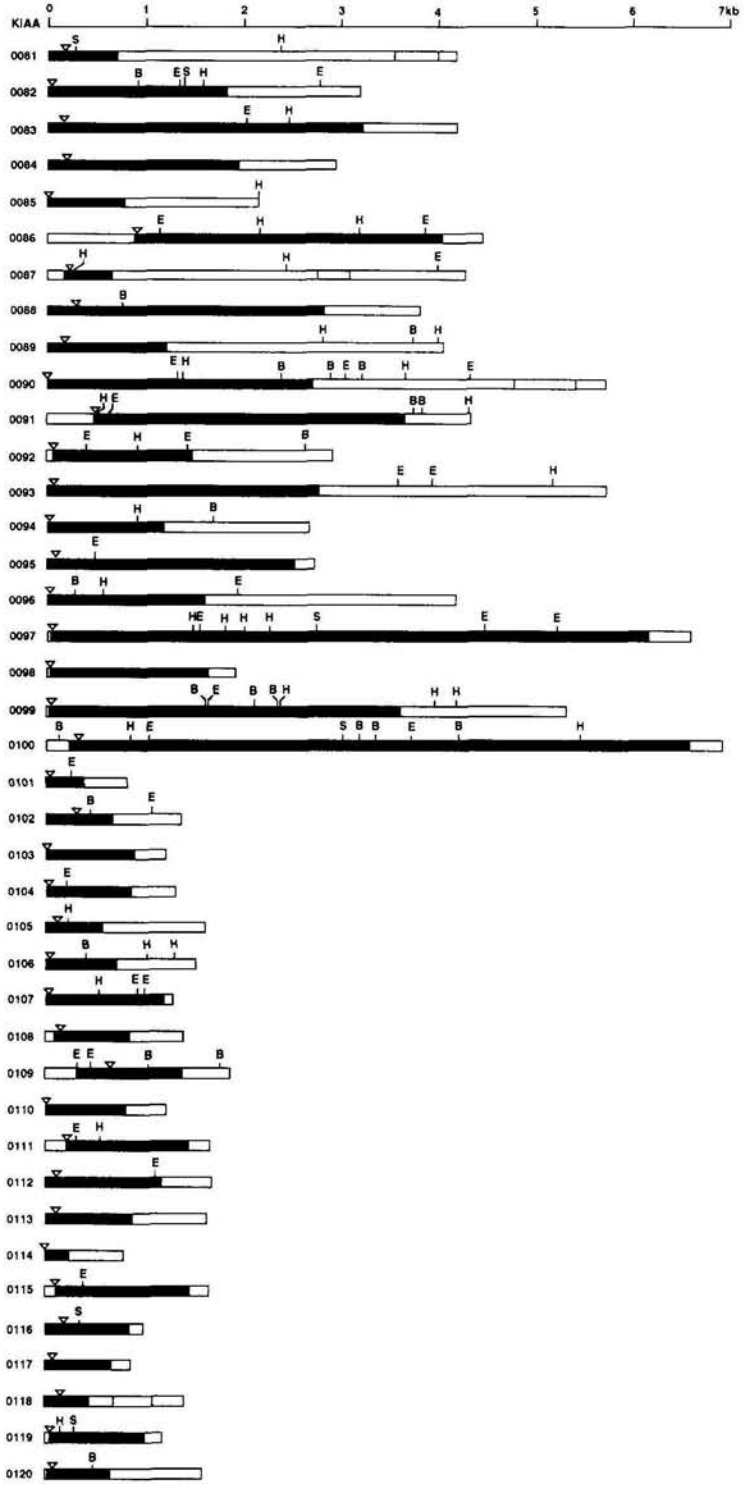

Figure 1. Physical maps of the 40 cDNA clones analyzed. The horizontal scale represents the cDNA length in kb, and gene numbers are given on the left. Open reading frames (ORFs) within coding regions, untranslated regions and repetitive sequences are indicated by solid, open and dotted boxes, respectively. The positions of the first ATG codon in each ORF are represented by triangles. The repetitive sequences observed are shown by hatched boxes. The major restriction sites are indicated above the sequences by the following abbreviations: B, *Bam*HI; E, *Eco*RI; H, *Hin*dIII; S, *Sal*I. The nucleotide sequence data reported in this paper were deposited in the GSDB, DDBJ, EMBL and NCBI nucleotide sequence databases under the accession numbers shown in Table 3.

Table 1. cDNA clones with similarities to the Genbank/EMBL database files.

Gene no. (KIAA)	Database files	Accession no.[a)]	Identity (%)	Overlap[b)] (amino acid residues)
0082	ORF69 (ACNPV)	L22858	30.5	174
0083	Chromosome III Cosmid 9986 (Sc)	U00027	41.0	458
0085	KIAA0108 (H)	D14696	31.3	261
0086	DNA repair protein SNM1 gene (Sc)	X64004	33.3	192
0088	Glucan 1,4-α-glucosidase (Sc)	Z36098	41.5	590
0089	Glycerol-3-phosphate dehydrogenase (H)	L34041	71.8	344
0090	Hypothetical protein YCL045C (Sc)	S19374[c)]	29.8	289
0091	Subtilisin (Ba)	A00970	30.4	303
0092	Smooth muscle myosin (H)	X69292	17.3	405
0093	NEDD4 gene (M)	D10714	84.4	488
0094	Methionine aminopeptidase 1 (Sc)	M77092	51.8	369
0095	Nucleoporin-interacting protein NIC96 gene (Sc)	X72923	22.4	840
0096	Protein kinase (M)	U11494	28.9	166
0098	Chaperonin containing TCP-1 (M)	Z31555	95.8	546
0099	Pumilio protein (D)	L07943	42.1	875
0100	el protein (M)	X81632	95.3	1182
0102	Signal peptidase complex SPC 25 (Do)	U12687	95.1	225
0106	B15C gene (Hv)	X76605	48.9	231
0108	KIAA0085 (H)	D42042	31.3	261
0109	Clathrin-associated protein (Ce)	L26290	81.0	253
0111	Translation initiation factor nuk34 (H)	X79538	99.5	411
0115	Oligosaccharyltransferase 48 kDa subunit (Do)	M98392	95.7	443
0116	75 kDa autoantigen (H)	M58460	26.7	225
0118	Rab B (Dd)	L21012	59.0	161
0119	Chlordecone reductase (H)	S68288	99.4	323
0120	Neuronal protein NP25 (R)	M84725	69.7	195

ACNPV, *Autographa california* nuclear polyhedrosis virus; Ba, *Bacillus amyloliquifacience*; C, chicken; Ce, *Caenorhabditis elegance*; D, *Drosophila melanogaster*; Dd, *Dictyostelium discoideum*; Do, dog; H, human; Hv, *Hordeum vulgare*; M, mouse; R, rat; Sc, *Saccharomyces cerevisiae*.

[a)] Genbank/EMBL database files are shown except KIAA0090. [b)] The size of regions which show similarities. [c)] PIR database file.

which were more than 90% of the length of the corresponding transcripts were selected and subjected to sequence analysis. After the sequence determination, the integrity of the clones was analyzed by Northern hybridization,[1] and the coding sequences of 40 genes were newly predicted. Among the genes predicted, half belonged to the medium-sized cDNA class and the other half to the small-sized cDNA class, and their average sizes were roughly 4 kb and 1.4 kb, respectively.

In Fig. 1, open reading frames are indicated by solid boxes and the first ATG codon by open triangles above the solid boxes. In-frame termination codons upstream of the first ATG codon were identified in 9 clones in the medium-sized cDNA class and 6 clones in the small-sized cDNA class. It is therefore likely that at least 40% of the clones analyzed harbor the complete coding region. As can be seen in the patterns in Fig. 1, the medium-sized cDNAs retain relatively long stretches of 3′-untranslated regions (UTRs). The biological significance of these long UTRs remains to be elucidated.

Computer analysis of the sequences was carried out by using the GCG software package. The results are shown in Tables 1 and 2 and also in the figures in the Supplement section. Sequence features noted are summarized as follows:

1. Sequences of 14 genes were unrelated to any reported genes in the GenBank/EMBL database files, while the remaining 26 genes carried some sequences with similarities to known genes (Table 1).
2. Protein motifs that matched those in the PROSITE motif database were found in 11 genes (Table 2).
3. Significant transmembrane domains were identified

Table 2. cDNA clones with regions that matched motifs in the PROSITE database.

Motifs	Description	Gene number (KIAA)	References
ATP GTP A	ATP/GTP-binding site motif A (P-loop)	0083, 0089	5
NAD G3PDH	NAD-dependent glycerol-3-phosphate dehydrogenase	0089	6
SUBTILASE HIS	Serine proteases, subtilase family, active site	0091	7
SUBTILASE SER	Serine proteases, subtilase family, active site	0091	7
RECEPTOR CYTOKINES 1	Growth factor and cytokine receptors family	0091	8
C2 DOMAIN	C2 domain	0093	9
PRENYLATION	Prenyl group binding site	0096	10
TCP1-1	Chaperonins TCP-1	0098	11
TCP1-2	Chaperonins TCP-1	0098	11
TCP1-3	Chaperonins TCP-1	0098	11
MITOCH CARRIER	Mitochondrial energy transfer proteins	0106	12
CLAT ADAPTOR M 2	Clathrin adaptor complexes medium chain	0109	13
DEAD ATP HELICASE	DEAD and DEAH box families ATP-dependent helicases	0111	14
ATP A	ATP/GTP-binding site motif A	0115	6
ALDOKETO REDUCTASE 1	Aldo/keto reductase family	0119	15

in 17 genes, 10 in which harbored multiple hydrophobic regions, as judged by the methods of Engelman et al.[3] and of Kyte and Doolittle[4] (see figures in the Supplement section).

4. Repetitive sequences were identified in the 3′-UTR of 4 genes.

5. Three genes (KIAA0096, 0099 and 0118) were related to signal transducing genes on the basis of sequence similarities and characteristic protein motifs. Particularly, it was noted that the KIAA0096 gene carried sequences with similarities to the genes in the protein kinase family and harbored a possible prenylation site, which is often observed in signal transducing membrane proteins.

6. The product of the KIAA0091 gene retained a possible leader peptide, a single transmembrane domain and a motif for a cytokine receptor, suggesting that the gene belongs to the cytokine receptor family.

3.2. Expression profiles in tissues

The expression profiles of the sequenced genes were examined in 16 different human tissues and in 2 cell lines including the KG-1 cell as a positive control. The results are summarized in Table 3. On the basis of the Northern hybridization profiles, the genes can be categorized into several types. Thirty genes in total, 11 genes in the medium-sized and 19 in the small-sized cDNA class, were expressed ubiquitously in all the cells and tissues examined. This indicates that the small cDNA class is more ubiquitously expressed in various tissues. The hybridization profiles also showed that expression of the small cDNA class is relatively abundant. Among the genes assigned to the ubiquitous class, 10 genes produced a few discrete bands which varied in size depending on the tissues (KIAA0084, HeLa: 0092, testis: 0093, skeletal muscle: 0094, testis: 0095, brain: 0097, KG-1: 0111, skeletal muscle: 0114, placenta: 0116, skeletal muscle: 0118, testis). This may be due to alternative splicing or alternative initiation of transcription. Typical expression patterns of the genes in this category are shown in Fig. 2E–G. In the remaining genes, the expression of 8 genes was tissue-specific and that of 2 genes (KIAA0083 and 0087) was specific for the KG-1 cell line, although slight expression of KIAA0083 was detected in the thymus. Typical expression profiles of the genes belonging to this category are shown in Fig. 2A–E, in comparison with that of the ubiquitously expressed β-actin gene (Fig. 2H).

Including our previously reported cDNA sequences,[1,2] we have predicted the coding sequences of 120 genes in total, and found that expression of 31 genes (26%) was cell- or tissue-specific. By continuation of this type of analysis, therefore, it is possible to find many new genes with tissue-specific expression.

The chromosomal location of these genes were determined using a panel of human-rodent hybrid cell lines (see Table 3).

Acknowledgments: This project was supported by grants from the Kazusa DNA Research Institute Foundation. We thank Dr. M. Takanami for his interest in this work.

Table 3. Summary of the cDNA sequence data and the expression patterns of the cloned genes in human tissues and cell lines.

Gene number (KIAA)	Total length of cDNA (bp)[a]	Amino acid residues	Expression[b]																	Chromosomal location	Accession[c] number
			KG-1	HeLa	He	Br	Pl	Lu	Li	Sk.m	Ki	Pa	Sp	Th	Pr	Te	Ov	Sm.i	Co	Pe.b	
0081[d]	4,169	233	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	15	D42039
0082[d,e]	3,186	607	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	6	D43949
0083[d,e]	4,206	1,076	+	+	−	−	−	−	−	−	−	−	−	±	−	−	−	−	−	−	10	D42046
0084	2,952	648	+	+	++	+	+	+	+	++	+	±	+	++	+	+	+	+	−	+	18	D42043
0085[d,e]	2,169	269	+	−	+	+	+	++	+	+	+	+	+++	+++	+	+	+	++	±	+++	1	D42042
0086[e]	4,468	1,040	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	10	D42045
0087	4,283	138	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	7	D42038
0088[d,e]	3,820	943	+	++	+	+	+	++	+	+	+	+	+	+	+	+	++	+	+	+	8, 11	D42041
0089[e]	4,043	411	++	++	+++	+	+	+	−	+++	+	+	+	+	++	+	+	+	+	+	3	D42047
0090[d,e]	5,726	905	+	+	+	+	+	+	−	+	+	+	+	+	+	+	+	+	+	+	1	D42044
0091[d,e]	4,338	1,052	+	+	++	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	16	D42053
0092[d,e]	2,913	474	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	11	D42054
0093[e]	5,749	927	+	+	+	−	+	+	+	+	+	+	+	+	+	+	+	+	+	+	15	D42055
0094[e]	2,671	394	+	+	+	+	+	+	+	++	+	+	+	+	+	+	+	+	+	+	4	D42084
0095[d,e]	2,739	819	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	16	D42085
0096[e]	4,165	527	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	±	++	3	D43636
0097[d]	6,628	2,032	+	+	+	+	+	+	±	++	+	+	+	+	+	+	+	+	+	−	11	D43948
0098[d,e]	1,938	546	++	++	++	++	++	++	++	++	++	+	+	+	+	++	+	+	+++	+	5, 6	D43950
0099[d,e]	5,319	1,186	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	1	D43951
0100[d,e]	6,962	2,092	+	+	+	+	+	+	+	++	+	+	+	+	+	+	+	+	+	+	17	D43947
0101	836	111	++	++	−	−	+	±	+	−	+	±	+	+++	−	±	+	+	+++	+	15	D14657
0102[d,e]	1,370	225	++	++	++	++	++	+++	+++	++	+++	+++	+	+	++	+	++	+	++	+	1, 11	D14658
0103	1,219	299	+	+	++	++	+	+++	+	+++	+++	+	+	+	++	+++	++	+	+++	+	8	D14659
0104	1,322	291	++	+++	+++	++	+	++	+++	+++	+++	+++	+	+	+	++	+	+	++	+	2	D14660
0105	1,622	192	++	++	++	+	+	++	++	++	+++	++	+	+	+	++	+	+	++	+	6	D14661
0106[e]	1,653	238	+++	+++	+++	++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	++	+++	+++	1	D14662
0107	1,308	397	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+	++	+	+++	++	+	++	++	3	D14663
0108[d,e]	1,402	233	++	+++	++	++	+++	+++	+++	+++	+++	+++	+++	++	+++	+++	+++	+++	+++	+++	2	D14696
0109[e]	1,870	251	+	+	++	+++	+	++	+	+	++	+	+	+	+	+	+	+	+	+	3	D30757
0110	1,233	275	+	+	++	+	+	+	+	+++	+++	+	+++	+++	+++	+++	+++	+++	+++	+++	6	D14811
0111[d,e]	1,682	411	++	++	+++	+++	+++	+++	+++	+++	+++	++	++	+++	+	+	+	+	++	±	17	D21853
0112	1,696	399	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	8	D25218
0113	1,658	296	+	++	++	+	+	++	++	+++	+	+	++	+	+	+	+	+	+	++	5, 6	D30755
0114	792	79	++	++	+++	+	+	+	+	+++	+++	+	+	+	+	+	++	+	++	+	4, Y	D28589
0115[d,e]	1,668	456	++	+	+	+	+	+	+	+	+	+	+	+	+	+	++	+	+	+	1	D29643
0116[e]	1,011	290	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	3	D29958
0117	881	227	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	1	D29955
0118[e]	1,413	161	+	+	++	+	+	+	+	++	+	+	+	+	+	+	+	+	+	+	12	D42087
0119[e]	1,204	323	+	+++	++	+	±	+++	+++	++	+++	++	+	++	++	++	++	+++	++	+	10	D17793
0120[e]	1,360	199	+++	+++	++	+	+++	+++	+++	++	+++	++	+++	+++	+++	+++	+++	+++	+++	+++	1, 8	D21261

He, heart; Br, brain; Pl, placenta; Lu, lung; Li, liver; Sk.m., skeletal muscle; Ki, kidney; Pa, pancreas; Sp, spleen; Th, thymus; Pr, prostate; Te, testis; Ov, ovary; Sm.i, small intestine; Co, colon; Pe.b, peripheral blood leukocytes. [a] Values excluding poly(A) sequences. [b] Expression of mRNA in indicated cells and human tissues (Clontech, USA) was examined by Northern hybridization, and the relative strength of the positive signals are indicated (±, +, ++, +++). [c] Accession number of GSDB, DDBJ, EMBL and NCBI nucleotide sequence databases. [d] The presence of possible transmembrane domains was revealed (see Supplemental pages). [e] Similarities to known genes were identified (see Table 1 and Supplemental pages).

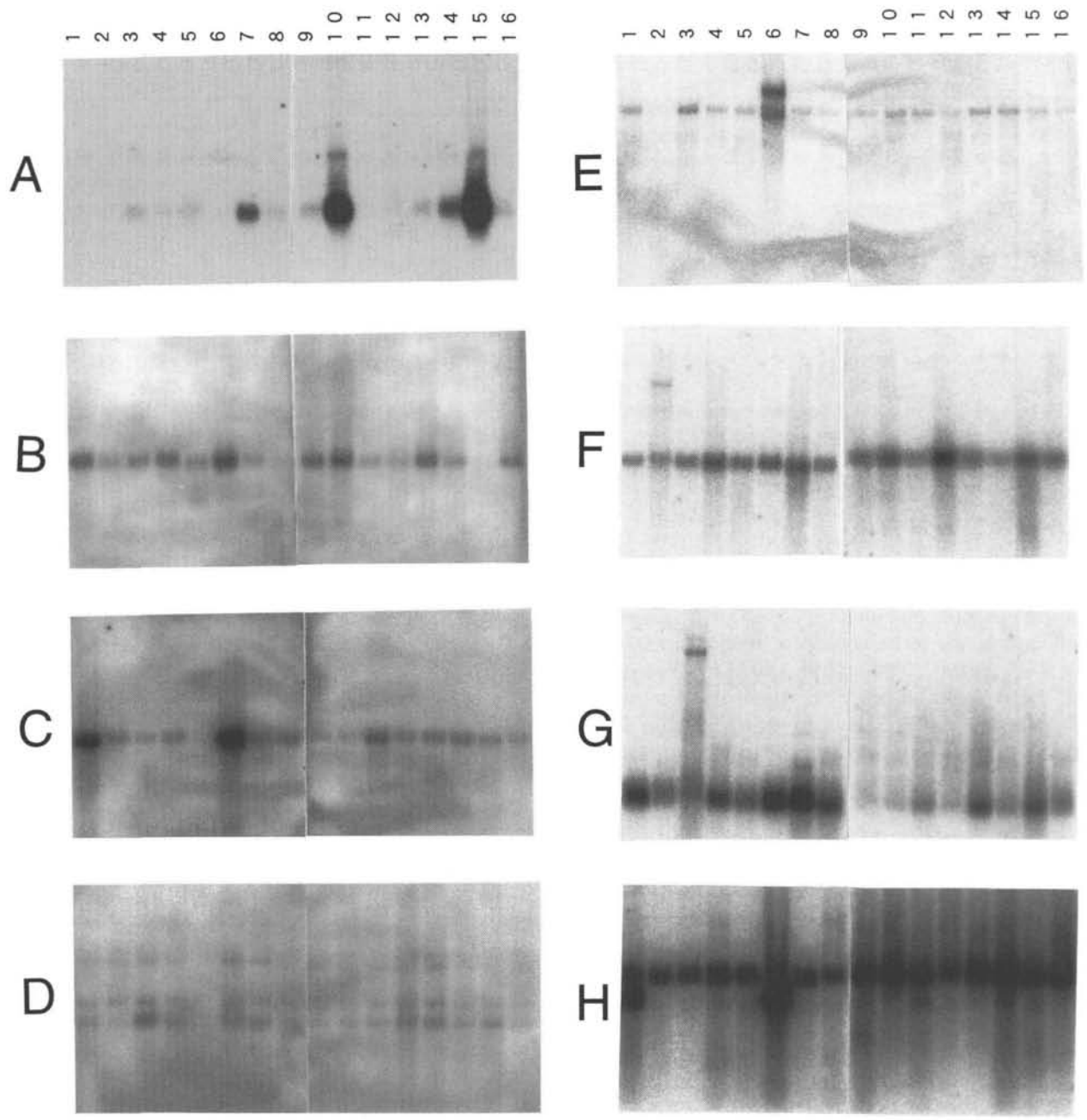

Figure 2. The expression patterns of representative clones. cDNA fragments were randomly labeled and hybridization was carried out as described previously. Human MTN blots were purchased from Clontech Laboratories. A, KIAA0101; B, 0084; C, 0089; D, 0090; E, 0093; F, 0095; G, 0114; H, β-actin gene. Lane 1, heart; 2, brain; 3, placenta; 4, lung; 5, liver; 6, skeletal muscle; 7, kidney; 8, pancreas; 9, spleen; 10, thymus; 11, prostate; 12, testis; 13, ovary; 14, small intestine; 15, colon; 16, peripheral blood leukocyte.

References

1. Nomura, N., Miyajima, N., Sazuka, T. et al. 1994, Prediction of the coding sequences of unidentified human genes. I. The coding sequences of 40 new genes (KIAA0001-KIAA0040) deduced by analysis of randomly sampled cDNA clones from human immature myeloid cell line KG-1, *DNA Res.*, **1**, 27–35.
2. Nomura, N., Nagase, T., Miyajima, N. et al. 1994, Prediction of the coding sequences of unidentified human genes. II. The coding sequences of 40 new genes (KIAA0041-KIAA0080) deduced by analysis of cDNA clones from human cell line KG-1, *DNA Res.*, **1**, 223–229.

3. Engelman, D. M., Steize, T. A., and Goldman, A. 1986, Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins, *Annu. Rev. Biophys. Biophys. Chem.*, **15**, 321–353.
4. Kyte, J. and Doolittle, R. F. 1982, A simple method for displaying the hydropathic character of a protein, *J. Mol. Biol.*, **157**, 105–132.
5. Walker, J. E., Saraste, M., Runswick, M. J., and Gay, N. J. 1982, Distantly related sequences in the α- and β-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold, *EMBO J.*, **1**, 945–951.
6. Otto, J., Argos, P., and Rossmann, M. G. 1980, Prediction of secondary structural elements in glycerol-3-phosphate dehydrogenase by comparison with other dehydrogenases, *Eur. J. Biochem.*, **109**, 325–330.
7. Barr, P. J. 1991, Mammalian subtilisins: The long-sought dibasic processing endoproteases, *Cell*, **66**, 1–3.
8. D'Andrea, A. D., Fasman, G. D., and Lodish, H. F. 1989, Erythropoietin receptor and interleukin-2 receptor β chain: A new receptor family, *Cell*, **58**, 1023–1024.
9. Perin, M. S., Fried, V. A., Mignery, G. A., Jahn, R., and Sudhol, T. C. 1990, Phospholipid binding by a synaptic vesicle protein homologous to the regulatory region of protein kinase C, *Nature*, **345**, 260–263.
10. Lowy, D. R. and Willumsen, B. M. 1989, New clue to ras lipid glue, *Nature*, **341**, 384–385.
11. Ellis, J. 1992, Cytosolic chaperonin confirmed, *Nature*, **358**, 191–192.
12. Nelson, D. R., Lawson, J. E., Klingerberg, M., and Douglas, M. G. 1993, Site-directed mutagenesis of the yeast mitochondrial ADP/ATP translocator, *J. Mol. Biol.*, **230**, 1159–1170.
13. Lee, J., Jongeward, G. D., and Sternberg, P. W. 1994, *unc*-101, a gene required for many aspects of *Canorhabditis elegans* development and behavior, encodes a clathrin-associated protein, *Genes Dev.*, **8**, 60–73.
14. Linder, P., Lasko, P., Ashburner, M. et al. 1989, Birth of the DEAD box, *Nature*, **337**, 121–122.
15. Bohren, K. M., Bullock, B., Wermuth, B., and Gabbay, K. H. 1989, The aldo-keto reductase family, *J. Biol. Chem.*, **264**, 9547–9551.
16. Nomura, N., Takahashi, M., Matsui, M. et al., 1988, Isolation of human cDNA clones of *myb*-related genes, A-*myb* and B-*myb*, *Nucleic Acids Res.*, **16**, 11075–11089.