DNA RESEARCH 5, 277–286 (1998)

Prediction of the Coding Sequences of Unidentified Human Genes. XI. The Complete Sequences of 100 New cDNA Clones from Brain Which Code for Large Proteins *in vitro*

Takahiro NAGASE, Ken-ichi ISHIKAWA, Mikita SUYAMA, Reiko KIKUNO, Nobuyuki MIYAJIMA, Ayako TANAKA, Hirokazu KOTANI, Nobuo NOMURA, and Osamu OHARA*

Kazusa DNA Research Institute, 1532-3 Yana, Kisarazu, Chiba 292-0812, Japan

(Received 1 October 1998)

Abstract

In our series of projects for accumulating sequence information on the coding sequences of unidentified human genes, we have newly determined the sequences of 100 cDNA clones from a set of size-fractionated human brain cDNA libraries, and predicted the coding sequences of the corresponding genes, named KIAA0711 to KIAA0810. These cDNA clones were selected according to their coding potentials of large proteins (50 kDa and more) *in vitro*. The average sizes of the inserts and corresponding open reading frames were 4.3 kb and 2.6 kb (869 amino acid residues), respectively. Sequence analyses against the public databases indicated that the predicted coding sequences of 78 genes were similar to those of known genes, 64% of which (50 genes) were categorized as proteins functionally related to cell signaling/communication, cell structure/motility and nucleic acid management. As additional information concerning genes characterized in this study, the chromosomal locations of the clones were determined by using human-rodent hybrid panels and the expression profiles among 10 human tissues were examined by reverse transcription-coupled polymerase chain reaction which was substantially improved by enzyme-linked immunosorbent assay.

Key words: large proteins; *in vitro* transcription/translation; cDNA sequencing; expression profile; chromosomal location; brain

1. Introduction

As a complement of human genome sequencing,[1] analysis of cDNAs is expected to provide indispensable information for the interpretation of genomic sequences. In particular, it is very advantageous that cDNA clones convey more unambiguous protein coding information than genomic clones and that they can be used as versatile reagents in functional studies of genes. Considering the importance of the cDNA analysis, we began sequencing the entire length of cDNAs in order to accumulate information on the coding sequences of unidentified human genes.[2] Recently, we have focused our sequencing efforts on the analysis of large cDNAs (> 4 kb) encoding large proteins (> 50 kDa) in brain, since these genes are likely to play an important role in mammals.[3,4] As an extension of the preceding reports, we herein present the entire sequences, expression profiles among 10 human tissues and chromosomal locations of 100 new cDNA clones. Furthermore, specific features of the newly predicted protein sequences are described on the basis of the homology/motif analysis.

2. Materials and Methods

2.1. Source and screening of cDNA clones

The size-fractionated human brain cDNA library Nos. 2 to 5 (average insert size = 3.9, 4.5, 5.3 and 6.1 kb, respectively)[3] were used as a source of cDNA clones. Most of the cDNA clones analyzed in this study were selected from library No. 2 (average insert size = 3.9 kb). cDNA clones were first screened according to their *in vitro* protein-coding potentials and then by single-pass sequencing of both termini as described previously.[3] The sequences thus determined were subjected to homology search against the GenBank database (release 102.0) excluding expressed sequence tags and genomic sequences. The clones with unidentified sequences at both ends were sequenced in their entirety as described.[3] As an exception, cDNA clones which are likely to contain much larger open reading frames (ORFs) than those already registered in the public databases were sequenced in their entirety.

Communicated by Michio Oishi

* To whom correspondence should be addressed. Tel. +81-438-52-3913, Fax. +81-438-52-3914, E-mail: ohara@kazusa.or.jp

2.2. *Gene expression profiles*

Expression patterns of newly identified genes in 10 human tissues were examined by reverse transcription-coupled polymerase chain reaction (RT-PCR) as described previously,[4] except that the detection and quantification of the PCR products were done by enzyme-linked immunosorbent assay (ELISA). For ELISA, the RT-PCR was modified to be conducted in the presence of digoxigenin (DIG)-11-dUTP (DIG PCR labeling mix; Boehringer Mannheim, Germany) while other conditions were unchanged. The nucleotide sequences of respective PCR primers are available upon request. The obtained DIG-labeled PCR products were subjected to quantification with a PCR ELISA kit from Boehringer Mannheim. Since the kit takes advantage of solution hybridization for specific detection of desired PCR products, authentic biotinylated products were prepared from the isolated cDNA clones by PCR in the presence of 0.1 mM biotin-14-dATP (Life Technologies, Inc., USA), purified on agarose gels, and then used as probes for the solution hybridization. The RT-PCR ELISA was performed exactly as described by the instructions provided with the kit, except for the following points: Because biotinylated PCR products were used as probes in place of oligomers, the hybridization was carried out in 210 μl of hybridization solution containing 20 ng of the probe at 65 °C for 16 hr; after the hybridization, biotin-labeled molecules were captured in a well of streptavidin-coated microtiter plate at 55 °C for 30 min. The color development with horseradish peroxidase and 2,2′-azino-bis(3-ethylbenzthiazoline-6-sulfonate) , the final detection step of the RT-PCR ELISA, was monitored by absorption at 405 nm in a kinetic mode with a SPECTRAmax 250 microtiter plate reader (Molecular Device, Co., Sunnyvale, CA) at 37 °C. The ELISA data were converted to the mRNA levels expressed as equivalent amounts of the cDNA plasmid on the basis of ELISA control curves using PCR products derived from serial dilutions of a known amount of the authentic plasmids with a software package, SOFTmax PRO (Molecular Device, Co.). The digitized mRNA levels were then displayed by color codes to facilitate survey of many gene expression profiles at a glance.

2.3. *Other methods*

DNA sequencing and homology search of the predicted protein-coding sequences were carried out as described previously,[3,4] except that most DNA sequencing reactions were performed using ABI PRISM™ dRhodamine terminator cycle sequencing ready reaction kit (Perkin-Elmer Co., USA). Plasmid DNAs for sequencing and *in vitro* transcription/translation were prepared by the Wizard *Plus* SV Minipreps DNA Purification system (Promega Corp., Madison, WI), except that the spin columns were replaced with MultiScreen-FB plates (Millipore Corp., Bedford, MA) for adapting the system in the 96-well format. When the possibility of spurious interruption of ORFs was noticed, a likely region which causes the interruption was amplified by RT-PCR and then examined by DNA sequencing as described previously.[5] Chromosomal locations of newly identified genes were determined using human-rodent hybrid panels, GeneBridge 4 (Research Genetics Inc., USA)[6] if their mapping data were not available in the UniGene database (http://www.ncbi.nlm.nih.gov/UniGene/index.html). For genes whose chromosomal locations were described in the UniGene database, we did not perform the radiation hybrid mapping experiments as a general rule. The actual primer sequences and the reaction conditions used for PCR-assisted chromosomal mapping are accessible through the World Wide Web at http://www.kazusa.or.jp.

3. Results and Discussion

3.1. *Sequence analysis and prediction of protein-coding regions of cDNA clones*

The criteria of cDNA clone selection were the same as reported in the previous studies; they must be uncharacterized and can direct synthesis of proteins larger than 50 kDa *in vitro*. One hundred clones thus selected were subjected to sequencing of entire inserts. Most of the cDNA clones (87 clones) were derived from library No. 2 (average insert size of 3.9 kb) in this study, since cDNA clones in this library had not been extensively characterized yet. As described previously,[5] some clones were found to carry spurious coding interruption(s): Four clones (KIAA0799-0801, and KIAA0810) were found to carry relatively long insertions, probably corresponding to intronic sequences; the ORFs in 5 clones (KIAA0803, KIAA0804, KIAA0807-0809) were frame-shifted by insertion or deletion of a small number of nucleotide residues; the ORFs in 3 clones (KIAA0802, KIAA0805 and KIAA0806) were interrupted by relatively short insertions (152, 94 and 188 bp, respectively). For those genes, the revised sequences by the RT-PCR experiments, not the actual cloned cDNA sequences, were deposited to GenBank/EMBL/DDBJ databases and used for prediction of protein-coding sequences. The sequence data revealed that the average sizes of these cDNA inserts and of their ORFs were 4.3 kb and 2.6 kb (corresponding to 869 amino acid residues), respectively. Physical maps of the 100 cDNA clones analyzed are shown in Fig. 1, where the ORFs and the first ATG codons in respective ORFs are indicated by solid boxes and triangles, respectively. The in-frame termination codons upstream of the first ATG codon were identified in 41 clones, among which 33 clones carried the ATG codon within the context of Kozak's rule.[7] In Fig. 1, short interspersed nu-

Figure 1. Physical maps of cDNA clones analyzed. The physical maps shown here were constructed on the basis of the sequence data of respective cDNA clones. The horizontal scale represents the cDNA length in kb, and the gene numbers corresponding to respective cDNAs are given on the left. The ORFs and untranslated regions are shown by solid and open boxes, respectively. The positions of the first ATG codons with or without the contexts of the Kozak's rule are indicated by solid and open triangles, respectively. RepeatMasker, which is a program that screens DNA sequences for interspersed repeats known to exist in mammalian genomes, was applied to detect repeat sequences in respective cDNA sequences (Smit, A. F. A. and Green, P., RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html). Short interspersed nucleotide elements (SINEs) including Alu and MIRs sequences and other repetitive sequences thus detected are represented by dotted and hatched boxes, respectively.

Table 1. Information of sequence data and chromosomal locations of the identified genes.

Gene number (KIAA)	Accession number[a)]	cDNA length (bp)[b)]	ORF length (amino acid residues)	Chromosomal location[c)]	Gene number (KIAA)	Accession number[a)]	cDNA length (bp)[b)]	ORF length (amino acid residues)	Chromosomal location[c)]
0711	AB018254	6,706	623	8	0761	AB018304	4,218	392	1
0712	AB018255	6,476	1,587	11	0762	AB018305	3,999	624	11
0713	AB018256	6,817	945	7	0763	AB018306	4,148	841	3
0714	AB018257	5,655	1,112	21	0764	AB018307	4,261	414	2*
0715	AB018258	5,029	914	5	0765	AB018308	4,168	594	20*
0716	AB018259	4,652	765	7	0766	AB018309	4,185	607	3
0717	AB018260	4,610	627	8	0767	AB018310	4,258	573	22*
0718	AB018261	3,742	742	7	0768	AB018311	4,150	872	4*
0719	AB018262	4,017	608	3	0769	AB018312	4,341	684	11*
0720	AB018263	4,066	864	1	0770	AB018313	4,239	731	15*
0721	AB018264	4,098	433	6	0771	AB018314	4,250	948	14*
0722	AB018265	4,120	784	12	0772	AB018315	3,879	468	9
0723	AB018266	3,803	847	3*	0773	AB018316	4,164	300	7
0724	AB018267	3,893	963	1	0774	AB018317	4,021	1,163	13*
0725	AB018268	3,911	573	8	0775	AB018318	4,121	615	17*
0726	AB018269	4,300	968	12	0776	AB018319	4,174	799	6
0727	AB018270	4,174	674	17	0777	AB018320	4,118	1,100	4*
0728	AB018271	3,964	1,065	6	0778	AB018321	4,184	640	1*
0729	AB018272	4,143	1,196	2*	0779	AB018322	3,743	320	3*
0730	AB018273	4,318	1,004	13	0780	AB018323	4,182	1,100	9
0731	AB018274	4,095	1,096	10	0781	AB018324	4,171	637	11*
0732	AB018275	4,193	849	17*	0782	AB018325	4,130	1,179	11
0733	AB018276	3,479	528	6*	0783	AB018326	4,231	888	7*
0734	AB018277	4,518	1,186	16	0784	AB018327	4,282	1,073	20*
0735	AB018278	4,948	683	15	0785	AB018328	4,485	694	2
0736	AB018279	4,353	742	1	0786	AB018329	4,345	1,021	1*
0737	AB018280	4,174	621	14	0787	AB018330	4,427	550	12
0738	AB018281	4,076	921	7	0788	AB018331	4,257	1,324	17*
0739	AB018282	4,079	1,130	12	0789	AB018332	4,480	369	12
0740	AB018283	4,239	532	10*	0790	AB018333	4,469	1,319	6
0741	AB018284	4,177	1,220	2*	0791	AB018334	4,200	1,332	5
0742	AB018285	4,309	1,236	2	0792	AB018335	4,074	807	1
0743	AB018286	4,149	1,061	14*	0793	AB018336	3,997	1,054	2*
0744	AB018287	4,238	590	7	0794	AB018337	4,656	490	3*
0745	AB018288	4,228	909	8	0795	AB018338	4,273	465	3*
0746	AB018289	4,086	1,029	4*	0796	AB018339	3,900	1,080	6*
0747	AB018290	4,026	1,072	12	0797	AB018340	4,128	1,084	6*
0748	AB018291	4,183	383	12*	0798	AB018341	4,517	652	19
0749	AB018292	3,628	678	12	0799[d)]	AB018342	5,613	712	5*
0750	AB018293	3,885	1,124	11	0800[d)]	AB018343	5,984	1,507	3
0751	AB018294	3,854	1,188	8*	0801[d)]	AB018344	3,740	1,032	5
0752	AB018295	4,332	334	8	0802[d)]	AB018345	5,252	1,353	18
0753	AB018296	4,424	967	17*	0803[d)]	AB018346	5,639	1,708	7
0754	AB018297	4,044	701	1	0804[d)]	AB018347	4,216	1,211	3
0755	AB018298	3,988	1,032	4	0805[d)]	AB018348	4,204	1,327	14
0756	AB018299	4,226	836	1	0806[d)]	AB018349	4,015	1,065	1*
0757	AB018300	3,970	520	20*	0807[d)]	AB018350	4,049	1,265	1
0758	AB018301	4,353	986	6	0808[d)]	AB018351	4,133	526	8*
0759	AB018302	4,006	673	14	0809[d)]	AB018352	3,901	1,241	3
0760	AB018303	4,194	1,206	16	0810[d)]	AB018353	4,047	824	7

a) Accession numbers of DDBJ, EMBL and GenBank databases.
b) Values excluding poly(A) sequences.
c) Chromosome numbers identified by using GeneBridge 4 radiation hybrid panel unless specified. The chromosomal locations highlighted by asterisks were fetched from the UniGene database.
d) cDNA and ORF lengths were revised by direct analysis of the RT-PCR products.

Table 2. Functional classifications of the gene products based on homologies to known proteins and sequence motifs.

Functional category[a]	Gene number (KIAA)	Similarity class[b]	Homologous entry in the database[c]	Accession no.[d]	Identities (%)[e]	Overlap (amino acid residues)[f]
Cell signaling/communication	0715	R	probable calcium-transporting ATPase 3 (Sc)	P39524	31.4	475
	0716	R	KIAA0299 (H)	AB002297	58.6	563
	0717	W	RhoA (H)	P06749	44.2	113
	0718	H	diacylglycerol kinase, β (R)	P49621	95.8	742
	0720	W	Lsc oncogene (M)	U58203	31.3	300
	0722	W	unc-51-like kinase ULK1 (M)	AF053756	86.8	788
	0735	H	synaptic vesicle protein 2 form B (R)	A47382	94.9	683
	0736	H	transport protein-like protein p87 (R)	S27263	99.1	742
	0739	R	sodium bicarbonate cotransporter HNBC1 (H)	AF007216	56.5	953
	0740	W	ras-related C3 boturinum toxin substrate 1 rac1 (H)	P15154	38.7	199
	0741	I	putative GTP-binding protein (H)	AJ006412	100.0	1221
	0743	H	neurexin III-α membrane-bound type 2 precursor (R)	A48218	99.3	833
	0746	W	sel-1 protein (Ce)	S68303	26.4	515
	0747	R	S/T kinase receptor type1 fhh (Fr)	AF056116	46.4	1103
	0751	R	rab3 effector RIM (R)	AF007836	58.1	573
	0754	W	megakaryocyte stimulating factor (H)	U70136	26.5	1411
	0756	H	neurofascin (R)	U81035	92.1	861
	0758	W	latrophilin-related protein 1 precursor (R)	U78105	22.8	448
	0762	H	F-spondin precursor (R)	P35446	96.8	624
	0763	R	KIAA0522 (H)	AB011094	62.4	867
	0768	R	latrophilin-related protein 1 precursor (R)	U78105	58.4	637
	0769	W	SH2/SH3 adaptor protein Nck (X)	U85781	28.0	164
	0771	R	Bcl2, p53 binding protein Bbp/53BP2 (H)	U58334	37.7	951
	0777	R	SH3-containing protein p4015 (R)	AF026505	87.9	1103
	0778	I	Na^+/K^+-transporting ATPase α-2 chain (H)	P50993	100.0	640
	0782	R	KIAA0580 (H)	AB011152	40.6	955
	0786	R	latrophilin-related protein 1 precursor (R)	U78105	62.6	1037
	0787	H	Ca^{2+}/calmodulin-dependent protein kinase kinase β chain (R)	JC5669	92.9	532
	0793	R	CDEP (H)	AB008430	54.2	1039
	0803	R	A-kinase anchoring protein AKAP120 (Oc)	U26360	76.0	871
	0806	R	glial cell membrane glycoprotein LIG-1 precursor (M)	A58532	52.9	910
	0807	R	S/T protein kinase MAST205 (M)	A54602	86.0	1254
Cell structure/motility	0711	W	ring canal protein (D)	Q04652	28.6	217
	0719	W	mitochondrial precursor protein import receptor (Nc)	P23231	28.4	559
	0725	W	cochlear (M)	Z78156	72.2	108
	0727	H	myosin I myr 4 (R)	A53933	98.1	674
	0728	W	spectrin α chain (R)	P16086	22.7	661
	0750	W	α-actinin 2, skeletal muscle isoform (H)	P35609	30.4	181
	0770	W	vacuolar assembly protein VPS39 (Sc)	Q07468	22.3	443
	0791	H	nuclear pore complex protein NUP155 (R)	P37199	94.1	1332
	0795	R	ring canal protein (D)	Q04652	37.5	461
	0796	W	utrophin (H)	P46939	20.3	340
	0799	H	myosin X (B)	U55042	94.5	637
Nucleic acid management	0742	H	probable finger protein (R)	S28499	90.9	1209
	0760	R	Olf-1/EBF associated Zn finger protein Roaz (R)	U92564	82.7	1160
	0780	R	putative 90.2 kD zinc finger protein (Sc)	P39956	49.3	211
	0788	R	putative RNA helicase (H)	AJ223948	45.8	1038
	0798	R	zinc finger protein 84 (H)	P51523	61.2	469
	0801	H	helicase HEL117 (R)	A57514	98.7	1032
	0809	R	helicase II RAD54L (H)	U09820	43.2	213
Cell division	0810	W	SAD1 (Sp)	Q09825	31.9	191
Metabolism	0772	R	oxysterol-binding protein (Rb)	P16258	38.5	327
Unclassified	0714	W	probable membrane protein YMR247c (Sc)	S56061	32.4	296
	0721	R	testis-specific protein TSPY (H)	U58096	39.5	228
	0723	I	matrin 3 (H)	P43243	99.0	417
	0724	W	hypothetical 111.5 kD protein C22G7.02 (Sp)	Q09796	18.9	741
	0726	W	Caenorhabditis elegans cosmid B0034 (Ce)	U23528	29.0	462
	0729	W	Murr2 (M)	D85434	65.7	201
	0731	R	Caenorhabditis elegans cosmid R144 (Ce)	U23515	36.9	366
	0734	I	clone 23595 (H)	AF038191	100.0	303
	0737	R	CAGF9 (H)	U80736	35.9	329
	0738	W	125K surface antigen M17 precursor (Eh)	JH0284	21.8	326
	0744	R	KIAA0288 (H)	AB006626	47.7	491
	0745	R	Caenorhabditis elegans cosmid C35A5 (Ce)	Z71185	39.7	655
	0749	R	dendrin (R)	P50617	74.6	677
	0755	R	KIAA0079 (H)	P53992	55.4	985
	0759	R	angel (D)	X85743	36.7	237
	0761	R	unknown transmembrane protein (X)	X92871	51.1	231
	0776	W	Caenorhabditis elegans cosmid C06G3 (Ce)	U61947	22.4	508
	0779	W	TEX28 (H)	U93720	35.0	203
	0783	W	maf10 (M)	AF010135	32.9	243
	0789	R	KIAA0523 (H)	AB011095	46.3	216
	0794	W	S.pombe chromosome I cosmid c2C4 (Sp)	Z99259	27.2	375
	0797	W	S.pombe chromosome I cosmid c17A5 (Sp)	Z98849	40.7	135

Table 2. Continued.

	0800	R	Caenorhabditis elegans cosmid K01H12 (Ce)	Z68218	31.2	911
	0804	W	Caenorhabditis elegans cosmid C42C1 (Ce)	AF043695	22.3	1107
	0805	R	pecanex protein (D)	P18490	56.1	319
	0808	R	CAGF9 (H)	U80736	47.1	238
No homology	0712		none			
	0713		none			
	0730		none			
	0732		none			
	0733		none			
	0748		none			
	0752		none			
	0753		none			
	0757		none			
	0764		none			
	0765		none			
	0766		none			
	0767		none			
	0773		none			
	0774		none			
	0775		none			
	0781		none			
	0784		none			
	0785		none			
	0790		none			
	0792		none			
	0802		none			

a) Classifications based on the annotations of their homologous protein entries in the databases.
b) The gene products were grouped into four similarity classes according to the sequence identities obtained by the GAP program: I, identical to known human gene products (sequence identity, > 90%); H, homologous to known non-human gene products (sequence identity, > 90%); R, related to some known gene products (sequence identity, 30 to 90%); W, very weakly related to known gene products (sequence identity, < 30%).
c) Organisms in which these entries were identified are given in parentheses: B, bovine; Ce, *Caenorhabditis elegans*; D, *Drosophila melanogaster*; Eh, *Entamoeba histolytica*; Fr, *Fugu rubripes*; H, human; M, mouse; Nc, *Neurospora crassa*; Oc, *Oryctolagus cuniculus*; R, rat; Rb, rabbit; Sc, *Saccharomyces cerevisiae*; Sp, *Schizosaccharomyces pombe*; X, *Xenopus laevis*.
d) Accession numbers of homologous entries in DDBJ/EMBL/GenBank/OWL/SWISS-PROT/PIR database are shown.
e) The values were obtained by the FASTA program.

cleotide elements (Alu and MIRs sequences) and other repetitive sequences detected by using RepeatMasker program are also displayed by dotted and hatched boxes, respectively. Table 1 lists the gene codes (KIAA numbers), the accession numbers of the nucleotide sequences in GenBank/EMBL/DDBJ databases, the sizes of the cDNA inserts and the identified ORF, and the chromosomal locations of the respective genes. The chromosomal locations of 37 genes, which are highlighted by asterisks, were fetched from the UniGene database while the remaining 63 chromosomal locations were experimentally determined in this study.

3.2. Functional classification of predicted gene products

By homology and motif searches against DNA, protein, and protein-motif databases [GenBank (release 108.0), OWL (release 30.3), and PROSITE (release 15.0) databases] using Wisconsin Sequence Analysis Package™ (version 8; Genetics Computer Group, Inc. USA), the predicted coding sequences of 78 genes were found to exhibit significant similarities to those of known genes, and 64% of them were classified into protein groups functionally related to cell signaling/communication, cell structure/motility and nucleic acid management. The results of the functional classification of these newly identified genes on the basis of this homology/motif analysis are summarized in Table 2.

Interesting features to be noted are summarized below.

1. Except for the C2H2-type zinc finger protein family, 21 newly identified genes constitute 18 independent paralogous groups together with the genes characterized in our cDNA project (Table 3). In this case, genes which exhibit significant similarities throughout the protein-coding sequences, not in distinct domains or motifs, have been assigned as those with a paralogous relationship. Among these paralogous groups, genes in 7 groups were "uncharacterized."

2. Several gene products exhibited similarities to synaptic proteins. Three of them (products of KIAA0735, KIAA0736 and KIAA0743) were human counterparts of rat synaptic vesicle protein 2B,[18] transport protein-like protein p87 (synaptic vesicle protein 2)[19] and Neurexin III,[11] respectively. In addition, KIAA0768 and KIAA0786 may play an important role in neurosecretion because their gene products have overall similarities to rat latrophilin-related protein 1 (LPH1),[14]

Table 3. The newly identified genes in paralogous relationship with genes characterized by our cDNA project.

New gene	Paralogous gene	Accession no.[a)]	Identities (%)[b)]	Corresponding gene product
KIAA0715	KIAA0566	AB011138	49.8	calcium-transporting ATPase
KIAA0716	KIAA0209	D86964	37.3	DOCK180 protein[8]
	KIAA0299	AB002297	55.8	DOCK180 protein
KIAA0717	KIAA0740[c)]	AB018283	66.6	uncharacterized
KIAA0722	KIAA0623	AB014523	45.6	protein kinase ULK1[9]
KIAA0728	KIAA0465	AB007934	71.5	spectrin α[10]
KIAA0737	KIAA0808[c)]	AB018351	43.2	uncharacterized
KIAA0743	KIAA0578	AB011150	77.2	neurexin III[11]
KIAA0744	KIAA0288	AB006626	52.8	uncharacterized
KIAA0751	KIAA0340	AB002338	60.0	rab3 effector RIM[12]
KIAA0755	KIAA0079	D38555	55.9	uncharacterized
KIAA0756	KIAA0343	AB002341	48.6	neurofascin[13]
KIAA0763	KIAA0522	AB011094	64.6	uncharacterized
KIAA0768	KIAA0786[c)]	AB018329	61.2	latrophilin-related protein 1[14]
KIAA0782	KIAA0580	AB011152	42.9	uncharacterized
KIAA0795	KIAA0132	D50922	35.7	ring canal protein[15]
KIAA0805	KIAA0435	AB007895	47.4	pecanex protein[16]
KIAA0807	KIAA0303	AB002301	58.5	serine/threonine protein kinase MAST205[17]
	KIAA0561	AB011133	60.7	serine/threonine protein kinase MAST205
KIAA0810	KIAA0668	AB014568	38.0	uncharacterized

a) Accession numbers of paralogous genes in DDBJ/EMBL/GenBank database are shown.
b) The values of the overall identities of amino acid residues were obtained by the GAP program.
c) These genes are reported in this paper.

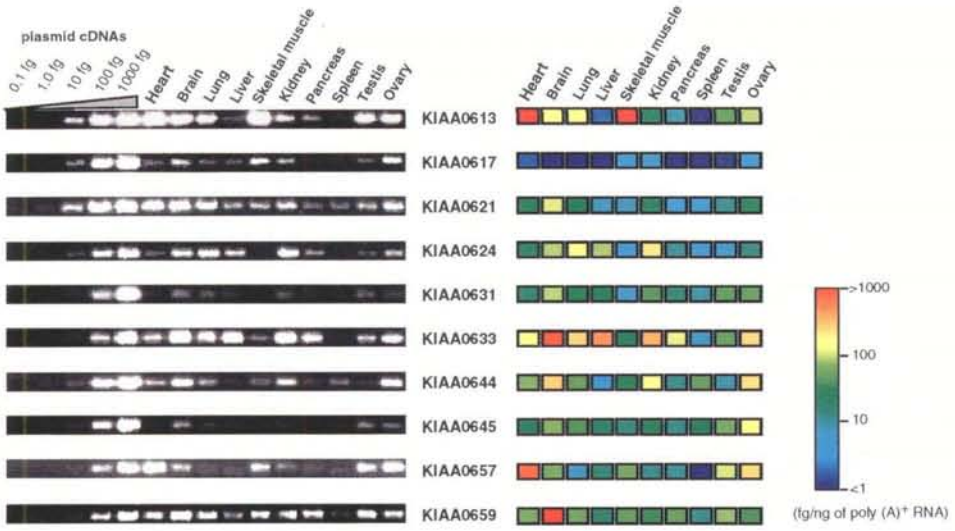

Figure 2. Comparison of RT-PCR ELISA method with conventional RT-PCR method coupled with gel electrophoresis. RT-PCR products of 10 previously reported genes[4] were analyzed by gel electrophoresis and RT-PCR ELISA. Ten microliters of RT-PCR products were analyzed on 2.5% NuSeive GTG agarose gels (FMC BioProducts, USA) and stained with ethidium bromide as described previously.[4] In each of the gel images, the first 5 lanes were used for external control reaction products, which allows us to estimate the PCR amplification efficiency. The RT-PCR ELISA data were expressed as amounts (fg) of the corresponding cDNA plasmids in 1 ng of starting poly(A)$^+$ RNAs by color codes using the conversion panel shown in the left side of this figure. RT-PCR products independently prepared were used for gel electrophoresis and ELISA. KIAA gene names are given between the gel images and the ELISA color codes.

which is a member of the secretin family of G protein-coupled receptors. Besides these genes and KIAA0686 previously reported, KIAA0758 gene product also exhibited weak similarity to rat LPH1 in the seven hydrophobic transmembrane segments conserved in this family. KIAA0751 gene product exhibited structural similarity in the overall region of rat RIM, which is a putative Rab3 effector in regulating synaptic-vesicle fusion.[12] In accord with their possible neuron-specific functions, KIAA0735, KIAA0736, KIAA0743, KIAA0751 and KIAA0768 genes expressed more abundant in brain

Figure 3. Expression profiles of 100 newly identified genes in 10 different tissues examined by RT-PCR ELISA. The tissue expression levels of 100 human genes newly identified in this study were analyzed by the RT-PCR ELISA. Gene names are given as KIAA numbers at the left side of each set of color codes. Tissue names are indicated on above the top sets of color codes.

than other human tissues as revealed by RT-PCR ELISA experiments described below.

3. Protein motif search against the PROSITE database revealed that protein kinase signatures are present in two genes (KIAA0787 and KIAA0807) and that two genes (KIAA0760 and KIAA0798) contained multiple C2H2-type zinc finger domains in the deduced coding regions. The Dbl homology domain[20] is present in the KIAA0720-encoded protein in addition to 10 such genes previously reported (KIAA0006, KIAA0142, KIAA0294, KIAA0337, KIAA0362, KIAA0380, KIAA0382, KIAA0424, KIAA0521 and KIAA0651).

3.3. Expression profiles of predicted genes

In this series of human cDNA analyses, we have used RT-PCR for exploring mRNA expression patterns of newly identified genes. Although this method allows us to detect even a trace amount of mRNA with minimum consumption of poly(A)$^+$ RNA, an obvious drawback of the method is the low quantity produced particularly when the signals are analyzed by gel electrophoresis followed by staining with a fluorescent dye. To address this problem, we newly introduced an ELISA-based procedure for quantification of specific PCR products in place of the gel electrophoresis-based one. Figure 2 compares the ELISA outputs of RT-PCR analyses of 10 previously identified genes with the gel images we had reported before.[4] Although both procedures gave essentially the same expression patterns for these 10 genes in a rough sense, it was evident that quantitative characteristics of tissue expression patterns were more clearly seen by the RT-PCR ELISA. Therefore, we decided to obtain the expression patterns of genes reported in this study by the RT-PCR ELISA method. Although the introduction of the ELISA-assisted procedure in the detection of PCR products made the quantitative characteristics of the data more evident, the expression profiles thus obtained still may possess the artifacts present in conventional RT-PCR based method. Since we could monitor the amounts of only a relatively short cDNA region flanked by PCR oligomers, alterations in transcript structure such as alternative splicing or alternative poly(A) addition could not be discriminated form changes in transcription levels. In addition, since the measurements were not multiplicated due to limitations of cost and labor, run-to-run variation could not completely be excluded. Taking these possibilities into account, we confined these expression profiles to use as important clues for the search of biologically interesting genes.

Figure 3 shows the expression patterns of the 100 newly identified genes reported in this study. By using color codes instead of numerals for displaying the expression levels, the screening of genes according to their tissue specificity was greatly facilitated. Since the expression levels are given as equivalent weights of the corresponding plasmid cDNAs, it is possible, at least in principle, to roughly compare the expression levels of a gene among 10 different tissues as well as those of genes within a particular tissue. Since these genes were identified in a human brain cDNA library, it is not surprising that dark blue color codes were hardly seen in the column of the brain. In contrast, the pancreas and the spleen columns contained many dark blue color codes, which suggested that a number of the genes actively expressed in the brain were dormant in these tissues. These expression profiles offer another line of information required for discovering biologically important genes characterized in this project.

Acknowledgments: This project was supported by grants from the Kazusa DNA Research Institute. We thank Tomomi Tajino, Keishi Ozawa, Tomomi Kato, Kazuhiro Sato, Akiko Ukigai, Emiko Suzuki, Kazuko Yamada, Kiyoe Sumi, Takashi Watanabe, Naoko Suzuki, Kozue Kaneko, Naoko Shibano and Taneaki Tsugane for their technical assistance.

References

1. Rowen, L., Mahairas, G., and Hood, L. 1997, Sequencing the human genome, *Science*, **278**, 605–607.
2. 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.
3. Ohara, O., Nagase, T., Ishikawa, K.-I. et al. 1997, Construction and characterization of human brain cDNA libraries suitable for analysis of cDNA clones encoding relatively large proteins, *DNA Res.*, **4**, 53–59.
4. Ishikawa, K.-I., Nagase, T., Suyama, M. et al. 1998, Prediction of the coding sequences of unidentified human genes. X. The complete sequences of 100 new cDNA clones from brain which can code for large proteins *in vitro*, *DNA Res.*, **5**, 169–176.
5. Ishikawa, K.-I., Nagase, T., Nakajima, D. et al. 1997, Prediction of the coding sequences of unidentified human genes. VIII. 78 new cDNA clones from brain which code for large proteins *in vitro*, *DNA Res.*, **4**, 307–313.
6. Gyapay, G., Schmitt, K., Fizames, C. et al. 1996, A radiation hybrid map of the human genome, *Hum. Mol. Genet.*, **5**, 339–346.
7. Kozak, M. 1996, Interpreting cDNA sequences: some insights from studies on translation, *Mammalian Genome*, **7**, 563–574.
8. Hasegawa, H., Kiyokawa, E., Tanaka, S. et al. 1996, DOCK180, a major CRK-binding protein, alters cell morphology upon translocation to the cell membrane, *Mol. Cell. Biol.*, **16**, 1770–1776.
9. Yan, J., Kuroyanagi, H., Kuroiwa, A. et al. 1998, Identification of mouse ULK1, a novel protein kinase structurally related to C. elegans UNC-51, *Biochem. Biophys. Res. Commun.*, **246**, 222–227.
10. Hong, W. J. and Doyle, D. 1989, Cloning and analysis

of cDNA clones for rat kidney alpha-spectrin, *J. Biol. Chem.*, **264**, 12758–12764.
11. Ushkaryov, Y. A. and Sudhof, T. C. 1993, Neurexin III alpha: extensive alternative splicing generates membrane-bound and soluble forms, *Proc. Natl. Acad. Sci. U.S.A.*, **90**, 6410–6414.
12. Wang, Y., Okamoto, M., Schmitz, F., Hofmann, K., and Sudhof, T. C. 1997, Rim is a putative Rab3 effector in regulating synaptic-vesicle fusion, *Nature*, **388**, 593–598.
13. Davis, J. Q., Lambert, S., and Bennett, V. 1996, Molecular composition of the node of Ranvier: identification of ankyrin-binding cell adhesion molecules neurofascin (mucin+/third FNIII domain-) and NrCAM at nodal axon segments, *J. Cell Biol.*, **135**, 1355–1367.
14. Lelianova, V. G., Davletov, B. A., Sterling, A. et al. 1997, α-Latrotoxin receptor, latrophilin, is a novel member of the secretin family of G protein-coupled receptors, *J. Biol. Chem.*, **272**, 21504–21508.
15. Xue, F. and Cooley, L. 1993, kelch encodes a component of intercellular bridges in *Drosophila* egg chambers, *Cell*, **72**, 681–693.
16. LaBonne, S. G., Sunitha, I. and Mahowald , A. P. 1989, Molecular genetics of pecanex, a maternal-effect neurogenic locus of *Drosophila melanogaster* that potentially encodes a large transmembrane protein, *Dev. Biol.*, **136**, 1–16.
17. Walden, P. D. and Cowan, N. J. 1993, A novel 205-kilodalton testis-specific serine/threonine protein kinase associated with microtubules of the spermatid manchette, *Mol. Cell. Biol.*, **13**, 7625–7635.
18. Bajjalieh, S. M., Peterson, K., Linial, M., and Scheller, R. H. 1993, Brain contains two forms of synaptic vesicle protein 2, *Proc. Natl. Acad. Sci. U.S.A.*, **90**, 2150–2154.
19. Gingrich, J. A., Andersen, P. H., Tiberi, M. et al. 1992, Identification, characterization, and molecular cloning of a novel transporter-like protein localized to the central nervous system, *FEBS Lett.*, **312**, 115–122.
20. Chan, A. M.-L., McGovern, E. S., Catalano, G., Fleming, T. P., and Miki, T. 1994, Expression cDNA cloning of a novel oncogene with sequence similarity to regulators of small GTP-binding proteins, *Oncogene*, **9**, 1057–1063.