DNA RESEARCH 5, 169–176 (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*

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

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

(Received 26 May 1998)

Abstract

As an extension of our cDNA analysis for deducing 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 KIAA0611 to KIAA0710. *In vitro* transcription-coupled translation assay was applied as the first screening to select cDNA clones which produce proteins with apparent molecular mass of 50 kDa and over. One hundred unidentified cDNA clones thus selected were then subjected to sequencing of entire inserts. The average size of the inserts and corresponding open reading frames was 4.9 kb and 2.8 kb (922 amino acid residues), respectively. Computer search of the sequences against the public databases indicated that predicted coding sequences of 87 genes were similar to those of known genes, 62% of which (54 genes) were categorized as proteins related to cell signaling/communication, cell structure/motility and nucleic acid management. The expression profiles in 10 human tissues of all the clones characterized in this study were examined by reverse transcription-coupled polymerase chain reaction and the chromosomal locations of the clones were determined by using human-rodent hybrid panels.

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

1. Introduction

Although the human genome is estimated to contain 50,000 to 100,000 genes, the sequences of only 10,000 genes are currently deposited in public databases. Sequencing of novel cDNAs is critical not only for identifying transcription units along the genome but also for predicting primary structures of gene products. So far, we have characterized the sequences of more than 600 human cDNA clones for a total of 3.1 Mb. In order to accumulate information on the coding sequences of unidentified human genes, a project of sequencing the entire length of cDNAs derived from relatively long transcripts has been initiated.[1] In particular, we have recently focused on analysis of genes encoding large proteins which are likely to play an important role in mammals.[2,3] As an extension of the preceding reports, we herein present the sequences of 100 new cDNA clones with the potential to code for relatively large proteins *in vitro* and discuss their biological roles predicted from the sequence data.

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. Materials and Methods

2.1. The source and screening of cDNA clones

cDNA clones were randomly selected from size-fractionated human brain cDNA libraries No. 2 to 5 (average insert size=3.9, 4.5, 5.3 and 6.1 kb) which we had constructed previously.[2] Almost half 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 both of the termini were sequenced as described previously.[4] The sequences thus determined were subjected to homology search against GenBank database (release 100.0) excluding expressed sequence tags and genomic sequences. The clones sorted to carry unidentified sequences at both ends were sequenced in their entirety as described.[2]

2.2. Other methods

DNA sequencing, homology search of the predicted protein-coding sequences, expression analysis of the sequenced cDNA by reverse transcription-coupled polymerase chain reaction (RT-PCR) and chromosomal

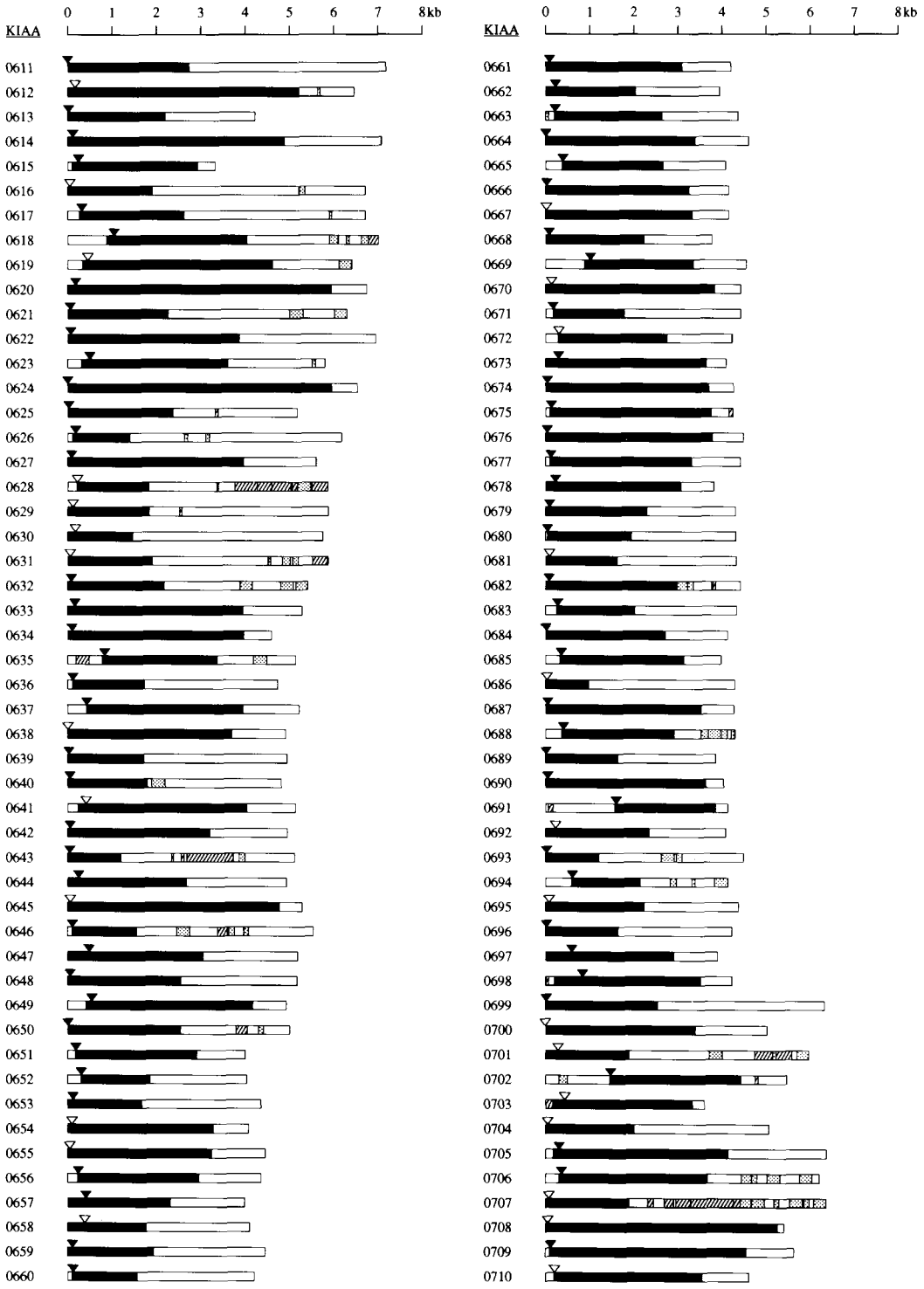

Figure 1. Physical maps of cDNA clones analyzed. The physical maps shown here were constructed on the basis of the sequence data of respective cDNA clones. The horizontal scale represents the cDNA length in kb, and the gene numbers corresponding to respective cDNAs are given on the left. The ORFs and untranslated regions are shown by solid and open boxes, respectively. The positions of the first ATG codons, with or without the contexts of 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 displayed by dotted and hatched boxes, respectively.

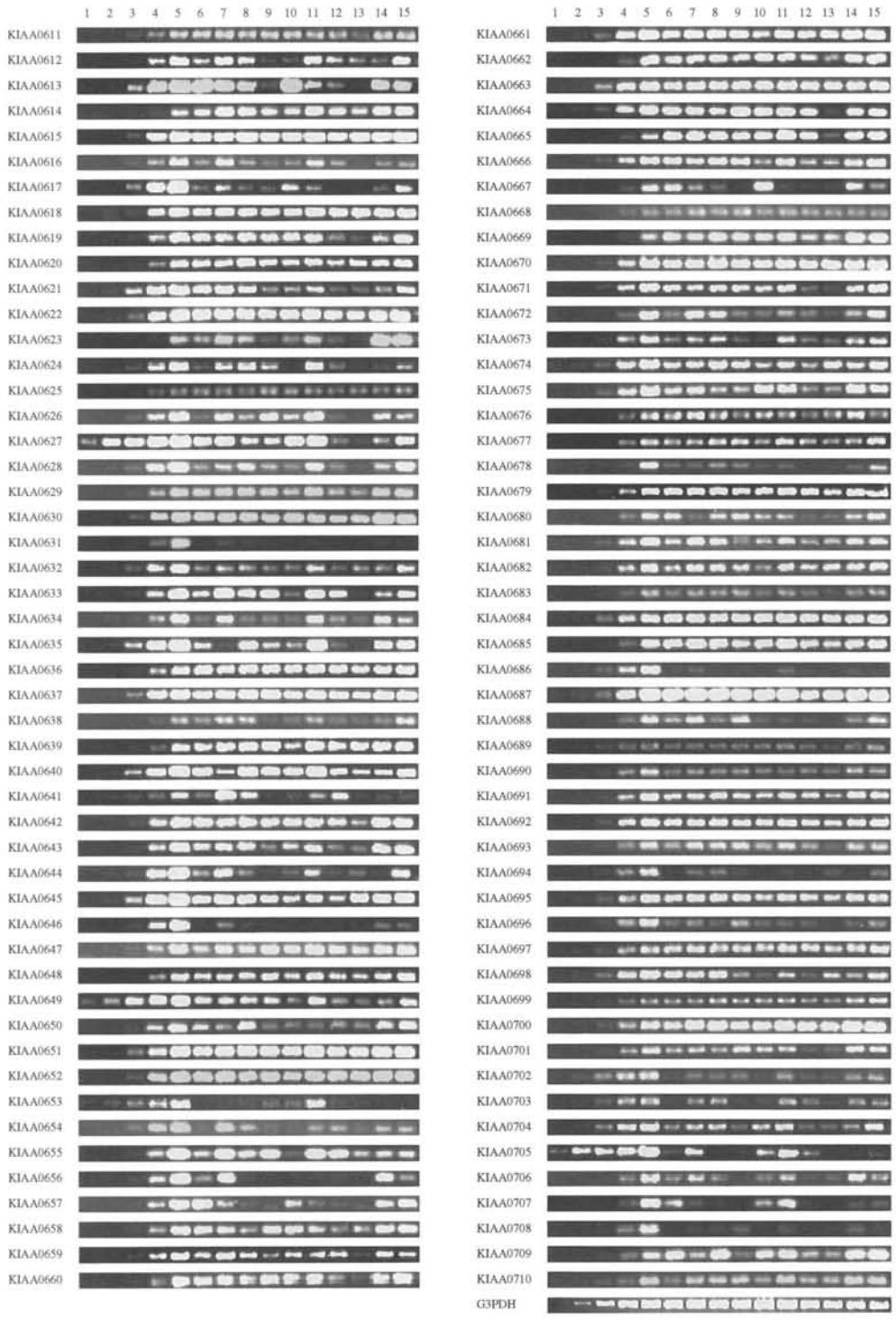

Figure 2. Expression profiles of 100 newly identified genes in 10 different tissues examined by RT-PCR. Electrophoretically resolved bands of the PCR products for individual genes are shown. Gene numbers are given on the left and G3PDH gene expression was analyzed as a positive control (at the bottom of the right column). In each set of electrophoretic patterns, lanes 1 to 5 show the PCR products derived from serial ten fold dilutions of a cDNA clone of interest (from 0.1 fg to 1 pg) for the estimation of the PCR amplification efficiency. Lanes 6 to 15 are electrophoretic patterns of the RT-PCR products originated from mRNAs of 10 different tissues: lane 6, heart; lane 7, brain; lane 8, lung; lane 9, liver; lane 10, skeletal muscle; lane 11, kidney; lane 12, pancreas; lane 13, spleen; lane 14, testis; lane 15, ovary.

mapping were carried out as described previously,[2,4] except that the RT-PCR analyses were done using poly(A)$^+$ RNAs from 10 different human tissues. The actual primer sequences and the reaction conditions used for PCR are available upon request, and are accessible through the World Wide Web at http://www.kazusa.or.jp.

As the nucleotide sequences of original 13 clones for KIAA0698 -KIAA0710 harbored two relatively long open reading frames (ORFs), the sequences between the two ORFs was re-determined by direct analysis of the RT-PCR products as described previously.[5] For clones which carried internal *Not* I site, the nucleotide sequences were determined after subcloning of the internal *Not* I-digested fragment as described before.[3]

3. Results and Discussion

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

By using *in vitro* protein-coding potentiality assay, we selected clones bearing the coding potential for large proteins from the size-fractionated human brain cDNA libraries with an average insert size between 3.9 and 6.1 kb.[2] Approximately half of the cDNA clones were derived from library No. 2 (average insert size=3.9 kb), because cDNA clones within this insert size range have not been extensively characterized yet by our cDNA analysis. One hundred clones thus selected were subjected to sequencing of entire inserts. As described previously, nucleotide sequences of the clones suspected to contain artifacts were further examined by direct sequencing of RT-PCR products. Four clones (KIAA0698, KIAA0700, KIAA0703 and KIAA0710) were found to carry insertions which appeared to be introns and the ORFs in 9 clones (KIAA0699, KIAA0701, KIAA0702, KIAA0704-0709) were artificially interrupted.The final sequence data revealed that the average size of these cDNA inserts and of their ORFs was 4.9 kb and 2.8 kb (corresponding to 922 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. It should be noted that the sequence information for KIAA0698-KIAA0710 shown in Fig. 1 and Table 1 were revised according to the results obtained by RT-PCR analyses. The in-frame termination codons upstream of the first ATG codon were identified in 42 clones, among which 34 clones carried the ATG codon within the context of Kozak's rule.[6] Short interspersed nucleotide elements (SINEs) including Alu and MIRs sequences and other repetitive sequences found using RepeatMasker program are indicated by dotted and hatched boxes, respectively. Table 1 lists the size of the inserts, as well as the ORF, the molecular masses of the largest *in vitro* products, and the chromosomal locations of the respective clones analyzed here.

3.2. Functional classification of predicted gene products

By computer-assisted analysis of the sequences with DNA, protein, and protein-motif databases [GenBank (release 104.0), OWL (release 30.1), and PROSITE (release 14.0) databases] using Wisconsin Sequence Analysis Package^{TM} (version 8; Genetics Computer Group, Inc. USA), the predicted coding sequences of 87 genes were found to exhibit significant similarities to known genes, and 62% of them were assigned to functional categories of proteins 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 homology/motif analysis are summarized in Table 2. The most homologous sequences in the public databases were aligned along with the newly identified ones with the GAP program and the results are also given in Table 2, where the degree of the overall similarity is indicated as "identical" (>90% to entries of human proteins), "homologous" (>90% to non-human protein entries), "related" (30–90% to entries of any organisms), or "weakly related" (<30% to entries of any organisms).

Other features to be noted are summarized below.

1. Except for the C2H2 type zinc finger protein family, 14 newly identified genes constitute 12 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 5 groups were judged as "uncharacterized."
2. Protein motif search against the PROSITE database revealed that protein kinase signatures are present in three genes (KIAA0619, KIAA0623 and KIAA0641) and that a C3HC4 type zinc finger signature exists in KIAA0646 and KIAA0708. Interestingly, C2H2-type zinc finger domain was identified only in one gene (KIAA0628) out of the 100 newly identified genes while the prevalence of genes encoding this domain in previously analysis was 7% (19 out of 278 genes in the previous reports[3–5]). The low occurrence of C2H2-type zinc finger domain seems to be linked to the smaller insert size of cDNA clones analyzed in this study than those analyzed before. LIM signature, which is a novel cysteine-rich zinc-binding domain and present in the protein family related to developmental regulation,[14] was identified in KIAA0613.
3. Several gene products exhibited local structural similarities to those we characterized previously. The

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)	Apparent molecular mass (kDa)[c]	Chromosomal location[d]
0611	AB014511	7,176	912	>100	20
0612	AB014512	6,458	1,736	>100	17
0613[g]	AB014513	4,223	734	>100	10
0614	AB014514	7,084	1,630	>100	12
0615[g]	AB014515	3,319	896	>100	16
0616	AB014516	6,718	634	75	19
0617	AB014517	6,721	768	99	2
0618	AB014518	7,011	999	>100	7
0619	AB014519	6,409	1,388	>100	2
0620	AB014520	6,754	1,985	>100	3
0621	AB014521	6,300	753	87	5
0622	AB014522	6,951	1,289	>100	2
0623	AB014523	5,808	1,036	>100	17
0624	AB014524	6,542	1,983	>100	11
0625	AB014525	5,176	791	75	9
0626	AB014526	6,184	409	51	4
0627	AB014527	5,614	1,324	>100	3
0628	AB014528	5,871	536	62	8
0629	AB014529	5,883	612	80	13
0630	AB014530	5,761	490	50	1
0631	AB014531	5,878	634	73	15
0632	AB014532	5,413	727	96	7
0633	AB014533	5,289	1,316	>100	7
0634	AB014534	4,591	1,321	>100	9
0635	AB014535	5,138	846	>100	4
0636	AB014536	4,737	537	65	8
0637	AB014537	5,217	1,171	>100	22
0638	AB014538	4,915	1,234	>100	11
0639	AB014539	4,950	574	87	8
0640	AB014540	4,824	603	66	11
0641	AB014541	5,140	1,207	>100	17[e]
0642	AB014542	4,959	1,069	>100	2
0643	AB014543	5,127	403	68	16
0644	AB014544	4,933	811	85	7
0645	AB014545	5,292	1,572	>100	22
0646	AB014546	5,545	485	62	6
0647	AB014547	5,183	1,016	>100	17
0648	AB014548	5,177	851	>100	4
0649	AB014549	4,932	1,209	>100	9
0650	AB014550	5,003	848	98	18
0651	AB014551	4,005	910	>100	1
0652	AB014552	4,040	517	67	11
0653	AB014553	4,358	558	60	21
0654	AB014554	4,081	1,093	>100	19
0655	AB014555	4,457	1,083	>100	12
0656	AB014556	4,369	907	>100	6
0657	AB014557	3,985	771	72	2
0658	AB014558	4,103	589	67	11
0659	AB014559	4,459	647	70	11
0660	AB014560	4,210	482	64	4
0661	AB014561	4,199	1,001	>100	16
0662	AB014562	3,944	677	>100	9
0663	AB014563	4,365	810	>100	1
0664	AB014564	4,611	1,134	>100	17
0665	AB014565	4,080	756	>100	16
0666	AB014566	4,153	1,085	>100	14
0667	AB014567	4,157	1,111	>100	3
0668	AB014568	3,767	742	75	22
0669	AB014569	4,550	780	>100	3
0670	AB014570	4,425	1,280	>100	14
0671	AB014571	4,423	536	66	2
0672	AB014572	4,227	818	>100	17
0673	AB014573	4,096	1,215	>100	1
0674	AB014574	4,263	1,234	>100	9
0675	AB014575	4,246	1,208	>100	3
0676	AB014576	4,491	1,262	>100	5
0677	AB014577	4,417	1,064	>100	1
0678	AB014578	3,811	1,021	>100	3
0679	AB014579	4,303	767	>100	10
0680	AB014580	4,318	634	95	6
0681	AB014581	4,323	538	57	20
0682	AB014582	4,422	960	>100	12
0683	AB014583	4,337	584	74	16
0684	AB014584	4,124	903	92	1
0685	AB014585	3,981	927	>100	22
0686	AB014586	4,299	326	58	5
0687	AB014587	4,266	1,175	>100	2
0688	AB014588	4,301	837	81	1
0689	AB014589	3,859	547	60	5
0690	AB014590	4,038	1,214	>100	10
0691	AB014591	4,133	753	>100	19
0692	AB014592	4,088	783	80	12
0693	AB014593	4,493	404	60	20
0694	AB014594	4,132	513	62	2
0695	AB014595	4,376	717	70	10
0696	AB014596	4,230	550	67	5
0697	AB014597	3,900	968	86	4
0698[f]	AB014598	4,227	891	79	X
0699[f]	AB014599	6,329	847	88	9
0700[f]	AB014600	5,020	1,130	>100	19
0701[f]	AB014601	5,976	630	51	12
0702[f]	AB014602	5,474	985	62	15
0703[f]	AB014603	3,600	963	63	16
0704[f]	AB014604	5,066	667	83	7
0705[f]	AB014605	6,379	1,278	>100	7
0706[f]	AB014606	6,217	1,103	80	17
0707[f]	AB014607	6,359	630	81	1
0708[f,g]	AB014608	5,404	1,753	>100	6
0709[f]	AB014609	5,641	1,479	>100	17
0710[f]	AB014610	4,607	1,115	>100	12

a) Accession numbers of DDBJ, EMBL and GenBank databases.
b) Values excluding poly(A) sequences.
c) Approximate molecular masses of the *in vitro* products estimated by SDS-PAGE.
d) Chromosome numbers identified by using GeneBridge 4 radiation hybrid panel unless specified.
e) Chromosome number determined by using CCR human-rodent hybrid panel.
f) cDNA and ORF lengths were revised by direct analysis of the RT-PCR products.
g) Nucleotide sequences were determined after subcloning of the internal *Not* I-digested fragment.

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)[e]
Cell signaling/communication	0613	R	enigma (H)	A55050	35.6	410
	0614	W	guanine nucleotide exchange factor p532 (H)	U50078	26.3	361
	0617	I	Hs-cul-3 (H)	U58089	100.0	577
	0619	H	serine/threonine-specific protein kinase (B)	S70633	97.8	1388
	0620	R	plexin 1 (M)	JC4980	26.7	1958
	0621	H	Rho-Gap protein (C)	U36309	91.6	510
	0623	R	unc-51 (Ce)	Z38016	37.2	573
	0629	R	A-kinase anchoring protein AKAP 220 (R)	U48288	67.9	582
	0630	R	serine/threonine protein kinase (H)	AF004849	32.2	295
	0636	R	copine I (H)	U83246	65.7	537
	0641	R	apoptosis associated tyrosine kinase AATYK (M)	AF011908	70.3	1222
	0644	R	insulin-like growth factor binding complex acid labile chain (M)	JC6128	35.9	312
	0647	R	KIAA0371 (H)	AB002369	45.1	1035
	0651	I	guanine nucleotide regulatory factor LFP40 (H)	U72206	92.3	801
	0658	H	photolyase/blue-light receptor homolog2 (M)	AB003433	95.6	569
	0660	R	ras-GTPase-activating SH3-domain binding protein G3BP (M)	U65313	60.6	452
	0669	W	TSC-22 variant (C)	D82364	27.5	534
	0671	H	SOCS-5 (M)	AF033187	94.6	536
	0672	R	SH3-binding protein 3BP-1 (M)	P55194	34.0	571
	0685	W	SIT4-associated protein SAP190 (Sc)	P36123	24.9	305
	0686	W	latrophilin-related protein 1 precursor (R)	U78105	27.0	267
	0687	R	NIK (M)	U88984	72.5	958
	0688	R	secretory protein containing thrombospondin motifs (M)	D67076	53.8	656
	0695	R	Hs-cul-4A (H)	U58090	87.4	414
	0698	R	ceruloplasmin precursor (H)	P00450	36.3	468
	0702	R	Na+/Ca2+,K+-exchanging protein (B)	S20969	55.8	893
	0703	R	calcium-transporting ATPase 1(Sc)	P13586	48.0	773
	0709	R	lectin lambda (M)	U56734	89.0	1481
Cell structure/motility	0618	R	nuclear envelope pore membrane protein POM 121 (R)	P52591	57.3	970
	0626	R	microfibril-associated glycoprotein 3 precursor (H)	P55082	57.4	291
	0634	R	astrotactin (M)	U48797	50.1	821
	0635	W	nonmuscle myosin heavy chain-B MYH10 (H)	M69181	17.1	847
	0639	W	kinesin-73 (D)	U81788	37.0	246
	0640	W	myosin heavy chain (D)	P05661	21.4	215
	0649	W	collagen alpha 1(VII) chain precursor (H)	Q02388	20.7	478
	0654	R	LAR-interacting protein LIP1b (H)	S55553	51.8	1117
	0655	R	huntingtin interacting protein HIP1 (H)	U79734	44.9	927
	0656	H	clathrin assembly protein AP180 short form (R)	S36326	83.2	924
	0657	W	connectin/titin (C)	D83390	20.1	523
	0662	W	UNC-89 (Ce)	U33058	21.1	199
	0664	R	150-kD protein cluA (Dd)	U49332	42.9	296
	0706	I	kinesin-like motor protein KIF1C (H)	U91329	99.3	1103
Nucleic acid management	0625	W	tRNA-splicing endonuclease positive effector (Sc)	Q00416	30.8	533
	0628	R	finger protein HZF10, Krueppel-related (H)	S47072	58.6	396
	0646[f]		none			
	0667	R	TIP120 (R)	D87671	52.1	912
	0670	W	U1 small nuclear ribonucleoprotein 70 KD (H)	P08621	31.1	183
	0677	W	putative 90.2 KD zinc finger protein (Sc)	P39956	25.9	536
	0681	R	transcriptional repressor protein Scm (D)	U49793	35.9	259
	0682	R	hypothetical protein YPR112c (Sc)	S59777	29.8	786
	0689	R	cleavage stimulation factor, 64 KD subunit (H)	P33240	67.0	461
	0699	R	Bic-D protein (H)	U90028	64.7	821
	0700	H	mSin3B (M)	L38622	90.0	710
	0708[f]	R	KIAA0076 (H)	Q14999	60.2	862
	0710	W	poly(A)-binding protein dependent poly(A)-ribonuclease subunit PAN2 (Sc)	U39204	34.3	251
Metabolism	0611	H	putative E1-E2 ATPase (M)	AF011336	98.1	785
	0631	R	putative long-chain-fatty-acid-CoA ligase (Mt)	Q10776	37.0	359
	0704	R	oxysterol-binding protein (H)	P22059	35.9	409
Cell division	0666	R	KIAA0381 (H)	AB002379	67.7	870
	0668	W	spindle pole body associated protein SAD1 (Sp)	Q09825	32.3	195
Unclassified	0612	R	KIAA0318 (H)	AB002316	63.4	93
	0615	R	KIAA0323 (H)	AB002321	61.4	184
	0622	W	hypothetical 117.3 KD protein (Ce)	P32744	24.3	535
	0627	W	hypothetical 117.3 KD protein (Ce)	P32744	22.6	680
	0633	R	cordon-bleu (M)	U26967	70.6	279
	0637	W	Caenorhabditis elegans cosmid C10A4 (Ce)	U23454	20.8	178
	0638	H	LL5 protein (R)	S37032	66.5	653
	0642	R	glutamine rich protein (C)	U90567	84.3	875
	0645	W	Caenorhabditis elegans cosmid T08A11 (Ce)	Z50875	41.1	353
	0648	W	bimD protein (En)	S52957	18.3	840

Table 2. Continued.

	0652	W	hypothetical 49.8 KD protein (Ce)	P34379	30.5	200
	0653	W	CD80-like protein precursor (C)	Y08823	32.9	298
	0659	W	Caenorhabditis elegans cosmid F42G9 (Ce)	U00051	33.3	129
	0661	W	hypothetical 97.1 KD protein (Ce)	P34537	29.0	376
	0665	W	Caenorhabditis elegans cosmid F55C12 (Ce)	U41107	28.2	287
	0673	W	Caenorhabditis elegans cosmid R13H4 (Ce)	Z81579	21.4	252
	0674	W	yotiao (H)	AF026245	20.0	385
	0675	W	TPR repeat protein D (H)	P53804	26.0	227
	0676	W	MIC1 protein (Sc)	P53258	32.5	455
	0678	R	Caenorhabditis elegans cosmid F18C12 (Ce)	Z75536	41.2	1041
	0679	W	Caenorhabditis elegans cosmid T20B5 (Ce)	U28742	32.4	370
	0680	W	Caenorhabditis elegans cosmid F26H9 (Ce)	Z81516	41.5	224
	0684	W	NOSA (D)	AF044255	27.3	704
	0690	W	probable membrane protein YPL012w (Sc)	S52519	24.4	501
	0691	W	S.pombe chromosome I cosmid c1B3 (Sp)	Z98598	32.5	382
	0693	R	SYT (M)	X93357	41.0	393
	0696	R	beta-transducin repeat-containing protein (X)	B48088	88.0	482
	0705	W	Caenorhabditis elegans cosmid K01A6 (Ce)	Z68750	24.2	392
No homology	0616		none			
	0624		none			
	0632		none			
	0643		none			
	0650		none			
	0663		none			
	0683		none			
	0692		none			
	0694		none			
	0697		none			
	0701		none			
	0707		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%). The gene products in class I (>90%) include alternative splicing products of reported genes.
c) Organisms in which these entries were identified are given in parentheses: B, bovine; C, chicken; Ce, *Caenorhabditis elegans*; Cr, *Chlamydomonas reinhardtii*; D, *Drosophila melanogaster*; Dd, *Dictyostelium discoideum*; En, *Emericella nidulans*; H, human; M, mouse; Mt, *Mycobacterium tuberculosis*; Oc, *Oryctolagus cuniculus*; R, rat; Sc, *Saccharomyces cerevisiae*; Sp, *Schizosaccharomyces pombe*; X, *Xenopus laevis*.
d) Accession numbers of homologous entries in DDBJ/EMBL/GenBank/OWL/SWISS-PLOT/PIR database are shown.
e) The values were obtained by the FASTA program.
f) Classifications based on the sequence motifs.

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
KIAA0612	KIAA0318	AB002316	41.2	uncharacterized
KIAA0615	KIAA0323	AB002321	34.9	uncharacterized
KIAA0617	KIAA0695[c)]	AB014595	41.0	cullin[7]
KIAA0620	KIAA0315	AB002313	36.6	MET-hepatocyte growth factor receptor family[8]
	KIAA0407	AB007867	38.5	MET-hepatocyte growth factor receptor family[8]
	KIAA0463	AB007932	34.9	MET-hepatocyte growth factor receptor family[8]
KIAA0622	KIAA0627[c)]	AB014527	73.1	uncharacterized
KIAA0634	KIAA0289	AB006627	51.5	astrotactin[9]
KIAA0647	KIAA0371	AB002369	49.5	putative tyrosine phosphatase family[10]
KIAA0651	KIAA0521	AB011093	37.9	uncharacterized
KIAA0666	KIAA0381	AB002379	69.5	diaphanous[11]
KIAA0687	KIAA0551	AB011123	73.2	Ste20-related kinase[12]
KIAA0688	KIAA0366	AB002364	29.9	metalloproteinase-disintegrin family[13]
KIAA0708	KIAA0076	D38548	57.5	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.

Hect domain, which is known to be present in the carboxy-terminal regions of several ubiquitine-protein ligases,[15] has been frequently found in the predicted gene products in our cDNA project. In addition to 8 previously identified genes (KIAA0010, KIAA0032, KIAA0045, KIAA0093, KIAA0312, KIAA0317, KIAA0322, and KIAA0439), KIAA0614 also encodes a protein with the *Hect* domain. KIAA0639 has similarity to KIAA0291 in a region including the CAP-Gly domain signature, which appeared in some cytoskeleton-associated proteins.[16]

3.3. Expression profiles of the predicted genes

In order to obtain further information as to biological significance of these newly identified genes, we examined their expression profiles in 10 different human tissues by RT-PCR. Figure 3 shows the results of the RT-PCR analyses for KIAA0611 to KIAA0710. The first 5 lanes are controls (respective cDNA plasmids), which allowed us to evaluate the efficiency of the PCR amplification we employed. The expression profiles of these newly identified genes can be categorized into the following classes: genes expressed (11 genes) or suppressed (56 genes) in a limited number of tissues, and genes expressed ubiquitously (33 genes). We excluded genes specifically expressed in brain in this study, which are known to occur approximately 10% of the randomly sampled clones.[4,5] They are being further analyzed in depth for their possible biological significance and the results will be reported elsewhere.

Acknowledgments: This project was supported by grants from the Kazusa DNA Research Institute. We thank Dr. M. Oishi and Dr. M. Takanami for their continuous support and encouragement. Thanks are also due to Tomomi Tajino, Keishi Ozawa, Tomomi Kato, Kazuhiro Sato, Akiko Ukigai, Emiko Suzuki, Kazuko Yamada, Naoko Suzuki, Kozue Kaneko, and Naoko Shibano for their technical assistance.

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. 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.
3. Nagase, T., Ishikawa, K.-I., Miyajima, N. et al. 1998, Prediction of the coding sequences of unidentified human genes. IX. The complete sequences of 100 new cDNA clones from brain which can code for large proteins *in vitro*, *DNA Res.*, **5**, 31–39.
4. Nagase, T., Ishikawa, K.-I., Nakajima, D. et al. 1997, Prediction of the coding sequences of unidentified human genes. VII. The complete sequences of 100 new cDNA clones from brain which can code for large proteins *in vitro*, *DNA Res.*, **4**, 141–150.
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. Kozak, M. 1996, Interpreting cDNA sequences: some insights from studies on translation, *Mammalian Genome*, **7**, 563–574.
7. Kipreos, E. T., Lander, L. E., Wing, J. P., He, W. W., and Hedgecock, E. M. 1996, *cul-1* is required for cell cycle exit in C. elegans and identifies a novel gene family, *Cell*, **85**, 829–839.
8. Maestrini, E., Tamagnone, L., Longati, P. et al. 1996, A family of transmembrane proteins with homology to the MET-hepatocyte growth factor receptor, *Proc. Natl. Acad. Sci. USA*, **93**, 674–678.
9. Zheng, C., Heintz, N., and Hatten, M. E. 1996, CNS gene encoding astrotactin, which supports neuronal migration along glial fibers, *Science*, **272**, 417–419.
10. Laporte, J., Hu, L. J., Kretz, C. et al. 1996, A gene mutated in X-linked myotubular myopathy defines a new putative tyrosine phosphatase family conserved in yeast, *Nature Genet.*, **13**, 175–182.
11. Castrillon, D. H. and Wasserman, S. A. 1994, Diaphanous is required for cytokinesis in Drosophila and shares domains of similarity with the products of the limb deformity gene, *Development*, **120**, 3367–3377.
12. Jiahuai, Y. S., Han, J., Xu, S., Cobb, M., and Skolnik, E. Y. 1997, NIK is a new Ste20-related kinase that binds NCK and MEKK1 and activates the SAPK/JNK cascade via a conserved regulatory domain, *EMBO J.*, **16**, 1279–1290.
13. Kuno, K., Kanada, N., Nakashima, E., Fujiki, F., Ichimura, F., and Matsushima, K. 1997, Molecular cloning of a gene encoding a new type of metalloproteinase-disintegrin family protein with thrombospondin motifs as an inflammation associated gene, *J. Biol. Chem.*, **272**, 556–562.
14. Sánchez-García, I. and Rabbitts, T. H. 1994, The LIM domain: a new structural motif found in zinc-finger-like proteins, *Trends Genet.*, **10**, 315–320.
15. Huibregtse, J. M., Scheffner, M., Beaudenon, S., and Howley, P. M. 1995, A family of proteins structurally and functionally related to the E6-AP ubiquitin-protein ligase, *Proc. Natl. Acad. Sci. USA*, **92**, 2563–2567.
16. Riehemann, K. and Sorg, C. 1993, Sequence homologies between four cytoskeleton-associated proteins, *Trends Biochem. Sci.*, **18**, 82–83.