Prediction of the Coding Sequences of Unidentified Human Genes. XVI. The Complete Sequences of 150 New cDNA Clones from Brain Which Code for Large Proteins *in vitro*

Takahiro NAGASE,* Reiko KIKUNO, Ken-ichi ISHIKAWA, Makoto HIROSAWA, and Osamu OHARA

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

(Received 28 January 2000)

Abstract

We have carried out a human cDNA sequencing project to accumulate information regarding the coding sequences of unidentified human genes. As an extension of the preceding reports, we herein present the entire sequences of 150 cDNA clones of unknown human genes, named KIAA1294 to KIAA1443, from two sets of size-fractionated human adult and fetal brain cDNA libraries. The average sizes of the inserts and corresponding open reading frames of cDNA clones analyzed here reached 4.8 kb and 2.7 kb (910 amino acid residues), respectively. From sequence similarities and protein motifs, 73 predicted gene products were functionally annotated and 97% of them were classified into the following four functional categories: cell signaling/communication, nucleic acid management, cell structure/motility and protein management. Additionally, the chromosomal loci of the genes were assigned by using human-rodent hybrid panels for those genes whose mapping data were not available in the public databases. The expression profiles of the genes were also studied in 10 human tissues, 8 brain regions, spinal cord, fetal brain and fetal liver by reverse transcription-coupled polymerase chain reaction, products of which were quantified by enzyme-linked immunosorbent assay.

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

We have been making efforts to accumulate information on the coding sequences of unidentified human genes.[1,2] Especially, recent our interest is focused on the unidentified genes encoding large proteins in human brain since these gene products are likely to play important roles in the central nervous system.[2,3] To identify such genes, we constructed a set of strictly size-fractionated cDNA libraries from human brain and *in vitro* transcription/translation system have been applied to select the cDNA clones coding for large proteins prior to the determination of their entire sequence.[3] As an alternative method for clone selection, we have recently introduced a computer-based approach using GeneMark analysis for picking up cDNA clones with a high probability of coding for protein.[4] This new approach would be expected to minimize the risk of overlooking important cDNA clones which fail to produce proteins *in vitro*.

The sequences of more than 1200 cDNA clones have been reported by our project and the total length of the determined sequences exceeds 6.3 Mb[1–3] and the average length of gene products deduced from the cDNAs from brain is over 900 amino acid residues.[2,3] As an extension of the preceding reports, we herein report the coding sequence features of 150 new cDNA clones which have the potential to code for large proteins *in vitro*. In addition to the specific features of the newly predicted protein sequences annotated by the database search, the expression profiles and the chromosomal locations of these 150 new genes are also described. The information regarding these newly identified genes would greatly increase our understanding of the biological functions of human genes at the molecular level.

1. Sequence Analysis and Prediction of Protein-Coding Regions in cDNA Clones

cDNA clones to be entirely sequenced were selected according to the following criteria: (1) novelties of their single-pass sequences of both the cDNA ends; (2) potentialities of their protein coding. The latter criterion was critical for us to conduct our cDNA project efficiently, because there are many cDNA clones which apparently do not possess a protein-coding region in the

Communicated by Michio Oishi

* To whom correspondence should be addressed. Tel. +81-438-52-3930, Fax. +81-438-52-3931, E-mail: nagase@kazusa.or.jp

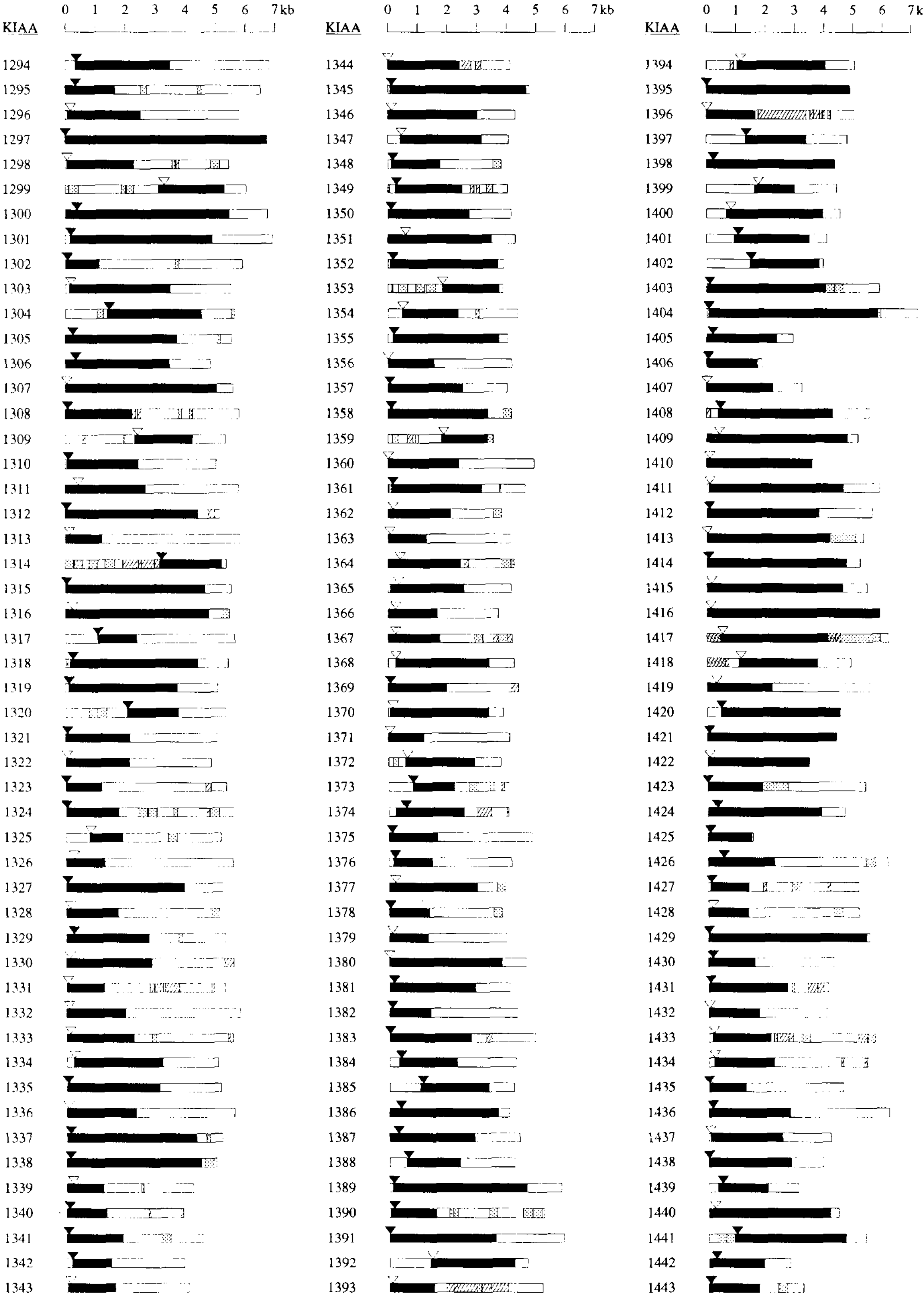

Figure 1. Physical maps of cDNA clones analyzed. The physical maps shown here were constructed from the sequence data of respective cDNA clones or, when necessary, from the combination of cDNA clones and RT-PCR products. The horizontal scale represents the cDNA length in kb, and the gene numbers corresponding 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, 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)[c)]	Chromosomal location[d)]	Gene number (KIAA)	Accession number[a)]	cDNA length (bp)[b)]	ORF length (amino acid residues)[c)]	Chromosomal location[d)]
1294	AB037715	6,816	1,051	10*	1369	AB037790	4,391	653	7*
1295	AB037716	6,524	550	5*	1370	AB037791	3,863	1,107	15*
1296	AB037717	5,796	815	10*	1371	AB037792	4,096	395	4
1297	AB037718	6,726	2,242	2	1372	AB037793	3,771	773	11
1298	AB037719	5,463	738	12*	1373	AB037794	4,052	463	10
1299	AB037720	6,043	730	16	1374	AB037795	4,044	764	3
1300	AB037721	6,747	1,820	15	1375	AB037796	4,823	546	3*
1301	AB037722	6,926	1,581	2	1376	AB037797	4,131	437	5*
1302	AB037723	5,904	375	11*	1377	AB037798	3,916	988	11
1303	AB037724	5,538	1,119	17*	1378	AB037799	3,815	451	4*
1304	AB037725	5,633	1,051	12*	1379	AB037800	3,956	434	6*
1305	AB037726	5,553	1,238	14*	1380	AB037801	4,614	1,265	10*
1306	AB037727	4,832	1,154	16*	1381	AB037802	4,052	961	17*
1307	AB037728	5,601	1,678	1	1382	AB037803	4,312	462	12*
1308	AB037729	5,796	745	9*	1383	AB037804	4,904	907	1
1309	AB037730	5,331	639	9	1384	AB037805	4,261	652	18#
1310	AB037731	5,028	794	2*	1385	AB037806	4,193	768	14*
1311	AB037732	5,774	889	5*	1386	AB037807	4,030	1,214	7*
1312	AB037733	5,139	1,471	3	1387	AB037808	4,385	950	2
1313	AB037734	5,818	398	X	1388	AB037809	4,220	599	16#
1314	AB037735	5,369	681	18	1389	AB037810	5,801	1,514	1*
1315	AB037736	5,526	1,545	6*	1390	AB037811	5,222	505	1
1316	AB037737	5,477	1,590	14	1391	AB037812	5,901	1,194	11*
1317	AB037738	5,646	435	5	1392	AB037813	4,654	950	4
1318	AB037739	5,425	1,418	X	1393	AB037814	5,164	500	14
1319	AB037740	5,073	1,208	1*	1394	AB037815	5,065	1,003	11#
1320	AB037741	5,321	567	6*	1395	AB037816	4,886	1,628	19
1321	AB037742	5,058	714	17	1396	AB037817	5,041	551	19
1322	AB037743	4,832	702	4	1397	AB037818	4,810	686	6
1323	AB037744	5,356	396	18	1398	AB037819	4,372	1,456	7
1324	AB037745	5,567	580	1	1399	AB037820	4,450	452	2*
1325	AB037746	5,155	354	4*	1400	AB037821	4,554	1,093	4
1326	AB037747	5,563	424	14	1401	AB037822	4,107	853	17
1327	AB037748	5,205	1,310	4*	1402	AB037823	3,970	788	17*
1328	AB037749	5,097	574	18	1403[e)]	AB037824	5,897	1,337	15*
1329	AB037750	5,287	907	4*	1404[e)]	AB037825	7,204	1,925	20*
1330	AB037751	5,577	945	15*	1405[e)]	AB037826	2,945	791	1*
1331	AB037752	5,273	412	3*	1406[e)]	AB037827	1,876	571	9
1332	AB037753	5,788	651	1*	1407[e)]	AB037828	3,247	744	3*
1333	AB037754	5,534	741	14*	1408[e)]	AB037829	5,568	1,298	10*
1334	AB037755	5,043	989	5*	1409[e)]	AB037830	5,160	1,597	14
1335	AB037756	5,123	1,026	20*	1410[e)]	AB037831	3,604	1,201	3
1336	AB037757	5,591	766	2*	1411[e)]	AB037832	5,900	1,522	6*
1337	AB037758	5,181	1,438	1	1412[e)]	AB037833	5,664	1,274	9*
1338	AB037759	4,994	1,495	15*	1413[e)]	AB037834	5,361	1,399	4
1339	AB037760	4,217	409	7	1414[e)]	AB037835	5,242	1,586	2*
1340	AB037761	3,876	441	12*	1415[e)]	AB037836	5,480	1,539	20
1341	AB037762	4,544	620	15*	1416[e)]	AB037837	5,901	1,967	8
1342	AB037763	3,910	426	18*	1417[e)]	AB037838	6,206	1,217	6*
1343	AB037764	4,083	520	1*	1418[e)]	AB037839	4,896	889	2*
1344	AB037765	4,135	806	14*	1419[e)]	AB037840	5,560	738	11#
1345	AB037766	4,790	1,532	4#	1420[e)]	AB037841	4,516	1,349	1
1346	AB037767	4,309	999	21*	1421[e)]	AB037842	4,391	1,463	15
1347	AB037768	4,075	918	3	1422[e)]	AB037843	3,456	1,151	9
1348	AB037769	3,830	545	16	1423[e,g)]	AB037844	5,390	616	6*
1349	AB037770	4,055	752	17	1424[e,g)]	AB037845	4,655	1,286	6
1350	AB037771	4,153	911	4	1425[e,g)]	AB037846	1,543	495	1#
1351	AB037772	4,307	1,163	10*	1426[g)]	AB037847	6,140	758	16
1352	AB037773	3,893	1,212	5*	1427[g)]	AB037848	5,145	439	11*
1353	AB037774	3,877	640	1	1428[g)]	AB037849	5,148	458	3*
1354	AB037775	4,352	632	9*	1429[g)]	AB037850	5,507	1,795	8*
1355	AB037776	4,036	1,189	1	1430[g)]	AB037851	4,282	527	4
1356	AB037777	4,183	519	2	1431[g)]	AB037852	4,076	891	19
1357	AB037778	4,022	836	6*	1432[g)]	AB037853	4,024	571	9*
1358	AB037779	4,183	1,123	7	1433[g)]	AB037854	5,671	652	2
1359	AB037780	3,550	517	3	1434[g)]	AB037855	5,443	677	20*
1360	AB037781	4,944	796	12*	1435[g)]	AB037856	4,574	415	2*
1361	AB037782	4,620	1,005	17*	1436[g)]	AB037857	6,160	924	1*
1362	AB037783	3,842	699	12	1437[g)]	AB037858	4,161	811	9*
1363	AB037784	4,116	430	3*	1438[g)]	AB037859	3,907	934	22
1364	AB037785	4,261	811	22	1439[f,g)]	AB037860	3,063	561	1*
1365	AB037786	4,150	831	1	1440[f,g)]	AB037861	4,434	1,377	7
1366	AB037787	3,716	550	17*	1441[f,g)]	AB037862	5,378	1,258	1*
1367	AB037788	4,196	579	14*	1442[f,g)]	AB037863	2,782	627	20
1368	AB037789	4,250	1,049	5*	1443[f,g)]	AB037864	3,239	573	14

a) Accession numbers of DDBJ, EMBL and GenBank databases. b) Values excluding poly(A) sequences. c) Values were calculated from the number of amino acid residues between two termination codons in the case where the in-frame termination codon exists upstream of the first ATG codon. d) Chromosome numbers were identified by using GeneBridge 4 radiation hybrid panel unless specified. The actual primer sequences and the PCR conditions used for the radiation hybrid mapping are accessible through the World Wide Web at http://www.kazusa.or.jp/huge. The chromosomal locations highlighted by asterisks were fetched from the UniGene database. The chromosomal locations highlighted by sharp were referred from the GenBank database because the sequences of the cDNA clones could be found in the genomic sequences whose chromosome numbers were assigned. e) cDNA and ORF lengths were revised by direct analysis of the RT-PCR products. f) Nucleotide sequences were determined after subcloning of the internal *Not* I-digested fragment. Therefore, cDNA length of these genes represented those of *internal Not* I-digested fragment. g) cDNA clones were selected by analysis of 5′-end single-pass sequences using the GeneMark analysis.

Table 2. Functional classifications of the gene products.

2-1. Predicted function based on homology search[a)]

Function[b)]	Gene product	aa.res.	OWL ID	aa.res.	% identity	% coverage[c)]	Definition
Cell signaling/communication	KIAA1296	815	AF078667	714	82	96	ponsin-1, complete cds. - mouse
	KIAA1297	2242	P53355	1431	35	13	death-associated protein kinase 1 - human
	KIAA1299	730	JC5887	670	93	92	signaling mediator variant - mouse
	KIAA1304	1051	P98171	946	48	72	rho-GAP hematopoietic protein C1 - human
	KIAA1308	745	Q03385	852	81	73	guanine nucleotidedissociation stimulator ralGDS form A - mouse
	KIAA1312	1471	D67076	951	44	46	secretory protein containing thrombospondin motifs, complete cds. - mouse
	KIAA1314	681	Y00661	1227	30	30	bcr - human
	KIAA1322	702	U81500	438	39	50	phgA gene, complete cds. - *Dictyostelium discoideum*
	KIAA1327	1310	T03730	1567	61	100	antigen containing epitope to monoclonal antibody MMS-85/12 - mouse
	KIAA1338	1495	M20487	1020	35	31	protein kinase GCN2, complete cds. - *S.cerevisiae*
	KIAA1342	426	P50232	425	90	100	synaptotagmin IV - rat
	KIAA1347	918	A42764	919	97	100	Ca2+-transporting ATPase (EC 3.6.1.38) - rat
	KIAA1348	545	AF062741	530	84	97	pyruvate dehydrogenase phosphatase isoenzyme 2, complete cds. - rat
	KIAA1356	519	P08104	1951	97	100	sodium channel protein, brain I alpha subunit - human
	KIAA1361	1005	AF084205	1001	99	100	serine/threonine protein kinase TAO1, complete cds. - rat
	KIAA1366	550	U41662	836	98	100	neuroligin 2, complete cds. - rat
	KIAA1368	1049	AF030430	888	93	84	semaphorin VIa, complete cds. - mouse
	KIAA1369	653	AF028808	619	83	95	hemin-sensitive initiation factor 2 alpha kinase, complete cds. - mouse
	KIAA1385	768	Q03555	736	100	96	gephyrin (putative glycine receptor-tublin linker protein) - rat
	KIAA1389	1514	AF090989	1783	53	96	putative GAP protein alpha, complete cds. - human
	KIAA1400	1093	U88549	896	97	80	OL-protocadherin, complete cds. - mouse
	KIAA1422	1151	AF089730	1237	94	91	potassium channel subunit (Slack), complete cds. - rat
	KIAA1424	1286	U02289	1439	48	17	GTPase-activating protein (CEGAP), partial cds. - *C. elegans*
	KIAA1427	439	P46096	421	32	61	synaptotagmin I - mouse
	KIAA1436	924	Q62786T	879	89	95	prostaglandin F2-alpha receptor regulatory protein precursor - rat
Nucleic acid management	KIAA1339	409	AF020591	715	45	61	zinc finger protein, complete cds. - human
	KIAA1341	620	A56704	435	90	73	regulatory protein Myef-2 - mouse
	KIAA1349	752	Q05481	1191	56	88	zinc finger protein 43 - human
	KIAA1367	579	Q10568	782	99	100	cleavage and polyadenylation specificity factor, 100 kD subunit - bovine
	KIAA1380	1265	Q63679	1214	46	66	testis specific protein A - rat
	KIAA1388	599	Q05481	1191	39	83	zinc finger protein 91 - human
	KIAA1396	551	P52742	469	59	83	zinc finger protein 135 - human
	KIAA1416	1967	X86691	1912	42	34	218kD Mi-2 - human
	KIAA1431	891	P10078	614	75	64	zinc finger protein ZFP28 - mouse
	KIAA1439	561	P09414	509	100	91	nuclear factor 1 (NF-I) - rat
	KIAA1442	627	U92704	551	77	83	Olf-1/EBF-like-2(0S) transcription factor, complete cds. - mouse
	KIAA1443	573	JC4863	873	35	35	homeotic protein protein zhx-1 - mouse
Protein management	KIAA1301	1581	P46934	927	36	49	KIAA0322, partial cds. - human
	KIAA1320	567	AF037454	854	45	61	ubiquitin protein ligase, complete cds. - mouse
	KIAA1346	999	T00017	951	82	95	ADAMTS-1 protein - mouse
	KIAA1352	1212	Q09996	1198	56	97	probable leucyl-tRNA synthetase (EC 6.1.1.4) - *C. elegans*
Metabolism	KIAA1363	430	A58922	398	43	94	esterase/N-deacetylase (EC 3.5.1.-), 50K hepatic - rabbit
Cell structure/motility	KIAA1294	1051	P26044	583	32	24	radixin - pig
	KIAA1306	1154	S22697	464	33	18	extensin - *Volvox carteri*
	KIAA1309	639	AF059569	593	30	85	actin binding protein MAYVEN, complete cds. - human
	KIAA1354	632	AF059569	593	30	86	actin binding protein MAYVEN, complete cds. - human
	KIAA1357	836	S22697	464	35	25	extensin - *Volvox carteri*
	KIAA1362	699	AF038388	766	33	64	actin-filament binding protein Frabin, complete cds. - rat
	KIAA1365	831	U66707	1495	93	100	densin-180, complete cds. - rat
	KIAA1378	451	AF059569	593	36	95	actin binding protein MAYVEN, complete cds. - human
	KIAA1405	791	AF009624	242	91	30	KIF3-related motor protein, partial cds. - human
	KIAA1410	1201	U03975	1125	77	68	dynein heavy chain isotype 6, partial cds. - sea urchin
	KIAA1437	811	U66707	1495	30	38	densin-180, complete cds. - rat

a) Homology search was performed by Smith-Waterman algorithm, using BioView Toolkit and GeneMatcher (revision 3.3, Paracel Inc. USA) against OWL database (release 31.4). The homologous protein with the highest score was listed, when it satisfied the following conditions, i) the protein was functionally annotated, ii) the aligned region exceeded 200 amino acid residues, and iii) percent identity in the algined region was 30% or greater. b) Function was classified based on the annotation of the entry of the homologous protein in the database. c) The values mean the ratio of the length of aligned region to the original length of the query sequence, in percentage.

cDNA libraries derived from tissue poly(A)$^+$ RNA. To screen cDNA clones according to their protein-coding capability, we have used an *in vitro* expression system and recently introduced a computer-based method called GeneMark analysis for minimizing the risk of overlooking important cDNA clones.[2,4] In this report, 21 cDNA clones were selected by GeneMark analysis and 129 cDNA clones were selected by the *in vitro* expression system. These cDNA clones were isolated from the size-fractionated human adult brain cDNA libraries Nos. 2 to 5 (insert sizes ranging from 4 to 6 kb) and the size-fractionated human fetal brain cDNA libraries Nos. 4 and 6 (insert sizes ranging from 4 to 7 kb) previously constructed.[2,3] The clones with unidentified sequences at both ends were chosen by single-pass sequencing and a homology search was performed against the GenBank database (release 113.0) excluding expressed sequence tags and genomic sequences.[3] A total of 35 cDNA clones (KIAA1389-KIAA1402, KIAA1415-KIAA1422, KIAA1424, KIAA1425 and KIAA1433-KIAA1443) were selected from the adult brain libraries and the remaining 115 cDNA clones were obtained from the fetal brain cDNA libraries. Entire sequencing of these clones was performed according to the methods previously described in detail.[2,3] Twenty-three clones (KIAA1403-KIAA1425) seemed to carry spurious coding interruption caused by errors of the reverse transcriptase or by retained intron sequences. For these cases, the sequences of the regions causing interruption of an open reading frame (ORF) were reexamined by direct se-

Table 2. Continued.

2-2. Predicted function by motif search[a)]

Function[b)]	Gene product	aa res.	Pfam ID	E-value[c)]	Definition
Cell signaling/communication	KIAA1295	550	PF00018	4.80E-06	SH3 domain
			PF00018	1.30E-04	SH3 domain
	KIAA1298	738	PF00782	2.10E-34	Dual specificity phosphatase, catalytic domain
	KIAA1330	945	PF00047	4.10E-02	Immunoglobulin domain
	KIAA1355	1189	PF00041	1.50E-09	Fibronectin type III domain
			PF00041	1.80E-08	Fibronectin type III domain
			PF00047	5.70E-01	Immunoglobulin domain
			PF00047	4.20E-12	Immunoglobulin domain
			PF00047	3.60E-08	Immunoglobulin domain
			PF00047	5.50E-05	Immunoglobulin domain
			PF00047	9.00E-06	Immunoglobulin domain
	KIAA1360	796	PF00069	3.00E-07	Eukaryotic protein kinase domain
	KIAA1391	1194	PF00169	9.30E-01	PH domain
			PF00620	7.30E-30	RhoGAP domain
	KIAA1406	1876	PF00888	4.00E-01	Cullin family
	KIAA1415	1539	PF00610	1.70E-10	Domain found in Dishevelled, Egl-10, and Pleckstrin
	KIAA1428	458	PF00169	5.10E-04	PH domain
			PF00640	8.70E-04	Phosphotyrosine interaction domain
Nucleic acid management	KIAA1311	889	PF00076	5.90E-02	RNA recognition motif
			PF00642	3.50E-02	Zinc finger C-x8-C-x5-C-x3-H type
	KIAA1343	520	PF00249	1.80E-08	Myb-like DNA-binding domain
			PF00249	4.10E-06	Myb-like DNA-binding domain
			PF01448	3.30E-12	ELM2 domain
	KIAA1384	652	PF00651	2.60E-24	BTB/POZ domain
			PF01344	4.10E-02	Kelch motif
			PF01344	7.60E-03	Kelch motif
			PF01344	5.10E-15	Kelch motif
			PF01344	5.20E-06	Kelch motif
			PF01344	5.90E-05	Kelch motif
			PF01344	1.20E-01	Kelch motif
	KIAA1425	495	PF00249	9.20E-01	Myb-like DNA-binding domain
	KIAA1441	1258	PF00096	3.10E-02	Zinc finger, C2H2 type
			PF00096	6.50E-02	Zinc finger, C2H2 type
			PF00096	9.80E-04	Zinc finger, C2H2 type
			PF00096	2.30E-02	Zinc finger, C2H2 type
			PF00096	5.20E-03	Zinc finger, C2H2 type
Cell structure/motility	KIAA1364	811	PF00307	8.60E-18	Calponin homology (CH) domain
			PF00412	3.30E-06	LIM domain containing proteins
Protein management	KIAA1333	741	PF00632	2.20E-01	HECT-domain
	KIAA1350	911	PF00443	6.30E-01	Ubiquitin carboxyl-terminal hydrolase family 2
	KIAA1372	773	PF00442	4.10E-13	Ubiquitin carboxyl-terminal hydrolases family 2
			PF00443	9.10E-20	Ubiquitin carboxyl-terminal hydrolase family 2
	KIAA1414	1586	PF00298	1.40E-01	Ribosomal protein L11
Metabolism	KIAA1315	1545	PF00389	3.50E-01	D-isomer specific 2-hydroxyacid dehydrogenases

a) Motif search was performed by HMMER2.1.1 against Pfam database (release 4.4). b) Function was classified based on the annotation of the Pfam entry which was hit in the query sequence. c) Only the entries possesing the expectation value (E-value) less than 1.0 were presented.

quencing of the major reverse transcription-coupled polymerase chain reaction (RT-PCR) products to precisely predict protein-coding sequences.[5] This examination revealed spurious interruptions in the following clones: ORFs in 7 clones (KIAA1403, KIAA1405, KIAA1409, KIAA1410, KIAA1415, KIAA1424 and KIAA1425) were found to carry single- or multiple-insertions most of which probably corresponded to intronic sequences; ORFs in 7 clones (KIAA1411, KIAA1412, KIAA1413, KIAA1416, KIAA1418, KIAA1420 and KIAA1421) were frame-shifted by single- or double-short insertions or single-deletion (< 5 nucleotide residues); ORFs in 4 clones (KIAA1404, KIAA1408, KIAA1417 and KIAA1423) were found to carry single- or double-deletions; ORFs in 4 clones (KIAA1406, KIAA1407, KIAA1414 and KIAA1422) were divided into some portions by a combination of spurious interruptions including insertions/deletions; KIAA1419 carried a nonsense mutation in the ORF. 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 analyses in this study including prediction of their protein-coding sequences unless otherwise stated. The results of the comparison between the cloned DNA and the revised DNA sequences are available through the World Wide Web site at http://www.kazusa.or.jp/huge. The actual primer sequences and the PCR conditions used for the RT-PCR experiment are accessible through the web site at http://www.kazusa.or.jp/~hirosawa/interruption/entrance.html. Notably, clones for eight genes (KIAA1297, KIAA1395, KIAA1398, KIAA1410,

Table 3. Homologues of the newly identified genes found in various databases.[a]

Database[b]	New gene	aa. res.[c]	ID in database	aa. res.	% Identity	%coverage[d]	Comment[e]
HUGE and new genes	KIAA1294	1051	KIAA1013	1062	51	90	
	KIAA1301	1581	KIAA0322	1562	56	98	
	KIAA1304	1051	KIAA0456	1095	68	99	
	KIAA1306	1154	KIAA1139	1124	34	100	
	KIAA1309	639	KIAA1354	632	92	97	
		639	KIAA1129	625	30	86	
	KIAA1316	1590	KIAA1414	1586	56	98	
	KIAA1346	999	KIAA0688	849	49	81	
	KIAA1347	918	KIAA0703	1051	68	96	
	KIAA1349	752	KIAA1141	914	45	81	
	KIAA1354	632	KIAA1129	625	30	86	
	KIAA1361	1005	KIAA0881	1064	70	100	
	KIAA1366	550	KIAA0951	679	61	100	
	KIAA1378	451	KIAA0795	465	35	96	
	KIAA1396	551	KIAA0798	682	50	88	
	KIAA1431	891	KIAA0065	848	41	86	
	KIAA1441	1258	KIAA0211	1317	34	92	
yeast	KIAA1347	918	SW-ATC1_YEAST	950	50	95	Ca2+-transporting ATPase (EC 3.6.1.38)
	KIAA1352	1212	SW-SYLC_YEAST	1090	46	84	leucyl-tRNA synthetase, cytoplasmic (EC 6.1.1.4)
	KIAA1401	853	S67595	788	30	90	hypothetical protein YDL060w
C.elegans	KIAA1347	918	ZK256.1a	922	59	94	CECC42
			K11D9.2b	1004	36	81	CEK11D91
			K11D9.2a	1059	36	81	CEK11D92
			B0365.3	996	30	85	CEB03652
			C01G12.8	1049	30	82	CEC01G126
	KIAA1352	1212	R74.1	1186	56	97	probable leucyl-tRNA synthetase (EC 6.1.1.4)
	KIAA1361	1005	T17E9.1	982	37	89	serine/threonine-protein kinase sulu (EC 2.7.1.-)
	KIAA1374	764	F38G1.1	759	41	95	che-2 protein
	KIAA1378	451	R12E2.1	531	39	81	CELR12E214
	KIAA1401	853	F10G7.1	785	39	91	CELF10G79
	KIAA1422	1151	F08B12.3b	1107	46	83	CEF08B122
			F08B12.3a	1119	46	83	CEF08B123
	KIAA1434	677	T05H10.7	796	33	90	hypothetical 90.8 kd protein t05h10.7 in chromosome II
			K10B3.6	757	30	93	CELK10B31
	KIAA1435	415	D2013.2	415	40	95	hypothetical 46.2 kd trp-asp repeats containing protein d2013.2 in chromosome II
OWL	KIAA1296	815	AF078667	714	82	96	ponsin-1, complete cds. - mouse
	KIAA1299	730	JC5887	670	93	92	signaling mediator variant - mouse
	KIAA1301	1581	KIAA0322	1562	56	98	KIAA0322, partial cds. - human
	KIAA1303	1119	SPAC57A710	1313	38	98	S.pombe chromosome I cosmid c57A7. - fission yeast
	KIAA1304	1051	KIAA0456	1095	68	99	KIAA0456, partial cds. - human
	KIAA1309	639	AF059569	593	30	85	actin binding protein MAYVEN, complete cds. - human
	KIAA1327	1310	T03730	1567	61	100	antigen containing epitope to monoclonal antibody MMS-85/12 - mouse
	KIAA1341	620	S35532	729	45	89	hnRNA-binding protein M4 - human
	KIAA1342	426	SYT4_RAT	425	90	100	synaptotagmin IV - rat
	KIAA1346	999	T00017	951	82	95	ADAMTS-1 protein - mouse
	KIAA1347	918	A42764	919	97	100	Ca2+-transporting ATPase (EC 3.6.1.38) - rat
	KIAA1348	545	AF062741	530	84	97	pyruvate dehydrogenase phosphatase isoenzyme 2, complete cds. - rat
	KIAA1349	752	ZN43_HUMAN	803	54	84	zinc finger protein 43 - human
	KIAA1352	1212	SYLC_CAEEL	1198	56	97	probableleucyl-tRNA synthetase (EC 6.1.1.4) - *C. elegans*
	KIAA1354	632	AF059569	593	30	86	actin binding protein MAYVEN, complete cds. - human
	KIAA1356	519	CIN1_HUMAN	423	91	81	sodium channel protein, brain I alpha subunit - human
	KIAA1361	1005	AF084205	1001	99	100	serine/threonine protein kinase TAO1, complete cds. - rat
	KIAA1363	430	A58922	398	43	94	esterase/N-deacetylase (EC 3.5.1.-), 50K hepatic - rabbit
	KIAA1368	1049	AF030430	888	93	84	semaphorin VIa, complete cds. - mouse
	KIAA1369	653	AF028808	619	83	95	hemin-sensitive initiation factor 2 alpha kinase, complete cds. - mouse
	KIAA1373	463	HSU73522	424	57	87	AMSH, complete cds. - human
	KIAA1374	764	CEL011523	760	41	98	CHE-2 protein. - *C. elegans*
	KIAA1376	437	S73591B	391	41	89	brain-expressed HHCPA78 homolog - human
	KIAA1378	451	KIAA0795	465	35	96	KIAA0795, partial cds. - human
	KIAA1379	434	AF104402	441	96	100	syndapin I, complete cds. - rat
	KIAA1381	961	AF109377	980	82	99	ldlBp (LDLB), complete cds. - mouse
	KIAA1382	462	HSU49082	504	57	98	transporter protein (g17), complete cds. - human
	KIAA1385	768	GEPH_RAT	736	100	96	gephyrin (putative glycine receptor-tublin linker protein) - rat
	KIAA1388	599	Z184_HUMAN	726	38	82	zinc finger protein 184 - human
	KIAA1389	1514	AF090989	1783	53	96	putative GAP protein alpha, complete cds. - human
	KIAA1393	500	JC4255	475	33	82	met-10+ protein - *Neurospora crassa*
	KIAA1396	551	Z135_HUMAN	469	59	83	zinc finger protein 135 - human
	KIAA1398	1456	A56734	1534	83	95	ribosome receptor, 180k - dog
	KIAA1400	1093	MMU88549	896	97	80	OL-protocadherin, complete cds. - mouse
	KIAA1401	853	CELF10G79	785	39	91	Caenorhabditis elegans cosmid F10G7. - *C. elegans*
	KIAA1422	1151	AF089730	1237	94	91	potassium channel subunit (Slack), complete cds. - rat
	KIAA1431	891	Z184_HUMAN	726	45	83	zinc finger protein 184 - human
	KIAA1433	652	AF053768	630	39	94	brain specific cortactin-binding protein CBP90, partial cds. - rat
	KIAA1434	677	YRS7_CAEEL	796	33	90	hypothetical 90.8 KD protein T05H10.7 in chromosome II - *C. elegans*
	KIAA1435	415	YLN2_CAEEL	415	40	95	hypothetical 46.2 kd trp-asp repeats containing protein d2013.2 in chromosome II - *C. elegans*
	KIAA1436	924	FPRP_RAT	879	89	95	prostaglandin F2-alpha receptor regulatory protein precursor - rat
	KIAA1439	561	NFIL_RAT	509	100	91	nuclear factor I (NF-I) - rat
	KIAA1441	1258	D86966	1267	34	92	KIAA0211, complete cds. - human
	KIAA1442	627	MMU92704	551	77	83	Olf-1/EBF-like-2(OS) transcription factor, complete cds. - mouse

a) The definition of homologues used here was the proteins found in the databases satisfying the following conditions: i) the length ranged from 80% to 125% of the query sequence; ii) the ratio of the length of aligned region to that of the original sequence of the query was 80% or greater; iii) percent identity was 30% or greater. The method of homology search was the same to that explained in Table 2-1. b) The following databases were used. HUGE, our cDNA-encoded protein database (http://www.kazusa.or.jp/huge); yeast, non redundant peptide database from genome-ftp.stanford.edu: /pub/yeast/yeast_protein/yeast_nrpep.fasta.Z; *C. elegans*, protein database deduced from *C. elegans full genome* sequence (ftp.sanger.ac.uk:/pub/databases/C.elegans_sequences/C_elegans_proteins_1998-10-16.pep) and the entries derived from *C. elegans* of OWL, and OWL (release 31.4). In the case of database search against OWL, only the homologue with the highest score to each query was listed. c) The number of amino acid residues of the gene produt. d) The values mean the ratio of the length of aligned region to the original length of the query sequence, in percentage. e) For entries from databases, yeast and OWL, the annotations were listed. For *C. elegans*, IDs of OWL were listed, when sequences identical to the entries from the full genome were registered in OWL.

KIAA1416, KIAA1420, KIAA1421 and KIAA1422) seemed to lack regions encoding C-terminal portions due to the presence of a *Not* I site in their coding regions because cDNAs were digested with *Not* I before ligation into vector. In contrast, clones for five genes (KIAA1439-KIAA1443) were found to lack 5′-portions of the sequences due to the presence of an internal *Not* I site in their sequences. For these five genes, the nucleotide sequences of only the region between two *Not* I sites were determined, since their original clones were most likely to harbor two intermolecularly ligated independent cDNAs.[6] After these revisions, the average size of the cDNA sequences became 4.8 kb and that of the ORFs corresponded to approximately 910 amino acid residues. Physical maps of the 150 cDNA sequences 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. Repeat sequences are also shown in Fig. 1. Comparing the predicted protein-coding sequence for KIAA1299 with those of mouse and rat homologues,[7,8] this cDNA clone seems to encode a complete protein although it possessed an unusually long 5′ non-coding sequence expanding more than 3 kb. Table 1 lists the lengths of inserts, the ORF lengths and the chromosomal locations of the respective clones. Chromosomal loci of 66 newly identified genes were assigned using human-rodent hybrid panels, GeneBridge 4 (Research Genetics Inc., USA),[9] since their mapping data were not available in the public databases. The chromosomal locations of the 78 genes, which are highlighted by asterisks in Table 1, were fetched from the UniGene database (http://www.ncbi.nlm.nih.gov/UniGene). The chromosomal locations of the remaining six genes, which are highlighted in Table 1, were obtained from the GenBank database because the sequences of the cDNA clones were already assigned to chromosome numbers.

2. Functional Classification of Predicted Gene Products

The gene products predicted from the cDNA sequences were classified by homology and/or motif search against the following public databases: protein sequence database, OWL (release 31.4),[10] databases of predicted protein sequences from yeast[11] and *C. elegans*[12] genomes [genome-ftp.stanford.edu:/pub/yeast/yeast_protein/ yeast_nrpep.fasta.Z, ftp.sanger.ac.uk:/pub/databases/C. elegans_sequences/C_elegans_proteins_1998-10-16.pep], protein domain database, Pfam (release 4.4),[13] and our own database, HUGE[14] (http://www.kazusa.or.jp/ huge). As shown in Table 2, the 73 gene products were classified into five functional categories. Among them, 53 gene products indicated significant sequence similarity to functionally annotated proteins (Table 2-1). The functions of the other 20 gene products were predicted based on the presence of functional motifs/domains, since they did not show sequence similarity to functionally annotated proteins (Table 2-2). In total, 63 gene products (86.3% of genes functionally annotated here) were suggested to have functions relating to cell signaling/communication, nucleic acid management or cell structure/motility. Of the 12 genes in functional class of nucleic acid management, 5 coded for DNA binding proteins carrying C_2H_2-type zinc finger domains. The average number of these domains among these gene products was about 15. Since the majority of zinc finger proteins in yeast contain only two domains per polypeptide, multiple appearance of C_2H_2-type zinc finger domains in a single polypeptide might be a specific character of large proteins in multicellular organisms. To find the genes conserved in other species, we tentatively defined "homologues" as genes sharing at least 30% of protein sequence identity spanning almost the entire region (more than 80% coverage against the query protein sequence). As shown in Table 3, 48 KIAA gene products were found to have the "homologues" in the databases. Homologues to 9 of the 48 KIAA proteins were found in *C. elegans* and 3 (KIAA1347, KIAA1352 and KIAA1401) were found in both yeast and *C. elegans.* KIAA1347 and KIAA1352 were similar to Ca^{2+}-transporting ATPase and leucyl-tRNA synthetase, respectively, though KIAA1401 had no similarity to any functionally known genes.

3. Expression Profiles of Predicted Genes

The expression profiles of the genes newly identified in this study are shown in Fig. 2 by using color codes.[15] KIAA1379 was homologous to rat synaptic dynamin-associated protein I (Syndapin I)[16] and predominantly expressed in hippocampus. The gene expression levels of KIAA1341 and KIAA1366, which were similar to mouse transcriptional suppressor of the myelin basic protein gene[17] and rat neuroligin 2,[18] respectively, were relatively high in all brain regions examined. KIAA1346 and KIAA1434 were predominantly expressed in spinal cord. KIAA1312, KIAA11315 and KIAA1417 were expressed very poorly in all regions examined, but their mRNAs were detected. These expression profiles also provide us important information for identifying biologically important genes characterized in this project.

Acknowledgements: 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, Kozue Kaneko, Chie Mori, Naoko Shibano, Hisako Takahashi, Satomi Nakagaki, and Kazutaka Mitsui for their technical assistance.

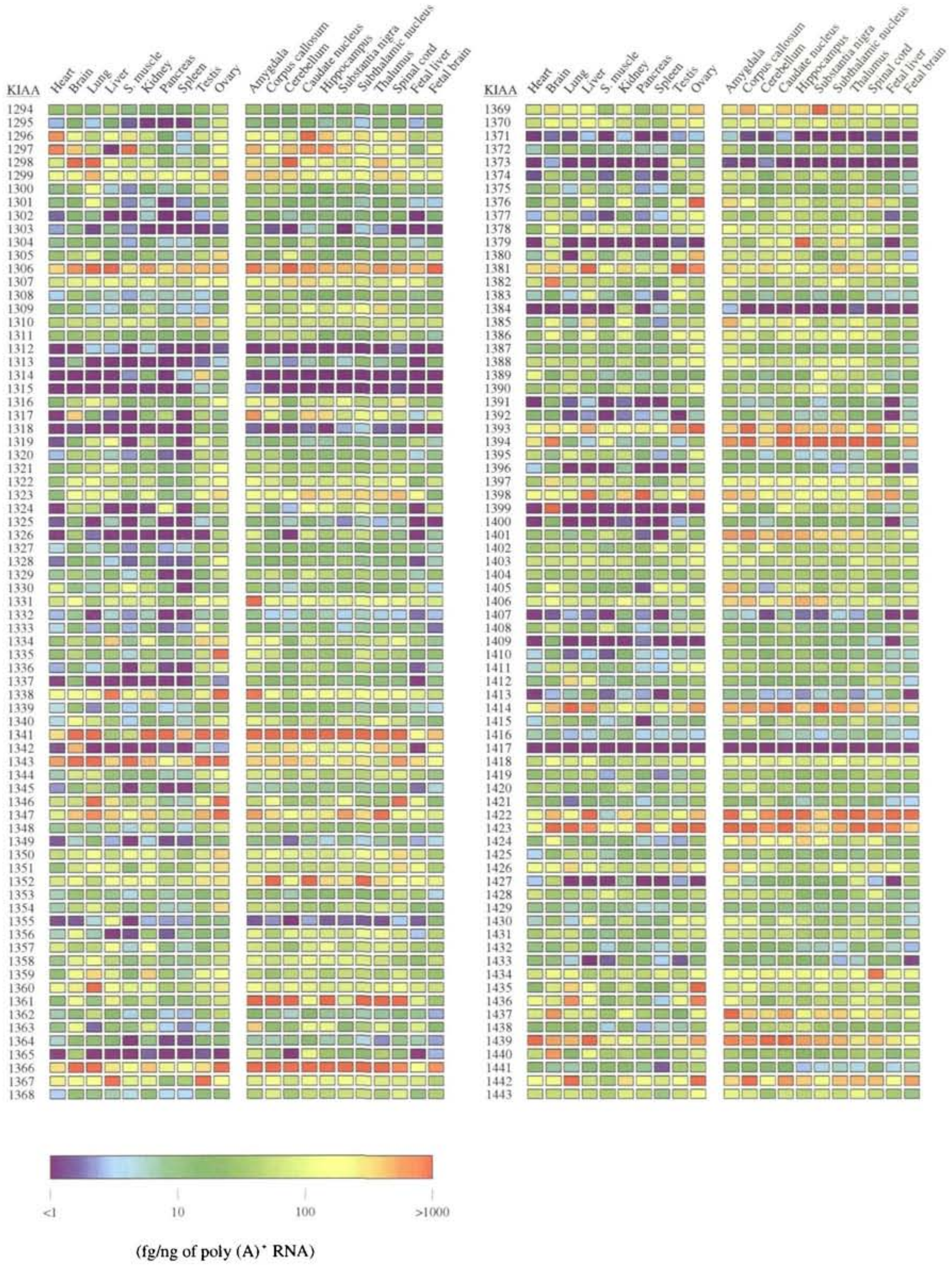

Figure 2. Expression profiles of 150 newly identified genes examined by RT-PCR ELISA. The tissue expression levels of the 150 human genes were analyzed by using the RT-PCR ELISA according to methods previously described.[15] Gene names are given as KIAA numbers at the left side of each set of color codes. Tissue and brain region names are indicated above the top sets of color codes. A color conversion panel shown at the bottom was used for displaying mRNA levels as color codes. The mRNA levels are expressed in equivalent amounts (fg) of the authentic cDNA plasmids in 1 ng of starting poly(A)$^+$ RNAs. Besides 10 tissues, 9 regions of the adult central nervous system (amygdala, corpus callosum, cerebellum, caudate nucleus, hippocampus, substantia nigra, subthalamic nucleus, thalamus, and spinal cord) and fetal brain were included in the expression profiling. As a control, mRNA levels in fetal liver were also examined.

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. Nagase, T., Ishikawa, K.-I., Kikuno, R. et al. 1999, Prediction of the coding sequences of unidentified human genes. XV. The complete sequences of 100 new cDNA clones from brain which code for large proteins *in vitro*, *DNA Res.*, **6**, 337–345.
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. Hirosawa, M., Nagase, T., Ishikawa, K.-I., Kikuno, R., Nomura, N., and Ohara, O. 1999, Characterization of cDNA clones selected by the GeneMark analysis from size- fractionated cDNA libraries from human brain, *DNA Res.*, **6**, 329–336.
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. Nagase, T., Ishikawa, K.-I., Miyajima, N. et al. 1998, Prediction of the coding sequences of unidentified human genes. IX. 100 new cDNA clones from brain which can code for large proteins *in vitro*, *DNA Res.*, **5**, 31–39.
7. Riedel, H., Wang, J., Hansen, H., and Yousaf, N. 1997, PSM, an insulin-dependent, pro-rich, PH, SH2 domain containing partner of the insulin receptor, *J. Biochem.*, **122**, 1105–1113.
8. Rui, L., Mathews, L. S., Hotta, K., Gustafson, T. A., and Carter-Su, C. 1997, Identification of SH2-Bbeta as a substrate of the tyrosine kinase JAK2 involved in growth hormone signaling, *Mol. Cell Biol.*, **17**, 6633–6644.
9. Gyapay, G., Schmitt, K., Fizames, C. et al. 1996, A radiation hybrid map of the human genome, *Hum. Mol. Genet.*, **5**, 339–346.
10. Bleasby, A. J., Akrigg, D., and Attwood, T. K. 1994, OWL – a non-redundant composite protein sequence database, *Nucleic Acids Res.*, **22**, 3574–3577.
11. Goffeau, A., Barrell, B. G., Bussey, H. et al. 1996, Life with 6000 genes, *Science*, **274**, 546–567.
12. The C. elegans Sequencing Consortium. 1998, Genome sequence of the nematode, *C. elegans*: A platform for investing biology, *Science*, **282**, 2012–2018.
13. Bateman, A., Birney, E., Durbin, R. et al. 1999, Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins, *Nucleic Acids Res.*, **27**, 260–262.
14. Kikuno, R., Nagase, T., Suyama, M., Waki, M., Hirosawa, M., and Ohara, O. 2000, HUGE: a database for human large proteins identified in Kazusa cDNA sequencing project, *Nucleic Acids Res.*, **28**, 331–332.
15. Nagase, T., Ishikawa, K.-I., Suyama, M. et al. 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*, *DNA Res.*, **5**, 277–276.
16. Qualmann, B., Roos, J., DiGregorio, P. J., and Kelly, R. B. 1999, Syndapin I, a synaptic dynamin-binding protein that associates with the neural Wiskott-Aldrich syndrome protein, *Mol. Biol. Cell*, **10**, 501–513.
17. Steplewski, A., Haas, S., Amini, S., and Khalili, K. 1995, Regulation of mouse myelin basic protein gene transcription by a sequence-specific single-stranded DNA-binding protein in vitro, *Gene*, **154**, 215–218.
18. Ichtchenko, K., Nguyen, T., and Sudhof, T. C. 1996, Structures, alternative splicing, and neurexin binding of multiple neuroligins, *J. Biol. Chem.*, **271**, 2676–2682.