Prediction of the Coding Sequences of Mouse Homologues of FLJ Genes: The Complete Nucleotide Sequences of 110 Mouse FLJ-Homologous cDNAs Identified by Screening of Terminal Sequences of cDNA Clones Randomly Sampled from Size-Fractionated Libraries

Noriko Okazaki,1 Reiko Kikuno,1 Reiko Ohara,1 Susumu Inamoto,2,3 Haruhiko Koseki,4,5 Shuichi Hiraoka,4 Yumiko Saga,6 Hiroshi Kitamura,5 Tomoko Nakagawa,5 Takahiro Nagase,1 Osamu Ohara,1,5 and Hisashi Koga1,3,*

Kazusa DNA Research Institute, 2-6-7 Kazusa-Kamatari, Kisarazu, Chiba 292-0818, Japan,1 Institute of Research and Innovation, 1201 Takada, Kashiwa, Chiba 277-0861, Japan,2 Chiba Industry Advancement Center, 2-6 Nakase, Mihama-ku, Chiba 261-7126, Japan,3 Department of Molecular Embryology, Graduate School of Medicine, Chiba University, 1-8-1 Inohana, Chuo-ku, Chiba 260-8670, Japan,4 RIKEN Research Center for Allergy and Immunology, 1-7-22 Suehiro, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan,5 and Division of Mammalian Development, National Institute of Genetics, Yata 1111, Mishima 411-8540, Japan6

(Received 30 January 2004)

Abstract

We have been conducting a mouse cDNA project to predict protein-coding sequences of mouse KIAA-homologous genes since 2001. As an extension of this project, we also started to accumulate mouse cDNA clones homologous to the human FLJ cDNA clones which are another long cDNA resource produced in our institute. We have isolated the cDNA clones from size-fractionated cDNA libraries derived from five different mouse tissues and natural killer T-cells. Although the human FLJ cDNA clones were originally derived from human spleen libraries, one-third of their mouse homologues were obtained from the brain library. We designated these homologues “mFLJ” plus a 5-digit number and herein characterized 110 mFLJ cDNA clones. We assigned an integrity of the CDSs from the comparison of the 110 cDNA clones with the corresponding human FLJ cDNA clones. The average size of the 110 mouse cDNA sequences was 3.8 kb and that of the deduced amino acid sequences from their longest CDS in each cDNA was 663 amino acid residues. Homology and/or motif search against public databases revealed new domains and/or motifs in 26 mFLJ gene products which provide additional speculation regarding the function of FLJ genes.

Key words: mFLJ; mouse; cDNA sequencing; large proteins; orthologue; novel genes; protein-coding region

Introduction

Highly evolved organisms seem to accumulate many large multi-domain proteins having several protein-protein interaction modules at the specific subcellular loci.1 These large proteins subsequently tend to construct big protein complexes and are thereby related to higher-order cell functions. Although the importance of the functional analysis of large proteins is widely accepted, the functions of many large proteins remain to be elucidated because the analyses are technically difficult. In this regard, we have achieved the technical breakthrough and have been conducting human cDNA sequencing projects to accumulate sequence information of long cDNAs (> 4 kb) encoding large proteins since 1994.2 At present, the total number of human cDNAs designated using “KIAA” plus a four-digit number and “FLJ” plus a five-digit number have exceeded 2000 and 300, respectively.3–5 We have deposited the sequence data of these clones in the DDBJ/GenBank/EMBL database, and the information regarding KIAA and FLJ genes has been shared with the research community through the HUGE (http://www.kazusa.or.jp/huge) and NEDO.
plasmic RNA Reagent (Invitrogen, USA). The cDNA clones to be entirely sequenced were selected according to the procedures previously described.\textsuperscript{10,13,14} After careful comparison with the sequence information already available in the public databases, we only deposited the 82 newly identified mouse cDNA sequences to the DDBJ/EMBL/GenBank databases (accession numbers are given in Table 1).\textsuperscript{10,13,14} Among the 110 cDNA clones reported here, 27 were derived from the adult brain library, 16 from the fetal brain library, 8 from the adult thymus library, 19 from the adult spleen library, 35 from the embryonic tail library and 5 from the NKT cell library. Although the human FLJ cDNA clones were originally derived from a human spleen library, one-third of their mouse homologues were obtained from the brain libraries. It is difficult to exclude the possibility that these clones are derived from the mRNA of circulating leukocytes in the brain vessels, however, these FLJ genes may also have some physiological functions even in the brain. We designated these mouse homologous cDNAs as “mFLJ” plus the same five-digit number assigned to the corresponding human FLJ cDNAs. The sequence features of the 110 cDNA clones are listed in Table 1 and their structural features are shown in Figs. 1–3 together with the corresponding human FLJ cDNA clones.

We first identified CDSs in the obtained cDNA sequences by GeneMark analysis.\textsuperscript{16} When the GeneMark analysis did not detect any reliable CDSs in the respective cDNA sequence, CDS was tentatively assigned as the longest open reading frame (ORF). The average size of the 110 mouse cDNA sequences was 3.8 kb and that of the deduced amino acid sequences from their longest CDS in each cDNA was 663 amino acid residues. Multiple CDSs were predicted in 32 mouse cDNA sequences (Fig. 2), whereas the remaining 78 cDNAs carried single CDSs, as predicted by GeneMark analysis (Fig. 1). Such multiple predicted CDSs in a single cDNA sequence are most likely the result of spurious CDS splits caused by errors of reverse transcription, retained intron(s), or other cloning artifacts.\textsuperscript{17} Further confirmatory experiments are required to determine whether these predicted CDS interruptions are spurious or not. However, if a CDS split was predicted only in mouse cDNA clone and the corresponding region was assigned to encode a continuous single CDS in human cDNA, we considered it very likely that the predicted CDS interruption in the mouse cDNA clone was spurious. Using this assumption, 13 out of 32 mFLJ cDNAs were evaluated that they were containing spurious CDS interruption(s). These spurious CDS interruptions were classified into the following two categories: 1) 6 mFLJ cDNA clones appeared to contain frame-shift errors (mFLJ00021, mFLJ00040, mFLJ00182, mFLJ00226, mFLJ00246 and mFLJ00385); and 2) 8 mFLJ cDNA clones retained intron(s) (mFLJ00002, mFLJ00068, mFLJ00070, mFLJ00153, mFLJ00193, mFLJ00217, mFLJ00307 and mFLJ00376).
| Gene name | Accession numbers | cDNA length (bp) | CDS length (a.a.) | Chromosomal number |
|-----------|-------------------|-----------------|-------------------|--------------------|
| mFLJ0002  | AK131108          | 5652            | 1487              | 17                 |
| mFLJ0006  | BC018197*         | 2725            | 766               | 15                 |
| mFLJ0007  | BC057367*         | 4223            | 687               | 18                 |
| mFLJ0012  | AK131109          | 6360            | 454               | 7                  |
| mFLJ0018  | AK131110          | 4880            | 1372              | 7                  |
| mFLJ0019  | AK131111          | 3834            | 962               | 7                  |
| mFLJ0021  | AK131112          | 4329            | 460               | 11                 |
| mFLJ0022  | AK131113          | 4516            | 340               | 2                  |
| mFLJ0023  | AK131114          | 5992            | 164               | 6                  |
| mFLJ0024  | AK079514*         | 3991            | 297               | 2                  |
| mFLJ0025  | AK131115          | 3986            | 1204              | 6                  |
| mFLJ0034  | BC042576*         | 4000            | 1255              | 7                  |
| mFLJ0037  | AK131116          | 4464            | 136               | 15                 |
| mFLJ0040  | AK131117          | 4355            | 638               | 7                  |
| mFLJ0041  | AK131118          | 4759            | 337               | 17                 |
| mFLJ0043  | AF305089*         | 1575            | 431               | 19                 |
| mFLJ0044  | AK131119          | 3186            | 733               | 8                  |
| mFLJ0045  | AK131120          | 4162            | 612               | 1                  |
| mFLJ0052  | AK045475*         | 1591            | 364               | 5                  |
| mFLJ0055  | AK131121          | 1962            | 442               | 2                  |
| mFLJ0057  | AK131122          | 4547            | 1081              | 10                 |
| mFLJ0062  | AK131123          | 5179            | 958               | 15                 |
| mFLJ0064  | AK131124          | 2033            | 401               | 2                  |
| mFLJ0065  | AK040622*         | 4554            | 252               | 2                  |
| mFLJ0068  | AK131125          | 5557            | 998               | 8                  |
| mFLJ0069  | AK0797947*        | 3016            | 968               | 17                 |
| mFLJ0070  | AK131126          | 5729            | 1166              | 19                 |
| mFLJ0085  | AK077341*         | 1485            | 479               | 1                  |
| mFLJ0086  | AK131127          | 3225            | 888               | 17                 |
| mFLJ0087  | AK131128          | 3219            | 981               | 17                 |
| mFLJ0098  | AK131129          | 5082            | 1091              | 13                 |
| mFLJ0107  | AK131130          | 5034            | 1371              | 18                 |
| mFLJ0108  | AK131131          | 1404            | 289               | 5                  |
| mFLJ0110  | AK131132          | 1687            | 244               | 5                  |
| mFLJ0111  | BC046981*         | 3459            | 442               | 4                  |
| mFLJ0114  | AK131133          | 4031            | 1188              | 7                  |
| mFLJ0120  | AK131134          | 4112            | 1173              | 5                  |
| mFLJ0123  | BC041780*         | 2315            | 490               | 11                 |
| mFLJ0124  | AF217090*         | 973             | 285               | 7                  |
| mFLJ0127  | AK131135          | 2066            | 641               | 8                  |
| mFLJ0128  | BC014817          | 4844            | 1517              | 15                 |
| mFLJ0132  | BC023900*         | 2654            | 870               | 19                 |
| mFLJ0136  | YX263158*         | 2836            | 857               | 11                 |
| mFLJ0137  | AK131137          | 4896            | 1396              | 11                 |
| mFLJ0138  | AK034662*         | 2345            | 449               | 3                  |
| mFLJ0144  | AK004778*         | 2811            | 413               | 6                  |
| mFLJ0147  | AK131138          | 4215            | 982               | 7                  |
| mFLJ0150  | AK131139          | 3974            | 1174              | 2                  |
| mFLJ0153  | AK131140          | 4632            | 504               | 9                  |
| mFLJ0157  | AK131141          | 4582            | 770               | 13                 |
| mFLJ0158  | AK131142          | 911             | 236               | X                  |
| mFLJ0161  | AK050100*         | 835             | 265               | 8                  |
| mFLJ0164  | AK131143          | 3411            | 1012              | 5                  |
| mFLJ0169  | AK035383*         | 1181            | 233               | 5                  |
| mFLJ0175  | AK131144          | 6660            | 2118              | 6                  |

a) Accession numbers of DDBJ, EMBL, and GenBank databases. The accession numbers with asterisk are those deposited by other groups because these cDNAs encoded the identical proteins to those reported previously.
b) Values excluding poly(A) sequences.
c) Deduced amino acids length of the CDS in each cDNA. CDSs were identified according to the result of GeneMark analysis. When multiple CDSs were predicted, the sum total value of the length of multiple CDSs more than 50 amino acids residues were shown.
d) Chromosome numbers were determined from the results of BLAT2 search of cDNA clones against the mouse draft sequence (ftp://ftp.ensembl.org/pub/mouse-7.3a/data/golden_path/).11
Figure 1. Schematic comparison of structures of mFLJ and FLJ cDNAs which had a single predicted CDS. A shotgun method was applied to determine the entire sequences of cDNA clones according to the method previously reported. The mFLJ gene numbers corresponding to respective cDNAs are given on the left. The horizontal scales represent the cDNA length in kb. The predicted CDSs and untranslated regions are shown by dark blue and open boxes, respectively. The positions of the first ATG codon with or without the contexts of Kozak’s rule are illustrated by solid and open triangles, respectively. SINEs and other repetitive sequences are displayed by dotted and hatched boxes, respectively. The start and end points of the aligned region(s) between mFLJ and FLJ CDSs are connected by thin lines on the basis of results of FASTA search of amino acid sequences predicted from corresponding CDSs. A canonical polyadenylation signal sequence (AATAAA) and its single-nucleotide variants are represented by red lines when they existed in the same or close (within 5 bases) positions on the aligned sequences between mFLJ and FLJ cDNAs and at least one of them was in the 35-bp upstream region of the 3′-extreme end. If either the canonical polyadenylation signal sequence or its single-nucleotide variants were found upstream from the possible polyadenylation signal sequence thus detected by the alignment of mFLJ and FLJ cDNA sequences, these upstream latent polyadenylation signals are represented by orange lines as an indication that these hexamer sequences were likely to be for alternative polyadenylation. On the other hand, the polyadenylation signal hexamer sequences found at the 3′-end are represented by green lines when their positions differed by more than 5 bases on the aligned sequences of mFLJ and FLJ cDNAs or they were present only in one of the mFLJ and FLJ cDNAs. The sequence alignments of mouse and human FLJ cDNAs represented in this report are available through our web site (http://www.kazusa.or.jp/rouge/).
(Fig. 2). These two categories of CDS interruptions in the cDNAs are shown in Fig. 2 using the symbols $ and # for categories 1 and 2, respectively. Most of the frame-shift errors were caused by a one- or two-nucleotide insertion/deletion which is frequently found in regions with homopolymeric runs and is most likely due to errors in reverse transcription. In these cases (frame-shift errors and retained introns) we have deposited the hypothetically joined CDS sequence of these clones as their valid protein in the DDBJ/GenBank/EMBL database and our ROUGE database. When the hypothetically joined CDSs in mFLJ and mKIAA cDNAs are deposited, their features are described in the DDBJ/GenBank/EMBL database as follows; molecular type = “pre-RNA”, CDS join (the start and end positions of the all joined coding sequence). In the ROUGE database, the start and end positions of the all joined CDSs in the cDNA are also enumerated. If we combined multiple CDSs of more than 50 amino acid residues in each cDNA clone with the longest CDS, the average size of the deduced amino acid
sequences of 110 mFLJ cDNAs became 708 amino acid residues. When the CDS interruptions were observed in both mouse and human cDNA clones, we could not confidently conclude the authenticity of the CDS interruption in these mouse FLJ cDNA clones without any other lines of evidence.

2. Chromosomal Loci of mKIAA Genes

The currently available draft sequence of the mouse genome (ftp://ftp.ensembl.org/pub/mouse-7.3a/data/golden_path/),\(^7\) enabled us to predict the genomic structures of mouse genes by comparing the cDNA sequences with their corresponding genomic sequence.\(^{13,14}\) To assign the chromosomal localization of genes generating the 110 cDNAs identified in this study, the cDNA se-
sequences were subjected to BLAST search against the mouse genome draft sequence, and the genome sequences that satisfied either of the following two conditions were selected: E-value = 0.0 and sequence identity is 90% or greater; or E-value $\leq 1e^{-10}$ and sequence identity is 99% or greater. Then they are aligned with the cDNA sequences by SIM4.25 As shown in Table 1, we could successfully map all 110 mouse FLJ cDNAs on the genome under these conditions.

3. Relationships of mFLJ cDNAs Clones with the Corresponding FLJ and its Redundant FLJ cDNA Clones

In our previous study, we already mentioned that some human FLJ genes were found to be redundant whereas the clone selection was done based on the different end sequence information.4,5 However each representative clone among the redundant human FLJ clones was chosen based on the length and continuity of the CDS. We also noted that these redundant clones do not always have identical structures in the overlapping regions with their corresponding clones. To identify the functional difference among these redundant clones, it is helpful to analyze the expression of each clone in different organs and/or species. In this regard, we compared the structures of mouse mFLJ cDNAs clones with human FLJ cDNAs and their redundant clones (Fig. 3). FLJ00001 is a redundant clone of FLJ00022 and FLJ00031, however the predicted CDS in FLJ00001 could not observed either in FLJ00022 or in mFLJ00031 and no homologous protein sequences have been registered in the public databases. Furthermore, mFLJ00022 was derived from the fetal brain library, thus the encoded 155 amino acids of FLJ00001 may have a specific role in human spleen. Alternatively, we could not exclude the possibility that the predicted CDS in FLJ00001 was falsely predicted in silico. In this regard, further experimental validation might answer this question.

![Figure 3](https://example.com/figure3.png)

**Figure 3.** The relationship of mFLJ cDNAs with the corresponding FLJ cDNAs and their redundant clones. The structures of mFLJ cDNAs with the corresponding FLJ cDNAs are shown together with their redundant clones, which are illustrated according to the formula described in the legend of Fig. 1. The FLJ and the mFLJ numbers are given on the left side of the schematic illustrations of the cDNA structures.
Table 2. List of the domains and the motifs that were newly identified in mFLJ gene products.\(^a\)

| Function                         | Gene product | Pfam ID    | E-value\(^c\) | Definition                                      |
|----------------------------------|--------------|------------|---------------|------------------------------------------------|
| Cell structure/motility          | FLJ00114     | PF00092    | 1.90E-57      | von Willebrand factor type A domain             |
|                                  | PF01839      | 8.10E-06   |               | FG-GAP repeat                                   |
|                                  | PF01839      | 1.90E-05   |               | FG-GAP repeat                                   |
|                                  | PF01839      | 2.20E-05   |               | FG-GAP repeat                                   |
|                                  | FLJ00150     | PF00038    | 0.075         | Intermediate filament protein                   |
|                                  | FLJ00205     | PF00652    | 0.00068       | QXW lectin repeat = ricin B lectin domain       |
|                                  | PF00652      | 0.17       |               | QXW lectin repeat                               |
|                                  | FLJ00238     | PF01576    | 0.0073        | Myosin tail                                     |
|                                  | FLJ00279     | PF03148    | 0.2           | Tektin family                                   |
|                                  | FLJ00307     | PF02318    | 0.0058        | Rabphilin-3A effector domain                    |
|                                  |              |            |               |                                                 |
|                                  | FLJ00006     | PF02759    | 3.00E-57      | RUN domain                                      |
|                                  | PF00018      | 3.90E-15   |               | SH3 domain                                      |
|                                  | FLJ00123     | PF00522    | 1.90E-19      | SPRY domain                                     |
|                                  | FLJ00189     | PF02204    | 2.30E-06      | Vacular sorting protein 9 (VPS9) domain         |
|                                  | FLJ00207     | PF00501    | 0.0033        | AMP-binding enzyme                              |
|                                  | FLJ00227     | PF00168    | 1.20E-26      | C2 domain                                       |
|                                  | PF00168      | 6.40E-16   |               | C2 domain                                       |
|                                  | FLJ00238     | PF02759    | 0.053         | RUN domain                                      |
|                                  | FLJ00248     | PF00560    | 0.96          | Leucine Rich Repeat                             |
|                                  | PF00560      | 0.62       |               | Leucine Rich Repeat                             |
|                                  | FLJ00279     | PF00612    | 0.0017        | IQ calmodulin-binding motif                     |
|                                  | FLJ00332     | PF00566    | 1.60E-27      | TBC domain                                      |
|                                  | FLJ00369     | PF00621    | 1.70E-19      | RhoGEF domain                                   |
|                                  | PF00169      | 0.28       |               | PH domain                                       |
|                                  | FLJ00395     | PF00018    | 1.40E-18      | SH3 domain                                      |
| Nucleic acid management          | FLJ00019     | PF00651    | 0.072         | BTB/POZ domain                                  |
|                                  | FLJ00086     | PF00097    | 0.032         | Zinc finger, CCHC4 type (RING finger)           |
|                                  | FLJ00127     | PF00651    | 1.20E-35      | BTB/POZ domain                                  |
|                                  | FLJ00191     | PF00249    | 8.60E-08      | Myb-like DNA-binding domain                     |
|                                  | PF00569      | 0.0031     |               | Zinc finger, ZZ type                            |
|                                  | FLJ00216     | PF01704    | 2.90E-47      | UTP-glucose-1-phosphate uridylytransferase      |
|                                  | FLJ00238     | PF01363    | 1.60E-11      | FYVE zinc finger                               |
|                                  | FLJ00307     | PF00097    | 0.91          | Zinc finger, CCHC4 type (RING finger)           |
| Protein management              | FLJ00069     | PF00004    | 1.30E-05      | ATPase family associated with various cellular activities (AAA) |
|                                  | FLJ00147     | PF00443    | 0.012         | Ubiquitin carboxyl-terminal hydrolase family 2  |
|                                  | FLJ00334     | PF00326    | 1.50E-07      | Prolyl oligopeptidase family                    |
|                                  | FLJ00373     | PF00004    | 1.00E-44      | ATPase family associated with various cellular activities (AAA) |
|                                  | PF01202      | 0.075      |               | Shikimate kinase                                |
|                                  | FLJ00394     | PF00789    | 0.035         | UBX domain                                     |
| Metabolism                       | FLJ00202     | PF01553    | 0.00083       | Aciyltransferase                                |

\(^a\) Motif search was performed by HMMER2.1.1 against Pfam database (release 9.0).

\(^b\) Function was classified based on the annotation of the Pfam entry which was hit in the query sequence.

\(^c\) Only the entries possessing the expectation value (E-value) less than 1.0 were presented.

4. Functional Annotation of FLJ Genes from Newly Identified Domains and Motifs in mFLJ cDNA Clones

In the previous studies, human FLJ gene products were functionally classified by homology and/or motif search against public databases such as a non-redundant amino acid sequence database; databases of predicted protein sequences from the genome sequences of yeast, nematoda and the fruit fly; a protein domain database; and our own database.\(^3\) According to this strategy, we performed motif searches of 110 mFLJ gene products against a protein domain database (Pfam, release 9.0).\(^18\) Since 44 mFLJ cDNAs contain CDSs that are longer than or cannot be aligned with their corresponding human FLJ cDNAs, we expected to find new functional annotation to these mFLJ genes. We newly found the domains and/or motifs in 26 mFLJ gene products, and list them in Table 2. For example, we found a BTB/POZ domain at the C-terminal region of the mFLJ00019 gene product; mFLJ00019 protein may produce a homomeric or heteromeric dimer and function to mediate transcriptional repression and to interact with components of histone deacetylase co-repressor complexes including nuclear receptor co-repressor N-CoR and SMRT. This kind of comparative analysis of mouse and human gene products is promising to provide further functional insight to these hypothetical genes.

Acknowledgements: We are grateful to Dr. Masaru Taniguchi (RIKEN Research Center for Allergy and Immunology and Department of Molecular Immunology, Graduate School of Medicine, Chiba University) and his colleagues for kindly providing the purified NKT cells. We thank Hiroshi Kohga for construction of a computer system for experimental management and we also thank Tomomi Kato, Tomomi Tajino, Keishi Ozawa, Kazuhiro Sato, Akiko Ando, Takashi Watanabe, Kiyoe Sumi, Kyoko Watanabe, Hiroko Kinoshita, Noriko...
Utsumi, Nobue Kashima and Masaki Takazawa for their technical assistance. This study was supported by the following grants: the CREATE Program (Collaboration of Regional Entities for the Advancement of Technological Excellent) from JST (Japan Science and Technology Corporation); a grant from Special Coordination Funds and a grant from Organized Research Combination System of the Ministry of Education, Culture, Sports, Science and Technology, the Japanese Government; and a grant from the Kazusa DNA Research Institute.

References

1. Apic, G., Huber, W., and Teichmann, S. A. 2003, Multi-domain protein families and domain pairs: comparison with known structures and a random model of domain recombination, J. Struct. Funct. Genomics, 4, 67–78.

2. Nomura, N., Miyajima, N., Suzuka, 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. Nagase, T., Kikuno, R., and Ohara, O. 2001, Prediction of the coding sequences of unidentified human genes. XXII. The complete sequences of 50 new cDNA clones which code for large proteins, DNA Res., 8, 319–327.

4. Hattori, A., Okumura, K., Nagase, T., Kikuno, R., Hirosawa, M., and Ohara, O. 2000, Characterization of long cDNA clones from human adult spleen, DNA Res., 7, 357–366.

5. Jikuya, H., Takano, J., Kikuno, R. et al. 2003, Characterization of long cDNA clones from human adult spleen. II. The complete sequences of 81 cDNA clones, DNA Res., 10, 49–57.

6. Kikuno, R., Nagase, T., Nakayama, M. et al. 2004, HUGE: a database for human KIAA proteins, a 2004 update integrating HUGEepi and ROUGE, Nucleic Acids Res., 32, D502–D504.

7. Waterston, R. H., Lindblad-Toh, K., Birney, E. et al. 2002, Initial sequencing and comparative analysis of the mouse genome, Nature, 420, 520–562.

8. Straussberg, R. L., Feingold, E. A., Grouse, L. H. et al. 2002, Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences, Proc. Natl. Acad. Sci. U.S.A., 99, 16899–16903.

9. VanBuren, V., Piao, Y., Dudekula, D. B. et al. 2002, Assembly, verification, and initial annotation of the NIA mouse 7.4K cDNA clone set, Genome Res., 12, 1999–2003.

10. Okazaki, Y., Furuno, M., Kasukawa, T. et al. 2002, Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs, Nature, 420, 563–573.

11. Koga, H., Shimada, K., Har, Y. et al. 2004, A comprehensive approach for establishment of the platform to analyze functions of KIAA proteins: Generation and evaluation of anti-mKIAA antibodies, Proteomics, in press.

12. Har, Y., Shimada, K., Kohga, H., Ohara, O., and Koga, H. 2003, High-throughput production of recombinant antigens for mouse KIAA proteins in Escherichia coli: computational allocation of possible antigenic regions, and construction of expression plasmids of glutathione-S-transferase-fused antigens by an in vitro recombination-assisted method, DNA Res., 10, 129–136.

13. Okazaki, N., Kikuno, R., Ohara, R. et al. 2002, Prediction of the coding sequences of mouse homologues of KIAA gene: I. The complete nucleotide sequences of 100 mouse KIAA-homologous cDNAs identified by screening of terminal sequences of cDNA clones randomly sampled from size-fractionated libraries, DNA Res., 9, 179–188.

14. Okazaki, N., Kikuno, R., Ohara, R. et al. 2003, Prediction of the coding sequences of mouse homologues of KIAA gene: II. The complete nucleotide sequences of 400 mouse KIAA-homologous cDNAs identified by screening of terminal sequences of cDNA clones randomly sampled from size-fractionated libraries, DNA Res., 10, 35–48.

15. Ohara, O., Nagase, T., Mitsui, G. et al. 2002, Characterization of size-fractionated cDNA libraries generated by the in vitro recombination-assisted method, DNA Res., 9, 47–57.

16. Borodovsky, M., McIninch, J. D., Koonin, E. V., Rudd, K. E., Medigue, C., and Danchin, A. 1995, Detection of new genes in a bacterial genome using Markov models for three gene classes, Nucleic Acids Res., 23, 3554–3562.

17. Hirosawa, M., Ishikawa, K., Nagase, T., and Ohara, O. 2000, Detection of spurious interruptions of protein-coding regions in cloned cDNA sequences by GeneMark analysis, Genome Res., 10, 1333–1341.

18. Bateman, A., Coin, L., Durbin, R. et al. 2004, The Pfam protein families database, Nucleic Acids Res., 32, D138–D141.

19. Ahmad, K. F., Engel, C. K., and Prive, G. G. 1998, Crystal structure of the BTB domain from PLZF, Proc. Natl. Acad. Sci. U.S.A., 95, 12123–12128.

20. Yoon, H. G., Chan, D. W., Reynolds, A. B., Qin, J., and Wong, J. 2003, N-CoR mediates DNA methylation-dependent repression through a methyl CpG binding protein Kaiso, Mol. Cell., 12, 723–734.

21. Kozak, M. 1996, Interpreting cDNA sequences: some insights from studies on translation, Mamm. Genome, 7, 563–574.

22. Brenner, S. E., Chothia, C., and Hubbard, T. J. 1998, Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. PG - 6073-8, Proc. Natl. Acad. Sci. U.S.A., 95, 6073–6078.