Construction of Expression-ready cDNA Clones for KIAA Genes: Manual Curation of 330 KIAA cDNA Clones

Daisuke Nakajima, Noriko Okazaki, Hisashi Yamakawa, Reiko Kikuno, Osamu Ohara, and Takahiro Nagase*

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

(Received 13 June 2002)

Abstract

We have accumulated information on protein-coding sequences of uncharacterized human genes, which are known as KIAA genes, through cDNA sequencing. For comprehensive functional analysis of the KIAA genes, it is necessary to prepare a set of cDNA clones which direct the synthesis of functional KIAA gene products. However, since the KIAA cDNAs were derived from long mRNAs (> 4 kb), it was not expected that all of them were full-length. Thus, as the first step toward preparing these clones, we evaluated the integrity of protein-coding sequences of KIAA cDNA clones through comparison with homologous protein entries in the public database. As a result, 1141 KIAA cDNAs had at least one homologous entry in the database, and 619 of them (54%) were found to be truncated at the 5′ and/or 3′ ends. In this study, 290 KIAA cDNA clones were tailored to be full-length or have considerably longer sequences than the original clones by isolating additional cDNA clones and/or connected parts of additional cDNAs or PCR products of the missing portion to the original cDNA clone. Consequently, 265, 8, and 17 predicted CDSs of KIAA cDNA clones were increased in the amino-, carboxy-, and both terminal sequences, respectively. In addition, 40 cDNA clones were modified to remove spurious interruption of protein-coding sequences. The total length of the resultant extensions at amino- and carboxy-terminals of KIAA gene products reached 97,000 and 7216 amino acid residues, respectively, and various protein domains were found in these extended portions.

Key words: large proteins; cDNA sequencing; manual curation; protein production

Even though the human genome draft sequence is now available, it is still difficult to obtain a complete catalog of human proteins for further comprehensive analysis of gene functions. A collection of cDNA clones entirely sequenced serves as an important genomic resource for functional analysis of human genes and provides valuable information to complement the draft genome sequence. Therefore, we have been conducting a human cDNA sequencing project to accumulate information regarding the protein-coding sequences (CDSs) of unidentified genes since 1994. In particular, we have focused our sequencing efforts on long cDNAs (> 4 kb) encoding large proteins (> 400 amino acid residues) because we are most interested in cDNA clones encoding multidomain proteins. We have already reported the cDNA sequences of more than 2000 genes to date and systematically designate them using "KIAA" plus a four-digit number. The average length of the cDNA inserts and gene products deduced from the cDNAs is 4.8 kb and 866 amino acid residues, respectively.

In our project, entire cDNA sequencing analysis is done on a single-clone-per-single-gene basis due to limited sequencing capacity. Under this situation, all the analyzed cDNA clones are not necessarily expected to be a perfect copy of major mRNA species; some cDNA clones might be copies of immature or non-functional mRNA or partial copies of the mRNAs, or they might contain some artifacts generated during reverse transcription and cloning. This possibility is highly problematic particularly when we consider a set of cDNA clones as reagents for functional analysis of genes. Thus, we previously developed a computer-based approach to detect spurious interruption of protein-coding regions in cloned cDNA sequences. Once a potential spurious interruption site is detected by computer analysis, we experimentally examined whether or not the observed interruption was present in the major form of the transcripts of interest. However, it is still difficult to evaluate conclusively the integrity of cDNA sequences at the 5′- and 3′-ends from the sequence information alone. In particular, the evaluation of the integrity of cDNA sequences at the 5′-end is more difficult than...
that at the 3′-end. Although there are many methods
that purport to isolate full-length cDNAs, none of them
can guarantee that all the clones in their cDNA libraries
harbor copies of the complete mRNAs. In fact, the rate of
occurrence of full-length cDNA clones has been described
to be only 40–60% in their reports.7,8 Thus, unless a
new efficient experimental method emerges for identi-
fying transcription start sites of human genes, it is almost
impossible to experimentally evaluate the integrity of the
5′-end of cDNAs in a high-throughput manner. However,
because our goal is to evaluate whether or not a cDNA
clone can direct the synthesis of functional gene products
of interest, one practical alternative for us is to examine
whether a cDNA clone contains a translation, not tran-
scription, start site or not. Because CDSs are well con-
served in homologues in a close evolutionary relationship,
comparison of the homologous protein sequences gives
helps to identify the translation start site. This in silico
method enables us to examine the integrity of CDS en-
coded by KIAA cDNAs in a high-throughput manner.

As we enter the endgame of discovery of human genes,
preparation of a set of expression-ready cDNA clones be-
comes more urgent than before. In general, long cDNAs
are difficult molecules to prepare in an expression-ready
form. Thus, once we could successfully prepare a set
of KIAA cDNAs which produce functional KIAA gene
products, this set must serve as a highly useful genomic
resource for the research community. Based on the eval-
uation results of the integrity of CDSs in KIAA cDNAs, we
thus tailored the cloned KIAA cDNAs to be full-length,
to have longer CDSs, or to remove spurious CDS inter-
ruption by a combination of currently available methods.
Consequently, the average size of the tailored KIAA pro-
teins in this study increased from 937 to 1288 amino acid
residues.

1. Evaluation of Integrity of CDSs Encoded by
KIAA Genes

CDSs encoded by cDNAs may be truncated due to the
fact that cloned cDNAs are not always complete copies
of mature mRNAs. In fact, there are many possible ways
for cDNA clones to become truncated: (1) some clones
might retain an intron at the 5′- and/or 3′-end, which
results in wrong assignment of CDS,9 (2) the 3′-end of
some cloned cDNAs might be truncated by internal prim-
ing of cDNA synthesis from an internal long A stretch in
mRNA;10 (3) if cDNAs contain an internal Not I site,
they can be cloned only in a 3′-end truncated form be-
cause Not I digestion of cDNAs is unavoidable during
the conventional cloning method we employed.5 In addi-
tion, long mRNAs (> 4 kb) are generally difficult templates
to copy in their entirety as cDNAs due to limitations of
performing in vitro reverse transcription. Thus, the
occurrence frequency of partial cDNA clones, which usu-
ally have truncated 5′-ends, increases with mRNA size.

Even if sophisticated full-length cDNA cloning meth-
ods are employed, the rates of occurrence of full-length
clones in randomly sampled clones have been reported to
be 40–60%.7,8 Since we have focused our sequencing ef-
forts on long cDNAs (> 4 kb) encoding large proteins
(> 400 amino acid residues), we have selected cDNA
clones in size-fractionated cDNA libraries on a random
sampling basis according to the capability of coding for
large proteins, but not to the fullness of the cDNA clones.
Thus, our first goal was to clarify how many KIAA cDNA
clones could produce full-length protein products.

To evaluate the fullness of CDSs in KIAA cDNA
clones, we compared the predicted protein sequences
with their homologues in the public databases. In addi-
ton, although all of KIAA cDNAs had been un-
known when we deposited their sequences in the pub-
lic databases, some KIAA cDNA sequences have been
updated in the databases by other groups. These up-
dated sequences were longer than our deposited cDNA
sequences in most cases, because they carefully studied
the structure and function of each KIAA gene on the
basis of our reported sequence information. Thus,
homology searching is expected to detect not only ho-
logous proteins in other organisms but also KIAA
protein sequences updated by other groups. Homol-
ogy search of the predicted KIAA proteins was per-
formed using BLAST against the SPTR protein database
(SWISS-PROT release 40.18, TrEMBL release 20.7).11,12

Figure 1 summarizes the results of the BLAST search:
57% (1141 out of 2001 cDNAs) of KIAA proteins have
more than one corresponding homologue or updated en-
tries in the SPTR protein database (Fig. 1A). The align-
ment of KIAA protein sequences with their homologous
or updated ones enabled us to detect the truncation of
CDSs of KIAA cDNAs as shown in Fig. 1. Figure 1B
shows that 52% (595 cDNAs) of the KIAA proteins had
the same or extended amino-terminals compared with
their homologues, although the start sites of CDSs of
their homologues were not always experimentally deter-
mined. Although a considerable number of the previ-
ously predicted KIAA proteins (573 of 595 cDNAs) were
apparently longer in the amino-terminal than their ho-
logues, this was simply because we tentatively reported
predicted protein sequences by open reading frames de-
tected by GeneMark analysis. In fact, 77% of cDNAs ap-
parently encoding proteins with longer amino-terminals
than those in the public database (442 cDNAs) were
found to contain the first methionine residue at the
same positions as homologous entries in the SPTR pro-
tein database. Interestingly, the CDSs predicted by
110 KIAA cDNAs had upstream methionine(s) in the
extended sequences. This indicated that all the protein
sequence entries in the SPTR protein database were not
full-length, and thus the evaluation of the integrity of
amino-terminals of KIAA proteins on the basis of these
data was not conclusive. The results also clarified that
Evaluation of the amino- or carboxy-terminal ends of KIAA proteins by comparison with the hit proteins. Homology search of the predicted KIAA proteins was performed by BLAST against the SPTR protein database (SWISS-PROT release 40.18, TrEMBL release 20.7). The hit proteins satisfying the following conditions were selected: i) the length was less than 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 70% or greater. (A) Integrity of 1141 KIAA proteins in comparison with the hit proteins analyzed by homology search against the SPTR database was evaluated. The states of KIAA proteins were indicated as follows: black bar, KIAA proteins which covered the entire region of the hit protein; dotted bar or hatched bar, KIAA proteins which had a truncated amino-terminal or carboxy-terminal end, respectively; open bar, KIAA proteins which had both truncated ends. Species of hit proteins are shown at the left of the graph. ‘Others’ included Canis familiaris, Oryctolagus cuniculus, Cricetulus griseus, Xenopus laevis, Gallus gallus, Brachydanio rerio, Cricetulus longicaudatus, Fugu rubripes, Mesocricetus auratus, and Spalax galilii. The number of KIAA proteins is horizontally indicated as bars in the respective species. (B) The relative amino-terminal positions of KIAA proteins to those of hit proteins are shown. The differences in their end positions are indicated at the left of the graph as the number of amino acid residues. Negative or positive numbers mean deficits or increases in the number of amino acid residues in KIAA proteins. The number of KIAA proteins is horizontally indicated as bars. If the amino-terminal position of KIAA protein is more extended than hit protein, the number of KIAA proteins of which the first methionine positions are coincident with those of the hit proteins are overwrote as a black bar. (C) Relative carboxy-terminal positions of KIAA proteins to those of the hit proteins are shown as described in (B).

As for carboxy-terminal ends of KIAA gene products, 82% (931 cDNAs) of them were found to have common carboxy-terminal ends to those of homologues in the databases (Fig. 1C). However, 148 KIAA cDNAs encoded proteins with carboxy-terminal truncation. The observed truncation of these KIAA cDNAs at the carboxy-terminal seemed to be caused by a retained intron, alternative splicing, internal dT-priming of cDNA synthesis, or internal digestion of CDSs with Not I. For example, the carboxy-terminal end of KIAA1471 protein (predicted by a cloned cDNA, fj02383) was truncated by internal digestion with Not I during cDNA synthesis. On the other hand, a cloned cDNA for KIAA1506 (hk04442) seemed to be synthesized by internal dT-priming because the sequence on the genome adjacent to the 3′-end of the cDNA contained a region relatively rich in A residues. As a result, their predicted CDSs were truncated and they had no in-frame stop codon up to the end.

To know the occurrence rate of the aberrant 3′-end of...
cDNA clones, we also evaluated their integrity by checking for the presence of a canonical poly(A) addition signal sequence within 30 nucleotides upstream of the 3′-end of the KIAA cDNAs having a respective hexamer horizontally indicated as dotted bars. cDNA clones having six or more consecutive adenesines, or eight adenesines within a 10-nucleotide region on the genomic sequences adjacent to the 3′-end of cDNA sequences were tentatively defined as “internal-priming” despite the poly(A) addition signals. When the 3′-end of cDNA had a Not I site without a poly(A) sequence, the cDNA was defined as “Internal Not I.” The number of KIAA cDNAs defined as an internal-priming or an internal Not I is also indicated by dotted bars.

2. Manual Curation of KIAA cDNA Clones

In the next step, we tailored previously isolated KIAA cDNA clones to have longer CDSs according to the evaluation results described above. For this purpose, we applied three different approaches as shown in Fig. 3. In approaches A and B, we selected another clone to be sequenced or used for subcloning a single KIAA gene in the accumulated cDNA clones analyzed by single-pass sequencing at their ends (>220,000 sequences). In approach C, we prepared a missing portion by PCR coupled with reverse transcription (RT-PCR) or by oligonucleotide synthesis and then connected it to the original cDNA clone by conventional methods. Consequently, 265, 8, and 17 predicted CDSs of KIAA cDNA clones were increased in the amino-, carboxy- and both terminal sequences, respectively (see supplement, Table 1), and the increases in the number of amino acid residues were 97,000 and 7216 in amino- and carboxy-terminal sequences, respectively (Fig. 4). The physical maps of the manually cured KIAA cDNAs whose gene products were predicted to be more than 300 amino acid residues longer than previously predicted are shown in Fig. 5. Table 2 shows Pfam domains revealed in the newly identified sequences in the manually cured KIAA cDNAs. For example, in the case of KIAA32623, ankyrin repeats were newly identified in the revised protein sequences. The methods of revisions for each KIAA cDNAs were described in the supplement (Table 1) and the original cDNA sequences were updated in the public databases. The comparisons of the original and revised cDNA clones are available at HUGE (http://www.kazusa.or.jp/huge).

Furthermore, spurious interruption of CDSs were also cured in 75 cDNA clones in this study. Twenty-five cDNA clones for KIAA genes were newly sequenced in their entireties to replace their original cDNAs having spurious splits of CDSs. Twelve cDNA clones were revised by subcloning of DNA fragments prepared from other cDNA clones without spurious coding interruption. Among these 37 clones, 35 cDNA clones were revised not only to remove spurious coding interruptions but also to extend amino- and/or carboxy-terminals of the encoded proteins as described above. In addition, 38 cDNA clones were revised by replacement of the regions having spurious CDS interruption by the homologous recombination reaction in yeast with the major RT-PCR products amplified using gene-specific primers. After confirmation of the sequences of the revised regions in respective clones, their sequences were deposited in the public databases as revisions of the previously deposited KIAA cDNAs. The list of these revised KIAA genes is available upon request.
Figure 3. Methods used in searching for DNA fragments compensate for truncated region of KIAA cDNA clones. For the clustering analysis using the expressed sequence tags (ESTs) of human cDNA clones we have analyzed (Kazusa-ESTs), more than 220,000 Kazusa-ESTs were applied to mask the repetitive sequences by RepeatMasker. The resultant sequences were subjected to BLAST search against a merged database which contained the end sequences themselves and human mRNA sequences excluding ESTs from GenBank (release 127.0). When a query end sequence shared more than 90% sequence identity without masked nucleotides and showed a score of more than 235 with any other sequences in the database, the hit sequences were grouped with the query. If a sequence was found in common with any of the different groups, the groups were merged into a single cluster. CAP2 was applied to assemble the sequences in each cluster and align them to make contigs (Fig. 3A). If each of the terminal sequences of a cDNA did not exist in the same contig, the positional relationship between the EST and the consensus sequence of the contig was elucidated along the genomic sequences (Fig. 3B). When the cDNA sequences which had CDSs longer than those of KIAA cDNAs have been deposited in the public database as an updated entry, the synthetic oligonucleotides or RT-PCR products which cover the truncated sequences were prepared (Fig. 3C). On the basis of the results of clustering analysis, either the cDNA clones having the extended region of KIAA cDNAs were sequenced in their entireties or the obtained DNA fragments having the extended regions were cloned into the original KIAA cDNAs by conventional methods. The gene-specific primers used in PCR and the amplified DNA fragment are indicated as solid triangle and dotted line, respectively.

3. Concluding Remarks

In this study, we evaluated KIAA cDNA sequences by comparing them with sequences of their updated entries or homologues in the SPTR database and manually cured 330 cDNA clones. Consequently, 604 clones became expression-ready in total among 1141 KIAA cDNA clones which had their homologous protein in the SPTR protein database. Because we have recently started a cDNA sequencing project of mouse KIAA homologues, we will be able to evaluate the integrity of all the KIAA cDNAs in this way in the near future. However, this preliminary study revealed that there are many reasons why it is not always straightforward to make KIAA cDNAs expression-ready. Although there seems to be no simple solution for this problem, we must conduct manual curation of KIAA cDNA clones one by one because the preparation of a set of expression-ready cDNA clones serves as an urgent and highly critical step toward functional analysis of KIAA genes.
Figure 5. Physical maps of revised KIAA cDNA clones. The physical maps shown here were constructed from the sequence data of respective revised 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 open reading frames 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 including Alu and MIRs sequences and other repetitive sequences thus detected are displayed by dotted and hatched boxes, respectively. The regions corresponded to the original KIAA proteins are partitioned by thin lines under the respective boxes.
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, Kazuko Yamada, Kiyoe Sumi, Takashi Watanabe, Kozue Yamane, Naoko Kato, Kazuhiro Sato, Akiko Ukigai, Kazuko Yamada, and Sachiko Minorikawa for their technical assistance.

References

1. Venter, J. C., Adams, M. D., Myers, E. W. et al. 2001, The sequence of the human genome, *Science*, 291, 1304–1351.

2. International Human Genome Sequencing Consortium. 2001, Initial sequencing and analysis of the human genome, *Nature*, 409, 860–921.

3. 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.

4. 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 from human immature myeloid cell line KG-1, *DNA Res.*, 1, 27–35.

5. 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.

6. Hirosawa, M., Ishikawa, K.-I., 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.

7. Nishikawa, T., Ota, T., Kawai, Y. et al. 2001, Comparison of sequences of cDNA clones obtained from oligo-capping cDNA libraries with those from unigene, *DNA Res.*, 31, 255–262.

8. The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium. 2001, Functional annotation of a full-length mouse cDNA collection, *Nature*, 409, 685–690.

9. Kozak, M. 1996, Interpreting cDNA sequences: some insights from studies on translation, *Mamm. Genome*, 8, 563–574.

10. Nam, D. K., Lee, S., Zhou, G. et al. 2002, Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription, *Proc. Natl. Acad. Sci. USA*, 99, 6152–6156.

11. Altschul, S. F., Madden, T. L., Schaffer, A. A. et al. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, 25, 3389–3402.

12. Bairoch, A. and Apweiler, R. 2000, The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000, *Nucleic Acids Res.*, 28, 45–48.

13. Beaudoin, E., Freier, S., Wyatt, J. R., Claverie, J.-M., and Gautheret, D. 2000, Patterns of variant polyadenylation signal usage in human genes, *Genome Res.*, 10, 1001–1010.

14. Kikuno, R., Nagase, T., Waki, M., and Ohara, O. 2002, HUGE: a database for human large proteins identified in the Kazusa cDNA sequencing project, *Nucleic Acids Res.*
15. Raymond, C. K., Pownder, T. A., and Sexson, S. L. 1999, General method for plasmid construction using homologous recombination, BioTechniques, 26, 134–141.
16. Li, C. and Evans, R. M. 1997, Ligation independent cloning irrespective of restriction site compatibility, Nucleic Acids Res., 25, 4165–4166.
17. Bateman, A., Birney, E., Durbin, R. et al. 2000, The Pfam families database, Nucleic Acids Res., 28, 263–266.
18. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Rapp, B. A., and Wheeler, D. L. 2000, GenBank, Nucleic Acids Res., 28, 15–18.
19. Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M., and Miller, W. 1998, A computer program for aligning a cDNA sequence with a genomic DNA sequence, Genome Res., 8, 967–974.