D502–D504 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh035

HUGE: a database for human KIAA proteins, a 2004 update integrating HUGEppi and ROUGE

Reiko Kikuno[1,*], Takahiro Nagase[1], Manabu Nakayama[1], Hisashi Koga[1,2], Noriko Okazaki[1], Daisuke Nakajima[1] and Osamu Ohara[1,3]

[1]Kazusa DNA Research Institute, 2-6-7 Kazusa-Kamatari, Kisarazu, Chiba, 292-0818, Japan, [2]Chiba Industry Advancement Center, 2-6 Nakase, Mihama-ku, Chiba, 261-7126, Japan and [3]RIKEN Research Center for Allergy and Immunology, 1-7-22 Suehiro, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan

Received September 12, 2003; Accepted September 16, 2003

ABSTRACT

We have been developing a Human Unidentified Gene-Encoded (HUGE) protein database (http://www.kazusa.or.jp/huge) to summarize results from sequence analysis of human novel large (>4 kb) cDNAs identified in the Kazusa cDNA sequencing project. At present, HUGE contains 2031 cDNA entries (KIAA cDNAs), for each of which a gene/protein characteristic table has been prepared. Since we have been shifting our research attention from the identification and cloning of novel cDNAs to the functional analysis of the proteins encoded by these cDNAs (KIAA proteins), we have not substantially increased the number of cDNA entries in HUGE for some time. Instead, we have manually curated 451 KIAA cDNAs in order to prepare a set of genetic resources to facilitate the functional analysis of KIAA proteins. In addition, we have updated the contents of the corresponding gene/protein characteristic tables in HUGE and have constructed two subsidiary databases, HUGEppi (http://www.kazusa.or.jp/huge/ppi) and ROUGE (http://www.kazusa.or.jp/rouge), to make available the results from our study of KIAA protein function. HUGEppi shows detailed information on protein–protein interactions detected between 84 pairs of KIAA proteins by yeast two-hybrid screening. ROUGE summarizes the results of computer-assisted analyses of ~1000 mouse homologues of human large cDNAs that we identified.

INTRODUCTION

We have been involved in the Kazusa cDNA sequencing project, which focuses on sequencing long cDNA clones (>4 kb) with the aim of identifying and characterizing previously unidentified human genes that can encode large proteins (>50 kDa) (1,2). Over 2000 novel human genes have been characterized in this project. The genes are systematically designated 'KIAA' plus a four-digit number. We have been developing the HUGE database to provide a summary view of the results of experiments and computer-assisted analyses of the KIAA cDNAs and the proteins that they are predicted to encode (3,4). We have updated the results of *in silico* genome mapping of the KIAA cDNAs and sequence comparison between the cDNA and the genome to the HUGE database on a regular basis since 2001, when the human draft genome sequence was made available to the public (5). The results of these *in silico* experiments have helped us evaluate the completeness and the accuracy of cDNA clones.

Our ultimate goal is to identify the physiological functions of the proteins encoded by KIAA genes. For this purpose, we have been redirecting our efforts into the next stage of the project, which focuses on the functional analysis rather than the large-scale isolation, of human novel large cDNA clones. The next stage of the project includes (i) manual curation of KIAA cDNA clones to obtain full-length cDNAs, (ii) functional assays such as protein–protein interaction analysis and DNA and protein microarrays, and (iii) the establishment of model animal experimental systems. We have been proceeding with each approach in parallel and, in some cases, include the information obtained from the experimental approaches in the HUGE database.

Following the addition of 95 new KIAA cDNAs since our last report, in January 2002 (4), the total number of cDNA entries in the HUGE database has reached 2031. In addition 451 KIAA cDNA sequences have been revised by manual curation (6). In addition, we introduce the two new subsidiary databases, HUGEppi (http://www.kazusa.or.jp/huge/ppi) and ROUGE (http://www.kazusa.or.jp/rouge) to complement the HUGE database. HUGEppi was constructed to show the results of protein–protein interaction analysis for 84 pairs of KIAA proteins identified using the yeast two-hybrid system (7). ROUGE contains the results of computer-assisted sequence analysis of mouse homologues of KIAA cDNA (mKIAA cDNA) that we isolated (8–10). Presently, the ROUGE database contains ~1000 mKIAA cDNA entries. It has the same basic architecture as the HUGE database. In this report, we will describe a recent update of the HUGE database and the creation of two subsidiary databases, HUGEppi and ROUGE.

*To whom correspondence should be addressed. Tel: +81 438 52 3932; Fax: +81 438 52 3931; Email: kikuno@kazusa.or.jp

ORGANIZATION OF THE HUGE AND THE ROUGE DATABASES

Each cDNA entry in the HUGE database has been designated 'KIAA' plus a four-digit number, e.g. KIAA0001. Accordingly, the mouse homologue of each KIAA cDNA has been entered into the ROUGE database with the designation 'mKIAA' plus the same four-digit number as its human counterpart. The HUGE and the ROUGE databases have the same basic organization. Each cDNA entry has its own gene/protein characteristic table, in which the results from computer-assisted analysis of the cDNA sequence and the deduced amino acid sequences are summarized. The table includes the predicted 5′- and 3′-integrity of the cDNA sequence, a report on the potential for spurious coding region interruption and N-terminal truncation by GeneMark analysis (11), the results of various database searches and genome mapping, and a comparison between the cDNA and the corresponding genomic sequences (5,12).

MANUAL CURATION OF KIAA CDNA SEQUENCES

Since the KIAA cDNAs were derived from long mRNAs, it is unlikely that all are full length. Furthermore, large cDNAs are more likely to contain artifacts than shorter cDNAs due to an increased likelihood of reverse transcriptase error and/or the retention of intron sequence(s) in the template mRNA. Distinctively, the cDNA clones carrying the artifacts have truncated and/or interrupted protein coding sequence (CDS). This was problematic, as it was anticipated that the entire set of cDNA clones would be used as reagents for functional analysis of KIAA gene products (KIAA proteins). Accordingly, we first calculated coding potential using the GeneMark program to examine whether or not the CDS was interrupted; subsequently, we predicted the 5′- and 3′-integrity of the CDS by sequence comparison with those proteins in the public databases that are closely related to each KIAA cDNA. Then, the cDNAs that were found either to carry artifacts or to be extended were subjected to manual curation to obtain more precise, and/or longer, CDS information (6).

Since we opened the HUGE database to the public (3), we have tailored 394 and 36 KIAA cDNAs for the 5′- or 3′-terminal extension, respectively, and revised 60 KIAA cDNAs for spurious CDS interruptions. In total 22% of the KIAA cDNA entries have been manually curated at least once. The HUGE database provides the revision history of the update, which shows the result of sequence comparison between the original and updated KIAA cDNA sequences, to indicate both the extent of tailored sequence and any nucleotide differences between the two sequences. The updated cDNA sequences and the revised CDS information were submitted to the DDBJ/EMBL/GenBank database. The date of the last update in DDBJ/EMBL/GenBank is shown in the 'List of Gene/Protein Characteristic Table', which is linked from the top page of the HUGE database site.

PROTEIN–PROTEIN INTERACTIONS

Large proteins, such as proteins encoded by KIAA genes (KIAA proteins) frequently display multiple domains, and they are probably involved in various interactions with other molecules *in vivo*. However, little information on protein–protein interactions between large proteins has been gathered, mainly due to technical limitations. In this respect, comprehensive study of protein interactions, focusing on KIAA proteins, would better characterize the functions of KIAA proteins and would allow the detection of previously unknown protein interactions. This type of study was carried out by Nakayama *et al.* (7) who used yeast two-hybrid screening to identify 84 submembranous protein–protein interactions between KIAA proteins. The HUGEppi database was constructed to present detailed information on the protein–protein interactions between KIAA proteins that are reported in these studies. To indicate which portions of KIAA proteins were involved in the interaction, we show the positions of bait and prey(s) in diagrams that also show the positions of protein motifs predicted by InterProScan (13), and of transmembrane regions assigned by SOSUI (14), for each interaction pair of KIAA proteins. In addition, information about those preys that gave positive signals are presented with the results of inspection of the coding frames and directions. This information helps us to determine the essential interaction sites as well as to evaluate whether or not the interactions observed were real positives.

ACCUMULATION OF MOUSE HOMOLOGUES OF KIAA CDNAS

To elucidate functional roles for KIAA proteins in biological processes, physiological, developmental and genetic studies are necessary. We intend to establish model animal systems to accumulate experimental data for the characterization of KIAA proteins *in vivo*, thereby circumventing the legal and ethical restrictions on the use of human materials for these studies. The first step of this project began in 2001 with the collection and characterization of cDNAs encoding mouse counterparts of human KIAA proteins (8). As CDSs in genes with orthologous relationships are generally highly conserved between human and rodent (15), obtaining mouse orthologues of KIAA cDNAs (mKIAA cDNAs) will also help us to evaluate whether or not the KIAA cDNA sequences contained full-length and correct CDSs. We have already reported ~1000 mKIAA cDNA sequences, and deposited novel sequences to DDBJ/EMBL/GenBank.

The ROUGE database has been developed to show the results from this study. The organization of the ROUGE database is fundamentally similar to that of the HUGE database. All of the computational analyses used to characterize KIAA cDNA sequences and the deduced amino acid sequences have been applied and the results are shown in the ROUGE database. Additionally we have displayed the results of sequence comparison between KIAA and mKIAA cDNAs at the nucleotide and amino acid levels. To assign a CDS for each mKIAA cDNA, we applied GeneMark analysis, as in the case of KIAA cDNAs, and predicted the longest ORF as the CDS when there was no CDS interruption alert from the GeneMark program. When a CDS split was reported only in mKIAA cDNA, the corresponding regions were assigned as encoding a continuous single CDS in KIAA cDNA; we considered that the predicted CDS interruption in the mKIAA cDNA was spurious. When each of the split CDSs was at least 50 amino acid residues in length, and shared sequence identity

of ⩾50%, we merged each CDS into a single consecutive CDS *in silico*, and produced the amino acid sequence from the merged CDS. Further computer analysis at the amino acid sequence level was performed on the merged CDSs.

FUTURE DIRECTIONS

Sequence comparison between KIAA and mKIAA cDNAs helps us to evaluate the completeness and correctness of the CDSs. To report this evaluation, we presented the sequence alignment of CDSs and the extent of sequence differences at both DNA and amino acid levels in the ROUGE database. We also presented a 3′-UTR sequence alignment to determine whether the polyadenylation signal sequences had a conserved position. Using our data we can examine the authenticity of 3′-UTR sequences, or possibly predict the different spliced forms of 3′-UTRs, between KIAA and mKIAA cDNAs. We are planning to predict the authenticity of KIAA cDNA translation start sites more precisely by integrating human and mouse genome sequence data into the comparative analysis of the KIAA and mKIAA cDNA sequences, i.e. in addition to the database searches against known protein sequences, which we have already completed during the tailoring of KIAA cDNA clones. Furthermore, as we obtain functional information on KIAA/mKIAA proteins from various ongoing experimental approaches, such as protein–protein interaction analyses, DNA and protein microarrays and gene knockout experiments, we will incorporate them into the HUGE/ROUGE databases.

ACKNOWLEDGEMENTS

We are grateful to Masaki Takazawa and Nobue Kashima for their excellent technical assistance. This work was supported by the Kazusa DNA Research Institute Foundation.

REFERENCES

1. Ohara,O., Nagase,T., Ishikawa,K.-I., Nakajima,D., Ohira,M., Seki,N. and Nomura,N. (1997) Construction and characterization of human brain cDNA libraries suitable for analysis of cDNA clones encoding relatively large proteins. *DNA Res.*, **4**, 53–59.
2. 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 brain which code for large proteins *in vitro*. *DNA Res.*, **8**, 319–327.
3. Suyama,M., Nagase,T. and Ohara,O. (1999) HUGE: a database for human large proteins identified by Kazusa cDNA sequencing project. *Nucleic Acids Res.*, **27**, 338–339.
4. 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.*, **30**, 166–168.
5. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
6. Nakajima,D., Okazaki,N., Yamakawa,H., Kikuno,R., Ohara,O. and Nagase,T. (2002) Construction of expression-ready cDNA clones for KIAA genes: manual curation of 330 KIAA cDNA clones. *DNA Res.*, **9**, 99–106.
7. Nakayama,M., Kikuno,R. and Ohara,O. (2002) Protein–protein interactions between large proteins: two-hybrid screening using a functionally classified library composed of long cDNAs. *Genome Res.*, **12**, 1773–1784.
8. Okazaki,N., Kikuno,R., Ohara,R., Inamoto,S., Hara,Y., Nagase,T., Ohara,O. and Koga,H. (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.
9. Okazaki,N., Kikuno,R., Ohara,R., Inamoto,S., Aizawa,H., Yuasa,S., Nakajima,D., Nagase,T., Ohara,O. and Koga,H. (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.
10. Okazaki,N., Kikuno,R., Ohara,R., Inamoto,S., Koseki,H., Hiraoka,S., Saga,Y., Nagase,T., Ohara,O. and Koga,H. (2003) Prediction of the coding sequences of mouse homologues of KIAA gene: III. The complete nucleotide sequences of 500 mouse KIAA-homologous cDNAs identified by screening of terminal sequences of cDNA clones randomly sampled from size-fractionated libraries. *DNA Res.*, **10**, 167–180.
11. Hirosawa,M., Isono,K., Hayes,W. and Borodovsky,M. (1997) Gene identification and classification in the *Synechocystis* genomic sequence by recursive gene mark analysis. *DNA Seq.*, **8**, 17–29.
12. Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
13. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
14. Hirokawa,T., Boon-Chieng,S. and Mitaku,S. (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, **14**, 378–379.
15. Makalowski,W. and Boguski,M.S. (1998) Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl Acad. Sci. USA*, **95**, 9407–9412.