Identification of Novel Transcribed Sequences on Human Chromosome 22 by Expressed Sequence Tag Mapping

Makoto HIROSAWA,1,* Takahiro NAGASE,1 Yasuaki MURASHI,2 Reiko KIKUNO,1 and Osamu OHARA1

Kazusa DNA Research Institute, 1532-3 Yana, Kisarazu, Chiba 292-0812, Japan1 and Higeta Shouyu Co., LTD. 2-8 Chuo-cho, Choshi, Chiba 288-8680, Japan2

(Received 7 December 2000; revised 31 January 2001)

Abstract

To identify sequences on the human genome that are actually transcribed, we mapped expressed sequence tags (ESTs) of long cDNAs ranging from 4 kb to 7 kb along a 33.4-Mb sequence of human chromosome 22, the first human chromosome entirely sequenced. By the EST mapping of 30,683 long cDNAs in silico, 603 cDNA sequences were found to locate on chromosome 22 and classified into 169 clusters. Comparison of the genomic loci of these cDNA sequences with 679 genes already annotated on chromosome 22q revealed that 46 clusters represented newly identified transcribed sequences. To further characterize these sequences, we sequenced 12 cDNAs in their entirety out of 46 clusters. Of these 12 cDNAs, 6 were predicted to include a protein-coding region while the remaining 6 were unlikely to encode proteins. Interestingly, 3 out of the 12 cDNAs had the nucleotide sequences of the opposite strands of the genes previously annotated, which suggested that these genomic regions were transcribed bi-directionally. In addition to these newly identified 12 cDNAs, another 12 cDNAs were entirely sequenced since these cDNAs were likely to contain new information about the predicted protein-coding sequences previously annotated. In the cases of KIAA1670 and KIAA1672, these single cDNA sequences covered two separately annotated transcribed regions. For example, the sequence of a clone for KIAA1670 indicated that the CHKL and CPT1B genes were co-transcribed as a contiguous transcript without making both the protein-coding regions fused. In conclusion, the mapping of ESTs derived from long cDNAs followed by sequencing of the entire cDNAs provided indispensable information for the precise annotation of genes on the genome together with ESTs derived from short cDNAs.

Key words: long cDNA; single-pass sequence; chromosome 22; cDNA sequencing; brain

1. Introduction

Although it had been thought to take a long time to complete the human genome sequencing 10 years ago, the international human genome sequencing consortium announced last year that the “working draft” of the human genome had almost been completed. The next target is obviously to list up and analyze all the genes encoded by the human genome. For this purpose, the processes of gene annotation on chromosomes 21 and 22, the complete sequences of which are already known,1,2 have given us lessons to learn about how to annotate genes on the human genome. In general, most current procedures of gene annotation are done combining ab initio exon prediction programs with pre-existing experimental data such as cDNA sequences (entire sequences or terminal sequences) and protein sequences. Terminal sequence data of human cDNAs, called expressed sequence tags (ESTs), are overwhelmingly used as the experimental data for complementing in silico gene prediction because of the enormous number of known ESTs. However, it is evident that sequence information of entire cDNAs is also eventually required for accurate gene annotation due to the complex nature of exon/intron organization of human genes. Considering this, we have conducted a cDNA sequencing project for entire regions of cDNAs since 1994.3,4 Our project is distinct from others in that we focus our sequencing efforts on long cDNAs.5 The total number of nucleotide residues determined by our cDNA project has reached nearly 10 Mb to date4 and these data were, in fact, widely used for gene annotation along the human genome. Because most publicly available ESTs originate from relatively short cDNAs (average size of 0.9–1.5 kb),6,7 they represent only partial regions of long transcripts. This might cause serious problems in gene annotation, since these public ESTs are apt to fail to provide any experimental evidence of continuity of

* Communicated by Michio Oishi
* To whom correspondence should be addressed. Tel. +81-438-52-3919, Fax. +81-438-52-3918, E-mail: hirosawa@kazusa.or.jp
long transcripts in most cases. In this respect, analysis of ESTs derived from long cDNAs is obviously the most straightforward way to solve the problems in annotation of long transcribed regions.

In this study, ESTs from long cDNAs were mapped on chromosome 22q in silico to identify overlooked transcribed sequences in the previous gene annotation. For this purpose, we used the 5′ and 3′ terminal sequence data of long cDNA clones previously accumulated in our comprehensive cDNA project. Based on the mapping data of terminal sequences of these long cDNAs, we selected cDNA clones for entire sequencing as follows. Among cDNAs identified to be derived from unannotated regions in the previous study, 12 cDNA clones were selected and entirely sequenced. Furthermore, an additional 12 cDNA clones which cover much longer genomic regions than previously annotated genes were also sequenced. Consequently, we found that regions allocated for consecutive annotated genes were transcribed as a single contiguous transcript in some cases, which indicated that they were incorrectly annotated. Interestingly, we identified transcribed sequences in the opposite strand of the previously annotated genes, which implied that these regions were bi-directionally transcribed. Taking into consideration that the number of ESTs analyzed in this study was much smaller than that of publicly available ESTs, we conclude that the EST mapping of long cDNA clones is a powerful and efficient complement to conventional gene annotation methods.

2. Materials and Methods

2.1. cDNA library construction and DNA sequencing

ESTs of 30,683 cDNA clones from the size-fractionated human brain cDNA libraries were used for analysis. These cDNAs were prepared according to the method previously described. In brief, cDNA fragments were synthesized by the use of reverse transcriptase (SuperScript II, GibcoBRL, USA), a (dT)$_{15}$ primer carrying a Not I site and poly(A)$^+$ RNAs of human whole adult or fetal brain (Clontech, USA) according to the instruction manual. After ligation of the SalI adaptor to the 5′-end of the cDNAs, Not I-digested cDNAs were separated on agarose gel to remove cDNA fragments less than 3 kb. The recovered cDNA fragments larger than 3 kb were ligated with the SalI-Not I-digested pBlue-script II SK$^+$ vector. Moreover, the recovered plasmids from bacterial colonies in a covalently closed circular form were size-fractionated by agarose gel electrophoresis. For DNA sequencing, dye-primer or dye-terminator cycle sequencing reactions were performed using ABI PRISM cycle sequencing kits (Applied Biosystems, USA) and the products were analyzed with ABI 373 or 377 DNA sequencers. The entire sequencing of cDNA inserts was performed under the shotgun strategy according to the procedures previously described.

2.2. Computational assignment of transcribed regions using ESTs of long cDNAs on chromosome 22

The contiguous sequence of 33.4 Mb of the long arm of the chromosome 22 (22q) including 11 gaps was fetched from the public database [ftp://ftp.sanger.ac.uk/pub/human/chr22/sequences/Chr_22/complete_sequence/Chr_22_analysis_version_22-10-1999.fa]. The cDNA clones used for the EST mapping were isolated from the size-fractionated human adult brain cDNA library Nos. 2 and 6 (insert sizes ranging from 4 to 7 kb, 14,123 cDNA clones) and fetal brain cDNA library Nos. 4 to 6 (insert sizes ranging from 5 to 7 kb, 16,560 cDNA clones) previously constructed as described above. The 5′- and 3′-end single-pass sequences of 30,683 cDNA clones were subjected to BLAST search against the contiguous sequence of 22q with a threshold E-value of 1.0 e$^{-50}$. If both of the 5′- and 3′-end single-pass sequences matched the genomic sequence of 22q with sequence identity of 80% or greater, the cDNA clones were considered as candidates derived from 22q. This threshold value of sequence identity was used for the first screening of the candidate sequences allocated on the genomic sequences of 22q to prevent us from overlooking the sequences with gap region(s) caused by splicing events. When both single-pass sequences of a clone could be aligned to the genomic sequence with identity of more than 90% in a colinear manner, the clone was automatically mapped. Finally, manual inspection was performed to curate the BLAST search results.

2.3. Reverse transcription-coupled polymerase chain reaction (RT-PCR)

To identify the sequences of the major transcripts for KIAA1670, we carried out the polymerase chain reaction following reverse transcription (RT-PCR) using the poly(A)$^+$ RNA of human adult brain and the gene-specific primers: 5′-GAAGATTTGCTGTAGAAGTC-3′ and 5′-GTTGATCCCAGACAGGTAGAC-3′. The cDNA template for RT-PCR was synthesized according to the method previously described. An aliquot of the cDNA template mixture (corresponding to 2 ng of the starting poly(A)$^+$ RNA) was subjected to PCR using LA Taq DNA polymerase (0.25 units, Takara Shuzo Co., Ltd., Japan) in 10 µl with a DNA Thermal Cycler P9600 (Applied Biosystems, USA). The thermal cycling conditions used were as follows: the first denaturation at 95°C for 1 min; 30 cycles of 30-sec denaturation at 95°C, 60-sec primer annealing at 60°C and 90-sec polymerization at 72°C. After confirming that PCR worked well by agarose gel electrophoresis, the PCR products were subjected to direct sequencing using the gene-specific oligomers shown above as sequencing primers.
3. Results and Discussion

3.1. In silico mapping of ESTs of long cDNAs on chromosome 22

Among 30,683 cDNA clones, ESTs of 603 cDNA clones were allocated to 22q by the BLAST search and classified into 169 clusters. The ratio of the number of cDNA clones allocated on 22q over the total number of cDNA clones (603/30,683 = 2.0%) was close to that of the length of chromosome 22 over that of the total genome (1.6–1.8%). The average genomic size corresponding to the allocated cDNAs is 34.5 kb (median 5.5 kb) and obviously longer than that of the previous statistics of the sizes of the annotated genes on 22q (19.2 kb, median 3.7 kb)². This means that the EST mapping of longer cDNA is more informative for understanding the gene organization of long transcripts. The mapped genomic loci of these cDNA clones were then compared with those of the genes annotated on 22q by Dunham et al., namely, ‘known’ genes as genes identical to known human sequences (247), ‘related’ genes as genes containing a region similar to known sequences (150), ‘predicted’ genes as sequences homologous to only ESTs (148), ‘pseudo’ genes as sequences homologous to a known sequence but with a disrupted protein coding sequence (CDS) (134).² From the comparison of our EST mapping data with the previous annotation, 46 clusters were newly identified as transcribed sequences which do not overlap with any of the 679 annotated genes reported by Dunham et al. while the genomic loci of the remaining 123 clusters partially overlapped with those of the annotated genes. Seven clusters could also be mapped in regions in each of which multiple genes were located in the previous report.² In addition, some clusters were allocated on regions where ‘related’ or ‘predicted’ genes, which have not been experimentally verified as human genes yet, were mapped. Thus, to further characterize cDNA sequences identified, we sequenced entire cDNA inserts of some clones. The cDNA clones entirely sequenced in this study are summarized as follows: newly identified transcribed sequences whose corresponding genomic lengths were more than 6.4 kb, 12 cDNAs (KIAA1644, KIAA1645, KIAA1648, KIAA1650, KIAA1652, KIAA1654-KIAA1656, and KIAA1658–KIAA1661); transcribed sequences allocated to the regions overlapping with the ‘related,’ ‘predicted,’ or ‘pseudo’ genes but covering most of the regions of the genes, 8 cDNA clones (KIAA1646, KIAA1647, KIAA1665–KIAA1669, and KIAA1671); 2 cDNA clones (KIAA1662 and KIAA1663) having transcribed sequences assigned on the regions overlapping with ‘known’ genes and having longer 5′-upstream genomic sequences; and 2 cDNA clones (KIAA1670 and KIAA1672) having transcribed sequences almost completely covering two ‘known’ or ‘predicted’ genes. In total, 24 cDNA clones were entirely sequenced in this study.

3.2. Entire sequences of newly identified cDNAs

As described above, 46 cDNA clusters represented novel transcribed sequences which did not overlap with the annotated genes in the previous report.² Thus, we determined the entire sequences of 12 cDNA clones shown in Table 1, whose corresponding genomic lengths ranged from 6 kb to 42 kb. We selected these cDNA clones simply because cDNAs consisting of multiple protein-coding exons frequently occupy larger genomic regions than those derived from a small number of exons. In fact, in the other 34 cDNA clusters, most of whose corresponding genomic sizes were less than 6.4 kb, no CDS was predicted in the corresponding genomic regions of 24 clusters by an exon prediction program, GENSCAN,¹² and the average length of CDSs predicted in those of the remaining 10 clusters was 163 amino acid residues. Table 1 lists the sequence information of the CDSs entirely sequenced, the lengths of cDNAs and the corresponding genomic regions, the lengths of the longest CDS predicted in cDNA by a CDS prediction program GeneMark,¹³,¹⁴ and the lengths of CDSs predicted in the genomic regions by GENSCAN. When CDSs were predicted only in the opposite strand of the genomic sequence entry, their lengths are shown with the prefix “-”.

Regarding 3 genes (indicated as “NewR” in Table 1), no CDS was predicted in the same strand in both the cDNA and their genomic sequences. The CDS lengths actually found in the cDNAs were essentially in agreement with those predicted from the corresponding genomic regions. These results thus indicate that EST mapping enabled us to obtain information on the possible CDSs of the cDNAs without sequencing in their entirety. Moreover, from the EST mapping, 3′-integrity of the CDSs was also confirmed by presence of canonical polyadenylation signals (AAUAAA or AUUAAA) usually observed within 30 nucleotides upstream of the 3′-end of cDNA.¹⁵ Furthermore, the possibility of CDSs generated by internal priming with dT primer could be assessed by the appearance of a poly(A) stretch within 10 nucleotides just downstream of the 3′-end of the cDNA on the genome. In fact, cDNAs for seven genes (KIAA1652, KIAA1654, KIAA1655, KIAA1658–KIAA1661) were found to be synthesized by internal priming since canonical poly(A) signals were not observed and since 10 or more consecutive A residues flanking the 3′-end of the cDNAs were present in the genome (Table 1). In the sequence data of these 12 cDNAs, CDSs were detected in 6 cDNAs (KIAA1644, KIAA1645, KIAA1648, KIAA1650, KIAA1654, and KIAA1658) by the GeneMark program.¹³,¹⁴ Regarding the remaining 6 cDNAs, (KIAA1652, KIAA1655, KIAA1656, and KIAA1658–KIAA1660), no possible CDSs were detected and most of them seemed to be synthesized by internal priming (Table 1).
Table 1. Information of sequence data of the identified genes.

| Gene number (KIAA) | Accession number | Class of cDNA | cDNA length (bp) | Predicted CDS length in cDNA (aa) | Lengths of genomic regions (bp) | Predicted CDS length in genomic region (aa) | Integrity of 3′-end of cDNA | poly(A) addition signal | Internal priming |
|-------------------|-----------------|---------------|-----------------|-----------------------------------|---------------------------------|-----------------------------------------------|--------------------------|---------------------|----------------|
| KIAA1644          | AB051431        | New           | 6,440           | 132                               | 42,069                          | 70 AUA AAA                                   |                         |                     |                 |
| KIAA1645          | AB051432        | New           | 6,645           | 386                               | 38,155                          | 664-977                                     |                         |                     |                 |
| KIAA1646          | AB051433        | Overlap       | 4,171           | 481                               | 35,801                          | 422 AUA AAA                                  |                         |                     |                 |
| KIAA1647          | AB051434        | Overlap       | 5,477           | 159                               | 23,996                          | 378 AUA AAA                                  |                         |                     |                 |
| KIAA1648          | AB051435        | NewR          | 5,851           | 112                               | 16,113                          | -521 AUA AAA                                 |                         |                     |                 |
| KIAA1650          | AB051437        | New           | 4,292           | 797                               | 12,563                          | 780 AUA AAA                                  |                         |                     |                 |
| KIAA1652          | AB051439        | New           | 6,381           | N                                 | 10,139                          | 155 none                                     | yes                      |                     |                 |
| KIAA1654          | AB051441        | NewR          | 5,147           | 91                                | 7,730                           | -125 none                                    | yes                      |                     |                 |
| KIAA1655          | AB051442        | New           | 6,419           | N                                 | 7,242                           | 106 none                                     | yes                      |                     |                 |
| KIAA1656          | AB051443        | New           | 6,970           | N                                 | 6,973                           | -617 AUA AAA                                 |                         |                     |                 |
| KIAA1658          | AB051445        | NewR          | 5,879           | N                                 | 6,859                           | -140 none                                    | yes                      |                     |                 |
| KIAA1659          | AB051446        | New           | 6,855           | N                                 | 6,841                           | 0 none                                       | yes                      |                     |                 |
| KIAA1660          | AB051447        | New           | 6,799           | N                                 | 6,790                           | 133 none                                     | yes                      |                     |                 |
| KIAA1661          | AB051448        | New           | 6,475           | 90                                | 6,465                           | 77+62 none                                   | yes                      |                     |                 |
| KIAA1662          | AB051449        | Overlap       | 4,042           | 1,231                             | 48,112                          | 600+686 none                                  |                         |                     |                 |
| KIAA1663          | AB051450        | Overlap       | 4,314           | 347                               | 13,285                          | 344 AUA AAA                                  |                         |                     |                 |
| KIAA1665          | AB051452        | Overlap       | 4,311           | 217                               | 18,519                          | 264 AUA AAA                                  |                         |                     |                 |
| KIAA1666          | AB051453        | Overlap       | 5,477           | 1,003                             | 5,475                           | 1,569 AUA AAA                                 |                         |                     |                 |
| KIAA1667          | AB051454        | Overlap       | 6,118           | 708                               | 30,144                          | 664 AUA AAA                                  |                         |                     |                 |
| KIAA1668          | AB051455        | Overlap       | 5,779           | 791                               | 34,503                          | 819 none                                     | yes                      |                     |                 |
| KIAA1669          | AB051456        | Overlap       | 5,605           | 1,405                             | 26,671                          | 2,206 AUA AAA                                 |                         |                     |                 |
| KIAA1670          | AB051457        | Overlap       | 4,906           | 598                               | 14,097                          | 1,244 AUA AAA                                 |                         |                     |                 |
| KIAA1671          | AB051458        | Overlap       | 6,123           | 364                               | 127,226                         | 18+120+4+44+246 AUA AAA                     |                         |                     |                 |
| KIAA1672          | AB051459        | Overlap       | 5,168           | 395                               | 120,288                         | 714+77+194 AUA AAA                            |                         |                     |                 |

a) Accession numbers of DDBJ, EMBL, and GenBank databases.
b) New, Overlap, and NewR signify cDNA newly identified as transcribed sequence, cDNA overlapping with annotated genes in the same direction, and cDNA overlapping with annotated genes in the opposite direction, respectively.
c) Values excluding poly(A) sequences.
d) The length of the longest CDS (in amino acids) is shown even if the GeneMark analysis predicted multiple CDSs in the cDNA. Detailed information is available through our web site HUGE (http://www.kazusa.or.jp/huge). N, not predicted.
e) Lengths of genomic regions covered by the corresponding cDNAs.
f) CDS length (in amino acids) predicted by GENSCAN in the genomic region. When multiple CDSs were predicted, these lengths are concatenated with ‘+’.
g) Polyadenylation signal observed within a 30-nucleotides upstream of the 3′-end of cDNA was indicated. The potential internal priming of the cDNA was described as ‘yes’ if a polyA stretch existed more than six contiguously or more than 80% on the genome within a 10-nucleotides just downstream of the 3′-end of the cDNA.

3.3. Entire sequences of cDNAs overlapping with annotated genes

In addition to the cDNAs originating from the newly identified transcribed genomic regions described above, additional 12 cDNAs (indicated as “Overlap” in Table 1), which partially overlapped with the annotated genes on the genomic sequence of 22q, were entirely sequenced since these cDNAs are likely to contain considerable amounts of new information regarding gene structures already annotated. As shown in Table 1, these cDNAs had possible CDSs although KIAA1668 and KIAA1672 seem to be synthesized by internal priming. Table 2 shows the locus names of the annotated genes, the genomic loci assigned for both the ends of cDNAs, orientations of transcription, and genomic lengths covered by cDNAs. The genomic lengths of 12 cDNAs (indicated as “Overlap” in Table 2) were much larger than those of the annotated genes and information on newly predicted amino acid sequences was obtained. Regarding KIAA1662 and KIAA1663, sequence information on the 5′-upstream genomic regions of the ‘known’ genes was obtained. Although the CDS of KIAA1663 was identical to that of Tob2 (GenBank accession numbers: D64109 and AB035207) the 5′-end position of the KIAA1663 cDNA was more than 9 kb upstream to that of the Tob2 gene on the genomic sequence. This result revealed that the transcription start site of the Tob2 gene is located further upstream to the genomic position previously reported. The two cDNA sequences for KIAA1670 and KIAA1672 almost completely covered the genomic regions in each
Table 2. Transcribed sequences identified in the genomic regions where genes are previously annotated.

| Gene number (KIAA) | Class of cDNA | Gene categories/locus name | End position 1 (bp) | End position 2 (bp) | Orientation | Length of genomic regions (bp) |
|--------------------|--------------|-----------------------------|---------------------|---------------------|-------------|-------------------------------|
| KIAA1646           | Overlap      | Predic/ bK29P11.1           | 30464493            | 30500293            | R           | 35,801                        |
| KIAA1647           | Overlap      | Pseudo/AC007326.1           | 2837790             | 2861785             | D           | 23,966                        |
| KIAA1662           | Overlap      | Known/J37E16.4              | 21549664            | 21597775            | D           | 48,112                        |
| KIAA1663           | Overlap      | Known/TROB2                 | 25227146            | 25270430            | R           | 13,285                        |
| KIAA1665           | Overlap      | Related/dJ347E13.5          | 25349474            | 25367992            | R           | 18,519                        |
| KIAA1666           | Overlap      | Predict/ AP000552.5         | 5328924             | 5334398             | D           | 5,475                         |
| KIAA1667           | Overlap      | Predict/ bK1048E9.4         | 10437834            | 10467977            | R           | 30,144                        |
| KIAA1668           | Overlap      | Predict/ dJ1014D13.2        | 21736023            | 21770525            | D           | 34,503                        |
| KIAA1669           | Overlap      | Predict/ dJ402G11.6         | 34025453            | 34052123            | R           | 26,671                        |
| KIAA1670           | Overlap      | Known/CPT1B                 | 34330742            | 34348383            | R           | 14,097                        |
| KIAA1671           | Overlap      | Known/CHKL                 | 34331000            | 34340300            | R           | 9,301                         |
| KIAA1672           | Overlap      | Predict/ bK221G9.4          | 9105955             | 9183157             | D           | 127,203                       |
| KIAA1673           | NewR         | Known/KIAA1043             | 27379191            | 27499478            | R           | 120,288                       |
| KIAA1674           | NewR         | Known/BIK                  | 26929239            | 26936968            | R           | 7,730                         |
| KIAA1675           | NewR         | Known/SEC14L2              | 14318974            | 14325832            | R           | 6,859                         |

a) Overlap and NewR signify cDNA overlapping with annotated genes in the same direction and cDNA overlapping with annotated genes in the opposite direction, respectively.
b) Information on the annotated genes was fetched from the original paper (ref. 2).
c) End positions indicated the genomic positions matched to end sequences of the cDNAs. End positions of the annotated genes except for KIAA1043 were fetched from the original paper (ref. 2).
d) Orientation of the genes were indicated as D in the same or R in the opposite direction to the genome.
e) Lengths of genomic regions covered by the corresponding cDNAs.

of which two genes were assigned previously. KIAA1672 spanned two 'predicted' genes from EST data. Therefore, it is possible that these ESTs were derived from a single transcript. Interestingly, the cDNA sequences for KIAA1670, on the other hand, included the nucleotide sequences of choline kinase isolog (CHKL) and membrane carnitine palmitoyltransferase (CPT1B) in this order, although several intronic sequences remained in this cDNA clone (Fig. 1A). To examine whether or not such a fused transcript is present in vivo, RT-PCR and direct sequencing experiments were applied using gene-specific primers as described in “Materials and Methods.” As shown in Fig. 1B, PCR products of 568 bp and 458 bp were predominantly amplified by gene-specific primers.
Table 3. Transcribed sequences which covered multiple annotated genes.

| Cluster number | Locus name | Gene categories or cDNA clone ID | End position 1 (bp) | End position 2 (bp) | Orientation | Genomic size (bp) | Title |
|----------------|------------|----------------------------------|---------------------|---------------------|-------------|------------------|-------|
| cluster_#103   | hg03996    | Gene                             | 2987533             | 3011290             | R           | 14,368           | DiGeorge Syndrome gene H |
|                | AC004471.2 | Gene                             | 2999970             | 3000970             | R           | 1,101            | DiGeorge Syndrome gene 1 |
|                | AC004471.1 | Gene                             | 3001210             | 3011920             | R           | 10,711           |                                 |
| cluster_#207   | fgl02029   | Gene                             | 7568747             | 7647574             | R           | 78,828           | Immunoglobulin lambda-like polypeptide 2 |
|                | IGL1.2     | Gene                             | 7568750             | 7571020             | R           | 2,271            | Argininosuccinate lyase-like |
|                | ASLL       | Gene                             | 7578042             | 7283010             | R           | 4,191            |                                 |
|                | AP000346.2 | Pseudo                           | 7600790             | 7600960             | R           | 171              | GGT related |
|                | AP000347.1 | Pseudo                           | 7644450             | 7647570             | R           | 3,121            | Similar to BETA-GLUCURONIDASE PRECURSOR |
| cluster_#317   | hh11755    | Gene                             | 135445E3            | 13639626            | D           | 95,014           | Neurofibromatosis 2 |
|                | NF2        | Gene                             | 13544800            | 13636110            | D           | 91,301           | Similar to TrO43767, human RIBULOSE-5-PHOSPHATE-EPIMERASE |
|                | AC005529.8 | Pseudo                           | 13550000            | 13550500            | D           | 501              | Matches EST cluster |
|                | AC005529.3 | Predict                          | 13639300            | 13639600            | D           | 301              |                                 |
| cluster_#549   | hh04658    | Pseudo                           | 25205613            | 25222964            | D           | 17,352           | Thyrotrophic embryonic factor |
| TEF            | bK223H9.3  | Gene                             | 25205700            | 25219700            | D           | 14,001           | Matches EST cluster |
| cluster_#602   | hq02515    | Predict                          | 27379288            | 27499478            | R           | 120,271          | Matches EST sequences |
| (KIAA1672)     | dJ345P10.2 | Predict                          | 27427440            | 27443300            | R           | 15,901           | Matches EST sequences |
|                | dJ345P10.3 | Predict                          | 27496300            | 27499600            | R           | 3,301            |                                 |
| cluster_#674   | hq04251    | Gene                             | 34330746            | 34348333            | R           | 14,088           | Carnitine palmitoyltransferase 1, muscle |
| (KIAA1670)     | CPTIB      | Gene                             | 34331000            | 34340300            | R           | 9,301            | Choline kinase-like |
|                | CHKL       | Gene                             | 34340800            | 34344800            | R           | 4,001            |                                 |
| cluster_#682   | hq04543    | Predict                          | 34544511            | 34563850            | D           | 21,340           | Matches EST cluster |
|                | AC02055.4  | Predict                          | 34549600            | 34550000            | D           | 401              | Matches EST cluster |
|                | AC02055.5  | Predict                          | 34553200            | 34553800            | D           | 601              | Matches EST cluster |
|                | AC02055.2  | Pseudo                           | 34559400            | 34559800            | D           | 401              | Similar to TrF29316, Human 60S RIBOSOMAL PROTEIN L23A |

a) Information on the annotated genes was fetched from the original paper (ref. 2).
b) End positions indicated the genomic positions matched to end sequences of the cDNAs. End positions of the annotated genes were fetched from the original paper (ref. 6).
c) Orientation of the genes were indicated as D in the same or R in the opposite direction to the genome.
d) Lengths of genomic regions covered by the corresponding cDNAs.
e) cDNA clone_ID used in our project.
Figure 1. Relationship of KIAA1670, CHKL, and CPT1B genes. A: physical maps of CHKL, CPT1B, and KIAA1670 genes. The predicted CDSs and untranslated regions are shown by solid and open boxes, respectively. The numerals represent the end positions of the respective cDNAs in bp. The KIAA1670-specific primers used in the RT-PCR experiment are indicated by arrowheads. B: the RT-PCR experiment using KIAA1670-specific primers and the first-strand cDNA template derived from poly(A)+ RNA of human whole brain. The RT-PCR products analyzed by electrophoresis using 2% agarose gel are indicated by arrows on the right. The positions of linear DNA size markers are indicated in bp on the left. C: schematic representation of the relationship between the RT-PCR products along the genome. The numerals indicate positions in the sequence of KIAA1670 for the RT-PCR product 2 and in that of chromosome 22q13 BAC clone CIT987SK–384D8 (GenBank accession number, U62317) for the genome. The positions of the stop codon for CHKL and the initiation codon for CPT1B are indicated by arrows.

designed to amplify the region between the CDSs of these two genes. The structures of the transcribed sequences derived from this genomic region were deduced from the results of direct sequencing of the PCR products and are schematically represented along the genome in Fig. 1C. From these results, we conclude that CHKL and CPT1B generated a long transcript in which these two genes are contiguously transcribed. There were at least two splicing variants in this type of the fused transcript since all gaps detected between the PCR products and the genomic sequence were deleted in accordance with the GT-AG rule (Fig. 1C). In each variant, the CDSs of CHKL and CPT1B were not fused. Although this strange gene organization has been already suggested by van der Leij et al. and Yamazaki et al., this is the first report of the actual isolation of a cDNA clone completely covering these two genes. It is very unlikely that this cDNA clone was an artifact since several clones like this were isolated from different libraries. These results suggest that a new type of regulatory mechanism controlling gene expression might govern the functions of CHKL and CPT1B. In addition, other clusters identified in this study, which mostly covered multiple annotated genes, are listed in Table 3. For cluster #100 and #207 in addition to cluster #674 (KIAA1670), two known genes were covered by our single cDNA clone, hh05996 and fg02029, respectively. Regarding the remaining clusters, all included ‘predicted’ genes derived from clustered EST data.
3.4. Transcripts expressed on the opposite strand of \textit{known} genes

As shown in Table 2, two cDNA clones, for KIAA1654 and KIAA1658, were transcribed in the opposite direction to the annotated genes, BCL2-interacting killer and SEC14-like 2, respectively, although they shared the same genomic regions with the corresponding annotated genes. These cDNA clones are unlikely to be accidentally derived from genomic fragments during the cDNA library construction because their sequences appeared to be already spliced although they did not show conspicuous CDSs. In addition, a cDNA clone for KIAA1648 was also found to be transcribed from the opposite strand of the KIAA1043 which was previously identified by us.\textsuperscript{8} Interestingly, the cDNA sequence for KIAA1648 had five introns spliced out according to the GT-AG rule and was polyadenylated with a canonical poly(A) addition signal. Figure 2 represents the relationship of these overlapping genes, KIAA1648 and KIAA1043, in terms of their genomic organization. Notably, the genomic regions used as exons in the respective genes hardly overlap with each other. We cannot at present, however, exclude the possibility that these overlapping genes expressed from different genomic regions duplicated outside of chromosome 22q. Although we do not know whether or not these transcripts act as functional molecules \textit{in vivo}, stable mRNA-like transcripts were reported to be spliced and polyadenylated without defined CDSs in cells such as Xist, H19 and His-1.\textsuperscript{1,8} The functions of these bidirectionally transcribed genes have yet to be determined. As for the cDNA clone for KIAA1656, a relatively large unidentified CDS was predicted only in the opposite direction although the cDNA was polyadenylated with a canonical poly(A) addition signal.

3.5. Concluding remarks

We could experimentally identify some additional transcribed regions on 22q through mapping of ESTs of 30,683 long cDNA clones. The results indicate that the EST mapping of long transcripts well complemented the information obtained by conventional EST mapping, in which relatively short cDNA clones are used. In particular, the information provided by long cDNAs was indispensable for identification of long genes in some cases because long genes have a risk of being incorrectly annotated either by \textit{ab initio} exon prediction programs or by EST mapping of short cDNAs. The results also indicate that randomly sampled cDNAs allocated to 22q did not always contain CDSs. However, if the genomic sequence corresponding to a cDNA is available, it is safer to analyze the genomic sequence prior to entire cDNA sequencing \textit{in silico}, because it could allow us to avoid non-productive sequencing of cDNA clones without any CDSs. The integration of cDNA sequence data with the genomic data, and vice versa, would offer us great benefits for further characterization of genes.

\textbf{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, Kiyo Suni, Takashi Watanabe, Kozue Kaneko, Naoko Shibano, Mina Waki, and Nobue Kashima for their technical assistance.

\textbf{Figure 2.} Relationship between KIAA1648 and KIAA1043 genes along the genome. KIAA1648 and KIAA1043 are mapped to BAC clone, gi4165236 (GenBank accession number, AL033538). Both their cDNA sequences are schematically represented along the genomic sequences. Highly homologous regions between the cDNAs and the BAC sequence are shown in gray. CDSs predicted by GeneMark in respective cDNAs are shown in black bars. Transcription directions of cDNAs are indicated by arrows. The horizontal scales represent the length of the cDNAs or genomic sequences in kb.
References

1. Hattori, M., Fujiyama, A., Taylor, T. D. et al. 2000, The DNA sequence of human chromosome 21, *Nature*, 405, 311–319.

2. Dunham, I., Shimizu, N., Roe, B. A. et al. 1999, The DNA sequence of human chromosome 22, *Nature*, 402, 489–495.

3. 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.

4. Nagase, T., Kikuno, R., Nakayama, M., Hirosawa, M., and Ohara, O. 2000, Prediction of the coding sequences of unidentified human genes. XVIII. The complete sequences of 100 new cDNA clones from brain which code for large proteins *in vitro*, *DNA Res.*, 7, 273–281.

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. Adams, M. D., Kerlavage, A. R., Fleischmann, R. D. et al. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. 1995, *Nature*, 377, supplement, 3–17.

7. Aaronson, J. S., Eckman, B., Blevins, R. A. et al. 1996, Toward the development of a gene index to the human genome: an assessment of high-throughput EST sequence data, *Genome Res.*, 6, 829–845.

8. Kikuno, R., Nagase, T., Ishikawa, K.-I. et al. 1999, Prediction of the coding sequences of unidentified human genes. XIV. The complete sequences of 100 new cDNA clones from brain which code for large proteins *in vitro*, *DNA Res.*, 6, 197–205.

9. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, 25, 3389–3402.

10. Nagase, T., Ishikawa, K.-I., Nakajima, D. et al. 1997, Prediction of the coding sequences of unidentified human genes. VII. The complete sequences of 100 new cDNA clones from brain which code for large proteins *in vitro*, *DNA Res.*, 4, 141–150.

11. Morton, N. E. 1991, Parameters of the human genome, *Proc. Natl. Acad. Sci. USA*, 88, 7474–7476.

12. Burge, C. and Karlin, S. 1997, Prediction of complete gene structures in human genomic DNA, *J. Mol. Biol.*, 268, 78–94.

13. 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.

14. Hirosawa, M., Isono, K., Hayes, W. S., and Borodovsky, M. 1997, Gene Identification and classification in the *Synechocystis* genomic sequence by recursive Gene Mark Analysis, *DNA Seq.*, 8, 17–29.

15. Beaudouing, 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.

16. van der Leij, F. R., Takens, J., van der Veen, A. Y., Terpstra, P., and Kuipers, J. R. G. 1997, Localization and intron usage analysis of the human CPT1B gene for muscle type carnitine palmitoyltransferase I, *Biochim. Biophys. Acta*, 1352, 123–128.

17. Yamazaki, N., Shinohara, Y., Kajimoto, K., Shindo, M., and Terada, H. 2000, Novel expression of equivocal messages containing both regions of choline/ethanolamine kinase and muscle type carnitine palmitoyltransferase I, *J. Biol. Chem.*, 275, 31739–31746.

18. Erdmann, V. A., Szymanski, M., Hochberg, A., de Groot, N., and Barciszewski. J. 1999, Collection of mRNA-like non-coding RNAs, *Nucleic Acids Res.*, 27, 192–195.