Vector-Capping: A Simple Method for Preparing a High-Quality Full-Length cDNA Library

Seishi Kato,1,* Kuniyo Ohtoko,1 Hideki Ohtake,1 and Tomoko Kimura2

Department of Rehabilitation Engineering, Research Institute, National Rehabilitation Center for Persons with Disabilities, 4-1 Namiki, Tokorozawa, Saitama 359-8555, Japan1 and Hitachi Instruments Service Co., Ltd., 4-28-8 Yotsuya, Shinjuku, Tokyo 160-0004, Japan2

(Received 11 November 2004; revised 17 December 2004)

Abstract

Full-length cDNAs play an essential role in identifying genes and determining their promoter regions. Here we describe a simple method for constructing a full-length cDNA library, which has the following advantages: (i) it consists of only three steps including direct ligation between a vector and a cDNA strand using T4 RNA ligase, (ii) it contains neither a PCR process generating mutations nor restriction enzyme treatment causing truncation of cDNA, (iii) the intactness of cDNA is assured due to the presence of an additional dGMP at its 5′ end, (iv) approximately 95% of cDNA clones are full-length when cultured cells or fresh tissues are used, (v) several micrograms of total RNA without mRNA purification is sufficient for preparation of a library containing >105 independent clones, and (vi) a long-sized full-length cDNA up to 9.5 kbp can be cloned. This method will accelerate comprehensive gene analysis in a variety of eukaryotes.

Key words: full-length cDNA; cDNA synthesis; transcription start site; T4 RNA ligase

1. Introduction

Genome sequencing projects of a variety of eukaryotes including human1,2 have been completed during the last decade. The next challenge is to decode whole genetic information recorded by these genomes. In order to achieve this aim, it is necessary to determine the whole loci of genes. A robust way to identify the exon-intron structure and promoter region of each gene is to determine the full sequence of mRNA transcribed from each gene locus and to map each sequence on the genome. Technology for synthesizing complementary DNA (cDNA) from mRNA with reverse transcriptase enabled us to analyze the full sequence of mRNA. The cDNA used for this purpose should be a genuine full-length cDNA that possesses the entire sequence of intact mRNA starting with a cap site and ending with a poly(A) tail.

The most popular method3 for preparing a cDNA library could not provide a full-length cDNA possessing a capped site sequence, because the deletion of terminal sequences occurred during a terminal trimming process after first-strand cDNA synthesis. This defect was overcome by one of two ways: the protection of the 3′ end of the first-strand cDNA or the modification of a cap structure at the 5′ end of intact mRNA. Protection of the 3′ end of the first-strand cDNA was carried out by adding a homo-oligomer tail using terminal deoxynucleotidyl transferase (TdTase)4–6 or by adding a synthetic oligonucleotide using T4 RNA ligase.7 Furthermore, the finding that multiple dCTPs (Cs) are preferentially added to the 3′ end of the first strand of a full-length cDNA by the TdTase activity of reverse transcriptase led to the development of the CapFinder method,8 in which the multiple Cs were used as a base-pairing partner for joining of a primer. The second way is based on the replacement of the cap structure by a synthetic oligomer.9,10 The oligo-capping method has successfully been applied to mass production of full-length cDNA clones10–12 and determination of transcription start sites.13 One defect of the oligo-capping methods, however, is the need for many steps including enzymatic treatment of mRNA, during which the degradation of mRNA frequently occurs and this results in the production of truncated cDNAs.

Any improved method can produce full-length cDNA clones. However, it is unavoidable that the library contains truncated cDNA clones derived from degraded mRNA. Thus, the inability to determine whether or not the library has a genuine full-length cDNA possessing a capped site sequence is a problem common to all methods, because there is no criterion to distinguish full-length cDNAs from truncated ones by their 5′-end sequence data only. Recently, we have found that the
anchor ligation method produces cDNA having an additional dGMP (G) at its 5′ end when capped mRNA but not cap-free mRNA is used as a template, and that the base of the added nucleotide is complementary to the base of the cap.\textsuperscript{14} This finding means that we can guarantee the intactness of the cDNA by the presence of additional G at the 5′ end. So far, the anchor ligation method has been limited to cloning specific genes and has not been used for large-scale analysis of cDNA clones.

In this report, we describe a simple and efficient method for preparing a high-quality cDNA library. This method owes its simplicity to the finding that the 3′ end of the first-strand cDNA in an mRNA/cDNA heteroduplex ligated to the blunt end of a vector DNA using T4 RNA ligase. As a result, full-length cDNAs had an additional G at the 5′ end in the same manner as those obtained by the anchor ligation method.\textsuperscript{14} As the ligation process produces the cDNA capped with the vector, the present method is named ‘vector-capping.’ We assessed this method by preparing and analyzing the cDNA libraries of cultured cells and mouse tissue. The results demonstrated that this method has remarkable advantages in terms of simplicity, efficiency, a full-length rate, and quality of cDNA as compared with conventional methods.

2. Materials and Methods

2.1. RNA preparation

Model mRNAs with cap analogues were described in a previous paper.\textsuperscript{14} The full-length cDNA encoding \textit{EEF1A1} was transcribed using an \textit{in vitro} transcription kit (MEGAscript\textsuperscript{TM}, Ambion) in the presence or absence of a cap analogue, m\textsuperscript{7}G-5′-ppp-5′-G or A-5′-ppp-5′-G (Ambion).

Human retinal pigment epithelium cell line ARPE-19 was obtained from American Type Culture Collection. ARPE-19 cells were cultured in Dulbecco’s modified Eagles medium: nutrient mixture F-12 (Invitrogen) containing 10% fetal bovine serum. Total RNA was isolated from 10\textsuperscript{8} cells of ARPE-19 by the acid phenol-guanidinium thiocyanate-chloroform extraction method.\textsuperscript{15} The A\textsubscript{260}/A\textsubscript{280} ratio of the extracted RNA was 1.6.

One gram of liver dissected from a male BALB/c mouse was homogenized in 10 ml of Isogen (Nippon Gene) with a homogenizer. The homogenate was extracted with 2 ml of chloroform. Total RNA was precipitated from aqueous phase by adding 5 ml of isopropanol. The pellet was washed with 1 ml of 75% ethanol solution and dissolved in 1 ml water. The A\textsubscript{260}/A\textsubscript{280} ratio of the extracted RNA was 2.0.

2.2. Vector primer

Two cloning vectors, pKA1U5 and pGCAP1, were prepared by modifying an original vector pKA1.\textsuperscript{10} pKA1U5 was constructed by substituting the origin of pKA1 with that of pUC19 and the \textit{BstXI} site sequence CCACCCCTGTTG with CCAGGGGGGTGG. pGCAP1 was constructed by substituting the cloning site \textit{EcoRI}-\textit{BstXI}-\textit{EcoRV}-\textit{KpnI} of pKA1U5 with \textit{EcoRI}-\textit{AflII}-\textit{SwaI}-\textit{KpnI}. Plasmids pKA1U5 and pGCAP1 were digested with \textit{KpnI} and tail with about 60 nucleotides of dT using TdTase (TaKaRa) according to the previous report.\textsuperscript{14} After digestion with \textit{EcoRV} (pKA1U5) or \textit{SwaI} (pGCAP1), the dT-tailed plasmid vector was separated on agarose gel and used as a vector primer. The nucleotide sequences of pKA1U5 and pGCAP1 are available from GenBank/EMBL/DDBJ under accession nos. AB191256 and AB191257, respectively.

2.3. cDNA synthesis

A mixture of an RNA sample and a vector primer (0.3 μg of model RNA and 0.15 μg of pGCAP1-derived vector primer, 5 μg of total RNA of ARPE-19 and 0.3 μg of pKA1U5-derived vector primer, or 10 μg of total RNA of mouse liver and 0.15 μg of pGCAP1-derived vector primer) was incubated at 65°C for 5 min. The first-strand cDNA was synthesized in 20 μl (or 40 μl for mouse liver) of a reaction mixture containing 50 mM Tris-HCl (pH 8.3); 75 mM KCl; 3 mM MgCl\textsubscript{2}; 5 mM DTT; 1.25 mM each dATP, dCTP, dGTP, and dTTP; 0.03 U/μl of RNasin (TaKaRa); and 1.2 U/μl of SuperScriptII\textsuperscript{TM} (Invitrogen). The reaction mixture was incubated at 42°C for 1 hr (model RNA) or 3 hr (total RNA). After phenol/chloroform extraction followed by ethanol precipitation, the pellet was dissolved in water.

Self-ligation was performed in 100 μl of a reaction mixture containing 25 mM Tris-HCl (pH 7.5), 5 mM MgCl\textsubscript{2}, 10 mM β-mercaptoethanol, 2 mM DTT, 0.5 mM ATP, 25% polyethylene glycol (PEG6000, Nakarai), 0.4 U/μl of RNasin (TaKaRa), and 1.2 U/μl of T4 RNA ligase (TaKaRa). The reaction mixture was incubated at 20°C overnight. After phenol/chloroform extraction followed by ethanol precipitation, the pellet was dissolved in water.

Second-strand synthesis was performed in 100 μl of a reaction mixture containing 20 mM Tris-HCl (pH 7.5), 4 mM MgCl\textsubscript{2}, 10 mM β-mercaptoethanol, 0.1 mM each dATP, dCTP, dGTP, and dTTP, 0.03 U/μl of RNaseH (TaKaRa), 0.3 U/μl of \textit{Escherichia coli} DNA polymerase I (New England Biolabs), and 0.6 U/μl of \textit{E. coli} DNA ligase (TaKaRa). The reaction mixture was incubated at 12°C for 5 hr. After two-times phenol/chloroform extraction followed by ethanol precipitation, the pellet was dissolved in 50 μl of TE.
2.4. Construction of cDNA library

One microliter of a cDNA sample was mixed with 20 µl of electrocompetent cell DH12S (Invitrogen), and a single electrical pulse was applied using an electroporation apparatus (MicroPulser, BioRad). Transformants were suspended in 1 ml of an SOC medium and incubated with gentle rotation at 37°C for 1 hr. The electroporated cell solution was plated onto LB agar. All colonies grown on the plates were picked and suspended in LB medium in 96-well plates. After incubation and subsequent addition of glycerol, the original plates were stored at −80°C.

2.5. Plasmid extraction and sequencing

Plasmid DNA was extracted with a plasmid isolation system (MG768, Hitachi Koki). The 5′ sequence of cDNA was determined with a capillary DNA sequencer ( Applied Biosystems 3730x/DNA analyzer) using a BigDye Terminator Cycle Sequencing FS Ready Reaction Kit. The full sequences of long-sized cDNAs were determined by a primer walking method. The determined sequences were deposited to GenBank/EMBL/DDBJ under accession nos. AB191258–AB191264. Similarity search was performed on the web site of the National Center for Biotechnology Information using BLAST search.

3. Results

3.1. Vector-capping protocol

The present method comprises only three steps as shown in Fig. 1: (i) a process for synthesizing first-strand cDNA with reverse transcriptase using a vector primer that has a dT tail at one end of a linear vector, (ii) a process for joining the first-strand cDNA to the other end of the vector primer with T4 RNA ligase, and (iii) a process for replacing mRNA by DNA.

The first process is essentially the same as described in the Okayama-Berg method.⁴ The use of the vector primer results in a directional insertion of cDNA, providing an advantage in sequencing and expression of the cDNA. The vector primer can be prepared from any plasmid vector having the following restriction enzyme sites: a 3′ protruding site for dT tailing (Kpn I in this experiment) and its proximal site for removal of the dT tail of one end (EcoRV or Swa I in this experiment). The critical factor for preparing a high-quality cDNA library is the length of the dT tail and the purity of the vector primer. The optimum length of the dT tail is around 60 nucleotides as shown in the original report.⁴ Uncut and untailed vectors, increasing the background “noise” of the library, should be removed as much as possible.

Both poly(A)+ RNA and total RNA can be used as a template RNA. Owing to the difficult purification of poly(A)+ RNA and to the degradation of RNA during the purification step, total RNA is preferable. The optimum amount of total RNA for synthesis depends on the content of poly(A)+ RNA and the amount of the vector primer used. The ratio of total RNA to the vector primer should be examined in pilot experiments. At the optimum conditions, no band of the vector primer was observed by agarose gel electrophoresis after first-strand cDNA synthesis. In the present experiment, libraries containing 10⁵–10⁶ independent clones were prepared using 5–10 µg of total RNA and 0.15–0.3 µg of the vector primer.

The characteristic and key process of the present method is the second step, the circularization of a cDNA-vector primer duplex with T4 RNA ligase. T4 RNA ligase has been known to catalyze the ligation between single-stranded DNAs¹⁶,¹⁷ as well as single-stranded RNAs. However, there is no report that it catalyzes the ligation between a DNA/RNA heteroduplex and a DNA duplex. Especially, our case was unexpected, because the terminus of the mRNA/cDNA heteroduplex has a protruding cap structure that may hinder the ligation between the
Table 1. The 5′-terminal sequences of EEF1A1 cDNAs prepared from model RNAs.

| Type                      | Sequence of junction between vector and cDNAa | Number |
|---------------------------|-----------------------------------------------|--------|
| m7G-capped RNA            |                                               |        |
| Full-length clones        | . . . gaattccttaagtttaaatgtgGCTTTTTTCGCAA . . | 38     |
|                           | . . . gaattccttaagtt-------GCTTTTTTCGCAA . . | 2      |
|                           | . . . gaattccttaagatt-------GCTTTTTTCGCAA . . | 2      |
| Truncated clones          |                                               | 5      |
| Total                     |                                               | 43     |
| A-capped RNA              |                                               |        |
| Full-length clones        | . . . gaattccttaagtttaaatgtgAGCTTTTTTCGCAA . . | 30     |
|                           | . . . gaattccttaagtt-------AAGCTTTTTTCGCAA . . | 1      |
|                           | . . . gaattccttaagatt-------AGCTTTTTTCGCAA . . | 27     |
| Truncated clones          |                                               | 7      |
| Total                     |                                               | 37     |
| Cap-free RNA              |                                               |        |
| Full-length clones        | . . . gaattccttaagatt-GCTTTTTTCGCAA . .       | 29     |
|                           | . . . gaattccttaagattT9GCTTTTTTCGCAA . .      | 28     |
| Truncated clones          |                                               | 9      |
| Total                     |                                               | 38     |

aThe sequences of the vector and cDNA are indicated in lower and upper cases, respectively. Underlined letters represent an inserted nucleotide.

3′ end of the cDNA and the 5′ end of the DNA duplex. Our results show that the addition of an oligonucleotide to the 3′ end of cDNA with reverse transcriptase allows these duplexes to join. On the hand, the use of T4 DNA ligase resulted in a high-background library composed of vectors with no insert.

The ligation reaction requires a high concentration of polyethyleneglycol, as shown in the other reactions using T4 RNA ligase. The ligation efficiency was so low that we could not detect the ligation products by agarose gel electrophoresis. Nevertheless, the number of transformants was nearly sufficient for ordinary uses of the cDNA library. The present experiments use the conditions for RNA ligation. When the optimized conditions for single-stranded DNA ligation are used, no transformant was obtained. T4 RNA ligase intrinsically catalyzes the joining of single strands. If the blunt end of the vector is converted to the 5′ protruding end by restriction enzyme treatment, the ligation efficiency is expected to increase. When an EcoRI digestion process was added after the first-strand cDNA synthesis, slight increase of the yield and reduction of the “background” clones with no insert were observed.

The third step is a procedure to replace an mRNA strand by DNA using RNase H, E. coli DNA polymerase I and E. coli DNA ligase. If the ligation products are not stored for a long time, this third step can be omitted because the repair reaction proceeds in the E. coli cells after transformation. The yield of transformants depends on the transformation efficiency. An electroporation method is preferentially used, and a transformation efficiency of >10^9 colonies per microgram of control plasmid DNA is necessary. To avoid the generation of copied clones, we did not amplify the library in liquid medium. Instead, a small portion of the final reaction product was used to transform E. coli cells and all independent colonies grown on agar plates were picked for glycerol stock of the cDNA library.

3.2. Cap-dependent nucleotide addition
The vector-capping protocol was tested using model mRNAs that were prepared by in vitro transcription of human eukaryotic translation elongation factor 1 alpha 1 (EEF1A1) cDNA in the presence of a cap analog such as 7-methylG(m7G)-5′-ppp-5′-G and A-5′-ppp-5′-G. Using 0.3 µg of mRNA and 0.15 µg of a vector primer, 5 × 10^4 transformants were obtained. There was no significant difference in yield among three libraries constructed from m7G-capped, A-capped, and cap-free mRNA. Table 1
shows the sequence of the junction between a cDNA insert and the vector. The end of the vector should be a blunt-ended Swa I site (..attt), but some clones have an additional vector sequence up to the blunted Kpn I site (..atttaaatgtg), which may be generated by incomplete digestion with Swa I during preparation of the vector primer. When cap-free RNA was used, 28 out of 38 clones contained a full-length cDNA starting with the 5′ end of the transcript. This means that the blunt end of a cap-free RNA/cDNA duplex joined to the blunt-ended Swa I site of the vector primer. On the other hand, when a cap-free RNA/cDNA duplex joined to the blunt-ended Swa I site during preparation of the vector primer. When cap-free RNA was used, 28 out of 38 clones contained a full-length cDNA starting with the 5′ end of the transcript. This means that the blunt end of the transcript. This means that the blunt end of the transcript. This means that the blunt end of a cap-free RNA/cDNA duplex joined to the blunt-ended Swa I site of the vector primer. On the other hand, when m7G-capped RNA and A-capped RNA were used as a template, the full-length cDNAs had an additional G and A at their 5′ end, respectively. The truncated cDNA may be derived from the degraded RNA. It is noted that most truncated cDNAs had no additional nucleotide at the 5′ end. These results are consistent with those of the anchor ligation method described in the previous paper, suggesting that cap-dependent oligonucleotide addition due to the TdTase activity of reverse transcriptase occurred during the first-strand cDNA synthesis.

### 3.3. Construction of cDNA library using total RNA

The vector-capping method was applied to construct cDNA libraries from total RNA of human retinal pigment epithelium cell line ARPE-19 and mouse liver. The cDNA library of ARPE-19 was prepared using 5 µg of total RNA and 0.3 µg of a pKA1U5 vector primer. On the other hand, the cDNA library of mouse liver was prepared using 10 µg of total RNA and 0.15 µg of a pGCAP1 vector primer. After transformation of E. coli cells, transformants were grown on agar plates without amplification in liquid medium. The libraries were composed of approximately 10⁶ and 10⁷ independent colonies, respectively. From the two libraries, 4800 and 1152 colonies were randomly picked, respectively. Sequencing the 5′ end of these clones, we obtained readable sequences of 4084 and 791 cDNA inserts from ARPE-19 and mouse liver, respectively.

### 3.4. Quality of ARPE-19 cDNA library

The model experiment using m7G-capped RNA of EEF1A1 showed that a full-length cDNA has an additional G at the 5′ end. To determine the proportion of full-length cDNAs having an additional G in the ARPE-19 library, we analyzed the 5′-end sequences of abundant cDNA clones. Figure 2A and 2B show the 5′ sequence alignment of top two abundant genes, glyceraldehyde 3-phosphate dehydrogenase (GAPD, 55 clones, 1.3% in content) and actin gamma 1 (ACTG1, 44 clones, 1.1%), together with the corresponding genome sequences. All 55 GAPD clones had a full-length cDNA insert possessing an additional G or NG (where N indicates extra nucleotides that contain T and/or G in most cases) that does not exist in the genome sequence. The transcription start sites of both genes were diverse, which is consistent with the results obtained by the oligo-capping method. The most frequent sequence of each gene, which would be a main transcript, started with one additional G. Some ACTG1 clones started with G existing in the genome sequence so that we can not determine whether these GS originate from the transcript or the added G. However, GS located near the starting sites of the other full-length clones would certainly result from the added nucleotide.

Table 2 summarizes the 5′-terminal nucleotides of 14 abundant genes with >20 clones (>0.5% in content). The 5′-end sequences of their clones are shown in Supplementary Table 1 (http://www.dna-res.kazusa.or.jp/12/1/05/supplement/supplement11.html). Only 8 clones out of total 455 were truncated cDNAs so that the full-length rate was 98.2%. Of the full-length cDNA clones, 429 clones (96.0%) started with G or NG at the 5′ end, in which 390 ends were derived from an additional G not existing in the genome sequence. On the other hand, 3 out of 8 truncated clones started with G existing in the genome sequence. Consequently, the content of a full-length cDNA in total clones starting with G or NG was 99.3% (429 out of 432). Minor clones carrying a full-length cDNA possessed an additional A (one clone) or T (four clones) not existing in the genome sequence. These clones may result from non-templated nucleotide addition.

### 3.5. Quality of mouse liver cDNA library

In general, it is easy to isolate intact mRNA from cultured cells as compared with that from tissues. To examine the quality of a cDNA library prepared from tissue, we analyzed the 5′-end sequences of mouse liver cDNA clones. Table 3 shows the distribution of the 5′-terminal nucleotides of 13 abundant genes with >8 clones (>1% in content). The full-length rate was 98.4% (246 clones out of 250) which is comparable to that of the ARPE-19 library. Of the full-length cDNA clones, 230 clones (93.5%) started with G or NG at the 5′ end, in which 185 ends were derived from an additional G not existing in the genome sequence. Every gene showed the diversity of the transcription start site (Supplementary Table 2, http://www.dna-res.kazusa.or.jp/12/1/05/supplement/supplement2.html).

The most abundant clone (70 clones, 8.8% in content) encoded major urinary protein 1, in which only one was a truncated clone. Almost all full-length clones started with G or NG. The second abundant gene was albumin 1 (35 clones, 4.4% in content), 33 clones of which had a full-length insert. Their 5′-end sequences are aligned in Fig. 2C. The transcription of 32 clones started with A, and 19 clones of them had an extra G or NG at the 5′ end, but 12 clones had no additional nucleotide, unlike the other genes. The absence of an addition of nucleotide was
observed in transthyretin (Ttr) cDNA and several cDNAs obtained from the ARPE-19 library (Table 2). The common feature of these genes is that the capped site sequence starts with a nucleotide sequence AC (Ttr, ACTB and MT2A) or a pyrimidine-rich sequence (EEFIA1 and RPS3), suggesting that the 5' end sequence of mRNA may affect the addition or elimination of a cap structure in the cells.
3.6. Long-sized full-length cDNA clones

One of requirements for a high-quality cDNA library is to contain long-sized full-length cDNA clones. The average length of 14 abundant genes isolated from the ARPE-19 library was 1.1 kbp (Table 2). To examine the effect of a cDNA insert size on a full-length rate, we analyzed 93 clones of 15 genes with >2.5 kbp in length (average 3.2 kbp) and >0.1% in content (>4 clones) (Table 4). The full-length rate was 95.7% (89 clones), suggesting that the cDNA insert size slightly affected the full-length rate.

This result was confirmed by the presence of 8 full-length cDNAs with a length of >7 kbp in the ARPE-19 cDNA library prepared without a size selection process. The full sequences of these clones were determined and are listed in Table 5. All clones started with G, and 7 clones had an extra G not existing in the genome sequence. Only agrin produced a truncated clone. Filamin A and myosin heavy polypeptide 9 generated two full-length clones with different 5′-end sequences, which may reflect the mRNA population. The size differences of the cDNA and encoded protein are likely due to different transcription start sites, alternative splicing, and/or alternative poly(A) addition. The clones coding for filamin A and B, and fibronectin 1 are novel splicing variants. The agrin clone is a full-length cDNA reported for the first time. It is noted that their sequences registered in the reference sequence data (RefSeq) were produced by joining the cDNA fragments. These results demonstrate that the present method is also powerful for long-sized full-length cDNA cloning.
Table 4. The 5′-terminal nucleotides of abundant long-sized clones isolated from ARPE-19 cDNA library.

| Gene symbol | RefSeq | Size (bp) | Number | Full-length | Insertion | Insertion? | No insertion | Sum | Truncated |
|-------------|--------|-----------|--------|-------------|-----------|-----------|-------------|------|-----------|
| PABPC1      | NM_002586.2 | 2554 | 18 | 18 | 0 | 0 | 0 | 0 | 18 | 0 |
| SERPINE1    | NM_006602.1 | 2876 | 8 | 7 | 0 | 1 | 0 | 0 | 8 | 0 |
| OK/SW-cl.S6 | NM_178014.2 | 2510 | 8 | 4 | 1 | 0 | 0 | 1 | 7 | 1 |
| PSAP        | NM_002778.1 | 2767 | 7 | 5 | 1 | 0 | 0 | 0 | 6 | 1 |
| FSTL1       | NM_007085.3 | 3705 | 6 | 4 | 0 | 0 | 0 | 0 | 4 | 2 |
| HSPCB       | NM_007355.2 | 2567 | 5 | 3 | 0 | 1 | 0 | 1 | 5 | 0 |
| NAP1L1      | NM_139207.1 | 3582 | 5 | 4 | 0 | 0 | 1 | 0 | 5 | 0 |
| VIL2        | NM_003379.3 | 3166 | 5 | 4 | 0 | 0 | 1 | 0 | 5 | 0 |
| TGFB1       | NM_000358.1 | 2691 | 5 | 2 | 1 | 1 | 1 | 0 | 5 | 0 |
| TGFBI       | NM_199512.1 | 3981 | 5 | 4 | 0 | 0 | 1 | 0 | 5 | 0 |
| APP         | NM_000484.2 | 3641 | 5 | 5 | 0 | 0 | 0 | 0 | 5 | 0 |
| TRA1        | NM_000392.1 | 2760 | 4 | 4 | 0 | 0 | 0 | 0 | 4 | 0 |
| SOSTM1      | NM_003900.3 | 2868 | 4 | 0 | 0 | 0 | 0 | 0 | 4 | 0 |
| MANT2       | NM_002380.2 | 3496 | 4 | 4 | 0 | 0 | 0 | 0 | 4 | 0 |
| ACTN1       | NM_001102.2 | 3398 | 4 | 4 | 0 | 0 | 0 | 0 | 4 | 0 |

*The clones with >2.5 kbp in length were selected. The average length is 3.2 kbp. These Gs exist in genome sequence.

Table 5. Characterization of long-sized full-length cDNA clones isolated from ARPE-19 cDNA library.

| Gene name, non-erythrocyt | Accession | cDNA (bp) | Protein (aa) | Genome and 5′-terminal sequence of cDNA* |
|---------------------------|-----------|-----------|--------------|--------------------------------------|
| Filamin, beta             | AC114399.2 | 9421      | 2502         | gcgcggcgcagccagcagccgcgcatccgattcga |
| Reo23004                  | NM_001457.1 | 9405      | 2519         | gcgcggcgcagccagcagccgcgcatccgattcga |
| Filamin, alpha            | BX936346.4 | 8368      | 2647         | actctcggcgcagccagcagccgcgcatccgattcga |
| Reo06005                  | NM_001456.1 | 8212      | 2612         | gcgcggcgcagccagcagccgcgcatccgattcga |
| Fibronectin 1             | AC0124625  | 8272      | 2296         | gcgcggcgcagccagcagccgcgcatccgattcga |
| Reo25009                  | NM_212476.1 | 7753      | 2265         | gcgcggcgcagccagcagccgcgcatccgattcga |
| Spectrin, alpha, non-erythrocyt | NM_003127.1 | 7787      | 2472         | gcgcggcgcagccagcagccgcgcatccgattcga |
| Reo45002                  | NM_003126.1 | 7791      | 2452         | gcgcggcgcagccagcagccgcgcatccgattcga |
| Myosin, non-muscle         | CR383703.1 | 7454      | 1960         | gcgcggcgcagccagcagccgcgcatccgattcga |
| Heavy polypeptide, 9, non-muscle | NM_002473.3 | 7436      | 1960         | gcgcggcgcagccagcagccgcgcatccgattcga |
| Agrin                     | AL654608.30 | 7302      | 2045         | gcgcggcgcagccagcagccgcgcatccgattcga |

*Underlined letters represent nucleotides not existing in the genome sequence. This sequence starts further 185 nucleotides upstream.

4. Discussion

Here we describe a simple method for preparing a high-quality cDNA library, solving many problems involved with the conventional methods. Our method has six advantages: (i) it consists of only three steps including direct ligation between a vector and a cDNA strand using T4 RNA ligase, (ii) it contains neither a PCR process generating mutations nor restriction enzyme treatment causing truncation of cDNA, (iii) the intactness of cDNA is assured due to the presence of an additional G at its 5′ end, (iv) approximately 95% of cDNA clones are full-length when cultured cells or fresh tissues dissected from an animal organ are used, (v) several micrograms of total RNA is sufficient for preparation of a library containing >10⁵ independent clones, and (vi) a long-sized full-length cDNA up to 9.5 kbp can be cloned.

The present method is based on our finding that...
T4 RNA ligase catalyzes the ligation between an mRNA/cDNA heteroduplex and a blunt-ended DNA duplex. The double-stranded vector DNA corresponds to a single-stranded oligonucleotide used in the anchor ligation method so that the two methods should provide cDNA showing the same 5′-end feature. In fact, experiments using model RNA templates showed that the 5′ end of a full-length cDNA synthesized by either method possessed a cap-dependent additional nucleotide, G when m7G-capped mRNA was used. Thus, the anchor ligation method also can be used to prepare the cDNA library. However, the present method using the vector primer has several advantages: (i) PCR, which frequently causes artificial mutations of the cDNA sequence, is not needed, (ii) definite unidirectional insertion of cDNA, (iii) no truncation by restriction enzyme digestion, (iv) a smaller bias by cDNA size when the cDNA is inserted into a vector, and (v) no chimeric products.

The cap-dependent G addition should occur when using the tailing method. Unfortunately, Okayama and Berg and Pruitt used a C tailing so that they missed the cap-dependent C addition to the 3′ end of cDNA, because they could not determine whether the C at the 3′ end of the cDNA was due to the cap-dependently added C or the tailed C. On the other hand, the Cap trapper method that used a G tailing would produce full-length cDNA with an additional G. In fact, Seki et al. reported that many cDNA clones obtained by the Cap trapper method had a G at the 5′ end, but they did not search further for its cause. This kind of nucleotide addition to the 3′ end of the first-strand cDNA would occur by TdTase activity of reverse transcriptase. The addition of multiple C has been reported under certain conditions. Using this multiple C addition, the CapFinder and CapSelect methods have been developed to trap the capped site sequence. These methods probably cause modification of the 5′ end of cDNA by adding artificial Gs derived from a linker oligonucleotide so that it is impossible to determine whether the G at the 5′ end results from cap-dependent G addition or the linker sequence. The present method can avoid this problem by using a vector whose 5′ joining site to the 3′ of cDNA is not C.

To assess the present method, two cDNA libraries were constructed from cultured cells and mouse liver. The sequencing analyses of these libraries showed incredibly high content of full-length cDNA clones despite the absence of any selection step such as the modification of a cap structure used in conventional methods. One explanation of the high full-length rate is that the ligation step can be regarded as a full-length selection step, because the mRNA/cDNA heteroduplex produced by premature termination of reverse transcriptase reaction would not serve as the substrate for self-ligation owing to the steric hindrance of a protruding mRNA strand. Nevertheless, the truncated cDNA should be produced from degraded mRNAs especially derived from long mRNAs. However, the size of mRNA little affected the full-length rate. Besides, full-length cDNAs (7–9.5 kbp) were cloned without a size selection procedure. On the other hand, when in vitro transcribed model RNAs were used, 10–25% of the obtained clones had a truncated cDNA derived from a degraded RNA product. These results suggest that most mRNA molecules extracted from fresh cells have an intact form rarely contaminated by degraded products probably because of the high degradation rate of unnecessary mRNA in the cells. The use of total RNA and the absence of an mRNA modification process would also minimize degradation of mRNA.

The availability of a small amount of total RNA instead of poly(A)⁺ RNA was an unexpected feature of the present method as well as the ligation between an mRNA/cDNA heteroduplex and a blunt-ended DNA duplex with T4 RNA ligase. When we used a few micrograms of mRNA according to the ordinary conditions described in the conventional methods, the resulting library mainly contained short-sized cDNA clones. On the contrary, the library prepared using a few micrograms of total RNA unexpectedly contained >10⁵ independent clones including many long-sized full-length cDNA clones (up to 9.5 kbp in length). The use of a small amount of substrate mRNA may contribute to the effective synthesis of long-sized cDNAs, because reverse transcriptase may not be consumed by synthesis of abundant short mRNA-derived cDNAs.

The present method solves many problems with conventional methods, but it has still two limitations. One is that this method has no experimental process for selecting the full-length cDNA clones so that it is difficult to construct a high-quality cDNA library from an RNA sample containing highly degraded mRNA. This disadvantage may partly be compensated by the ability to select full-length cDNA clones by looking for the presence of an additional G at the 5′ end. However, it may not guarantee the full-length cDNA in the case of a low-quality cDNA library containing a lot of truncated cDNA clones, because a portion of truncated cDNAs might start with G including an extra G added by non-template nucleotide addition. The second problem is the low efficiency of the ligation reaction with T4 RNA ligase. This is not so serious because the present library size is sufficient for ordinary use. Improving the ligation efficiency will enable us to prepare a cDNA library from a smaller amount of RNA sample.

The vector-capping method is now being applied to construct cDNA libraries from various tissues of various species in many laboratories. A great advantage of this method is that only 1/100th of the amount of total RNA necessary for conventional methods is sufficient for the construction of a high-quality cDNA library. Thus, this method is effective when only a small amount of total RNA sample is available. Furthermore, this method
would be useful for the precise analysis of an expression profile because of its low bias in representing the RNA population in the cell. Consequently, this method is a powerful tool for analyzing gene and protein networks in the post genome-sequencing era.

Acknowledgements: This work was partly supported by a grant from the Ministry of Health, Labor and Welfare of Japan.

References
1. Lander, E. S., Linton, L. M., Birren, B. et al. 2001, Initial sequencing and analysis of the human genome, Nature, 409, 860–921.
2. Venter, J. C., Adams, M. D., Myers, E. W. et al. 2001, The sequence of the human genome, Science, 291, 1304–1351.
3. Gubler, U. and Hoffman, B. J. 1983, A simple and very efficient method for generating cDNA libraries, Gene, 25, 263–269.
4. Okayama, H. and Berg, P. 1982, High-efficiency cloning of full-length cDNA, Mol. Cell. Biol., 2, 161–170.
5. Pruitt, S. C. 1988, Expression vectors permitting cDNA cloning and enrichment for specific sequences by hybridization/selection, Gene, 66, 121–134.
6. Carninci, P., Kvam, C., Kitamura, A. et al. 1996, High-efficiency full-length cDNA cloning by biotinylated CAP trapper, Genomics, 37, 327–336.
7. Apte, A. N. and Siebert, P. D. 1993, Anchor-ligated cDNA libraries: a technique for generating a cDNA library for the immediate cloning of the 5′ ends of mRNAs, Biotechniques, 15, 890–893.
8. Clontech Laboratories 1996, CLONTECniques, 11, 2–4.
9. Maruyama, K. and Sugano, S. 1994, Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides, Gene, 138, 171–174.
10. Kato, S., Sekine, S., Oh, S. W. et al. 1994, Construction of a human full-length cDNA bank, Gene, 150, 243–250.
11. Suzuki, Y., Yoshitomo-Nakagawa, K., Maruyama, K., Suyama, A., and Sugano, S. 1997, Construction and characterization of a full-length-enriched and a 5′-end-enriched cDNA library, Gene, 200, 149–156.
12. Ota, T., Suzuki, Y., Nishikawa, T. et al. 2004, Complete sequencing and characterization of 21,243 full-length human cDNAs, Nature Genet., 36, 40–45.
13. Suzuki, Y., Tsunoda, T., Sese, J. et al. 2001, Identification and characterization of the potential promoter regions of 1031 kinds of human genes, Genome Res., 11, 677–684.
14. Ohtake, H., Ohtoko, K., Ishimaru, Y., and Kato, S. 2004, Determination of the capped site sequence of mRNA based on the detection of cap-dependent nucleotide addition using an anchor ligation method, DNA Res., 11, 305–309.
15. Chomczynski, P. and Sacchi, N. 1987, Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction, Anal. Biochem., 162, 156–159.
16. Snopek, T. J., Sugino, A., Agarwal, K. L., and Cozzarelli, N. R. 1976, Catalysis of DNA joining by bacteriophage T4 RNA ligase, Biochem. Biophys. Res. Commun., 68, 417–424.
17. Tessier, D. C., Brousseau, R., and Vernet, T. 1986, Ligation of single-stranded oligodeoxyribonucleotides by T4 RNA ligase, Anal. Biochem., 158, 171–178.
18. Suzuki, Y., Taira, H., Tsunoda, T. et al. 2001, Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites, EMBO Rep., 2, 388–393.
19. Chen, D. and Patton, J. T. 2001, Reverse transcriptase adds nontemplated nucleotides to cDNAs during 5′-RACE and primer extension, Biotechniques, 30, 574–582.
20. Seki, M., Carninci, P., Nishiyama, Y., Hayashizaki, Y., and Shinozaki, K. 1998, High-efficiency cloning of Arabidopsis full-length cDNA by biotinylated CAP trapper, Plant J., 15, 707–720.
21. Schmidt, W. M. and Mueller, M. W. 1999, CapSelect: a highly sensitive method for 5′ CAP-dependent enrichment of full-length cDNA in PCR-mediated analysis of mRNAs, Nucleic Acids Res., 27, e31.