doi:10.1093/dnares/dsn010

# Fine Expression Profiling of Full-length Transcripts using a Size-unbiased cDNA Library Prepared with the Vector-capping Method

Mio OSHIKAWA[1,2], Yoshiko SUGAI[1,2], Ron USAMI[2], Kuniyo OHTOKO[1], Shigeru TOYAMA[1], and Seishi KATO[1,*]

*Department of Rehabilitation Engineering, Research Institute, National Rehabilitation Center for Persons with Disabilities, 4-1 Namiki, Tokorozawa, Saitama 359-8555, Japan[1] and Department of Biological Applied Chemistry, Graduate School of Engineering, Toyo University, Kujirai 2100, Kawagoe, Saitama 350-8585, Japan[2]*

(Received 24 December 2007; accepted on 10 April 2008; published online 16 May 2008)

## Abstract

**Recently, we have developed a vector-capping method for constructing a full-length cDNA library. In the present study, we performed in-depth analysis of the vector-capped cDNA library prepared from a single type of cell. As a result of single-pass sequencing analysis of 24 000 clones randomly isolated from the unamplified library, we identified 19 951 full-length cDNA clones whose intactness was confirmed by the presence of an additional G at their 5' end. The full-length cDNA content was >95%. Mapping these sequences to the human genome, we identified 4513 transcriptional units that include 36 antisense transcripts against known genes. Comparison of the frequencies of abundant clones showed that the expression profiles of different libraries, including the distribution of transcriptional start sites (TSSs), were reproducible. The analysis of long-sized cDNAs showed that this library contained many cDNAs with a long-sized insert up to 11 199 bp of golgin B, including multiple slicing variants for filamin A and filamin B. These results suggest that the size-unbiased full-length cDNA library constructed using the vector-capping method will be an ideal resource for fine expression profiling of transcriptional variants with alternative TSSs and alternative splicing.**

**Key words:** full-length cDNA; expression profile; transcriptional start site; alternative splicing; antisense transcript

## 1. Introduction

The Human Genome Project disclosed sequence information on an entire set of elements that construct and regulate each human cell, tissue, and organ.[1–3] Our next challenge is to decode the instructions encoded by three billion bases on the genome and to understand the molecular basis of human biology and pathology. In the last decade, post-sequencing projects have taken two approaches: (i) to collect an entire set of full-length transcripts (transcriptome) and in turn map them on the genome to identify their promoter region, and (ii) to compare the expression profiles of genes in different types of cells or tissues at different biological conditions. To achieve these purposes, many technologies have been developed: practical methods for synthesizing full-length cDNA,[4–7] hybridization-based microarray,[8,9] and tag-sequencing-based technologies[10–14] for analyzing an expression profile.

Large-scale analyses using these technologies have produced an enormous amount of data on human full-length transcripts,[15–17] as well as on expression profiles,[18–20] and have revealed the following facts. (i) Multiple transcripts of different sizes are generated

Edited by Minoru Ko

* To whom correspondence should be addressed. Tel. +81 4-2995-3100. Fax. +81 4-2995-3132. E-mail: seishi@rehab.go.jp

from the same gene locus by alternative promoter usage,[21] diverse transcriptional initiation,[22] alternative splicing,[23] and alternative polyadenylation.[24] (ii) There are an unexpectedly large amount of sense–antisense pairs of transcripts.[25,26] (iii) Many non-coding RNAs are transcribed.[15,16] (iv) A considerable amount of transcripts estimated by expressed sequence tags (EST) of the UniGene database[27] or by genome tiling array[28,29] still remain to be annotated because of their low abundance.

These data sets were accumulated using diverse cell types or tissues, because the first phase of the project was aimed at comprehensive gene collection. Hereafter, a second phase of investigation requires in-depth analysis of transcripts in a single type of cell population to elucidate the intracellular transcriptional regulatory network in a given cell. However, the full-length cDNA libraries constructed using the conventional methods do not meet the requirements for in-depth analyses because of the low complexity of the constituents. Furthermore, the presence of transcriptional variants raises a serious problem regarding the use of the conventional method for expression profiling of genes. If multiple variants transcribed from the same gene locus have a different function, we need to determine the expression profile of each variant to elucidate their role in the regulation network. However, the conventional methods that measure the amount of the limited region of each transcript by counting the number of sequence tags or by quantifying the hybridization signal cannot distinguish these variants without the precedent information of the full sequence of all variants expressed in a given cell.

The two problems described above, the low complexity of the full-length cDNA library and the presence of multiple variants, can be solved by analyzing all variants that comprise a bias-free full-length cDNA library constructed from a single type of cell. Unfortunately, the conventional methods for synthesizing full-length cDNA were unsuitable for constructing a bias-free cDNA library that reflects an expression level of each transcript in the cell, because they have many processes during which the ratio of components may change. It is especially difficult to construct a bias-free library containing rare or long-sized full-length cDNA clones without changing their content.

Recently, we have developed the vector-capping (V-capping) method for synthesizing full-length cDNA.[30] This method is expected to be suitable to construct a bias-free cDNA library, because it consists of only three steps: the first-strand cDNA synthesis using a vector primer, self-ligation of the cDNA-vector construct, and the replacement of mRNA by the second-strand cDNA. The previous paper showed that we were able to construct the cDNA libraries containing full-length cDNA clones of >95% content without any selection procedure for full-length cDNAs. The further advantage of this library is that we can validate full-length cDNA by the presence of an additional G at its 5' end.

In this paper, we performed in-depth analysis of the cDNA libraries constructed using the V-capping method from the total RNA isolated from human retinal pigment epithelial cell line ARPE-19, and demonstrated that the constructed library was useful as a starting resource not only for the comprehensive collection of full-length cDNA clones, but also for fine expression profile analysis of transcripts expressed in a single type of cell, including variants generated by alternative promoter usage, alternative transcriptional initiation, alternative splicing, and alternative polyadenylation.

## 2. Materials and methods

### 2.1. Cell culture and RNA preparation

Human retinal pigment epithelium (RPE) cell line ARPE-19 was obtained from American Type Culture Collection (Manassas, VA, USA). ARPE-19 cells were cultured in Dulbecco's modified eagle's medium: nutrient mixture F-12 (Invitrogen, Carlsbad, CA, USA) containing 10% fetal bovine serum. The cells were incubated for 4 days to confluence in a humidified atmosphere of 5% $CO_2$ and 95% air at 37°C. The cells were harvested by trypsinization. Total RNA was isolated using ISOGEN (NIPPON GENE, Tokyo, Japan).

### 2.2. Vector primer

A plasmid vector pGCAP10 was constructed by substituting the cloning site *Eco*RI–*Afl*II–*Swa*I–*Kpn*I of pGCAP1[30] with *Swa*I–*Eco*RI–*Fse*I–*Eco*RV–*Kpn*I. The nucleotide sequence of pGCAP10 is available from GenBank/EMBL/DDBJ under accession no. AB371573. The plasmid pGCAP10 was digested with *Kpn*I and tailed with ~60 nucleotides of dT using terminal deoxynucleotidyl transferase (Takara Bio, Ohtsu, Shiga, Japan) according to the Okayama and Berg method.[31] After digestion with *Eco*RV, the dT-tailed plasmid vector was purified on agarose gel and used as a vector primer. A cDNA insert can be cut out from the vector with two 8-nucleotide restriction enzymes, *Swa*I and *Not*I.

### 2.3. cDNA synthesis with the V-capping method

Two libraries, Lib-1 (ARe) and Lib-2 (ARi), were constructed using the V-capping method.[30] Lib-1 has already been described in the previous paper.[30] Lib-2 was different from Lib-1 in terms of the source of total RNA, a pGCAP10-derived vector primer, and a modified protocol. The experimental conditions were the same as described in the previous paper. A mixture of 5 μg of total RNA and 0.15 μg of pGCAP10-derived vector primer was incubated at 65°C for 5 min. The first-strand cDNA was synthesized using SuperScript III$^{TM}$ reverse transcriptase (Invitrogen). The reaction mixture was incubated at 45°C for 3 h. After phenol/chloroform extraction followed by ethanol precipitation, the pellet was dissolved in water. The next step in the original protocol is self-ligation with T4 RNA ligase (Takara Bio). The present protocol includes an *Eco*RI digestion step before self-ligation. The *Eco*RI digestion was performed in 200 μL of a reaction mixture containing 50 mM Tris–HCl (pH 7.5), 10 mM $MgCl_2$, 1 mM dithiothreitol, 100 mM NaCl, and 0.2 U/μL of *Eco*RI (Takara Bio). The reaction mixture was incubated at 37°C for 1.5 h. After phenol/chloroform extraction followed by ethanol precipitation, the pellet was dissolved in water. Self-ligation and second-strand synthesis were performed in the same way as in the previous paper.[30]

### 2.4. *Construction of cDNA library*

Transformation of *Escherichia coli* cells DH12S was performed using an electroporation method as previously described.[30] Transformants were plated on LB agar without amplification. Colonies grown on the plates were picked manually or using a Flexys Colony Picker (Genomic Solutions, Ann Arbor, MI, USA) and suspended in 96-well or 384-well plates. After incubation and the addition of 50% glycerol, the original plates were stored at −80°C.

### 2.5. *Plasmid isolation and sequencing*

The isolated plasmid DNA or DNA amplified using the illustra TempliPhi$^{TM}$ DNA amplification kit (GE Healthcare, Uppsala, Sweden) was used as a template for sequencing. DNA sequencing from the 5' end of the cDNA insert was carried out with a capillary DNA sequencer (Applied Biosystems Inc., Foster City, CA, USA) using a BigDye$^{TM}$ Terminator Cycle sequencing FS Ready reaction kit. The full sequence of the cDNA insert was determined by a primer walking method.

### 2.6. *BLAST search and annotation*

First, the 5'-end sequences were used to query our custom database for human full-length cDNA clones (Homo-Protein cDNA bank)[4] with a software GENETYXR-PDB (GENETYX Co., Tokyo, Japan). Most of the abundant genes, ribosomal RNAs, and mitochondria-derived sequences were identified by this search. Sequences not matching to entries in our custom database were used to query the NCBI Human Genome database (National Center for Biotechnology Information, Bethesda, MD, USA) with the BLAST algorithm.[32] Each search was carried out manually, and the sequence alignment and map shown on the NCBI's Map Viewer were checked visually by us. Most sequences were mapped to the first exon of a known gene locus. If the query sequence was mapped to the upstream region of a known gene locus in the same direction, the sequence was assigned to that gene. Through the websites linked to the Map Viewer, including Entrez Gene[33] and UniGene,[27] we retrieved information on gene name, gene symbol, gene ID, chromosomal location, and RefSeq[34] accession number. Sequences not mapped to the known gene locus were BLAST-searched against the NCBI database, including non-redundant nucleotide sequences and ESTs. EST sequences not included in Entrez Gene and the determined full sequences of long-sized cDNAs were deposited in GenBank/EMBL/DDBJ under accession numbers AB371430–AB371572 and AB371574–AB371588, respectively.

### 2.7. *Estimation of the total number of genes composing libraries*

The total number of genes constituting the library was estimated according to two approaches used for species richness estimation: non-sampling-based extrapolation and statistical sampling approaches.[35] The former was performed by curve fitting to a gene-accumulation curve using asymptotic models, including negative exponential models and hyperbolic models.[35] The curve fitting was carried out using software KaleidaGraph (Synergy Software, Reading, PA, USA). The latter approach used an abundance-based coverage estimator model ACE-1, a modified ACE for highly heterogeneous communities.[36] The calculation was done using the SPADE (Species Prediction and Diversity Estimation) algorithm.[37]

### 2.8. *Quantitative real-time PCR*

First-strand cDNA was synthesized with oligo$(dT)_{30}$ as a primer from 20 μg of total RNA using SuperScript III$^{TM}$ reverse transcriptase (Invitrogen), and then purified by a Wizard PCR Preps DNA Purification System (Promega, Madison, WI, USA). Real-time PCR was performed using TaqMan Universal Master Mix (Applied Biosystems) on an ABI PRISM 7000 Sequence Detection System (Applied Biosystems) according to the manufacturer's instructions. One microlitter of diluted cDNA, equivalent to 300 ng of the initial total RNA template, was used in each reaction. Probes and primers designed by TaqMan Gene Expression Assays (Applied Biosystems) were used for the assays of ACTB (Hs99999903_m1), CFL1 (Hs00830568_g1), FLNA (Hs99999905_m1), FLNB (Hs00181698_m1), GAPDH (Hs99999905_m1), GUK1 (Hs00176133_m1), MYH9 (Hs00159522_m1),

**Table 1.** Summary of single-pass sequencing analysis of libraries

| Library Name | Lib-1 | Lib-2 | |
|---|---|---|---|
| Clone Name | ARe/ARf | ARi | ARiS |
| Vector | pKA1U5 | pGCAP10 | pGCAP10 |
| *Eco*RI cut | No | Yes | Yes |
| Total | 10 176 | 6528 | 7296 |
| Unreadable | 940 | 340 | 588 |
| Readable | 9236 | 6188 | 6708 |
| Insert-free vector | 305 | 244 | 221 |
| dT tail | 177 | 10 | 9 |
| Mitochondria | 86 | 65 | 69 |
| rRNA | 3 | 2 | 4 |
| cDNA insert | 8665 | 5867 | 6405 |
| Full-length | 8275 | 5586 | 6090 |
| Truncated | 310 | 243 | 271 |
| Poly(A) | 80 | 38 | 44 |

and RAI14 (Hs00210238_m1). The expression level was calculated based on a standard curve prepared for each gene using a plasmid with each cDNA as a template.

## 3. Results

### 3.1. cDNA Library

Two cDNA libraries, Lib-1 and Lib-2, were constructed using the V-capping method from the total RNA isolated from ARPE-19. The construction of Lib-1 and part of its analysis were described in a previous paper.[30] A total of 10 176 clones from Lib-1 were randomly picked, cultured, and stored as a glycerol stock in 96-well plates. The clones were named ARe and ARf. Lib-2 was prepared from a different lot of total RNA using a slightly modified method that included an *Eco*RI digestion step after the first-strand cDNA synthesis. Transformation of the *E. coli* cells by cDNA vectors for Lib-2 was carried out in two batches at different times. The colonies picked from each batch were incubated and stored in 96-well plates (ARi, 6528 clones) and 384-well plates (ARiS, 76 800 clones), respectively. In this study, all of the ARe/ARf and ARi clones and part of ARiS (7296 clones) were analyzed by single-pass sequencing of the 5' end of cDNA.

Table 1 shows the contents of each library classified by single-pass sequencing analysis. More than 90% clones provided the high-quality sequence data necessary for sequence analysis. The unreadable sequence may result from (i) deletion of a sequencing primer site on the vector, (ii) mixing of different clones, or (iii) failure of template DNA preparation. Many cases were attributed to the first reason because they showed no sequencing signals and could not be cut by restriction enzymes adjacent to the upstream of the cDNA insert. Each library contained insert-free vectors (3.3−3.9% in content), which may result from uncut vectors escaping from removal during the vector primer preparation process.

Lib-1 contained clones carrying a dT tail at the 5' end (1.9% in content). Inspection of the downstream sequence of these clones showed that they lacked a poly(A) tail and contained the inversely inserted cDNA whose 5' end was joined to the *Kpn*I-cut end of the vector where the 3'-protruding bases were deleted. These clones may be generated by use of an aberrant vector primer that has a dT tail at the opposite end from the intended one, implying that the dT-tail addition to only one end of the vector plasmid occasionally occurred at the dT-tailing step with terminal deoxynucleotidyl transferase. Although the dT tail added to the opposite end should be removed by *Eco*RV digestion after the dT-tailing reaction, some must remain uncut. These clones starting with the dT tail could be removed by adding an *Eco*RI digestion step before self-ligation, as shown in the modified protocol. In fact, it worked so well that these artifacts drastically decreased to 0.16% in the ARi library and to 0.13% in the ARiS library.

The same mechanism can also explain the addition of a vector-derived sequence, ATCCTG in the case of using pKA1U5 as a vector primer, adjacent to the 5' end of $\sim$5% of cDNA clones isolated from Lib-1. In this case, the opposite end might not have been dT-tailed and in addition might have escaped from *Eco*RV digestion. Although the *Eco*RI digestion step in the modified protocol was expected to remove this kind of additional sequence, 2.8% of clones in Lib-2 still had a residual sequence (CGGCCGGCCGAT) derived from the vector sequence located between the *Eco*RI and *Eco*RV sites because of incomplete digestion with *Eco*RI.

### 3.2. Assessment of full-length cDNA

The 5'-terminal sequences of cDNA inserts were used for the database search. Clones having only poly(A) or the sequence of several bases with poly(A) were classified into a truncated cDNA, because these clones might be derived from degraded mRNA. The sequence similarity was searched against our custom human full-length cDNA database, the NCBI RefSeq database, and the human genome database using the BLAST algorithm. The libraries contained cDNA clones for mitochondria genome-derived transcripts ($\sim$1% of cDNA-carrying clones) and rRNA ($<$0.06%). Except for mitochondrial clones, all sequences were able to map on the human genome.

Most of the full-length cDNA clones identified as a known gene have a transcriptional start site (TSS) near to that of RefSeq. However, some sequences did not match the 5'-terminal sequence of RefSeq, presumably because our cDNA had a longer 5' UTR than RefSeq or was transcribed by the usage of an alternative promoter. These sequences were mapped on the human genome and the location of the sequence was determined. If the query sequence was located near the upstream region of the first exon of RefSeq or could be linked to the RefSeq via other mRNA or EST sequences that partly shared the query sequence, the clone was assigned to the gene for the corresponding RefSeq even though there was no shared sequence between the query and RefSeq.

Consequently, 8275 clones from the ARe/ARf library, 5586 from ARi, and 6090 from ARiS were identified as a full-length cDNA. Most clones (93.6% for Lib-1, 92.3% for Lib-2) had an additional G or NG (N: T, TT, G, etc.) at their 5' end, which is a requisite for full-length cDNA starting from the cap site when using the V-capping method. The previous paper showed that some full-length cDNAs had no additional G.[30] Thus, if the cDNA not having a 5'-end G started at the upstream region of the first exon of the known gene, it was admitted to be a full-length cDNA for the corresponding gene. Although some cDNAs contained a repetitive sequence such as an Alu repeat, they were also precisely mapped on the genome. Most of the truncated cDNAs existed as a short form with a poly(A) tail. Even such short cDNA was assigned to full-length when it was mapped to the region that was not a known gene locus and had an additional G at the 5' end. Consequently, the content of the full-length clone in all clones carrying a cDNA insert [including only a poly(A) insert] was calculated to be 95.5% for ARe/ARf, 95.2% for ARi, and 95.1% for ARiS.

Out of 19 951 full-length cDNA clones, 1123 clones (5.6%) lacked an additional 5'-end G. These 5'-G-free genes were classified into two groups. One group (625 clones) consisted of clones starting with a nucleotide A. Another group (309 clones) was a 5'-terminal oligopyrimidine tract (5'-TOP) gene family that started with a pyrimidine-rich sequence, including predominantly ribosomal proteins. The genes that contained more than three 5'-G-free clones (G −) with the same TSS were listed in Supplementary Table 1. It should be noted that G-added clones (G +) corresponding to each G− clone were obtained except for NDUFB11. The content of G− clones was 0.04–0.31 for 16 kinds of 5'-TOP genes. On the other hand, the G− content for 12 out of 15 A-starting genes was 0.38–1.0, higher than for 5'-TOP genes. Although we could not find any conserved sequence in the 5' end of the G-free A-starting genes, the following finding suggests that the 5'-end sequence affected the addition or elimination of a cap structure. We obtained 82 clones starting with 5'-ACCACGCACG… for MT2A, out of which 44 had an additional G and 38 were G-free as shown in Supplementary Table 1. In addition, we obtained 43 clones starting with the fourth nucleotide A of the previous MT2A clone, i.e. 5'-ACGCACG…. Interestingly, all 43 clones had an additional G, suggesting that the presence of 5'-end three-nucleotide sequence resulted in the production of G-free clones.

### 3.3. *Gene annotation*

Mapping a total of 19 951 full-length cDNA sequences to the genome revealed that they were classified into 4513 kinds of transcriptional units (i.e. genes), of which 4370 (96.8%) were included in Entrez Gene. All genes are listed in the order of GeneID in Supplementary Table 2. The list contains the symbol, GeneID, and name, which were retrieved from Entrez Gene, and in addition the accession number of the RefSeq, mRNA size, chromosomal location, and number of clones obtained from each library. Of 4370 genes, 4271 had an accession number with prefix "NM_" that indicates mRNA. The remaining 99 genes (2.3%) were possible non-coding genes.

### 3.4. *Expression profile*

Fig. 1 shows the frequency distribution of abundant full-length genes obtained from three libraries. The most abundant genes were GAPDH and FTH1 (each 248 clones, 1.2% in content) followed by ACTB, ACTG1, EEF1A1, VIM, RPL41, MT2A, RPL1, CRYAB, TMSB10, RPS3, and RPL10, each of which gave ≥100 clones (0.5% in content). The number of abundant genes with ≥0.05% content (10 clones) was 310 (6.9% of identified genes). The major components were ribosomal proteins (79 kinds). Of the total genes, 2221 (49.2%) were obtained as a non-redundant transcript.

In order to examine the presence of bias on the expression frequency in different libraries, we compared the expression profiles of different libraries. Fig. 2A shows the comparison between frequencies of abundant genes with ≥0.1% content identified from the different pools (ARi and ARiS) of the same library (Lib-2). Although the plots scattered owing to the small sampling number, the correlation between two expression profiles was good (the correlation coefficient = 0.94). As listed in Supplementary Table 2, there were many genes for which several clones were obtained from one library but no clone from another library, suggesting that we should take

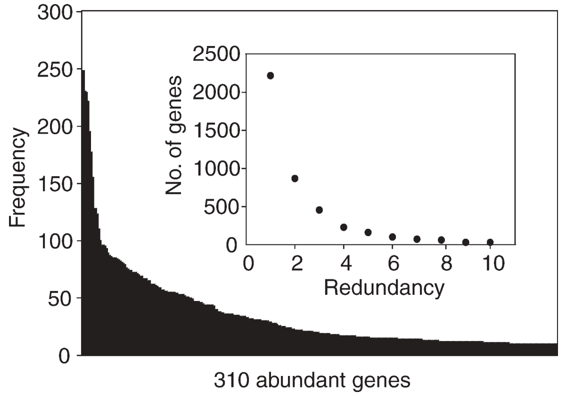

**Figure 1.** Frequencies of 310 kinds of abundant genes with $\geq$0.05% content (10 clones) obtained from three libraries. The inlet shows the number of low-redundant genes with $\leq$0.05% content at each redundancy.

the extent of sampling bias into account when we analyze a small number of samples.

Fig. 2B shows the comparison between frequencies of the different libraries, Lib-1 and Lib-2. Although the order of top 10 genes was different, the correlation was good on the whole (the correlation coefficient = 0.87). These two libraries were constructed from different lots of RNA using slightly different protocols. The variation may result from the difference of RNA lots rather than the difference of protocol. For example, it was reported that the expression level of CRYAB largely varied depending on the conditions of cell culture, such as heat stress[38] or hypertonic stress.[39] GAPDH was most abundant in Lib-1, but fifth in Lib-2. The expression level of GAPDH also varied with calcium,[40] insulin,[41] and oxygen[42] in the culture medium. Thus, the difference of culture conditions may induce a different expression level and result in the library-dependent variation in the expression level of these genes.

### 3.5. *Estimation of the total number of genes composing the library*

The cumulative number of gene occurrences was plotted as a function of the number of analyzed clones as shown in Fig. 3A. The cumulative number asymptotically increased but did not saturate within the analyzed range. The curve fitting was carried out using six asymptotic models, including negative exponential models and hyperbolic models. As a result, the best fitting was obtained by use of a hyperbolic curve, $D_t = St^{\alpha}/(\beta + t^{\alpha})$, where $D_t$ denotes the cumulative number of genes for the accumulated number $t$ of sequenced clones, $S$ is an asymptotic value, and $\alpha$ and $\beta$ are parameters to be estimated from data. The obtained $S$ was 14 348 for Lib-1 and 11 563 for Lib-2. Using the value for Lib-2, the cumulative gene number at 24 000 analyses was calculated to be 4578, which is similar to the observed value, 4513. Even if analyzing 48 000 and 96 000 clones, the cumulative gene number could merely reach 6177 (53.4%) and 7717 (66.7%), respectively.

Ida et al.[43] estimated the number of total gene clusters in mouse retina EST libraries using the abundance-based coverage estimator model ACE. We applied its modified model ACE-1 to estimate the total number of genes composed of Lib-1 and Lib-2. When the number of genes containing 10 or fewer clones was used, the total number was estimated to be 8019 for Lib-1 and 8469 for Lib-2, as shown in Fig. 3B. These values were smaller than those estimated using the cumulative curves. Consequently, the total number of genes comprising the present library was estimated to be between

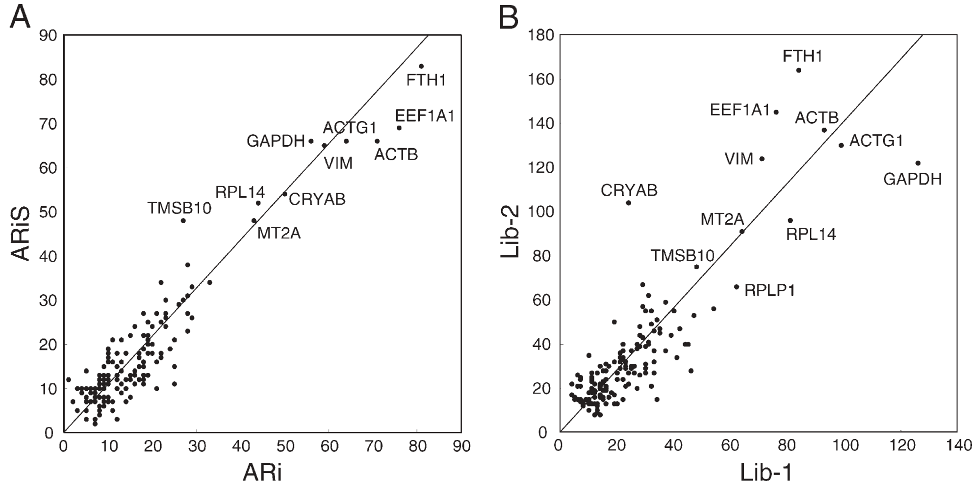

**Figure 2.** Comparison of frequencies of abundant genes. (**A**) Between different pools from the same library, Lib-2. (**B**) Between different libraries, Lib-1 and Lib-2. The genes with $\geq$0.1% content were plotted. The top 10 genes are designated by gene symbols. The line represents the case without bias.

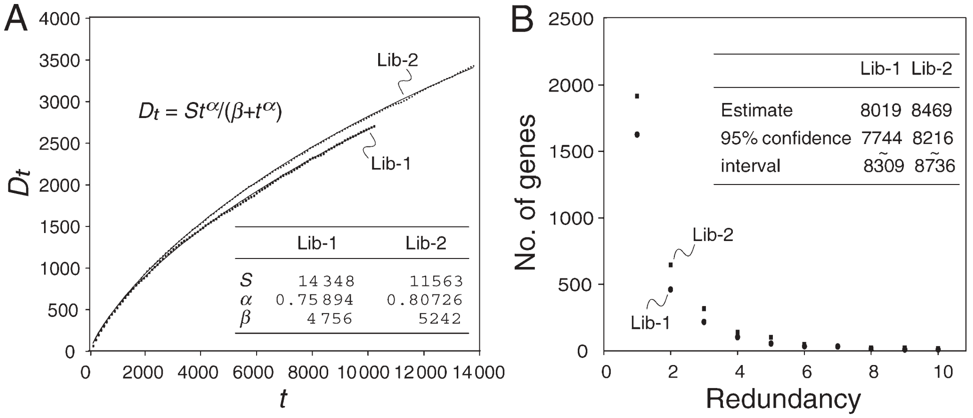

**Figure 3.** Estimation of the total number of genes composed of each library. (**A**) The cumulative number of gene occurrences, $D_t$, was plotted for $t$, the accumulated number of sequenced clones. In this study, novel gene occurrence per 96 clones was counted. The best fitting curves were obtained using the hyperbolic equation described. The asymptotic value, $S$, represents the estimate of the total number of genes. (**B**) Abundance-based coverage estimator model ACE-1. The number of genes containing 10 or fewer clones was used to calculate the gene richness.

8000 and 14 000, and thus many genes would remain unidentified.

### 3.6. *TSSs of abundant transcripts*

The 5'-terminal sequence analysis of full-length cDNA clones has the advantage of massive production of TSSs.[22] In particular, the full-length content of the vector-capped cDNA library is so high that the distribution of TSSs of transcriptional initiation variants can be determined for abundant transcripts obtained from each library. The full-length content of the top 11 abundant genes was 94.0–100% for each gene cluster. In addition, truncated cDNAs were easily distinguishable, because most of them started at the position in the last exon and had no additional G at the 5' end.

The distributions of TSSs for 11 genes, which were obtained from two libraries and DBTSS,[44] were compared as shown in Supplementary Table 3. The most frequent TSS for each gene, except for RPL1 and CRYAB, was identical among three distributions. EEF1A1 showed almost only one TSS, but generally multiple preferential TSSs were observed in other genes. Fig. 4 shows examples of comparison among distributions of TSSs for GAPDH, ACTG1, and CRYAB. The preferential TSSs formed a cluster in the region of ~10 nt, presumably owing to the presence of the

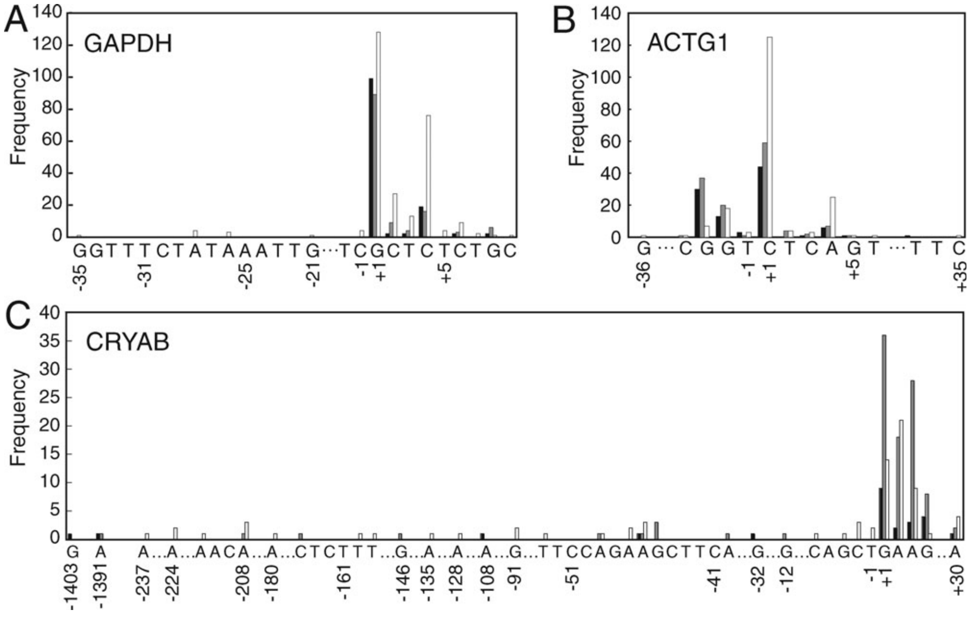

**Figure 4.** Comparison among distributions of transcriptional start sites. Black bar, Lib-1; Gray bar, Lib-2; White bar, DBTSS. Position 1 is defined as a major TSS.

TATA box upstream of these TSSs. CRYAB also showed widely scattered rare TSSs at the upstream region. In every gene, the distribution patterns of TSSs obtained from two libraries and DBTSS were similar. It should be noted that the data of DBTSS was obtained from various tissues, suggesting that the pattern of TSS distribution shows no tissue specificity in the case of abundant housekeeping genes. DBTSS represented TSSs generated by alternative promoter usage.[44] The TSS observed in the first intron of GAPDH and VIM, and two TSSs identified at the position ~1400 nt upstream of the main TSS cluster of CRYAB, may result from transcripts given by alternative promoter usage.

### 3.7. *Long-sized transcripts*

On the basis of the mRNA size of RefSeq for genes with GeneID in Supplementary Table 2, the gene-based and clone-based average lengths of cDNA inserts of our full-length gene collection were calculated to be 2.46 kb (4378 genes) and 1.68 kb (19 758 clones), respectively. The previous paper showed that the V-capping method could synthesize a long-sized full-length cDNA.[30] The present libraries also contained many long-sized full-length clones with a cDNA insert of ≥7 kb, as listed in Table 2. The size of the mRNA described in the RefSeq data was often different from the insert size of our cDNA clone because of size variants generated by alternative splicing or alternative polyadenylation. Therefore, cDNA clones that correspond to RefSeq derived from >6 kb mRNA were selected, and then the real size of the insert was examined by restriction enzyme digestion followed by agarose gel electrophoresis. Table 2 shows the determined size together with the mRNA size of RefSeq. Some clones were fully sequenced and their precise length was listed with an accession number.

The longest cDNA of 11 199 bp encoded golgin B1 (GOLGB1), which is a Golgi integral membrane protein originally named 'giantin' owing to its huge size of ~400 kDa.[45] When compared with RefSeq, the coding region of this cDNA had two insertions of 15 bp each that resulted in the insertion of a total of 10 amino acid residues, implying an alternative-splicing variant. A total of nine single-nucleotide variations were observed, one of which was the insertion of one nucleotide A to an A stretch at position 2958–2965 in the coding region, causing a frame-shift. This insertion may result from misreading of reverse transcriptase during first-strand synthesis, because sequencing of the GOLGB1 locus of the ARPE-19 genome showed the absence of such an insertion.

Redundant cDNA clones in the long-sized cDNAs of ≥8 kb were filamin A (FLNA) and filamin B (FLNB). As a result of full sequencing of a total of eight FLNA clones, three splicing variants were identified, whose exon–intron structures are depicted in Fig. 5A. V1 was the longest variant and three clones showed identical TSS. The V2 clones started at the 6th nucleotide downstream of the TSS of V1, and lacked the 29th exon, which caused deletion of eight amino acid residues. V3 had the same TSS as V2, and lacked the region from the middle of exon 36 to the middle of exon 41 of V1, which caused deletion of 305 amino acid residues. RefSeq seems to correspond to V2, but it is doubtful that it actually dose because RefSeq is constructed using multiple sequences reported by different researchers. Surprisingly, all four FLNB clones showed different splicing patterns, as shown in Fig. 5B.

### 3.8. *Correlation between the number of isolated full-length clones and mRNA content*

The high content of the very long-sized cDNA clones leads us to expect that the present library is unbiased by mRNA size. In order to assess the extent of bias, the mRNA contents for eight genes with different mRNA sizes (1.1–9.5 k) were measured by real-time PCR. As shown in Fig. 6, the total number of full-length clones obtained from the present libraries had good correlation with the content of each mRNA. GUK1 (1.1 k), RAI14 (5.0 k), and FLNB (9.5 k) showed a similar content and were represented by full-length cDNA clones with the number of the same order of magnitude independent of their mRNA sizes. Thus, the bias is expected to be low up to 9.5 kb, explaining the fact that a small-sized library composed of only 20 000 clones contained 48 long-sized full-length clones with a cDNA insert of ≥7 kb (Table 2).

### 3.9. *Unannotated transcripts*

Of the 4513 full-length transcripts identified, 143 (3.2%) have not been included in Entrez Gene, but 79 out of them hit rare ESTs registered in the UniGene EST database. Although 19 clones did not hit any EST sequence, it did not necessarily mean they were novel genes because our clone may result from a partial sequence not overlapping with known ESTs. The unannotated transcripts can be classified into two groups. One group included an intergenic transcript that was mapped to the gene-unoccupied space between known gene loci. Many of them were transcribed from just upstream of TSS of the known gene toward the opposite direction, presumably owing to the presence of a bi-directional promoter.[46] The second group (36 genes) was composed of transcripts whose location overlapped the exon of the known gene, generating an antisense transcript

**Table 2.** Long-sized full-length cDNA clones with >7 kb insert

| | HP ID | Symbol | Name | RefSeq | mRNA (bp) | Protein (aa) | Clone ID | Ac. No. | cDNA (bp) | Protein (aa) |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | HP08164 | GOLGB1 | Golgi autoantigen, golgin subfamily b, macrogolgin (with transmembrane signal) 1 | NM_004487.3 | 11 185 | 3259 | ARiS161G17 | AB371588 | 11 198 | 3269[a] |
| 2 | HP07459 | N4BP2 | Nedd4 binding protein 2 | NM_018177.2 | 6616 | 1770 | ARiS023B20 | AB371584 | 9736 | 1690[b] |
| 3 | HP08032 | ACACA | Acetyl-Coenzyme A carboxylase alpha | NM_198836.1 | 9585 | 2346 | ARiS088K16 | AB371587 | 9534 | 2346 |
| 4 | HP04958 | FLNB | Filamin B, beta (actin binding protein 278) | NM_001457.2 | 9463 | 2602 | ARe23D04 | AB191258 | 9405 | 2591 |
| 5 | HP04958 | FLNB | | | | | ARe77D06 | AB371580 | 8059 | 2633 |
| 6 | HP04958 | FLNB | | | | | ARe89D09 | AB371581 | 9366 | 2578 |
| 7 | HP04958 | FLNB | | | | | ARi12F08 | AB371582 | 7973 | 2602 |
| 8 | HP07616 | FLNC | Filamin C, gamma (actin binding protein 280) | NM_001458.3 | 9146 | 2725 | ARi57A02 | AB371585 | 9156 | 2725 |
| 9 | HP07744 | SPTBN1 | Spectrin, beta, non-erythrocytic 1 | NM_003128.2 | 10 238 | 2364 | ARiS088A21 | AB371586 | 8443 | 2364 |
| 10 | HP00079 | FLNA | Filamin A, alpha (actin binding protein 280) | NM_001456.2 | 8278 | 2639 | ARe06F05 | AB191259 | 8212 | 2612 |
| 11 | HP00079 | FLNA | | | | | ARe27E03 | AB191260 | 8241 | 2620 |
| 12 | HP00079 | FLNA | | | | | ARi13C12 | AB371574 | 8242 | 2620 |
| 13 | HP00079 | FLNA | | | | | ARi37B09 | AB371575 | 8212 | 2612 |
| 14 | HP00079 | FLNA | | | | | ARi47G07 | AB371576 | 8243 | 2620 |
| 15 | HP00079 | FLNA | | | | | ARi50A09 | AB371577 | 7321 | 2315 |
| 16 | HP00079 | FLNA | | | | | ARi66B08 | AB371578 | 8212 | 2612 |
| 17 | HP00079 | FLNA | | | | | ARiS088J13 | AB371579 | 8214 | 2612 |
| 18 | HP06504 | COL5A1 | Collagen, type V, alpha 1 | NM_000093.3 | 8439 | 1838 | ARe79B07 | AB371586 | 8139 | 1838 |
| 19 | HP07532 | GCN1L1 | GCN1 general control of amino-acid synthesis 1-like 1 (yeast) | NM_006836.1 | 8699 | 2671 | ARi43H04 | — | 8.0 k | |
| 20 | HP06485 | PTPRF | Protein tyrosine phosphatase, receptor type, F | NM_002840.2 | 7718 | 1897 | ARe76H09 | — | 8.0 k | |
| 21 | HP05449 | SPTAN1 | Spectrin, alpha, non-erythrocytic 1 (alpha-fodrin) | NM_003127.1 | 7787 | 2472 | ARe45C02 | AB191262 | 7791 | 2452 |
| 22 | HP00124 | FN1 | Fibronectin 1 | NM_212476.1 | 8272 | 2296 | ARe05G09 | AB191261 | 7753 | 2265 |
| 23 | HP00124 | FN1 | | | | | ARi53A03 | — | 7.7 k | |
| 24 | HP00124 | FN1 | | | | | ARiS087I09 | — | 7.9 k | |
| 25 | HP06896 | PCM1 | Pericentriolar material 1 | NM_006197.3 | 8788 | 2024 | ARi07G10 | — | 7.5 k | |
| 26 | HP04727 | GLIS3 | GLIS family zinc finger 3 | NM_001042413.1 | 7656 | 930 | ARe06E05 | — | 7.5 k | |
| 27 | HP04890 | BAT2 | HLA-B associated transcript 2 | NM_080686.1 | 6750 | 2157 | ARe17F01 | — | 7.5 k | |
| 28 | HP04890 | BAT2 | | | | | ARiS122P21 | — | 7.0 k | |
| 29 | HP04715 | ABCA7 | ATP-binding cassette, sub-family A (ABC1), member 7 | NM_019112.2 | 6704 | 2146 | ARe05B12 | — | 7.5 k | |
| 30 | HP07625 | WWC2 | WW, C2 and coiled-coil domain containing 2 | NM_024949.4 | 6492 | 987 | ARi58B08 | — | 7.5 k | |
| 31 | HP04667 | MYH9 | myosin, heavy polypeptide 9, non-muscle | NM_002473.3 | 7274 | 1960 | ARe01B05 | AB191263 | 7436 | 1960 |

*Continued*

**Table 2.** Continued

| | HP ID | Symbol | Name | RefSeq | mRNA (bp) | Protein (aa) | Clone ID | Ac. No. | cDNA (bp) | Protein (aa) |
|---|---|---|---|---|---|---|---|---|---|---|
| 32 | HP04667 | MYH9 | | | | | ARe12G09 | AB191263 | 7450 | 1960 |
| 33 | HP04667 | MYH9 | | | | | ARe58E01 | — | 7.5 k | |
| 34 | HP04667 | MYH9 | | | | | ARi26G12 | — | 7.5 k | |
| 35 | HP04667 | MYH9 | | | | | ARi63B09 | — | 7.5 k | |
| 36 | HP04667 | MYH9 | | | | | ARi63H03 | — | 7.0 k | |
| 37 | HP04667 | MYH9 | | | | | ARiS045L21 | — | 7.5 k | |
| 38 | HP04680 | AGRN | Agrin | NM_198576.2 | 7319 | 2045 | ARe02B04 | AB191264 | 7319 | 2045 |
| 39 | HP04680 | AGRN | | | | | ARiS085N10 | — | 7.4 k | |
| 40 | HP07482 | CDC42BPB | CDC42 binding protein kinase beta (DMPK-like) | NM_006035.2 | 6782 | 1711 | ARiS023M02 | — | 7.2 k | |
| 41 | HP07424 | MAP9 | microtubule-associated protein 9 | NM_001039580.1 | 7333 | 647 | ARiS109F14 | — | 7.1 k | |
| 42 | HP07506 | ROD1 | ROD1 regulator of differentiation 1 (S. pombe) | NM_005156.4 | 7230 | 552 | ARi40D10 | — | 7.0 k | |
| 43 | HP07574 | TRAM2 | Translocation associated membrane protein 2 | NM_012288.3 | 7065 | 370 | ARi50G08 | — | 7.0 k | |
| 44 | HP06927 | COL4A1 | Collagen, type IV, alpha 1 | NM_001845.4 | 6549 | 1669 | ARe94G08 | — | 7.0 k | [c] |
| 45 | HP06927 | COL4A1 | | | | | ARi60G12 | — | 8.2 k | |
| 46 | HP07685 | ARHGAP23 | Rho GTPase activating protein 23 | XM_290799.7 | 6475 | 1684 | ARi70H06 | — | 7.0 k | |
| 47 | HP02917 | PLXNB2 | Plexin B2 | XM_371474.4 | 6434 | 1901 | ARe49A01 | — | 7.0 k | [c] |
| 48 | HP06882 | GPAM | Glycerol-3-phosphate acyltransferase, mitochondrial | NM_020918.3 | 6386 | 828 | ARi06G08 | — | 7.0 k | |

[a] The frame shift owing to the insertion of one nucleotide A in the GOLGB1 clone was corrected.

[b] RefSeq terminates with a short A-stretch in a 3’-untranslated region.

[c] In addition, 8 and 5 clones with >6 kb were obtained for COL4A1 and PLXNB2, respectively.

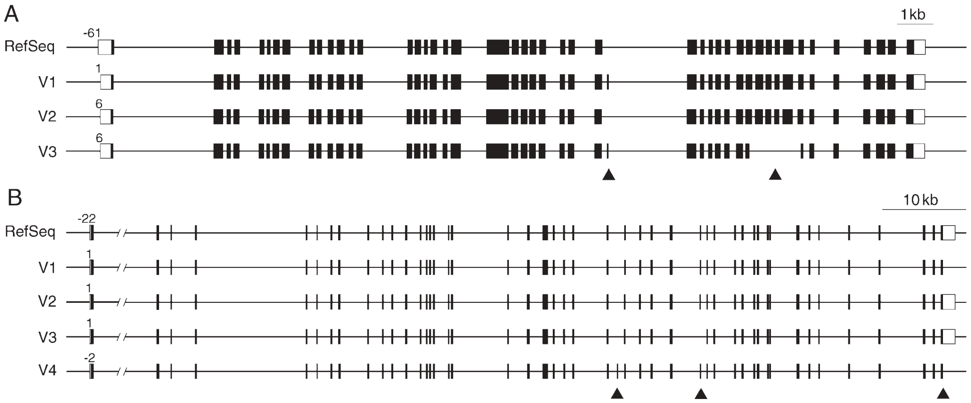

**Figure 5.** The exon–intron structure of splicing variants for FLNA (**A**) and FLNB (**B**). Arrow heads represent exon inclusion or deletion.

against the known exon as listed in Supplementary Table 4. The antisense transcript against the first exon of the known gene is especially intriguing because of possible involvement in transcriptional regulation. The corresponding ESTs often existed in the UniGene EST database, but they were assigned to the sense strand of the known gene.

## 4. Discussion

In-depth analysis of the full-length cDNA libraries constructed using the V-capping method has revealed that these libraries contained full-length cDNA clones with a wide range of sizes up to 11 kb, suggesting that they meet the requirement for size-unbiased libraries. This success may be attributed to the following reasons. (i) Total RNA was used as a template without purifying poly(A) RNA. (ii) The protocol of cDNA synthesis did not include an intact mRNA selection process such as modification of the cap structure.[4–6] These purification and modification processes might cause mRNA degradation. (iii) A small amount of mRNA was used as a template for reverse transcription. When a large amount of mRNA is used, the synthesis of short abundant cDNAs may consume reverse transcriptase and nucleotide substrates, and rare or long-sized cDNA synthesis may be suppressed. (iv) Size fractionation of cDNAs was not carried out. (v) Restriction enzyme digestion of cDNA was not included. (vi) PCR amplification was not used. (vii) Insertion of cDNA into a small-sized vector of 3.4 kb was done by self-ligation. One conceivable size bias may result from incomplete extension of the first-strand cDNA synthesized from a long-sized mRNA template. The presence of GOLB1 cDNA with an 11.2 kb insert and multiple FLNA cDNAs with an 8.2 kb insert suggested that the termination of reverse transcriptase was not problematic under the present conditions.

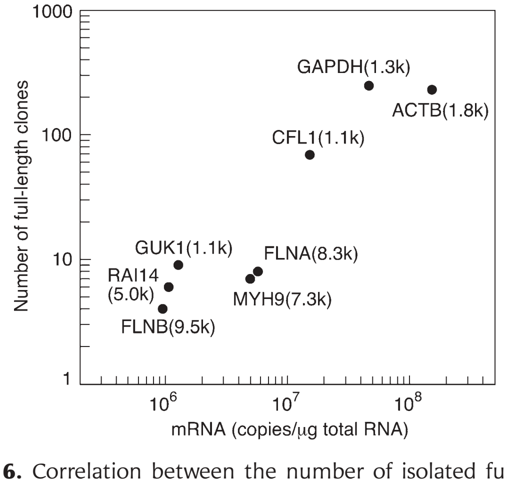

**Figure 6.** Correlation between the number of isolated full-length clones and the mRNA content. The size of mRNA was described in parentheses following each gene symbol.

The present analysis also revealed that the full-length content was unexpectedly high despite lacking a full-length selection process. The full-length content of abundant transcripts was 94.0–100%. The most probable explanation for this high content is that most mRNA molecules are intact in the cell. However, the degradation of mRNA, especially long-sized mRNA, seemed to occur during the RNA isolation process. The overall full-length content was 95%, and analysis of the remaining 5% incomplete cDNAs showed that ∼10% of them were derived from the degradation product of the long-sized mRNA of >6 kb (data not shown). Although a size bias due to RNA degradation during the course of its isolation is inevitable, this method can faithfully reflect the composition of a given purified RNA sample.

The high full-length content and size-unbiased feature of the library enabled us to determine the fine distribution of TSS of genes expressed in a single type of cell. Recently, methods such as CAGE[12] or 5'SAGE[13] have been developed to determine the distribution of TSS. For these analyses, a vector-capped library prepared using a modified vector primer for

CAGE or 5'SAGE analysis can serve as a starting material. The additional G at the 5' end could ensure the intactness of TSS and provide more precise results.

The vector-capped cDNA library also enabled us to analyze the presence of a cap structure at the 5' end of mRNA. The large-scale TSS analysis revealed that a small part of full-length clones lacked an additional 5'-end G, indicating the absence of the cap structure of the mRNA. The common feature of the 5'-end sequence of these clones was a pyrimidine-rich sequence or A-starting sequence. Further in-depth analysis of G-free clones suggested that the 5'-end sequence affected the addition or elimination of the cap structure. This finding may partially be explained by the recent results that human decapping enzyme Dcp2 preferentially binds to a subset of mRNAs and identifies sequences at the 5' terminus of the mRNA as a specific substrate.[47] Further investigation is required to elucidate the mechanism of the cap-free mRNA generation observed in the present study and its biological meaning.

We identified 4513 kinds of genes out of 19 951 full-length clones isolated from unamplified ARPE-19 cDNA libraries. The total number of genes consisting of the present library was estimated to be 8000–14 000, which is consistent with ~10 000 estimated for the transcriptome of RPE using UniGene clusters, SAGE tags and ESTs by Swaroop and Zack.[48] The expression profile analyses using a microarray with 12 600 gene probes identified 5634 ± 65 genes for ARPE-19 and 5580 ± 84 genes for human RPE.[49] These results suggest the low expression level of unidentified genes that may include non-coding and antisense genes. In order to obtain the remaining rare genes, further large-scale analyses in combination with subtraction/normalization are required.

Recent investigations have shown that an unexpectedly large amount of alternative splicing variants exist[23] and that some variants were expressed in a cell-specific manner.[50] In order to analyze a transcriptional network in a single type of cell, it is necessary to measure the expression levels of these variants. When alternative splicing occurs at the multiple sites, it is difficult to quantify each variant by the RT–PCR method, which measures only the expression level of a limited region of mRNA. The most precise method for determining the expression level of a splicing variant is to count the number of full-length cDNA clones for each variant. This is a particular requirement for long-sized full-length transcripts. The present study demonstrated that multiple splicing variants for long-sized genes such as FLNA and FLNB were expressed in a single type of cell and that the V-capping method was able to provide unique genuine full-length cDNA to each variant. The physiological role of each variant remains to be solved.

One of the characteristics of the V-capping method is the use of a vector primer with a relatively long dT tail, which has the following merits compared with the use of an oligo dT primer by conventional methods: (i) the unidirectional insertion of the cDNA is guaranteed and makes it easy to identify an antisense transcript against the known gene; (ii) it does not prime a short A-stretch in the mRNA, not generating 3'-truncated cDNAs that were occasionally observed in the RefSeq database (e.g. N4BP2 in Table 2). However, we should keep in mind the possibility that the vector primer could fail to capture mRNA with a short poly(A) tail.

The remaining challenge to analyzing the vector-capped cDNA library is to develop a method for high-throughput single-pass sequencing. When we intend to analyze >100 000 clones, the present single-pass sequencing strategy is not realistic because of high cost and because of the amount of work involved. However, considering that the analysis of the present high-quality full-length cDNA library is expected to enable discovery of rare or long-sized full-length cDNA clones and fine expression profiling of various variants, it makes sense to analyze more than a million clones corresponding to total clones composed of one library. Thus, developing a novel method for massive single-pass sequencing is desired, for example, by applying recently developed technologies.[51] Furthermore, in order to achieve a comprehensive collection of full-length transcripts, the removal of abundant clones should be carried out by subtraction.

In conclusion, a full-length cDNA library constructed with the V-capping method has been shown to meet requirements for a size-unbiased library, and thus is expected to be suitable for the comprehensive collection of full-length transcripts and their fine expression profiling. Recently, Miura et al.[52] analyzed full-length cDNA libraries constructed from budding yeast using the V-capping method and identified novel full-length transcripts, including splicing variants and antisense transcripts.[52] Even for well-characterized organisms such as yeast, in-depth analysis of the vector-capped cDNA library has been shown to still bring discovery of novel full-length transcripts. To analyze the transcriptome of not only uncharacterized organisms but also well-characterized ones such as human and mouse, the V-capping method will become the first choice for constructing a full-length cDNA library.

**Acknowledgement:** We thank Takumi Nonami for technical assistance for full sequencing of long-sized clones.

**Supplementary Data:** Supplementary data are available online at www.dnaresearch.oxfordjournals.org.

## Funding

This work was partly supported by a grant for the Genome Network Project from the Ministry of Education, Culture, Sports, Science and Technology. Funding for open access charge: Open access charge will be paid by Seishi Kato.

## 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. International Human Genome Sequencing Consortium. 2004, Finishing the euchromatic sequence of the human genome, *Nature*, **431**, 931–945.
4. Kato, S., Sekine, S., Oh, S.-W., et al. 1994, Construction of a human full-length cDNA bank, *Gene*, **150**, 243–250.
5. Carninci, P., Kvam, C., Kitamura, A., et al. 1996, High-efficiency full-length cDNA cloning by biotinylated CAP trapper, *Genomics*, **37**, 327–336.
6. 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.
7. Zhu, Y. Y., Machleder, E. M., Chenchik, A., Li, R. and Siebert, P. D. 2001, Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction, *Biotechniques*, **30**, 892–897.
8. Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. 1995, Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science*, **270**, 467–470.
9. Lockhart, D. J., Dong, H., Byrne, M. C., et al. 1996, Expression monitoring by hybridization to high-density oligonucleotide arrays, *Nat. Biotechnol.*, **14**, 1675–1680.
10. Okubo, K., Hori, N., Matoba, R., et al. 1992, Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression, *Nat. Genet.*, **2**, 173–179.
11. Velculescu, V. E., Zhang, L., Vogelstein, B. and Kinzler, K. W. 1995, Serial analysis of gene expression, *Science*, **270**, 484–487.
12. Shiraki, T., Kondo, S., Katayama, S., et al. 2003, Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage, *Proc. Natl. Acad. Sci. USA*, **100**, 15776–15781.
13. Hashimoto, S., Suzuki, Y., Kasai, Y., et al. 2004, 5'-end SAGE for the analysis of transcriptional start sites, *Nat. Biotechnol.*, **22**, 1146–1149.
14. Harbers, M. and Carninci, P. 2005, Tag-based approaches for transcriptome research and genome annotation, *Nat. Methods*, **2**, 495–502.
15. Ota, T., Suzuki, Y., Nishikawa, T., et al. 2004, Complete sequencing and characterization of 21,243 full-length human cDNAs, *Nat. Genet.*, **36**, 40–45.
16. Imanishi, T., Itoh, T., Suzuki, Y., et al. 2004, Integrative annotation of 21,037 human genes validated by full-length cDNA clones, *PLoS Biol.*, **2**, e162.
17. Gerhard, D. S., Wagner, L., Feingold, E. A., et al. 2004, The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC), *Genome Res.*, **14**, 2121–2127.
18. Kawamoto, S., Yoshii, J., Mizuno, K., et al. 2000, BodyMap: a collection of 3' ESTs for analysis of human gene expression information, *Genome Res.*, **10**, 1817–1827.
19. Edgar, R., Domrachev, M. and Lash, A. E. 2002, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, *Nucleic Acids Res.*, **30**, 207–210.
20. Brazma, A., Parkinson, H., Sarkans, U., et al. 2003, ArrayExpress—a public repository for microarray gene expression data at the EBI, *Nucleic Acids Res.*, **31**, 68–71.
21. Kimura, K., Wakamatsu, A., Suzuki, Y., et al. 2006, Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes, *Genome Res.*, **16**, 55–65.
22. 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.
23. Modrek, B., Resch, A., Grasso, C. and Lee, C. 2001, Genome-wide detection of alternative splicing in expressed sequences of human genes, *Nucleic Acids Res.*, **29**, 2850–2859.
24. Beaudoing, 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.
25. Yelin, R., Dahary, D., Sorek, R., et al. 2003, Widespread occurrence of antisense transcription in the human genome, *Nat. Biotechnol.*, **21**, 379–386.
26. Chen, J., Sun, M., Kent, W. J., et al. 2004, Over 20% of human transcripts might form sense-antisense pairs, *Nucleic Acids Res.*, **32**, 4812–4820.
27. Schuler, G. D. 1997, Pieces of the puzzle: expressed sequence tags and the catalog of human genes, *J. Mol. Med.*, **75**, 694–698.
28. Kampa, D., Cheng, J., Kapranov, P., et al. 2004, Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22, *Genome Res.*, **14**, 331–342.
29. Bertone, P., Stolc, V., Royce, T. E., et al. 2004, Global identification of human transcribed sequences with genome tiling arrays, *Science*, **306**, 2242–2246.
30. Kato, S., Ohtoko, K., Ohtake, H. and Kimura, T. 2005, Vector-capping: a simple method for preparing a high-quality full-length cDNA library, *DNA Res.*, **12**, 53–62.
31. Okayama, H. and Berg, P. 1982, High-efficiency cloning of full-length cDNA, *Mol. Cell. Biol.*, **2**, 161–170.

32. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. 1990, Basic local alignment search tool, *J. Mol. Biol.*, **215**, 403–410.
33. Schuler, G. D., Epstein, J. A., Ohkawa, H. and Kans, J. A. 1996, Entrez: molecular biology database and retrieval system, *Methods Enzymol.*, **266**, 141–162.
34. Pruitt, K. D., Tatusova, T. and Maglott, D. R. 2007, NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, *Nucleic Acids Res.*, **35**, D61–D65.
35. Chao, A. 2005, Species richness estimation. In: Balakrishnan, N., Read, C. B. and Vidakovic, B., eds. Species Estimation and Applications. Encyclopedia of Statistical Sciences, **Vol. 12**. Wiley: New York, pp. 7907–7916.
36. Chao, A. and Lee, S.-M. 1992, Estimating the number of classes via sample coverage, *J. Am. Stat. Assoc.*, **87**, 210–217.
37. Shen, T.-J., Chao, A. and Lina, C.-F. 2003, Predicting the number of new species in further taxonomic sampling, *Ecology*, **84**, 798–804.
38. Klemenz, R., Frohli, E., Steiger, R. H., Schafer, R. and Aoyama, A. 1991, Alpha B-crystallin is a small heat shock protein, *Proc. Natl. Acad. Sci. USA* **88**, 3652–3656.
39. Dasgupta, S., Hohman, T. C. and Carper, D. 1992, Hypertonic stress induces alpha B-crystallin expression, *Exp. Eye Res.*, **54**, 461–470.
40. Chao, C. C., Yam, W. C. and Lin-Chao, S. 1990, Coordinated induction of two unrelated glucose-regulated protein genes by a calcium ionophore: human BiP/GRP78 and GAPDH, *Biochem. Biophys. Res. Commun.*, **171**, 431–438.
41. Nasrin, N., Ercolani, L., Denaro, M., Kong, X. F., Kang, I. and Alexander, M. 1990, An insulin response element in the glyceraldehyde-3-phosphate dehydrogenase gene binds a nuclear protein induced by insulin in cultured cells and by nutritional manipulations in vivo, *Proc. Natl. Acad. Sci. USA*, **87**, 5273–5277.
42. Graven, K. K., Troxler, R. F., Kornfeld, H., Panchenko, M. V. and Farber, H. W. 1994, Regulation of endothelial cell glyceraldehyde-3-phosphate dehydrogenase expression by hypoxia, *J. Biol. Chem.*, **269**, 24446–24453.
43. Ida, H., Boylan, S. A., Weigel, A. L., et al. 2004, EST analysis of mouse retina and RPE/choroid cDNA libraries, *Mol. Vis.*, **10**, 439–444.
44. Yamashita, R., Suzuki, Y., Wakaguri, H., Tsuritani, K., Nakai, K. and Sugano, S. 2006, DBTSS: DataBase of human transcription start sites, progress report 2006. *Nucleic Acids Res.*, **34**, D86–D89.
45. Linstedt, A. D. and Hauri, H. P. 1993, Giantin, a novel conserved Golgi membrane protein containing a cytoplasmic domain of at least 350 kDa, *Mol. Biol. Cell*, **4**, 679–693.
46. Trinklein, N. D., Aldred, S. F., Hartman, S. J., Schroeder, D. I., Otillar, R. P. and Myers, R. M. 2004, An abundance of bidirectional promoters in the human genome, *Genome Res.*, **14**, 62–66.
47. Li, Y., Song, M. G. and Kiledjian, M. 2008, Transcript-specific decapping and regulated stability by the human Dcp2 decapping protein, *Mol. Cell. Biol.*, **28**, 939–948.
48. Swaroop, A. and Zack, D. J. 2002, Transcriptome analysis of the retina, *Genome Biol.*, **3**, reviews1022.1–1022.4.
49. Cai, H. and Del Priore, L. V. 2006, Gene expression profile of cultured adult compared to immortalized human RPE, *Mol. Vis.*, **12**, 1–14.
50. Xu, Q., Modrek, B. and Lee, C. Genome-wide detection of tissue-specific alternative splicing in the human transcriptome, *Nucleic Acids Res.*, **30**, 3754–3766.
51. Metzker, M. L. 2005, Emerging technologies in DNA sequencing, *Genome Res.*, **15**, 1767–1776.
52. Miura, F., Kawaguchi, N., Sese, J., et al. 2006, A large-scale full-length cDNA analysis to explore the budding yeast transcriptome, *Proc. Natl. Acad. Sci. USA*, **103**, 17846–17851.