2. Synthesis of Full-Length cDNA Using the Vector-Capping Method

We started our investigation 20 years ago by focusing on how to obtain whole human proteins. The strategy we took was to collect all proteins as a form of cDNA. At that time, the Human Genome Project had been launched and one of the projects was to analyze ESTs. However, ESTs composed of cDNA fragments were not suitable to obtain proteins. To achieve our purpose, we needed to obtain a full-length cDNA that contains an intact open reading frame (ORF) to produce the encoded protein. Thus, we developed a novel method to synthesize full-length cDNA based on replacing the cap structure of mRNA by a DNA-RNA chimeric oligonucleotide (Kato et al., 1994). This method enabled us to effectively synthesize full-length cDNAs, but it had drawbacks: it required a lot of mRNA and many reaction steps.

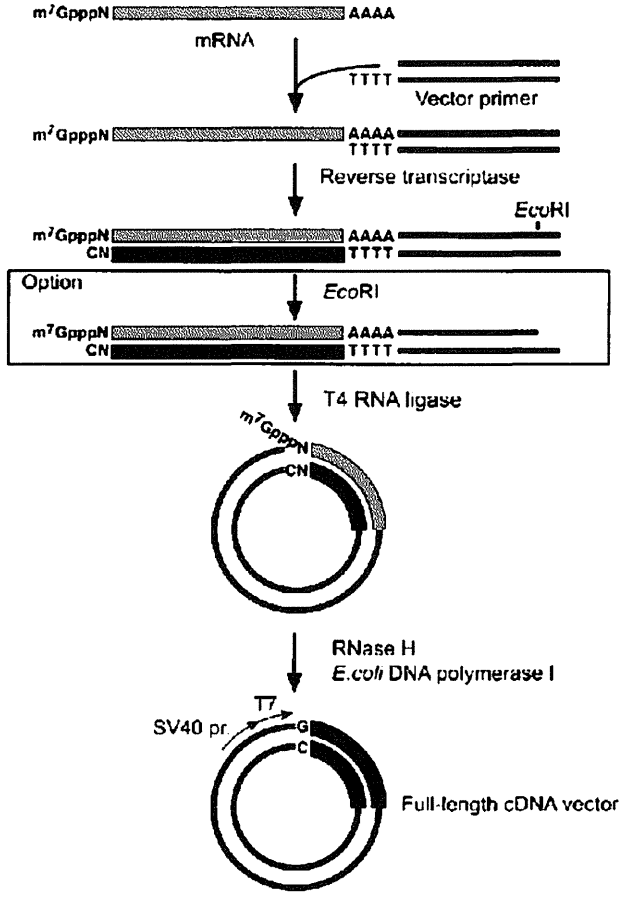

Figure 1. Schematic procedure for the vector-capping method. Several micrograms of total RNA is enough as a starting material. A vector primer has an approximately 60-nucleotide dT tail. The EcoRI digestion step can be omitted. Since the full-length cDNA vector has an SV40 promoter, the encoded protein can be produced in the mammalian cells by introducing the vector into the cells.

When improving this method, we succeeded in developing the vector-capping method shown in Figure 1 (Kato et al., 2005; Kato et al., 2011). Its process is very simple: the first-strand cDNA is synthesized using a vector primer, and then the vector-cDNA conjugate is circularized. The development of this method is attributed to the discovery of an unexpected reaction: the 3' end of the first-strand cDNA can be ligated to the 5' end of the vector primer using "RNA ligase". Furthermore, we found that the full-length cDNA possesses an additional dG at the 5' end. This additional dG is derived from dC that added to the 3' end of the first-strand cDNA by terminal deoxynucleotidyl transferase activity of reverse transcriptase only when the template mRNA has a cap structure. Thus, the presence of the additional dG at the 5' end assures the intactness of the 5'-end capped site sequence of the cDNA.

The full-length cDNA library constructed using the vector-capping method has the following advantages compared with those by conventional methods:

(i) the library is composed of genuine full-length cDNA clones of > 95% content,
(ii) we can identify full-length cDNA by the presence of an additional dG at the 5' end,
(iii) artificial mutation or deletion seldom occurs because the procedure contains neither PCR nor the restriction enzyme treatment step,
(iv) the library contains full-length clones for rare or long-sized genes,
(v) we can easily identify the cDNA for an antisense gene due to the use of the vector primer.

Thus, the resulting cDNA library seems to provide us with the full-length cDNA clones ideal for identifying the alternative TSS, AS, and alternative polyadenylation.

3. Analysis of Full-Length cDNA Clones

3.1. Retina-Derived Full-Length cDNA Libraries

We have been searching for genes responsible for retinitis pigmentosa, which is the major cause of visual impairment among patients visiting our center. To identify a novel candidate gene causing this disease, we identified genes specifically expressed in the retina by analyzing the full-length cDNA libraries that were constructed from human retinal pigment epithelium cell line ARPE-19 and human retinoblastoma cell line Y79 using the vector-capping method (Kato et al., 2005; Oshikawa et al., 2008; Oshikawa et al., 2011). We randomly picked up 100,000 clones from each library and stored them as glycerol stocks. By sequencing the 5' end of approximately 24,000 clones from each library, we identified a total of 39,643 full-length cDNA clones that were classified into 7,067 genes (Oshikawa et al., 2011). In this section, I describe the examples of novel AS variants obtained from the above libraries. Most of full-length clones and the other full-length cDNA libraries remain not fully analyzed: ARPE-19, 52,800 clones; Y79, 52,800 clones; embryonal pluripotent carcinoma cell line NT2/D1, 76,800 clones; human testis, 76,800 clones. If researchers are interested in particular genes, they may find new AS variants by full sequencing of those clones. All

clones are available from RIKEN BioResource Center DNA Bank (http://dna.brc.riken.jp/en/NRCDhumen.html).

3.2. Characterization of Eye-Specific Genes

3.2.1. Aryl Hydrocarbon Receptor Interacting Protein-Like 1 (AIPL1)

Since retinoblastoma cell line Y79 is derived from cone progenitor cells (Xu et al., 2009), Y79 cells expressed various photoreceptor-specific genes. One of abundant eye-specific genes found in our Y79 full-length cDNA libraries was *AIPL1*. *AIPL1* has been identified as a gene responsible for Leber congenital amaurosis (LCA), a severe early-onset retinopathy that leads to visual impairment in infants (Sohocki et al., 2000).

Fifteen clones (C1-C15) encoding AIPL1 were fully sequenced, and their exon-intron structures were determined as shown in Figure 2. Since TSSs were distributed within the range of position -14 to position 13, the same promoter seemed to be used. We identified seven AS variants by the shift of a splicing site, skipping of exon 3, and the alternative use of a polyadenylation signal. The encoded proteins were classified into five isoforms: 384-amino acid (aa) isoform (V1 and V4, 9 clones); 345-aa isoform (V3, 2 clones); 321-aa isoform (V2 and V5, 2 clones); 270-aa isoform (V6, 1 clone); 262-aa isoform (V7, 1 clone). The difference between V1 and V4 and the difference between V2 and V5 were due to the alternative use of the polyadenylation signal. V3 showed the 58-bp downstream shift of the 3' splice site of exon 1 (designated by #7 in Figure 2), resulting in the shift of the initiation codon of the longest ORF from exon 1 to exon 2. This shift causes the loss of the N-terminal 39-aa residues. V2 and V5 lacked exon 3, resulting in the 63-aa deletion from the middle part of the protein. V6 and V7 lacked exon 6 because of the shift of the polyadenylation site, resulting in the deletion of the C-terminal 114-aa residues. Furthermore, V7 showed the 24-bp downstream shift of the 5' splice site of exon 4 (designated by #8), causing the corresponding 8-aa deletion. It should be noted that the Y79 cell-derived transcripts showed five single nucleotide polymorphisms (#1, #2, #4, #5, #6) and 2-bp deletion (#3). As a result, the clones were classified into two haplotypes, and their allelic origin was identified. Half of the clones (C3, C7, C9, C11, C12, C13, C14) were assigned to one haplotype.

The identified variants were compared with RefSeq in GenBank provided by the National Center for Biotechnology Information (NCBI), which is a collection of taxonomically diverse, non-redundant and richly annotated sequences representing naturally occurring molecules of DNA, RNA, and protein (Pruitt et al., 2009). Three RefSeqs were constructed as the transcripts of an *AIPL1* gene. RefSeq1 and RefSeq2 correspond to V1 and V2, respectively. Our collection did not contain the clone corresponding to RefSeq3 that is an AS variant skipping exon 2. In NCBI's GenBank, 13 sequences except for our 14 sequences were registered as *AIPL1* mRNA with an ORF. However, these mRNA sequences had no polyadenylation signal, whereas our clones had a canonical polyadenylation signal, AATAAA (V1-V6) and AGTAAA (V7) followed by a poly(A) tail. The mRNA sequences without a polyadenylation signal were terminated before an A-stretch (A_{15}) at position 1386 of RefSeq1 or an A-rich region ($A_6CA_4CA_4CA_5$) at position 2131. These clones seem to have been synthesized by priming of the oligo(dT) primer to these A-stretch sites during the first-strand cDNA synthesis. Of these 13 mRNAs, five clones correspond to V1, two clones to V2,

and one clone to V6. The remaining five clones were novel variants, suggesting that there would be more variants for *AIPL1* transcripts to be identified.

AIPL1 has been reported to interact with various proteins including NUB1 (Akey et al., 2002), FAT10, FAT10nylated protein, UBA6 (Bett et al., 2012), and the catalytic subunit (alpha) of rod cGMP phosphodiesterase (PDE6A) (Kolandaivelu et al., 2009). AIPL1 is necessary for the proper assembly of functional rod PDE6 subunits (Kolandaivelu et al., 2009) that is a key phototransduction enzyme. AIPL1 is composed of three functional domains: an FKBP-like prolyl peptidyl isomerase (FKBP) domain, a tetratricopeptide (TPR) domain, and Pro-rich domain (PRD).

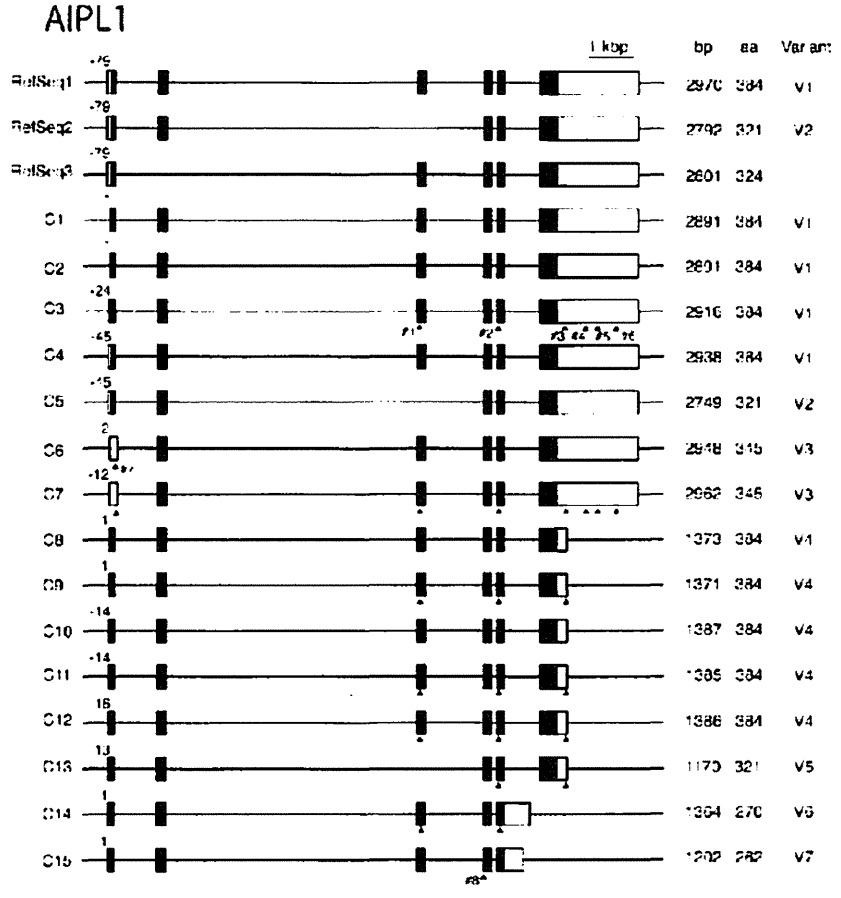

Figure 2. The exon-intron structure of alternative splicing variants for *AIPL1*. Fifteen clones (C1-C15) were classified into seven variants (V1-V7). The relative position of a transcription start site was indicated on exon 1. Arrowheads represent the following sequence variations: #1, SNP (A>G); #2, SNP (A>G); #3, 2-nt deletion (AA>--); #4, SNP (A>G); #5, SNP (A>G); #6, SNP (C>T); #7, downstream splicing site shift (58 bp); #8, downstream splicing site shift (24 bp). RefSeq1, RefSeq2, and RefSeq3 correspond to GenBank Accession No. NM_014336.3, NM_001033054.1, and NM_001033055.1, respectively. Our clones correspond to AB593062.1 - AB593067.1.

V3 encodes an isoform lacking the N-terminal 39-aa residues whose function is unclear. V6 and V7 encode an isoform lacking the C-terminal 114-aa residues corresponding to the

PRD carrying a chaperone activity (Li et al., 2013). This isoform might lose function as a chaperone, because the Trp278X mutant lacking the C-terminal 107-aa residues causes LCA (Sohocki et al., 2000). V2 and V5 encode an isoform lacking the FKBP domain. To elucidate why these isoforms lacking a functional domain are produced in the cell, it is necessary to know the role of each isoform in the AIPL1 regulation system by investigating their localization in the cell or binding activity with other proteins. Although the expression level of each minor variant is low, the total level of minor variants reaches the same level as main variants (V1 and V4), suggesting that isoforms encoded by these minor variants play their own roles in the AIPL1-related system. Since the EST database in GenBank (dbEST) contains novel variants, we would find more novel variants by analyzing the full-length cDNA libraries.

3.2.2. LIM Homeobox 3 (LHX3)

The Y79 libraries contained many clones encoding various eye-specific transcription factors. A transcription factor LHX3 is a member of a large protein family that carries a LIM domain, a Cys-rich zinc-binding domain, and is required for pituitary development and motor neuron specification. Two variants encoding human LHX3 (isoform a and isoform b) have been cloned from pituitary cDNA libraries (Sloop et al., 1999), and they were adopted as RefSeq for *LHX3* genes. The mutation of this gene caused combined pituitary hormone deficiency (CPDH) (Netchine et al., 2000). There is no report for *LHX3* expressed in the eye except for ESTs obtained from eye and pineal gland libraries in dbEST.

Eight clones for *LHX3* were obtained from the Y79 cDNA libraries and four variants were identified as shown in Figure 3. Five clones designated by V1 encoded a 397-aa isoform a. V2 using an alternative promoter had a different exon 1 from V1 and encoded a 402-aa isoform b. As a result, the N-terminal 25-aa sequence of isoform a was different from the N-terminal 30-aa sequence of isoform b. V3 was a novel variant containing an unspliced intron between exon 2 and exon 3, resulting in shortening of the first ORF followed by a longer ORF. The upstream short ORF encoded the N-terminal 89-aa sequence of the isoform a and the downstream long ORF encoded the C-terminal 264-aa sequence of isoform a. V4 using another promoter had a novel exon 1, and encoded a novel 386-aa isoform whose N-terminal 15-aa sequence was different from those of isoform a and b.

V1 and V2 correspond to RefSeq1 and RefSeq2, respectively. V3 and V4 are novel variants. In GenBank two mRNA sequences containing ORF were registered except for our six sequences. These mRNAs correspond to isoform a and isoform b that were cloned from human pituitary cDNA libraries (Sloop et al., 1999). These two sequences had no polyadenylation signal maybe due to the use of a random primer in synthesizing cDNA. The dbEST contained 13 sequences: nine from eye, two from pineal grand, and two from brain. All sequences lacked a sequence corresponding to exon 1. The cDNA libraries used for cloning these ESTs were prepared using conventional methods that contained a NotI or XhoI treatment step. Since the sequence of the full-length cDNA contained NotI and XhoI sites, the fragmentation of synthesized cDNA might occur. Although dbEST contained many ESTs isolated from the Y79 full-length cDNA libraries prepared using the oligo-capping method, there was no EST for *LHX3* in spite of an abundant gene. This was explained by the presence of an SfiI site in the *LHX3* gene, because the oligo-capping protocol contained a step using linkers with the SfiI site (Suzuki and Sugano, 2001).

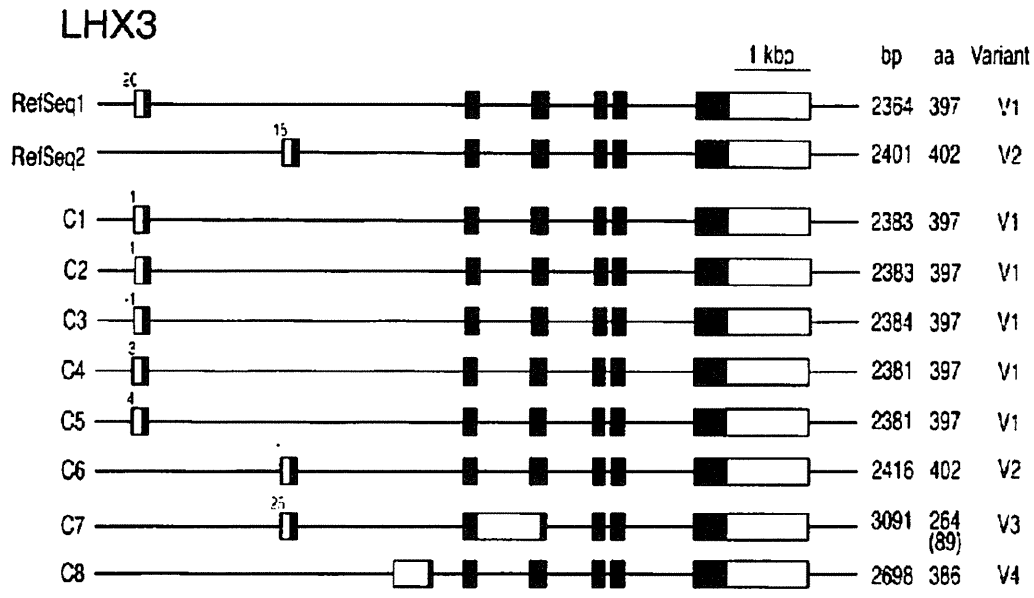

Figure 3. The exon-intron structure of alternative splicing variants for *LHX3*. Eight clones (C1-C8) were classified into four variants (V1-V4). The relative position of a transcription start site was indicated on exon 1. RefSeq1 and RefSeq2 correspond to GenBank Accession No. NM_178138.4 and NM_014564.3, respectively. Our clones correspond to AB593042.1 - AB593055.1.

LHX3 has two LIM domains, a homeodomain, and a C-terminal LHX3-specific domain. Three isoforms for human LHX3 (isoform a, isoform b, short isoform) have been reported (Sloop et al., 2001). Isoforms a and b differ in their N-terminal sequence. The N-terminal sequence of isoform b has been shown to inhibit the binding of LHX3 to DNA. Furthermore, isoform a produced a 264-aa short isoform starting at Met-134 and thus lacking LIM domains. This short isoform showed a transcription factor activity owing to the downstream region including a homeodomain. Interestingly, the longest ORF of the novel variant V3 encoded this short isoform. It is also interesting whether the N-terminal sequence of the novel isoform encoded by V4 affects the activity of LHX3. In the pituitary gland, the expression pattern of V1 and V2 differed between cell lines, suggesting that these variants play a different role in the regulation of gene expression during development of each cell type (Sloop et al., 2001). The role of these variants in development of the retina remains to be solved.

3.2.3. Neural Retina-Specific Leucine Zipper Protein (NRL)

NRL is a basic motif-leucine zipper transcription factor that plays an essential role in the differentiation of photoreceptor cells (Swaroop et al., 1992). The Y79 libraries contained seven clones encoding NRL, which were classified into five variants as shown in Figure 4. V1 corresponded to RefSeq1, and V2 had a shortened 3'-untranslated region (3'-UTR) due to the alternative use of a polyadenylation signal. V3, V4, and V5 used an alternative promoter located between exon 1 and exon 2. Although the 3'-end splice sites of a new exon 1 of these three variants were slightly different, these variants had the same ORF in exon 2, which encoded a 98-aa sequence starting from Met-140 in the isoform encoded by V1. When the

expression vectors of these variants were introduced into cultured cells, the corresponding 98-aa short protein was produced (unpublished data). Since this short isoform lacked a minimal transactivation domain (Friedman et al., 2004), it showed no transcriptional activity as expected. The Leu zipper domain has been reported to interact with a CRX homeodomain (Mitton et al., 2000). The short isoform carrying only the Leu zipper domain may be involved in the regulation of complex formation through this domain.

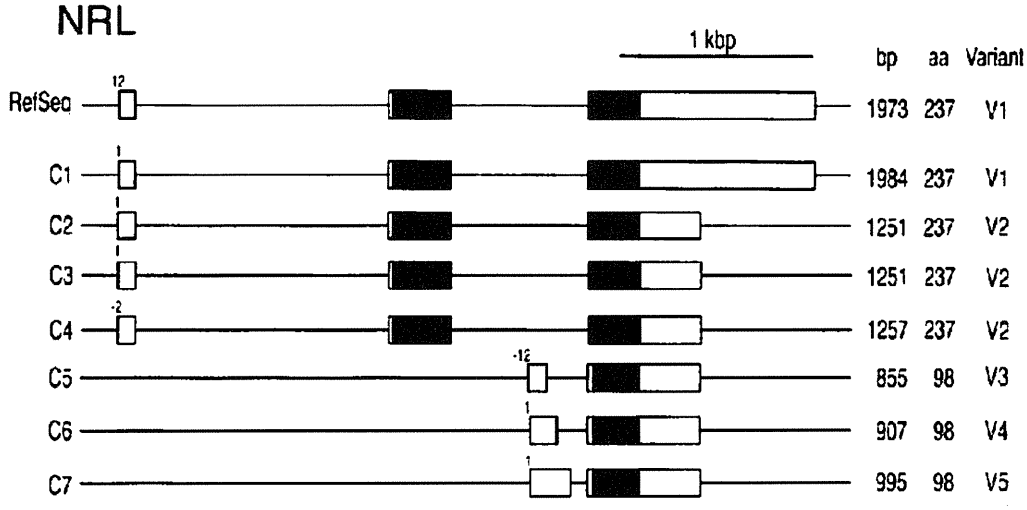

Figure 4. The exon-intron structure of alternative splicing variants for *NRL*. Seven clones (C1-C7) were classified into four variants (V1-V5). The relative position of a transcription start site was indicated on the first exon. RefSeq correspond to GenBank Accession No. NM_006177.3. Our clones correspond to AB593102.1 - AB593104.1.

GenBank contained three mRNAs (one corresponding to V1 and two to V2) except for the six sequences we registered. The dbEST contained six V4 sequences and one V5. Although one research group cloned the cDNA corresponding to V4, the authors could not judge whether it was a full-length or truncated one (Wistow et al., 2002). Like this case, when only one cDNA different from known ones is cloned, it is difficult to judge its intactness. Even in such case, our clone can be identified to be a full-length clone by the presence of an additional dG at the 5' end. Since the V1 sequence had SfiI sites, the Y79 cDNA libraries prepared using the oligo-capping method missed cloning the full-length cDNA for *NRL*.

3.2.4. OTX2 Antisense RNA 1 (OTX2-AS1)

Our libraries contained many novel non-coding RNAs including rare variants. As an example of an eye-specific non-coding RNA, we obtained five clones for *OTX2-AS1* from the Y79 libraries. These clones showed a variety of structures as shown in Figure 5. The length of cDNA varied from 303 bp of V2 to 2900 bp of V3 and the splicing pattern varied from clone to clone. Only a part of exon 1 was shared within all variants. The sequences in dbEST are also rich in variety. *OTX2-AS1* is a gene transcribed in the opposite direction at the upstream region of the locus of orthodenticle homeobox 2 (OTX2) that is a transcription factor involved in the development of brain and sensory organs (Alfano et al., 2005). Our Y79 libraries also contained three clones for *OTX2*, each of which is an AS variant (data not

shown). The exon 1 of a variant of mouse *Otx2-as1* overlapped with the antisense strand of the exon 1 of *Otx2* (Alfano et al., 2005), but there was no human variant whose exon 1 overlapped to *OTX2*. Since all ESTs for *OTX2-AS1* in dbEST were obtained from retina cDNA libraries, this gene might be involved in the development of retina. The presence of diverse AS variants implies the complex regulation system by this gene.

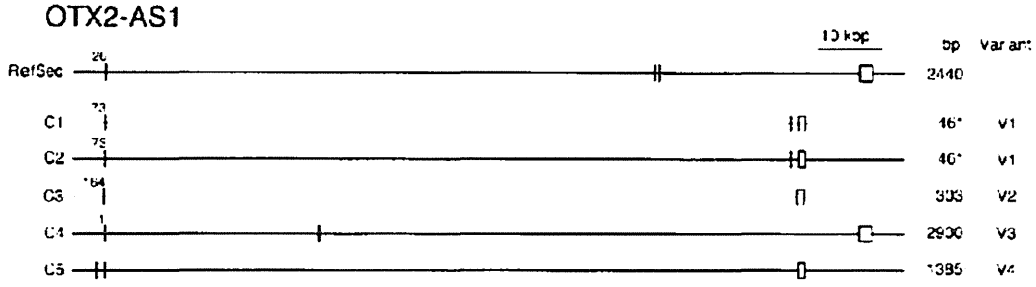

Figure 5. The exon-intron structure of alternative splicing variants for *OTX2-AS1*. Five clones (C1-C5) were classified into four variants (V1-V4). The relative position of a transcription start site was indicated on the first exon. RefSeq correspond to GenBank Accession No. NR_029385.1. Our clones correspond to AB593038.1 - AB593041.1.

3.3. Characterization of Long-Sized Genes

3.3.1. Very-Long-Sized Genes

We succeeded to clone 82 full-length cDNAs with >7kbp from the libraries prepared using the vector-capping method (Oshikawa et al., 2008; Oshikawa et al., 2011). The ARPE-19 libraries contained full-length cDNA clones encoding golgin B1 (*GOLGB1*, 11.2kbp), NEDD4 binding protein 2 (*N4BP2*, 9.7 kbp), acetyl-CoA carboxylase alpha (*ACACA*, 9.5 kbp), filamin B, beta (*FLNB*, 8.0-9.4 kbp), filamin C, gamma (*FLNC*, 9.2 kbp), spectrin, beta, non-erythrocytic 1 (*SPTBN1*, 8.4 kbp), filamin A, alpha (*FLNA*, 8.2 kbp), collagen, type V, alpha 1 (*COL5A1*, 8.1 kbp), spectrin, alpha, non-erythrocytic 1 (*SPTAN1*, 7.8 kbp), fibronectin 1 (*FN1*, 7.8 kbp), myosin, heavy chain 9, non-muscle (*MYH9*, 7.4 kbp), and agrin (*AGRN*, 7.3 kbp). The Y79 libraries contained full-length cDNAs encoding Dmx-like 1 (*DMXL1*, 12.8 kbp), GOLGB1 (11.1 kbp), SEC16 homolog A (*SEC16A*, 9.0 kbp), FLNA (8.4 kbp), eyes shut homolog (*EYS*, 8.0 kbp). Out of these genes, four genes having multiple AS variants were selected and their structures were analyzed below.

3.3.2. Golgin B1 (GOLGB1)

GOLGB1 is a huge integral membrane protein located in Golgi, originally named giantin (Linstedt et al., 1993). Two research groups cloned approximately 10-kbp cDNA encoding a protein that reacts with autoantibody contained in sera of patients with chronic rheumatism: mRNA1, 10,295-bp cDNA encoding 3,225-aa protein (Sohda et al., 1994); mRNA2, 10,300-bp cDNA encoding 3,259-aa protein (Seelig et al., 1994). These clones were not derived from a single mRNA. The full sequence was constructed by combining the sequences of cDNA fragments. Thus, it is doubtful whether the sequence reflects the true structure of the AS variant.

Our libraries contained two full-length cDNA clones for *GOLGB1* (V1 from ARPE-19, 11.2 kbp; V2 from Y79, 11.1 kbp). The exon-intron structures of the above four clones were different as shown in Figure 6. In GenBank, RefSeqs seem to be constructed by referring to registered mRNAs including our clones: RefSeq1 to V1; RefSeq2 to mRNA2; RefSeq3 to mRNA1; RefSeq4 to V2. Although exon 1 was shared within all clones, they were all different AS variants encoding the protein with the different number of aa residues. In V1 and mRNA1, the 28-bp downstream shift of the 3' splice site of exon 2 (designated by #1 in Figure 6) caused a frame shift, and thus the initiation codon in exon 3 was used. As a result, the N-terminal sequence was shortened by 39 aa compared with the isoform for V1. Furthermore, V2 lacked a 41-aa sequence corresponding to exon 7 by exon skipping. mRNA2 lacked a 5-aa sequence by the 15-bp downstream shift of the 5' splice site of exon 7 (#2). V1 had 5-aa insertion by the 15-bp downstream shift of the 3' splice site of exon 18 (#3). The dbEST contained ESTs carrying not only these four variations but also other variations including the shift of splice site or skipping of exon 4, 6, 10, 11, 12, 15-21, suggesting the presence of diverse AS variants of *GOLGB1*.

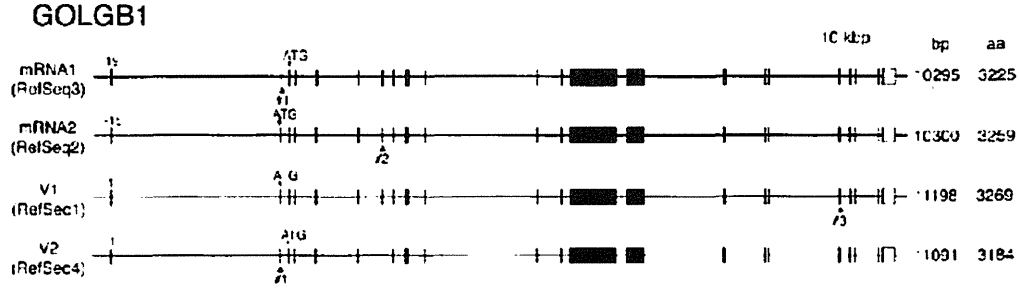

Figure 6. The exon-intron structure of alternative splicing variants for *GOLGB1*. C1 was cloned from the ARPE-19 cDNA library and C2 from the Y79 cDNA library. The relative position of a transcription start site was indicated on exon 1. Arrowheads represent the following sequence variations: #1, the downstream shift of the 3' splice site (28 bp); #2, the downstream shift of the 5' splice site (15bp); #3, the downstream shift of the 5' splice site (15 bp). mRNA1 and mRNA2 correspond to GenBank Accession No. D25542.1 and X75304.1, respectively. Our clones correspond to AB371588.1 and AB593126.1.

GOLGB1 is an integral membrane protein involved in linkage between a Golgi membrane and a COPI vesicle (Sönnichsen et al., 1998). This protein has no N-terminal secretory signal sequence, but has a C-terminal transmembrane domain. Most of the cytoplasmic part is composed of a coiled-coil structure in which the AS variants had deletion or insertion. This structure is thought to be involved in regulation of retrograde trafficking to the endoplasmic reticulum in Golgi apparatus through binding of small GTPase such as Rab6 and Rab1 (Rosing et al., 2007). Thus, each AS variant may play a role in the regulation of this trafficking. To elucidate the detailed mechanism, further investigation is necessary using these AS variants.

3.3.3. Filamin A, Alpha (FLNA)

FLNA was most abundantly found in our libraries as a long-sized gene with >7kbp. FLNA is an actin-binding protein involved in change in cell shape and migration through crosslinking of actin filaments and linking actin filaments to membrane glycoproteins. ARPE-

19 and Y79 libraries contained eight clones (7.3 – 8.2 kbp) and one clone (8.4 kbp), respectively. These clones were classified into four variants as shown in Figure 7. V1 and V2 were main components in ARPE-19 cells. V2 lacked exon 30, resulting in deletion of an 8-aa sequence. V3 lacked exon 38-41 because of AS between the middle splice site in exon 37 and the middle splice site in exon 42. Y79-originated V4 started from a 135-bp upstream TSS compared with V1, resulting in the generation of a novel initiation codon that caused 27-aa extension of the N-terminal sequence.

Figure 7. The exon-intron structure of alternative splicing variants for *FLNA*. Eight clones obtained from ARPE-19 were classified into three variants (V1-V3). V4 was obtained from Y79. The relative position of a transcription start site was indicated on the first exon. "No." represents the number of obtained clones. Arrowheads represent the following sequence variations: #1, the 96-bp upstream shift of the 3' splice site of exon 37; #2, the 72-bp downstream shift of the 5' splicing site of exon 42. RefSeq1 and RefSeq2 correspond to GenBank Accession No. NM_001456.3 and NM_001110556.1, respectively. Our clones correspond to AB191259.1 - AB191260.1, AB371574.1- AB371579.1 and AB593010.1.

GenBank contained 3 mRNAs with a full ORF except for the nine clones we registered. The sequence of mRNA1 (X53416) was constructed by combining seven fragments cloned from a human endothelial cell (Gorlin et al., 1990). mRNA2 (AK090427) originated from a single mRNA and had the same initiation codon as V4, but lacked exon 30. Although this clone seems to be near full-length cDNA, the registrant regarded it as a truncated clone maybe because of no evidence for intactness of the 5' end of the cDNA. mRNA3 (GU727643) was synthesized with RT-PCR based on RefSeq. RefSeq1 corresponding to mRNA1 has exon 1, which uses an upstream alternative promoter. Its ORF starts from the same initiation codon with V4. Our nine clones had no exon 1. There were 14 sequences having exon 1 in dbEST. RefSeq2 was constructed based on our clone V1 except for exon 1. The dbEST contained a sequence (CN421698) that has the same deletion as V3.

FLNA has a rod-like structure composed of 24 repeats of the beta-pleated sheet unit: an actin-filament binding domain (Rod1, repeats 1-15), a partner protein binding domain (Rod2, repeats 16-23), a self-assembly domain (repeat 24), and a hinge linking the domains (Hinge-1 and Hinge-2) (Nakamura et al. 2011). V2 lacked the 8-aa residues that were located in the last repeat 15 of Rod1. V3 lacked the 114-aa sequence that was the part of repeats 18 and 19 of Rod2. Since these repeats are known to bind to several partner proteins, these isoforms may lose a function borne by the corresponding repeat.

3.3.4. Filamin B, Beta (FLNB)

FLNB and FLNC as well as FLNA are a member of a filamin family. The ARPE-19 libraries contained four *FLNB* clones (8-9 kbp) and one *FLNC* clone (9156 bp). The *FLNC* clone corresponded to RefSeq (data not shown). All clones for FLNB were different AS variants as shown in Figure 8. Four RefSeqs are constructed based on our four clones. They had a similar TSS. V4 had a total of 47 exons and the other three variants lacked exon 26 (93 bp, 31 aa). V2 lacked 11-aa residues due to the 33-bp upstream shift of 3' splice site of exon 31 (designated by #1 in Figure 8). V1 and V4 had a shorter exon 47 due to the use of an alternative polyadenylation signal. GenBank contained two mRNA sequences except for our clones. These two sequences corresponding to V1 were constructed by combining the sequences of cDNA fragments (Takafuta et al., 1998; Xu et al., 1998). Xu et al. (Xu et al., 1998) have reported near full-length cDNA clone (9.5 kb) for V3 derived from a single mRNA. The dbEST contained four clones possessing a shortened exon 31 found in V2 and two clones possessing exon 26 found in V4.

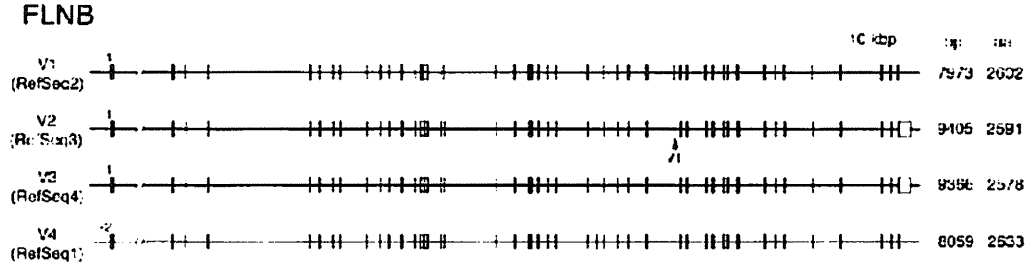

Figure 8. The exon-intron structure of alternative splicing variants for *FLNB*. Four clones were all different variants. Four RefSeqs are constructed based on our four clones as shown in parenthesis. The relative position of a transcription start site was indicated on the first exon. Arrowhead #1 represents the 33-bp upstream shift of the 3' splice site of exon 31. Our clones correspond to AB191258.1 and AB371580.1- AB371582.1.

FLNB has a domain structure similar to FLNA. V4 encoded an isoform possessing 31-aa insertion in the middle part of repeat 31 due to the insertion of exon 26. The isoform encoded by V3 lacked Hinge-1 of 24-aa residues corresponding to exon 31. V2 encoded an isoform lacking the 11-aa C-terminal half of Hinge-1. The deletion of these aa sequences might affect the function of each isoform through the change of binding ability to the partner proteins as well as FLNA. For example, van der Flier et al. showed that an FLNB fragment lacking a part of repeat 19-20 or C-terminal repeat 24 obtained using RT-PCR had a different binding ability to integrin beta subunit (van der Flier, 2002). Furthermore, they showed that the expression pattern of these variants varied from tissue to tissue and during myogenesis. We have to keep in mind that this kind of experiment using RT-PCR shows the expression level of only a partial sequence of transcript and the expression pattern does not reflect the change of the full-length transcript.

3.3.5. Eyes Shut Homolog (EYS)

EYS is an extracellular matrix specifically produced in photoreceptor cells (Abd El-Aziz, 2008; Collin et al., 2008). Recently, we showed that one-third of Japanese patients with retinitis pigmentosa had founder mutations in the *EYS* gene. RefSeq1 for the *EYS* gene

comprises 43 exons as shown in Figure 9 and the length of mRNA is 11 kb. In GenBank, there are two short RefSeqs terminating by exon 11. RefSeq2 has a long 3'-UTR. RefSeq3 uses an alternative promoter located between exon 2 and exon 3, resulting in the formation of a new exon 1. The Y79 libraries contained two clones only for short variants. Although V1 was a very-long-sized clone with an insert of 7,898 bp, it terminated by exon 11 containing a long 3'-UTR and encoded the N-terminal 594-aa sequence as well as RefSeq2. The exon 11 of RefSeq2 was split due to splicing. V2 terminated by exon 4 and encoded a short isoform of 318 aa. The EYS protein comprises 27 EGF-like domains and 5 laminin G-like domains. Thus, the isoform encoded by V1 terminated at the middle of the sixth EGF-like domain, and the isoform for V2 at the middle of the third EGF-like domain. The function of these short forms of the EYS protein remains to be solved.

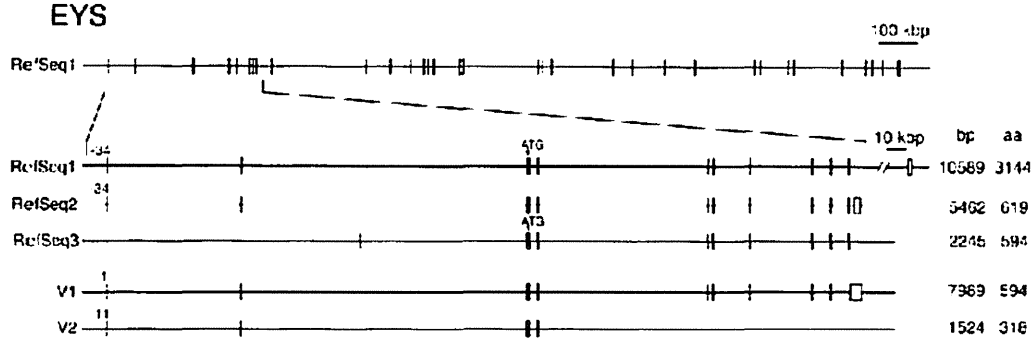

Figure 9. The exon-intron structure of alternative splicing variants for *EYS*. Two variants were cloned from the Y79 cDNA library. The relative position of a transcription start site was indicated on the first exon. RefSeq1, RefSeq2 and RefSeq3 correspond to GenBank Accession No. NM_001142800.1, NM_001142800.2, and NM_198283.1, respectively. Our clones correspond to AB593114.1 and AB593112.1.

4. Full-Length cDNA Libraries Derived from Other Species

The power of the vector-capping method was first demonstrated by the transcriptome analysis of budding yeast. Miura et al. performed a large-scale analysis of full-length cDNA libraries prepared from budding yeast cells growing exponentially in a minimal medium and meiotic cells (Miura et al., 2006). They identified 11,575 TSSs associated with 3,638 genes, suggesting that most yeast genes have two or more TSSs. They also identified 45 previously undescribed introns, including those spliced alternatively. Furthermore, they found 667 transcripts in the intergenic region and 367 transcripts derived from antisense strands of known genes. These results suggest that many genes remain unidentified even in an intensively analyzed simple organism such as budding yeast.

Since the vector-capping method was published in 2005 (Kato et al., 2005), it has been adopted by various research projects to construct cDNA libraries from various tissues of various species: plants such as burma mangrove (Miyama et al., 2006), miniature tomato (Aoki et al., 2010), Chinese cabbage (Abe et al., 2011), rubber tree (Suzuki et al., 2012);

mammals such as macaque monkey (Osada et al., 2009), pig (Uenishi et al., 2012), common marmoset (Tatsumoto et al., 2013); parasites such as *Haemaphysalis* (Zhou et al., 2006), *Echinococcus* (Watanabe et al., 2007), *Babesia* (Aboge et al., 2008); hagfish (Uchida et al., 2010); *Bombyx mori* nucleopolyhedrovirus (Katsuma et al., 2011). In many cases, it seems to have been difficult to obtain a large amount of starting material. The vector-capping method requires only several micrograms of total RNA. This seems to be one reason why this method was adopted to prepare the libraries from the above samples. The cloning ability of a long-sized cDNA was confirmed by the cloning of very-long-sized genes (9.1kb and 9.8kb) that encode egg case silk from a wasp spider (Zhao et al., 2006).

5. Problems on Identification of AS Variants

5.1. AS Variants of Rare or Long-Sized Genes

In the above sections, I have described some problems in identifying AS variants using the conventional methods. A serious problem occurred in the case of a long-sized gene. The combination of alternative promoter usage, multiple alternative splicing sites, and alternative polyadenylation can produce diverse forms of transcripts. The partial sequence analyses using RT-PCR or RNA-seq do not disclose this combination. To determine the precise structure of the AS variant, it is necessary to determine the full sequence of a single mRNA. One solution for this requirement is to determine the full sequence of a full-length cDNA derived from a single mRNA.

Another problem is related to the intactness of the full-length cDNA. In the case of an abundant gene, many cDNA clones can be obtained. If these cDNAs are shown to start at the similar site by comparing their 5'-end sequences, we could regard them as a full-length or near full-length cDNA having a capped site sequence. However, a rare gene may give only one cDNA clone, thus we cannot judge the intactness of this cDNA. The same problem occurs in the case of very short or very long genes. It may be difficult to judge whether the cDNA are derived from intact mRNA or degraded mRNA. The vector-capping method solves these problems. We can judge the intactness of the cDNA by inspecting the presence of the additional dG at the 5' end of the cDNA.

5.2. Synthesis of Full-Length cDNA

It has been difficult to obtain a full-length cDNA for a rare or long-sized gene using conventional methods. Here, the problems are shown with regard to each step of cDNA synthesis.

(1) Oligo(dT) Priming

The conventional methods usually use an oligo(dT) primer of ~20 nt to synthesize the first-strand cDNA. When mRNA has a short A stretch, the oligo(dT) primer can accidentally hybridize to this site and be used for cDNA synthesis, resulting in missing the downstream part of mRNA to a poly(A) tail. A good example is an *AIPL1* gene shown in section 3.2.1.

We observed several other examples of such mispriming (data not shown). The vector-capping method uses a vector primer possessing approximately 60-nt dT at one end of the vector. The long dT tail may rarely prime a short A stretch in mRNA.

(2) Reaction Conditions

According to our experience, the amount of template mRNA, reverse transcriptase and substrate nucleotides seem to be essential factors that are related to biases by the expression level or size of mRNA. Usually the first-strand cDNA synthesis is carried out using several micrograms of poly(A)$^+$RNA. In these reaction conditions, most reverse transcriptase and substrate nucleotides seem to be consumed to synthesize cDNA mainly from abundant or short-sized mRNAs, causing biases by the expression-level and size of mRNA. In the vector–capping method, total RNA is used as a template in place of poly(A)$^+$RNA to synthesize the first-strand cDNA under the same reaction conditions. Thus, the amount of enzyme and substrate might be enough to synthesize cDNA from rare or long-sized mRNAs. Omitting mRNA purification steps also may help to reduce these biases.

(3) PCR Step

Some conventional methods including the oligo-capping method contain a PCR step in the procedure for preparing the cDNA library. The amplification step by PCR may cause bias by the expression level and the size of mRNA. In fact, when the full-length cDNA libraries prepared from monkey liver and kidney by the oligo-capping method were compared with those prepared by the vector-capping method, the redundancy of the vector-capped libraries is lower than those of the oligo-capped libraries (Osada et al., 2009). To synthesize rare or long-sized cDNAs, the PCR step should be avoided.

(4) Restriction Enzyme Treatment

The conventional methods contain a linker attachment step, in which a oligonucleotide linker with a restriction enzyme site (e.g. NotI, EcoRI, SalI, XhoI, SfiI et al.) are ligated to the double-stranded cDNA and then after cutting by restriction enzyme the cDNA are introduced into a vector. If the cDNA has the same restriction enzyme site as the linker, it is difficult to obtain full-length cDNA or any cDNA in some cases. The examples are shown in the above sections on *LHX3* and *NRL*. Many clones registered in dbEST seem to be a truncated cDNA that was generated due to this step.

(5) Size Fractionation

Some protocol contains a size fractionation step to remove short cDNA fragments. This step should be avoided because there are many short transcripts having a poly(A) tail. We observed such short full-length cDNAs with < 100 bp (data not shown).

5.3. Vector-Capping Method

The vector-capping method solves all the above problems. Thus, this will be the most effective method to synthesize genuine full-length cDNAs at present. However, this has one limitation. It is difficult to obtain full-length cDNA clones from a low-quality RNA sample

containing highly degraded mRNA, because this protocol does not contain a step for experimentally selecting full-length cDNAs, such as a cap-dependent linker ligation in the oligo-capping method. In addition, it requires a lot of labor and cost to search novel AS variants from the vector-capped libraries. This is the case particularly when the target cell expresses genes with low complexity. In that case, we should use a subtraction or normalization protocol together. If the target gene has been decided, we may isolate in advance target cDNA using a probe for the target gene.

Conclusion

Here I have demonstrated that the vector-capping method provides us with a high-quality cDNA library composed of genuine full-length cDNA clones derived from a single mRNA and that the obtained clones can be used to effectively identify AS variants. This library contains many full-length cDNA clones for rare or long-sized genes whose intactness is guaranteed. By analyzing these clones, we can identify novel AS variants for rare or long-sized genes that have been difficult to obtain using conventional methods. These results suggest that comprehensive, in-depth analysis of full-length cDNA clones isolated from the vector-capped libraries is the most effective way to identify an entire set of AS variants. Furthermore, these full-length cDNA clones can be used as a resource for producing the encoded proteins. I hope that the vector-capping method will be widely used for analyzing full-length AS variants derived from various tissues of various species.

References

Abd El-Aziz MM, Barragan I, O'Driscoll CA, et al. EYS, encoding an ortholog of Drosophila spacemaker, is mutated in autosomal recessive retinitis pigmentosa. *Nat Genet.* 2008;40(11):1285-1287.

Abe H, Narusaka Y, Sasaki I, Hatakeyama K, et al. Development of full-length cDNAs from Chinese cabbage (Brassica rapa Subsp. pekinensis) and identification of marker genes for defense response. *DNA Res.* 2011;18(4):277-289.

Aboge GO, Jia H, Terkawi MA, et al. Cloning, expression, and characterization of Babesia gibsoni dihydrofolate reductase-thymidylate synthase: inhibitory effect of antifolates on its catalytic activity and parasite proliferation. *Antimicrob Agents Chemother.* 2008;52(11):4072-4080.

Akey DT, Zhu X, Dyer M, et al. The inherited blindness associated protein AIPL1 interacts with the cell cycle regulator protein NUB1. *Hum Mol Genet.* 2002;11(22):2723-2733.

Alfano G, Vitiello C, Caccioppoli C, et al. Natural antisense transcripts associated with genes involved in eye development. *Hum Mol Genet.* 2005;14(7):913-923.

Aoki K, Yano K, Suzuki A, et al. Large-scale analysis of full-length cDNAs from the tomato (Solanum lycopersicum) cultivar Micro-Tom, a reference system for the Solanaceae genomics. *BMC Genomics.* 2010;11:210.

Beaudoing E, Freier S, Wyatt JR, Claverie JM, Gautheret D. Patterns of variant polyadenylation signal usage in human genes. *Genome Res.* 2000;10(7):1001-1010.

Bett JS, Kanuga N, Richet E, et al. The inherited blindness protein AIPL1 regulates the ubiquitin-like FAT10 pathway. *PLoS One.* 2012;7(2):e30866.

Clamp M, Fry B, Kamal M, et al. Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci U S A.* 2007;104(49):19428-19433.

Collin RW, Littink KW, Klevering BJ, et al. Identification of a 2 Mb human ortholog of Drosophila eyes shut/spacemaker that is mutated in patients with retinitis pigmentosa. *Am J Hum Genet.* 2008;83(5):594-603.

Djebali S, Davis CA, Merkel A, et al. Landscape of transcription in human cells. *Nature.* 2012;489(7414):101-108.

ENCODE Project Consortium. A User's Guide to the Encyclopedia of DNA Elements (ENCODE). *PLoS Biol.* 2011;9(4):e1001046.

Friedman JS, Khanna H, Swain PK, et al. The minimal transactivation domain of the basic motif-leucine zipper transcription factor NRL interacts with TATA-binding protein. *J Biol Chem.* 2004;279(45):47233-47241.

Gorlin JB, Yamin R, Egan S, et al. Human endothelial actin-binding protein (ABP-280, nonmuscle filamin): a molecular leaf spring. *J Cell Biol.* 1990;111(3):1089-1105.

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature.* 2004;431(7011):931-945.

Kato S, Ohtoko K, Ohtake H, Kimura T. Vector-capping: a simple method for preparing a high-quality full-length cDNA library. *DNA Res.* 2005;12(1):53-62.

Kato S, Oshikawa M, Ohtoko K. Full-length transcriptome analysis using a bias-free cDNA library prepared with the vector-capping method. *Methods Mol Biol.* 2011;729:53-70.

Kato S, Sekine S, Oh SW, Kim NS, et al. Construction of a human full-length cDNA bank. *Gene.* 1994;150(2):243-250.

Katsuma S, Kang W, Shin-i T, et al. Mass identification of transcriptional units expressed from the Bombyx mori nucleopolyhedrovirus genome. *J Gen Virol.* 2011;92(Pt 1):200-203.

Kimura K, Wakamatsu A, Suzuki Y, et al. Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res.* 2006;16(1):55-65.

Kolandaivelu S, Huang J, Hurley JB, Ramamurthy V. AIPL1, a protein associated with childhood blindness, interacts with alpha-subunit of rod phosphodiesterase (PDE6) and is essential for its proper assembly. *J Biol Chem.* 2009;284(45):30853-30861.

Li J, Zoldak G, Kriehuber T, et al. Unique Proline-Rich Domain Regulates the Chaperone Function of AIPL1. *Biochemistry.* 2013;52 (12):2089–2096.

Linstedt AD, Hauri HP. Giantin, a novel conserved Golgi membrane protein containing a cytoplasmic domain of at least 350 kDa. *Mol Biol Cell.* 1993;4(7):679-693.

Mitton KP, Swain PK, Chen S, Xu S, Zack DJ, Swaroop A. The leucine zipper of NRL interacts with the CRX homeodomain. A possible mechanism of transcriptional synergy in rhodopsin regulation. *J Biol Chem.* 2000;275(38):29794-29799.

Miura F, Kawaguchi N, Sese J, et al. A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. *Proc Natl Acad Sci U S A.* 2006;103(47):17846-17851.

Miyama, M., Shimizu H., Sugiyama, M. and Hanagata N. Sequencing and analysis of 14,842 expressed sequence tags of burma mangrove, Bruguiera gymnorrhiza. Plant Science. 2006;171(2):234-241.

Modrek B, Resch A, Grasso C, Lee C. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.* 2001;29(13):2850-2859.

Nakamura F, Stossel TP, Hartwig JH. The filamins: organizers of cell structure and function. *Cell Adh Migr.* 2011;5(2):160-169.

Netchine I, Sobrier ML, Krude H, et al. Mutations in LHX3 result in a new syndrome revealed by combined pituitary hormone deficiency. *Nat Genet.* 2000;25(2):182-186.

Osada N, Hirata M, Tanuma R, et al. Collection of Macaca fascicularis cDNAs derived from bone marrow, kidney, liver, pancreas, spleen, and thymus. *BMC Res Notes.* 2009;2:199.

Oshikawa M, Sugai Y, Usami R, Ohtoko K, Toyama S, Kato S. Fine expression profiling of full-length transcripts using a size-unbiased cDNA library prepared with the vector-capping method. *DNA Res.* 2008;15(3):123-136.

Oshikawa M, Tsutsui C, Ikegami T, et al. Full-length transcriptome analysis of human retina-derived cell lines ARPE-19 and Y79 using the vector-capping method. *Invest Ophthalmol Vis Sci.* 2011;52(9):6662-6670.

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet.* 2008;40(12):1413-1415.

Pruitt KD, Tatusova T, Klimke W, Maglott DR. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.* 2009;37(Database issue):D32-36.

Rosing M, Ossendorf E, Rak A, Barnekow A. Giantin interacts with both the small GTPase Rab6 and Rab1. *Exp Cell Res.* 2007;313(11):2318-2325.

Seelig HP, Schranz P, Schröter H, Wiemann C, Renz M. Macrogolgin--a new 376 kD Golgi complex outer membrane protein as target of antibodies in patients with rheumatic diseases and HIV infections. *J Autoimmun.* 1994;7(1):67-91.

Sloop KW, Dwyer CJ, Rhodes SJ. An isoform-specific inhibitory domain regulates the LHX3 LIM homeodomain factor holoprotein and the production of a functional alternate translation form. *J Biol Chem.* 2001;276(39):36311-36319.

Sloop KW, Meier BC, Bridwell JL, Parker GE, Schiller AM, Rhodes SJ. Differential activation of pituitary hormone genes by human Lhx3 isoforms with distinct DNA binding properties. *Mol Endocrinol.* 1999;13(12):2212-2225.

Sohda M, Misumi Y, Fujiwara T, Nishioka M, Ikehara Y. Molecular cloning and sequence analysis of a human 372-kDA protein localized in the Golgi complex. *Biochem Biophys Res Commun.* 1994;205(2):1399-1408.

Sohocki MM, Bowne SJ, Sullivan LS, et al. Mutations in a new photoreceptor-pineal gene on 17p cause Leber congenital amaurosis. *Nat Genet.* 2000;24(1):79-83.

Sönnichsen B, Lowe M, Levine T, Jämsä E, Dirac-Svejstrup B, Warren G. A role for giantin in docking COPI vesicles to Golgi membranes. *J Cell Biol.* 1998;140(5):1013-1021.

Suzuki N, Uefuji H, Nishikawa T, et al. Construction and analysis of EST libraries of the trans-polyisoprene producing plant, Eucommia ulmoides Oliver. *Planta.* 2012;236(5):1405-1417.

Suzuki Y, Sugano S. Construction of full-length-enriched cDNA libraries. The oligo-capping method. *Methods Mol Biol.* 2001;175:143-153.

Swaroop A, Xu JZ, Pawar H, Jackson A, Skolnick C, Agarwal N. A conserved retina-specific gene encodes a basic motif/leucine zipper domain. *Proc Natl Acad Sci USA.* 1992;89(1):266-270.

Takafuta T, Wu G, Murphy GF, Shapiro SS. Human beta-filamin is a new protein that interacts with the cytoplasmic tail of glycoprotein Ibalpha. *J Biol Chem.* 1998;273(28):17531-17538.

Takeda J, Suzuki Y, Nakao M, et al. Large-scale identification and characterization of alternative splicing variants of human gene transcripts using 56,419 completely sequenced and manually annotated full-length cDNAs. *Nucleic Acids Res.* 2006;34(14):3917-3928.

Tatsumoto S, Adati N, Tohtoki Y, et al. Development and Characterization of cDNA Resources for the Common Marmoset: One of the Experimental Primate Models. *DNA Res.* 2013;20(3):255-262.

Uchida K, Moriyama S, Chiba H, et al. Evolutionary origin of a functional gonadotropin in the pituitary of the most primitive vertebrate, hagfish. *Proc Natl Acad Sci U S A.* 2010;107(36):15832-15837.

Uenishi H, Morozumi T, Toki D, Eguchi-Ogawa T, Rund LA, Schook LB. Large-scale sequencing based on full-length-enriched cDNA libraries in pigs: contribution to annotation of the pig genome draft sequence. *BMC Genomics.* 2012;13:581.

van der Flier A, Kuikman I, Kramer D, et al. Different splice variants of filamin-B affect myogenesis, subcellular distribution, and determine binding to integrin [beta] subunits. *J Cell Biol.* 2002;156(2):361-376.

Wakamatsu A, Kimura K, Yamamoto J, et al. Identification and functional analyses of 11,769 full-length human cDNAs focused on alternative splicing. *DNA Res.* 2009;16(6):371-383.

Wang ET, Sandberg R, Luo S, et al. Alternative isoform regulation in human tissue transcriptomes. *Nature.* 2008;456(7221):470-476.

Watanabe J, Wakaguri H, Sasaki M, Suzuki Y, Sugano S. Comparasite: a database for comparative study of transcriptomes of parasites defined by full-length cDNAs. *Nucleic Acids Res.* 2007;35(Database issue):D431-438.

Wistow G, Bernstein SL, Wyatt MK, et al. Expressed sequence tag analysis of human RPE/choroid for the NEIBank Project: over 6000 non-redundant transcripts, novel genes and splice variants. *Mol Vis.* 2002;8:205-220.

Xu Wf, Xie Z, Chung DW, Davie EW. A novel human actin-binding protein homologue that binds to platelet glycoprotein Ibalpha. *Blood.* 1998;92(4):1268-1276.

Xu XL, Fang Y, Lee TC, et al. Retinoblastoma has properties of a cone precursor tumor and depends upon cone-specific MDM2 signaling. *Cell.* 2009;137(6):1018-1031.

Zhao AC, Zhao TF, Nakagaki K, et al. Novel molecular and mechanical properties of egg case silk from wasp spider, Argiope bruennichi. *Biochemistry.* 2006;45(10):3348-3356.

Zhou J, Liao M, Hatta T, Tanaka M, Xuan X, Fujisaki K. Identification of a follistatin-related protein from the tick Haemaphysalis longicornis and its effect on tick oviposition. *Gene.* 2006;372:191-198.