ACADEMIC PRESS

Biochemical and Biophysical Research Communications 293 (2002) 356–365

BBRC

www.academicpress.com

Identification and characterization of *C6orf37*, a novel candidate human retinal disease gene on chromosome 6q14

P.S. Lagali,[a] L.E. Kakuk,[b] I.B. Griesinger,[b] P.W. Wong,[a] and R. Ayyagari[b,*]

[a] *Departments of Biological Sciences, Ophthalmology, and Medical Genetics, University of Alberta, Edmonton, Alta., Canada*
[b] *Kellogg Eye Center, University of Michigan, 100 Wall Street, Room 325, Ann Arbor, MI 48105, USA*

Received 22 March 2002

Abstract

We have identified a novel human gene, *c*hromosome *6 o*pen *r*eading *f*rame *37* (*C6orf37*), that is expressed in the retina and maps to human chromosome 6q14, a genomic region that harbors multiple retinal disease loci. The cDNA sequence contains an open reading frame of 1314 bp that encodes a 437-amino acid protein with a predicted molecular mass of 49.2 kDa. Northern blot analysis indicates that this gene is widely expressed, with preferential expression observed in the retina compared to other ocular tissues. The C6orf37 protein shares homology with putative proteins in *R. norvegicus*, *M. musculus*, *D. melanogaster*, and *C. elegans*, suggesting evolutionary conservation of function. Additional sequence analysis predicts that the *C6orf37* gene product is a soluble, globular cytoplasmic protein containing several conserved phosphorylation sites. Furthermore, we have defined the genomic structure of this gene, which will enable its analysis as a candidate gene for chromosome 6q-associated inherited retinal disorders.

Keywords: Chromosome 6q14; Retinal degeneration; Retinal gene; Fovea; Developmental regulation; Conserved gene; Fetal tissue; Brain transcript

Human chromosome 6 harbors a wide array of retinal disease loci (http://www.sph.uth.tmc.edu/RetNet). At least seven different loci for retinal diseases alone have been localized to the long arm of human chromosome 6 in the region 6q14–q16 [1–9]. The gene associated with one of these genetic lesions has recently been identified [10], however, those underlying the remaining retinal diseases mapped to this region have yet to be defined. Within the last decade and a half, a number of retinal disease genes have been identified through a straightforward positional cloning approach [11–16]. Rapid advances in the elucidation and annotation of the human genome sequence have enabled a complementary approach to the cloning of disease genes, involving both classical genetic linkage studies combined with sequence database mining for the identification of putative candidate genes in a genomic region of interest. This has led to the discovery in recent years of several genes implicated in hereditary retinopathies [10,17–20]. Given the high density of retinal disease loci that have been mapped to human chromosome 6q14–q16, our goal is to define potential candidate genes in this interval that can be tested for mutations for each genetic disorder. A feasible approach is to identify retina-expressed genes within the 6q14–q16 region and then to assess them for the presence of mutations in affected individuals in families with retinal diseases mapping to this interval. Toward this end, we describe here the isolation and characterization of a novel human gene located on chromosome 6q14 that is expressed in the retina. This gene encodes a novel protein that appears to be evolutionarily conserved across a variety of species. The putative gene product is likely a soluble, globular, cytosolic factor that is predicted to have several phosphorylation sites that are shared among orthologs. We have named this gene *C6orf37* (*C*hromosome *6 o*pen *r*eading *f*rame *37*) and have tested it as a candidate gene for retinal disorders mapping to 6q14–q16.

* Corresponding author. Fax: +1-734-936-7231.
E-mail address: ayyagari@med.umich.edu (R. Ayyagari).

0006-291X/02/$ - see front matter
PII: S0006-291X(02)00228-0

Materials and methods

Identification and analysis of ESTs localized to 6q14–q16. To identify putative genes present in the genomic interval corresponding to the adMD1/STGD3 critical region [6,7], we searched for ESTs mapping to the genomic region flanked by microsatellite markers D6S1596 and D6S1644 using human genome resources (http://www.ncbi.nlm.nih.gov/). Fine mapping of the ESTs was established by screening for homologous sequences using the human chromosome 6 Blast server (http://www.sanger.ac.uk/HGP/Chr6/Chr6_blast_server.shtml). Map positions of ESTs were confirmed by PCR amplification of YAC clones comprising a physical contig spanning the 6q14 region (unpublished data). A total of 148 non-overlapping ESTs localized to the region of interest was purchased from Research Genetics. Clones corresponding to the ESTs were spotted onto nylon membranes for dot blot analysis. Total RNA was isolated from human peripheral retina and human fovea tissue samples using the Trizol reagent (Life Technologies). Five micrograms of total RNA derived from each tissue was used for cDNA synthesis by reverse transcription using the Life Technologies cDNA Synthesis Kit. (α-^{32}P)dCTP-labeled total cDNA probes were generated by random priming using the multiprime DNA labeling system (Amersham) followed by purification with G-50 Sephadex spin columns (Eppendorf-5 Prime). Hybridization of the dot blots was carried out in Hybrisol II hybridization solution (Intergen) at 65 °C for 16 h using 10^6 cpm/ml of radioactively labeled, purified probes. The membranes were then washed twice for 15 min in 2× SSC at 65 °C, once for 30 min in 2× SSC, 0.1% SDS at 65 °C, and once for 10 min in 0.1× SSC, 0.1% SDS at 65 °C, followed by exposure to Kodak X-Omat film at −70 °C for 24 h.

Isolation of C6orf37 cDNA. A BLAST search [22] for genomic fragments containing the nucleotide sequence for dot blot hybridization-positive EST clone H65844 identified human chromosome 6 genomic PAC clone dJ991C6 (GenBank Accession No. AL078599). Analysis of the dJ991C6 sequence for other ESTs in the region of H65844 yielded 17 ESTs (AW243203, BF025838, AW951915, R74266, AW391612, AW246673, AI022319, AW952387, R78250, AV647023, BG620844, T73543, BE782072, BG622030, BG434626, BG621371, and AI990503) that formed a minimal overlapping contig of 5591 bp, defining the presumed entire *C6orf37* mRNA transcript-encoding region. An open reading frame (ORF) of 1329 bp was found within the predicted *C6orf37* mRNA transcript sequence using the ORF Finder program (http://www.ncbi.nlm.nih.gov/gorf/). Five ESTs that encompassed the putative protein-encoding region (AA551498, R69202, R26344, H65844, and AA564797) were purchased from Research Genetics and subjected to automated sequencing using an ABI Prism 377 Sequencer (Perkin–Elmer). Sequence assembly was carried out using the Sequencer program (Gene Codes Corporation) and ORF detection was performed with the DNASTAR software package. The 5′- and 3′-ends of the *C6orf37* cDNA were defined by identifying transcription initiation [23,24] and transcript cleavage motifs [25,26], respectively, in the PAC sequences flanking the EST contig sequence (see Results).

C6orf37 DNA and protein sequence analysis. Alignment of the *C6orf37* cDNA sequence with the genomic sequence contained in the PAC clone dJ991C6 revealed the location of intron–exon boundaries. PCR primers flanking the proposed exons were designed and used to amplify and sequence the corresponding sequences from human genomic DNA to confirm the predicted *C6orf37* gene structure. The following primers were generated: Exon1F, 5′-AGTCGTCCCGC ACTAACTGC-3′; Exon1R, 5′-AAGCGCAAGCGAGAGTCCCG-3′; Exon2F, 5′-CGCTCCTTTCCCTTTCTTTT-3′; Exon2R, 5′-TTAAC CGATGCCCCTCTG-3′; Exon3F, 5′-TTTCTAGGTTTGCTCATT TTGTTTT-3′; Exon3R, 5′-GCTTCCCCACAGGACATTT-3′.

DNA and protein homology searches for *C6orf37* sequences were performed using the BLASTN and BLASTP search programs at the NCBI, respectively [22]. Additional protein sequence characterization was carried out using bioinformatics tools available through the ExPASy server (http://www.expasy.org/tools/) and the Canadian Bioinformatics Resource (http://www.cbr.nrc.ca/).

Zooblot analysis. A commercially available zooblot containing *Eco*RI-digested genomic DNA from various species (human, monkey, rat, mouse, dog, cow, rabbit, chicken, and yeast) was purchased from Clontech. Hybridizations and washes were performed at 42 °C using the above procedure and a radioactively labeled probe corresponding to the *C6orf37* cDNA sequence. The washed blot was exposed to X-ray film at −70 °C overnight.

Expression analysis by Northern blotting. Northern blots were prepared using 5 μg samples of total RNA isolated from human ocular tissues (retina, retinal pigment epithelium, and choroid). Human fetal and adult multitissue Northern blots containing two micrograms of poly(A)$^+$ RNA derived from various tissues (heart, brain, placenta, lung, liver, skeletal muscle, kidney, pancreas, spleen, thymus, prostate, testis, ovary, small intestine, colon, and peripheral blood leukocyte) were purchased from Clontech. Hybridizations and washes of all blots were performed at 65 °C using radioactively labeled *C6orf37* and human β-actin cDNA probes. Washed blots were exposed to X-ray film at −70 °C for 1–4 days.

In vitro translation. EST clone AA564797 was used as the template for in vitro translation using the TNT Coupled Reticulocyte Lysate System and ^{35}S-methionine (Promega). Reaction products were subjected to SDS–PAGE followed by enhanced autoradiography.

Mutational analysis of C6orf37 in adMD1/STGD3 pedigrees. PCR reactions to amplify the C6orf37 protein-encoding exons were performed using the Exon2F, Exon2R, Exon3F, and Exon3R primers described above and 100 ng of genomic DNA extracted from blood samples obtained from adMD1/STGD3 family members [6,7] as template in 20-μl reactions. PCR cycling parameters were as follows: 1 cycle of denaturation at 94 °C for 3 min; 35 cycles of denaturation at 94 °C for 30 s, primer annealing at 55 °C for 30 s, primer extension at 72 °C for 40 s; 1 cycle of extension at 72 °C for 3 min. The amplification products were sequenced using the PCR primers as well as two internal primers for exons 2 and 3 (Exon2INT, 5′-TGAGGT CCAGGTCCTTGTAG-3′; Exon3INT, 5′-TGAAGGGTCATGAG ATACTC-3′) to detect the presence of sequence alterations in affected individuals from adMD1/STGD3 families. Sequencing reactions were prepared using the Thermo Sequenase Radiolabeled Terminator Cycle Sequencing Kit (USB). Reaction products were electrophoresed on 6% polyacrylamide gels followed by autoradiography at room temperature.

Results

Isolation of human C6orf37 cDNA

To identify genes within the adMD1/STGD3 critical region on chromosome 6q14–q16 that may represent putative candidate genes for the retinal diseases that map to this genomic interval [1–9], we screened 148 non-overlapping ESTs that have also been reported to be localized to this region for retinal expression. Dot blot analysis of EST H65844 with human peripheral retina and human fovea total cDNA probes indicated expression in both retinal tissues (data not shown). Fine mapping of H65844 on a YAC contig of the 6q14–q16 region confirmed its localization between microsatellite markers D6S463 and D6S1455 (Fig. 1A–C). Bioinformatic analysis of the H65844 EST sequence resulted in

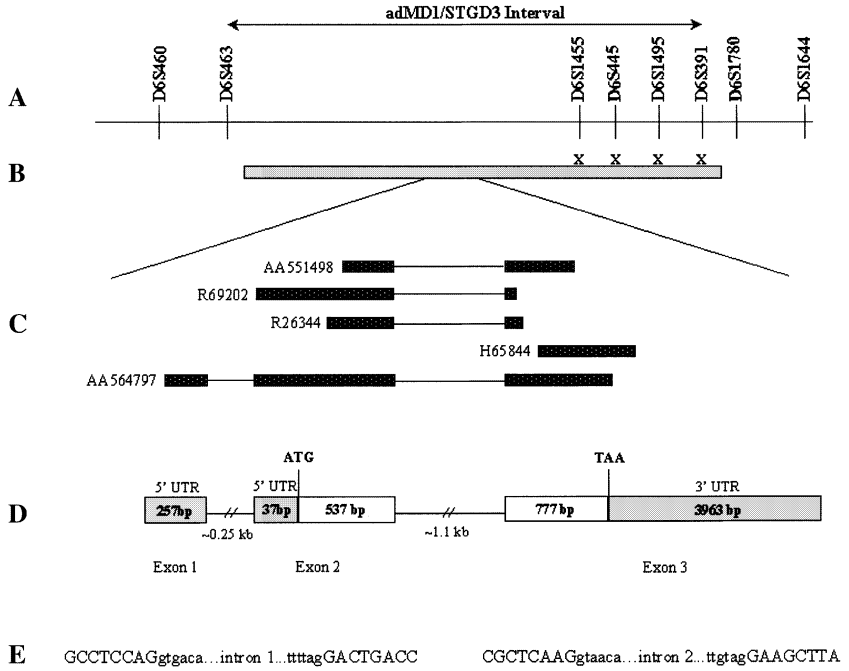

Fig. 1. Cloning strategy used to identify and define the human *C6orf37* gene (see text for details). (A) Representation of the adMD1/STGD3 interval relative to human chromosome 6q microsatellite marker positions. (B) Genomic PAC clone dJ991C6 (GenBank Accession No. AL078599) mapping to the adMD1/STGD3 interval. Xs denote the presence of the corresponding chromosome 6q microsatellite markers shown in Panel A. (C) Human ESTs spanning the *C6orf37* ORF. Stippled black rectangles represent the EST-spanning regions, while intervening horizontal lines indicate the locations of gaps in the alignment of the EST sequences with the genomic PAC sequence. GenBank accession numbers for each of the EST clones are given. (D) Genomic organization of the *C6orf37* gene. Rectangles represent exons and intervening horizontal lines represent introns. Open rectangles correspond to the ORF, while shading denotes untranslated regions (UTR). The lengths of exon sequences are given within the rectangles and approximate sizes of the introns are indicated below the intervening horizontal lines.The positions of the start (ATG) and stop (TAA) codons of the ORF are shown. (E) *C6orf37* exon–intron boundary sequences. Nucleotides comprising intron-flanking exon sequences are given in upper case, while terminal portions of the intron sequences are given in lower case.

the identification of numerous ESTs that spanned a 5591-bp segment contained within the chromosome 6q genomic PAC clone dJ991C6. The sequence of 17 of these ESTs (AW243203, BF025838, AW951915, R742 66, AW391612, AW246673, AI022319, AW952387, R78 250, AV647023, BG620844, T73543, BE782072, BG622 030, BG434626, BG621371, and AI990503) could be assembled to form a minimal contig comprising the entire 5591-bp expressed sequence. Analysis of the transcript sequence revealed the presence of a 1329-bp ORF. DNA sequencing and sequence assembly of five ESTs (AA551498, R69202, R26344, H65844, and AA564797) that encompassed the predicted protein-encoding region of the newly identified gene, which we named *C6orf37*, were subsequently performed. Minor discrepancies were detected between the contig assembled by sequencing the five EST clones and the mRNA sequence predicted from the reported PAC clone sequence, including the absence of a 15-nucleotide (and therefore 5-amino acid) repeat and four occurrences of G ↔ C switched nucleotide assignments resulting in two amino acid changes. However, the reading frame was not altered and no significant changes to the overall protein sequence were noted. Additional changes found correspond to exon 1 sequence alterations that do not affect the translation product, such as the absence of three individual nucleotides and a T ↔ C reversal within the 5′-UTR. The sequence discrepancies detected may represent polymorphisms between different tissue sources or may simply reflect the existence of multiple sequence variants. Our results thus indicate a 1314-bp full-length ORF contained within the deduced 5571-bp cDNA sequence for *C6orf37* (Fig. 2).

In addition to the constituent EST sequence boundaries, the 5′-end of the complete cDNA was defined by identifying a canonical TATA box promoter element [23] at nucleotide positions −21 to −15 and a eukaryotic transcription inititator (Inr) sequence [24] in the genomic PAC clone just upstream of the putative *C6or37* transcription start site (+1 position). The 3′-end of the cDNA was delineated by the presence of a poly(A)$^+$ signal at position 5546–5551 just upstream of a consensus mRNA cleavage site [26] in the genomic sequence. Analysis of the complete cDNA sequence with gene prediction software located the ATG start codon of the 1314-bp continuous ORF at position 295 and the TAA stop codon at position 1606. The sequence surrounding the proposed initiation codon

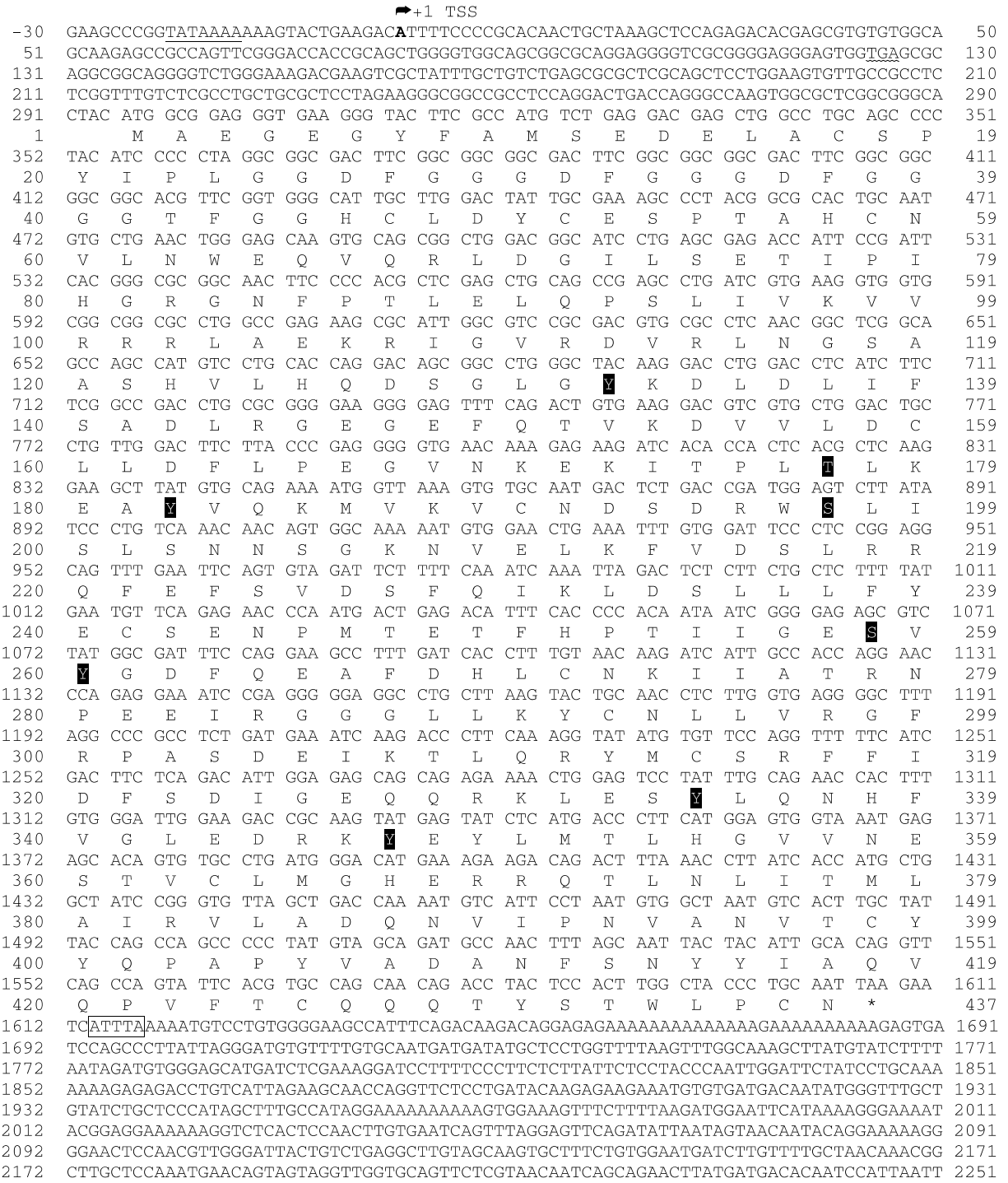

Fig. 2. Nucleotide and deduced translation product sequences of human *C6orf37* cDNA. The sequences are numbered starting at the predicted transcription start site (TSS; boldface nucleotide) and initiation codon of the gene and protein, respectively. A putative TATA box in the proximal promoter region is underlined with a solid line. The terminal stop codon of the amino acid sequence is indicated by an asterisk. The in-frame stop codon in the 5′-UTR is underlined with a wavy line. Conserved potential phosphorylation sites are highlighted in black. RNA destabilization elements in the 3′-UTR are outlined with boxes. A consensus polyadenylation signal is underlined with a double solid line. The sequence of human *C6orf37* has been deposited in GenBank under Accession No. AF350451.

represents a suitable context for translation initiation in vertebrate mRNAs [27]. An in-frame TGA stop codon was found at position 124, 171 nucleotides upstream of the predicted start codon, suggesting that the 1314-bp sequence does in fact represent a complete ORF.

```
2252  CCAGCTGCGTGCATAGATCACATTTTTAAAATGTAAAAATGCAAGCAAAAACAGCTGTAACAAAGAAAGTGTGCTCAAGG 2331
2332  ACCAAAGATTTAACAGATAAAAATACCCAATTAGAAGAGATATAGTAGACTATATGAAGAGAGATTATATTTGTTACACA 2411
2412  CCAATATACATCAAAGTGCCTGTTGCCTTCTGAAAATTTGAAGTGGCAAAATTATTTTATGGTTTAATGATTATTTTATT 2491
2492  TTATCAGGGACTGCCTCAAGAAGAAAATAACATAAGCTTGTGAATGGTGGAGAAAATGCCCTATTTTTTCTTGCAAATAC 2571
2572  TTGTATAAAGTTAACATTTGTTGATCTGATATTATCATAGGTACATGTGTATGTGTGTATAAATTATATGTGTGTGTGTA 2651
2652  TATATACATTTTATATATACATTTTATATGTATATATACACAGTAGATTGACTATGATCTAGAATAATGTCTCAAATAGG 2731
2732  AAATGTTTAAATACTGTGTGTTTTTATGTTTTCAACAGGATAACATGAGACGTGGGCATATTGCAATGATGAATTAAATC 2811
2812  CACATCTAAAAAAATTAAATGAAGGAGGGAACCAAGTAATATATTTCATAGGAAGAGCAGAAATTATACTGTTTTAGTGG 2891
2892  GATTTTTTTTTCTTTTTTTTTTTTCTTTGGTGAGCCATAAAATTCCACAAATGGGAGAATATTTGTTTGGCAGAGCACT 2971
2972  CTTTTTTATATTGAACTGCCATTTTGACAGTTGGAACCCATTTATTAAAAAAAAAATTGCATTCCTCTATGATGTTTAAT 3051
3052  CTAGTGGATCATGGATCAGTAATAGGCTACTTAAATCCCTGACTGCTAAAAAGGATTTCCGGTGATCTAAACACTACTTG 3131
3132  CTAATGTTTAAATGAATTTTAATGAATGCATTCTGCATTTCTGGACCACTAGAATTTAGTAATGTGAAATGACCCTTTTT 3211
3212  ACAGAATATTTGCACAATTGCTTAAAATTTATATATGAGATATATATTATATATAACATTTTATAAATCATGTCAATATG 3291
3292  AAACATCTTTGATCTGGTTGTCACACTGCATTTAAATATTTAGTACTGTACTTTAAATCGCTTTCCATTAAATCAAATCC 3371
3372  AACTTTATTTTCTTTCTTACAAAAATACCAGTTATACCTTTGTGAAATGAACTGGCATTACTATTTCAGTTCAATAACAG 3451
3452  CTAATCCTAAAACCACCCTTTCTCCTAGCCAGTAGTTCCTCTAGATACTGGTCTCTGAAAATGCATTTGTTAAAAACAAA 3531
3532  ACAAAACTAACACATAAGAACCTTCCCTTTGTGTTGTGAAACAACCACATAATCTCCACAACCTTAGTGGATGACTGCTT 3611
3612  GCTATGATAATTCCTCGAAGACCCAATTAGAAGATTTTCATCATCAGTTAAAGAGAGACCACGGGAGAAAAAAATATCCT 3691
3692  CCTGTTGGCAGTATAATTTGTTTGTTTGTTTATCTAGGGATCCTCAGATGCTTAGTGCTAGGTTAATCCAGGTTAATCCG 3771
3772  TCTGGACTACCTTTTGTGCATCTTTCTTTGAAGCCTTAATGGGAACCTGATGGGTTTGCTGTAGCAGCTTCCTTGTGAAT 3851
3852  TCTGTCAGAGCTGCAACAGCCGCTGCACTGCCACTCAGTTTTCTAAGGAACTCCTCCTACTACCATCTTGGCTCAGTCTC 3931
3932  CCTCACTTAAGCCCTGGGTTTGAAAAATTAATTGCAACTTCCCAGGAAACATTGTTCAGTTTGCAGATTAAGCCTGGCAC 4011
4012  TCACCTATCAGAAACCAGAGCTCCGCCTGCTTAGTTGTTTCAAAGTTTTCTGAAAGAAAACTAGGGGAGCACTTGTGAAC 4091
4092  ACAGGAGCAGCTGGTGATCTGCTTTCTTACCCTAACTCTTGACAAATGAGTCGTCTACTATTTTAAAGAGTCTGGAGGTC 4171
4172  TCTGACTCTGCCATAACAATAACCTGCTGTTAATTTATAACACAGATTTTTGTTTGGAAGAGCCTTATTTGAAATACACT 4251
4252  TTGATTTATTTTTCTTAAATATTTATATTCTTTTCTTGCTTACTTCAGGGTTGGTAGCTTAGTTGGAAGTGCCAGCACCTG 4331
4332  GCACCTATTCATATAGAACAGGCTGTACTCAAGACAACTTCTAGCATTTACTTTAAGACTTATATAATTTATTTCTATTT 4411
4412  TGTGTGTACTATAGTCTTGTGCATATGTAGTTGAACACACAGTGAAATATATGTCTCTCTTTGTGGATGTGCGGCCTAAA 4491
4492  AATTTGAATGTCTGGTGAGAGAGAGCCATGTGTATAGGTCAGAGAAAAGAACAGCTCCCGACTCCCTATTAGCGCCTGTG 4571
4572  ATTTGTTTCCTTTTGTGTTTATCTGGCCTAGTGTGCTGTTTCTTTAAACCAGGAAGAAGTTTTGTCTTTTGGAGGCTCTT 4651
4652  CTCACCTGTCCAGCCTGGCATGTCAGAGAACACATAGCCTGTGACAATGCCGTTTTTAAAGGTTTACTTAATTTGCAGTA 4731
4732  AATCCAGCTGCCTCAAGAACTCCTACACCAAGATGGACATTTCCTTTCCAGAAATGGGATCAAGTATCTGCTCACTTTGG 4811
4812  TATTGGATGGACTAATAATGTAGCTCCAAAAATGCAAGGATGGAAGAATATGTGTAATCCAAACCAAGGAAGGAAATGAA 4891
4892  AAGTGAACGTACTGTTTTTTACCACCCCTTTCTGTTTGCTTATTGTTGGTTGCTTCACTGTGCATAAAGTTGTTTTCAATG 4971
4972  CAACGCTTGTTAAATAAATATTGTGAACTATTTTGTAAATGAAATGTATTATGTTGAAAGCTGTCAGTTCAAAAATAAGC 5051
5052  TTTTTTGTTGTTGTTGAAGATGAAGTGTGTTAGGTGAAACCAAAAAGCCAAAAAAAGTAATTTCATATATAGCATCTATT 5131
5132  TGAATATAATCTTTCTTTAAAATTTCTTTTAGCATAGCATTTTCAGTGCTAAGAAAGAATCTCTATGTTATATTTTGTTA 5211
5212  AAATAATGGCTTTCTAACAAAGCAAATGGTAAAGTACAAAGTTGGAAGATGTCAAGTTAACGAGACTTGCTGCAAAGCCT 5291
5292  TGCAGAACGGAGGAGGCTCGCCTGCTGGCTGTCTCTCCCTCCAACCTCTCTACAATCATGCCTGCTTTGAGGTGTTCTGT 5371
5372  TGCAGCAAGCTGCACCTTGGGTCACTCTTTTGGAATATTTTGACTATAGGCTGCGTCACAGGCAGAAAAGGAGTTGATGG 5451
5452  AAAATGGACTAAAAAACTGACATGTTTGAATCAGTGCTAGAGGGAACAGATTGTGAATTTTGTTTACAGCATCCAATATT 5531
5532  TGGATTTTTTTGTAAATAAAAAAGTTATTTTTTTCTATTG                                          5571
```

Fig. 2. *(Continued)*

Determination of C6orf37 genomic structure

Alignment of the *C6orf37* expressed sequence with the genomic PAC clone dJ991C6 sequence reveals that the gene is organized into three exons. Exon 1 comprises the first 257 nucleotides of the 5′-UTR. Exon 2 contains 574 bp, consisting of the terminal 37 bp of the 5′-UTR followed by the initial 537 bp of the protein-encoding region of the gene. Exon 3 contains 4740 bp, 777 bp of which completes the coding region and the remaining 3963 bp corresponds to the 3′-UTR (Fig. 1D). The exons are separated by two intervening intron sequences that are 252 bp and 1118 in length, respectively. All splice junctions conform to consensus sequences for splice donor and acceptor sites [29]. The genomic organization of the *C6orf37* gene was confirmed by using primers that flank the proposed exons to amplify the putative gene regions from human genomic DNA. The resulting amplicons were sequenced and found to match the deduced *C6orf37* gene sequence.

Sequence conservation of C6orf37

Searches of non-redundant DNA and protein databases with the BLAST algorithms did not identify any significant homologies between *C6orf37* and previously characterized human genes or proteins. Similarities of the proposed *C6orf37* translation product to putative proteins in mouse (72%), *D. melanogaster* (60%), and *C. elegans* (40%) were identified (Fig. 3), however, the function of these proteins is as of yet unknown. Sequence assembly and translation of six overlapping *R. norvegicus* ESTs (AA998465, AA926118, AA933175, AW502512, BF558308, and BI294811) displaying significant similarity to the *C6orf37* cDNA sequence also revealed the existence of a rat homolog that shares 91% identity with the human protein (Fig. 3). In addition,

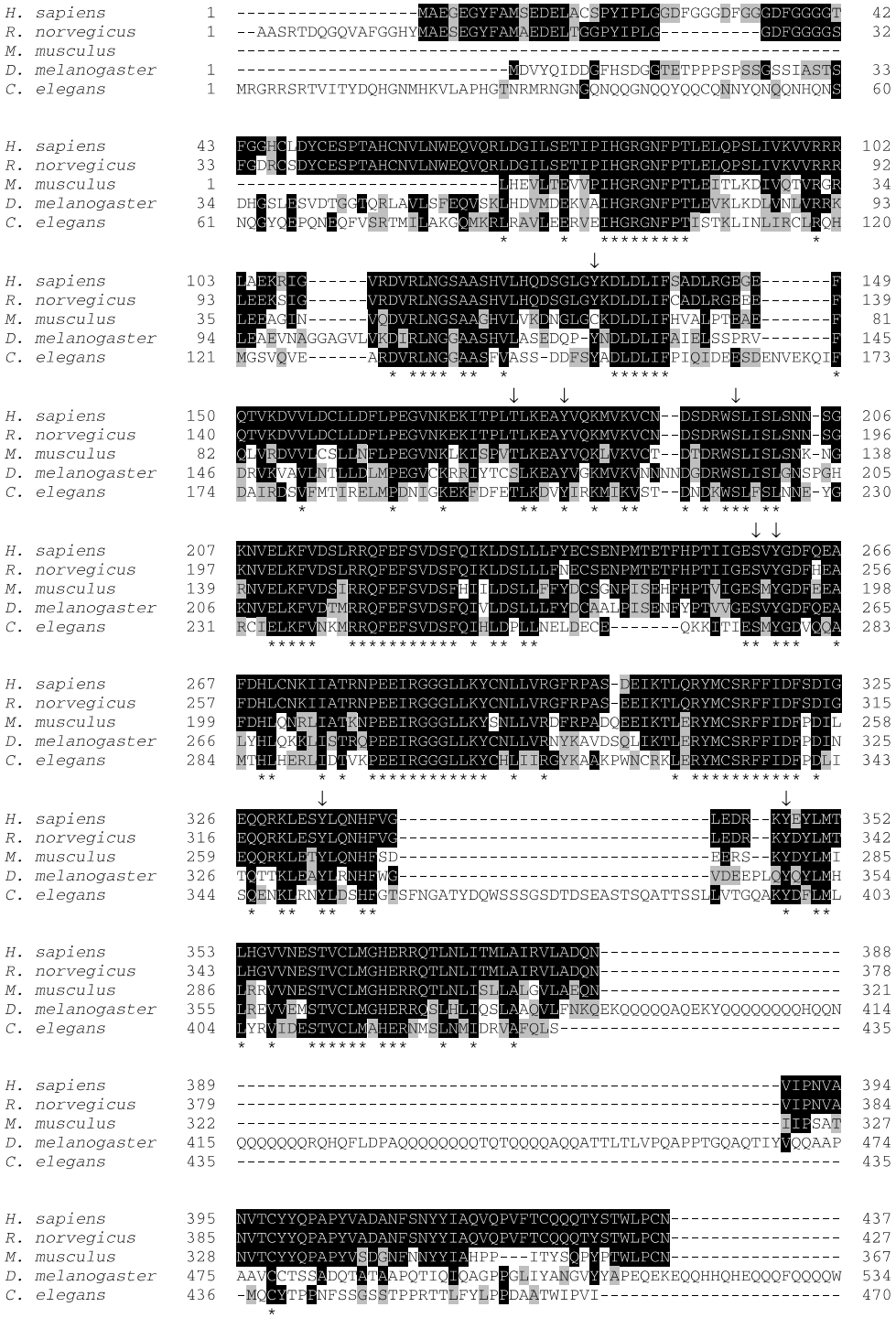

Fig. 3. Sequence alignment of the human C6orf37 protein and homologs identified in other species. The amino acid sequences of the *H. sapiens* (GenBank Accession No. AF350451), *R. norvegicus* (deduced from translation of overlapping ESTs), *M. musculus* (Accession No. BAB29768), *D. melanogaster* (Accession No. AAF59198), and *C. elegans* (Accession No. T15196) homologs were aligned using the Clustal W program (version 1.81). Hyphens represent gaps introduced to maintain the alignment. The amino acid residues are numbered to the left and right of the sequences. Identical residues among the sequences in the alignment are shown on a black background and conservative amino acid substitutions are shaded in gray. Residues identical in all five homologs are marked with asterisks. The conserved potential phosphorylation sites predicted are indicated by arrows.

zooblot analysis showed hybridization of *C6orf37* cDNA probes to genomic DNA extracted from a number of other species including chicken, rabbit, cow, and the yeast *Saccharomyces cerevisiae* (data not shown), further suggesting the evolutionary conservation of this gene in both mammalian and non-mammalian species.

```
H. sapiens              ------------------------------------------------------------
R. norvegicus           ------------------------------------------------------------
M. musculus             ------------------------------------------------------------
D. melanogaster   535   SEDALGATLLAETETEPPLATASIAPPPPSSQPPSSPSTTAPRTQVQNQTRSTTINSSNR  594
C. elegans              ------------------------------------------------------------

H. sapiens              ------------------------------------------------------------
R. norvegicus           ------------------------------------------------------------
M. musculus             ------------------------------------------------------------
D. melanogaster   595   KKKNKAKTQSKQQTQTATEAESNCNEKPTTATMTRTTTKKSSRIKQMAAAAAATEACAAK  654
C. elegans              ------------------------------------------------------------

H. sapiens              ------------------------------------------------------------
R. norvegicus           ------------------------------------------------------------
M. musculus             ------------------------------------------------------------
D. melanogaster   655   GGAGGITGSLVLTPQQINGLPVYNQLDDGGRSAAGTYSGQFAYFYPHDGSAPSMFLLGED  714
C. elegans              ------------------------------------------------------------

H. sapiens              ------------------------
R. norvegicus           ------------------------
M. musculus             ------------------------
D. melanogaster   715   SYLNVTEQQEQHQHQHQHHNHQFV  738
C. elegans              ------------------------
```

Fig. 3. *(Continued)*

C6orf37 protein and mRNA sequence analysis

The 1314-bp ORF encodes a protein of 437 amino acids with an apparent molecular weight of 49.2 kDa and an isoelectric point of 5.01. In vitro translation using the EST clone AA564797, which was found to encompass the entire *C6orf37* coding sequence, generated a protein product of approximately 49 kDa as predicted (Fig. 4). Analysis of the C6orf37 amino acid sequence with the NetPhos 2.0 program [30] predicts the presence of several potential phosphorylation sites that are found to be conserved among the different orthologs (Fig. 3). Protein structure prediction using the PredictProtein software [31] reveals the absence of any putative hydrophobic transmembrane segments and predicts a globular structure, suggesting that the C6orf37 protein is unlikely to be membrane bound. Subcellular localization of the protein predicted by the PSORTII program [32] indicates that it is likely to be found in the cytoplasm due to the lack of any known intracellular organelle targeting signals. Searches of the C6orf37 amino acid sequence against a variety of protein motif databases using the ProfileScan, PROSITE [33], PRINTScan [34], BLOCKS [35], pFAM [36], and ProDom [37] programs failed to find significant homology to any known protein domains.

Additional examination of the *C6orf37* cDNA sequence reveals the presence of several ATTTA motifs in the lengthy 3′-UTR (Fig. 2). These AU-rich elements (AREs) are believed to play a role in determining RNA stability in mammalian cells [28], suggesting that the *C6orf37* transcript may undergo rapid and selective decay.

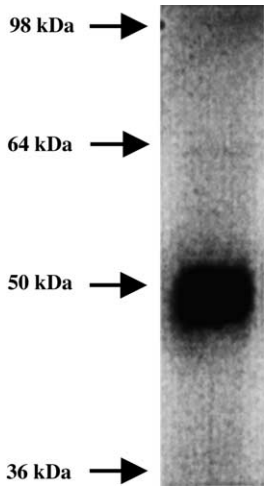

Fig. 4. In vitro translation product of *C6orf37* coding sequence. Positions of protein size markers are shown on the left.

Tissue distribution of C6orf37 mRNA

To determine the expression pattern of *C6orf37*, we performed Northern blot analysis of human fetal and adult tissues using the EST H65844 sequence as a probe (Fig. 5). Two transcript sizes were detected, the major one consistent with the proposed full-length 5571-bp *C6orf37* cDNA species, and a second, less abundant and slightly larger transcript possibly arising from alternative splicing. The predominant, approximately 5.6 kb message is expressed in most tissues tested. It is weakly detectable in adult skeletal muscle, thymus, liver, lung, and brain. Interestingly, the *C6orf37* transcripts are prominently expressed in fetal liver, lung, and brain. In addition we observed the *C6orf37* 5.6 kb transcript in adult heart but not in fetal heart tissue. These findings suggest that the *C6orf37* gene may be developmentally regulated in some tissues.

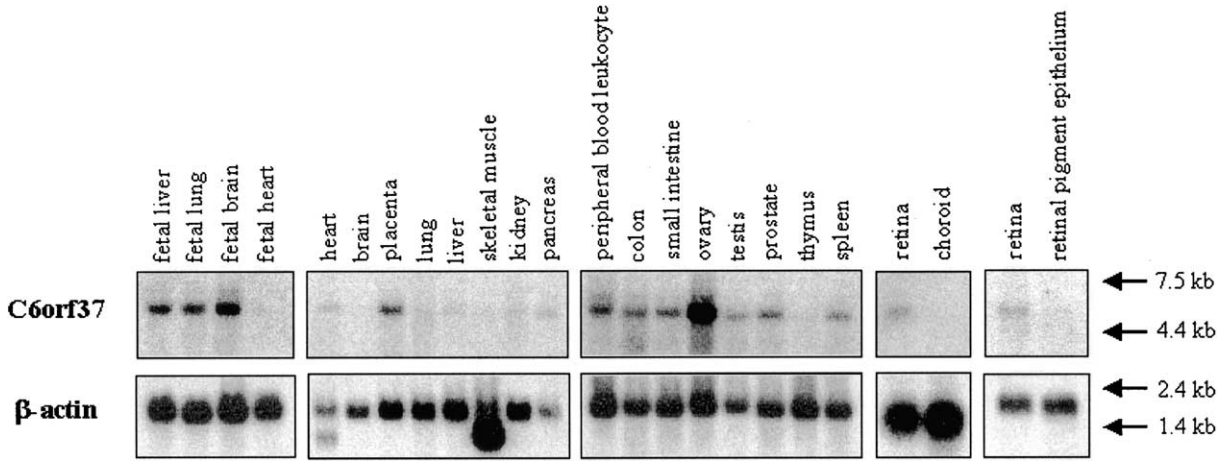

Fig. 5. Northern blots of human RNA samples probed with *C6orf37* and control (β-actin) cDNA probes. Retina, choroid, and retinal pigment epithelium lanes contain 5 μg of total RNA while all other lanes contain 2 μg of poly(A)$^+$ RNA. Positions of RNA size markers are shown on the right.

Within the immediate region of the retina, we detect the major it C6orf37 transcript in the neural retina but neither in the adjacent retinal pigment epithelium nor in the choroid (Fig. 5). The initial dot blot analysis performed for two non-overlapping ESTs corresponding to the *C6orf37* cDNA sequence (H65844 and H92739) indicates that the gene is expressed in both the foveal and peripheral regions of the adult human retina (data not shown).

Exclusion of C6orf37 as the adMD1/STGD3 causative gene

To test whether *C6orf37* represents the adMD1/STGD3 disease gene, we screened the protein-encoding exons for the presence of mutations among affected individuals of families diagnosed with autosomal dominant atrophic macular degeneration/Stargardt-like macular dystrophy in which the putative disease gene(s) has (have) been mapped to the 6q14–q16 region [6,7]. We did not find any sequence alterations in affected versus unaffected individuals (data not shown) and thus excluded *C6orf37* as the gene that underlies the condition seen in these families.

Discussion

Through a procedure of genomic and EST sequence database mining coupled with experimental analysis for tissue-selective expression status, we have described here a strategy for the identification of candidate genes for inherited diseases mapping to a genomic interval of interest. By this method, we identified a novel gene that maps to the human chromosome 6q14–q16 region, a genomic interval harboring multiple retinal disease loci. In this study we report the isolation and characterization of this gene, which we named *C6orf37*.

The *C6orf37* gene is composed of three exons and two introns, spanning approximately 7 kb of genomic DNA. A transcript length of 5571 nucleotides is predicted for the functional *C6orf37* message, encoding a putative protein of 437 amino acids. The gene product appears to share significant homology with predicted proteins in *R. norvegicus*, *M. musculus*, *C. elegans*, and *D. melanogaster*, however, the biological role of *C6orf37* or any of its orthologs is at present unknown. No homology of *C6orf37* with any protein of known function was detected and bioinformatic analysis of the amino acid sequence failed to reveal the presence of any previously characterized functional motifs. However, the extent of sequence conservation observed across a wide range of species suggests that *C6orf37* and its homologous counterparts may represent the first-identified members of a novel protein family with an important cellular role.

Clues to the nature of this role were gained from additional sequence analysis of the *C6orf37* gene. A number of conserved potential phosphorylation sites were detected in the primary structure of the C6orf37 protein, which may be involved in post-translational activation or regulation of its function. The lack of any transmembrane segments or subcellular organelle targeting signals and the prediction of its cytoplasmic localization are consistent with its possible role as a soluble intracellular signaling molecule that may serve as a target for the action of kinase and/or phosphatase enzymes. Control of *C6orf37* function may also be imparted at the level of mRNA turnover, as suggested by the presence of multiple ATTTA sequence motifs in the 3′-UTR that are known to act as RNA destabilizing elements [28]. Multiple *C6orf37* regulatory mechanisms

including potential protein phosphorylation or dephosphorylation and selective transcript degradation may indicate the importance of rapid signal transduction mediated through the *C6orf37* gene product in a particular functional pathway.

Northern blotting of RNA from human ocular tissues confirmed the retinal expression initially detected for *C6orf37* constituent ESTs by dot blot analysis and further demonstrated preferential expression of the gene within the neural retina over other eye tissues such as the retinal pigment epithelium and the choroid. However, retina-specific expression is not a requisite feature of retinal disease genes, as has been observed for a number of ubiquitously expressed genes that underlie various forms of retinal degeneration [13–16,38–40], and therefore should not be considered a strict requirement of potential candidates. Additional analysis of human non-ocular tissues suggests that the *C6orf37* gene is also expressed in a wide range of tissue types, consistent with the broad spectrum of tissue sources noted for derivative ESTs. Furthermore, a developmental role is suggested by the differential expression observed between the adult and fetal tissues examined. This may indicate that there are different functions of the C6orf37 protein in different tissues at various stages of development. Studies are underway to extend this finding to human ocular tissues as well as various tissues from non-human species. The Northern blots also reveal the presence of at least two *C6orf37* transcripts in several tissue types. The larger, less abundant transcript may be the product of a rare splicing event and raises the possibility that alternative splicing may be another means by which *C6orf37* gene expression is regulated.

Satisfying the criteria of mapping to the chromosome 6q14–q16 interval and clearly displaying retinal expression, we screened the *C6orf37* candidate gene for mutations in affected individuals of families with autosomal dominant atrophic macular degeneration (adMD1)/Stargardt-like macular dystrophy (STGD3) that we have previously described and for which we have mapped the disease loci to 6q14–q16 [6,7]. No mutations were detected within the protein-encoding exons of *C6orf37*, leading to its exclusion as the gene associated with the disease in these families. As it turns out, a 5-bp deletion in another gene that we subsequently localized to the same genomic region, *ELOVL4*, was found to define the genetic lesion in our families [10]. Given the mapping and expression status of *C6orf37* and *ELOVL4*, they should both be tested as putative candidates for the other retinal dystrophies that are known to localize within the 6q14–q16 region. Identification and characterization of additional retina-expressed genes located in this chromosomal interval will be useful in elucidating the myriad of possible mutations that give rise to the retinal disorders that map to this unique region of the human genome.

Acknowledgments

We thank Paul A. Sieving and Ian M. MacDonald for providing the patients with adMD1/STGD3. We also thank the patients and their families for their participation in these studies. The work was supported by The Foundation Fighting Blindness USA, NEI Vision Core Grant EY0-7003, RPB Harrington Scholar Award (RA), Foundation Fighting Blindness-Canada (PWW), E.A. Baker Foundation of the Canadian National Institute for the Blind (PWW, PSL), and Canadian Institutes of Health Research Doctoral Research Award (PSL).

References

[1] K.W. Small, J.L. Weber, A. Roses, F. Lennon, J.M. Vance, M.A. Pericak-Vance, North Carolina macular dystrophy is assigned to chromosome 6, Genomics 13 (1992) 681–685.
[2] E.M. Stone, B.E. Nichols, A.E. Kimura, T.A. Weingeist, A. Drack, V.C. Sheffield, Clinical features of a Stargardt-like dominant progressive macular dystrophy with genetic linkage to chromosome 6q, Arch. Ophthalmol. 112 (1994) 765–772.
[3] R.E. Kelsell, B.F. Godley, K. Evans, P.A. Tiffin, C.Y. Gregory, C. Plant, A.T. Moore, A.C. Bird, D.M. Hunt, Localization of the gene for progressive bifocal chorioretinal atrophy (PBCRA) to chromosome 6q, Hum. Mol. Genet. 4 (1995) 1653–1656.
[4] R.E. Kelsell, K. Gregory-Evans, C.Y. Gregory-Evans, G.E. Holder, M.R. Jay, B.H. Weber, A.T. Moore, A.C. Bird, D.M. Hunt, Localization of a gene (CORD7) for a dominant cone–rod dystrophy to chromosome 6q, Am. J. Hum. Genet. 63 (1998) 274–279.
[5] A. Ruiz, S. Borrego, I. Marcos, G. Antinolo, A major locus for autosomal recessive retinitis pigmentosa on 6q, determined by homozygosity mapping of chromosomal regions that contain gamma-aminobutyric acid-receptor clusters, Am. J. Hum. Genet. 62 (1998) 1452–1459.
[6] I.B. Griesinger, P.A. Sieving, R. Ayyagari, Autosomal dominant macular atrophy at 6q14 excludes CORD7 and MCDR1/PBCRA loci, Invest. Ophthalmol. Vis. Sci. 41 (2000) 248–255.
[7] P.S. Lagali, I.M. MacDonald, I.B. Griesinger, M.L. Chambers, R. Ayyagari, P.W. Wong, Autosomal dominant Stargardt-like macular dystrophy segregating in a large Canadian family, Can. J. Ophthalmol. 35 (2000) 315–324.
[8] S. Dharmaraj, Y. Li, J.M. Robitaille, E. Silva, D. Zhu, T.N. Mitchell, L.P. Maltby, A.B. Baffoe-Bonnie, I.H. Maumenee, A novel locus for Leber congenital amaurosis maps to chromosome 6q, Am. J. Hum. Genet. 66 (2000) 319–326.
[9] M. Kniazeva, E.I. Traboulsi, Z. Yu, S.T. Stefko, M.B. Gorin, Y.Y. Shugart, J.R. O'Connell, C.J. Blaschak, G. Cutting, M. Han, K. Zhang, A new locus for dominant drusen and macular degeneration maps to chromosome 6q14, Am. J. Ophthalmol. 130 (2000) 197–202.
[10] K. Zhang, M. Kniazeva, M. Han, W. Li, Z. Yu, Z. Yang, Y. Li, M.L. Metzker, R. Allikmets, D.J. Zack, L.E. Kakuk, P.S. Lagali, P.W. Wong, I.M. MacDonald, P.A. Sieving, D.J. Figueroa, C.P. Austin, R.J. Gould, R. Ayyagari, K. Petrukhin, A 5-bp deletion in ELOVL4 is associated with two related forms of autosomal dominant macular dystrophy, Nat. Genet. 27 (2001) 89–93.
[11] W.H. Lee, R. Bookstein, F. Hong, L.J. Young, J.Y. Shew, E.Y. Lee, Human retinoblastoma susceptibility gene: cloning, identification, and sequence, Science 235 (1987) 1394–1399.
[12] F.P. Cremers, D.J. van de Pol, L.P. van Kerkhoff, B. Wieringa, H.H. Ropers, Cloning of a gene that is rearranged in patients with choroideraemia, Nature 347 (1990) 674–677.
[13] A. Meindl, K. Dry, K. Herrmann, F. Manson, A. Ciccodicola, A. Edgar, M.R. Carvalho, H. Achatz, H. Hellebrand, A. Lennon, C.

Migliaccio, K. Porter, E. Zrenner, A. Bird, M. Jay, B. Lorenz, B. Wittwer, M. D'Urso, T. Meitinger, A. Wright, A gene (RPGR) with homology to the RCC1 guanine nucleotide exchange factor is mutated in X-linked retinitis pigmentosa RP3, Nat. Genet. 13 (1996) 35–42.

[14] R. Roepman, G. van Duijnhoven, T. Rosenberg, A.J. Pinckers, L.M. Bleeker-Wagemakers, A.A. Bergen, J. Post, A. Beck, R. Reinhardt, H.H. Ropers, F.P. Cremers, W. Berger, Positional cloning of the gene for X-linked retinitis pigmentosa 3: homology with the guanine-nucleotide-exchange factor RCC1, Hum. Mol. Genet. 5 (1996) 1035–1041.

[15] U. Schwahn, S. Lenzner, J. Dong, S. Feil, B. Hinzmann, G. van Duijnhoven, R. Kirschner, M. Hemberger, A.A. Bergen, T. Rosenberg, A.J. Pinckers, R. Fundele, A. Rosenthal, F.P. Cremers, H.H. Ropers, W. Berger, Positional cloning of the gene for X-linked retinitis pigmentosa 2, Nat. Genet. 19 (1998) 327–332.

[16] B.H. Weber, G. Vogt, R.C. Pruett, H. Stohr, U. Felbor, Mutations in the tissue inhibitor of metalloproteinases-3 (TIMP3) in patients with Sorsby's fundus dystrophy, Nat. Genet. 8 (1994) 352–356.

[17] C.G. Sauer, A. Gehrig, R. Warneke-Wittstock, A. Marquardt, C.C. Ewing, A. Gibson, B. Lorenz, B. Jurklies, B.H. Weber, Positional cloning of the gene associated with X-linked juvenile retinoschisis, Nat. Genet. 17 (1997) 164–170.

[18] K. Petrukhin, M.J. Koisti, B. Bakall, W. Li, G. Xie, T. Marknell, O. Sandgren, K. Forsman, G. Holmgren, S. Andreasson, M. Vujic, A.A. Bergen, V. McGarty-Dugan, D. Figueroa, C.P. Austin, M.L. Metzker, C.T. Caskey, C. Wadelius, Identification of the gene responsible for Best macular dystrophy, Nat. Genet. 19 (1998) 241–247.

[19] M.M. Sohocki, S.J. Bowne, L.S. Sullivan, S. Blackshaw, C.L. Cepko, A.M. Payne, S.S. Bhattacharya, S. Khaliq, S. Qasim Mehdi, D.G. Birch, W.R. Harrison, F.F. Elder, J.R. Heckenlively, S.P. Daiger, Mutations in a new photoreceptor-pineal gene on 17p cause Leber congenital amaurosis, Nat. Genet. 24 (2000) 79–83.

[20] D.Y. Nishimura, C.C. Searby, R. Carmi, K. Elbedour, L. Van Maldergem, A.B. Fulton, B.L. Lam, B.R. Powell, R.E. Swiderski, K.E. Bugge, N.B. Haider, A.E. Kwitek-Black, L. Ying, D.M. Duhl, S.W. Gorman, E. Heon, A. Iannaccone, D. Bonneau, L.G. Biesecker, S.G. Jacobson, E.M. Stone, V.C. Sheffield, Positional cloning of a novel gene on chromosome 16q causing Bardet–Biedl syndrome (BBS2), Hum. Mol. Genet. 10 (2001) 865–874.

[21] P. Deloukas, G.D. Schuler, G. Gyapay, E.M. Beasley, C. Soderlund, P. Rodriguez-Tome, L. Hui, T.C. Matise, K.B. McKusick, J.S. Beckmann, S. Bentolila, M. Bihoreau, B.B. Birren, J. Browne, A. Butler, A.B. Castle, N. Chiannilkulchai, C. Clee, P.J. Day, A. Dehejia, T. Dibling, N. Drouot, S. Duprat, C. Fizames, D.R. Bentley, et al., A physical map of 30,000 human genes, Science 282 (1998) 744–746.

[22] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.

[23] S.L. McKnight, R. Kingsbury, Transcriptional control signals of a eukaryotic protein-coding gene, Science 217 (1982) 316–324.

[24] R. Javahery, A. Khachi, K. Lo, B. Zenzie-Gregory, S.T. Smale, DNA sequence requirements for transcriptional initiator activity in mammalian cells, Mol. Cell. Biol. 14 (1994) 116–127.

[25] N. Proudfoot, Poly(A) signals, Cell 64 (1991) 671–674.

[26] F. Chen, C.C. MacDonald, J. Wilusz, Cleavage site determinants in the mammalian polyadenylation signal, Nucleic Acids Res. 23 (1995) 2614–2620.

[27] M. Kozak, Interpreting cDNA sequences: some insights from studies on translation, Mamm. Genome 7 (1996) 563–574.

[28] N. Xu, C.Y. Chen, A.B. Shyu, Modulation of the fate of cytoplasmic mRNA by AU-rich elements: key sequence features controlling mRNA deadenylation and decay, Mol. Cell. Biol. 17 (1997) 4611–4621.

[29] R.A. Padgett, P.J. Grabowski, M.M. Konarska, S. Seiler, P.A. Sharp, Splicing of messenger RNA precursors, Annu. Rev. Biochem. 55 (1986) 1119–1150.

[30] N. Blom, S. Gammeltoft, S. Brunak, Sequence and structure-based prediction of eukaryotic protein phosphorylation sites, J. Mol. Biol. 294 (1999) 1351–1362.

[31] B. Rost, PHD: predicting one dimensional protein structure by profile-based neural networks, Methods Enzymol. 266 (1996) 525–539.

[32] K. Nakai, M. Kanehisa, A knowledge base for predicting protein localization sites in eukaryotic cells, Genomics 14 (1992) 897–911.

[33] K. Hofmann, P. Bucher, L. Falquet, A. Bairoch, The PROSITE database, its status in 1999, Nucleic Acids Res. 27 (1999) 215–219.

[34] T.K. Attwood, M.D. Croning, D.R. Flower, A.P. Lewis, J.E. Mabey, P. Scordis, J.N. Selley, W. Wright, PRINTS-S: the database formerly known as PRINTS, Nucleic Acids Res. 28 (2000) 225–227.

[35] J.G. Henikoff, E.A. Greene, S. Pietrokovski, S. Henikoff, Increased coverage of protein families with the blocks database servers, Nucleic Acids Res. 28 (2000) 228–230.

[36] A. Bateman, E. Birney, R. Durbin, S.R. Eddy, K.L. Howe, E.L. Sonnhammer, The Pfam protein families database, Nucleic Acids Res. 28 (2000) 263–266.

[37] F. Corpet, F. Servant, J. Gouzy, D. Kahn, ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons, Nucleic Acids Res. 28 (2000) 267–269.

[38] E.M. Stone, A.J. Lotery, F.L. Munier, E. Heon, B. Piguet, R.H. Guymer, K. Vandenburgh, P. Cousin, D. Nishimura, R.E. Swiderski, G. Silvestri, D.A. Mackey, G.S. Hageman, A.C. Bird, V.C. Sheffield, D.F. Schorderet, A single EFEMP1 mutation associated with both Malattia Leventinese and Doyne honeycomb retinal dystrophy, Nat. Genet. 22 (1999) 199–202.

[39] A. Gal, Y. Li, D.A. Thompson, J. Weir, U. Orth, S.G. Jacobson, E. Apfelstedt-Sylla, D. Vollrath, Mutations in MERTK, the human orthologue of the RCS rat retinal dystrophy gene, cause retinitis pigmentosa, Nat. Genet. 26 (2000) 270–271.

[40] A.B. McKie, J.C. McHale, T.J. Keen, E.E. Tarttelin, R. Goliath, J.J. van Lith-Verhoeven, J. Greenberg, R.S. Ramesar, C.B. Hoyng, F.P. Cremers, D.A. Mackey, S.S. Bhattacharya, A.C. Bird, A.F. Markham, C.F. Inglehearn, Mutations in the pre-mRNA splicing factor gene PRPC8 in autosomal dominant retinitis pigmentosa (RP13), Hum. Mol. Genet. 10 (2001) 1555–1562.